Scientific Journals
Maritime University of Szczecin

Zeszyty Naukowe
Akademia Morska w Szczecinie

2014, 38(110) pp. 10–14
ISSN 1733-8670

2014, 38(110) s. 10–14

Hiding data in text environments and their parameters

O.Y. Afanasyeva[1], Jurii M. Korostil[2]

[1] Institute of environmental geological chemistry of National Academy of Sciences of Ukraine
[2] Maritime University of Szczecin, Institute of Marine Technology
70-500 Szczecin, ul. Wały Chrobrego 1/2, e-mail: j.korostil@am.szczecin.pl

Key words: hiding, steganography, dictionaries, semantic methods, extraction, message, digital text environment

Abstract
In this study, the parameters have been developed, characterized by the invisibility level of the message, which is embedded in a text environment, and methods which determine the values of these parameters. Also a review of the system composition of steganography, which is oriented to the use of text digital environments. Also included is the concept of semantic vocabulary of digital environments and text messages. Were also made analysis methods for implementing messages, depending on the fit of these dictionaries.

Introduction

Use of digital text environments for hiding messages (V_i), is one of perspective directions in data protection sphere. Development of this direction is directly linked to use of digital information systems for saving and transmitting text data [1, 2]. Methods of hiding separate messages in digital environments of various types is widely researched and developed in sphere of steganography methods of their protection [3]. Those methods are based on use of semantic excessiveness which appears in reflection of relevant environments in form of perceptible images. Thus, the level of invisibility messages in digital environments is closely connected with the peculiarities of perception of the semantic content of the image by the system of human perception (SSL) information, represented by this type of images.

Basic Definitions

Depending on type of image, rendered to user by digital media, *SSL* in one or another way uses different data perception organs and various mechanisms of transformation of those data into information, which we will call interpretational transformations, formally written down as:

$$j(x_i) = F^j(d_i, q_i) \quad (1)$$

where: $j(x_i)$ – interpretational description of data d_i, which are represented in image like q_i. Interpretational description we will represent in text form. It is common to distinguish following types of digital images: graphical, audio, text, numeric etc. Other image types which can be formed in digital representation systems, are derivative from mentioned above basic types, as an example could be multimedia images which generally are the synthesis audio and graphics images, animation images, which are dynamic form of images, and so on [4].

In this case we will review text digital images. Thus we introduce the definition of the parameters, helping to create a basic definition of invisibility level or the level of covering the message in the digital text environment (*TCS*), which we call the confidentiality level (u) messages in *TCS*. Level of confidentiality of the message in the text environment is determined by the following parameters:

- visibility level (μ);
- recognition level (η);
- forecast level (λ);
- audible sensitivity ($\varkappa$).

Definition 1. Level of visibility μ of hidden message in text representation of some image (*TO*), is defined by the level of recognition of semantic nature *TO*, describing hidden message in the envi-

ronment of representation of relevant text fragment mt_i as text image to_i.

Level μ means the possibility of detection of representation of semantics of hidden message during perception *SSL* of open text by user. An example, illustrating this parameter could be recognized in text environment of some words of hidden message. Formally, this level is defined by the following correlation:

$$\mu = \alpha \left\{ \left[\sum_{i=1}^{m} \gamma_i (y_{i-1}, x_i, y_i) \right] / \sum_{j=1}^{k} y_j \right\} \qquad (2)$$

where: α – proportional coefficient, x_i – word from hidden message, m – size of hidden message, y_i – word of open text image, k – number of words in *TO*, γ_i – function, defining level of semantic coherence of word x_i and neighbor words y_{i-1} and y_i from *TO*.

Definition 2. Level of recognition η defines level of detection of graphical anomalies in *TO*, which is caused by introduction of messages.

Level η means the possibility of detection of text anomalies in *TO*, which is caused by introduction of V_i in *TO*. Such anomalies can show themselves in following. Each *TO* is formed according to grammar rules of language $\Gamma_i(x_1, \cdots, x_n)$, which is used to form *TO*. Each $\Gamma_i(\gamma_i)$ defines some structure of corresponding text and other peculiarities of forming texts in selected language. Corresponding anomalies can be of following types:

- structural grammar anomalies (a^{Γ});
- dictionary language anomalies (a^{C});
- semantic excessiveness (a^{S});
- phonetic anomaly (a^{F}).

