A Hybrid Approach to Web Forum Interactional Coherence Analysis

Tianjun Fu, Ahmed Abbasi, and Hsinchun Chen
Artificial Intelligence Lab, Department of Management Information Systems, The University of Arizona, Tucson, AZ 85721. E-mail: futj@email.arizona.edu, aabbasi@email.arizona.edu, hchen@eller.arizona.edu

Despite the rapid growth of text-based computer-mediated communication (CMC), its limitations have rendered the media highly incoherent. This poses problems for content analysis of online discourse archives. Interactional coherence analysis (ICA) attempts to accurately identify and construct CMC interaction networks. In this study, we propose the Hybrid Interactional Coherence (HIC) algorithm for identification of web forum interaction. HIC utilizes a bevy of system and linguistic features, including message header information, quotations, direct address, and lexical relations. Furthermore, several similarity-based methods including a Lexical Match Algorithm (LMA) and a sliding window method are utilized to account for interactional idiosyncrasies. Experiments results on two web forums revealed that the proposed HIC algorithm significantly outperformed comparison techniques in terms of precision, recall, and F-measure at both the forum and thread levels. Additionally, an example was used to illustrate how the improved ICA results can facilitate enhanced social network and role analysis capabilities.

Introduction

Computer-Mediated Communication (CMC) is any form of communication between two or more individuals who interact and influence each other via computer-supported media. Text-based modes of CMC include e-mail, listservs, forums, chatrooms, instant messaging, and the World Wide Web (Herring, 2002). There is no doubt that the popularity of CMC is continuing to grow. E-mail, Web forums, newsgroups, and chatrooms have already become essential parts of our daily lives, providing a communication medium for various activities (Meho, 2006; Radford, 2006). Although the ubiquitous nature of CMC provides a convenient mechanism for communication, it is not without its shortcomings. The fragmented, ungrammatical, and interactionally disjointed nature of CMC discourse, attributable to the limitations of the CMC media, has rendered CMC highly incoherent (Hale, 1996).

Beaugrande and Dressler (1996) defined coherence in linguistics as a "continuity of senses" and "the mutual access and relevance within a configuration of concepts and relations." For Web discourse, coherence defines the macro-level semantic structure (Barzilay & Elhadad, 1997). Barzilay and Elhadad further pointed out that "coherence is represented in terms of coherence relations between text segments, such as elaboration, cause and explanation." Coherence of online discourse, correspondingly, is represented in terms of the reply-to relations between CMC messages. The reply-to relationships can serve several functions, such as elaborating or complementing previous postings, greeting fellow users, answering questions, or oppugning previous messages.

Computer-Mediated Interaction (CMI) refers to the social interaction between CMC users (Walther, Anderson, & Park, 1994). Such social interaction is built through the reply-to relationships between messages. Therefore, we also refer to the reply-to relationship as the interaction relationship between messages. A social interaction in online discourse happens if a user posts a message that has a reply-to relation with other users' messages. Occasionally, a user may interact with other users without specifying the messages he or she responds to. Common greeting messages like "Hi Jatt" are examples. But we can build fake reply-to relationships between such messages with the addressed user's nearest message. This method does not affect the social interaction relationships between the users.

Since the reply-to relations between CMC messages can be used to build the social interaction between users, coherence of CMC is also called CMC interactional coherence in previous studies (e.g., Herring, 1999). However, current CMC media suffer the "disrupted turn adjacency" problem and the existed system functionalities do not contain sufficient reply-to information. In light of the incoherent and fragmented nature of text-based Web discourse, many researchers have pointed out the importance of automatically identifying CMC interactional coherence. Te'eni (2001) claimed that interactional coherence information is particularly important "when

Received April 26, 2007; revised July 6, 2007; accepted November 12, 2007

 ● Published online 24 March 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20827

there are several participants" and "when there are several streams of conversation and each stream must be associated with its particular feedback." Users of CMC systems cannot safely assume that they will receive a response to their previous message because of the lack of interactional coherence (Herring, 1999). Accurate interaction information is also important to researchers for a plethora of reasons. User interaction in text-based CMC represents one of the fundamental building block metrics for analyzing cyber communities. Interaction-related attributes help identify CMC user roles and user's social and informational value, as well as the social network structure of online communities (Smith & Fiore, 2001; Fiore, Tiernan, & Smith, 2002; Barcellini, Detienne, Burkhardt, & Sack, 2005). Moreover, interactional coherence is invaluable for understanding knowledge flow in electronic communities and networks of practice (Osterlund & Carlile, 2005; Wasko & Faraj, 2005).

Interactional coherence analysis (ICA) attempts to accurately identify the reply-to relationships between CMC messages so that we can reconstruct CMC interactional coherence and present the social interaction between CMC users. Previously used ICA features include system generated attributes such as quotations and message headers as well as linguistic features such as repetition of keywords across postings (Sack, 2000; Spiegel, 2001; Yee, 2002). Although considerable efforts have been devoted to improving interaction representations using ICA, previous studies suffer from several limitations. Most used a couple of specific features, whereas effective capture of interaction cues entails the use of a larger set of system and linguistic attributes (Nash, 2005). Furthermore, the techniques incorporated often ignored noise issues such as typos, misspellings, nicknames, etc., which are prevalent in CMC (Nasukawa & Nagano, 2001). In addition, there has been little emphasis on Web forums, a major form of asynchronous online discourse. Previous work has focused on e-mail-based newsgroups and chatrooms. Web forums differ from e-mail and synchronous forms of electronic communication in terms of the types of salient coherence cues, user behavior, and communication dynamics (Hayne, Pollard, & Rice, 2003).

In this study, we propose the Hybrid Interactional Coherence (HIC) algorithm for Web forum interactional coherence analysis. HIC attempts to address the limitations of previous studies by utilizing a holistic feature set which is composed of both linguistic coherence attributes and CMC system features. The HIC algorithm incorporates finite state automation, where each stage captures interaction based on different feature types, for improved performance. The technique utilizes several similarity-based methods such as a sliding window algorithm and a Lexical Match Algorithm (LMA) in order to identify interaction based on message content cues irrespective of the various facets of CMC noise (e.g., incorrect system feature usage, misspellings, typos, nickname usage). Collectively, HIC's ability to consider a larger set of diverse coherence features while also accounting for noise elements allows an improved representation of CMC interaction.

The remainder of this article is organized as follows. The Related Work section presents a review of previous interactional coherence analysis (ICA) research. The Research Gaps and Questions section highlights important research gaps and questions. The System Design: Hybrid Interactional Coherence System section presents a system design geared towards addressing the research questions, including the use of the HIC algorithm with an extended set of system and linguistic features. It also provides details of the various components of our HIC algorithm. Experimental results based on evaluations of the HIC algorithm in comparison with previous techniques are described in the Evaluation section. The Conclusion section concludes with closing discussions and future directions.

Related Work

CMC interactional coherence is crucial for both researchers and CMC users. Interaction information can be used to identify user roles, messages' values, as well as the social network pertaining to an online discussion. Example applications that can benefit from accurate online discourse interaction information include analyzing the effectiveness of e-mail-based interviewing (Meho, 2006) and chat-based virtual reference services (Radford, 2006). Interactional coherence analysis provides users and researchers a better understanding of specific online discourse patterns. Unfortunately, deriving interaction information from online discourse can be problematic, as discussed below.

Obstacles to CMC Interactional Coherence

Two properties of the CMC medium are often cited as obstacles to CMC interactional coherence (Herring, 1999): lack of simultaneous feedback and disrupted turn adjacency. Most CMC media are text-based so they lack audio or visual cues prevalent in other communication mediums. Furthermore, text-based messages are sent in their entirety without any overlap. These two characteristics result in a lack of simultaneous feedback. However, advanced CMC media have already provided simple solutions to address this concern. For example, newer versions of instant messaging software include audio and video capabilities in addition to the standard text functionality. These tools also show whether a user is typing a response, thereby providing response cues allowing interaction in a manner more similar to face-to-face communication. Since those solutions perform quite well, lack of simultaneous feedback is no longer a severe problem for CMC interactional coherence.

In contrast, resolving the disrupted turn adjacency problem remains an arduous yet vital endeavor. Disrupted turn adjacency refers to the fact that messages in CMC are often not adjacent to the postings to which they are responding. Disrupted adjacency stems from the fact that CMC is "turn-based." As a result, the conversational structure is fragmented, that is, a message may be separated both in time and place from the message it responds to (Herring, 1999). Both synchronous

DOI: 10.1002/asi

User	Feature	Message
Ashna	Direct Address	Hi jatt
Dave-G	Direct Address	Kally I was only joking around
Jatt	Direct Address	Ashna: hello?
Kally	Substitution	I don't think so.
Ashna	Direct Address & Co-reference	How are u jatt
LUCKMAN	N/A	SSa all
Dave-G	Co-reference & Conjunction	Therefore we need to talk
Jatt	Lexical relation & Co-reference	Do we know each other? I'm ok how are you

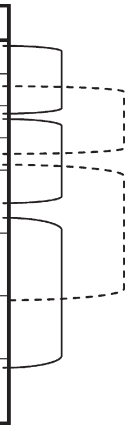

FIG. 1. Example of disrupted adjacency.

