Journal of the Korea Industrial Information Systems Research Vol. 21 No.3, June. 2016 :7-12
http://dx.doi.org./10.9723/jksiis.2016.21.3.007 ISSN:1229-3741

유전 알고리즘 기반 한글 텍스트 스테가노그래피의 연구

(A Study of Hangul Text Steganography based on Genetic Algorithm)

지 선 수[1)]
(Seon-Su Ji)

요 약 인터넷의 적대적인 환경에서 보안성을 향상시키기 위해 스테가노그래피는 커버 매체 내부에 비밀 메시지를 숨기는데 초점을 두고 있다. 즉 암호화의 보완이다. 이 논문에서 한글을 이용한 텍스트 스테가노그래피 기법을 제안한다. 보안 수준을 높이기 위해 비밀 메시지는 유전 알고리즘 연산자 교차를 통해 암호화한다. 커버 매체의 특성과 구조 변화가 없는 스테고 텍스트 형태를 만들기 위한 커버 텍스트로 메시지를 삽입한다. 커버 매체에 3.69% 삽입 용량을 유지하기 위해, 스테고 텍스트의 크기가 14%로 증가되는 것을 확인할 수 있다.

핵심주제어 : 교차, 유전 알고리즘, 스테고 텍스트, 정보은닉, 텍스트 스테가노그래피

Abstract In a hostile Internet environment, steganography has focused to hide a secret message inside the cover medium for increasing the security. That is the complement of the encryption. This paper presents a text steganography techniques using the Hangul text. To enhance the security level, secret messages have been encrypted first through the genetic algorithm operator crossover. And then embedded into an cover text to form the stego text without changing its noticeable properties and structures. To maintain the capacity in the cover media to 3.69%, the experiments show that the size of the stego text was increased up to 14%.

Key Words : Crossover, Genetic algorithm, Hidden Data, Stego Text, Text Steganography

1. 서 론

인터넷은 현대를 살아가는 우리 모두에게 의식주와 더불어 기본적인 필수 요소가 되었다. 정보의 검색과 수집에서부터 전자상거래와 업무처리에 이르기까지 편리한 삶을 보장해 주는 다양한 순기능이 있다. 인터넷 사용이 급속히 증가하면서 나타난 사이버 중독증과 폭력에서부터 시공간을 초월한 개인정보 침해와 시스템 파괴에 이르기까지의 역기능 또한 매우 심각한 수준에 이르렀다. 따라서 인터넷 상에서 송수신되는 정보의 무결성을 보호하기 위해 정보보안 및 보호는 가장 중요하고 필수적인 핵심적 가치로 자림 매김되었다. 현재 보안성과 견고성 측면에서 전달하려는 비밀 메시지를 암호화하여 숨기려는 스테가노그래피는 암호화보다 더 많은 관심을 받는 효

※ Corresponding Author : ssji@gwnu.ac.kr
Manuscript received Apr 6, 2016 / revised May 10 ,2016 / accepted June 28, 2016
1) 강릉원주대학교 정보기술공학과

과적인 기술이다. 스테가노그래피의 목적은 텍스트, 이미지, 오디오 및 비디오 매개체에 비밀 메시지가 숨겨진 스테고 매체의 존재 자체를 은폐하여 통신 채널을 통해 허가된 수신자에게만 무결성 자료를 효율적이고, 안전하게 전송하는 것이다. 일반적으로 정보를 숨기는 과정에서 다루어야 할 중요한 3가지 요소는 삽입 용량, 보안성, 공격자가 숨겨진 정보를 파괴하기 전에 스테고 매체가 견딜 수 있는 수정되는 양으로 설명되는 견고성 등이다. 텍스트는 텍스트 파일에서 다른 매개체보다 중복된 여분의 정보가 부족하기 때문에 삽입 용량과 보안성이 떨어지고, 적용면에서 어려운 기법중의 하나이다[1-2]. 그러나 인터넷에서 송수신되는 대부분의 자료는 텍스트 매체이고, 또한 문서의 구조와 형태가 다양하기 때문에 작은 메모리를 점유하는 비밀 정보를 합법적인 사용자에게만 송신할 때 언어적인 가시성을 유지하면서 보안성을 강화하는 새로운 텍스트 스테가노그래피 기법의 연구가 필요하다.