Anomaly a^{Γ} appears when structure of fragment in *TO* does not correspond to any structure, provided by grammar $\Gamma_i(x_1, \cdots, x_n)$. Such value is defined by difference of sequence of use of words x_i with some grammatical factors from sequence, defined by grammar $\Gamma_i(x_1, \cdots, x_n)$. Value of a^{Γ} parameter is defined by number of violations in structure mt_i with *TO* according to correlation:

$$a^{\Gamma}(TO_i) = \alpha^{\Gamma} \sum_{i=1}^{m} \beta_i (x_i, x_{i+1}) \qquad (3)$$

where: α^{Γ} is a coefficient of coherence of value a^{Γ}, $\beta_i(x_i, x_{i+1})$ – element of structural anomaly, which exists between neighbor symbols x_i and x_{i+1} in fragment $mt_i \in TO$.

Dictionary anomaly a^{C} means use of words in *TO*, which are not common for corresponding plot type TO_i. As an example of plot type TO_i could be text description of technical object, other type could be description of landscape etc. A dictionary of anomaly concerns key words. For its detection are used thematic thesauruses (Tz_i). Value a^{C} is defined by the following correlation:

$$a^{C} = \alpha^{C} \sum_{i=1}^{m} \urcorner x_i^k \qquad (4)$$

where: α^{C} – coefficient of correlation of value a^{C}.

Semantic excessiveness a^{S} appears in case when in $mt_i \in TO$ is used x_i with near, or equal semantic values $\sigma^{Z}(x_i^k)$. Value a^{S} depends on number of words, used in mt_i, which are semantically excessive. Formally, this value is defined by the following correlation:

$$\left\{ \sum_{i=1}^{m-1} \left\{ \left[\left[\sigma^{Z}(x_i) - \sigma^{Z}(x_{i+1}) \right] \le \delta\sigma^{Z} \right] \rightarrow \left(a^{S} = a^{S} + 1 \right) \right\} \right\} \qquad (5)$$

where: $\delta\sigma^{Z}$ – affordable threshold of differences between $\sigma^{Z}(x_i)$ and $\sigma^{Z}(x_{i+1})$, while x_i and x_{i+1} are accepted as semantically excessive to each other.

a^{F} is a phonetic anomaly mostly related to parameter of audible sensitivity. So, a^{F} we will link to parameter $\varkappa$. As an example of visible the phonetic anomaly could be poetic form of description of text fragment $mt_i \in TO$.

Definition 3. Level of audible sensitivity $\varkappa(\psi_i)$ is defined by the level of phonetic coherence of separate phrases φ_i, or sentences ψ_i, which relate to one fragment of text from text image, or whole text *TO*.

Level of audible sensitivity, by its nature, is defined by the level of coherence of sequential pairs of words, which is ensured by use of corresponding endings of the first word x_i of words pair $\langle x_i * x_{i+1} \rangle$ and use, if necessary, of appropriate preposition in x_{i+1}. Level of audible sensitivity $\varkappa$ gets its maximum value, if in framework of separate $mt_i \in TO$ is implemented such coherence, which allows to corresponding fragment to have rhyme. Level $\varkappa$ is formed according to requirements of orthography of relevant grammar $\Gamma_i(x_1, \cdots, x_n, \gamma_1, \cdots, \gamma_m)$. This parameter is basic in case of text analysis, which is audibly perceived or when we talk about analysis of language sounds, during insonification of corresponding *TO*. In that case we will limit ourselves by texts, displayed by visual electronic devices.

Definition 4. Level of predictability of current phrases, or fragments of text $\lambda(\varphi_i, \varphi_{i+1})$ is defined by the level of interpretational equality of two sequent or current phrases.

