BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC DÂN LẬP HẢI PHÒNG
-------o0o-------

ĐỒ ÁN TỐT NGHIỆP

NGÀNH CÔNG NGHỆ THÔNG TIN

HẢI PHÒNG 2013

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC DÂN LẬP HẢI PHÒNG

-------o0o-------

KỸ THUẬT GIẤU TIN TRONG TỆP VĂN BẢN

ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY

Ngành: Công nghệ Thông tin

HẢI PHÒNG - 2013

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC DÂN LẬP HẢI PHÒNG

-------o0o-------

KỸ THUẬT GIẤU TIN TRONG TỆP VĂN BẢN

ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY

Ngành: Công nghệ Thông tin

Sinh viên thực hiện: Đinh Tiến Hương
Giáo viên hướng dẫn: TS. Hồ Thị Hương Thơm
Mã số sinh viên: 111330

HẢI PHÒNG - 2013

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC DÂN LẬP HẢI PHÒNG

CỘNG HÒA XA HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

-------o0o-------

NHIỆM VỤ THIẾT KẾ TỐT NGHIỆP

Sinh viên: Đinh Tiến Hương

Mã SV: 111330

Lớp: CT1201

Ngành: Công nghệ Thông tin

Tên đề tài:Kỹ thuật giấu tin trong tệp văn bản

NHIỆM VỤ ĐỀ TÀI

1. Nội dung và các yêu cầu cần giải quyết trong nhiệm vụ đề tài tốt nghiệp

a. Nội dung

- Tổng quan về giấu tin trong dữ liệu đa phương tiện
- Tìm hiểu kỹ thuật giấu tin trong tệp văn bản
- Cài đặt, thử nghiệm chương trình..

b. Các yêu cầu cần giải quyết

- Lý thuyết
  + Nắm được tổng quan về kỹ thuật giấu tin trong đa phương tiện.
  + Hiểu và nắm rõ một kỹ thuật giấu tin trong văn bản.
- Thực nghiệm (chương trình)
  + Cài đặt được kỹ thuật giấu bằng Matlap, thử nghiệm trên một tập văn bản nào đó với độ dài bất kỳ nào đó.

2. Các số liệu yêu cầu để thiết kế, tính toán

……..……
……
……
……

3. Địa điểm thực tập

……..……
……
……
……..……
……

CÁN BỘ HƯỚNG DẪN ĐỀ TÀI TỐT NGHIỆP

Người hướng dẫn thứ nhất:

Họ và tên: Hồ Thị Hương Thơm

Học hàm, học vị: Tiến Sĩ

Cơ quan công tác: Trường Đại Học Dân Lập Hải Phòng

Nội dung hướng dẫn:

…….

……

……

Người hướng dẫn thứ hai:

Họ và tên: ……..

Học hàm, học vị:…….

Cơ quan công tác: …….

Nội dung hướng dẫn:

…….

…….

……

Đề tài tốt nghiệp được giao ngày tháng năm 2013

Yêu cầu phải hoàn thành trước ngày tháng năm 2013

Đã nhận nhiệm vụ: Đ.T.T.N	Đã nhận nhiệm vụ: Đ.T.T.N
Sinh viên	Cán bộ hướng dẫn Đ.T.T.N

TS. Hồ Thị Hương Thơm

Hải Phòng, ngày .....tháng.....năm 2013

HIỆU TRƯỞNG

GS.TS.NGƯT Trần Hữu Nghị

PHẦN NHẬN XÉT TÓM TẮT CỦA CÁN BỘ HƯỚNG DẪN

1. Tinh thần thái độ của sinh viên trong quá trình làm đề tài tốt nghiệp:

.....

.....

.....

.....

.....

.....

.....

.....

2. Đánh giá chất lượng của đề tài tốt nghiệp (so với nội dung yêu cầu đã đề ra trong nhiệm vụ đề tài tốt nghiệp)

.....

.....

.....

.....

.....

.....

.....

3. Cho điểm của cán bộ hướng dẫn:

(Điểm ghi bằng số và chữ)

.....

.....

.....

