Structural Analysis of Chat Messages for Topic Detection

Haichao Dong, Siu Cheung Hui and Yulan He *

School of Computer Engineering, Nanyang Technological University,
Nanyang Ave, Singapore 639798

Abstract

Purpose - This paper studies the characteristics of chat messages from analyzing a collection of 33,121 sample messages gathered from 1700 sessions of conversations of 72 pairs of MSN Messenger users over 4-month duration from June to September of 2005. The primary objective of chat message characterization is to understand the properties of chat messages for effective message analysis such as message topic detection.

Methodology/Approach - From the study on chat message characteristics, an indicative term-based categorization approach for chat topic detection is proposed. In the proposed approach, different techniques such as sessionalization of chat messages and extraction of features from icon texts and URLs are incorporated for message pre-processing. And Naïve Bayes, Associative Classification, and Support Vector Machine are employed as classifiers for categorizing topics from chat sessions.

Findings - Indicative term-based approach is superior than the traditional document frequency based approach for feature selection in chat topic categorization.

Originality/Value - This paper studies the characteristics of chat messages and proposes an indicative term-based categorization approach for chat topic detection. The proposed approach has been incorporated into an instant message analysis system for both online and offline chat topic detection.

Preprint submitted to Online Information Review *2 May 2006*

Keywords: Chat message characteristics, Instant messaging, Topic detection

Paper type: Research paper

1 Introduction

Instant Messaging (IM) is a peer-to-peer service for remote users to communicate with each other, which typically comprises many client-based chat programs and a centralized server. The client programs allow IM users to communicate with direct connections while the server broadcasts the availability of users. There are many IM systems available in the market. Some of the most popular ones are Microsoft's MSN Messenger (MSN.com, 2006), Yahoo Messenger (Yahoo.com, 2006), and America Online's ICQ (ICQ.com, 2006). The popularity of these IM systems is greatly attributed to its anonymity, privacy and convenience.

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 (Timothy, 2003), online bully (Wolf, 2003), and sensitive or confidential information stealing or leaking have been great threats to the daily life of people (Thomas, 2001), 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. Therefore, some kinds of control measures such as IM monitoring are highly desirable in order to fight against the misuse of IM. To support IM monitoring, it is necessary to have effective techniques for analyzing the recorded chat messages. Topic detection (Kolenda, Hansen, and Larsen, 2001; Elnahrawy, 2002; Bingham, Kab, and Girolami, 2003) is one of the impor-

* School of Computer Engineering, Nanyang Technological University,
Nanyang Ave, Singapore 639798

Email address: {DONG0006,asschui,asylhe}@ntu.edu.sg (Haichao Dong, Siu Cheung Hui and Yulan He).

tant research areas on chat message analysis, which aims to analyze the chat content for identifying the topics that are under discussion among users.

This paper studies the characteristics of chat messages from analyzing a collection of 33,121 sample messages gathered from 1700 sessions of conversations of 72 pairs of MSN Messenger users over 4-month duration from June to September of 2005. These users are randomly selected from graduate/undergraduate students of Nanyang Technological University and National University of Singapore, Singapore. The primary objective of chat message characterization is to understand the properties of chat messages for effective message analysis such as message topic detection.

In this paper, we propose an indicative term-based categorization approach for chat topic detection. In the proposed approach, different techniques such as sessionalization of chat messages and the extraction of features from icon text and URLs are incorporated for message pre-processing. And Naïve Bayes (Tzeras and Hartman, 1993), Associative Classification (Antonie and Zaiane, 2002), and Support Vector Machine (Joachims, 1998) are employed as classifiers for categorizing topics from chat sessions. The proposed approach has been incorporated into an instant message analysis system for both online and offline chat topic detection.

The rest of the paper is organized as follows. Section 2 analyzes chat messages based on chat language, icons, hyperlinks, message length, and dynamic content. The proposed topic detection approach is discussed in Section 3. Performance evaluation and results are presented in Section 4. Section 5 discusses briefly the instant message analysis system. Finally, Section 6 concludes the paper.

2 Chat Message Characterization

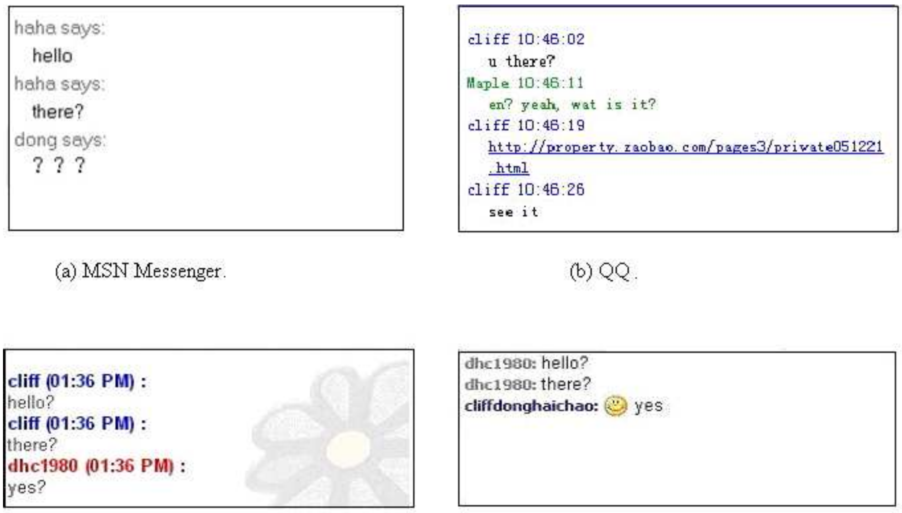
This section first reviews the general conversational formats for the most popular IM messaging systems including MSN Messenger, QQ, ICQ and Yahoo

Messenger. QQ (Tencent.com, 2006) is currently the most popular IM system for Chinese communities. It then analyzes chat messages in terms of chat language, icons, hyperlinks, message length, and dynamic content. Finally, message characteristics that are important for effective message analysis are summarized.

2.1 Conversational Format

Figure 1 shows the conversational formats of some of the most popular IM systems such as MSN Messenger, QQ, ICQ and Yahoo Messenger. In general, the conversational format consists of the following three components:

- *Chat participants.* They are IM users participated in the current session of conversation identified by their respective nicknames. For example, haha and dong in Figure 1(a) are the two participants of the current session of conversation in MSN Messenger. Similarly, chat participants of the other IM systems can be identified easily.
- *Optional information.* Optional information can be attached at the end of the nicknames of the chat participants. For example, in QQ and ICQ, a timestamp is attached at the end of the participant's nickname. QQ has the timestamp in the format of hour:minute:second (see Figure 1(b)), while ICQ displays the timestamp in a different format (see Figure 1(c)). Another example of optional information is shown in Figure 1(a) in which MSN Messenger attaches the word "says" after the participant's nickname, while Yahoo Messenger does not attach any optional information following the participant's nickname.
- *Chat messages.* The conversational contents of chat messages are displayed or typed after the participant's nickname and the associated optional information. In most IM systems except Yahoo Messenger, each chat message starts with a new line. Yahoo Messenger displays chat messages immediately after the participant's nickname.



```
haha says:
  hello
haha says:
  there?
dong says:
  ? ? ?
```

(a) MSN Messenger.

```
cliff 10:46:02
  u there?
Maple 10:46:11
  en? yeah, wat is it?
cliff 10:46:19
  http://property.zaobao.com/pages3/private051221
  .html
cliff 10:46:26
  see it
```

(b) QQ.

```
cliff (01:36 PM) :
hello?
cliff (01:36 PM) :
there?
dhc1980 (01:36 PM) :
yes?
```

(c) ICQ.

```
dhc1980: hello?
dhc1980: there?
cliffdonghaichao: 🙂 yes
```

(d) Yahoo Messenger.

Fig. 1. IM conversational formats.

2.2 Message Characteristics

IM systems are originally designed only for text-based communications. However, in additional to textual contents, contemporary IM systems also support the insertion of icons and hyperlinks into message contents. Moreover, the chat language used by most chat users is also quite different from conventional written English. In this section, we discuss the characteristics of chat messages based on the collected set of 33,121 chat messages.

2.2.1 Chat Language

Chat language is basically written English. However, due to the real-time and informal conversational environment of IM systems, chat messages are written in a very different way from conventional English. Some of the common usage features in chat language include *acronyms*, *short forms, polysemes, synonyms* and *mis-spelling* of terms.