이 논문에서는 보안성과 견고성을 강화하기 위해 암호화되고, 비트화된 비밀 메시지를 가지고 유전 알고리즘(genetic algorithm)의 교차(crossover)를 적용하는 한글 텍스트 스테가노그래피의 적용 기법을 제시한다.

2. 관련 연구

텍스트 스테가노그래피는 텍스트 매체에 비밀 자료의 조각을 숨기는 것으로써 전송된 스테고 텍스트의 의미가 바뀌지 않는 범위 내에서 문서의 구조를 조금씩 변경하여 비밀 메시지를 은닉하는 기법이다. 스테가노그래피에 비밀 메시지를 숨기는 일반적인 과정은 (1)식으로 표시한다[3].

$$s = f(c, m, k)$$
$$m = f^{-1}(s, k) \quad (1)$$

여기에서 s는 비밀 메시지를 포함한 스테고 매체이며, c는 커버 매체, m은 비밀 메시지, k는 메시지를 숨기거나 추출하는 암호키이다.

텍스트 스테가노그래피는 언어적 스테가노그래피와 기술적인 스테가노그래피로 분류된다. 언어적인 스테가노그래피는 비밀 메시지를 숨기기 위해 자연언어를 커버 매체로 이용하는 기술이다. 또한 텍스트 스테가노그래피는 형식기반, 임의적 통계적 생성 방법, 언어적 방법으로 구분된다.

2.1 Format-based

비밀 메시지를 숨길 수 있는 커버 텍스트의 서식을 변경하는 방법을 이용한다. 즉 삽입 공간, 글꼴 크기, 텍스트 전체에 분산된 맞춤법 오류 등은 텍스트 스테가노그래피에서 사용되는 다양한 형식 기반 방법 중의 하나이다. 예를 들어 하나의 공간을 '0'으로 정하며, 두 개의 연속적인 공간은 '1'로 해석되어 적용할 수 있다. 커버 텍스트에 적은 양의 정보가 숨겨질 수 있지만 숨겨진 정보의 존재를 드러내지 않으면서 모든 종류의 텍스트에 적용될 수 있다. 단어와 문단 사이에서 다양한 변화가 보통의 텍스트에서 일반적으로 나타나기 때문에 이러한 변화는 해석하기 어렵다는 장점이 있다[4-5]. 그러나 현시점에서 컴퓨터가 대중화되어 노출될 위험성이 높다.

2.2 Random and Statistical methods.

언어의 통계적 특성에 따라 특정 자연 언어에서 자동적으로 커버 텍스트를 생성하는 데 이용된다. 특정한 문자 시퀀스 안에 정보를 숨기는 것으로 문자에 임의의 시퀀스에서 나타나는 정보를 끼워 넣는 방법이다. 문자 생성에 대한 두 번째 접근법은 주어진 언어에서 실제 단어와 같은 통계적 속성이 나타나는 단어를 만들기 위해 단어 길이와 문자 주파수의 통계적 특성을 갖는다. PCFG(probabilistic context-free grammar)는 문맥 자유 문법의 각 변환 규칙이 연관된 확률을 갖는 언어 모형으로 이용된다. PCFG는 근원 노드에서 시작하여 반복적으로 임의 선택 규칙을 적용하여 단어 순서를 생성하는데 사용될 수 있다[4-5].

2.3 Linguistic method

언어적 방법은 이것을 수정하기 위해 텍스트의 언어적 확률과 특성을 이용한다. 즉 정보를 숨길 장소로 메시지의 언어 구조를 사용한다. 구문 방법은 쉼표, 마침표와 같은 문장 부호가 비밀 메시지를 포함하는 문서의 적절한 장소에 배치되며, 기호를 삽입할 수 있는 장소의 적절한 확인이 필요하다. 스테가노그래피 자료는 구문 구조 자체에 숨길 수 있다. 이때 어휘, 문법, 의미적인 검사에 저항할 수 있는 견고성을 만족해야 한다[4-5].