Ngày.....tháng.....năm 2013

Cán bộ hướng dẫn chính

(Ký, ghi rõ họ tên)

PHẦN NHẬN XÉT ĐÁNH GIÁ CỦA CÁN BỘ CHẤM PHẢN BIỆN

ĐỀ TÀI TỐT NGHIỆP

1. Đánh giá chất lượng đề tài tốt nghiệp (về các mặt như cơ sở lý luận, thuyết minh chương trình, giá trị thực tế, ...)

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

2. Cho điểm của cán bộ phản biện

(*Điểm ghi bằng số và chữ*)

.....

.....

Ngày.....tháng.....năm 2013

Cán bộ chấm phản biện

(Ký, ghi rõ họ tên)

LỜI CẢM ƠN

Trước tiên em xin được bày tỏ sự trân trọng và lòng biết ơn đối với cô giáoTS Hồ Thị Hương Thơm - giảng viên Khoa Công nghệ thông tin – Trường Đại học Dân Lập Hải Phòng. Trong suốt thời gian học và làm đồ án tốt nghiệp, cô đã dành rất nhiều thời gian quí báu để tận tình chỉ bảo, hướng dẫn, định hướng cho em trong việc nghiên cứu, thực hiện đồ án

Em xin được cảm ơn các thày cô giáo Trường Đại học Dân lập Hải phòng đã giảng dạy em trong quá trình học tập, thực hành, làm bài tập, đọc và nhận xét đồ án của em, giúp em hiểu thấu đáo hơn lĩnh vực mà em nghiên cứu, những hạn chế mà em cần khắc phục trong việc học tập, nghiên cứu và thực hiện bản đồ án này.

Xin cảm ơn bạn bè và nhất là các thành viên trong gia đình đã tạo mọi điều kiện tốt nhất, động viên, cổ vũ tôi trong suốt quá trình học và làm đồ án tốt nghiệp.

Hải Phòng, Tháng 4 năm 2013

Đinh Tiến Hương

MỤC LỤC

DANH MỤC BẢNG – HÌNH

LỜI MỞ ĐẦU

Cuộc cách mạng thông tin kỹ thuật số đã đem lại những thay đổi sâu sắc trongxã hội và trong cuộc sống của chúng ta, tạo môi trường mở và tiện nghi. Đồng thời mặt trái của nó là sự xuất hiện những vấn nạn, tiêucực như nạn ăn cắp bản quyền, nạn xuyên tạc thông tin, truy nhập thông tin trái phép v.v... Tìm giải pháp cho những vấn đề nêu trên không chỉ tạo điều kiện đi sâu vào lĩnh vực công nghệ phức tạp đang phát triển rất nhanh này mà còn dẫn đến những cơ hội phát triển kinh tế.Che giấu thông tinlà việc ẩn dữ liệu trongmột thông báo công khai và thựchiện nó làm cho người khác khó phát hiện ra.

Khi thế giới lo lắng về việc sử dụng bí mật thông tin liên lạc và các quy định được tạo ra bởi chính phủ để hạn chế mã hóa và vai trò của che giấu thông tin là nổi bật**.**

Mạng Internet toàn cầu đã biến thành một xã hộiảo nơi diễn ra quá trình trao đổi thông tin trong mọi lĩnh vực chính trị, quân sự, quốc phòng, kinh tế, thương mại… Và chính trong môi trường mở và tiện nghi như thếxuất hiện những vấn nạn, tiêu cực đang rất cần đến các giải pháp hữu hiệu cho vấn đềan toàn thông tin như nạn ăn cắp bản quyền, nạn xuyên tạc thông tin, truy nhập thông tin trái phép v.v.. Đi tìm giải pháp cho những vấn đềnày khôngchỉgiúp ta hiểu thêm vềcông nghệphức tạp đang phát triển rất nhanh này mà còn đưa ra những cơ hội kinh tếmới cần khám phá. Do đó trong đồán này tìm hiểu phương pháp giấu tin trongtệph văn bản. Nôi dung gồm 3 chương chính sau:

- Chương 1. Tổng quan về giấu tin.

Giới thiệu vềmột sốđịnh nghĩa giấu thông tinmôi trường giấu tin, sơ lược vềmô hình giấu tin cơ bản. đặc điểm về giấu tin trong văn bản.

