自然言語テキストを用いた秘密分散法

滝澤 修　　山村明弘

独立行政法人通信総合研究所
{taki, aki}@crl.go.jp

1. はじめに

昨今、コンピュータ技術およびネットワーク技術の進歩により、デジタル形式による画像、音声、テキストなどのコンテンツの流通が爆発的に増大している。そのような状況下で、デジタルコンテンツの著作権の主張や配布先の特定、あるいは情報伝送における盗聴防止のためのカムフラージュとするために、コンテンツの中に不可視な情報を隠蔽して埋め込む「情報ハイディング技術」の重要性が高まっている。情報ハイディング技術の一つとして、複数のメンバーが分散して保有する情報を合わせた場合にのみ秘密情報を復号できる秘密分散法(secret sharing scheme)がある[1][2]。コンテンツを対象とした秘密分散法の応用の一つとしてNaorら[3]によって提案された視覚復号型秘密分散法(Visual Cryptography または Visual Secret Sharing Scheme, 以下VSSS)は、複数の半透明なスライドを重ね合わせた場合にのみ意味のある画像が現れる技術であり、計算機を使わず人間の目視によって復号可能な新しい暗号技術として注目されている[4][5]。

筆者らは、自然言語テキストを埋め込み媒体とする秘密分散法(Text Secret Sharing Scheme, 以下TSSS)の実現の可能性について検討している[6][7]。本稿では、その簡単な実現方法の一つを提案する。

2. テキスト秘密分散法の考え方

TSSSは、秘密テキスト(secret text)を複数の分散テキスト(share text)に分散して隠蔽し(暗号化)、分散テキストを“重ね合わせて”秘密テキストを復号する手法、と定義できる。VSSSにおいては、分散画像はノイズのような無意味画像が使われることが多い[8]。従ってTSSSの場合も分散テキストが無意味な文字列であっても構わないが、そうすると秘密分散を使っていることを見破られる懸念があり、それは解読される脅威となりえることなので、できるだけ自然な分散テキストを合成することが望ましいといえる。

ギリシャ時代にスパルタで用いられていた「スキュタレー暗号」は、ある太さのドラムに紙テープを螺旋状に巻きつけ、ドラムの軸方向に文字を書き込んだ後にほどき、その紙テープを伝送し、受信者は同じ太さのドラムに巻きつけて復号するものであった。つまりドラムの太さに応じた一定間隔で文字列にスクランブルをかける暗号方式であった。TSSSにおいて、VSSSにおける“重ね合わせ”に対応する処理としてこのスキュタレー暗号の原理を応用する。即ち、複数枚の分散テキストをそれぞれ1行ずつ横書きに展開し、冒頭文字の位置を合わせた際に、ある位置において縦に並んだ文字列の中に秘密テキストが現れるようにする。図1に例を示す。この場合、スキュタレー暗号における一巻き分の紙テープが、各分散テキスト(各行)に相当することになる。

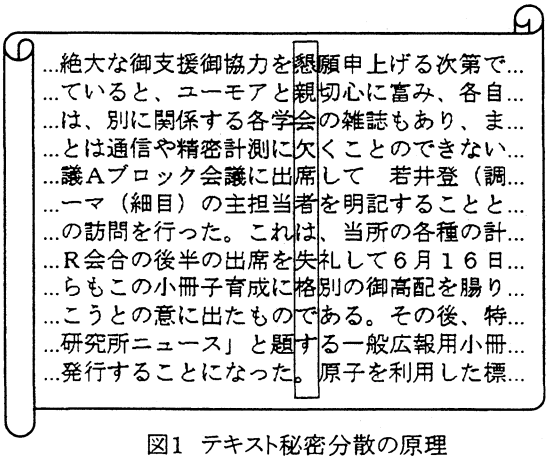

図1 テキスト秘密分散の原理
各分散テキストを1行ずつ横方向に展開して積み重ねたもの。四角で囲った列が秘密テキスト。

この方法は、VSSSとは異なり、重ね合わせる行為自体に計算機の補助が必要になる。また秘密テキストが隠蔽されている個所を目視で見つけて復号することはできなくはないが、計算機を援用すれば、より迅速に復号できる。以下の節では、分散テキスト生成と復号支援の実装についてそれぞれ検討する。

3. 提案手法の実装

本節では、第2節の方法を実現した処理アルゴリズムについて述べる。

3. 1 分散テキストの生成方法

第2節で述べたように、分散テキストは自然な文章であることが望ましい。単語を組み合わせることによって意味的に自然な文章を合成することは、計算機では膨大な知識と複雑なアルゴリズムを必要とする。そこで、句点までを単位とするテキストを大量にデータベース化しておき、そのテキストをつなぎ合わせることによって分散テキストを合成する方法を考える。ここで使うデータベースは、2つ以上のテキストを続けてもテキスト間の文脈が極端には不自然にならないように、同じ分野の内容についてのテキストによって構成されていることが重要と考えられる。

