Text Mining for Chat Message Analysis

Siu Cheung Hui
School of Computer Engineering
Nanyang Technological University
Singapore
asschui@ntu.edu.sg

Yulan He
Informatics Research Centre,
The University of Reading
Reading RG6 6BX, UK
yulan.he@hotmail.com

Haichao Dong
School of Computer Engineering
Nanyang Technological University
Singapore
dong0006@ntu.edu.sg

***Abstract*—Instant Messaging (IM) is a peer-to-peer service for remote users to communicate with each others. There are many IM systems such as MSN Messenger and Yahoo Messenger which are used by millions of users everyday. However, IM technology serves as a double edged sword and could be misused for illegitimate information exchange or committing crimes for its anonymity and completely uncontrolled chatting environment. To help enforce legitimate contents communicated in chat environments, IM monitoring systems have been developed for monitoring chat messages. Although most of these systems can provide good monitoring functions, they only provide simple message analysis features such as browsing and simple keyword-based searching of the recorded messages. In this paper, we propose a system, called IMAnalysis, that supports intelligent chat message analysis using text mining techniques. The IMAnalysis system provides functions on chat message retrieval, social network analysis and topic analysis. Chat message retrieval provides general browsing and retrieval support. Social network analysis discovers the social interactions of IM users with their contacts. And topic analysis detects automatically the topics that IM users are involved in.**

Keywords—Chat Message Analysis; Social Network Analysis; Topic Analysis; Machine Learning; Text Mining

I. INTRODUCTION

Recently, Instant Messaging (IM) has gained much popularity owing to its ability in supporting real-time communications of chat messages through private online channels. Some of the popular IM systems including Microsoft's MSN Messenger [1], Yahoo Messenger [2], and America Online's ICQ [3] are used by millions of users everyday. However, IM technology serves as a double edged sword and could be misused for illegitimate information exchange or committing crimes for its anonymity and completely uncontrolled chatting environment. Sexual solicitation [4], online bully [5], and confidential information stealing or leaking have been great threats to IM users [6], especially for children and youngsters. In addition, IM can also be used by terrorists for making contacts, which pose a great danger for the safety of a society.

To help enforce legitimate contents communicated in chat environments, a number of commercial IM monitoring systems such as Stellar Internet IM [7], Spectro Pro [8] and Chat Watch [9] have been developed for monitoring chat messages. These systems record chat messages and provide facilities to help the monitoring authorities to analyze the chat messages. Although most of these systems can provide good monitoring functions, they only provide simple message analysis features such as browsing and keyword-based searching of the recorded chat messages. To enhance chat messages analysis capability, a few research works [10, 11, 12] have also been started on analyzing chat messages for detecting the topics which are under discussions among users.

In this research, we have developed an intelligent chat message analysis system called IMAnalysis using text mining techniques. In IMAnalsyis, we have provided three functions for chat message analysis: chat message retrieval, social network analysis and topic analysis. In this paper, we discuss the different chat message analysis techniques in IMAnalysis. The rest of the paper is organized as follows. Section 2 reviews some of the current IM monitoring systems. Section 3 introduces the IMAnalysis system and its supports for chat message retrieval and social network analysis. Section 4 discusses the topic analysis technique. Finally, section 5 concludes the paper.

II. INSTANT MESSAGING MONITORING SYSTEMS

Currently, there are a number of instant messaging monitoring systems available commercially. These systems can be divided broadly into two categories, namely network-based and client-based IM monitoring systems. The network-based monitoring systems mainly focus on monitoring IM activities from a centralized server within a corporate network. And multiple IM systems from the client PCs are monitored at the same time. On the other hand, the client-based monitoring systems are installed at PCs and monitor IM activities directly from the PCs. In this section, we review some of the most popular commercial IM monitoring systems including Stellar Internet IM [7], Spector Pro [8], Chat Watch [9], Spybuddy [13] and Invisible Keylogger [14]. The Stellar Internet IM is a network-based monitoring system, while the others are client-based monitoring systems. In this section, we compare the different IM monitoring systems in terms of their chat monitoring and chat analysis capabilities.