- Chương 2. Giấu tin trong tệp văn bản.

Giới thiệu về tệp văn bản. Trình bày kỹthuật giấu tin trong tệp văn bản. Các đặc điểm về tệp văn bản, các ý tưởng và các thuật toán giấu tin trong văn bản.

- Chương 3. Cài đặt và thử nghiệm.

Đưa ra môi trường cài đặt, giới thiệu giao diện chương trình vàchạy thửnghiệm trên tệp văn bản.

Chương 1 : TỔNG QUAN VỀ GIẤU TIN

1.1. Giới thiệu về giấu tin

1.1.1. Khái niệm

Giấu thông tin là nhúng thông tin mật vào trong một nguồn đa phương tiện mà không để người khác nhận biết về sự tồn tại của thông tin được giấu.

Che giấu thông tin là kỹ thuật cất giấu một thông điệp bí mật trong dữ liệu công khai. Che giấu thông tin có nghĩa là được bảo vệ bằng văn bản và dữ liệu ẩn dưới 1 lớp dữ liệu khác được công khai. Với một người có một bí mật che giấu thông tin gửi đến người khác. Các tập tin lưu trữ, hoặc tin công khai là tập tin mà bất cứ ai có thể nhìn thấy. Đó là công khaidữ liệucó sẵnmàđược sử dụng để ẩn tin nhắn. Đó là công khai dữ liệu có sẵn mà được sử dụng để che giấu một tin nhắn.

1.1.2. Mục đích

- Bảo mật dữ liệu được che giấu.Thông tin mật được giấu kỹ trong một đối tượng khác sao cho người khác không phát hiện được.
- Bảo mật chính đối tượng được dùng để giấu dữ liệu vào, chẳng hạn các ứng dụng bảo vệ bản quyền, phát hiện xuyên tạc thông tin.

1.1.3. Tại sao phải ẩn giấu thông tin?

Cuộc cách mạng thông tin kỹ thuật số đã đem lại những thay đổi sâu sắc trongxã hội và trong cuộc sống của chúng ta, tạo môi trường mở và tiện nghi. Đồng thời mặt trái của nó là sự xuất hiện những vấn nạn, tiêucực như nạn ăn cắp bản quyền, nạn xuyên tạc thông tin, truy nhập thông tin trái phép v.v... Tìm giải pháp cho những vấn đề nêu trên không chỉ tạo điều kiện đi sâu vào lĩnh vực công nghệ phức tạp đang phát triển rất nhanh này mà còn dẫn đến những cơ hội phát triển kinh tế.Che giấu thông tinlà việc ẩn dữ liệu trongmột thông báocông khai và thựchiện nó làm cho người khác khó phát hiện ra.

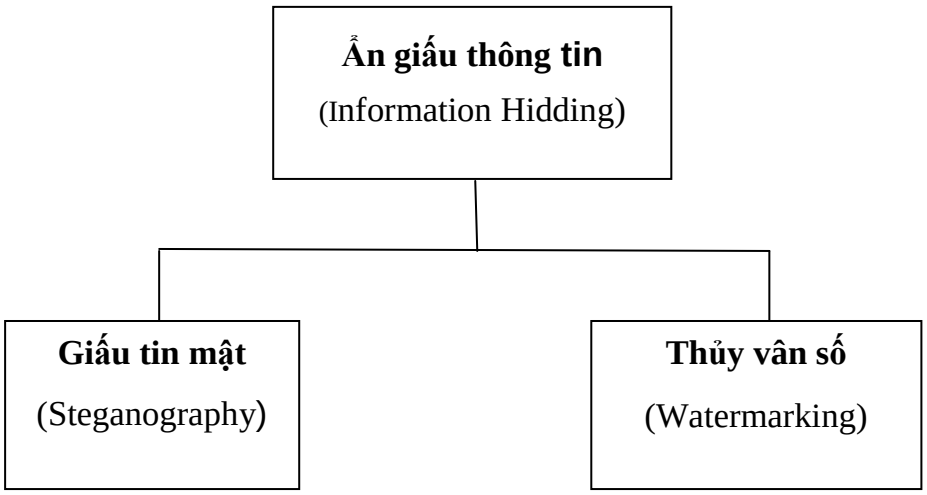
Khi thế giới lo lắng về việc sử dụng bí mật thông tin liên lạc và các quy định được tạo ra bởi chính phủ để hạn chế mã hóa và vai trò của che giấu thông tin là nổi bật.

