Instant Messages: A Framework for Reading Between the Lines

Jeffrey D. Campbell
UMBC
1000 Hilltop Circle
Baltimore, MD USA
1 410 455-3687
campbelljd@acm.org

ABSTRACT
A framework is described for analyzing keystroke level data from instant messages (IM). This is unlike other analyses of IM which employ server-based logs of messages. This framework can be used to identify metrics for evaluating the usability of IM during message composition. The current objective is evaluating awareness features. The model also identifies quantifiable factors that can be computed automatically during IM usage that could allow the system to adapt to different styles of IM usage. Data from a representative usability evaluation scenario is utilized to illustrate some results of using this framework. Computational aspects of the framework have been implemented in GLogger.

Categories and Subject Descriptors
H.5.3 [Information Interfaces and Presentation]: Group and Organization Interfaces – *computer supported cooperative work, synchronous interaction.* H.5.2 [Information Interfaces and Presentation]: User Interfaces – *evaluation/methodology.* H.4.3 [Information Interfaces and Presentation]: Communications Applications, *computer conferencing, teleconferencing, and videoconferencing.*

General Terms
Measurement, Design, Experimentation, Human Factors.

Keywords
Computer Mediated Communication, Usability Metrics, Awareness, Keystroke Analysis, Instant Messaging.

1. INTRODUCTION
Numerous studies of instant messaging (IM) message logs have investigated IM usage for both personal and professional purposes (e.g. [1, 5]). However, message level analysis is not sufficient to evaluate user interaction with the IM user interface. More than the final message is needed to address questions about interaction with other users and the awareness of their working status. Keystroke and mouse-click data can provide a fuller understanding of what was happening during message composition.

The volume of data and the difficulty of analyzing keystroke level data is well known. This paper presents a computational framework for analyzing this low level data from IM usage. The framework has been added to the GLogger log analysis tool. The GLogger data visualizations are described in [4] and the new framework enhancements are demonstrated in [2].

The framework was developed to perform quantitative analysis of IM usage. The framework identifies metrics for comparing usability of different user interfaces. A second use for the metrics is to identify individual differences in style. The metrics could also be used to identify differences in user behavior for different types of IM usage, for example, problem-solving versus social planning. A third potential benefit from the framework is identifying factors that the IM client can measure while the software is in use that can be used to adapt the user interface to the current situation. For example, it could activate spell checking for formal messages or adjust activity awareness indicators appropriately.

Obviously, the framework is not intended to be the only analysis of the data. For example, it is easy to calculate the time between various events at various levels but the interpretation is more difficult. One could compute the amount of time spent composing each message. One could draw conclusions from differences in composition length. However, those results must be interpreted in the context of the actual usage to avoid attributing improper causes to the quantified results. Further work is needed to validate the cause and effect relationship. With this caveat, the discussion of the framework will focus on the measurable results that can be calculated from the log files and potential interpretations that are consistent with those results.

Data has been collected from a series of laboratory-based usability evaluation scenarios designed to focus on awareness features. A typical awareness feature is an indicator that "X is typing" indicating that another person is active. Questions that stimulated the development of the framework included: To what extent are people typing messages at the same time? What is the impact of a message arriving while composing one? How often does the incoming message result in discarding the message being composed? Can a new awareness feature use this framework with real time analysis to help avoid any negative consequences?

Results from using the new framework to analyze more thoroughly data from a prior study [3] are used here to illustrate the types of results that can be obtained. The results described in [3] were based only on keystroke and message composition level

CSCW'04, November 6–10, 2004, Chicago, Illinois, USA.