(e.g., chatrooms, instant messaging) and asynchronous (e.g., e-mail, forums) forms of CMC suffer from disrupted turn adjacency. Several previous studies have observed and analyzed this phenomenon. Herring and Nix (1997) found that nearly half (47%) of all turns were "off-topic" in relation to the previous turn. Recently, Nash (2005) manually analyzed data from an online chat room and found that the gap between a message and its response can be as many as 100 turns.

Figure 1 shows an example of disrupted adjacency taken from Paolillo (2006). The disruption is obvious in the example and is attributable to the fact that two discussions are intertwined in a single thread. The lines to the right hand side indicate the interaction relations amongst postings: two different widths are used to differentiate the parallel discussions. There is also one message that is not related to any of the other messages, posted by the user "LUCKMAN." The middle column lists the linguistic features used in these messages, which will be introduced in CMC System Features section.

The objective of ICA is to develop techniques to construct the interaction relations such as those shown in the right hand side of the example. Such message interaction relations can be further used to construct the social network structure of CMC users, leading to a better understanding of CMC and its users and providing necessary information for improving ICA accuracy. A review of previous interactional coherence analysis research is presented in the following section.

CMC Interactional Coherence Analysis

Common interactional coherence research characteristics include domains, features, noise issues, and techniques. Table 1 presents a taxonomy of these vital CMC interactional coherence analysis characteristics. Table 2 shows previous CMC interactional coherence studies based on the proposed taxonomy. Header information and quotations (F1 and F2) are system features, whereas features three to six (F3-F6) are linguistic features. A dashed line is used to distinguish these feature categories. The taxonomy and related studies are discussed in detail below.

CMC interactional coherence domains. CMC interactional coherence research has been conducted on both synchronous and asynchronous CMC since both of these modes show a high degree of disrupted turn adjacency (Herring 1999). Synchronous CMC, which includes all forms of persistent conversation, suffers from multiple, intertwined topics of conversation (Khan, Fisher, Shuler, Wu, & Pottenger, 2002). In comparison, asynchronous CMC has a "thread" function, which is an effective method for categorizing forum postings based on a specific topic. However, the thread function is not perfect. First, it does not show message-level interactions, which are vital for constructing the social network structure of CMC users. Instead, it is just an effort to group related messages together. Second, even in a single thread, subtopics might be generated during the discussion. This phenomenon, which poses severe problems for Web forum information retrieval and content analysis, is called "topic decay/drift" (Herring, 1999; Smith & Fiore, 2001). Therefore, it is still necessary and important to apply interactional coherence analysis to asynchronous CMC.

Asynchronous CMC modes can be classified into two categories: SMTP-based and HTTP-based. SMTP-based modes (e.g., Usenet) use e-mail to post messages to forums, whereas HTTP-based methods use forms embedded in the Web pages. Previous research often focused on SMTP-based modes because the headers of posted messages contain what is referred to as "reply-to information" that specifically mentions the ID of the message being responded to. Loom (Donath, Karahalio, & Viegas, 1999), Conversation Map (Sack, 2000), and Netscan (Smith & Fiore, 2001) are all well-known tools that have been developed to show interaction networks of Usenet Newsgroups (SMTP-based). In contrast, HTTP-based modes such as Web forums and blogs do not contain such useful header information for constructing interaction networks. Consequently, there has been little work on HTTP-based CMC as illustrated by Table 2.

We also incorporate text documents into our taxonomy because they experience some problems similar to CMC incoherence, such as co-reference resolution (Bagga & Baldwin, 1998; Soon, Ng, & Lim, 2001) and text segmentation

DOI: 10.1002/asi

TABLE 1. A taxonomy of CMC interaction coherence research.

Category	Description	Label
Domain		
Synchronous CMC	Internet Relay Chat (IRC), MUD, IM, etc.	D1
SMTP -based Asynchronous CMC	Email, Newsgroups	D2
HTTP -based Asynchronous CMC	Web Forums/BBS, Web Blogs	D3
Text document	News, articles, text files, etc.	D4
Feature		
Header information	"Reply -to" information in the header or title	F1
Quotation	Copy previous related message in one's response	F2
Co- reference	Personal, demonstrative, comparative co -reference	F3
Lexical Relation	Repetition, synonymy, superordinate	F4
Direct Address	Mention username of respondent	F5
Other linguistic features	Substitution, ellipsis, conjunction	F6
Noise		
Typo, misspellings, nicknames, modified quotations		
Technique		
Manual	Manually identify the interaction	T1
Link -based method	Link messages by using CMC system features only	T2
Similarity -based method	Word match, VSM, SVM, lexical chain	T3

TABLE 2. Previous CMC interaction coherence studies.

		Features							
Previous Studies	Domains	F1	F2	F3	F4	F5	F6	Noise	Techniques
Xiong et al., 1998	SMTP-based	√						No	Link-based
Bagga et al., 1998	Text document			√				No	Similarity-based
Choi, 2000	Text document				√			No	Similarity-based
Smith et al., 2001	SMTP-based	√						No	Link-based
Sack, 2000	SMTP-based	√	√					No	Link-based
Spiegel, 2001	Synchronous				√	√		No	Similarity-based
Soon et al., 2001	Text document			√				No	Similarity-based
Newman, 2002	SMTP-based	√						Yes	Link-based
Yee, 2002	SMTP-based	√	√					No	Link-based
Barcellini et al., 2005	SMTP-based		√					—	Manual
Nash, 2005	Synchronous			√	√	√	√	—	Manual

(Choi, 2000). Techniques used for text document co-resolution, such as sliding windows (Hearst, 1994), lexical chains (Morris, 1988), and entity repetition (Kan, Klavans, & Mckeown, 1998) are applicable to all forms of text and can provide utility for CMC interactional coherence research.

CMC interactional coherence research features Two categories of features have been used by previous CMC researchers and system developers. The first category is system features, which are functionalities provided by the CMC systems. The second one is linguistic features, which are interpersonal language cues.

Nash (2005) defined explicit features as those that "make fewer assumptions about what information is activated for the recipients." Figure 2 shows features' relative explicit/implicit properties. Features on the left side are more explicit than those on the right side. Explicit features are generally easier to use for deriving interaction patterns. In contrast, implicit features such as conjunctions and ellipsis are far more difficult to accurately incorporate for interactional coherence analysis. The various features are described in detail in the next section.

CMC system features. CMC system features are usually only provided by asynchronous CMC systems. Header information and quotations are two kinds of CMC system features that can be used to construct interaction networks of asynchronous online discourse. Lewis and Knowles (1997) pointed out that SMTP-based asynchronous CMC systems will "automatically insert into a reply message two kinds of header information: unique message IDs of parent messages and a subject line of the parent (copied to the reply message's subject line)." Unique message IDs of the parent message are intuitively useful for interaction identification. In contrast, subject lines of messages are less useful because different conversations in the same thread may have similar subject lines. Unfortunately, for HTTP-based modes, only the second

DOI: 10.1002/asi

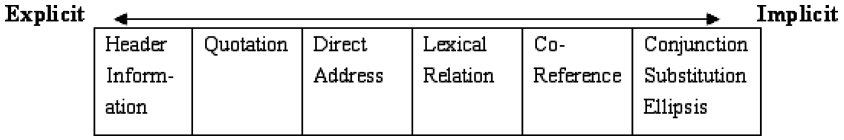

FIG. 2. Features' relative explicit/implicit properties.

type of header information is available. As shown in Table 2, most previous studies for SMTP-based asynchronous CMC systems relied on header information (F1 column) to construct interaction networks (e.g., Sack, 2000; Barcellini et al., 2005).

Quotations (F2 column), a context-preserving mechanism used in online discussions (Eklundh, 1994), are less frequently used to represent online conversations. Conversation Map (Sack, 2000) and Zest (Yee, 2002) are among the few previous studies that used automatic quotation identification to address disrupted adjacency. Barcellini et al. (2005) manually analyzed quotations and used them to identify participants' conversation roles.