- *Acronyms* are formed by extracting the first letters of a sequence of words. For example, ASAP is an acronym for “As Soon As Possible”. In the 33,121 chat messages we analyzed, there are altogether 156 acronyms. The top 12 most popular acronyms are listed in Table 1.

Table 1

Example of popular acronyms.

Acronym	*Equivalent Meaning*	*Acronym*	*Equivalent Meaning*
ASAP	As Soon As Possible	OTP	On The Phone
ASL	Age Sex Location	POS	Parent Over Shoulder
BRB	Be Right Back	TTYL	Talk To You Later
BF	Boy Friend	U2	You too
GF	Girl Friend	WTH	What The Heck
CU	See You	YW	You are Welcome

- *Short forms* refer to the case in which a lengthy word is replaced with a shorter alternative expression. Table 2 shows some example short forms. Unlike acronyms, it is observed that only some popular short forms have fixed expressions among different chat participants. Many short forms are highly subjective to the context of the conversation and chat users.

Table 2

Examples of short forms.

Short Form	*Equivalent Meaning*	*Short Form*	*Equivalent Meaning*
L8R	Later	Tom	Tomorrow
Nvr	Never	Btw	Between
Tat	That	Pic	Picture
Nvm	Never-mind	Wlm	Welcome
Frenz	Friends	Congrats	Congratulations
Sth	Something	Eg	Example

- *Polysemes* refer to terms that have more than one interpretation. In chat environment, a term can be either a word or a short form. For example, “comp” can refer to “company” or “computer” depending on the context.
- *Synonyms* refer to the case in which terms with similar or same meaning

are used interchangeably. For example, “network adaptor”, “network interface card”, and “NIC” can be used interchangeably during conversation on computer hardware and networking related topics.

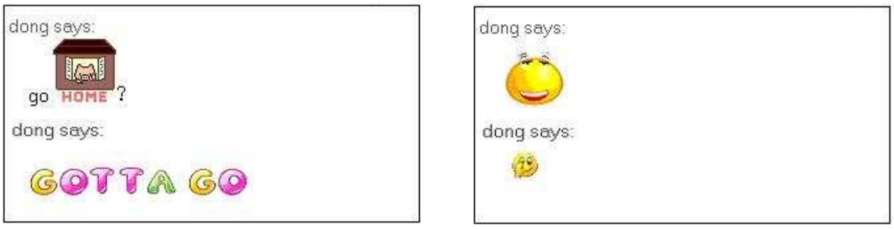
- *Mis-spelling* of terms occurs at a higher rate in chat conversations than in traditional published text documents. Due to the real-time and informal nature of chat conversation, mis-spelled words in chat messages often occur. There are also cases in which a chat participant purposely mis-spells a word to emphasize its meaning. A common case for mis-spelling is the use of duplicated vowels, such as “sooooo”, “noooo” and “thee” instead of “so”, “no” and “the” respectively. The number of duplications is not fixed. Another case is the substitution of similar pronounced letters. For example, “today” becomes “todae” and “ok” becomes “okie”.

In addition to the above language characteristics, we have also counted the number of words contained in the collection of sample messages. There are a total of 14,000 words or 6000 distinct words after stemming and conventional stop-word removal.

2.2.2 Icons

Icons are images inserted along with the text content. According to the graphical contents, icons can be divided into two groups: text and non-text. *Text icons* are images carrying graphical text. *Non-text icons* contain little textual content. For example, Figure 2(a) contains two chat messages composed with text icons. In the first message, the word “HOME” is part of an icon image. In the second message, the icon “GOTTA GO” is formed from images of characters ‘G’, ‘O’, ‘T’, and ‘A’. In another example, Figure 2(b) shows two smiley icons. The first icon mimics a silly laugh, whereas the second is a farewell icon. Both icons are just images without any textual information associated with it.

All IM systems under study have implemented shortcut texts for icons. Each icon is associated with a unique sequence of text as shortcut, which will be



dong says:

go HOME ?

dong says:

GOTTA GO

(a) Example text icons.

dong says:

dong says:

(b) Example non-text icons.

Fig. 2. Icons used in chat messages.

interpreted and replaced with the corresponding icon when typed. Graphical text is usually the same as the shortcut. For example, the first icon in Figure 2(a) is associated with the word "home". When the word "home" is typed, it is automatically replaced by the icon without any user actions. Non-text icons, on the other hand, are usually associated with annotations or semantically meaningless shortcuts specified by users. For example, the icons in Figure 2(b) are associated with the word "heihei" and the shortcut sequence ":8" respectively from our study.

2.2.3 Hyperlinks

In chat conversations, hyperlinks or URLs (Uniform Resource Locators) can be given to refer other chat participants to Web resources such as Web pages and files for information sharing. URL must be specified in an absolute form beginning with a scheme name followed by a network name which points to a host server. Table 3 gives the statistics on URL links from the collection of sample chat messages. Some popular URL scheme names used in the collection include *http* (HyperText Transfer Protocol), *ftp* (File Transfer Protocol), *mms* (Multimedia Message Service), and *rtsp* (Real Time Streaming Protocol). The "# of Occurrences" refers to the number of chat messages containing a URL. The percentage of occurrences is given in "% out of Total Occurrences" from the total of 251 URL occurrences. On the other hand, the "# of Sessions" shows the number of chat sessions containing one or more hyperlinks. The percentage of occurrences is given in "% out of Total Sessions" from the total of 1700 sessions.

Table 3

Statistics of URL links.

URL Scheme	*# of Occurrences*	*% out of Total Occurrences*	*# of Sessions*	*% out of Total Sessions*
http	219	87.25%	162	9.53%
ftp	27	10.76%	14	0.82%
mms	4	1.59%	2	0.12%
rtsp	1	0.40%	1	0.06%
Total	251	100.00%	179	10.53%

Amongst the 219 http links, 40 of them point to downloadable files such as *mp3*, while the rest are all Web pages. Http links are the most popular form of hyperlinks (87.25% out of all URLs) used in chat conversations and occur in 9.53% of the total 1700 sessions. On the other hand, *ftp* links are used for file sharing and occurred in 0.82% of total sessions in the data collection. Streaming media links such as *mms* and *rtsp* rarely occur in chat sessions.

2.2.4 Message Length

Table 4 shows the statistics on chat message and session length from the collection of chat messages. The message length is quite short with about 91.5% of chat messages less than 50 bytes. Most chat sessions have length greater than 50 bytes but less than 500 bytes. However, there are still 34.50% of chat sessions with less than 50 bytes in length. On the other hand, there are also 14.10% chat sessions containing more than 500 bytes but less than 5000 bytes of data.

Moreover, we have also found that the average time gap between two adjacent chat messages is around 20 seconds and a typical chat session has duration of 4-20 minutes for conversations greater than 50 bytes but less than 500 bytes.

Table 4

Chat message and session length.

Message Length *(bytes)*	*% out of* *Total Messages*	*Chat Session* *Length (bytes)*	*% out of* *Total Sessions*
0-10	34.40%	0-20	14.00%
11-20	22.40%	21-50	20.50%
21-50	34.70%	51-100	21.10%
51-100	6.90%	101-150	12.70%
101-500	1.40%	151-500	17.60%
501 and above	0.20%	501-5000	14.10%

2.2.5 Chat Topics

Due to the interactive and dynamic nature of the IM environment, chat topics (such as Games, Sports, Pornography, etc.) can be changed quite rapidly within a chat session. It is possible that each chat session of conversation contains multiple topics and a topic may also spread over several sessions of conversations. Table 5 shows the statistics on the number of topics discussed in the collected set of chat sessions. 67.24% of all sessions are observed to focus on a single topic. On the other hand, 13.59% of sessions contain two topics in a discussion and a total of about 20.3% of chat sessions contain two or more topics. Moreover, 12.46% of chat sessions are not dedicated to any meaningful topics. These chat sessions are mostly short sessions, which contain messages mainly on greetings.

2.3 Discussion

In this section, we have studied the conversational format and message characteristics of IM systems from the collection of 33,121 sample chat messages. We summarize the findings in relation to data analysis as follows:

Table 5

Chat session topics.