Formally, value $\lambda(\varphi_i, \varphi_{i+1})$ can be determined according to the following correlation:

$$\lambda(\varphi_i, \varphi_{i+1}) = \sum_{i=1}^{m} sg\left[a_i^{j}(\varphi_j) = a_i^{j+1}(\varphi_{j+1}) \right] \qquad (6)$$

where: $a_i^j(\varphi_j)$ – separate word a_i from text representation of interpretational definition of phrase φ_j, which is written down as $j(\varphi_j) = \langle a_{i1}*\cdots*a_{im}\rangle$. Obviously, parameter of such type can be reviewed also at the level of key phrases, if in framework of corresponding steganography system is used semantic dictionary S_C [5].

Value of parameter $\lambda(\varphi_i, \varphi_{i+1})$ for *TO*, in general, can vary in preset boundaries. Relatively to the text in general, parameter $\lambda(TO)$ defines text stylistics.

Methods of hidden messages embedding into text environment

Widely spread methods of embedding of V_i into *TO* are methods that use text structure [6, 7]. These include methods depend on text editors and methods do not depend on them. An example of the first type may be the method of using one or any number of spaces between words, or some other special character, supported by the editors. The semantic value of such characters is mostly minimal. For the methods of the second type are:

– hidden message consists of words, existing in text, but those words are used from definite positions in preset sequence.

Let's review method of message hiding, based on use of parameters μ, η and λ, which we will call the semantic secrecy level (*SMU*).

As far as hiding is performed based on the parameter that characterizes the semantic anomalies, which in some semantic correspondence between the separate words of text, based on parameter characterized the level of consistency from the point of view of the requirements of grammar to the parameter that characterizes the predictability of the following words or phrases in the message, then according to relevant parameters should set their thresholds limiting the modification of fragments of text. All mentioned parameters characterize semantics of text, to which message is being embedded and also semantics of message text. This methodic can be implemented in framework of following conditions.

Condition 1. Semantics of text environment should mostly match semantics of the message.

Condition 2. Semantic method should be based on use of semantic dictionaries.

The first condition is typical for steganography systems, orientated on use of digital environments of various types, because it supposes selection of digital environment, which would best fit for embedding the message in it. For example, in case of steganography systems, orientated on use of graphical digital environments, is solved task of selection of most suitable environment from the point of view of distortions of invisibility [8, 9]. In those cases sign of suitability of the environment is not connected to semantics of the message. To avoid necessity of fulfillment of condition 1 of embedding of V_i into text digital environment (*TCS*), we can use the following approaches:

– input separate syllables of V_i, during embedding of V_i into *TCS*;
– use semantic dictionaries $S_C(V_i)$ and $S_C(TCS_i)$ and semantic parameters of coherence of separate words in framework of a sentence.

The first approach is quite complex and we will not review it. The second approach can be implemented by method, based on use of accepted parameters, except $\varkappa$. Corresponding approach to hiding of V_i in *TCS* will be called to semantic method, or *SMU*.

Implementation of semantic method of hiding message in digital text environment

Implementation of a semantic method of hiding of V_i in *TCS* needs to be performed in framework of separate steganography system, which should contain the following components:

– semantic dictionary S_C of selected *TCS*, which we will call S_C^{TCS};
– semantic dictionary S_C for V_i, which we will call S_C^V;
– parameters of words selection, for implementation of V_i;
– means of selection of words from *TCS*, for identification of them as current word from V_i, or for replacement of it by the current word from V_i.

Semantic dictionary S_C^{TCS} is formed basing on text from *TCS* and functionally orientated thesaurus or encyclopedia. As in the Internet exist relevant thesauruses then in framework of semantic steganography system (*SSS*) is implemented software, which forms S_C^{TCS} according to basic words in *TCS* [10, 11]. The key words are all that are not complementary or services in the relevant grammar. Text descriptions in S_C^{TCS} are normalized and are by their nature thesauruses of corresponding *TCS*. Normalization of text descriptions in S_C^{TCS} is implemented basing on use of normalization rules, extending the corresponding grammar $\Gamma_i(x_1,...,x_n,\gamma_1,...,\gamma_m,\gamma_1^N,...\gamma_k^N)$, where γ_i^N – is normalization rule. An example of such rule could be replacement of synonym to basic word, or exclusion of word from description of word x_i, or from $j(x_i)$,

