# SMIRK: SMS Management and Information Retrieval Kit

Ibrahim Baggili*, Ashwin Mohan, and Marcus Rogers

Zayed University, Abu Dhabi, United Arab Emirates*
Purdue University, West Lafayette IN 47901, USA
baggili@gmail.com, {mohana,rogersmk}@Purdue.edu

**Abstract.** There has been tremendous growth in the information environment since the advent of the Internet and wireless networks. Just as e-mail has been the mainstay of the web in its use for personal and commercial communication, one can say that text messaging or Short Message Service (SMS) has become synonymous with communication on mobile networks. With the increased use of text messaging over the years, the amount of mobile evidence has increased as well. This has resulted in the growth of mobile forensics. A key function of digital forensics is efficient and comprehensive evidence analysis which includes authorship attribution. Significant work on mobile forensics has focused on data acquisition from devices and little attention has been given to the analysis of SMS. Consequentially, we propose a software application called: SMS Management and Information Retrieval Kit (SMIRK). SMIRK aims to deliver a fast and efficient solution for investigators and researchers to generate reports and graphs on text messaging. It also allows investigators to analyze the authorship of SMS messages.

**Keywords:** Cyber forensics, cellular phone forensics, forensic tools, SMS authorship attribution, post-hoc forensic analysis.

## 1 Introduction

Research has shown that on the Internet, people turn to some sort of textual communication setting to maintain their relationships (i.e., social networks) in a relatively safe environment. One such textual communication environment is SMS messaging on cellular phones. Mobile phones are widely used in the United States. In the first six months of 2006, the Cellular Telecommunication and Internet Association (CTIA) stated that there were 262.7 million U.S. wireless subscribers and wireless communication had penetrated more than 84% of the total U.S. population. The CTIA also reported that 75 billion SMS messages were sent per month, from 7.2 billion in the first six months of 2005 and 12.2 Million in 2000. This is an average of 300 messages per individual per month.

An SMS communication environment permits higher levels of visual anonymity when compared to face-to-face communication [1]. This anonymity can be misused by criminals to their own advantage. An example of the exploitation of visual anonymity is the case of Danielle Jones. Jones disappeared on the 18th of June in 2001

S. Goel (Ed.): ICDF2C 2009, LNICST 31, pp. 33–42, 2010.

when some text messages were sent from her phone. There was a suspicion among Law Enforcement officials that some of the messages were not actually written by her. In the Jones case, linguistic analysis concluded that the messages were more likely written by her uncle. Another case is the one of Jenny Nicholl, who disappeared in 2005. In the Nicholl case, linguistic analysis showed that text messages sent from her cellular phone were most likely to have been written by her ex lover [2][3].

The prevalence of SMS in social and business communications coupled with the need for visual anonymity justifies the imminent need for investigators to strongly consider SMS data for digital evidence. In research, little attention has been given to the analysis of SMS for use in investigations. Researchers and corporations gear their attention towards the acquisition and simple reporting of digital evidence from mobile phones as shown in the most encompassing NIST report on cell phone forensic tools [4]. Currently, there are no software packages from vendors that perform data analysis of SMS through linguistic techniques even though the need for this is apparent. In both the cases discussed, data analysis using linguistic techniques played a vital role in the prosecution and incarceration of individuals through authorship attribution.

Other forms of linguistic analysis have been discussed in literature in relation to SMS such as stemming and phonetic substitution [5]. However, these methods were not discussed in an applied forensic context. These types of linguistic tools can be useful during the analysis phase of forensic investigations. Therefore, we propose that SMIRK should include linguistic analysis tools for SMS.

## 2 Related Work

There has been related research work in computer based email forensics. This is of considerable importance since email, just like SMS, is a text-based communication medium.

