

# Information Hiding System for Online Conversation

Weihui Dai<sup>1</sup>, Yue Yu<sup>1</sup>, Yonghui Dai<sup>2</sup>, and Xingyun Dai<sup>1</sup>

<sup>1</sup> School of Management, Fudan University, Shanghai 200433, China

Email: {whdai, 072025045, 082025041}@fudan.edu.cn

<sup>2</sup> School of Software, Fudan University, Shanghai 200433, China

Email: dyh822@163.com

**Abstract**—In recent years, information hiding technology has developed from the focus of digital watermark into secret communication. This paper explores an information hiding system for online conversation based on the statistical relations between words. It establishes a rule system to ensure the secret information to be hid into text messages exactly, and provide enough flexibility to modulate those messages close to nature language with the chosen conversation topic.

**Index Terms**—information hiding, online conversation, intelligent composition, nature language

## I. INTRODUCTION

Information hiding is different from encrypting with its intention to avoid the intrusion or discovering to the hiding data, other than to restrict the data access [1]. Moreover, it should avoid any damages in the hiding data, with the capability of immunity from those regular operations, such as transformation and compression.

High transmission efficiency, low resource occupancy and intelligible meaning make text message as the most commonly used type of media in our daily communication, especially in online conversation. However, due to the restriction of redundant information as well as its alterability in manual operation, text message is difficult to hide secret information effectively and reliably [1]-[3].

Since 1990s, information hiding technology based on text media has developed from the focus of digital watermark into secret communication. A famous software system designed to hide encrypted data into text document was introduced by M. T. Chapman in 1997 [2]. In 2003, Chinese researchers X. X. Niu and Y.X. Yang presented a new algorithm for the communication of information hiding based on text media[3], which hides meaningful information into a stegotext by technical changing in its cover message, such as the character characteristics (size, space, color, front, attribute, intensity, etc.) [4]-[7], sentence structure [8] and other statistical characteristics [9]. Under the circumstances, the invader doesn't know the cover message whether hides other information. Even he knows, it is difficult to distill or wipes off the hidden information.

Up to now, various approaches and algorithms have already been explored in this field. Reference [10] summarized the current methods which utilize the space between words, rows and punctuations to realize information hiding. Reference [11] proposed an algorithm to hiding information of the cipher text by the changes in text front and text color. Other approaches and algorithms based on character front [4], character color [5], character intensity [6], structure of the natural language [7] [12], attributes of the HTML markup [8] and the statistical characteristics of characters [9] were presented as recent explorations. On the other side, steganalysis technology has been successfully developed to detect and find the hiding encrypted information in covertext by analyzing its redundancy [1][13].

By a comprehensive analysis of existed researches, we can draw a key problem that the capability of immunity from regular operations, such as formatting, compressing and sometimes manual altering operation is expected to be further explored so far in this field. At the same time, reduce of redundancy in covertext is to be improved to both ensure the transmission efficiency and antagonize the steganalysis.

This paper presents an information hiding system for online conversation to realize secret communication based on the relations between words in the conversation. It has the capability of immunity from regular operations, such as formatting, compressing and sometimes manual altering operation in text size, front, color and the space between words. With the help of the intelligent composition technology, this system can ensure the secret information to be hid into text messages exactly, and provide enough flexibility to modulate those messages close to nature language with the chosen conversation topic.

## II. INFORMATION HIDING BASED ON TEXT MESSAGE

Information hiding is a technology that hides meaningful information to a Cover C to get the Stego Cover S. In order to increase the offense difficulty, we can combine encrypting technology with information hiding technology. That is to encrypt the Message M to get the cryptograph information M', and then hide M' to Cover C. In this way, even though the invader wants to get the message, he should first detect the existence of the information, and know how to distill M' from the secret cover S, then decrypt M' to recovered message M.

---

This research was supported by Shanghai Leading Academic Discipline Project (No.B210).

Different from the analog signal method based on image, sound or video, the text cover doesn't use signal disposing model. It is more difficult to realize the information hiding based on text. The simplest method of information hiding is to select the cover first, adopt given rules to add the phraseological or spelling mistakes, or replace with synonymy words. For example, Textto [14] setups some sentence structure in advance, fills in the empty location by arranged words, and then the text doesn't have phraseological mistakes, but have some word changes or morphology mistakes.