A. Chat Monitoring

Table I gives a comparison between different IM monitoring systems in terms of IM monitoring capabilities. These systems can support popular IM systems such as ICQ, MSN Messenger and Yahoo Messenger. Network-based systems record chat messages from different users according to its message arrival order. Most client-based systems store chat messages as session-based logs. All IM monitoring systems record similar types of information such as the IM

978-1-4244-1674-5/08 /$25.00 ©2008 IEEE

system name, timestamp of the chat session and chat content. Additional information such as the names of participants and the titles of conversations are also recorded in some systems. All client-based monitoring systems are detectable by Microsoft Windows' Task Manager. They also provide password protection measure for protecting the system from being terminated and deleted. Moreover, most monitoring systems are able to monitor new versions of IM systems when the updates are installed. Some systems such as SpyBuddy and Invisible Keylogger are unable to deal with multilingual text messages.

TABLE I. IM MONITORING CAPABILITIES.

	Stellar Internet IM	Spector Pro	Chat Watch	Spybuddy	Invisible Keylogger
Supported IM Systems	MSN, ICQ, Yahoo	MSN, ICQ, Yahoo	MSN, ICQ, Yahoo	MSN, ICQ, Yahoo	Any IM systems
Chat Log Recording	Message-based	Session-based	Session - based	Session-based	Message-based
Detectable by Task Manager	NA	Yes	Yes	Yes	Yes
System Protection	NA	By access password	By access password	By access password	By access password
Adapting to New IM Versions	Yes, but update is required	Yes, but update is required	Yes, but update is required	Yes, but update is required	Yes, but update is required

B. *Chat Analysis*

Chat message analysis aims to browse through the vast amount of messages and filter out information that is of particular interest to the monitoring authorities. Most IM monitoring systems provide the following chat message analysis functions: chat log browsing, chat message retrieval, chat content search and simple statistical reporting.

Most IM monitoring systems support the viewing of the contents of chat log files. However, most IM monitoring systems only support constrained retrieval of chat messages. Stellar Internet IM performs chat message retrieval based on time or predefined network groups. Similarly, most client-based systems such as Spybuddy support retrieval of chat logs based on time. The constrained chat message retrieval capability helps users to focus only on a subset of chat logs. In chat content searching/filtering, most IM monitoring systems support the keyword-based chat content searching function. Stellar Internet IM also performs online chat content filtering based on a user predefined keyword list. Similarly, Chat Watch can also highlight sensitive keywords in chat logs. For statistical reporting, most IM monitoring systems, especially client-based systems, provide a very simple statistical report on IM monitoring including statistical information on the recorded chat messages. Network-based systems such as Stellar Internet IM are able to produce role-based reports based on predefined network user groups within organizations. Table II gives a comparison of existing IM monitoring systems based on message analysis capabilities.

In chat log analysis of most IM monitoring systems, only simple message analysis and visualization functions for displaying the recorded chat messages are provided. However, these functions are insufficient for the processing of a large quantity of chat messages. Human inspection of chat contents could be very tedious.

TABLE II. IM MESSAGE ANALYSIS CAPABILITIES.

	Stellar Internet IM	Spector Pro	Chat Watch	Spybuddy	Invisible Keylogger
Log Browsing	Yes	Yes	Yes	Yes	Yes
Message Retrieval	IP, time, network account	N.A.	N.A.	Time	N.A.
User Tracking	Nickname, client IP and network account	Windows account	Nickname	Windows account	N.A.
Content Searching	N.A.	Keyword based	Keyword based	N.A.	N.A.
Content Filtering	Keyword based (online)	N.A.	N.A.	N.A	N.A.
Statistical Reporting	By IP and network account	N.A.	N.A.	N.A	N.A.

III. CHAT MESSAGE ANALYSIS

The IMAnalysis system supports the following three major functions: chat message retrieval, social network analysis and topic analysis. In chat message retrieval, it allows users to browse and retrieve chat sessions. It also displays statistical data on chat activities of the monitored IM users. In social network analysis, it extracts sender-receiver pairs of chat messages and constructs the social interaction network of the monitored IM users. In topic analysis, it detects and analyzes session topics of chat messages communicated between users. In this section, we discuss chat message retrieval and social network analysis. And topic analysis will be discussed in the next section.