Viegas, Golder and Donath [6] developed *Themail*, an application that can visualize email archives along timelines by using content to portray individual relationships, both in terms of general trends, themes and detail oriented exploration. O. de Vel et al. in [7] used email document features like structural characteristics and linguistic patterns along with support vector machines to mine email content in aggregate and multi-topic email documents and used the information for author categorization. O.de Vel and others [8] also performed experiments using a corpus of email documents to attribute gender and language background for the authors.

EMT or Email Mining Toolkit [9] developed at the Columbia Intrusion detection lab is a popular data mining tool that has been used by law enforcement agencies. It provides features to look at content and flow of email attachments for individuals and aids in detecting anomalous behavior or common trends exhibited by a group of users or “social cliques”. The behavior of a single stationary user account or multiple similar ones can be modeled as histograms and may be supportive evidence to investigations.

## 3 Overview of Problems in SMS Evidence Analysis

Using the available literature and the authors’ experience of using various cellular phone forensic tools the following problems were identified when analyzing SMS data for evidence.

### 3.1 Problem 1: Proprietary File Formats

Most of the current forensic tools extract data from a cellular phone to a text file. The SMS messages in the cellular phone are stored with their content and attributes in this manner. However, these tools perform extraction in a number of different file formats. It was observed that some of these text files were tab delimited, others comma delimited and there were a few which did not conform to any delimiting standards and/ or included superfluous information in the form of header/footer data.

The researchers concluded a *post-hoc SMS data analysis tool should be capable of importing different text file formats without any significant effort.*

### 3.2 Problem 2: Lack of Linguistic Tools for Investigative Purposes

A major problem that exists in the construction of cellular phone forensic tools is that they do not incorporate linguistic analysis tools and techniques. This section outlines the problems related to linguistics when dealing with SMS as a source of digital evidence. If these problems are solved in a feasible manner, they could yield faster turn-around rates in investigations where SMS might be used as a source of digital evidence.

#### 3.2.1 Lack of SMS Authorship Attribution Tools

The introduction to the paper outlined the importance of using SMS authorship attribution in real investigative cases. To this end, the NIST report on Cellular Phone Forensic Tools [4] was reviewed along with the various software packages available on the market; it became apparent that none of the tools take into account analysis of SMS messages towards achieving the capability of identifying the author of an SMS message.

The researchers concluded a *post-hoc SMS data analysis tool should be capable of attributing authors to SMS messages.*

#### 3.2.2 SMS Written Language

Since SMS has gained worldwide popularity, cellular operators are providing users with the ability to type text messages in multiple languages. This is problematic when dealing with authorship attribution, since most authorship attribution systems are language dependent. In order to tackle this problem an N-gram based approach for authorship attribution was studied in a paper that is in publication by the authors. To discuss the method is beyond this paper's scope. However, the results from the study helped the authors in creating a method for SMS authorship in a language independent form.

The researchers concluded that a *post-hoc SMS data analysis tool should have the capability of attributing authors to SMS messages in a language independent manner.*

#### 3.2.3 Short Form Issues in SMS

If you have used SMS extensively, you notice that people do not write SMS messages the way they write e-mails, or other forms of written communication. Due to the limited size of SMS messages, people have started to replace words with numbers. For instance, the word “forgot” has been seen written as “4got” in SMS messages [5]. This is a problem because linguistic techniques for finding verbs and nouns depend on English language characteristics which get lost in the use of such short-hands. It then becomes important to be able to transform these short-hands back into their regular language counterparts. This may reveal information key to a case. For example, the word LHOS in short hand in an SMS message may imply “Let’s have online sex” [10], which could have significant implications in a child pornography case.

The researchers concluded that a *post-hoc SMS data analysis tool should be able to convert SMS messages to their natural English language equivalents.*

#### 3.2.4 Noun/Verb Detection

In real investigation cases, it would be useful to identify the subject or meaning of a series of SMS messages since it would help clarify the intent of the person under observation. A first step in this direction is to detect the nouns and verbs that constitute these messages since nouns provide object of interest (e.g., a Place like London) and verbs can signify the actions desired by the author (e.g., to move).