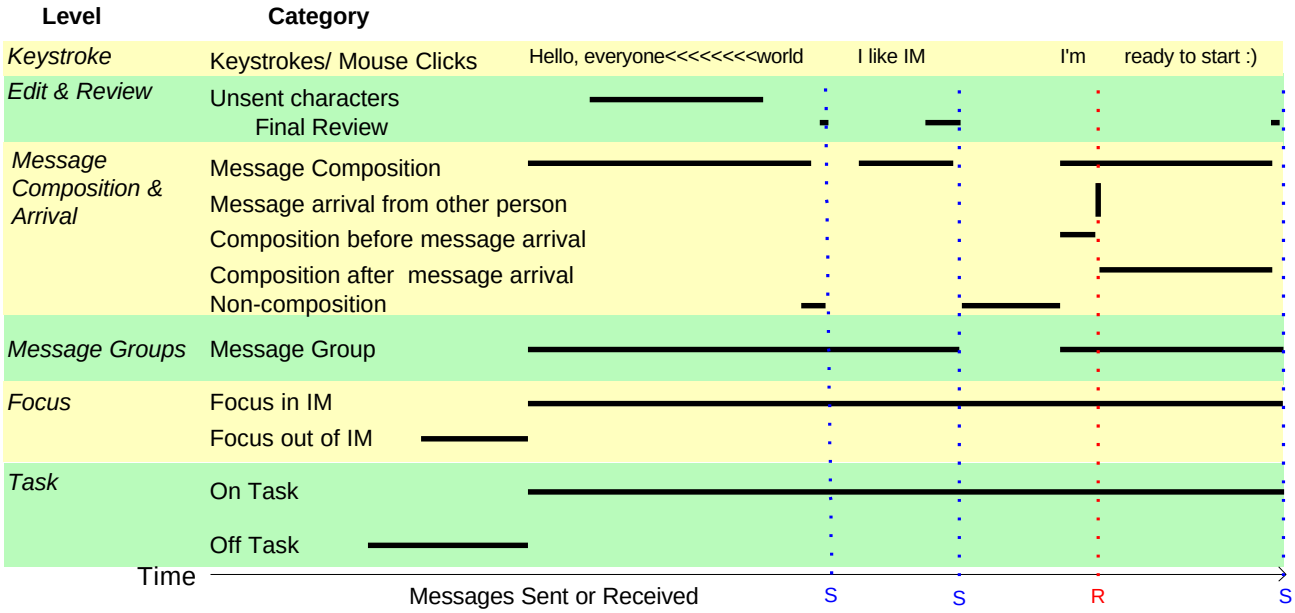

Figure 1 Instant Messaging analysis framework. Time increases to the right. Vertical dotted lines show timing of messages sent (S) and received (R). Dark horizontal lines represent classification at each level of the framework.

analysis. In that study, one participant was provided with a structure constructed from a dozen plastic toy blocks and the other had a larger pile of unassembled blocks. The objective was to build a replica of the original structure using IM as the only form of communication

The next section describes the framework and previews results from actual usability studies. Section 3 is a summary and discussion continuing research.

2. ANALYSIS FRAMEWORK

A methodology for analyzing the keystroke and mouse click data has been developed that identifies hierarchical groups of related actions. These groupings facilitate evaluation of the IM user interface by categorizing how the users are spending time while using IM. This can provide guidance in determining what aspects of IM should be improved to get the largest benefits.

Starting at the lowest level, the model consists of the following levels: Keystrokes, Edit and Review, Message Composition and Arrival (from another person), Message Groups, Focus, and Task. Each level is described in the following sections.

The framework is summarized in the time line shown in Figure 1. Alternating levels are shaded to show more clearly the categories in each level. The keystrokes are represented with time increasing to the right. For example, the gap between "I'm" and "ready" in the figure represents a pause in typing. The horizontal lines indicate the classification of a period of time at each level of the framework.

2.1 Keystroke Level

The keystroke level is the raw time stamped data about key presses and mouse actions. This includes information about characters deleted, cut, copied and pasted. Figure 1 represents backspaces with the "<" symbol.

Typing speed can be computed at the keystroke level and compared to normal typing performance. Variations in typing speed between messages or between IM conversations could be a factor in determining current usage style.

2.2 Edit and Review

The Edit and Review level is divided into two categories: Unsent Characters and Final Review. The first is used to examine any characters that were typed (or pasted) that were not in the final message sent to the other person. Unsent characters are a factor in efficiency. Those unsent characters can be classified in three categories based on the reason they were not sent. *Edits* are simple corrections to typographical or grammatical errors that use the original words. *Revisions* change the words but do not alter the meaning of the message. *Discards* make a significant change in the meaning of the message. Unsent character time is defined as the time spent creating the text that was removed and the actions (e.g. backspaces) to remove that text. This definition emphasizes the time spent on characters not sent as opposed to the time spent typing the replacement text. Alternatively, the definition could focus on time spent replacing characters and start with the removal and include the replacement characters. In the study, 57 percent of incidents in which one or more characters were unsent were the result of edits. 32 percent were revisions and 11 percent were discards.

Edit and Review data can provide insight into IM formality and awareness. If there is keystroke data available for "normal" typing by the person, the relative frequency of correcting errors can support the claims that "IM is informal" and "spelling and grammar do not matter." The techniques used in making edits and revisions could be interesting. For example, early analysis indicates that the vast majority of people backspace to the error and retype instead of using the arrow keys or mouse to just correct the erroneous characters. 98 percent of edits were made immediately, meaning that the correction was made by replacing all characters between

the error and the current position. This means that less than two percent of the edits involved leaving intervening text in place and going back to make a change. A larger percent (six) of revisions involved such delayed action. An intelligent user interface might be able to use this distinction to better accommodate the mode of current activity, perhaps handling all delayed changes as revisions instead of edits.