A. *Chat Message Retrieval*

Chat message retrieval supports the browsing and retrieval of chat session data archived in the chat log database. Chat message retrieval provides three sub-functions, namely statistics generation, chat message browsing and chat message searching.

In statistics generation, it gives the usage statistics on the average number of messages and average message length (in terms of number of words) that are generated from the specified chat sessions. In chat message browsing, it allows users to browse through the retrieved chat session data from chat log database. In addition, the user can also specify the keywords to be highlighted during browsing. In chat message searching, it allows users to search for chat sessions according

to their interest. A simple Boolean search is used for message searching. For example, in Figure 1, after specifying the target user and time duration, a search based on the keyword "basketball" is performed and the search results are then displayed on the screen.

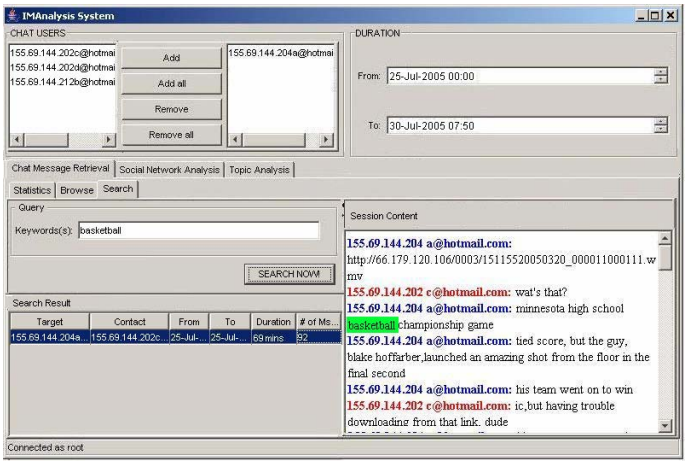


Figure 1. Chat message searching.

B. *Social Network Analysis*

Social network analysis [15] gives the social interactions between the target users and their contacts from the buddy lists. It extracts sender-receiver information of chat messages according to the specified criteria on the target user and time duration. The social interactions between the sender-receiver pair are described in terms of quantity and message direction. If there are many chat messages exchanged between the sender and receiver, it implies a close relationship or strong tie between the sender and receiver. The inbound or outbound message direction indicates the respective receiving or sending of messages of the target user. Moreover, the target user may also interact with contacts from the buddy list that may or may not be under monitoring.

Figure 2 shows the interface for social network analysis. After the user has selected the target user—155.69.144.204a@hotmail.com, a star-like social network of the target user is displayed. As shown in the figure, the target user has exchanged chat messages with two contacts. Chat users (both the target user and his contacts) are represented as nodes and the sender-receiver relationships (i.e., messages are exchanged between them) are represented as links. We use the thickness of links to indicate the amount of messages exchanged between the target user and the contact.

When the system user wants to know the statistical information of a social relationship between the target user and a contact, he can either click on a link or node. When a link is selected, the information for a social relationship between the target user and the contact is displayed. This information includes the total time spent, the total number of sessions occurred and the average time spent per session. When a node is selected, the information on the target user is then displayed. This information includes the number of connections, the number of inbound messages, and the number of outbound messages of the target user.

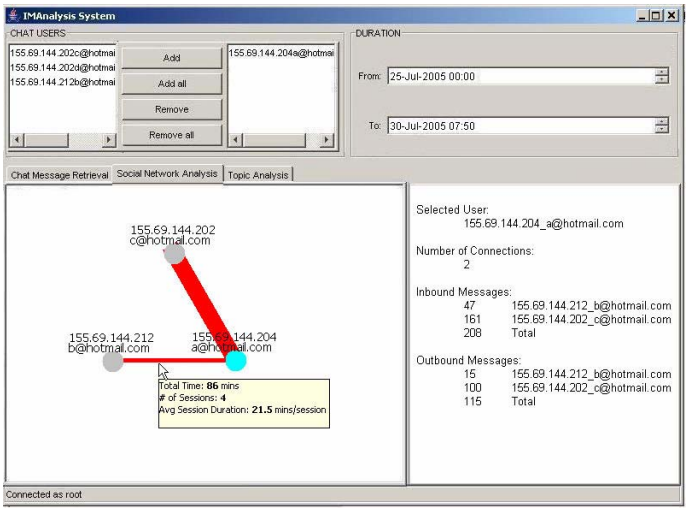


Figure 2. Social network analysis.

