Interactive Visualization and Content Analysis of Instant Messaging Networks

Preetha Appan  Bageshree Shevade  Hari Sundaram
Arts Media and Engineering Program, AME-TR-2004-14
e-mail: {preetha.appan, bageshree.shevade, hari.sundaram}@asu.edu

ABSTRACT
In this paper, we present our efforts towards allowing users to query and reflect upon their instant messaging histories. The problem is important since the volume of chat data is increasing at a significant rate, and traditional text based retrieval does not capture the social conventions in such systems, allowing no space for insightful exploration. In this work, we develop an analysis and interactive visualization framework, to provide insight to the user. In addition to the traditional text retrieval features like dominant concepts and their frequency, we propose to extract novel features specific to instant messaging domain like, usage of emoticons, use of exaggerations and emphasis words, number of foreign language words. These features are used to determine the key aspects of communication (a) between two users, (b) between a user and different categories/groups of people like friends, family, colleagues etc. Finally we develop two novel exploratory schemes to explore an archive of a group of users to help the users reflect on the nature of the communication (concepts, network temporal evolution, and communication density).

Author Keywords
Interactive Visualization, Content Analysis, instant messaging, Communication pattern.

INTRODUCTION
In this paper, we develop a system to analyze and visualize communication characteristics in instant messaging activity of a social network. The system incorporates novel features that are unique to the instant messaging domain and provides interactive visualizations that enable users to understand their communication activity in terms of these features.

There has been prior work in understanding the dynamics of online interactions through visual interfaces [2]. In [5,6] the authors use abstract shapes, color and size to indicate identity and activity of users. In [1] aspects of Usenet postings such as categories of interest, identifying leaders of a particular group are visualized. These systems provide novel visual metaphors for real-time analysis of network activity. However, since the analysis is done on generic chat room activity, social relationships, and the use of other languages (in English text) between members are not considered. In [4], the system analyzes online conversations and determines emerging social networks.

In our approach, we analyze the chat histories in a social network and develop interactive visualizations to allow the members to explore key concepts as well as social activity. We automatically parse and detect the following five features – (a) emoticons, (b) exaggerated / emphasized words, (c) foreign language words and (d) dominant concepts. We develop two novel exploratory visualizations that enable users to interactively browse their social activity over time, as well as their chat history for the extracted features. The pilot user studies indicate that the system was well liked.

The rest of the paper is organized as follows. We next discuss the design goals followed by a discussion on the messaging features. This is followed by a discussion on network visualizations and experiments.

DESIGN GOALS
We have built an interactive system to visualize and analyze user activity in the IM (Instant Messaging) domain. We formulated a set of design goals for the system, which are enumerated below:

▪ Content Analysis: Users are much more informal and expressive over IM, and this is not captured by traditional features for text information retrieval such as TF/IDF (Term Frequency – Inverse Document Frequency). We would like to develop a few novel features that capture the common instant messaging communication characteristics.

▪ Network chat visualization: The system would provide visualizations to enable users to understand various aspects of their chat activity including language use and group activity.

Our solution involves two steps – (a) feature extraction and (b) visualization interfaces. We will now describe our feature extraction procedure.

FEATURES OF CHAT COMMUNICATION
It is our goal to extract features from message logs that represent the unique aspects of Instant Messaging communication. Hence, in addition to text retrieval
features (e.g. TF/IDF), we propose the following features that capture the characteristics of the chat domain – emoticons, exaggerated words, emphasized words and foreign language words. Note that while these features are not exhaustive, we conjecture that they augment traditional features used in text retrieval and enable a better representation of the text communication in IM networks. In our work, the system initially uses a spell checker to rectify spelling mistakes. This is done to ensure that misspelled words do not get identified incorrectly as foreign language words.

Emoticons
Emoticons are graphical icons that usually represent facial expressions. These are created by a certain series of keystrokes, and are immensely popular with users of IM networks. All popular messaging networks provide a basic set of emoticons and in our system we detect 15 common emoticons (from our messaging application – MSN Messenger). These were chosen after determining the dominant emoticon usage and categorized by the authors. These are divided into five categories - “happy/laughter”, “sad/disappointment”, “naughty”, “angry” and “love”. We compute the frequency of emoticon usage for each person, and per every pair of links in the network. This computation is done per emoticon category. Once this is done, all emoticons are now removed from the message text, before extracting the next feature.