図1の原理により分散テキストを重ねて秘密テキストがうまく現れるようにするには、秘密テキストの文字数≦分散テキストの枚数、とする必要がある。この場合、十分な長さの秘密テキストを使えないか、もしくは大量の分散テキストを使わなければならない点で、実用上の制約があるが、とりあえずこの制約下で実装することにする。

【定義】

{ }は集合、< >は順序付き集合とする。英大文字は1テキスト(もしくはテキストの集合)、英小文字(添字を除く)は1文字を表す。

秘密テキストEは、文字列 $<e_1, e_2, \cdots, e_\varepsilon>$ から成る順序集合とする。図1の例の場合、

E=<懇,親,会,欠,席,者,は,失,格,で,す, 。>

となる。また、この例の場合、$\varepsilon=12$ となる。

テキストデータベースDは、テキストの集合 $\{T_1, T_2, \ldots\}$とする。

テキストデータベースの一要素 T_x は、文字列 $<t_{x1}, t_{x2}, \ldots>$ から成る順序集合とする。

なお、T_x の最後尾の文字は、句点“。”である。

【処理手順】

1. 秘密テキストの各文字を埋め込むテキストの抽出

$1 \leqq i \leqq \varepsilon$ であるすべてのiについて、文字e_i を要素に含むテキストT_{x_i}をDから抽出する。そして、$e_i = t_{x_i z_i}$ $(\in T_{x_i})$ とする。

その結果、

$$<e_1, e_2, \cdots, e_\varepsilon> \equiv <t_{x_1 z_1}, t_{x_2 z_2}, \cdots, t_{x_\varepsilon z_\varepsilon}>$$

となる。

2. 文字数合わせ処理

次に、

$$T_{w_i} = <t_{w_i 1}, t_{w_i 2}, \ldots, t_{w_i y_i}>$$

であり、

$$y_1+z_1 = y_2+z_2 = \ldots = y_\varepsilon+z_\varepsilon$$

を満たす$\{T_{w_i}\}$($\in$ D, $1 \leqq i \leqq \varepsilon$)を、Dから抽出する。但し、各$T_{w_i}$は一文またはそれ以上とする。

3. 分散テキストの合成

分散テキスト $<H_1, H_2, \ldots, H_\varepsilon>$を、

$H_1=<T_{w_1}, T_{x_1}>$

$=<t_{w_1 1}, t_{w_1 2}, \ldots, t_{w_1 y_1}, t_{x_1 1}, t_{x_1 2}, \ldots>$

$H_2=<T_{w_2}, T_{x_2}>$

$=<t_{w_2 1}, t_{w_2 2}, \ldots, t_{w_2 y_2}, t_{x_2 1}, t_{x_2 2}, \ldots>$

...

$H_\varepsilon=<T_{w_\varepsilon}, T_{x_\varepsilon}>$

$=<t_{w_\varepsilon 1}, t_{w_\varepsilon 2}, \ldots, t_{w_\varepsilon y_\varepsilon}, t_{x_\varepsilon 1}, t_{x_\varepsilon 2}, \ldots>$

とする。

(処理終)

上記の処理手順は、$1 \leqq i \leqq \varepsilon$ のすべての i について、e_i $(= t_{x_i z_i})$ の位置が文頭から数えて同じになるように $\{H_i\}$ を合成するものである。$\{T_{w_i}\}$は、文字数合わせのためにのみ挿入されることになる。

3. 2 秘密テキストの復号支援方法

第2節で述べた通り、提案方法は秘密テキストが隠蔽されている個所を見つけて復号することが目視では比較的面倒である。秘密テキストは意味を持つフレーズであるという前提に基づくと、意味を持たないフレーズは、1文字の形態素の並びになる場合が圧倒的に多い。そのため、2文字以上の形態素が多く現れる個所は意味のあるフレーズ、すなわち秘密テキストである可能性が高いことになる。この性質を復号に援用する。

3．3 具体的な処理の例

3．1項の分散テキストの生成機能、および3．2項の秘密テキストの復号支援機能をperlで実装した。分散テキストの生成のためのテキストデータベースとしては、通信総合研究所の広報紙「CRLニュース」の20年分の全記事[9]を使用した。このデータベースを採用した理由は、著作権上の懸念が無いことと、同じ分野の内容のテキストによって構成されているという条件を考慮したためである。データベースのサイズは約5MBである。また、秘密テキストの復号支援については、形態素解析器「茶筌」[10]を用いた。