Although header information and quotations are effective for identifying interaction and should result in high precision intuitively, in reality they suffer several drawbacks. From the systems' point of view, only asynchronous CMC systems contain such features. Moreover, header information provided by HTTP-based asynchronous CMC systems is of little value in many cases where the subject lines of all subsequent messages are similar or even identical. Furthermore, from the users' point of view, some participants do not use system features and others may not use system functions correctly (Lewis & Knowles, 1997; Eklundh & Rodriguez, 2004). For instance, interaction cues may appear in the message body. Finally, some messages can interact with multiple previous messages and system features may not be able to capture such multiple interactions. As a result, using system features alone fails to consider such idiosyncratic user behavior, resulting in an incomplete representation of CMC interaction.

As is shown in Table 2, previous research on SMTP-based asynchronous CMC relied mostly on system features to construct the interaction network. CMC systems incorporating system and linguistic features for identification of interaction patterns, such as the Conversation Map system proposed by Sack (2000), are a rarity. The Conversation Map system also constructs interaction networks primarily using system features but then uses the message content to construct semantic networks, which display the discussion themes for interacting messages (Sack, 2000).

The content of messages, which can be represented by various linguistic features, may be useful to complement system features in constructing CMC interactions and in many cases may be even more important (Nash, 2005). Therefore, our approach utilizes both CMC system and linguistic features to construct the interaction network with the intention of creating a more accurate representation of CMC interactional coherence and its social network structure. Important linguistic features are discussed in the following section.

Linguistic features. Linguistic features are interpersonal language cues and content-based features. Previous research on synchronous CMC systems had to rely on linguistic features to construct interaction networks since no system features were available. Several linguistic features for online communication have been identified by previous research. Three prevalent features are direct address, lexical relations, and co-reference (Halliday & Hasan, 1976; Herring, 1999; Spiegel, 2001; Nash, 2005).

Direct address takes place when a user mentions the username of another user whom he or she is addressing in the message. Coterie (Spiegel, 2001), a visualization tool for conversation within Internet Relay Chat, looks for direct addresses of specific people to construct the interaction network. It is important to note that addressing someone is different from referencing someone. Take the following sentence as an example: "John, take care of your brother Tom." The speaker is addressing (and interacting) with "John" only, although "Tom" is also referenced.

Lexical relations occur when a lexical item refers to another lexical item by having common meanings or word stems. Its most common forms are repetition and synonymy (Nash, 2005). Lexical relations have also been widely used in previous studies of synchronous CMC systems. For example, Choi (2000) used repetition of keywords to identify relationships between messages. Techniques that compare text similarities are often used for identifying lexical relations, where two messages are considered to have an interaction if their similarity is above some predefined threshold (Bagga & Baldwin, 1998).

Co-reference also occurs when a lexical item refers to another one; however, such a relationship can only be identified by the context instead of the word meanings or stems. Personal co-reference is most commonly used in CMC. For example, the word "you" is frequently used to refer to the person a message addresses. Other co-references include demonstrative co-reference, which is made on the basis of proximity, and comparative co-reference, which uses words such as "same," "similar," and "different" (Nash 2005).

Some other linguistic features identified by previous studies include: conjunctions (e.g., but, however, therefore), substitution (e.g., "I think so."), ellipsis (e.g., "Guess that would not be easy."), etc. (Nash, 2005). These features have rarely been incorporated in previous studies due to the difficulty in identifying such features and their lack of prevalence in online discourse. Figure 1 shows an example that includes most linguistic features mentioned here.

Looking back to Table 2, we can see that most previous studies only utilized one or two specific features. Only Nash

DOI: 10.1002/asi

(2005) manually identified multiple linguistic features for an online chatroom and found three of them to be dominant. Lexical relations covered 51% of the interaction pattern, whereas direct address and co-reference covered 28% and 15%, respectively.

Noise issues in ICA. In ICA, noise can be defined as obstacles to direct or exact match of various features. Noise can have a profound impact on the performance of automated approaches for identifying interaction patterns. It is highly prevalent in free text, diminishing feature extraction capabilities for text mining (Nasukawa & Nagano, 2001). Typos and misspellings are common types of noise for online discourse and they exist in both direct address and lexical relations. There are also some specific forms of noise for various features, which are discussed below.

In direct address, Nash (2005) pointed out that many CMC users use nicknames to address each other (e.g., "Martin" for user "MartinHilpert," "binary" for user "binarymike"). In addition, some usernames or their nicknames are common words; hence, we need to differentiate common usage of such words with their usage as a username. For example, the word "endless" can be used to mention user "EndlessEurope." However, "endless" might also be a common adjective in some messages. Consequently, simply comparing each word with all the usernames will not identify all the direct addresses.

In lexical relations, repetition of keywords has been used in previous research (Choi, 2000; Spiegel, 2001); but, morphological word changes often decrease its performance. Word stem repetition, an improved method, can be used to solve this problem (Reynar, 1994; Ponte & Croft, 1997). However, it still cannot alleviate the effect of typos and misspellings.

Even in quotations, which are generated by the system automatically, noise still exists. Newman (2002) noticed that sometimes there were differences between the line partitions in original messages as compared to the quoted versions. Moreover, users often engage in "partial quotation" where a specific portion or segment of the original message is quoted in the reply (Eklundh, 1998).

As is shown in Table 2, Newman's study (2002) is one of the few which addressed noise-related issues. He matched quotations based on sentences or sentence parts instead of matching them as a whole in order to compensate for partial quotation. In contrast, other studies failed to compensate for the existence of noise in CMC postings.

CMC interactional coherence analysis techniques. In light of the fact that several types of features can be used for interactional coherence analysis, many different techniques have previously been used to construct interaction patterns. These can be classified into three major categories: manual analysis, link-based techniques, and similarity-based techniques.

Eklundh and Rodriguez (2004) manually identified lexical relations, direct address, and co-reference for one specific online discussion. Similarly, Nash (2005) identified and extracted six linguistic features for an English chatroom. Barcellini et al. (2005) manually analyzed quotations and used them to identify participants' conversation roles. Manual analysis of CMC interactional coherence has the obvious advantage of accuracy. However, its disadvantage is also obvious: It is difficult to apply to large date sets and is labor intensive.

Link-based techniques construct interaction patterns using system features or rules based on message sequences. These techniques are highly prevalent in previous research because of their representational simplicity as compared to techniques that focus on linguistic features. Direct linkage techniques link messages based on header information and quotations. For residual messages unidentified by direct linkage, naïve linkage (Comer & Peterson, 1986) has been used. Naïve linkage is a rule-based technique that proposes that a message is related to all prior messages in the same discussion or the first message in the same discussion. The advantage of link-based techniques is that they are easy to implement. However, link-based techniques depend on the assumption that CMC users utilize system features correctly. Moreover, naïve linkage is of low accuracy and often overgeneralizes participation patterns due to its simplistic rule-based properties.

Similarity-based techniques typically use content similarity to construct interaction patterns. These techniques focus on uncovering interaction cues found in the message texts to provide insight into interactional coherence. The simplest method is exact match or direct match, which tries to identify repetition of words, word phrases, or even sentences (Choi, 2000; Spiegel, 2001). More advanced similarity-based techniques include Vector Space Model, which has been used for the cross-document co-reference solution (Bagga & Baldwin, 1998) as well as to identify quoted messages (Lewis & Knowles, 1997), and lexical chains, which are often created using WordNet for text summarization and interaction identification (Barzilay & Elhadad, 1997; Sack, 2000). Similarity-based techniques are effective for identifying certain linguistic features (e.g., lexical relations and direct address). Some have been successfully applied in research related to text documents. However, similarity-based techniques are susceptible to noise and require careful selection of parameters.

Research Gaps and Questions. Based on our review of previous literature, we have identified several important research gaps. First, little interactional coherence analysis has been conducted for HTTP-based asynchronous CMC. Previous research focused on USENET newsgroups and e-mail, the headers of which contain accurate interaction information, rendering the use of system features sufficient for accurately capturing a large proportion of the interaction patterns. However, many Web site and ISP forums (e.g., Yahoo, MSN) do not use the e-mail protocol. Relying only on system features for such CMC modes can result in a

DOI: 10.1002/asi

significant amount of neglected interaction information. Second, little previous research has implemented techniques that use both CMC system features and linguistic attributes for interactional coherence analysis. The use of a more holistic feature set comprised of features occurring in messages headers and bodies could greatly improve interaction recall. Finally, there has been little emphasis in previous research that takes into account the impact of noise in CMC interaction networks.

Based on the research gaps identified, we propose the following research questions:

1. How effectively can we analyze interactional coherence for HTTP-based Web forums using automated techniques?
2. How can techniques that use both CMC system and linguistic features improve interaction representational accuracy as compared to methods that only utilize a single feature category?
3. What impact do forum dynamics (i.e., users system usage behavior) exert on interaction representational accuracy?
4. How does noise affect the accuracy of automatically constructed CMC interaction networks?