# of Topics	*# of Sessions*	*% out of Total Sessions*
1	1143	67.24%
2	231	13.59%
3	59	3.47%
4	38	2.24%
5	17	1.00%
0	212	12.46%

- *Conversational format.* The conversational format preserves the correspondence between chat messages and participants. As such, statistical analysis and social network analysis based on the recorded chat messages are possible.
- *Chat language.* The chat language used for IM conversations contains acronym, short form, polyseme and mis-spelled words, which make data analysis difficult.
- *Hyperlinks* and *icons.* Hyperlinks and icons contain useful information for instant messaging. Both types of data are important for data analysis.
- *Message length.* Chat messages are short. Each chat session may have one or multiple chat messages. Instead of using each chat message as a unit for data analysis, chat messages can be organized as sessions for analysis.
- *Chat topics.* Each chat session may contain one or more topics. As such, topic detection should be able to identify multiple topics.

3 Topic Detection

In this section, we propose an indicative term-based chat topic detection approach for both online and offline topic analysis.

Topic detection, which identifies conceptual topics discussed among text documents, is a challenging problem in text mining (Joachims, 1998; Bingham, Kab, and Girolami, 2003; Young and Sycara, 2004). For the past few years, many research works (Masand, Linoff, and Waltz, 1992; Yang, 1994; Wiener, Pedersen, and Weigend, 1995; Yang and Liu, 1999) on topic detection have been conducted mainly based on conventional text documents such as news articles (e.g., Reuters-21578 news articles (NIST, 2006)), Web documents (e.g., Web pages from Open Directory Project (ODP, 2006)), and newsgroups (e.g., the 20 newsgroups (Lang, 2006)). These conventional text documents are written in a standard language (such as English) and usually have self-contained and coherent content. Recently, with the proliferation of Instant Messaging systems and the need for analyzing the contents of such systems, a few research works (Kolenda, Hansen, and Larsen, 2001; Elnahrawy, 2002; Bingham, Kab, and Girolami, 2003) have been started on analyzing conversational messages such as chat messages, which are also considered as written speech transcripts, for topic detection. The language structure of chat messages is quite different from that of conventional text documents in terms of chat language usage, incompleteness, shortness, interwoven topics, and multimedia context.

Topic detection approaches can be broadly classified into supervised and unsupervised. Supervised approaches require domain experts for training text documents on pre-defined conceptual topics, and prediction on topic labels can then be made on unknown data objects. Unsupervised approaches, on the other hand, cluster text documents into different groups according to the similarity of its contents without involving domain experts for the purpose of retrieving text documents of the same or similar topics.

Different classification techniques including regression models (Yang and Chute, 1994), nearest neighbors classification (Masand, Linoff, and Waltz, 1992; Yang, 1994), Bayesian probabilistic approaches (Tzeras and Hartman, 1993; Moulinier, 1997), decision trees (Apte, Damerau, and Weiss, 1998), inductive rule learning (Cohen and Singer, 1996; Moulinier, Raskinis, and Ganascia, 1996), neural networks (Ng, Goh, and Low, 1997), and Support Vector Machines (SVM) (Joachims, 1998) have been investigated for supervised topic detection. Elnhrawy (Elnahrawy, 2002) presented an offline topic categorization approach for analyzing chat conversation logs related to criminal activities. The logs are first pre-processed with stop-word removal and converted into term frequency weighted vectors. Classification techniques including k-NN, Naive Bayes and linear SVM are used for topic classification. Performance was evaluated based on Web chat logs using a measure on "average accuracy" and the results showed that the Naive Bayes classifier "significantly" outperformed k-NN and SVM. However, the evaluation is based only on a small data set and the single measure on "average accuracy", which is not clearly defined and could be biased.

Bengel *et al.* (Bengel, Gauch, Mittur, and Vijayaraghavan, 2004) also adopted a categorization approach for analyzing chat messages from Internet Relay Chat (IRC) (IRC.org, 2006). In this work, the archived chat messages are filtered based on time, chat room channel or chat message authors. The resultant collections of chat messages are grouped as "sessions" for processing and categorization. Each of these "sessions" is pre-processed with stop-word removal and stemming, and then represented using TFIDF weight scheme for classification. Instead of using chat messages, the classifier is trained based on Web pages obtained from ODP (Open Directory Project) for pre-defined topics. However, the performance of the categorization approach is not given in the paper.

Unsupervised approaches on conventional text documents are mainly based on clustering techniques such as K-Means (Jain and Dubes, 1998), Agglomerative Hierarchical Clustering (AHC) (Jain and Dubes, 1998), and Expectation Maximization (EM) (Dempster, Laird, and Rubin, 1977). On the other hand, unsupervised approaches for chat message topic detection are mostly based on signal processing techniques (Kolenda, Hansen, and Larsen, 2001; Bingham, Kab, and Girolami, 2003). These techniques treat text documents or chat sessions as a mixture of sources, i.e., topics. The textual terms are observations of signal sources. The objective is to separate the topic sources based on the observations of textual terms.

Kolenda *et al.* (Kolenda, Hansen, and Larsen, 2001) applied Independent Component Analysis (ICA) for chat room topic detection. In this approach, the chat messages are first pre-processed with stop-word removal, and then partitioned into "sessions" by overlapping fix-sized windows. Latent Semantic Indexing (LSI) is then applied to the "sessions" for dimension reduction before ICA is carried out for topic detection. This approach is said to be able to detect four highly relevant dynamic topics. However, as the performance is not given in the paper, it is not clear whether the proposed approach is effective. Further, human interpretation is required in order to label the topics with automatic generation of indicative terms.

Bingham *et al.* (Bingham, Kab, and Girolami, 2003) proposed a similar chat room topic detection approach to that of Kolenda *et al.* except that Complexity Pursuit is used instead of ICA. The Complexity Pursuit algorithm separates interesting components from a time series of chat message data and identifies the hidden topics. It is claimed to have the best performance compared with ICA-based approaches. However, the approach has encountered the same problem as that of ICA in terms of topic interpretation. Besides, the performance evaluation is conducted on the standard 20 newsgroups instead of chat messages based on a single measure on "total error", which could be biased.

The supervised approaches suffer from the major drawback that great effort is required for training classifiers for detecting topics from text documents. However, once the classifier has been trained, topic detection is efficient and effective. On the contrary, the unsupervised approaches detect all possible topical groups from text documents. However, unsupervised approaches simply group text documents discussing similar topics and present the overall context structure. Human effort is required for labeling the topics for interpretation.

Besides, the effectiveness of unsupervised approaches, especially online topic detection, on text documents is generally not satisfactory.

In chat topic detection, we aim to identify chat sessions discussing only a limited number of important topics with high accuracy. In addition, topic detection will also be incorporated into online monitoring, in which online topic detection is needed. Therefore, a supervised topic detection approach will be more suitable and is investigated here.

According to the review on topic classification approaches, Naive Bayes, associative classification, and SVM are some of the most commonly used classification techniques which can achieve good performance for conventional text documents. In this research, we adopt the three classification algorithms for our proposed chat message topic detection approach.

3.2 Proposed Topic Detection Approach

We propose a chat topic detection approach 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 the sample training data and are predefined for each topic. The use of indicative terms greatly reduce the inputs to various classification algorithms, thereby improving the efficiency of the categorization process. This is particularly important for online topic detection.

Before discussing the proposed approach, we first need to determine the topics that should be detected in chat conversations. In (Lee, 2003), the most popular topics chatted amongst most teenagers are identified. However, some of the popular topics are either very general (e.g. Gossip, Life in general, etc.) or not well-defined (e.g. Next weekend, The future, etc.). In this research, we focus on five topics for investigation which include Sports, Games (i.e., Video games/computers), Entertainment, Travel (i.e., Someone to date and weekend), and Pornography (i.e., Sex and Secret things). The first four topics are common topics amongst teenagers while the last one is an objectionable topic. The selection of these five topics is based on our major interest for analysing chat topics related to teenagers. Nevertheless, other topic categories of chat messages can also be defined.

Figure 3 shows the proposed classification-based approach for chat topic detection, which comprises the following four major components, Sessionalization, Feature Extraction, Feature Selection, and Topic Categorization. *Sessionalization* groups a collection of related chat messages into sessions for processing and categorization. *Feature Extraction* extracts features such as textual contents, icon text, and URL contents from chat sessions. *Feature Selection* selects chat features for categorization based on indicative terms stored in the Indicative Term Dictionary. *Topic Categorization* classifies chat sessions into one or more topic categories using topic classifiers based on Naive Bayes, Associative Classification, and Support Vector Machines.