This complex method is to produce secret text according to hiding information, and needn't select covers in advance. A more complex method is to use nature language disposal to make the secret text more nature. TextHide [15] hides the information in the manner of text overwriting and words' selection. NiceText [2] may imitate the given sample's writing style to create the text of approximate nature language, embed hiding information in the creating process. Another method is to read in all the character in its coding mode, these coding numbers exist in integer form, and haven't any redundancy, to express this bunch of number in its it stream, through some transform such as wavelet transform, FFT transform, DCT transform to get the signal that has redundancy, and then to disguise the text in the redundancy space.

### III. INFORMATION HIDING SYSTEM FOR ONLINE CONVERSATION

#### A. Statistical Relations in Nature Languages

The nature languages such as English, Chinese, French and German which we use in daily life are all information sources that composed by a group of signal collection. These signals are dependent, we can use statistical model to approach it. If we take English language as the example, the signal collect is selected as letters and spaces, and its output of information is a letter sequence. Because there is not the same probability of the English letters to buildup words, these letters have strict dependent relation. Table I shows the letters' probability [16].

The signal sequences have finite dependent relation of some information sources, that is, the probability of the signal at any times is related to some former signal. In most cases, the sequence of nature languages can be approximated as a signal from Markov Information Source, which has been applied to many areas of language analysis and machine training [17][18].

If we only consider the appearing frequency of the words and the dependent relation between words, we can use one-rank Markov information source to describe the source of English words. The creating sequence from Markov information source represents an actual English article. Reference [3] presented a concise method that formed approximate 2-rank Markov information source of English word sequence. In this paper, we utilize the statistical relations to hide secret information into the conversation sentences.

TABLE I. LETTERS' PROBABILITY TABLE

Letter	Probability	Letter	Probability
space	0.1859	N	0.0574
A	0.0642	O	0.0632
B	0.0127	P	0.0152
C	0.0218	Q	0.0008
D	0.0317	R	0.0484
E	0.1031	S	0.0514
F	0.0208	T	0.0796
G	0.0152	U	0.0228
H	0.0467	V	0.0083
I	0.0575	W	0.0175
J	0.0008	X	0.0013
K	0.0049	Y	0.0164
L	0.0321	Z	0.0005
M	0.0198		

#### B. Sample Text and Rule System

In online conversation, there are usually some topics to be focused. In order to produce the nature language close to specific topic, we select some sample text related to this topic to determine the statistical relations between their words, and then fix a statistic algorithm to create the same relation data. Based on those data, we establish a rule system to ensure the secret information to be hid into text messages exactly, and provide enough flexibility to modulate those messages close to nature language and related to the conversation topic. By this way; thereof, secret communication is set up between the sender and the receiver. They just need to confirm the sample text and the statistic algorithm before their conversation.

#### C. System Framework and Data Processing

The design objective of this system is to provide a software tool for information hiding in online conversation. Fig.1 shows the system framework and its data processing.

Before the conversation, the sender and receiver may select a certain sample text related to specific conversation topic and use it only for this conversation. Instead of this, they can also share the same rule system which has been established by pre-processing. In this conversation, secret information can be converted to binary data. After compressed and encrypted (with DES or 3DES algorithm), those data are hid into the text message by the sender. This process is supported by the intelligent composition module, which provides heuristic technique for the composition of text message and makes it close to the nature language. For example, if the secret information is "Meet you at East Gate on 13:00pm", and you have chosen the conversation topic as "University Student Life", the function of intelligent composition will be illustrated as followings:

Conversation topic: University Student Life  
Secret information:

“Meet you at East Gate on 13:00pm”

ASCII code:

```
4D 65 65 74 20 79 6F 75-20 61 74 20 45 61 73 74 Meet you at East
20 47 61 74 65 20 6F 6E-20 31 33 3A 30 30 70 6D Gate on 13:00pm
0D 0A 00 00 00 00 00 00-00 00 00 00 00 00 00 .....
```

Binary data (compressed and encrypted):

“7B13 6A84 07EA 9142 C571 52C2 E1C6 716C 1B41”

Keywords: “Class”, “Sport”, “Food”, “news”.....

If you select the keywords as “Class”, a group of coupled words will be displayed:

- “physics class”, (“Biology textbook”,.....)
- “optical experiment”, (“Subject Class”,.....)
- “third class”, (“taxonomic category”,.....)
- “scheduled time”, (“work of art”,.....)
- .....

The first coupled words can be freely replaced with any one in the “()” to provide flexibility for modulating your text messages. At last, you can composite the text messages, for example, as in the following conversation:

Sender: Hello!

Receiver: ☺

Sender: Have you attended the physics class? There is an optical experiment in the third class.

Receiver: No. I was called by our director to talk with a foreign professor.