System Design: Hybrid Interactional Coherence System

In order to address these research questions, we developed the Hybrid Interactional Coherence (HIC) algorithm to perform more accurate interactional coherence analysis, that is, to identify the reply-to relationships between CMC messages. The algorithm has three major components: system feature match, linguistic feature match, and residual match. System feature match and the direct address part of the linguistic feature matching component are used to identify interactions stemming from relatively more explicit features (such as headers, quotations, and direct addresses). The lexical relation match and rule-based module (which derive interaction patterns from relatively implicit cues), are only utilized when more explicit features are not present in a posting.

Several major types of noise have also been addressed. System features used in our implementation include both headers and quotations. With header information, unique IDs of parent messages are checked first. Message subject lines are also analyzed and used. With quotations, our algorithm can identify not only normal quotations but also two special types of quotation, that is, multiple quotations and nested quotation (Barcellini et al., 2005). The algorithm overcomes quotation noise by using a sliding window method, which compares part of the quotation to previous messages. The sliding window method has been successfully used in text similarity detection and authorship discrimination (Nahnsen, Uzuner, & Katz, 2005; Abbasi & Chen, 2006). Compared with the sentence-level matching approach adopted by Newman (2002), the sliding window is better at dealing with quotation modifications made by systems or users because it is a character-level method (i.e., it compares substrings).

With respect to linguistic features, our algorithm mainly uses direct address and lexical relations. For direct address, besides traditional simple name match, our algorithm uses Dice's equation to overcome noise such as typos, misspellings, and nicknames. Dice's equation uses character-level n-gram matching to identify semantically related pairs of words (Adamson & Boreham, 1974). We also differentiate common words and usernames by using a lexical database and automatically generated part-of-speech (POS) tags. For lexical relations, a Lexical Match Algorithm (LMA), developed based on the Vector Space Model, which is frequently used in information retrieval (Salton & McGill, 1986), is adopted.

A comprehensive residual matching mechanism is developed for the remaining messages. It improves the naïve linkage method (Comer & Peterson, 1986) by matching messages based on their context and co-reference features. Figure 3 shows our system design. Details of each component are presented below.

Data Preparation

The data preparation component is designed to extract messages and their associated meta data from Web forums. All relevant header information is extracted first. Then each message's quotation part and body text are separated using a parser program. The parser program was also designed to deal with two special types of quotation, nested quotation and multiple quotations. Nested quotation happens when a message which already contains quotations is quoted. The parser program only parses the quotation that is nearest to the message. Sometimes users respond to different quotations in one message, which is referred to as "multiple quotation." The parser program parses all the related quotations. The final step of data preparation is to extract other relevant information from each message, such as author screen names, date stamps, message subjects, etc.

HIC Algorithm: System Feature Match

Header information match. In header information match, unique message IDs of parent messages, if available, are used to identify interaction. Subject lines of messages in the same thread are often consistently similar with each other if they are automatically generated by CMC systems. However, if CMC users intentionally embed interaction cues within them, subject lines can be used to identify interaction patterns as well.

Quotation match. In quotation match, the quotation part of each message is compared with the body text of previous messages. As previously mentioned, CMC systems may modify the format of quotations (Newman, 2002), whereas CMC users may modify quotations to save space (Eklundh, 1998). Therefore, in our system the quotation part of each message is first searched for in the body text of all previous messages, referred to as "simple match." If simple match fails due to the various aforementioned forms of noise, a sliding window method is triggered.

A sliding window method breaks up a text into overlapping windows (substrings) and compares each window against

DOI: 10.1002/asi

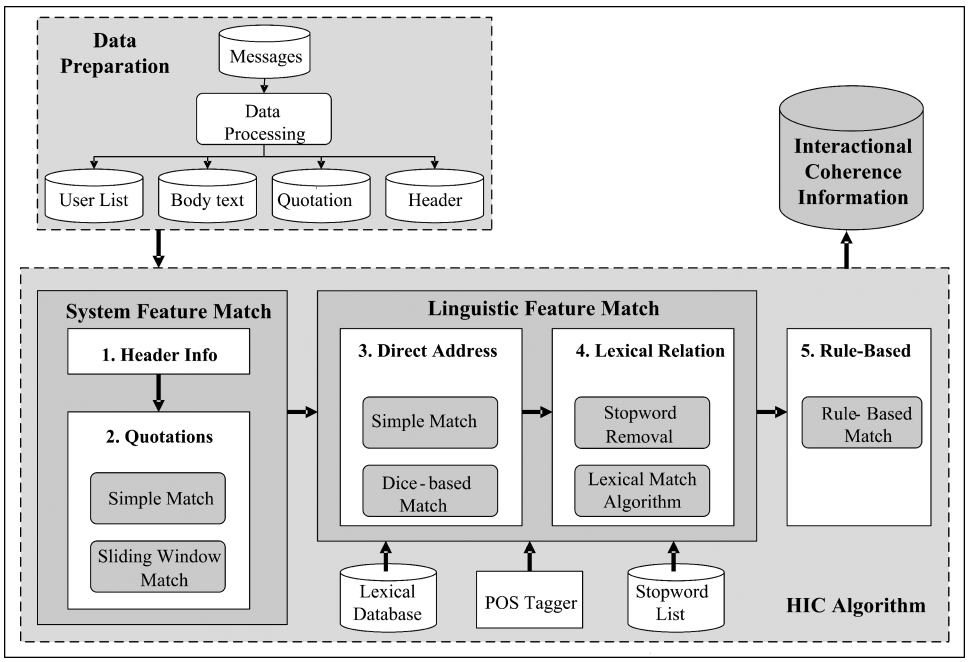

FIG. 3. HIC system design.

previous body texts (Kjell, Woods, & Frieder, 1994; Nahnsen et al., 2005; Abbasi & Chen, 2006). The system assigns the message (i.e., creates an interaction link) to the quoted message with the highest number of matching windows. The following example shows how a sliding window method with a window size of 10 characters and a jump interval of 2 characters can be used to identify modified quotations.

Original Message	Quoted Content	Message Text Windows	Quoted Text Windows
"What do you prefer?"	". . . do you prefer?"	"What do yo" "at do you" "do you pr?" **"o you pref"** **"you prefer"**	". . . do you" ".do you pr" **"o you pref"** **"you prefer"**

HIC Algorithm: Linguistic Feature Match

Linguistic features are used to complement system features in constructing CMC interaction patterns. Nash (2005) found that direct address, lexical relations, and co-reference were three dominant linguistic features. Therefore, our hybrid interactional coherence algorithm mainly uses direct address and lexical relations in linguistic feature match, whereas the co-reference feature is indirectly used in residual match.

Direct address match. In direct address match, each word of a message is compared to the screen names of previous messages' authors. By only considering authors that have appeared in prior postings within the same thread, we reduce the possibility of incorrectly considering username references to be direct addresses. For the previous example "John, take care of your brother Tom," if user "Tom" has not already appeared in the thread, an interaction between the current message's user and Tom will not be assigned. In situations where a direct address based interaction is found, the message containing the interaction cue is assumed to have a reply-to relation with the addressed users' most recent posting. Initially a simple match is performed in order to detect messages containing the exact same author screen names. If no simple matches are found, a Dice-based character-level n-gram matching technique is used to compensate for the effect of prevalent direct address noise in CMC such as typos, misspellings, and nicknames. The technique first uses the following Dice equation, which has been successfully used in identifying semantically related pairs of words (Adamson & Boreham, 1974; De Roeck & Al-Fares, 2000), to estimate the similarity between a word and an author's screen name:

$$\text{Dice Score} = \frac{2 \times (\textit{number of shared unique } n-\textit{grams})}{\textit{Total unique } n-\textit{grams}}$$

A pre-established experiment-based threshold is applied to improve the accuracy of direct address match. However, since many CMC users choose common English words as their screen names, word sense disambiguation methods need to be applied to differentiate common usages of a word with the use of a word as a screen name. Our HIC algorithm makes use of WordNet (Miller, 1990), which has already been widely used in word sense identification (Voorhees, 1993; Resnek, 1995), to identify the meaning of words, and a POS tagger (McDonald, Chen, Su, & Marshall, 2004) to

DOI: 10.1002/asi

generate the part-of-speech tags. Details of our direct address match are presented below:

1. For each screen name in the author list, query WordNet for meanings;
2. For each word in a message, do the following:
2.1 Use Dice equation to find the most similar screen name appeared before;
2.2 If the highest Dice score is greater than a predefined threshold, query WordNet for the meanings of the word and do the following:
2.2.1 If neither the word nor the screen name has meanings, assign direct address;
2.2.2 Else, get POS tag for the word. If the word is a noun or noun phrase, assign direct address;
2.2.3 Else, do not assign direct address for the word.