2.4 분석도구

텍스트 스테가노그래피에서 왜곡 정도를 알아보기 위해 유사성을 측정하며, 효율성을 위해 삽입 용량을 확인할 필요가 있다. 텍스트 스테가노그래피에서 비밀 메시지를 숨길 수 있는 삽입 용량(capacity)은 삽입율(embedding ratio)과 공간 저장율(saving space ratio)을 이용하며, 수식 (2)식 혹은 (3)식으로 각각 계산할 수 있다[4].

$$Cap = \frac{total\ number\ of\ hidden\ bits}{total\ bits\ of \exp ected\ stego\ text} \times 100 \qquad (2)$$

혹은

$$Cap = \frac{\sum_{i=1}^{p} (\ total\ number\ of\ hidden\ bits)_i}{\sum_{j=1}^{q} (total\ bits\ of\ cover\ text)_j} \times 100 \qquad (3)$$

여기에서 p는 숨기려는 문자수이며, q는 커버 매체에 포함된 전체 문자수를 의미한다.

커버 매체와 비밀 메시지가 삽입된 스테고 매체 사이의 유사성을 비교하기 위한 상관계수는 수식 (4)를 사용하여 측정할 수 있다. 여기에서 X는 커버 매체, Y는 스테고 매체 자료이며, $x_i \in X$, $y_i \in Y$이다. $\overline{x}$와 $\overline{y}$는 X와 Y의 각각의 평균을 의미한다.

$$Corr = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \qquad (4)$$

3. 제안된 방법

임의의 삽입 순서로 한번에 00, 01, 10, 11의 정보를 어구 길이와 공백을 기반으로 커버 매체의 적절한 위치에 삽입한다. 여기에서 한글 커버 텍스트의 삽입 과정을 검사하는 동안에 구두점과 특수 문자와 같은 서식은 무시한다. 보안 수준을 강화하기 위해 비밀 메시지는 높은 수준의 보안 기능을 갖춘 유전 연산의 교차를 이용하는 알고리즘을 이용하여 인코딩한다.

3.1 Crossover operators

유전 알고리즘은 유전자의 변화를 통해 좋은 방향으로 진화해 가는 자연 진화의 과정과 자연 환경에 잘 적응한 개체가 더 많은 자손을 만들게 되는 적자생존의 과정을 모방한 최적화 알고리즘이다. 유전 알고리즘에서 교차는 한 세대에서 다음 세대까지 염색체 혹은 염색체의 프로그래밍을 변경하기 위해 사용하는 유전 연산자이다. 이것은 재 생 및 교차와 유사하며, 유전적 알고리즘의 기반이 된다. 교차는 하나 이상의 부모(상위) 솔루션을 취하고, 그들로부터 자식(하위) 솔루션을 선택하는 과정이다. 많은 교차 기법은 그들 자체를 저장하기 위해 서로 다른 데이터 구조를 사용하는 유기체가 존재한다. 일반적으로 교차는 사용자가 정의한 교차 확률에 따라 진화하는 동안에 발생한다[6-7].

3.1.1 산술 교차

산술 교차(arithmetic crossover)는 두 부모의 가중산술평균을 이용하여 자손을 만든다. 자손은 선형 제약 조건과 범위에 대해 만들어 질 수 있다. 두 명의 부모가 있을 때 첫 번째 부모가 더 좋은 적합도 값을 가지며 자식을 구성한다. 이를

수식(5)로 표현할 수 있다[8-9].

$$C_i^{gen+1} = \alpha \cdot C_i^{gen} + (1-\alpha) \cdot C_j^{gen}$$
$$C_j^{gen+1} = (1-\alpha) \cdot C_i^{gen} + \alpha \cdot C_j^{gen} \quad (5)$$

여기에서 부모 C_i^{gen}, C_j^{gen}는 두 명의 자손, 즉 C_i^{gen+1}와 C_j^{gen+1}를 각각 만들 수 있으며, 이들의 부모는 선형조합을 이룬다. 또한 가중치(α)는 $0 < \alpha \leq 1$ 사이의 임의의 값이다.