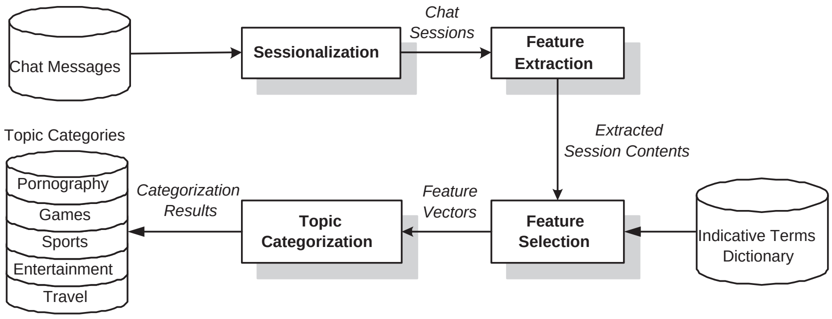


Fig. 3. The proposed classification-based approach for chat topic detection.

3.2.1 Sessionalization

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, a collection of chat messages are gathered as the 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. The assumption is that the dialog window will be closed when the conversation ends naturally. However, there are exceptions:

- A user may close a dialog window after a few messages have been exchanged.
- A user may also leave a chat window running for a long time without closing it even though all conversations have ended.

The above two situations cause problems to our initial definition on sessions. To resolve this, we refine the session definition with the following two heuristics:

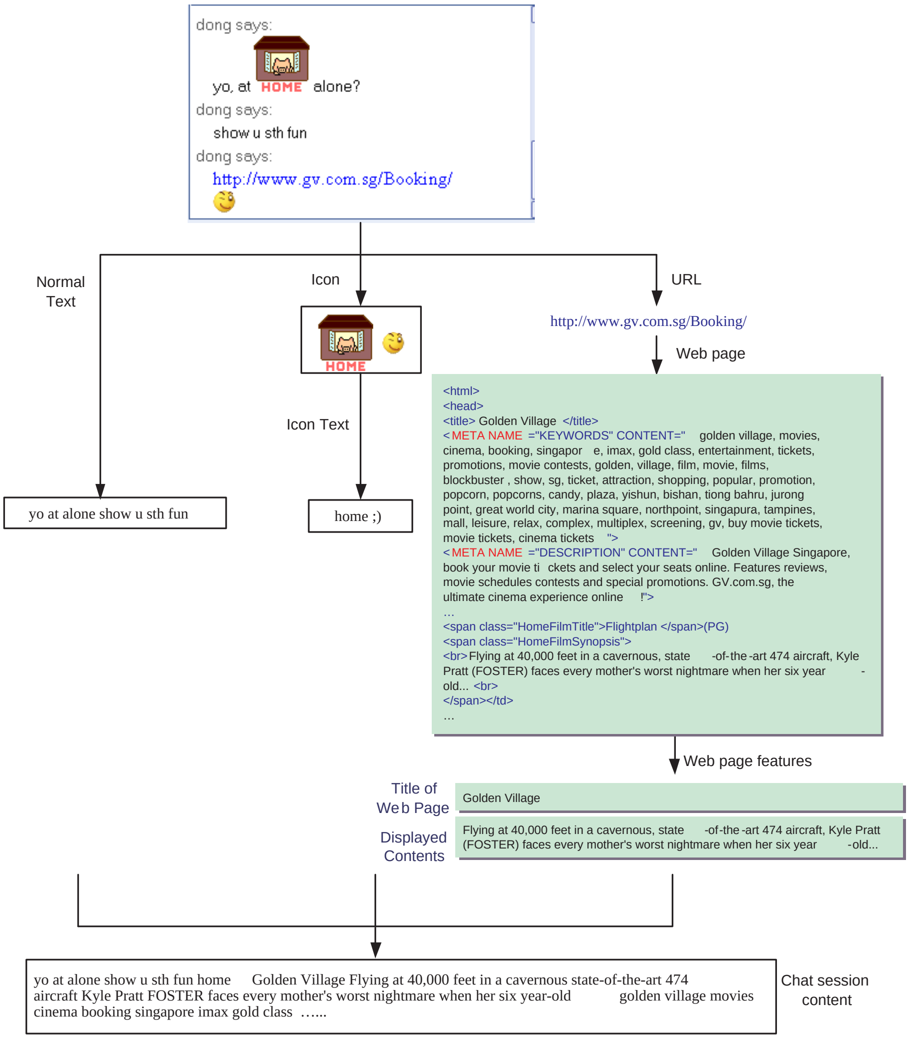
- Merge sessions of the same participants with temporal proximity for their potentially coherent content. In this research, we set the temporal proximity boundary to be 10 minutes empirically. In other words, two chat sessions occurred between the same participants with an interval less than 10 minutes will be merged.
- Split messages of a chat session that have a long time gap between them into two or more sessions. Typical chat sessions last between 4 to 20 minutes. If two chat messages have a time gap of more than 40 minutes, we will consider the gap as long and the chat messages will be divided into two sessions.

3.2.2 Feature Extraction

Apart from textual contents, a chat message may also contain icons and even URLs. *Feature Extraction* extracts both icon texts and Web page contents of the corresponding URLs appeared in chat sessions. For 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. Chat session contents will then consist of textual message contents, icon texts and Web page contents if there are any. Figure 4 shows an example for the *Feature Extraction* process.

3.2.3 Feature Selection

In *Feature Selection*, indicative terms stored in the Indicative Term Dictionary are used to select appropriate features for classification purposes. This is based



dong says:

yo, at HOME alone?

dong says:

show u sth fun

dong says:

http://www.gv.com.sg/Booking/

```
<html>
<head>
<title> Golden Village  </title>
< META NAME  ="KEYWORDS" CONTENT="    golden village, movies,
cinema, booking, singapor   e, imax, gold class, entertainment, tickets,
promotions, movie contests, golden, village, film, movie, films,
blockbuster , show, sg, ticket, attraction, shopping, popular, promotion,
popcorn, popcorns, candy, plaza, yishun, bishan, tiong bahru, jurong
point, great world city, marina square, northpoint, singapura, tampines,
mall, leisure, relax, complex, multiplex, screening, gv, buy movie tickets,
movie tickets, cinema tickets    ">
< META NAME  ="DESCRIPTION" CONTENT="    Golden Village Singapore,
book your movie ti   ckets and select your seats online. Features reviews,
movie schedules contests and special promotions. GV.com.sg, the
ultimate cinema experience online     !">
…
<span class="HomeFilmTitle">Flightplan </span>(PG)
<span class="HomeFilmSynopsis">
<br>Flying at 40,000 feet in a cavernous, state      -of-the -art 474 aircraft, Kyle
Pratt (FOSTER) faces every mother's worst nightmare when her six year         -
old... <br>
</span></td>
…
```

Normal Text
Icon
URL
HOME
Icon Text
Web page
yo at alone show u sth fun
home ;)
Web page features
Title of Web Page
Golden Village
Displayed Contents
Flying at 40,000 feet in a cavernous, state   -of-the -art 474 aircraft, Kyle Pratt (FOSTER) faces every mother's worst nightmare when her six year   -old...
Chat session content

yo at alone show u sth fun home   Golden Village Flying at 40,000 feet in a cavernous state-of-the-art 474 aircraft Kyle Pratt FOSTER faces every mother's worst nightmare when her six year-old   golden village movies cinema booking singapore imax gold class ……

Fig. 4. Feature extraction.

on our observation that chat conversations on a particular subject/discussion (called *topic*) usually contain a set of words, known as *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 are not limited to single word, it might be phrases as well. With indicative terms predefined as features for selection, it can also reduce the dimensionality of input features to the classifiers.

Figure 5 shows the *Feature Selection* process which consists of two steps, Tokenization and Indicative Terms Identification. *Tokenization* simply breaks the chat session content into a list of single words or tokens while preserving the relative ordering of the tokens. Each term is also converted into lower cases for processing. *Indicative Terms Identification* then selects a set of terms from the tokens for each topic category based on the Indicative Terms Dictionary. The selected indicative terms will then be incorporated into a feature vector which will be used as the input to the topic classifiers. The weight of each indicative term in the feature vector will be "1" or "0" depending on the appearance of the corresponding term in the chat session.