Lexical relations: The lexical match algorithm. Lexical relation match assumes an interaction between the two messages that are most similar. It calculates the lexical similarities among stopword-removed messages when more explicit interactional coherence features such as quotations and direct address are not found. The key to lexical relation match is to develop an appropriate formula to calculate the similarity score. We propose a Lexical Match Algorithm (LMA) for lexical relation match. The lexical matching algorithm (LMA) is designed to identify lexical relation based interactions between postings while taking into consideration the unique characteristics of CMC interaction, such as topic drift/decay and various forms of noise (e.g., misspellings, idiosyncrasies, etc.). The algorithm measures the similarity between messages based on the content as well as turn proximity and levels of inflection and idiosyncratic literary variation. LMA integrates the Vector Space Model with Dice's equation and a turn based proximity scoring mechanism.

Vector Space Model (VSM) is one of the most popular methods used to identify lexical similarities (Salton & McGill, 1986). By using word stems, VSM can also identify morphological word changes. However, in order to identify typos, misspellings, abbreviated references, and other forms of creative user behavior, the Dice equation (Adamson & Boreham, 1974; De Roeck & Al-Fares, 2000) is adopted in LMA to complement the traditional VSM.

Additionally, a high degree of topic decay/drift has been found in asynchronous CMC (Herring, 1999; Smith & Fiore, 2001). Nash (2005) also noticed that most CMC interactions happen within three turns. Therefore, CMC interactions represent a "closeness" characteristic, which means two closer messages are more likely to interact than two messages further away. A topic decay factor calculated by the distance (number of turns) between two messages is adopted in our LMA formula to address this "closeness" characteristic.

Here is our LMA formula for lexical similarity:

$$\left(\sum_{i=0}^{LenX} \sum_{j=0}^{LenY} \frac{Tf_{Xi} + Tf_{Yj}}{Df_{Xi} + Df_{Yj}}_{\,if\,(Dice(Xi,Yj)>0.55)} \times (LenX \times LenY)^{-1} \right) \times (Distance(X,Y) + C)^{-1}$$

X and Y are the two compared messages. LenX and LenY are the number of unique non-stopword terms in the two messages, Xi refers to the i^{th} non-stopword word in message X and Yj the j^{th} non-stopword term in message Y. Tf is the term frequency and Df is the document frequency. Distance(X, Y) refers to the number of turns or messages between two compared messages. If there are N messages between the two compared messages, their distance is N + 1. C is a constant used to control the impact of message proximity on the overall similarity between two messages.

In the formula, Dice(Xi, Yj) is used to compare two non-stopword terms. If their similarity is greater than 0.55, which is a predefined experiment-based threshold, a combined "tf-idf" score is calculated. $(LenX \times LenY)^{-1}$ is the length normalization factor and $(Distance(X, Y) + C)^{-1}$ is the topic decay factor mentioned before. If the highest score calculated by our formula is greater than 0.002, another threshold we use, an interaction is identified. Otherwise, residual match is used. The value of constant C and the two thresholds are developed based on a manually analysis of ten other threads in the LNSG forum. These 10 threads are not included in our evaluation.

HIC Algorithm: Residual Match

Residual match is used for messages which do not contain obvious clues for automatic interaction identification. It is utilized to help enhance interaction recall by assigning interactions based on common communication patterns. Prior residual matches have used variants of the naïve linkage method. One such implementation assigns each remaining posting (i.e., one with no identified interaction) to the first message in the thread (Comer & Peterson, 1986). Other versions of naïve linkage assign each posting to the preceding message. The intuition behind assigning each remaining post to the prior one is that messages are likely to interact with predecessors in close proximity, given the turn-based nature of CMC (Herring, 1999). Since residual matching techniques use very general assignment rules, they tend to have lower precision as compared to other techniques which use system and linguistic interaction cues. We propose a new rule-based residual match method which considers the message proximity as well as the conversation structure and context. The details for our residual match are provided below:

X: the residual message of author A
Y: previous message of author A
Z: messages of other authors which are posted between Y and X and are replies to messages of author A

1. If Y does not exist, X replies to the first message in the discussion;
2. If Y exists and Z exists, X replies to Z;
3. If Y exists and Z does not exist, X replies to what Y replies to.

DOI: 10.1002/asi

The first rule is to apply the improved naïve linkage method when the residual message is the first message the author has posted in the thread. The other two rules are generated based on two human communication characteristics, which can also be found in CMC. If people give feedback or raise questions to our proposed ideas and statements, it is natural for us to comment on the feedback or answer the questions, which is characterized by the second rule. On the other hand, even if no feedback is given, people tend to strengthen or make clear their previous statements, characterized by the third rule.

Evaluation

In order to evaluate the effectiveness of our HIC algorithm, two experiments were conducted. The first experiment compared the HIC algorithm against the link and similarity-based methods. The second experiment assessed the impact of noise compensation on interaction pattern identification performance. The test bed and experimental design are described in detail below.

Test Bed

Our test bed consisted of two Web forums. The first forum was the Sun Java Technology Forum (http://forum.java.sun.com), which is an electronic network of practice. Analysis of such forums can help examine their social capital and knowledge contribution (Wasko & Faraj, 2005). The second one was the Libertarian National Socialist Green Party (LNSG) Forum (http://www.nazi.org/community/forum/). Analysis of such social online communities is important in order to improve our understanding of these groups and organizations (Burris, Smith, & Strahm, 2000; Schafer, 2002; Chen, 2005). Furthermore, these two types of forums were selected because of their contrasting usage mechanisms and user behavior, which can help evaluate the impact of forum dynamics (e.g., user system usage behavior) on interaction patterns. Users of electronic networks of practice, particularly ones pertaining to technology, are likely to be more technically savvy and less interpersonal, whereas those of social forums are more personal and closely affiliated. For both forums, several of the longest threads were studied (shown in Table 3).

The threads in the Sun Java Technology forum were much longer than those of the LNSG forum. All seven threads were manually tagged first by a single annotator to identify their interactional coherence. A sample of one hundred messages from the annotator was also tagged by a second coder to check the accuracy of the tagging. Both independent annotators were graduate students with strong linguistic backgrounds. The annotators determined a correct interaction by looking for interaction cues in every message. The cues included features found in message headers (e.g., an "RE:" in the subject line), quoted content from another message, linguistic cues inherent in the message body (e.g., direct address and lexical relations) as well as those based on the thread context (i.e., residual rule matching based on previous postings and interaction). The annotators utilized the guidelines proposed by Nash (2005) for manually identifying linguistic interaction cues. Figure 1 provided examples of how interactional coherence could be derived using linguistic features. The inter-coder reliability across the one hundred messages had a kappa statistic of 0.88, which is considered to be reliable. The tagging results were used as our gold standard. The interaction feature breakdowns across threads based on the manual tagging are presented in Table 4. The difference in forum dynamics can be clearly seen. Quotations are much more prevalent in the Sun Java Technology Forum, most likely because its users are better at utilizing system functionalities. Moreover, using quotations in long threads helps readers understand the context of each message. In contrast, lexical relation is preferred in the LNSG Forum. Furthermore, the LNSG Forum members use direct address more often. This is likely attributable to the fact that people in such social groups know each other better. Finally, the high percentage of "other" features in the LNSG Forum implies that this forum's users are more likely to display idiosyncratic and/or creative usage of CMC systems.

TABLE 3. Details for data sets in test bed.

Forum	Thread no.	Thread subject	# of Messages	# of users
Sun Java forum	1	Java switch statement	429	31
	2	Double precision catastrophic	403	37
	3	Why use int over double?	453	36
LNSG forum	4	Idea for banner / icon	148	24
	5	Blue eyes, blond hair	62	22
	6	Greetings	85	14
	7	Race mixing	143	39

Experiment 1: Comparison of Techniques

Experiments setup. In the first experiment, we compared our HIC algorithm with a link-based method that relies on system features, as well as against a similarity-based method, which relies on linguistic features. These comparison techniques were incorporated since variations of the link-based method and similarity-based method have been adopted in previous studies (Spiegel, 2001; Soon et al., 2001; Newman, 2002; Yee, 2002). The purpose of this experiment was to study the effectiveness of the combined usage of system features and linguistic features, as done in the proposed HIC algorithm, over techniques mostly utilizing a single category of features.

The link-based method uses the quotations in the header information for interactional coherence identification (Yee, 2002). If a quotation exactly matches previous messages, the interaction is noted between the two postings. For remaining messages, the naïve linkage method is used, which assumes that the remaining messages are replies to the first message.

DOI: 10.1002/asi

TABLE 4. Interaction feature breakdowns across threads.