3.1.2 링 교차

링 교차(ring crossover) 연산자는 다음의 4단계로 구성된다. 1단계는 교차 과정을 위해 두 부모의 문자열(string)이 링 형태로 결합된다. 2단계는 링의 임의 위치에서 절단점이 결정된다. 3단계는 절단점을 참조하여 자손 중의 하나는 시계 방향으로 생성되고, 다른 하나는 반시계 방향으로 결정한다. 단계4는 링 교차에서 교환(swapping)과 역 공정(reversing process)이 수행된다. 링의 길이가 부모 전체의 길이와 같고, 자손은 링의 임의의 시점에 따라 생성되기 때문에 SPC(single point crossover), TPC(two point crossover) 연산에 따라 링 연산에 의해 많은 다양한 결과가 제공될 수 있다[6]. 링 교차 과정을 Fig. 1으로 표현할 수 있다.

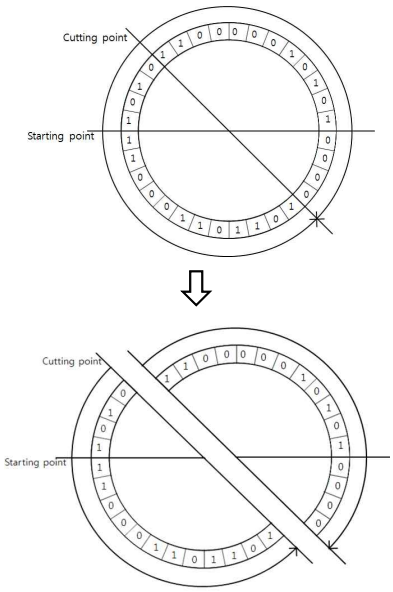

Fig. 1 Ring crossover process

3.2 Steganography Process

삽입 순서에 따라서 홀수 또는 짝수 크기의 단어와 여분의 빈 공간으로 정보를 삽입하는 Bhattacharyya 등(2010)이 제안한 텍스트 스테가노그래피 기법[3]과 링 교차 유전 알고리즘을 참고한다. 제안된 알고리즘을 Fig. 2로 나타낼 수 있으며, 그림에서 진하게 표현된 부분이 논문에서 제안된 부분이다.

1단계:(비밀 메시지 암호화)비밀 메시지를 암호화할 때 원주환치암호(columnar transposition cipher) 혹은 Playfair 암호와 유전 연산 교차의 기본 연산을 이용한다. 이때 암호화키와 링 교차 유전 알고리즘의 절단점을 이용한다.

2단계(비밀 메시지 삽입과정)전송하려는 비밀 메시지를 삽입하려는 커버 텍스트를 선택한다.

(1) 암호화된 비밀 메시지의 비트화된 정보 순서를 확인하고, 처음 두 비트 정보(MSG)를 선택한다.

(2) 커버 텍스트에서 임의의 삽입 시점을 결정한 후 비밀 메시지의 비트화된 정보를 삽입한다. 이때 선택한 커버 텍스트가 비밀 메시지를 삽입할 수 있는지 공간과 어구 구조를 확인한다.

(3) 어구의 크기가 홀수이고, 하나의 공간인 경우 MSG='00', 어구의 크기가 홀수이고, 두 개의 공간인 경우 MSG='01', 어구의 크기가 짝수이고, 하나의 공간인 경우 MSG='10', 어구의 크기가 짝수이고, 두 개의 공간인 경우 MSG='11'으로 정한다.

(4) 남아있는 비밀 메시지의 비트화된 정보를 삽입하기 위해 (1)부터 (3)의 과정을 반복한다.