The researchers concluded that a *post-hoc data analysis tool should have the ability to find the nouns and verbs in a set of SMS messages tagged as being part of communication of interest.*

### 3.3 Visualization and Reporting of SMS Analysis

Data analysis cannot be useful without providing adequate capability to report the results in textual or graphical forms. These methods of reporting are advantageous when investigators are searching the data for evidence. Most forensic tools provide techniques for reporting as it is an important step in the forensic process; however, they are not extensive and do not cater for SMS message analysis. To speed up the cellular phone forensic analysis process, it is important to empower investigators with the ability to generate reports, graphs and pie charts on SMS.

The researchers concluded that a *post-hoc data analysis tool would have the capability to generate reports on SMS and visualization aids in the form of pie charts and bar graphs for the results of the cellular phone forensic analysis and make these available for future use.*

## 4 Overview of SMIRK

SMIRK is a field and offline analysis software package for public agencies and the scientific community developed at Purdue University during the spring of 2009. The prototype and release versions were created using the Visual C# .NET programming language. Any computer capable of running the .NET Framework is capable of executing SMIRK. SMIRK will achieve the objectives which were outlined in Section 3 and have been summarized in the table below.

**Table 1.** Post-hoc SMS data analysis tool solutions to SMS messaging analysis problems

| Problem | Solution /Capabilities |
|---|---|
| Interoperability with different exported SMS data sources from various mobile acquisition tools | *Import different text file formats without any significant effort.* |
| Reporting of identifiable patterns in SMS message contents and attributes | *Generate reports on SMS and visualization aids in the form pie charts and bar graphs for the results of the cellular phone forensic analysis and to make these available for future use* |
| A graphical representation of SMS messaging patterns | |
| The prediction of SMS authors independent of language | *Attribute authors to SMS messages in a language independent manner* |
| Capability to morph between regular English corpuses and their SMS equivalent using phonetic substitution | *Convert SMS messages to their natural English language equivalent* |
| Noun and verb boundary detection and reporting for SMS messages | *Find the nouns and verbs in a set of SMS messages tagged as being part of communication of interest* |

These objectives served as the template to design the different functional modules of the application. It was assumed that SMIRK would be used in congruence with mobile forensics acquisition tools, which extract raw data from cell phones. An intuitive graphical user interface was created to enable users to import raw data into the application with minimal effort.

# 5 SMIRK Modules

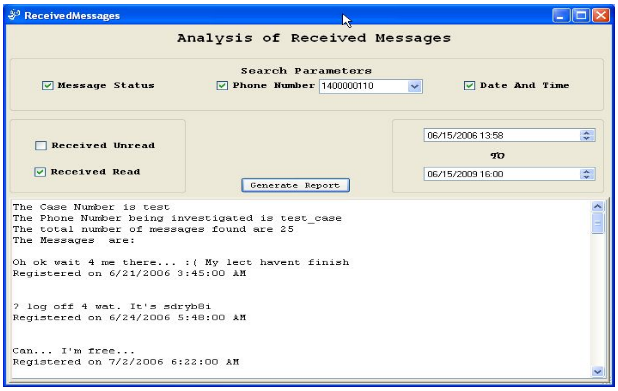
## 5.1 Importing

A graphical data import wizard was created as mentioned in Section 4. The wizard allows the user to import raw data stored in text files. It provides options to define the type of formatting used in the file. This can be one of the standard file delimited formats like tab, comma or space delimitation, or a specific user defined delimiting character. The user also enters the list of status values corresponding to the message status flags. This information along with the delimiting character is used to parse the data into a database stored in Random Access Memory (RAM).

SMIRK is independent of whichever mobile forensics acquisition tool is used for raw data extraction, as long as the user knows the format in which the data is stored.

## 5.2 Reporting

Importing data without proper means for analysis is not useful. SMIRK offers the ability to generate reports based on the imported data file. Reports in SMIRK focus on messaging patterns. These reports are available for both received and sent messages.