Forum	Thread no.	# of messages	Quotation	Direct address	Lexical relation	Others
Sun Java forum	1	429	**68.4%**	14.5%	9.1%	8.0%
	2	403	**70.3%**	7.8%	7.6%	14.3%
	3	453	**75.5%**	6.4%	8.0%	10.1%
	Overall	1285	**71.5%**	9.6%	8.3%	10.6%
	4	148	16.2%	16.2%	**41.9%**	25.7%
	5	62	9.7%	9.7%	**53.2%**	27.4%
LNSG forum	6	85	21.2%	24.7%	**35.3%**	18.8%
	7	143	33.6%	8.4%	**33.6%**	24.4%
	Overall	438	21.9%	14.4%	**39.5%**	24.2%

The similarity-based method consists of two parts: simple direct address match and Vector Space Model match (Bagga & Baldwin, 1998). The first part identifies interactional coherence when a word is an exact match with other authors' screen names. The second part uses the traditional "tf-idf" score to identify lexical similarity. Threshold 0.2, shown as the best threshold by Bagga and Baldwin (1998), is used for this traditional VSM match. Precision, recall, and F-measure at both the forum and thread level were used to evaluate the performance of these methods.

$$\text{Precision} = \frac{\textit{Number of Correctly Identified Interactions}}{\textit{Total Number of Identified Interactions}}$$

$$\text{Recall} = \frac{\textit{Number of Correctly Identified Interactions}}{\textit{Total Number of Interactions}}$$

$$\text{F- measure} = \frac{2 \times \textit{precision} \times \textit{recall}}{\textit{precision} + \textit{recall}}$$

Hypotheses. Given the presence of system and linguistic interaction cues in online discourse, we believe that interactional coherence identification techniques incorporating both feature types are likely to provide better performance. Therefore, we propose the following hypotheses:

H1a: The HIC algorithm will outperform the link-based method for Web forum interactional coherence analysis.

H1b: The HIC algorithm will outperform the similarity-based method for Web forum interactional coherence analysis.

Experimental results. Table 5 shows the experimental results for all three methods. Our HIC algorithm had the best performance on both the forums in terms of precision, recall, and F-measure. The linked-based method performed better than the similarity-based method for the Sun Java Technology forum, whereas its performance was worse on the LNSG forum.

TABLE 5. Experimental results for experiment 1.

Forum	Technique	Precision	Recall	F-measure
Sun Java forum	HIC Algorithm	**0.842**	**0.878**	**0.860**
	Link-based	0.793	0.756	0.774
	Similarity-based	0.691	0.719	0.705
LNSG forum	HIC Algorithm	**0.711**	**0.711**	**0.711**
	Link-based	0.560	0.551	0.555
	Similarity-based	0.584	0.678	0.625

Hypotheses results. Table 6 shows the *p*-values for the pair-wise *t*-tests conducted on the interactional coherence identification accuracies to measure the statistical significance of the results. Bolded values indicate statistically significant outcomes in line with our hypotheses. Both hypotheses, H1a and H1b, are supported.

H1a: The HIC algorithm outperformed the link-based method for both the Web forums ($p < 0.01$).

H1b: The HIC algorithm outperformed the similarity-based method for both the Web forums ($p < 0.01$).

Results discussion. The HIC algorithm performed better than both the link-based and similarity-based methods for our test bed. The F-measure was 8%–15% higher than the other two techniques. Such improved performance was consistent across all seven threads in our test bed, as depicted in Figure 4.

The enhanced accuracy of the HIC algorithm was attributable to the incorporation of both system and linguistic features and its ability to handle various forms of CMC noise. The link-based method performed better than the similarity-based method in the Sun Java Technology forum because quotation features were more prevalent in this forum as illustrated in Table 4. For the LNSG forum, lexical relations were more commonly used as interaction cues, resulting in the improved performance of the similarity

TABLE 6. *P*-values for pair-wise t-tests on accuracy for experiment 1.

Forum	Techniques	*P*-values[a]
Sun Java forum	HIC versus link based	**<0.001***
	HIC versus similarity based	**<0.001***
	Link based versus similarity based	**<0.001***
LNSG forum	HIC versus link based	**<0.001***
	HIC versus similarity based	**<0.001***
	Link based versus similarity based	**<0.001***

[a]*P*-values significant at alpha = 0.01

DOI: 10.1002/asi

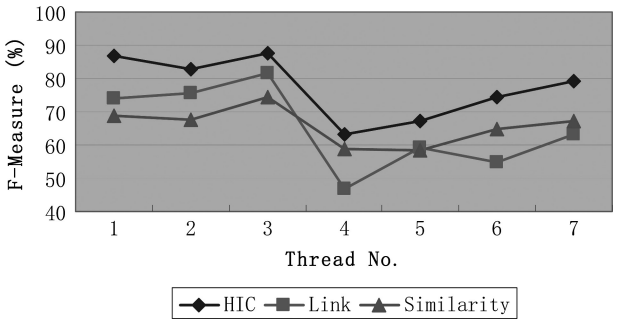

FIG. 4. Experiment 1 F-measure performance for each thread.

method over the link-based method on this forum. The LNSG forum members were less likely to utilize system features, which are heavily relied upon by the link-based method.

Experiment 2: Impact of Noise

Experiment setup. In the second experiment, we evaluated the effectiveness of the noise compensation components in the HIC algorithm. The HIC algorithm was compared against an implementation devoid of any noise compensation components. First, in quotation match, no sliding window was used to identify modified quotations. Second, in direct address match and lexical relation match, Dice's equation wasn't utilized. Thus, only simple direct address match and standard Vector Space Model for lexical relations were incorporated in the "no noise compensation" implementation. Again, precision, recall, and F-measure are used as our evaluation criteria.

Hypothesis. By not considering the noise issues, we suspect some CMC interactions cannot be detected. Since our HIC algorithm utilizes several similarity-based methods that are likely impacted by noise, we propose the following hypothesis:

H2: Addressing noise issues using our proposed HIC algorithm will improve the results of interactional coherence analysis as compared to not accounting for noise issues.

Experimental results. Table 7 shows the experimental results. Our HIC algorithm has better performance on both the forums.

Hypothesis results. Table 8 shows the *p*-values for the pair-wise *t*-tests conducted on the interactional coherence identification accuracies of the two methods. Our hypothesis H2 is supported based on the result. Addressing noise issues using the HIC algorithm improves the results of interactional coherence analysis as compared to not accounting for noise ($p < 0.001$).

Results discussion. The HIC algorithm's F-measure was around 6% higher than that of the implementation with no noise compensation. Figure 5 shows the F-measure performance of the two methods for the seven threads. The HIC algorithm outperformed HIC with no noise compensation in all seven threads. Noise had a slightly larger effect on the LNSG forum than on the Sun Java Forum. A possible explanation is that users of technology forums compose messages more carefully than users in social forums. The Sun Java forum members are computer programmers with greater technical prowess, while the LNSG forum members are more creative in terms of their usage of language and electronic communication media. The experimental results demonstrate the impact of noise on CMC interaction networks as well as the effectiveness of noise compensation components in the HIC algorithm.

TABLE 7. Experimental results for experiment 2.

Forum	Technique	Precision	Recall	F-measure
Sun Java Forum	HIC algorithm	**0.842**	**0.878**	**0.860**
	No noise compensation	0.798	0.807	0.802
LNSG Forum	HIC algorithm	**0.711**	**0.711**	**0.711**
	No noise compensation	0.653	0.640	0.646

TABLE 8. *P*-values for pair-wise t-tests on accuracy for experiment 2.

Forum	Techniques	*P*-values[a]
Sun Java forum	HIC versus no noise compensation	**<0.001***
LNSG forum	HIC versus no noise compensation	**<0.001***

[a]*p*-values significant at alpha = 0.01

Evaluating the Impact of Interaction Representation: An Example

Interaction networks can be used to generate the social network structure of CMC users. Inaccurate or incomplete interaction patterns have an obvious impact on overall network topology, and also on individual node metrics (e.g., degree and centrality). Such incorrect individual node statistics can affect participant role and interaction measures, which are important units of CMC content analysis (Henri, 1992; Rourke, Anderson, Garrison, & Archer, 2001).

In order to illustrate how the HIC algorithm can improve social network analysis metrics as compared to previous techniques, we present an example from the Java forum. A user called "krebsnet" from the Sun Java forum that initiated

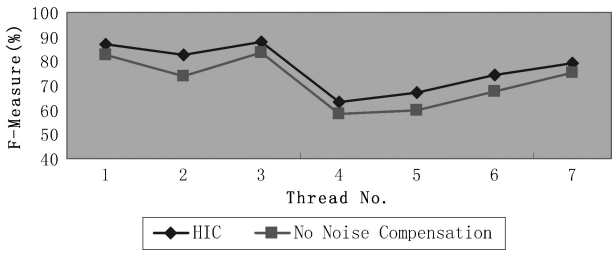

FIG. 5. Experiment 2 F-measure performance for each thread.