Discards are an interesting type of unsent characters. At the very least, the person has lost time typing part of a message that is now not used, so there is some loss of efficiency. It is possible that the removed material could have been significant to the discussion, but was lost, for example, as the topic quickly changed. During the study, about 12 percent of the occurrences of changes were classified as discards. The discard measure could be used as one aspect in comparing the effectiveness of these awareness indicators that help people know when another is active.

The second category at the Edit and Review level is Final Review. This is the time between the last keystroke or editing action and sending the message. In essence, this is the time the person may spend proofreading or reviewing the completed message. Obviously, the person may review before the end of the message, so this does not include all review time. It is also possible that the person was doing something else during this time. For example, they could have been interrupted by another task. Final review time for the 4500 messages in the sample analysis averaged 724 ms with a range from 0 (within the timer resolution) to 46 seconds.

This measure can provide insight into several behaviors. Final review time that approximates the difference between keystrokes implies the person is sending the message immediately, without much review. In such a case, there is little time for delay from spell checking or another processing intensive feature.

2.3 Message Composition and Arrival

2.3.1 Message Composition

Message composition is defined to be the time from the first action (keystroke, paste) of a message until the message is sent. Composition time continues even if the person removes the text written so far. The time between sending the message and the first action for the next message is non-composition time.

The main value in identifying composition and non-composition time blocks is the ability to analyze events that occur during those blocks. This includes the type and frequency of unsent characters which provides information about the amount of editing that occurred within a message. The duration of message composition compared to the predicted time needed for the keystrokes (total or just those sent) is another measure of editing. The number of people composing at the same time can be computed to obtain insight into parallel activity and for awareness purposes.

During the sample study, 36 percent of the time nobody was composing; 53 percent, one person was composing; and 11 percent, both people were composing simultaneously. Looking at just the composition time, 17 percent of the time when anyone was composing, they both were. Another IM experiment task had 25 percent overlap, so there seems to be some dependence on the type of task. If overlapping composition is an important aspect for improving awareness indicators, this variation could be significant.

2.3.2 Message Arrival

Messages can arrive from others during composition or non-composition time. Quantifying the impact of the arrival on message composition is an important component in evaluating IM usability. In particular, message composition blocks are divided into two segments – before and after message arrival. Comparison of measurements in the two segments for each message is the basis for analysis.

The rates of unsent characters for each message can be compared for the pre-arrival and post-arrival segments. An increase in discards in post-arrival segments would quantify the anecdotal observation that people often significantly change their message when a new message arrives. A higher frequency of cut or copy actions from the message in progress during the post-arrival segment would also be consistent with this observation. Again, awareness features that could help people reduce the apparent inefficiency due to these changes could be evaluated with these measurements.

A lower number of characters per second in the later segment would be consistent with the person pausing composition while reading the newly arrived message. The time from message arrival to such a pause would roughly measure how quickly the person looked at the new message. Of course, other factors could cause a pause, but a consistent pattern of longer pauses during the later segment of composition with message arrival than without message arrival could help quantify message reading behavior without using eye tracking equipment.

2.4 Message Groups

Message groups consist of multiple messages by the same person where the composition of the next message starts immediately after the previous message was sent. More formally, message groups are defined as the composition and intervening non-composition time when the non-composition time is less than a threshold. In the initial study, it was found that a threshold of 1 second appropriately grouped contiguous messages. It should be emphasized that such groups are based on the time composition starts for the next message, not when it is received, so these message groups cannot be identified from server logs. There were 682 message groups in the 4500 messages in the study. There were clear individual differences with some people never creating a message group and others with as many as ten (roughly fifteen percent of the messages they sent).

The primary purpose of message groups is to measure the tendency of people to quickly send a series of messages. For example, it appears that some people will divide a message into smaller chunks. This could be a stylistic difference between users. Recognizing this difference could be significant in designing the most appropriate user interface for different types of usage. Such chunking of messages could also be an attempt to send a message before someone else can change the topic which might make the current partial message moot or cause it to be discarded. Chunking could also be a way of making it clear to the other person that one is still active. Comparison of the word counts and durations of individual messages, messages within groups and the overall groups could help evaluate the situation. Lower word counts could be associated with avoiding being interrupted while higher word counts, especially within long messages groups, would be consistent with showing attention.