1.1.4. Phân loại các kỹ thuật giấu tin

Do kỹ thuật giấu thông tin số mới được hình thành trong thời gian gần đây nên xu hướng phát triển vẫn chưa ổn định. Nhiều phương pháp mới, theo nhiều

khía cạnh khác nhau đang và sẽ được đề xuất, bởi vậy chưa thểcó được một định nghĩa chính xác, một sự đánh giá phân loại rõ ràng.

Có thể chia lĩnh vực giấu dữ liệu ra làm hai hướng lớn, đó làThủy vân số (watermarking) và Giấu tin mật (steganography). Nếu như watermarking quan tâm nhiều đến các ứng dụng giấu các mẩu tin ngắn nhưng đòi hỏi độ bền vững lớn của thông tin cần giấu (trước các biến đổi thông thường của tệp dữ liệu môi trường) thì steganography lại quan tâm tới các ứng dụng che giấu các bản tin đòi hỏi mật độ và dung lượng càng lớn càng tốt.



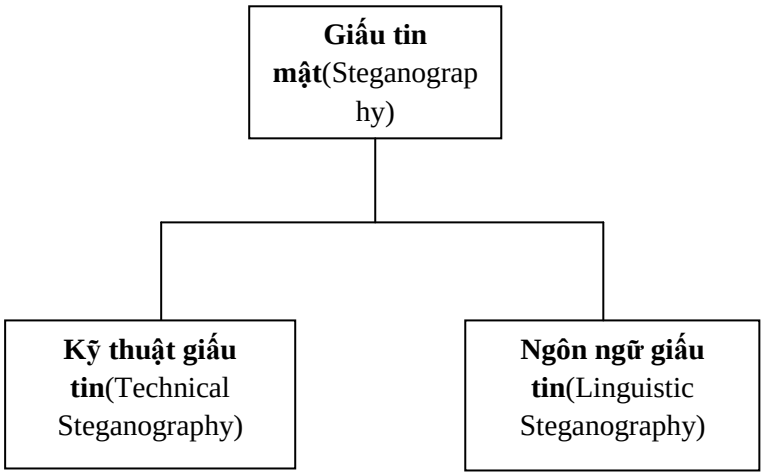
Hình 1.1. Phân loại các kỹ thuật giấu tin.

1.1.4.1. Ẩn thông tin(steganography)

Steganography là một kỹ thuật ẩn thông tin liên lạc, là quy trình giấu thôngtin cá nhân hay thông tin nhạy cảm vào những thứ mà không để lộ chúng theo dạngthông thường. Steganography có nguồn gốc từ tiếng Hy Lạp. Steganos (có nghĩa làphủ hoặc bí mật) và graphy (bằng văn bản hoặc bản vẽ). Che giấu thông tin mứcđộ đơn giản là ẩn chữ viết, cho dù nó bao gồm mực vô hình trên giấyhoặc bảnquyền thông tin ẩn trong một tập tin âm thanh.

Ngày nay, che giấu thông tin được thực hiện dữ liệu ẩn bên trong các dữ liệukhác trong một tập tin điện tử, việc ẩn dữ liệu thường gắn liền với công nghệcao.Ví dụ: bên trong một tài liệu Word có thể được ẩn một tập tin hình ảnh. Điềunày được thực hiện bằng cách thay thế các bit không quan trọng hoặc không cầnthiết nhất của dữ liệu trong bản gốc tập tin mà mắt và tai con người khó nhớ vớicác dữ liệu ẩn bit.

Mục tiêu của steganography là chuyển được một thông điệp thông qua vài phương tiện vô hại như: text, image, audio, video,.v.v.v. qua một kênh truyền thông nơi mà các thông điệp hiện hữu được che giấu.