DOI: 10.1002/asi

TABLE 9. Degree and centrality measures of user "krebsnet."

Technique	Centrality		
	Betweenness	Closeness	Degree
Actual (manual)	97.072	80.00	10
HIC algorithm	139.079	79.00	14
Linkage	206.377	68.00	25
Similarity match	212.969	64.00	28

thread #1 of our test bed is analyzed. The user's degree and centrality measures generated by the various methods are shown below, in comparison with the values generated based on the manual interaction tagging (which is once again deemed the gold standard).

As shown in Table 9, our HIC algorithm is most reflective of the user's actual involvement in the thread, with a more approximate measurement of centrality and degree. The other techniques exaggerate the user's degree and centrality, which is shown in Figure 6. Based on the thread-level interaction results from the three methods above, the networks shown in Figure 6 were generated using a spring layout algorithm, which places more central nodes near the middle. The circled point represents the user "krebsnet." Figure 6 shows that the linkage and similarity match methods tend to over-assign messages to this initial poster. This is evident based on the spatial location and number of links for "krebsnet" in the linkage and similarity match methods. The social networks generated using the prior methods have a percentage error of over 100% for the betweenness and degree measures for the example node provided. The comparison techniques are off by as much as 180% regarding the node's degree measure. In addition to differences in the absolute metric values, the degree and centrality ranking for the user (relative to other posters in the thread) is also greatly exaggerated by the link and similarity based methods. Both these comparison techniques rank "krebsnet" first in terms of degree, while the user is actually ranked 7th. The HIC algorithm ranks "krebsnet" sixth, closer to the poster's actual level of importance. For the linkage method, the disparity is attributable to the naïve linkage match incorrectly assuming that residual messages are likely to refer to the initial posting. For the similarity match method, the erroneous metric values occur because the initial message/posting contains many important keywords in the thread. The similarity scores for this initial message are consequently

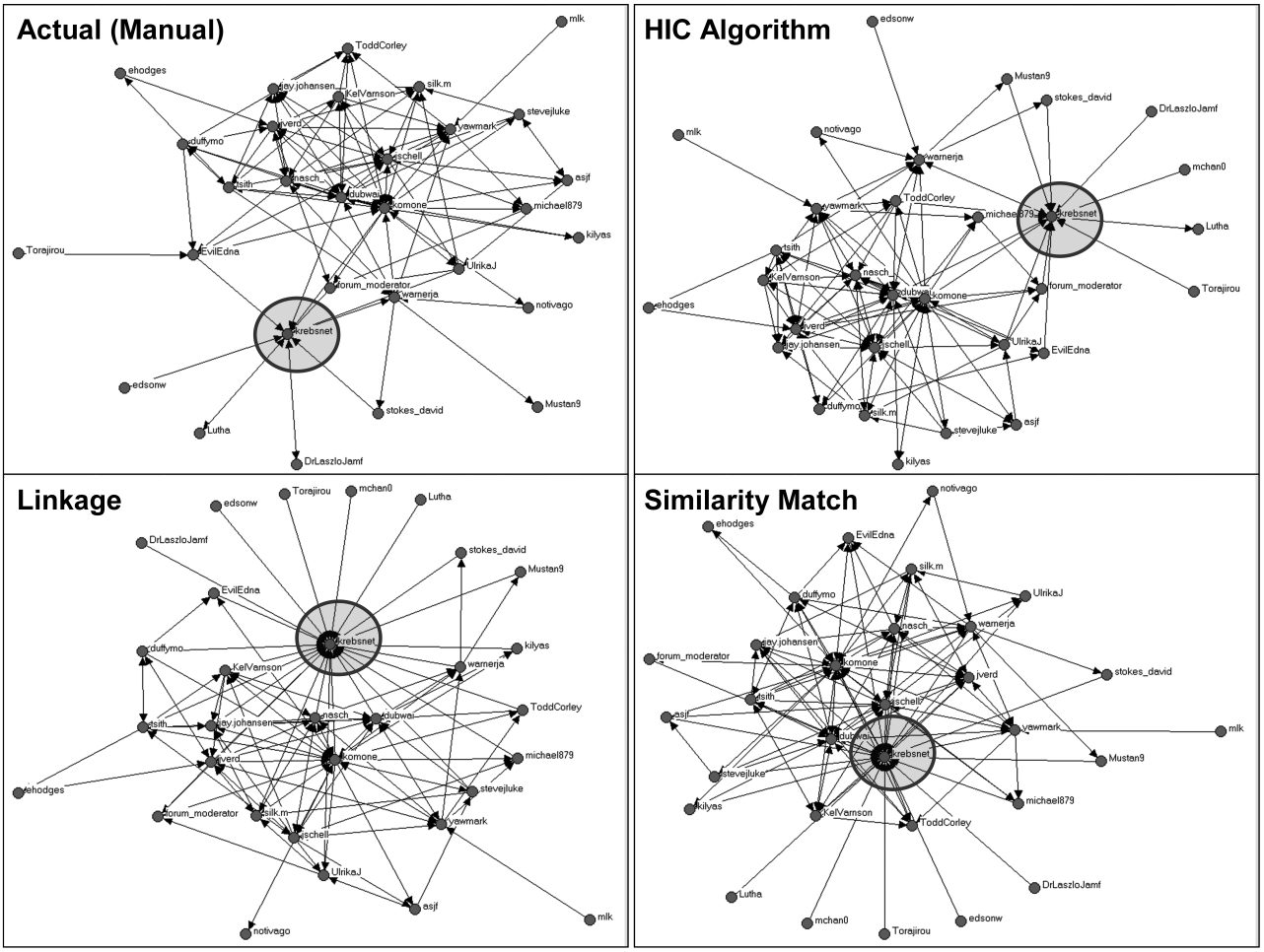

FIG. 6. Social network structure of users in thread #1.

DOI: 10.1002/asi

higher when comparing it against other messages in the thread. This results in a high level of false message assignments. The results suggest that an improved thread-level interaction network will result in a more accurate representation of the social network structure of CMC users, which is important for CMC content analysis.

Conclusion

In this study we applied interactional coherence analysis to Web forums. We developed a hybrid approach that uses both CMC system features and linguistic features for constructing interaction patterns from Web discourse. The results show that our approach outperformed traditional link-based and similarity-based methods due to the use of a robust set of interaction features. Furthermore, the HIC algorithm also incorporates a wide array of techniques to address various types of noise found in CMC. Noise analysis results show that accounting for noise considerably improves performance as compared to methods that do not consider noise. Finally, we show that an improved representation of interaction networks results in a more accurate representation of the social network structure of CMC users. This is especially crucial for effective content analysis of online discourse archives.

In the future, we will work on analyzing user roles in Web forums based on interaction networks generated by the HIC algorithm. We are also interested in identifying interaction across different forums so that we can understand the information dissemination patterns across multiple forums, and in exploring the effectiveness of using thread-level interaction networks to identify important threads in Web forums. Another attractive direction is to apply our techniques to other CMC modes such as Blogs and Chatroom discussion. Blogs have very similar system features with Web forums, including headers and quotations. Bloggers also share usage idiosyncrasies with Web forum posters, such as typos and misspellings. Chatrooms, however, usually do not have system features and the chat postings are often too short to provide useful lexical information. By applying our algorithm to these two types of dataset we may be able to identify the potential differences in their interactional coherence.

Acknowledgements

This research was funded in part by the following grant: NSF Information and Data Management. "SGER: Multilingual Online Stylometric Authorship Identification: An Exploratory Study," August 2006–August 2007.

References

Abbasi, A., & Chen, H. (2006). Visualizing authorship for identification. In the 4th IEEE Symposium on Intelligence and Security Informatics (ISI'06). New York, NY: Springer.

Adamson, G.W., & Boreham, J. (1974). The use of an association measure based on character structure to identify semantically related pairs of words and document titles. Information Storage and Retrieval, 10, 253–260.

Bagga, A., & Baldwin, B. (1998). Entity-based cross-document coreferencing using the vector space model. In Proceedings of the 36th annual meeting on Association for Computational Linguistics (Vol. 1, pp. 79–85). Morristown, NJ: ACL.

Barcellini, F., Detienne, F., Burkhardt, J., & Sack, W. (2005). A study of online discussions in an open-source software community: Reconstructing thematic coherence and argumentation from quotation practices. In Proceedings of the Communities and Technologies Conference (C&T'05) (pp. 301–320). New York, NY: Springer.