ReceivedMessages

Analysis of Received Messages

Search Parameters

☑ Message Status ☑ Phone Number 1400000110 ☑ Date And Time

□ Received Unread

☑ Received Read

06/15/2006 13:58

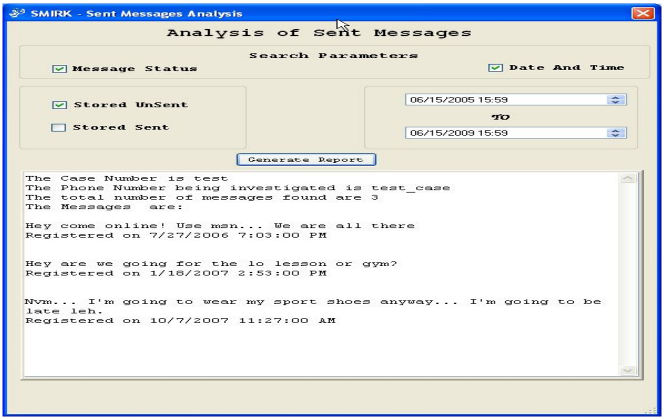
TO

06/15/2009 16:00

```
The Case Number is test
The Phone Number being investigated is test_case
The total number of messages found are 25
The Messages  are:

Oh ok wait 4 me there... :( My lect havent finish
Registered on 6/21/2006 3:45:00 AM


? log off 4 wat. It's sdryb8i
Registered on 6/24/2006 5:48:00 AM


Can... I'm free...
Registered on 7/2/2006 6:22:00 AM
```

**Fig. 1.** Received Messages Report

The interface for the reports permit the user to select options to refine processing based on the destination phone number (only for sent messages), the message status (read/unread for received messages and sent/not sent for sent messages) and date and time of the communication with respect to the number being investigated.



SMIRK - Sent Messages Analysis

Analysis of Sent Messages

Search Parameters

☑ Message Status ☑ Date And Time

☑ Stored UnSent

□ Stored Sent

06/15/2005 15:59

TO

06/15/2009 15:59

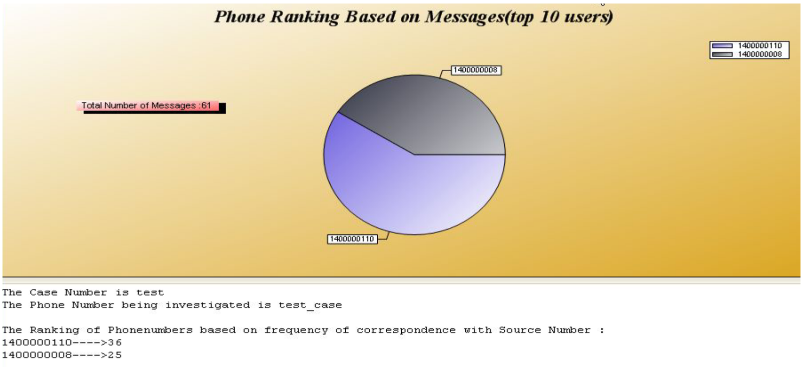
```
The Case Number is test
The Phone Number being investigated is test_case
The total number of messages found are 3
The Messages  are:

Hey come online! Use msn... We are all there
Registered on 7/27/2006 7:03:00 PM


Hey are we going for the lo lesson or gym?
Registered on 1/18/2007 2:53:00 PM


Nvm... I'm going to wear my sport shoes anyway... I'm going to be
late leh.
Registered on 10/7/2007 11:27:00 AM
```

**Fig. 2.** Sent Messages Report

A standard report consists of the case number, the investigated phone number, and each of the SMS communications along with their corresponding destination numbers, date, time and message content, as shown in Figure 1 and Figure 2.