3단계:(비밀 메시지 복호화)비밀 메시지를 복호화 하는 것은 수신측에서 실행되며, 송신측에서 수행한 과정의 역동작이다.

4단계:(비밀 메시지 추출과정)비밀 메시지가 삽입된 스테고 텍스트와 암호화 키, 교차 유전 알고리즘의 절단점 정보를 참조한다. 비밀 메시

지를 삽입하는 과정의 역 동작이다.

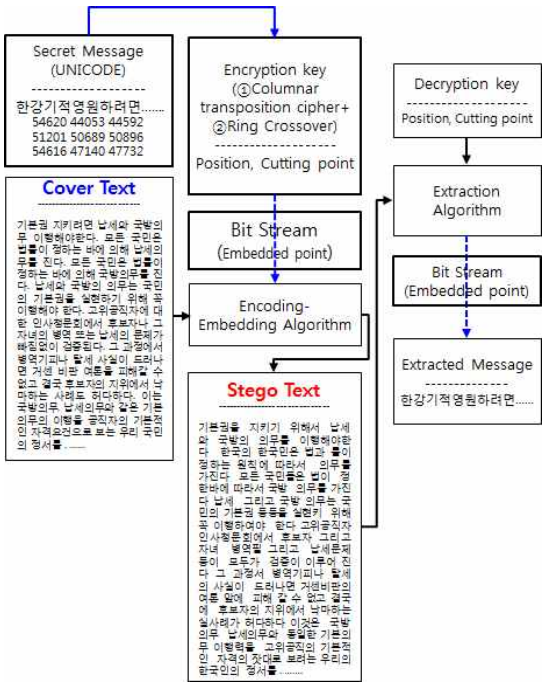

Fig 2. Text steganography model for the proposed implementation

3.3 적용 및 결과

논문에서 사용된 비밀 메시지의 크기는 8, 14, 26, 42, 72, 140바이트로 사용하였다. 비밀 메시지를 원주환치암호 기법을 적용하여 암호화한 후 비트 패턴으로 변환한다. 다음으로 링 교차를 이용하여 새로운 비트화된 정보를 얻는다. 이때 암호화키는 25314로 하며, 절단점으로 4를 적용하였다. 예를 들어 임의의 비밀 메시지 정보로부터 원주환치암호 기법을 적용하여 처음 두 글자의 비트화된 정보 강′과 원′을 계산하였다고 가정한다. 강′(44053, 1010110000010101)과 원′(50896, 1100011011010000)을 절단점이 4이고, 링 교차를 적용할 경우 강″(1100000101010000)과 원″(1100011011010101) 정보를 각각 얻을 수 있다. 다음으로 3.2절에서 제시한 알고리즘의 2단계 과정을 적용한다. 여기에서 사용한 한글 커버 텍스트의 크기는 3,333바이트이다. 알고리즘을 구현하는 과정은 J2SE을 이용하였다.

한글 문서인 경우 25% 내외가 공백으로 이루어졌으며 홀수 개 어구와 짝수 개 어구비율은 약 0.54:0.46로 구성되어 있어 좋은 불편성(unbiased)을 가지고 있다. 동일한 조건에서 한글 문서 100개를 가지고 각각의 비밀 메시지가 삽입될 때 비밀 메시지와의 상관계수는 근사식 $y = -0.00074x + 0.995$으로 구할 수 있음을 확인하였다. 여기에서 x는 비밀 메시지의 크기(byte)를 나타낸다.

Table 1 Results after the embedding process (cover text 3,333bytes)

Message (bytes)	Stego size (bytes)	Capacity(%)		Corr.
		eq.(2)	eq.(3)	
8	3,373	0.24	0.24	0.989
14	3,392	0.41	0.42	0.984
26	3,435	0.76	0.78	0.975
42	3,480	1.21	1.26	0.963
72	3,589	2.01	2.16	0.941
140	3,814	3.69	4.21	0.890

Table 1에서와 같이 커버 매체에 3.69% 삽입용량을 유지할 경우 스테고 텍스트의 크기가 14%로 증가되는 것을 확인할 수 있다. 대략적으로 텍스트를 커버 매개체로 사용할 경우 영문 및 아랍어 2.4%, 독일어 1.8% 내외의 삽입용량을 가질 때 보안성이 우수하며, 한글을 이용한 스테가노그래피의 경우 삽입용량 면에서 나쁘지 않음을 확인하였다.