IV. TOPIC ANALYSIS

Chat topic analysis is one of the important functions of the IMAnalysis system. Our topic analysis is based on topic detection techniques which can be broadly classified into supervised and unsupervised. Classification techniques have been investigated for supervised topic detection on chat messages. Elnhrawy [11] presented a topic categorization approach for analyzing chat conversation logs related to criminal activities. Classification techniques including k-Nearest Neigbour, Naïve Bayes and linear Support Vector Machine have been used for topic classification. On the other hand, for unsupervised topic detection, Kolenda *et al.* [10] applied Independent Component Analysis (ICA) for chat room topic detection. And Bingham *et al.* [12] proposed a similar chat room topic detection approach to that of [10] except that Complexity Pursuit is used instead of ICA.

A. *Classification-based Topic Detection*

For topic analysis in IMAnalysis, we aim to identify chat sessions limited only to a number of important topics. Therefore, we have adopted supervised topic detection approaches based on Naïve Bayes [16], associative classification [17] and Support Vector Machine (SVM) [18] which have been demonstrated with good performance for classifying text documents. In this research, we have developed a chat topic detection approach [19] using topic indicative terms to support multi-label categorization, i.e., chat sessions can be classified with multiple class labels. Topic indicative terms are identified by an experimental study on sample training data and are predefined for each topic. The use of indicative terms has greatly reduced the inputs to the various classification algorithms, thereby improving the effectiveness and efficiency of the categorization process. In IMAnalysis, we focus on detecting five topics which include Sports, Games, Entertainment, Travel and Pornography. The

first four topics are common topics amongst teenagers whilst the last one is an objectionable topic.

Figure 3 shows the classification-based approach for chat topic detection, which comprises the following four major components: Sessionalization, Feature Extraction, Feature Selection and Topic Categorization.

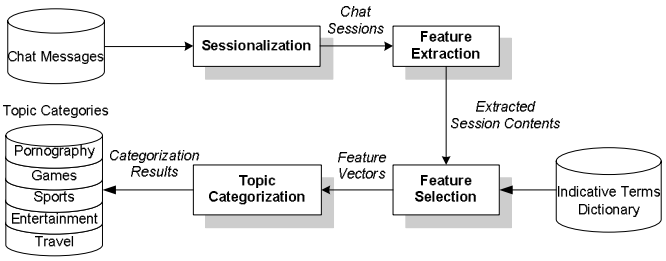


Figure 3. Classification-based approach for chat topic analysis.

Sessionalization aims to group a collection of related chat messages into sessions for processing and categorization. Due to the nature of chat messages that is short and concise, a single chat message is typically less than 10 words. This poses a big challenge for topic detection. To tackle this problem, we group a collection of chat messages into a basic processing and categorization unit or *session*. A session is defined as a sequence of chat messages exchanged within the lifespan of a chat dialog window.

Feature Extraction extracts features such as textual contents, icon text, and URL contents from chat sessions. For URL's web page contents, information displayed in the viewable body text and several other locations such as the title of the web page, and the meta-data of "description" and "keywords" are extracted. The textual contents, the icon text and the web page contents are combined together to form the chat session content. Figure 4 shows an example for the Feature Extraction process.

Feature Selection selects chat features for categorization based on indicative terms stored in the Indicative Term Dictionary. This is based on our observation that chat conversations on a particular topic usually contain a set of words, called *indicative terms* (or *topic keywords*) that characterize that particular topic. This set of indicative terms is considered to be highly representative for all conversations on the same topic. Therefore, indicative terms can be treated as a unique collection of features characterizing the chat contents belonging to a particular topic. Indicative terms can also reduce the dimensionality of input features to the classifiers. Table III gives some example indicative terms stored in the Indicative Term Dictionaries for the Games category. As shown in Table III, each row represents a unique indicative term indexed by the first column. Different entries from the same row represent all the possible variations of the unique term. During Feature Selection, any matches of the indicative terms in the same row will contribute to one occurrence of that unique feature represented by the row.