Barzilay, R., & Elhadad, M. (1997). Using lexical chains for text summarization. In Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization, (pp. 10–17). Morristown, NJ: ACL.

Beaugrande, R.A., & Dressler, W.U. (1996). Introduction to text linguistics (pp. 84–112). New York: Longman.

Burris, V., Smith, E., & Strahm, A. (2000). White supremacist networks on the Internet. Sociological Focus, 33(2), 215–234.

Chen, H. (2005). Introduction to the special topic issue: Intelligence and security informatics. Journal of the American Society for Information Science and Technology, 56(3), 217–220.

Choi, F.Y.Y. (2000). Advances in domain independent linear text segmentation. In Proceedings of the Meeting of the North American Chapter of the Association for Computational Linguistics (ANLP-NAACL-00), (pp. 26–33). San Francisco, CA: Morgan Kaufmann.

Comer, D., & Peterson L. (1986). Conversation-based mail. ACM Transactions on Computer Systems (TOCS), 4(4), 299–319.

De Roeck, A.N. and Al-Fares, W. (2000). A morphologically sensitive clustering algorithm for identifying Arabic roots. In Proceedings of the 38th Annual Meeting on Association for Computational Linguistics (pp. 199–206). Morristown, NJ: ACL.

Donath, J., Karahalios, K., and Viegas, F.B. (1999). Visualizing Conversation. In Proceedings of the 32nd Annual Hawaii international Conference on System Sciences (HICSS'99) (Vol. 2, pp 2023). Washington, DC: IEEE Computer Society.

Eklundh, K.S. (1998). To quote or not to quote: Setting the context for computer-mediated dialogues. Technical report TRITA-NA-P9807, IPLab-144, Royal Institute of Technology, Stockholm.

Eklundh, K.S., & Rodriguez, H. (2004). Coherence and interactivity in text-based group discussions around Web documents. In Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04) (pp. 40108.3). Washington, DC: IEEE Computer Society.

Fiore, A.T., Tiernan, S.L., & Smith, M.A. (2002). Observed behavior and perceived value of authors in Usenet newsgroups: Bridging the gap. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Changing our world, changing ourselves, (pp. 323–330). New York, NY: ACM.

Hale, C. (1996), Wired style: Principles of English usage in the Digital Age. San Francisco, CA: HardWired.

Halliday, MAK, & Hasan, R. (1976). Cohesion in English. London: Longman.

Hayne, S.C., Pollard, C.E., & Rice, R.E. (2003). Identification of comment authorship in anonymous group support systems. Journal of Management Information Systems, 20(1), 301–329.

Hearst, M.A. (1994). Multi-paragraph segmentation of expository text. In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL'94) (pp. 9–16). Morristown, NJ: ACL.

Henri, F. (1992). Computer conferencing and content analysis. In A.R. Kaye (Ed.), Collaborative learning through computer conferencing: The Najaden papers (pp. 115–136). New York, NY: Springer.

Herring, S.C., & Nix, C. (1997). Is "serious chat" an oxymoron? Academic vs. social uses of Internet Relay Chat. Presented at the American Association of Applied Linguistics, Orlando, FL.

Herring, S.C. (1999). Interactional coherence in CMC. Journal of Computer-Mediated Communication, 4(4).

Herring, S.C. (2002). Computer-mediated communication on the Internet. Annual Review of Information Science and Technology, 36(1), 109–168.

DOI: 10.1002/asi

Khan, F.M., Fisher, T.A., Shuler, L., Wu, T., & Pottenger, W.M. (2002). Mining chat-room conversations for social and semantic interactions. http://www3.lehigh.edu/images/userImages/jgs2/Page_3471/LU-CSE-02-011.pdf

Kan, M., Klavans, J.L., & Mckeown, K. R. (1998). Linear segmentation and segment significance. In Proceedings of the 6th International Workshop of Very Large Corpora (WVLC) (pp. 197–205). Morristown, NJ: ACL.

Kjell, B., Woods, W.A., & Frieder, O. (1994). Discrimination of authorship using visualization. Information Processing and Management, 30(1), 141–150.

Lewis, D.D., & Knowles, K.A. (1997). Threading electronic mail: A preliminary study. Information Processing and Management, 33(2), 209–217.

McDonald, D., Chen, H., Su, H., & Marshall, B. (2004). Extracting gene pathway relations using a hybrid grammar: The Arizona relation parser. Bioinformatics, 20(18), 3370–3378.

Meho, L. (2006). E-Mail interviewing in qualitative research: A methodological discussion. Journal of the American Society for Information Science and Technology, 57(10), 1284–1295.

Miller, G.A., (Ed.) (1990). WordNet: An on-line lexical database. International Journal of Lexicography, 3(4), 235–312.

Morris, J. (1988). Lexical cohesion, the thesaurus, and the structure of text. Technical Report CSRI 219, Computer System Research Institute, University of Toronto.

Nahnsen, T., Uzuner, O., & Katz, B. (2005). Lexical chains and sliding locality windows in content-based text similarity detection. CSAIL Memo, AIM-2005-017.

Nash, M.C. (2005). Cohesion and reference in English chatroom discourse. In Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05) (pp. 108.3). Washington, DC: IEEE Computer Society.

Nasukawa, T., & Nagano, T. (2001) Text analysis and knowledge mining system. IBM Systems Journal, 40(4), 967–984.

Newman, P.S. (2002). Exploring discussion lists: Steps and directions. In Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital libraries, (pp. 126–134). New York, NY: ACM.

Osterlund, C., & Carlile, P. (2005) Relations in practice: Sorting through practice theories on knowledge sharing in complex organizations. The Information Society, 21(2), 91–107.

Ponte, J.M., & Croft, B.W. (1997). Text segmentation by topic. In Proceedings of the First European Conference on research and advanced technology for digital libraries (pp 113–126). New York, NY: Springer.

Paolillo, J. C. (2006). Conversational codeswitching on Usenet and Internet Relay Chat. Computer-Mediated Conversation, S. Herring (Ed.), to appear.

Radford, M.L. (2006). Encountering virtual users: A qualitative investigation of interpersonal communication in chat reference. Journal of the American Society for Information Science and Technology, 57(8), 1046–1059.

Resnik, P. (1995). Disambiguating noun groupings with respect to WordNet senses. In Proceedings of the 3rd Workshop on Very Large Corpora (pp. 54–68). New York, NY: Springer.

Reynar, J.C. (1994). An automatic method of finding topic boundaries. In Proceedings of 32nd Annual Meeting of the Association for Computational Linguistics (student session) (pp. 331-333). Morristown, NJ: ACL.

Rourke, L., Anderson, T., Garrison, R., & Archer, W. (2001). Methodological issues in the content analysis of computer conference transcripts. International Journal of Artificial Intelligence in Education, 12, 8–22.

Sack, W. (2000). Conversation map: An interface for very large-scale conversations. Journal of Management Information Systems, 17(3), 73–92.

Salton, G., & McGill, M. J. (1986). Introduction to modern information retrieval. New York, NY: McGraw-Hill.

Schafer, J.A. (2002). Spinning the web of hate: Web-based hate propagation by extremist organizations. Journal of Criminal Justice and Popular Culture, 9(2), 69–88.

Smith, M.A., & Fiore, A.T. (2001). Visualization components for persistent conversations. In Proceedings of the SIGCHI conference on Human Factors in Computing Systems, (pp. 136–143). New York, NY: ACM.

Soon, W.M., Ng, H.T., & Lim D.C.Y. (2001). A machine learning approach to coreference resolution of noun phrases. Computational Linguistics, 27(4), 521–544.

Spiegel, D. (2001). Coterie: A visualization of the conversational dynamics within IRC. MIT Master's Thesis, http://alumni.media.mit.edu/~spiegel/thesis/Thesis.pdf.

Te'eni, D. (2001). Review: A cognitive-affective model of organizational communication for designing IT. MIS Quarterly, 25(2), 251–312.

Voorhees, E.M. (1993). Using Word Net to disambiguate word senses for text retrieval. In Proceedings of the 16th annual international ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 171–180). New York, NY: ACM.

Walther, J.B., Anderson, J.F., & Park, D.W. (1994). Interpersonal effects in computer-mediated interaction: A meta-analysis of social and antisocial communication. Communication Research, 21(4), 460–487.

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), 35–57.

Xiong, R., Smith, M.A., & Drucker, S.M. (1998). Visualizations of collaborative information for end-users. Technical Report MSRTR-98-52, Microsoft Research.

Yee, K. (2002). Zest: Discussion mapping for mailing lists. In CSCW 2002 Conference Supplement, (pp. 123–126). New York, NY: ACM.

DOI: 10.1002/asi