### 5.3 Graphing

A limited capability for graphical representation has been provided in SMIRK. There are two types of graphs employed. The first is a date-time bar chart for the reports on messaging patterns as described in section 5.2. The options provided to filter the results are the same as for those of the corresponding reports. The second type of graph is a pie chart that details frequency of message flow. Figure 3 is an example of a pie chart created by SMIRK on messaging patterns.

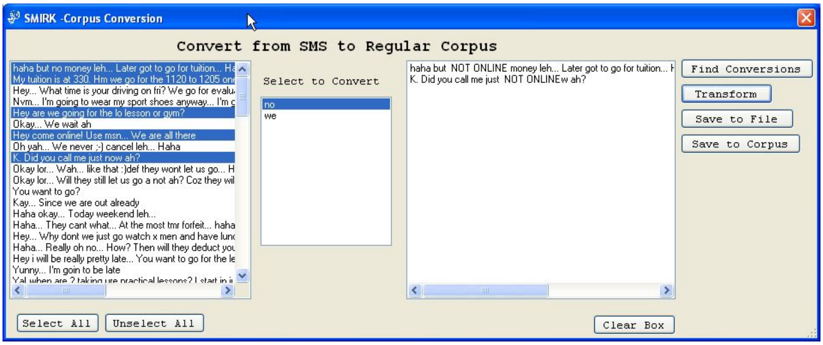


```
The Case Number is test
The Phone Number being investigated is test_case

The Ranking of Phonenumbers based on frequency of correspondence with Source Number :
1400000110---->36
1400000008---->25
```

**Fig. 3.** Text Messages Pie Chart

## 5.4 Dataset Conversion

SMRIK provides the capability to convert between an SMS corpus and its Standard English equivalent. SMIRK parses SMS content by applying a conversion algorithm to the target dataset based on phonetic similarity rules [11] assisted with a preordered list of some popular substitutions which are in common use. The user is allowed to decide which conversions they want to perform on the SMS data, based on any prior information they might have. The new corpus is saved in a database stored in RAM and can now also be used when performing noun/verb boundary detection which is discussed in section 5.6. Figure 4 illustrates the corpus conversion.



haha but no money leh... Later got to go for tuition... Ha
My tuition is at 330. Hm we go for the 1120 to 1205 one
Hey... What time is your driving on fri? We go for evalu
Nvm... I'm going to wear my sport shoes anyway... I'm g
Hey are we going for the lo lesson or gym?
Okay... We wait ah
Hey come online! Use msn... We are all there
Oh yah... We never ;-) cancel leh... Haha
K. Did you call me just now ah?
Okay lor... Wah... like that :)def they wont let us go... H
Okay lor... Will they still let us go a not ah? Coz they wil
You want to go?
Kay... Since we are out already
Haha okay... Today weekend leh...
Haha... They cant what... At the most tmr forfeit... haha
Hey... Why dont we just go watch x men and have lunc
Haha... Really oh no... How? Then will they deduct you
Hey i will be really pretty late... You want to go for the le
Yunny... I'm goin to be late
Yal when are ? taking ure practical lessons? I start in ir

no
we

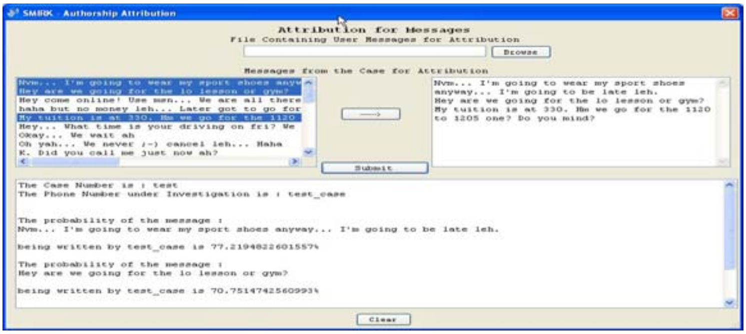
haha but NOT ONLINE money leh... Later got to go for tuition... H
K. Did you call me just NOT ONLINEw ah?