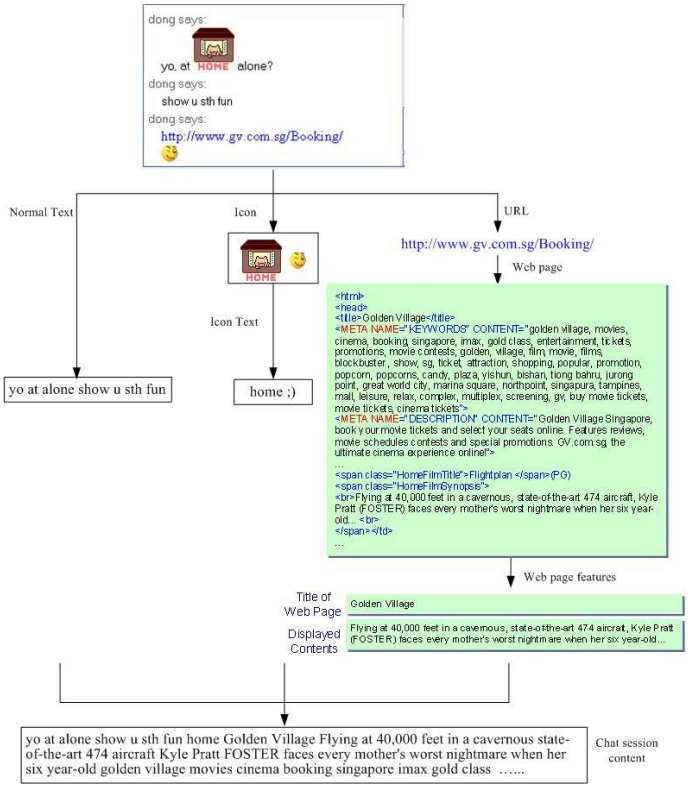


Figure 4. An example on feature extraction.

TABLE III. SAMPLE INDICATIVE TERMS FOR "GAMES".

Index	Term 1	Term 2	Term 3	…
1	cs	counter strike	counterstrike	
2	game	games	gaming	gamers
3	graphics card	graphics cards	gfx card	
4	multiplay	multiplayer	multi-players	multi-player
5	pc game	computer game	video game	computer games
	...			

Topic Categorization classifies chat sessions into one or more topic categories using topic classifiers based on Naïve Bayes, Associative Classification and SVM. Multiple chat topic classifiers have been built with one for each topic category. If more than one classifier votes positive, the chat session will be assigned with multiple chat topic labels. If none of the classifiers votes positive, the chat session will then be classified into the "Others" category. One of the major advantages of this multiple-voting process is that a new topic classifier can be easily incorporated if an additional topic is to be considered.

B. Performance Results

The performance of the classification-based chat topic detection approach has been evaluated based on the three classifiers Naïve Bayes (NB), Associative Classification (AC) and Support Vector Machine (SVM) using an Intel Pentium 4 3.0 Gigahertz machine with 1 Gigabytes of memory running the Microsoft Windows XP operating system. To evaluate the performance, we downloaded chat messages from several web chat sites including *UGroups.com* [20], *jolt.co.uk* [21] and *AdultfriendFinder* [22]. Five subsets of chat messages were collected, with each corresponding to one of the five topic categories under evaluation (i.e., Pornography, Sports, Games, Travel and Entertainment). The "Others" subset was also collected for training purposes. Each subset contained about 800 sessions for each category. The classifiers were undergone a training process before they were used for topic detection.

Figure 5 shows the topic categorization performance results of NB, AC and SVM on each category of the testing data set of chat messages. As shown in the figure, SVM outperformed the other two classifiers in precision, F-measure and accuracy. Among the five categories, all the three classifiers, SVM, NB and AC, give the best performance on the "Sports" category and the poorest performance on the "Entertainment" category. Table IV gives the global performance of the three classifiers based on the macro-average of all measures across the five categories. It can be seen that SVM has achieved the best performance in all measures. High precision (87.25%) and accuracy (92.14%) are obtained. NB has achieved better performance than AC. However, all classifiers have produced relatively lower recall values with 77.16%, 70.80% and 68.65% for SVM, AC and NB respectively.

Figure 6 shows the interface for topic analysis. After the user has specified the target monitored users and time duration, the identified session data and its detected topics are then displayed on the screen. The user can select to view the details of a chat session. In addition, the displayed information can also be sorted according to different attributes. This will enable users to sort the chat sessions according to topics.