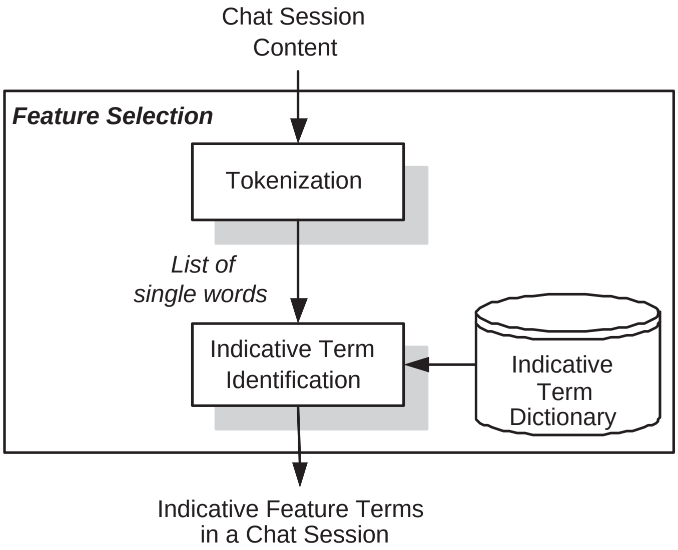

Chat Session Content
Feature Selection
Tokenization
List of single words
Indicative Term Identification
Indicative Term Dictionary
Indicative Feature Terms in a Chat Session

Fig. 5. Feature selection.

As only the indicative terms will be used for the categorization process, it is of the utmost importance to choose the most representative set of indicative terms for each topic category. The list should not be too long or too short. If the list is too long, noise (irrelevant words) and overheads will be introduced. On the other hand, if the list is too short, performance will be affected. After careful inspection and statistical analysis on the training sample data for each topic category, the sets of indicative terms for each of the five topic categories have been identified. The identification of indicative terms has taken into the consideration of chat message characteristics such as short-forms, acronyms, and polysemic words as discussed in Section 2.

Table 6 gives some example indicative terms stored in the Indicative Terms Dictionary for the topic category "Games". As shown in the figure, each row represents a unique indicative term indexed by the first column. Different entries in the same row represent all the possible variations of a particular term. During *Indicative Term Identification*, any matches of the entries in the same row will contribute to one occurrence of that unique term.

Table 6

A sample indicative dictionary for “Games”.

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

3.2.4 Topic Categorization

Topic Categorization classifies chat sessions into one or more topic categories. Multiple chat topic classifiers have been built with one for each topic category. If more than one classifier vote positive, the chat session will be assigned multiple chat topic labels. If none of the classifiers votes positive, the chat session will then be classified into “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.

4 Experiments

Three sets of classifiers have been built based on three different classification techniques, Naïve Bayes (NB), Associative Classification (AC), and Support Vector Machine (SVM), for topic categorization. All experiments have been conducted on an Intel Pentium 4 3.0 Gigahertz machine with 1 Gigabytes of memory running Microsoft Windows XP operating system.

4.1 Setup

Experiments were conducted on chat messages downloaded from several Web chat sites including *UGroups.com* (Ugroups.com, 2006), *jolt.co.uk* (online game forum, 2006) and *AdultfriendFinder* (AdultFriendFinder.com, 2006). Five subsets of chat messages were collected, with each corresponding to one of the five topic categories under evaluation. For each subset of the data set, the structural information of Web chats, such as reply, quote, and author, is removed to retain only the chat contents. In addition, the "Others" subset is also collected for training purposes. Each subset has about 800 sessions for each category. Table 7 shows the statistics on the data set of chat messages.

Table 7
The statistics on the data set of chat messages.

Category	*No. of chat sessions*	*Size (MB)*
Sports	792	1.1
Pornography	807	2.5
Games	789	1.8
Travel	784	3.2
Entertainment	753	2.5
Others	1118	5.6

The classifiers are required to undergo a training process before they can be used for topic detection. Therefore, the training data set was formed by taking approximately 70% of chat messages for each topical category of the collected data. The remaining 30% was then used as the testing data set for performance evaluation of topic categorization.

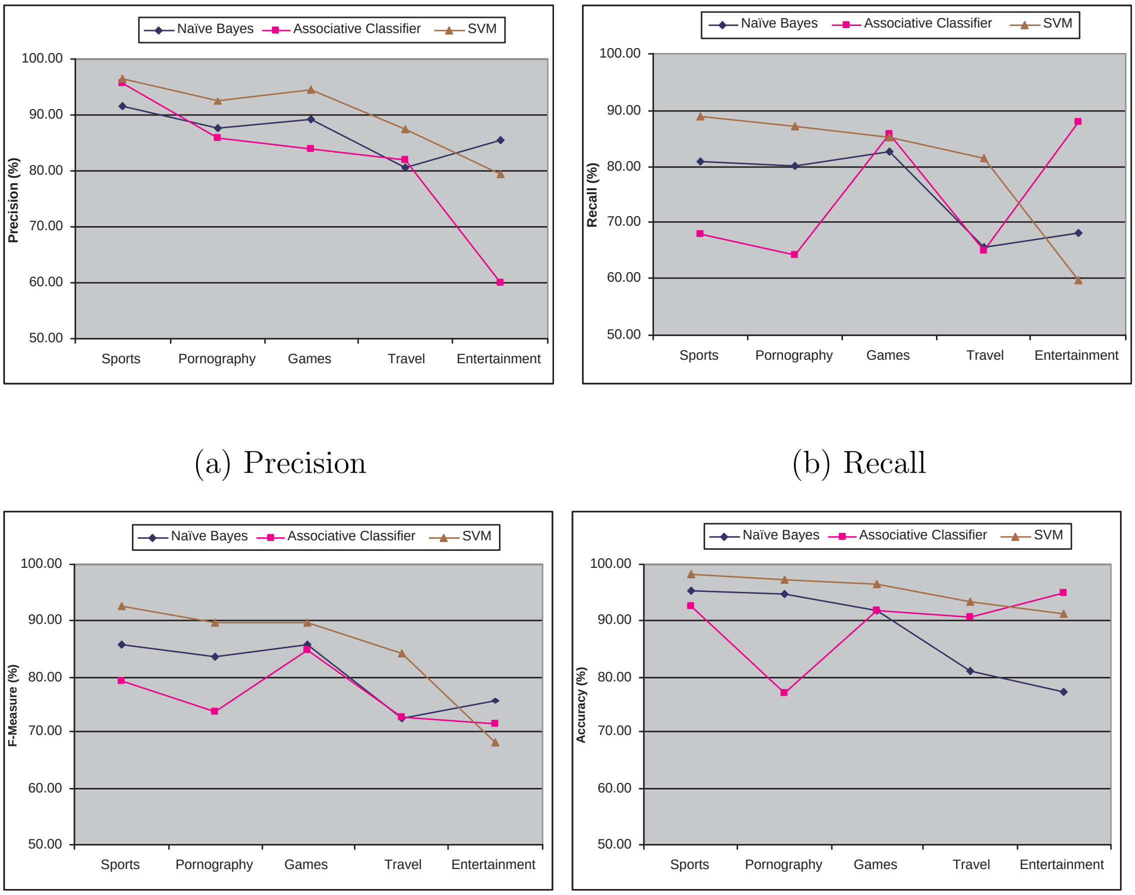
4.1.1 Evaluation Measures

In the proposed topic categorization approach, each binary classifier will determine whether a new chat session belongs to a particular topic category C_i. Training samples belonging to a particular topic category are positive samples, and the rest are negative samples. We count four values a, b, c and d for a topic category C_i, where a is the number of correctly predicted positive samples; b is the number of incorrectly predicted positive samples; c is the number of incorrectly predicted negative samples; and d is the number of correctly predicted negative samples.

To evaluate the performance for each topic category C_i, we use the measures *precision (p)*, *recall (r)*, *F-measure (F_1)* and *accuracy* (van Rijsbergen, 1979) where *precision* is defined as the proportion of correctly predicted positive samples in all positive samples $a/(a+b)$; *recall* is defined as the proportion of correctly predicted samples in all positive predicated samples $a/(a+c)$; *F-measure* measures the balance between precision and recall and is defined as $2rp/(r+p)$; *accuracy* is the proportion of total correct prediction in all samples and is defined as $(a+d)/(a+b+c+d)$. In addition, we also use *macro-average* (van Rijsbergen, 1979) to calculate the average value of all topic categories under evaluation for each measure in order to evaluate the performance across all topic categories.