Hình 1.2. Phân loại Steganography

Theohình 1.2 steganography là một kỹ thuật che giấu thông tinvà có thể được phân loại bao gồm ngôn ngữ steganography và kỹ thuật steganography. Ngôn ngữ giấu tin (Linguistic steganography) được định nghĩa bởi Chapman và đồng sự trong "the art of using written natural language to conceal secret messages".

1.1.4.2.Thủy vân số(Watermarking)

Thủy vân số (watermarking) là một trong những kỹ thuật giấu dữ liệu hiện đại, là quá trình chèn thông tin vào dữ liệu đa phương tiện nhưng bảo đảm không nhận biết được, nghĩa là chỉ làm thay đổi nhỏ dữ liệu gốc. Thông thường người ta chỉ đề cập đến nhúng thủy vân số. Một tập các dữ liệu số thứ cấp - gọi là mã đánh dấu bản quyền hay thủy vân (watermark), được nhúng vào dữ liệu số sơ cấp - gọi là dữ liệu bao phủ (ví dụ như văn bản, hình ảnh,âm thanh và phim số,.v.v.v.).

Nhúng thủy vân được ứng dụng trong nhiều lĩnh vực như bảo vệ quyền sở hữu, điều khiển việc sao chép, xác nhận giấy tờ, hay truyền đạt thông tin khác, …trong đó ứng dụng phổ biến là cung cấp bằng chứng về bản quyền tác giả của các dữ liệu số bằng cách nhúng các thông tin bản quyền.

Bảng 1.1. So sánh Giấu tin mật và Thủy vân số

	Giấu tin mật	**Thủy vân số**
Mục đích	- Che giấu sự hiện hữu của thông điệp. - Thông tin che giấu độc lập với vỏ bọc.	- Thêm vào thông tin bản quyền. - Che giấu thông tin gắn với đối tượng vỏ bọc.
Yêu cầu	- Không phát hiện được thông điệp bị che giấu. - Dung lượng tin được giấu.	- Tiêu chuẩn bền vững.
Tấn công thành công	- Phát hiện ra thông điệp bí mật bị che giấu.	- Thủy vân bị phá vỡ.