V. CONCLUSIONS

As instant messaging may be misused, it poses a great threat to IM users, especially children. In order to help protect online safety for chat users, an intelligent chat message analysis system, called IMAnalysis, has been developed. The IMAnalysis system provides three main functions, namely chat message retrieval, social network analysis and topic analysis, for analyzing chat messages for possible detection on the misuse of IM.

Chat message retrieval can provide statistical information that helps detect unusual chat usage of a user and the corresponding time periods. In addition, the chat message browsing and searching functions can also help to examine the potential leakage of personal information to others. The social network analysis can help display the social interactions between a target user and his contacts. As such, it can be used to detect whether a user is withdrawn from social relationships. Topic analysis can help identify the topics of chat sessions, thereby determining whether there are any inappropriate chat contents communicated in the chat conversations.

TABLE IV. CATEGORIZATION PERFORMANCE RESULTS BASED ON MACRO-AVERAGE MEASURE.

Methods	Precision (%)	Recall (%)	F-Measure (%)	Accuracy (%)
NB	84.10	68.65	75.14	86.91
AC	77.83	70.80	72.53	85.01
SVM	87.25	77.16	81.19	92.14

REFERENCES

[1] MSN Messenger. Available at http://messenger.msn.com/.

[2] Yahoo Messenger. Available at http://messenger.yahoo.com/.

[3] ICQ. Available at http://www.icq.com/.

[4] Timothy, "How to cheat on/exploit Bejeweled (theoretically)". Available at http://timothy.mess.be/Bejeweld-Exploit.html, 2003.

[5] M. Wolf, "Cyber Brats: Bullies Who Taunt Their Peers with the Click of a Mouse". Available at http://www.parenthood.com/articles.html?article id=4335.

[6] K. Thomas, "Kids need enduring smarts with instant messaging". Available at http://www.usatoday.com/tech/news/2001-07-10-instant-messaging-safety.htm.

[7] Stellar Technologies, "Stellar Technologies, Inc. – web, e-mail & IM management Stellar Internet global employee management". Available at http://www.stellarim.com/.

[8] Spectorsoft.com, "SpectorSoft.com – Spector Pro for Windows". Available at http://www.spectorsoft.com/products/SpectorPro_Windows/index.html.

[9] Zemerick Software Inc., "Zemerick software – home of Chat Watch". Available at http://www.zemericks.com/.

[10] T. Kolenda, L. K. Hansen and J. Larsen, "Signal detection using ICA: application to chat room topic spotting". In: 3rd International Conference on Independent Component Analysis and Blind Source Separation. pp. 540-545, 2001.

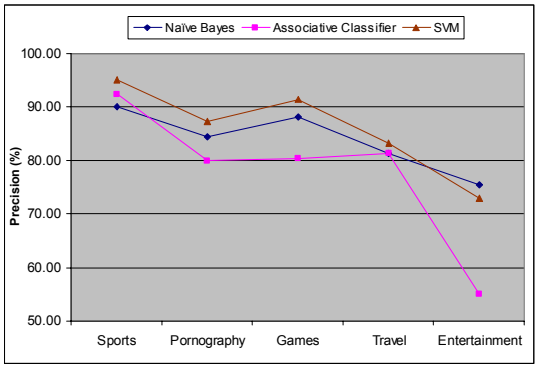
[11] E. Elnahrawy, "Log-based chat room monitoring using text categorization: a comparative study". In: Proceedings of the IASTED International Conference on Information and Knowledge Sharing (IKS 2002). St. Thomas, US Virgin Islands, 2002.

[12] E. Bingham, A. Kab and M. Girolami, "Topic identification in dynamical text by complexity pursuit". Neural Processing Letters 17, 69-83, 2003.