which is semantically excessive etc. Process of normalization is described by following correlation:

$$S_C^{TCS} = F^N\left[\Gamma_i\left(\gamma_1^N,\cdots,\gamma_k^N\right)\right]$$

where: F^N is a function of use of γ_i^N in selected fragment from $j(x_i)$. Basing on analysis of $S_C{}^{TCS}$ is set importance of $\sigma^Z(x_i)$ for each key word x_i from *TCS*. In framework of each phrase is set the range of change of function of semantic controversy between sequent words ($x_i * x_{i+1}$). The sequences consist of words of the same grammar type, example of which are subjects, verbs, adjectives etc. But this does not mean that sequences of words do form separate phrases according to grammar $\Gamma_i(x_1,...,x_n, \gamma_1,...,\gamma_m)$.

If words from *TCS* are used in V_i and, respectively are located in $S_C{}^{TCS}$, then embedding of V_i in *TCS* is implemented at the level of use of words from *TCS*. Unlike steganography systems, which use, for example, graphical environment, in which embedding is performed by modification of digital element of image with the aim to embed message data, in steganography system using *TCS* modification of environment is not performed, but are used words to form V_i, which is located in *TCS*. If subject areas $W(V_i)$ and $W(TCS)$ differ to such level, that their total word reserve does not much to such level that V_i cannot be modified so that word reserves of $W(V_i)$ and $W(TCS)$ match, then for implementation of V_i into *TCS* are used rules of selection of words in environment of *TCS* for their replacement to words from V_i. To such rules belong rules of construction of phrases and sentences γ_i, which are in $\Gamma_i(x_1,...,x_n,\gamma_1,...,\gamma_m,\gamma_1{}^N,...,\gamma_k{}^N)$. During that, values of parameters μ, η, λ are used as criteria for selection of words from *TCS*. For the case, when $W(V_i)$ and $W(TCS)$ are different, is introduced concept of semantic similarity between V_i and *TCS* at the level of separate words $x_i{}^V$ and $x_i{}^{TCS}$.

Definition 5. Semantic similarity $\pi(x_i{}^V, x_i{}^{TCS})$ is defined by the level of similarity $j(x_i{}^V)$ and $j(x_i{}^{TCS})$ with $S_C{}^{TCS}$ and $S_C{}^V$, which is formally described by following correlation:

$$\pi\left(x_i^V, x_i^{TCS}\right) = \sum_{i=1}^{k,m} sg\left(a_i^V = a_j^{TCS}\right) = \sum_{i,j}^{k,m} sg\left(\chi_{i,j}\right)$$

where: k, m – is a number of words in $j(x_i{}^V)$ and $j(x_i{}^{TCS})$, respectively, $a_i{}^V \in j(x_i{}^V)$, $a_j{}^{TCS} \in j(x_i{}^{TCS})$ and takes place following correlation:

$$\left[\left(a_i^V = a_j^{TCS}\right) \to \left[sg\left(\chi_{i,j}\right) = 1\right]\right] \vee$$
$$\left[\left(a_i^V \neq a_j^{TCS}\right) \to \left[sg\left(\chi_{i,j}\right) = 0\right]\right]$$

If $[\pi(x_i{}^V, x_i{}^{TCS}) = k]$ & $[k = m]$, then $x_i{}^V$ and $x_i{}^{TCS}$ are complete synonyms. If takes place $[\pi(x_i{}^V, x_i{}^{TCS}) = k]\&[k > m]$, then $x_i{}^V$ is a dominating synonym, which we will write down as $sd(x_i{}^V)$. If takes place $[\pi(x_i{}^V, x_i{}^{TCS}) = k]$ & $[k < m]$, then $x_i{}^V$ is called incomplete synonym and is written down as $sn(x_i{}^V)$. If takes place $[\pi(x_i{}^V, x_i{}^{TCS}) \neq k]$ & $[k \neq m]$, then $x_i{}^V$ is called a close synonym and is written down as $sb(x_i{}^V)$. During use of $sn(x_i{}^V)$ and $sb(x_i{}^V)$ takes place substitution of words from *TCS* by the words from V_i, and selection of words from *TCS*, for their replacement is implemented using parameters μ, η, λ i π.