**Fig. 4.** Corpus Conversion

## 5.5 Authorship Attribution

Identifying the author for a message can be of considerable importance for investigators and can play a major role in criminal proceedings. SMIRK implements an authorship identification algorithm based on the concept of N-grams [12]. When an unidentified message is presented by the user through the graphical interface, its n-gram tokens are parsed, and compared with those of the messages already existing in the SMS data file imported into the application. Based on similarity scoring techniques [13], it is displayed whether the unattributed messages have been written by the author of the messages in the imported data file and if so with what probability of a match.

The N-grams look at patterns in the user text rather than style pattern associated with a particular grammar, hence it is language independent. The assumption for attribution is that the messages in the imported corpus are tagged as being written by the individual whose cellular phone is being investigated, and there is no supposition otherwise. Figure 5 is an example of how authorship attribution is performed in SMIRK.


SMIRK - Authorship Attribution
Attribution for Messages
File Containing User Messages for Attribution
Browse
Messages from the Case for Attribution
--->
Submit
Clear

```
Nvm... I'm going to wear my sport shoes anyw
Hey are we going for the lo lesson or gym?
Hey come online! Use msn... We are all there
haha but no money leh... Later got to go for
My tuition is at 330. Hm we go for the 1120
Hey... What time is your driving on fri? We
Okay... We wait ah
Oh yah... We never ;-) cancel leh... Haha
K. Did you call me just now ah?
```

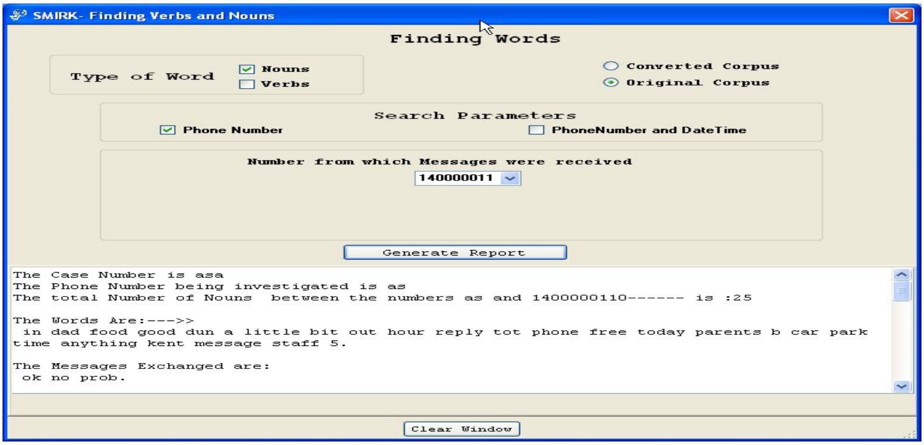
```
Nvm... I'm going to wear my sport shoes
anyway... I'm going to be late leh.
Hey are we going for the lo lesson or gym?
My tuition is at 330. Hm we go for the 1120
to 1205 one? Do you mind?
```

```
The Case Number is : test
The Phone Number under Investigation is : test_case


The probability of the message :
Nvm... I'm going to wear my sport shoes anyway... I'm going to be late leh.

being written by test_case is 77.2194822601557%

The probability of the message :
Hey are we going for the lo lesson or gym?

being written by test_case is 70.7514742560993%
```

**Fig. 5.** Authorship Attribution

### 5.6 Noun/Verb Boundary Detection

Nouns and Verbs in SMS can be useful to understand the meaning and context of a message. This is helpful when trying to construct motive behind a communication. The application provides the limited ability to perform boundary detection of verbs and nouns in messages. The messages are selected by the user from the imported dataset using options in the module. The user is allowed to choose the number from