4.1.2 Training Performance

Figure 6 shows the training performance results of NB, AC and SVM on each category of training data set of chat messages. SVM outperformed the other two classifiers in precision, F-measure and accuracy. It achieved very good performance in precision and accuracy across all five topic categories. However, all classifiers performed relatively poor in recall. Among all topic categories, classification of the “Sports” category achieved the best performance, while classification of the “Entertainment” and “Travel” categories generally gave poorer performance. This might be due to the difficulty faced to extract truly representable indicative terms from the sample data since both topics cover a broad range of sources from TV/movies to games and travel.



(a) Precision

(b) Recall

(c) F-measure

(d) Accuracy

Fig. 6. Training performance results based on the training data set.

Table 8 shows the global training performance results in macro-average measures. It can be observed that SVM achieved the best performance in all measures with high precision (90.02%) and accuracy (95.33%) across all five topic categories. SVM and NB gave relatively lower scores in recall (80.52% and 75.52% respectively) compared with other measures.

4.1.3 Categorization Performance

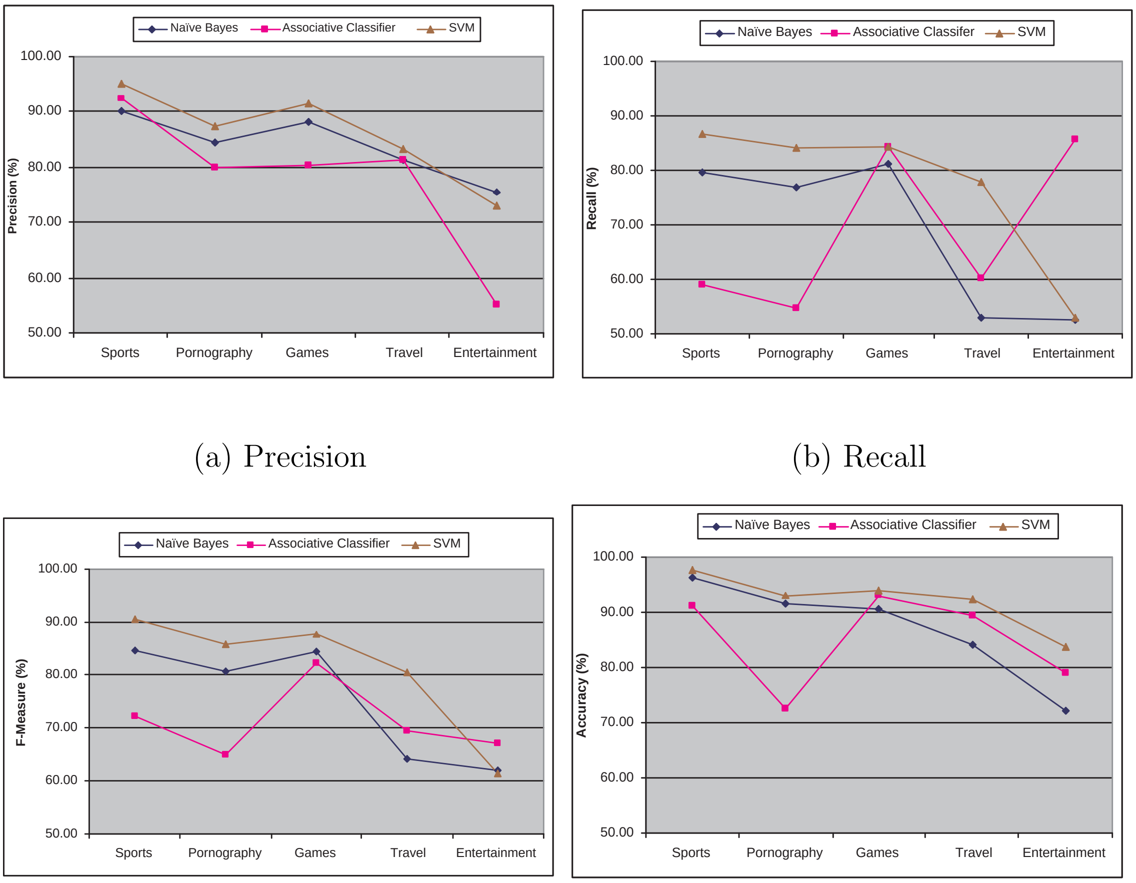
Figure 7 shows the topic categorization performance results of NB, AC and SVM on each category of the testing data set of chat messages. SVM outperformed the other two classifiers in precision, F-measure and accuracy. Among

Table 8

Training performance results based on macro-average measures.

Methods	*Precision (%)*	*Recall (%)*	*F-Measure (%)*	*Accuracy (%)*
NB	86.81	75.52	80.70	88.10
AC	81.51	74.19	76.33	89.44
SVM	90.02	80.52	84.88	95.33

the five categories, all the three classifiers, SVM, NB and AC, gave the best performance on the "Sports" category and the poorest performance on the "Entertainment" category. In general, the categorization results follow the same trend observed in the training performance.



(a) Precision

(b) Recall

(c) F-measure

(d) Accuracy

Fig. 7. Categorization performance results based on the testing data set.

Table 9 presents 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 achieved the best performance in almost all measures except recall with high precision (87.25%) and accuracy (92.14%), which are comparable to its training performance. NB achieved better performance than AC. However, all classifiers produced low recall values with 77.16%, 70.80% and 68.65% for SVM, AC and NB respectively.

Table 9

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

From the performance evaluation results, it can be observed that SVM generally outperforms the other two classifiers across all categories and NB gives better performance than AC. The recall values are generally lower than the precision values for all the classifiers. The construction of indicative terms for topic categories has great impact on the performance of the proposed approach. For example, the indicative terms for the "Pornography" category are generally easier to identify and more accurate than those of the "Entertainment" category. As a result, the performance on the "Pornography" category is better than that of the "Entertainment" category.

The overall performance of AC is not as promising as that given in (Antonie and Zaiane, 2002) using news articles in their evaluation. This is probably due to the fact that we have adopted indicative terms for forming a very small feature set, which do not contribute to the selection of high quality rules for classification. Nevertheless, the high precision and accuracy achieved by the classifiers of the proposed categorization approach across different categories are promising for topic detection.

4.1.4 Comparison with Document Frequency Based Approach

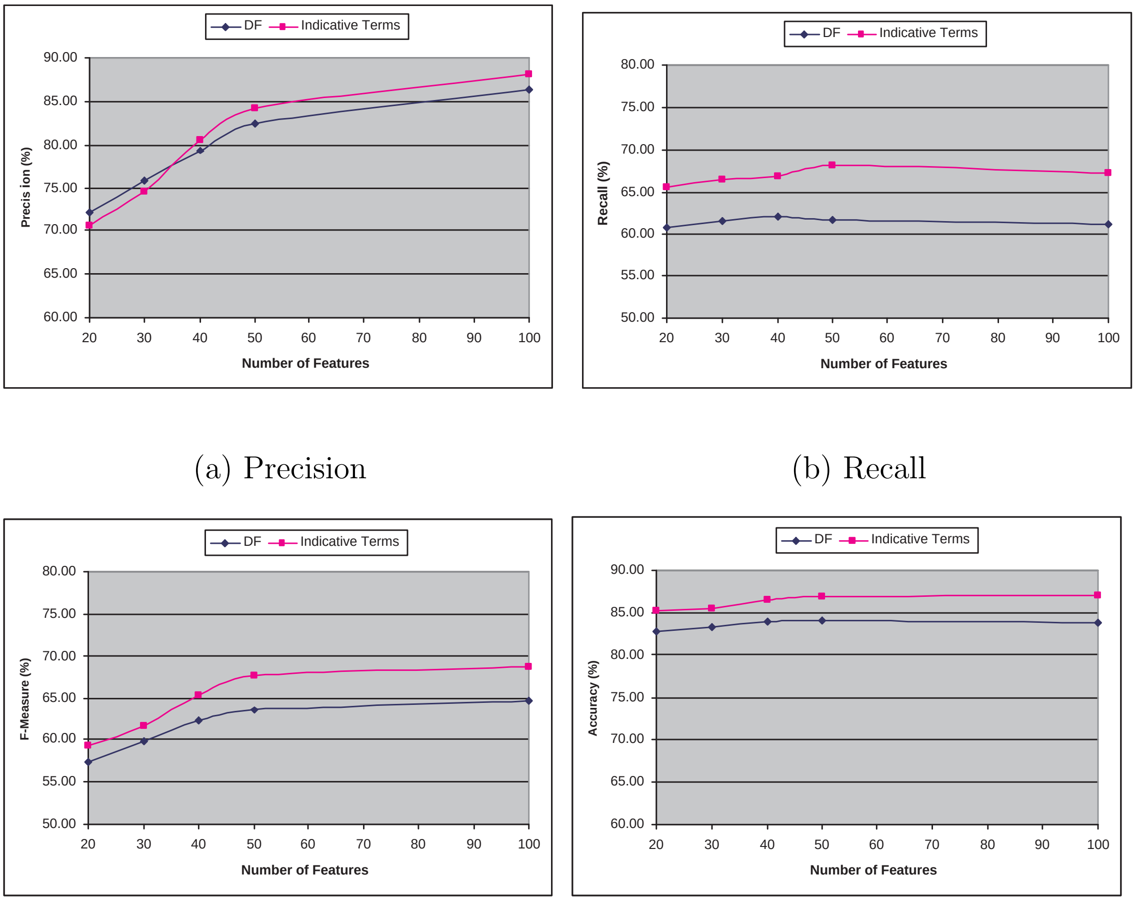
In this section, the performance of using *document frequency* (DF) thresholding feature selection (Apte, Damerau, and Weiss, 1994) approach is compared with our proposed indicative term-based approach for categorization. Document frequency is one of the best feature selection approaches for classification (Yang and Pedersen, 1997).