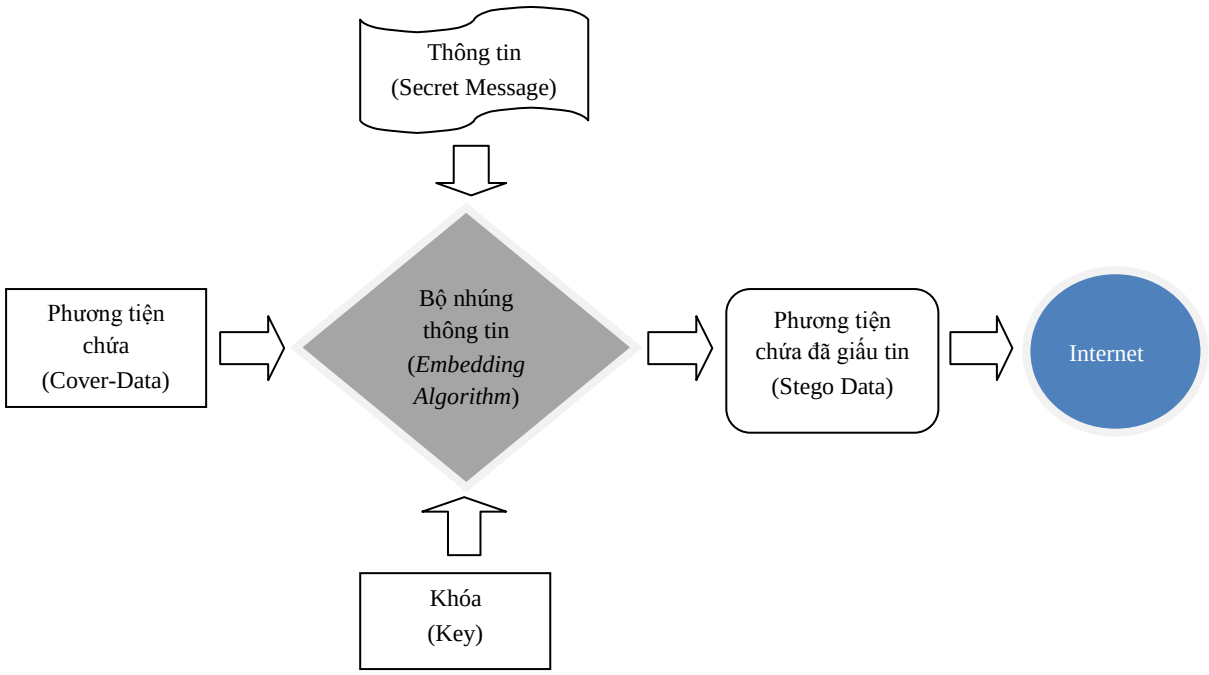
1.2. Giấu tin và tách tin

Các thành phần chính của một hệ giấu tin và tách tin trong ảnh số gồm:

- ***Thông tin (Secret Message):*** có thể là văn bản hoặc tệp ảnh hay bất kỳ một tệp nhị phân nào, vì quá trình xử lý đều chuyển chúng thành chuỗi các bit.
- ***Dữ liệu (hay dữ liệu gốc) (Cover Data):*** là dữ liệu được dùng để làm môi trường nhúng tin mật.
- ***Khoá bí mật K (Key):*** khoá viết mật tham gia vào quá trình giấu tin để tăng tính bảo mật.
- ***Bộ nhúng thông tin (Embedding Algorithm):*** Những chương trình, thuật toán nhúng tin.
- ***Bộ giải mã thông tin (RecoveringAlgorithm :*** Tách thông tin nhúng.
- ***Dữ liệu đã được nhúng thông tin (Stego Data):*** là dữ liệu khi đã nhúng tin mật vào đó.
- ***Kiểm định (Control):***kiểm tra thông tin sau khi được giải mã.

1.2.1. Giấu tin

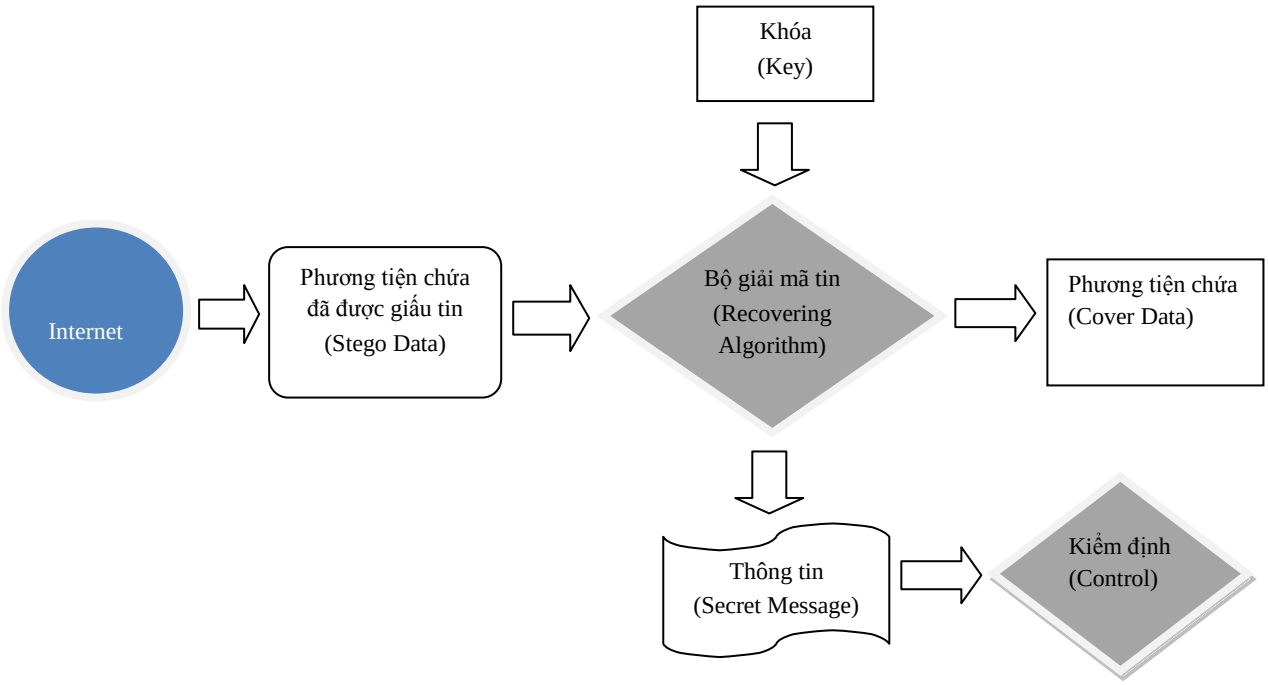
Thông tin sẽ được giấu vào trong phương tiện chứa nhờ một bộ nhúng, bộnhúng là những chương trình, triển khai các thuật toán để giấu tin và được thực hiện với một khoá bí mật giống như các hệ mật mã cổ điển. Sau khi giấu tin, ta thu được phương tiện chứa bản tin đã giấu và phân phối sử dụng trên mạng.