Exaggerated words
IM messages incorporate phrases that exaggerate the affect state of the participants. Examples of such phrases (or words) are “That’s cool!” “wow”, “awesome” etc. Their use is also changed by the familiarity of the relationship. However, the nature of the social relationship, while crucial for understanding the communication is rarely incorporated in text-retrieval scenarios. We developed a simple exaggerated phrase detector for each user, based on analysis of 10% of messages. We then detected such phrases for each user and then detected exaggerated words in other messages of the same user. We kept a frequency count per user, of the user’s exaggerated words. All such exaggerated words found are removed from the message text, before extracting the next feature.

Emphasized words
Emphasizing is a way of calling attention and conveying importance. In order to show emphasis for an otherwise unemphatic word or phrase, people use various techniques like underlining, highlighting, and printing in italics, changing fonts etc. Detecting emphasized words can convey interesting aspects like (a) important concepts/themes of a conversation, (b) concepts important to a particular user etc. This feature is again not captured in the traditional text retrieval system. In our work, emphasis is defined as follows: enclosing of words in quotes, or asterisk as well as capitalizing the whole word or the first letter of the word. For each category we store the phrase type and the number of words in that type.

Foreign language words
An important aspect of the global popularity of IM is the use of English-text to communicate in other languages (e.g. Indian languages such as Tamil and Hindi). These languages are also currently unsupported by messenger clients. The system initially removes all the stop words from a chat conversation. The system then performs stemming on the remaining words and searches these words in WordNet [3]. All words that are not found in WordNet after stemming, are treated as foreign language words. Note that at this stage, we have already removed emoticons so they would not be classified as foreign language words. We store the frequency count of the foreign words.

Dominant Concepts
Dominant concepts are a set of dominant non-stop words that do not contain emoticons, exaggerated / emphasized or foreign language words. Extracting dominant concepts and their frequency is similar to extracting term-frequency in traditional text retrieval systems. The system after spell-check, removes all the stop words and prepositions. The system also removes exaggerated words, emphasized words and emoticons. The system then removes the foreign language words and computes the frequency of the remaining concepts. The dominant set is just the top \( k \) most frequently occurring words in this pruned set.

NETWORK CHAT VISUALIZATIONS
We have built two interactive visualizations to enable users to gain insight into their IM activity. These visualizations are for a network of people, who contribute their message logs to the system. Our visualization enables users to answer the following queries about their chat communication.

1. Given a single user, what are the overall communication characteristics of this user, aggregated over all her conversations?
2. Given a pair or users, what are the key aspects of the communication between them?
3. How do the communication characteristics of users change, with respect to specific categories? (For e.g. friends, family and co-workers).
4. How did the communication between people in the network evolve over time?

We will now discuss the two visualizations in detail.

INTERACTIVE GRAPH VISUALIZATION
This visualization (ref. Figure 1) enables the user to execute the first three queries listed above. We first
present the user with a graph representing her chat communication network. Each node in the graph represents a member of the network and a link connects two nodes if the member’s are in each others friends list. We use different colors to indicate relationships between nodes. i.e. for each node, links to other members in the friends category are colored green, links to colleagues are colored blue and family members are colored red. This gives a quick visual representation of each member’s network of IM connections. The user can click on both nodes and links in this graph, to interact with the chat history of members of the network. We now describe how the system helps the user execute each of the above mentioned queries.

**Network Member Characteristics**

When the user chooses a particular node, the chat characteristics of that node in terms of the features extracted are aggregated over all conversations. Buttons are provided for each feature i.e. dominant concepts, emoticons, exaggerated words, emphasized words and foreign language words. However since each feature is about a unique aspect of the chat communication, the statistics collected are also presented differently.

**Dominant concepts**

To display dominant concepts of a selected node (i.e. person), we select the top 10 most frequent concepts, aggregated from all conversations of this person. These are displayed in the right part of the visualization as shown in Figure 2. The saturation of the color of the text varies according to the frequency; so the most important concept is brightly colored, while the ones with lower importance are more saturated, and appear almost white. The user can also click on any dominant concept. The system then displays the conversation between the currently selected node and another participant (node) such that the frequency of the dominant concept is highest in that conversation.