「懇親会欠席者は失格です。」（文字数12）を秘密テキストとした場合に生成した分散テキスト（12枚）のうち、例として2枚の一部を以下に示す。いずれも、自然言語テキストとして不自然ではないと思われる。なお、丸で囲った文字が秘密テキストの一部をなす。

「船舶や航空機といった移動体がその運航中、気象情報など航行に必要な各種のデータを無線通信を介して入手し、また航法電波を受信して測位を行い、さらにレーダによって障害物を自ら探知するなどして、日々安全航行を確保していることは、今や余りにも当然のこととなっているが、思えば、移動体から他に、あるいは他から移動体に、“無線”でアクセスするというこの分野は、電波の全くの独壇場であり、その機能は他の追随を許さぬものがある。昭和５２年の年頭にあたり御挨拶旁々、本年における当所の研究方針について抱負を述べ、職員各位の一層の御尽力をお願いするとともに、部外の関係各位に対しましては相変りませず、絶大な御支援御協力を懇願申上げる次第であります。」

「その後、特に重視されていたとう載テープレコーダの動作、センサブームの展開、長短２組の観測用アンテナの伸展等いずれも計画どおりに行われ、打ち上げ後約１か月間衛星の状態、機器の動作、特にミッション機器の動作のチェックが逐次実施され、各動作が正常であることを確認した上で電波研究所は４つのミッションすなわち、電離層臨界周波数の世界分布の観測（ＴＯＰ－Ａ、Ｂ）、電波雑音源の世界分布の観測（ＲＡＮ）、電離層上部の空間におけるプラズマ特性の測定（ＲＰＴ）、及び正イオン組成の測定（ＰＩＣ）について定常段階の観測を開始する予定であった。５１年度から研究計画書は新しい形式とし、特に研究サブテーマ（細目）の主担当者を明記することとした。」

また、分散テキストを重ね合わせて得られる文字列を茶筌の標準辞書によって形態素解析した結

育た	動詞-自立	2
」	記号-括弧閉	1
な	助動詞	1

力	名詞-一般	1
ア	未知語	1
各	接頭詞-名詞接続	1
測	動詞-自立	1
に	助詞-副詞化	1
担	未知語	1
こ	動詞-自立	1
席	名詞-一般	1
成	名詞-固有名詞-地域-一般	1
も	助詞-係助詞	1
とっ	動詞-自立	2

を	助詞-格助詞-一般	1
と	助詞-格助詞-引用	1
学	名詞-一般	1
に	助詞-格助詞-一般	1
出	動詞-自立	1
当れ	動詞-自立	2
を	助詞-格助詞-一般	1
に	動詞-自立	1
の	助詞-連体化	1
題	名詞-一般	1
た	助動詞	1
懇親	名詞-一般	2
会	名詞-接尾-一般	1
欠席	名詞-サ変接続	2
者	名詞-接尾-一般	1
は	助詞-係助詞	1
失格	名詞-サ変接続	2
です	助動詞	2
。	記号-句点	1
願	名詞-一般	1
切	名詞-形容動詞語幹	1
の	助詞-連体化	1
くし	名詞-一般	2
を	助詞-格助詞-一般	1
、	記号-読点	1
礼	名詞-一般	1
別	名詞-接尾-一般	1
ある	動詞-自立	2
原	接頭詞-名詞接続	1

申	名詞-一般	1
心	名詞-一般	1
雑	名詞-一般	1
こ	動詞-自立	1
て	助詞-接続助詞	1
明	名詞-固有名詞-人名-名	1
当	名詞-一般	1
し	動詞-自立	1
のる	動詞-自立	2
一子	名詞-一般	2

上	名詞-接尾-副詞可能	1
に	助詞-格助詞-一般	1
誌	名詞-一般	1
と	助詞-並立助詞	1
	記号-空白	1
記	名詞-サ変接続	1
所	名詞-接尾-一般	1

図2 分散テキストを重ね合わせて得られる文字列を形態素解析した結果
（右数字は形態素文字数）

果の一部を図2に示す。秘密テキストである「懇親会欠席者は失格です。」の一節（囲った部分）に、2文字形態素が多く現れており、適切な閾値を設定することにより、この部分を高い精度で機械的に切り出せることが示唆されている。