In this set of experiments, both feature selection approaches are evaluated based on the same specified numbers of different feature sets (i.e., 20, 30, 40, 50 and 100). For the DF-based approach, the features are selected according to the specified number of most representative words, whereas the specified number of most representative terms from the ranked list of indicative terms for each topic is selected as features for the indicative term-based approach. Naïve Bayes is used as the classifier here. The evaluation is based on the macro-average of precision, recall, F-measure and accuracy for all the categories.

Figure 8 shows the topic categorization performance results of Naïve Bayes using different number of feature sets for DF-based and indicative term-based feature selection approaches. As shown in the figure, the indicative term-based approach achieved better performance than the DF-based approach in almost all measures except precision. An important advantage of the indicative terms based feature selection approach is its relative high and stable performance with the limited number of features. This results in high computational efficiency while maintaining a satisfactory classification performance, which is especially important for online topic detection.

5 Instant Message Analysis System

The chat topic detection has been incorporated into an instant message analysis system called IMAnalysis. The IMAnalysis system is part of a client-server based instant message monitoring and analysis system called IMMonitor which is shown in Figure 9. The IMMonitor system comprises three major compo-



(a) Precision

(b) Recall

(c) F-measure

(d) Accuracy

Fig. 8. Performance results of indicative terms based and DF-based approaches.

nents, *IMRecorder*, *IMServer* and *IMAnalysis*. *IMRecorder* is located at each target client (or a target PC) from which IM chat messages are monitored and recorded and subsequently transmitted to *IMServer* in real-time. *IMServer* stores the chat message data into the Chat Log Database. *IMAnalysis*, which is installed at each monitoring client, performs online and offline analysis of the recorded chat message data. For online chat analysis, *IMServer* forwards the received chat data to *IMAnalysis*. For offline chat analysis, chat data from the Chat Log Database are accessed and analyzed by *IMAnalysis*.

Online monitoring allows users to monitor target IM users online at anytime, anywhere through a Web Browser Interface. Offline chat analysis supports the browsing and retrieval of chat session data archived in the Chat Log Database, visualization of social networks, and identification of chat topics. Browsing

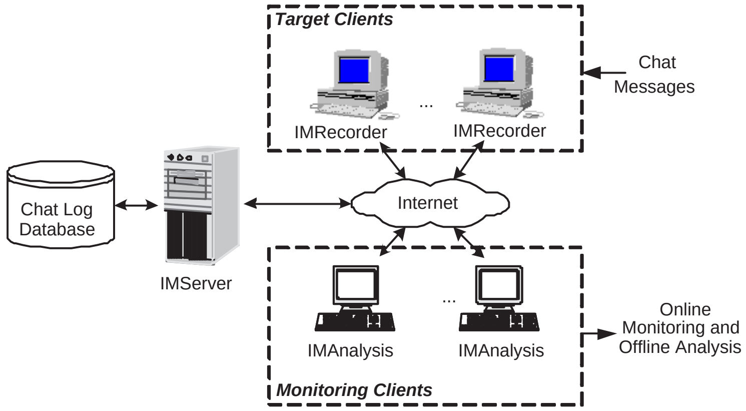


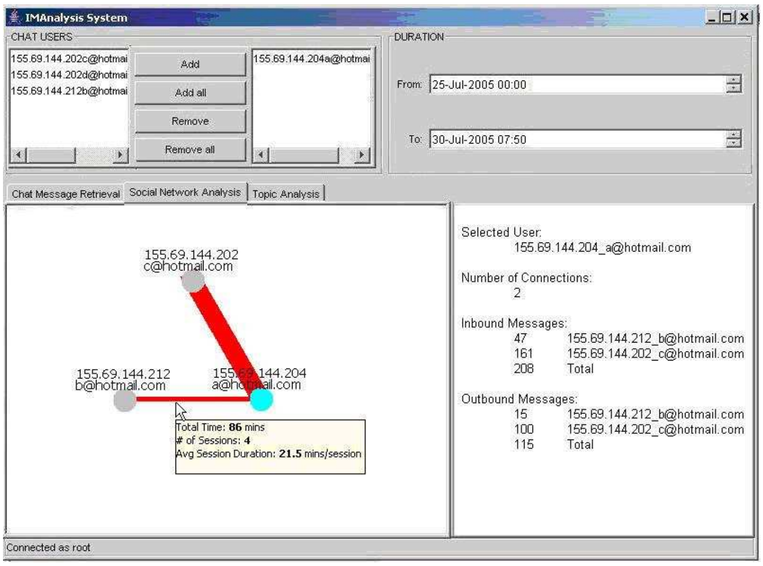
Fig. 9. Instant message monitoring and analysis.

and retrieval support statistical display of chat message data. Social network analysis (Resig, Dawara, Homan, and Teredesai, 2004; Camtepe, Goldberg, Magdon-Ismail, and Krishn, 2005) gives the social interactions between the target monitored users and their contacts from the buddy lists. An example of a social network is shown in Figure 10. After selecting a monitored 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 are represented as links. Thickness of links indicates the amount of messages exchanged between two persons.

Chat topic identification classifies chat sessions into one of the five predefined categories, namely Pornography, Sports, Games, Travel, and Entertainment. By specifying the monitored users and the time duration, the chat session data and its detected topics will be listed as shown in Figure 11. The user can also select one of the sessions to view the details.

6 Conclusions

In this paper, we have studied the conversational format and message characteristics of IM systems from the collection of 33,121 sample chat messages. Based on the chat message analysis results, we have proposed an indicative



From: 25-Jul-2005 00:00

To: 30-Jul-2005 07:50

Total Time: **86** mins
# of Sessions: **4**
Avg Session Duration: **21.5** mins/session

Selected User:
155.69.144.204_a@hotmail.com

Number of Connections:
2

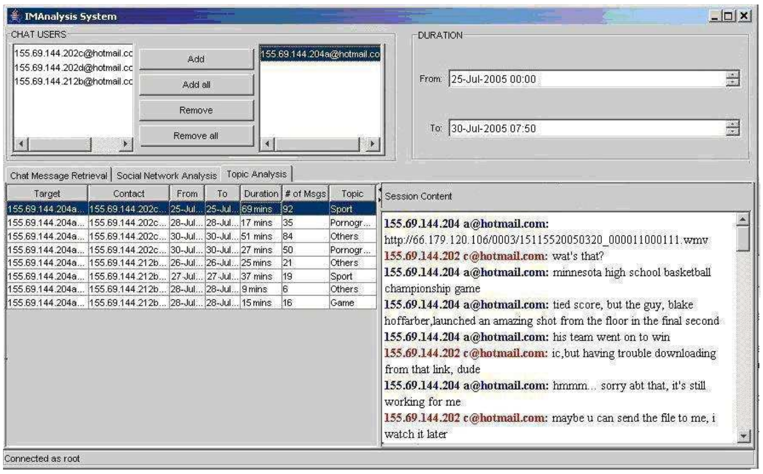
Inbound Messages:

47	155.69.144.212_b@hotmail.com
161	155.69.144.202_c@hotmail.com
208	Total

Outbound Messages:

15	155.69.144.212_b@hotmail.com
100	155.69.144.202_c@hotmail.com
115	Total

Fig. 10. Social network visualization.