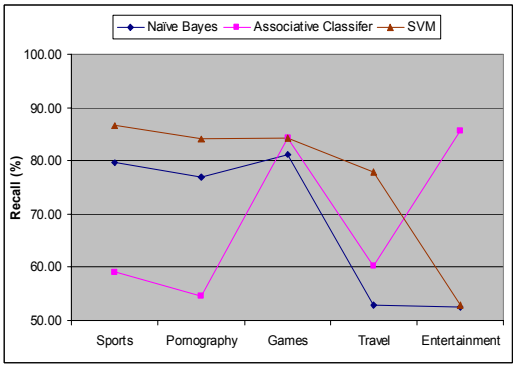
[13] Exploreanywhere.com, "SpyBuddy spy software". Available at http://www.exploreanywhere.com/sb-intro.php.

[14] InvisibleKeylogger.com, "Invisible Keylogger – the perfect stealth keylogger". Available at http://www.invisiblekeylogger.com/.

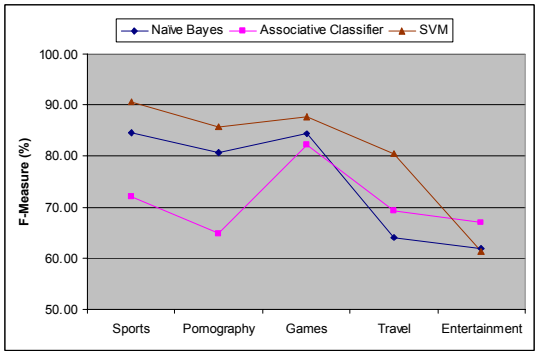
[15] J. Resig, S. Dawara, C. Homan and A. Teredesai, "Extracting social networks from Instant Messaging populations". In: KDD 04 Link Discovery Workshop (LinkKDD 2004), 2004.



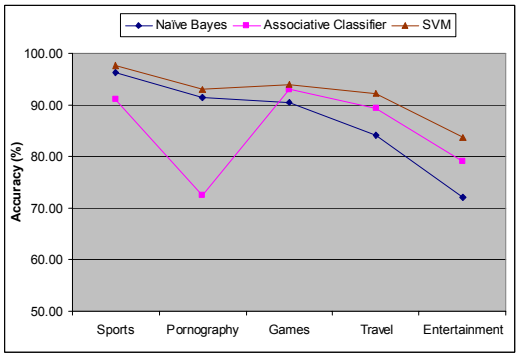
(a) Precision



(b) Recall



(c) F-measure



(d) Accuracy

Figure 5. Categroization performance results.

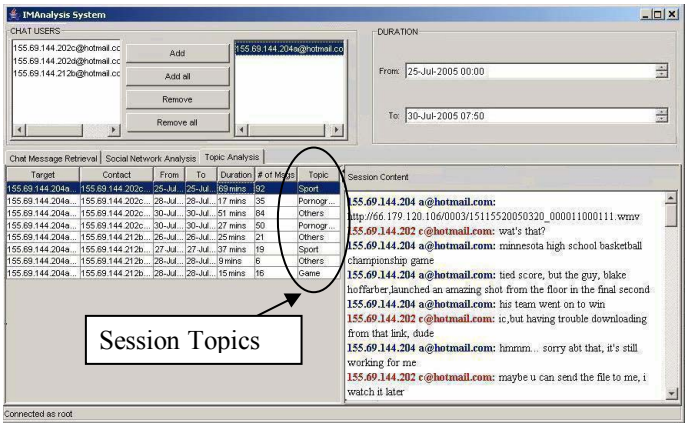


Figure 6. Topic analysis interface.

[16] K. Tzeras and S. Hartman, "Automatic indexing based on Bayesian inference networks". In: 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'93), pp. 22-34, 1993.

[17] M.-L. Antonie and O. R. Zaiane, "Text document categorization by term association". In: IEEE International Conference on Data Mining, pp. 19-26, 2002.

[18] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features". In: European Conference on Machine Learning (ECML), Berlin, pp. 137-142, 1998.

[19] H. Dong, S.C. Hui and Y. He, "Structural analysis of chat messages for topic detection". Online Information Review, 30(5), pp. 496-516, 2006.

[20] Ugroups.com, "UGroups: free web access to the Usenet newsgroups and forums". Available at http://www.ugroups.com.

[21] Jolt.co.uk online game forum, "jolt.com.uk public forums – powered by vBulletin". Available at http://forums.jolt.co.uk.

[22] AdultFriendFinder.com, "Adult friendfinder – the world's largest sex personal sites". Available at http://www.adultfriendfinder.com.