Sender: Oh, I have a scheduled time table for you to make up the experiment class on the open time of our North Building. Please take your manuscript paper designed for this experiment. By the way, bring a red pan with you.

Receiver: That’s great! Thank you very much!

The secret information “Meet you at East Gate on 13:00pm” have been successfully hid in the above text messages by the sender.

For the receiver, this system can distill the hid data with the help of rule system, and process those data by decrypting and decompressing. At last, the original information can be recovered from the decompressed and decrypted binary data.

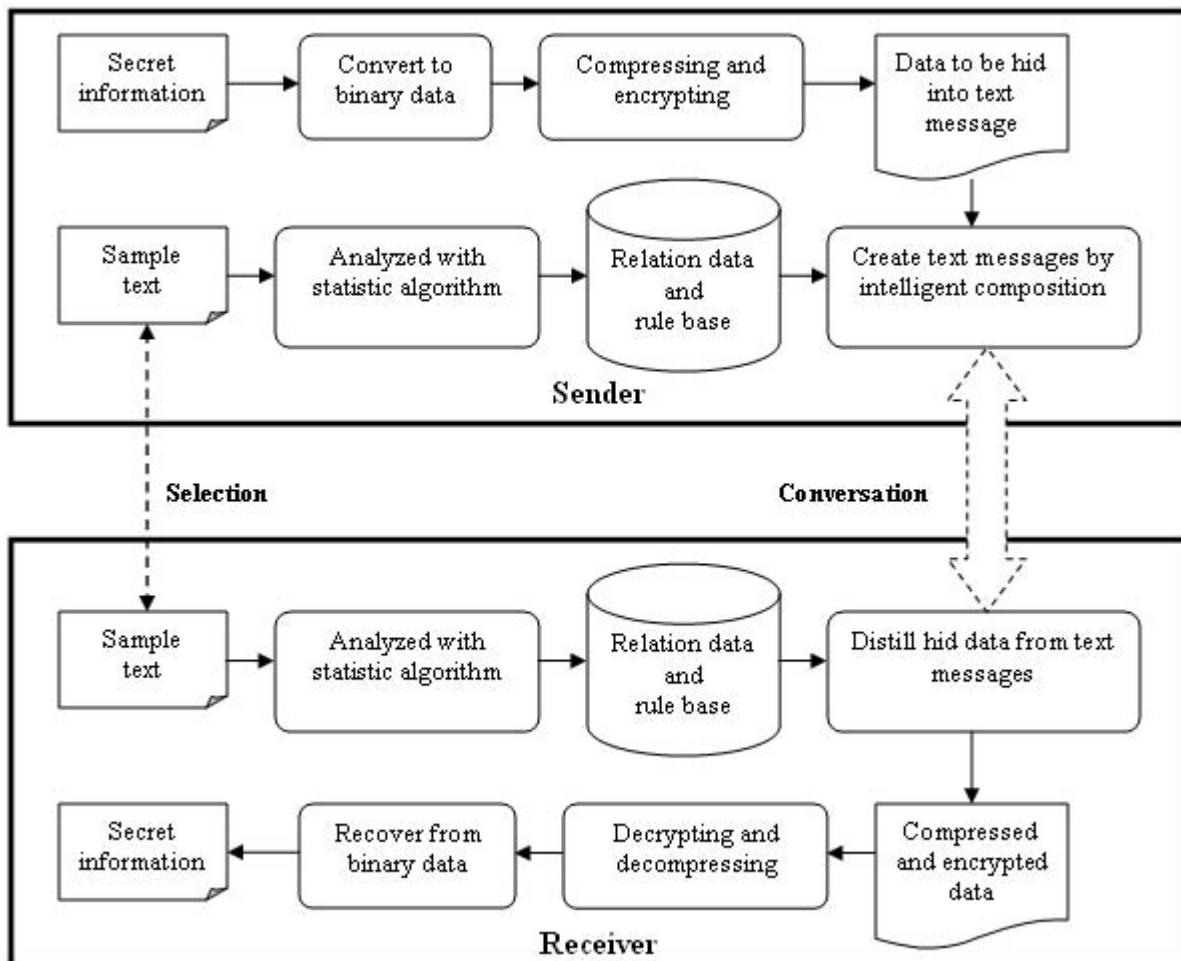


Figure 1. Information hiding system for online conversation

#### D. System Function Modules

This system is composed of four modules: interface, sample analysis, intelligent composition, and information recovery.

The user interface module deals with the problem of the alternation of the users which includes the operation such as sample setup, information input and display, hiding and recovery. The sample analysis module analyzes the sample text and creates their statistic relation data. Those data are applied to produce rules for the intelligent composition of text messages in conversation.