Type of Word ☑ Nouns ☐ Verbs

○ Converted Corpus
● Original Corpus

Search Parameters
☑ Phone Number ☐ PhoneNumber and DateTime

Number from which Messages were received: 140000011

```
The Case Number is asa
The Phone Number being investigated is as
The total Number of Nouns  between the numbers as and 1400000110------ is :25

The Words Are:--->>
 in dad food good dun a little bit out hour reply tot phone free today parents b car park
time anything kent message staff 5.

The Messages Exchanged are:
 ok no prob.
```

**Fig. 6.** Noun/Verb Detection

which the messages were received and the date and time it was sent. If a corpus conversion has been performed (refer to section 4.4), the user can choose to detect the verbs and nouns in the transformed corpus rather than in the original SMS file. A preordered list of commonly used verbs/nouns is used along with an algorithm based on lexical analysis to perform the detection. Figure 6 shows the Noun/Verb Detection implemented in SMIRK.

## 6 Conclusions/Future Work

The researchers aimed to create one of the first tools that could be used for the analysis of SMS messages. Using SMIRK, the researchers achieved the aforementioned goals. With SMIRK, investigators and researchers will be able to closely examine SMS content, messaging patterns and authorship attribution.

Currently, SMIRK has been inquired about by cyber crime investigators that are interested in using it in cases where SMS may be valuable to them. It is important to note that the next phase of SMIRK would be extensive field testing and feedback from the forensic community. The researchers also plan on adding more graphs representing the chronology of messages. Finally, more accurate methods and algorithms for authorship attribution of SMS messages are still being investigated.

## References

1. McKenna, K., Green, A., Gleason, M.: Relationship formation on the Internet: What's the big attraction. Journal of Social Issues 58(1), 9–31 (2002)
2. The Independent, Dr Tim Grant: How text-messaging slips can help catch murderers (retrieved November 29) (2008), http://www.independent.co.uk/opinion/commentators/dr-tim-grant-how-textmessaging-slips-can-help-catch-murderers-923503.html
3. Cellular-news, SMS as a tool in murder investigations (retrieved November 24) (2008), http://www.cellular-news.com/story/18775.php
4. Ayers, R., Jansen, W., Delaitre, A., Moenner, L.: Cell Phone Forensics Tools: An Overview and Analysis Update, NIST Interagency Report (IR) 7387 (February 2007)
5. Lee, F.: SMS Shortform Identification and Codec. National University of Singapore Thesis (2005), http://www.comp.nus.edu.sg/~rpnlpir/downloads/corpora/smsCorpus/mingFungLeeThesis.pdf (retrieved May 21, 2009 )
6. Viégas, F.B., Golder, S., Donath, J.: Visualizing Email Content: Portraying Relationships from Conversational Histories. Long paper, CHI 2006 (2006)
7. de Vel, O., et al.: Mining E-Mail Content for Author Identification Forensics. SIGMOD Record 30(4), 55–64 (2001)
8. de Vel, O., Corney, M., Anderson, A., Mohay, G.: Language and Gender Author Cohort Analysis of E-mail for Computer Forensics. In: Digital Forensic Research Workshop, Syracuse, NY, August 7-9 (2002)
9. Stolfo, S.J., Hershkop, S., Wang, K., Nimeskern, O., Hu, C.: Behavior Profiling of Email. In: Proc. of NSF/NIJ Symposium on Intelligence & Security Informatics (2003)

10. Netling, The netlingo list of acronyms & text messaging shorthand (retrieved May 21) (2009), http://www.netlingo.com/acronyms.php
11. UzZaman, N., Khan, M.: T12: An Advanced Text Input System with Phonetic Support for Mobile Devices. In: 2nd International Conference on Mobile Technology, Applications and Systems, pp. 1–7 (2005)
12. Shannon, C.E.: A Mathematical Theory of Communication. Bell System Technical Journal 27, 379–423, 623–656 (1948)
13. SimMetrics: Open Source Similarity Measure Library, http://www.dcs.shef.ac.uk/~sam/simmetrics.html