From: 25-Jul-2005 00:00

To: 30-Jul-2005 07:50

Target	Contact	From	To	Duration	# of Msgs	Topic
155.69.144.204a...	155.69.144.202c...	25-Jul...	25-Jul...	69 mins	92	Sport
155.69.144.204a...	155.69.144.202c...	28-Jul...	28-Jul...	17 mins	35	Pornogr...
155.69.144.204a...	155.69.144.202c...	30-Jul...	30-Jul...	51 mins	84	Others
155.69.144.204a...	155.69.144.202c...	30-Jul...	30-Jul...	27 mins	50	Pornogr...
155.69.144.204a...	155.69.144.212b...	26-Jul...	26-Jul...	25 mins	21	Others
155.69.144.204a...	155.69.144.212b...	27-Jul...	27-Jul...	37 mins	19	Sport
155.69.144.204a...	155.69.144.212b...	28-Jul...	28-Jul...	9 mins	6	Others
155.69.144.204a...	155.69.144.212b...	28-Jul...	28-Jul...	15 mins	16	Game

155.69.144.204 a@hotmail.com: http://66.179.120.106/0003/15115520050320_000011000111.wmv
155.69.144.202 c@hotmail.com: wat's that?
155.69.144.204 a@hotmail.com: minnesota high school basketball championship game
155.69.144.204 a@hotmail.com: tied score, but the guy, blake hoffarber,launched an amazing shot from the floor in the final second
155.69.144.204 a@hotmail.com: his team went on to win
155.69.144.202 c@hotmail.com: ic,but having trouble downloading from that link, dude
155.69.144.204 a@hotmail.com: hmmm... sorry abt that, it's still working for me
155.69.144.202 c@hotmail.com: maybe u can send the file to me, i watch it later

Fig. 11. Topic distribution display.

term-based categorization approach for chat topic detection which incorporated different techniques such as sessionalization of chat messages and the extraction of features from icon text and URLs for pre-processing. Different classification techniques such as Naïve Bayes, Associative Classification, and Support Vector Machines are employed as classifiers for detecting topics from chat sessions. The performance of the proposed approach has been evaluated based on precision, recall, F-measure and accuracy from the chat message data set collected from the Web. The experimental results have shown that SVM outperforms NB and AC and gives high precision and accuracy values. Moreover, the proposed approach is shown superior than the document frequency based approach and is highly computational efficient as it is able to achieve relatively high and stable performance with just limited number of features, which makes it suitable for online monitoring.

References

AdultFriendFinder.com, 2006. Adult friendfinder - the world's largest sex personal sites. Http://www.adultfriendfinder.com.

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

Apte, C., Damerau, F., Weiss, S., 1994. Automated learning of decision rules for text categorization. ACM Transactions on Information Systems 12 (3), 233–251.

Apte, C., Damerau, F., Weiss, S., 1998. Text mining with decision rules and decision trees. In: Proceedings of the Conference on Automated Learning and Discovery, Workshop 6: Learning from Text and the Web.

Bengel, J., Gauch, S., Mittur, E., Vijayaraghavan, R., June 2004. Chattrack: Chat room topic detection using classification. In: 2nd Symposium on Intelligence and Security Informatics. Tucson, Arizona.

Bingham, E., Kab, A., Girolami, M., 2003. Topic identification in dynamical text by complexity pursuit. Neural Processing Letters 17, 69–83.

Camtepe, S., Goldberg, M., Magdon-Ismail, M., Krishn, M., Feb. 2005. Detecting conversing groups of chatters: a model, algorithms, and tests. In: IADIS International Conference on Applied Computing. Algarve, Portugal.

Cohen, W., Singer, Y., 1996. Context-sensitive learning methods for text categorization. In: 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'96). pp. 307–315.

Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society 39, 1–38.

Elnahrawy, E., 2002. 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.

ICQ.com, 2006. ICQ - Community, People Search and Messaging Service. Http://www.icq.com.

IRC.org, 2006. IRC.org - home of IRC. Http://www.irc.org.

Jain, A. K., Dubes, R. C., 1998. Algorithms for Clustering Data. Prentice Hall.

Joachims, T., 1998. Text categorization with support vector machines: Learning with many relevant features. In: European Conference on Machine Learning (ECML). Springer, Berlin, Germany, pp. 137–142.

Kolenda, T., Hansen, L. K., Larsen, J., 2001. 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.

Lang, K., 2006. 20 Newsgroups Data Set. Http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/news20.html.

Lee, R., 2003. Teenage life online: the rise of the networked generation. In: Youth.Net Conference. Singapore, http://www.pewinternet.org/ppt/200320Conference.ppt.

Masand, B., Linoff, G., Waltz, D., 1992. Classifying news stories using memory based reasoning. In: 15th Ann International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'92). pp. 59–64.

Moulinier, I., 1997. Is learning bias an issue on the text categorization problem? technical report LAFORIA-LIP6, Universite Paris VI.

Moulinier, I., Raskinis, G., Ganascia, J.-G., 1996. Text categorization: a symbolic approach. In: 5th Annual Symposium on Document Analysis and Information Retrieval. pp. 87–99.

MSN.com, 2006. MSN Messenger version 7.5. Http://messenger.msn.com/.

Ng, H. T., Goh, W. B., Low, K. L., 1997. Feature selection, perception learning, and a usability case study for text categorization. In: 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97). pp. 67–73.

NIST, 2006. Reuters Corpora @ NIST. Http://trec.nist.gov/data/reuters/

reuters.html.

ODP, 2006. ODP - Open Directory Project. Http://dmoz.org/.

online game forum, J., 2006. jolt.com.uk public forums - powered by vBulletin. Http://forums.jolt.co.uk.

Resig, J., Dawara, S., Homan, C., Teredesai, A., Aug. 2004. Extracting social networks from instant messaging populations. In: Workshop on Link Analysis and Group Detection (LinkKDD 2004). Seattle, WA.

Tencent.com, 2006. Tencent homepage. Http://www.qq.com.

Thomas, K., 2001. Kids need enduring smarts with instant messaging. Http://www.usatoday.com/ tech/news/2001-07-10-instant-messaging-safety.htm.

Timothy, 2003. How to cheat on / exploit Bejeweled (theoretically). Http://timothy.mess.be/Bejeweld-Exploit.html.

Tzeras, K., Hartman, S., 1993. 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.

Ugroups.com, 2006. UGroups: free web access to the Usenet newsgroups and forums. Http://www.ugroups.com.

van Rijsbergen, C. J., 1979. Information Retrieval. Butterworths, London.

Wiener, E. D., Pedersen, J., Weigend, A., 1995. A neural network approach to topic spotting. In: Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval (SDAIR'95). Nevada, Las Vegas, pp. 317–332.

Wolf, M., 2003. Cyber Brats: Bullies Who Taunt Their Peers with the Click of a Mouse. Http://www.parenthood.com/articles.html?article_id=4335.

Yahoo.com, 2006. Yahoo! Messenger with Voice. Http://messenger.yahoo.com/.

Yang, Y., 1994. Expert network: effective and efficient learning from human decisions in text categorization and retrieval. In: 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94). pp. 13–22.

Yang, Y., Chute, C. G., 1994. An example-based mapping method for text categorization and retrieval. ACM transaction on Information Systems (TOIS) 12 (3), 252–277.

Yang, Y., Liu, X., 1999. A re-examination of text categorization methods. In: 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99). pp. 42–49.

Yang, Y., Pedersen, J., 1997. A comparative study on feature selection in text categorization. In: 14th International Conference on Machine Learning (ICML-97). pp. 412–420.

Young, W. S., Sycara, K., Jan. 2004. Text clustering for topic detection. technical report CMU-RI-TR-04-03, Robotics Institute, Carnegie Mellon University.