**Emoticons**

We classify emoticons into five categories (i.e. happiness, anger, naughty, sad and love). Icons representing these emoticons are used to indicate the emotion usage characteristics of the selected node. The size of the displayed icon is proportional to the frequency of its usage. This is shown in Figure 3.

**Exaggerated words and emphasized words**

We visualize the usage of both exaggerated and emphasized words relative to group usage. Figure 4 shows a detail view of this visualization that shows the relative usage of exaggerated words within a group of three users. We compare the selected user’s usage of these words with other nodes that have contributed their message archives, and depict the usage statistics graphically. A vertical bar is used to indicate the center (or average) of all users. If the selected node’s usage of exaggerated words is below the average it is indicated as a horizontal bar to the left, and if above average it is indicated as a horizontal bar to the right of the middle. The length of the bar is proportional to the difference in usage of these words between the selected node and the average. This is done with respect to three social categories – friends, family and colleagues.

**Foreign language words**

We display foreign language words, similar to dominant concepts as described above. i.e. we indicate the top 10 foreign language words of the selected user and vary their color to indicate frequency of usage. The user can click on one of these words to go to the conversation that has the maximum occurrence of that foreign language word. The system allows the user to understand the communication characteristics of the conversation.
between the two connecting nodes, when a particular link that connects two nodes is chosen by the user. The user can now determine the usage of dominant concepts, emoticons and foreign language words in the conversation.

**INSTANT MESSAGING NETWORK EVOLUTION**

![Figure 5](image)

**Figure 5:** Detail from our visualization, showing the evolution of instant messaging network over time. The thickness of the line is proportional to the activity. The user can control the visualization using a slider control.

We now present a simple visualization (ref. Figure 5) that allows the user to gain insight into the evolution of the communication network as well as the density of the communication. In order to understand the evolution, the conversations were grouped on a daily basis.

As shown in the figure, the system provides a slider that allows the user to navigate across time. The left most end of the slider indicates the earliest conversation (by day) in the network and the right most end indicates the latest conversation. As the user navigates using the slider, nodes and links appear as a graph, depending on whether a conversation took place between people representing the nodes on a given day that is indicated by the slider position. The thickness of the links also increases depending on the number of conversations that have occurred between the nodes till the currently selected day. Thus the visualization gives an intuitive understanding of how people joined the network as well as the density of communication between people.

**EVALUATION**

In order to evaluate the system, we have conducted an informal user study. The authors contributed their message logs, which ranged over approximately a year (from Jan 1 2004 to Dec 1 2004). The authors recorded their observations about the following aspects of the system: (a) Whether the features extracted captured their general chat communication behavior (b) Whether the interaction provided insight into their chat activity. We also asked three other users who were part of the network, to interact with the system and provide feedback. These users evaluated the system based on the above mentioned criteria and were asked for their opinion on the features and suggestions for improvement.

The authors found that the system reflected their communication pattern well. The dominant concepts extracted did depict the most frequent topics of communication. Also, other users liked to see variations in their emoticon usage across different categories of people. Another positive feedback was: “It was interesting to trace back the way you communicate with others as well as to know how other friends communicate”. Specifically, users found the emphasis and exaggeration features interesting. They also found that the visualizations did provide insight into important relationships (frequency of communication).

**CONCLUSIONS**

In this paper, we described a system that enables users to understand communication in an instant messaging/chat network. We extracted novel features that depict the common communication characteristics of IM users i.e. usage of emoticons, exaggerated words, emphasized words, foreign language words and dominant concepts. We also developed novel interactive visualizations that help understand the overall communication characteristics of – (a) a single user, (b) between a pair of users, (c) user with respect to a given category viz. friends, family and colleagues and (d) the evolution of network over time. Our initial user study indicates that the interfaces were well liked and the features captured the key communication characteristics. We plan to develop visualizations that help understand the evolution of concepts across users over time. We also plan to conduct extensive user studies for further system evaluation.

**REFERENCES**