Hình 1.3. Sơ đồ biểu diễn quá trình giấu tin.

1.2.2. Tách tin

Sau khi nhận được đối tượng phương tiện chứa có giấu thông tin, quá trình tách tin được thực hiện thông qua bộ nhúng thông tin cùng với khoá của quá trình nhúng. Kết quả thu được gồm phương tiện chứa gốc và bản tin mật đã được giấu. Bước tiếp theo bản tin mật thu được sẽ được xử lý kiểm định so sánh với thông tin giấu ban đầu



Hình 1.4.Sơ đồ biểu diễn quá trình giải mã

1.3. Môi trường giấu tin

Kỹthuật giấu tin đã được nghiên cứu và áp dụng trong nhiều môi trường dữ liệu khác nhau như trong dữ liệu đa phương tiện (văn bản, hình ảnh, âm thanh, phim), trong sản phẩm phần mềm và gần đây là những nghiên cứu trên lĩnh vực cơ sở dữ liệu quan hệ. Trong các dữ liệu đó, dữ liệu đa phương tiện là môi trường chiếm tỉ lệ chủ yếu trong các kỹ thuật giấu tin.

1.3.1. Giấu tin trong ảnh (image)

Giấu thông tin trong ảnh, hiện nay, là một bộ phận chiếm tỉ lệ lớn nhất trong các chương trình ứng dụng, các phần mềm, hệ thống giấu tin trong đa phương tiện do lượng thông tin được trao đổi bằng ảnh là rất lớn và hơn nữa giấu thông tin trong ảnh cũng đóng vai trò quan trọng trong hầu hết các ứng dụng bảo vệ an toàn thông tin như: xác thực thông tin, xác định xuyên tạc thông tin, bảo vệ bản quyền tác giả, điều khiển truy cập, giấu tin bí mật... Do đó vấn đề này đã nhận được sự quan tâm lớn của các cá nhân, tổ chức, trường đại học, và viện nghiên cứu trên thế giới. Thông tin sẽ được giấu cùng với dữ liệu ảnh nhưng chất

lượng ảnh ít thay đổi và không ai biết được đằng sau ảnh đó mang những thông tin có ý nghĩa.

Ngày nay, khi ảnh số đã được sử dụng phổ biến, giấu thông tin trong ảnh đã đem lại nhiều những ứng dụng quan trọng trên nhiều lĩnh vực trong đời sống xã hội. Ví dụ đối với các nước phát triển, chữ kí tay đã được số hoá và lưu trữ sử dụng như hồ sơ cá nhân của các dịch vụ ngân hàng và tài chính, nó được dùng để xác thực trong các thẻ tín dụng của người tiêu dùng. Phần mềm WinWord của MicroSoft cũng cho phép người dùng lưu trữ chữ kí trong ảnh nhị phân rồi gắn vào vị trí nào đó trong file văn bản để đảm bảo tính an toàn của thông tin. Tài liệu sau đó được truyền trực tiếp qua máy fax hoặc lưu truyền trên mạng. Theo đó, việc xác thực chữ kí, xác thực thông tin đã trở thành một vấn đề quan trọng khi việc ăn cắp thông tin hay xuyên tạc thông tin bởi các tin tặc đang trở thành một vấn nạn đối với bất kì quốc gia nào, tổchức nào. Hơn nữa có nhiều loại thông tin quan trọng cần được bảo mật như những thông tin về an ninh, thông tin về bảo hiểm hay các thông tin về tài chính, các thông tin này được số hoá và lưu trữ trong hệ thống máy tính hay trên mạng. Chúng dễ bị lấy cắp và bị thay đổi bởi các phần mềm chuyên dụng. Việc xác thực cũng như phát hiện thông tin xuyên tạc đã trở nên vô cùng quan trọng, cấp thiết.