Closely related to message groups is the concept of turns – the change in author of the most recently received message. This ignores the time between messages and is different from message groups. The duration and number of turns can be computed with either keystroke data or server message logs. Sixty-five percent of the 4500 messages were from a user different from the sender of the prior message.

2.5 Focus

Focus indicates which application is the currently active application and will receive user input. Most simply, the focus is either on IM or not. A richer analysis could categorize non-IM focus time by the application that did have the focus. For example, in an experiment using IM to coordinate web searching to write a document using a collaborative text editor, keystrokes and focus changes were logged for the editor in addition to IM. This supports analysis of text copied and pasted between the web browser, IM and the editor.

Identifying focus and non-focus blocks is helpful in interpreting responses to messages that arrive. It can also help evaluate the effectiveness of any notification techniques in attracting attention to the new messages. A weakness of measuring only focus is that non-focus does not mean not visible, so someone could still be reading messages in a visible IM window without focus. Visibility could be added as the next higher level in the hierarchy, but at present, visibility information is not available from the keystroke logs. Visibility is difficult to log automatically since a window can be partially visible, but an analyst could record that information contemporaneously or from screen capture video.

2.6 Task

Whether or not the person is working on the specified task is the top level of the framework. At least in a lab-based usability evaluation with a set task structure, it would be possible to annotate the logs to indicate which task was being performed. This could facilitate analysis based on task and to factor out any extraneous activity. For example, if there are multiple tasks in a random or counterbalanced sequence, the task level would assist in comparing the same task across evaluation sessions.

2.7 Other Features

There are a variety of additional features that could be included in the framework to address specific research objectives. The state of UI options (e.g. audio notification of message arrival) could span tasks or be changed within a task. Other actions, such as file transfer or using an electronic whiteboard would seem to be within a task. Analysis can be performed based on these events, for example, comparing the time to reply to a message with different notification options. If the usage of the received message list (top portion of most IM clients) is of interest, the scrolling position in that list can be logged. Another level could be added to the framework identifying time when the most recent message was visible and when only old messages were displayed. This could then lead to analysis comparing composition and non-composition time depending upon visibility of the latest message in the message list.

3. SUMMARY

The purpose has been to describe a framework for analysis of keystroke level data from IM. The intention is that the framework can be used to select the appropriate factors to analyze. Not every level needs to be analyzed for every usability study. The framework is being used to analyze a series of lab evaluations of IM software which will serve as a base line for comparison when changes are made to the user interface. A log analysis tool has been developed [4] [2] that automatically identifies all of the hierarchical categories and basic calculations in about five minutes per log file. Manual classification of edit/revise/discard requires about as much time as the log duration.

The application of this framework to allow the software to evaluate the current usage style would seem to offer great potential. For example, it could interpret consistently unusually long final review times to conclude that the messages are relatively formal and automatically activate spell checking. Further work is needed to compare the metrics to observed behavior. For example, comparing think aloud protocols to the framework analysis or using eye tracking to confirm the timing of reading incoming messages instead of just implying reading from the keystroke timings. While developed for IM, a similar framework could apply to email and perhaps other text creation applications. In particular style clues that suggest formality in email could be similar to those in IM.

4. ACKNOWLEDGMENTS

Early testing of the framework by Enrique Stanziola is greatly appreciated.

5. REFERENCES

[1] Bradner, E. and Mark, G., Why Distance Matters: Effects on Cooperation, Persuasion and Deception. In *CSCW 2002*, (New Orleans, LA, 2002), ACM Press, 226-235.

[2] Campbell, J.D., GLogger: A Tool for Collaborative Usability Study Analysis and Visualization. Demonstration at *CSCW 2004*, (Chicago, IL, 2004).

[3] Campbell, J.D., Stanziola, E. and Feng, J., Instant Messaging: Between the Messages. In *IEEE Systems, Man and Cybernetics Conference*, (Washington, DC, 2003), IEEE Press, 2193 - 2198.

[4] Campbell, J.D., Stanziola, E. and Sears, A., Data Analysis and Visualization for Usability Evaluation for Collaborative Systems. In *HCI International*, (Crete, Greece, 2003), Lawrence Erlbaum Associates, 869 - 873.

[5] Isaacs, E., Walendowski, A., Whittaker, S., Schiano, D.J. and Kamm, C., The Character, Functions, and Styles of Instant Messaging in the Workplace. *In CSCW 2002*, (New Orleans, LA, 2002), ACM Press, 11-20.