4. 考察

提案した秘密分散法は、重ね合わせる順序も鍵になっているので、順序を入れ替えることで別の秘密テキストも隠蔽することが原理的には可能である。

提案した秘密分散法は、自然言語テキストを秘密情報としているため、分散テキストが完全に揃っていない場合、秘密テキストは一部の文字が歯抜けになるものの、読解における補間により、大まかに解読できてしまうという本質的懸念があり、対策を今後検討する必要がある。但しパスワードなど1文字でも欠けたら無効な文字列を秘密テキストとする場合には、本秘密分散法は有効である。

提案した処理手法では、形態素解析を援用して、秘密テキストの場所を見つけやすくする処理を施している。これはVSSSにおいて、分散画像の重ね合わせによりコントラストが低下した秘密原画像を、輪郭線強調などを援用してより目視しやすくすることに相当する処理と言える。

秘密テキスト自体が無意味な文字列（例えばパスワードなど）の場合には、形態素解析による復号支援処理は機能しない。その場合は、無意味な文字列の前後を意味ある文字列ではさむ等の対応をする必要がある。VSSSの場合でも、無意味な画像を秘密原画像とした場合には、目視によって抽出することは困難であるので、TSSSに特有な課題ではないといえる。また、意味を持つ文字列であっても、例えば「山の木を切る。」のように、偶然に一文字形態素の連なりになる場合があるので、提案した復号支援処理は万能とはいえない。

秘密テキストを複数の部分秘密テキストにちぎって、「部分秘密テキストの最大文字数≦分散テキストの枚数」とし、分散テキスト上で置く位置を散らせることによって、埋め込める秘密テキストの文字数を増やすことが可能である。但しちぎる際に形態素を分断するような変なちぎり方をすると、形態素解析を用いた復号支援処理が困難になるので、工夫が必要である。また、置く位置を散らせた状態を構成するようにうまく分散テキストを合成することは制約条件が厳しいため難しく、データベースのテキストを短くするなどの工夫が必要であろう。

5. TSSSの応用

一般に暗号は、鍵の所有者一人が裏切れば直ちに安全性が脅かされるが、秘密分散は鍵を分散共有し、全員の協力が無い限り復号できない原理であるため、一部の者が裏切っても安全性は保たれる性質がある。従ってTSSSの応用としては、お互いに信頼できない、ゆきずりの者による共同作業における秘密メッセージ交換などが考えられる。また、出会い系のゲームなどのエンターティメントに面白い応用が考えられるかもしれない。

なお、提案した処理手法は特許出願中である[11]。

【謝辞】

本研究のきっかけを与えて下さった、PuKyong National University の Prof.Ji-Hwan Park に感謝する。また、自然言語テキストを埋め込み媒体として適用する方法に関して、横浜国立大学の松本勉教授、東京大学の中川裕志教授、三菱総合研究所の村瀬一郎、井上信吾、牧野京子の各氏から有益な助言を賜っていることに感謝する。最後に、本研究を日頃ご支援下さる通信総合研究所の大野浩之グループリーダに感謝する。

【参考文献】

[1] A.Shamir, "How to share a secret", Communications of the ACM, 612-613, 1979.

[2] G.Blakley, "Safeguading cryptographic keys", Proceedings of AFIPS National Computer Conference, 313–317, 1979.

[3] M.Naor and A.Shamir, "Visual Cryptography", Advances in Cryptology-Eurocrypt'94 , 1-12, 1994.

[4] 加藤拓, 今井秀樹, "視覚復号型秘密分散法の拡張構成方式", 信学論, Vol.J79-A, No.8, 1344-1351, 1996年8月.

[5] 凸版印刷株式会社, 視覚復号型暗号製品「あわすとでーる」, http://www.toppan.co.jp /aboutus/release/article463.html, 2001年4月.

[6] 滝澤修, 山村明弘, "自然言語文を用いた秘密分散の提案", 情報処理学会 コンピュータセキュリティシンポジウム 2001, pp.343-348, ISSN 1344-0640, 2001年11月.

[7] 山村明弘, 滝澤修, "テキスト秘密分散", 電子情報通信学会 暗号と情報セキュリティシンポジウム, 11A-3, pp.787-792, 2002 年 1 月.

[8] Moon-Soo Kim, Seong-Han Shin, Ji-Hwan Park, "New Construction for Multiple Visual Secret Sharing", 電子情報通信学会 暗号と情報セキュリティシンポジウム, 2000 年 1 月.

[9] 通信総合研究所, "CRL ニュース", 創刊号～第 238 号, ISSN 1346-8626, 1976 年～1995 年.

[10] 奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座(松本研究室), "日本語形態素解析システム茶筌 version 2.0 for Windows", 1999.

[11] 滝澤修, 山村明弘, "隠蔽文章抽出方法及び装置", 特願 2001-335566.