The intelligent composition module converts the secret information into binary data, and then compresses and encrypts those data so that they can be hid effectively with enough safety. Especially, this module provides heuristic technique for the hiding composition of text message and makes it close to the nature language, according to the rules decided by the statistic relation data of sample text.

The information recovery module distills the binary data from text message and converts it to original information.

By this system, the conversation parties can reach concealed communication even through a public online channel.

#### IV. DISCUSSION AND CONCLUSION

This system uses the statistical relations between the words of sample text to achieve the purpose of information hiding, so it has the capability of immunity from regular operations, such as formatting, compressing and sometimes manual altering operation in text size, front, color or the space between words, and can work reliably.

In order to improve the language quality of text messages, we explore a man-machine cooperation approach by adopting the technique of heuristic composition, and provide enough flexibility to modulate those messages close to nature language with the chosen conversation topic. But affected by the manual operation on the composition of conversation message, system efficiency probably restricts its application to hide short information. It is worth of our further research that a reasonable trade-off between the system efficiency and language quality is expected to be explored.

#### ACKNOWLEDGMENT

This research is supported by Shanghai Leading Academic Discipline Project (No.B210).

#### REFERENCES

[1] J. Jin., *Research on Data Hiding in Text Documents*, Shantou: Shantou University, 2008.

[2] M. T. Chapman, *Hiding the Hidden: A Software System for Concealing Ciphertext as Innocuous Text*, Milwaukee: University of Wisconsin-Milwaukee, 1997.

[3] X. X. Niu, and Y.X. Yang, "Research on the algorithm of text steganography," *ACTA Electronica Sinica*, vol.31(3), pp.402-405, March, 2003.

[4] F. Chen, and B. Wang, "An algorithm of text information hiding based on font," *Computer Technology and Development*, vol. 16 (1), pp. 20-22, January, 2006.

[5] P. Chen, S. W. Guo, and H. L. Chen, "Color-based information hiding algorithm for text documents," *Science Technology and Engineering*, vol.7 (14), pp.3544-3546, July, 2007.

[6] L. Ou, X. M. Sun, and Y. L. Liu, "Adaptive algorithm of text information hiding based on character intensity," *Application Research of Computers*, vol. 24(5), pp.130-132, May, 2007.

[7] D. C. Han, *Text Information Hiding Algorithm Based on the Layered Structure of the Natural Language*, Changsha: Hunan Science and Technology University, 2008.

[8] S. W. Xu, and D. J. Xu, "New hypertext steganography method based on attribute redundancy," *China Science and Technology Information*, vol. 2007 (19), pp.111-113, 2007.

[9] P. Chen, and F. Zhang, "Research on text information hiding techniques based on statistical characteristics of characters," *Journal of Pingdingshan Institute of Technology*, vol.16 (4), pp. 16-18, pp.26, July, 2007.

[10] C.Y. Ye, Y. S. Bi, X. S. Zhang, and J. Y. Qi., "An algorithm of text steganography," *China Information Security*, vol.2005(11),pp.106-108, November, 2005.

[11] P. Chen, and L. H. Zhang, "Research on information hiding techniques based on text," *Journal of Chongqing University of Science and Technology (Natural Sciences Edition)*, vol. 9(4), pp.107-109, pp.115, December, 2007.

[12] Z. X. DAI, F. Hong, and J. Dong, "Algorithm of Text Information Hiding Based on Huffman Coding," *Computer Engineering*, vol.33(15), pp.147-147, pp.151, August, 2007.

[13] G. Luo, and X. M. Sun, "Steganalysis for stegotext based on text redundancy," *Journal on Communications*, vol. 30(6), pp.19-25, June, 2009.

[14] K. Maher. TEXTO. URL: <ftp://ftp.funet.fi/pub/crypt/steganography/texto.tar.gz>, May.21, 2008.

[15] P. Grosse. TextHide. URL: <http://www.compris.com/TextHide/en/>, June 6, 2009.

[16] S. F. Wu, *Researches on Information Hiding Technology*, Hefei: China Science and Technology University, 2003.

[17] H. T. Liu, Z. W. Zhao, and G. L. Sheng, "Hidden markov models and its application to natural language process," *Microprocessors*, vol. 2009(3), pp.74-76, June, 2009.

[18] J. D. Yu, X. Z. Fan, and J. H. Yin, "Application of hidden markov model in natural language processing," *Computer Engineering and Design*, vol.28(22), pp.5514-5516, November, 2007.