4. 결 론

인터넷 공간에서 자료가 송신 및 수신될 경우 내부적인 자료측면에서 볼 때 보안성이 중요시되며, 외부적인 자료측면에서 무결성이 중요시 된다. 인터넷에서 송수신되는 대부분의 자료가 텍스트이고, 구조와 형태가 다양하기 때문에 텍스트 스테가노그래피는 보안성이 강화된 작은 용량의 비밀 메시지를 은닉하는데 적절한 도구이다. 한글 텍스트 스테가노그래피에서 커버 텍스트의

3.6% 내외로 비밀 메시지를 은닉할 수 있으며, 효율성과 보안성을 위해 원주환치암호기법을 이용한 암호화와 교차 유전 알고리즘의 절단점을 이용하였으며, 2% 내외의 삽입 용량, 6.4% 파일 크기 변화를 유지하기 위해서는 커버 텍스트와 스테고 텍스트에서 95% 이상의 상관성을 유지할 필요가 있다는 것을 확인하였다. 산술교차 유전 알고리즘을 이용한 이미지 스테가노그래피 영역은 향후 연구해야할 부분이다.

References

[1] S. S. Ji, “Advanced LSB technique for hiding messages in audio steganography”, KIISC, Vol 17, No. 5, pp. 37-42, 2014.

[2] C. K. Mulunda, P. W. Wagacha and L. O. Adede, “Genetic algorithm based model in text steganography”, The African Journal of Information Systems, Vol 5, Issue 4, Article 2, pp. 131-144, 2013.

[3] S. Bhattacharyya, I. Banerjee and G. Sanyal, “A novel approach of secure text based steganography model using word mapping method(WMM)”, International Journal of Computer and Information Engineering Vol 4, No. 2, pp. 97-103, 2010.

[4] I. Banerjee, S. Bhattacharyya and G. Sanyal, “A procedure of text steganography using Indian regional language”, I. J. Computer Network and Information Security, Vol 8, pp. 65-73, 2012.

[5] M. Nosrati and R. Karimi, “A survey on usage of genetic algorithms in recent steganography researches”, World Applied Programming, Vol 2, No. 3, pp. 206-210, March 2012.

[6] Y. Kaya, M. Uyar and R. Tekin, “A novel crossover operator for genetic algorithms: ring crossover”, AWERProcedia Information Technology&Computer Science, Vol 1, pp. 1286-1292, 2012.

[7] I. B. S. Bhattacharyya and G. Sanyal, “Design and implementation of a secure text based steganography model”, In Proceedings of 9th annual Conference on Security and Management under The 2010 World Congress in Computer Science, Computer Engineering, and Applied Computing, LasVegas, USA, July 12-15, 2010.

[8] T. Yalcinoz, H. Altun and M. Uzam, “Economic dispatch solution using a genetic algorithm based on arithmetic crossover”, Proto Power Tech Conference 2001, 2001.

[9] B. Osman, R. Din and M. R. Idrus, “Capacity performance of steganography method text based domain” ARPN Journal of Engineering and Applied Sciences, Vol 10, NO. 3, pp. 1345-1351, 2015.

지 선 수(Seon-Su Ji)

- 종신회원
- 1984년 충남대학교 계산통계학과(학사)
- 1986년 중앙대학교 응용통계학과(석사)
- 1993년 중앙대학교 응용통계학과(박사)
- 2006년 명지대학교 컴퓨터공학과(박사수료)
- (현)강릉원주대학교 정보기술공학과 교수
- 관심분야 : 정보보안(암호키, 정보은닉), 스테가노그래피