Conclusions

Use of the mentioned in this work parameters, characterizing *SSS* enabled embedding text V_i into *TCS* with preset level of its secrecy or invisibility. As far as *TCS* is a totality of defined according to relevant standards codes, then modification of those codes can lead only to distortion of a separate symbol and for its substitution by another one [12]. This happens because symbol codes are not excessive and their modification leads to invisibility of corresponding symbol.

Are reviewed cases, when dictionary of the message is embedded into dictionary of *TCS*, or $S_C{}^V \subset S_C{}^{TCS}$ and then the message is being embedded by selection in *TCS* text of word sequence in predefined places, if there is more than one same word there. Obviously, during this may not be fulfilled condition of harmonization between words in V_i, required by grammar $\Gamma_i(x_1,...,x_n,\gamma_1,...,\gamma_m,\gamma_1{}^N,..., \gamma_k{}^N)$. If the absence of such harmonization does not distort semantics of V_i, then embedding is complete. If $S_C{}^V$ partially match or totally mismatch $S_C{}^{TCS}$, then in *TCS* is implemented substitution of fords from *TCS* to V_i. Such substitution is implemented basing on use of parameters μ, η, λ and π, for which are set allowable values. Words from *TCS*, for their substitution by words from V_i, are selected in such way, that rules of construction of sentences and phrases, defined by grammar $\Gamma_i(x_1,...,x_n,\gamma_1,..., \gamma_m,\gamma_1{}^N,...,\gamma_k{}^N)$ are executed with accuracy, defined by values of given parameters.

For extraction of the message, hidden with *SSS*, is used a session key, which is sequence of words in *TCS*, which form V_i. Such key is passed to recipient via separate protected channel, such key can be encrypted with one of cryptographic algorithms.

References

1. SERGEEV A.P.: Office local networks. 2003.
2. SOLOMATIN N.M.: Information semantic systems. 1989.

3. COX J., MILLER M.L., BLOOM J.A.: Digital watermarking. Morgan Kaufman Publizhers, 2002.
4. ROMANETS Y.V., TIMOFEEV P.A., SHANGIN V.F.: Security of information in computer systems and networks. Radio i svyaz, 1999.
5. AFANASYEVA O.Y., DURNIAK B.V., KOROSTIL Y.M.: Methods of representation of technical parameters of image in semantic dictionary of steganography system. Digest of scientific works of the Institute of Problems of Modelling in Power Engineering (IPME of NAS of Ukraine), Issue 46, 2008, 151–156.
6. AFANASYEVA O.Y., OLESHKO T.I.: Information model of steganography system. Digest of scientific works of the Institute of Problems of Modelling in Power Engineering (IPME of NAS of Ukraine), Issue 48, 2008, 151–156.
7. AFANASYEVA O.Y.: Methods of semantic transformations in steganography systems. Modelling and information technologies: digest of scientific works (IPME of NAS of Ukraine), Issue 56, 2010, 188–196.
8. AFANASYEVA O.Y.: Method of hiding of messages in graphical digital environment, ensuring JPEG standard stability. Modelling and information technologies: digest of scientific works (IPME of NAS of Ukraine), Issue 30, 2005, 162–165.
9. AFANASYEVA O.Y.: Analysis of parameters of steganography system, orientated on use of graphical digital environments. Modelling and information technologies: digest of scientific works (IPME of NAS of Ukraine), Issue 50, 2009, 48–57.
10. DURNIAK B.V., SHEVCHENKO O.V.: Analysis of development of information technologies. Digest of scientific works of the Institute of Problems of Modelling in Power Engineering (IPME of NAS of Ukraine), Issue 66, 2013, 169–176.
11. DURNIAK B.V., SABAT V.I.: Semantic security of information in document workflow systems. Ukrainian Printing Academy, Lviv 2010.
12. SHEVCHENKO O.V.: Analysis of data transmission channels, used in communication networks. Digest of scientific works of the Institute of Problems of Modelling in Power Engineering (IPME of NAS of Ukraine), Issue 46, 2008, 199–206.