Một đặc điểm của giấu thông tin trong ảnh là thông tin được giấu trong ảnh một cách vô hình, tương tự cách truyền thông tin mật cho nhau mà người khác không thể biết được bởi sau khi giấu thông tin chất lượng ảnh gần như không thay đổi.

1.3.2. Giấu tin trong âm thanh(audio)

Giấu thông tin trong âm thanh mang những đặc điểm riêng khác với giấu thông tin trong các đối tượng đa phương tiện khác. Một trong những yêu cầu cơ bản của giấu tin là đảm bảo tính chất ẩn của thông tin được giấu đồng thời không làm ảnh hưởng đến chất lượng của dữ liệu gốc. Để đảm bảo yêu cầu này, kỹ thuật giấu thông tin trong ảnh phụ thuộc vào hệ thống thị giác của con người -HVS còn kỹ thuật giấu thông tin trong âm thanh lại phụ thuộc vào hệ thống thính giác - HAS.

Một vấn đề khó khăn là hệ thống thính giác của con người nghe được các tín hiệu ở các dải tần rộng và công suất lớn nên đã gây khó khăn đối với các phương pháp giấu tin trong âm thanh. Nhưng hệ thống thính giác của con người lại kém trong việc phát hiện sự khác biệt các dải tần và công suất, điều này có

nghĩa là các âm thanh to, cao tần có thể che giấu được các âm thanh nhỏ thấp một cách dễ dàng. Các mô hình phân tích tâm lí đã chỉ ra điểm yếu trên và thông tin này sẽ giúp ích cho việc chọn các âm thanh thích hợp cho việc giấu tin.

Vấn đề khó khăn thứ hai đối với giấu thông tin trong âm thanh là kênh truyền tin. Kênh truyền hay băng thông chậm sẽ ảnh hưởng đến chất lượng thông tin sau khi giấu. Giấu thông tin trong âm thanh đòi hỏi yêu cầu cao vềtính đồng bộ và tính an toàn của thông tin. Các phương pháp giấu thông tin trong âm thanh đều lợi dụng điểm yếu trong hệ thống thính giác của con người.

1.3.3. Giấu tin trong phim (video)

Giấu tin trong phim cũng được quan tâm và được phát triển mạnh mẽ cho nhiều ứng dụng như điều khiển truy cập thông tin,xác thực thông tin và bảo vệ bản quyền tác giả. Ví dụ các hệ thống chương trình trả tiền xem theo đoạn với các đoạn phim (pay per view application). Các kỹ thuật giấu tin trong phimcũng được phát triển mạnh mẽ và cũng theo hai khuynh hướng là thuỷ vân số và giấu thông tin. Nhưng phần này chỉ quan tâm tới các kỹ thuật giấu tin trong phim.

Một phương pháp giấu tin trong phim được Cox đưa ra là phương pháp phân bố đều. Ý tưởng cơ bản của phương pháp là phân phối thông tin giấu dàn trải theo tần số của dữ liệu gốc. Nhiều nhà nghiên cứu đã dùng những hàm cosin riêng và các hệ số truyền sóng riêng để giấu tin. Trong các thuật toán đầu tiên thường các kỹ thuật cho phép giấu các ảnh vào trong phim nhưng thời gian gần đây các kỹ thuật cho phép giấu cả âm thanh và hình ảnh vào phim.

1.3.4. Giấu thông tin trong văn bản dạng(text)

Giấu tin trong văn bản dạng text khó thực hiện hơn do có ít các thông tin dư thừa, để làm được điều này người ta phải khéo léo khai thác các dư thừa tự nhiên của ngôn ngữ. Một cách khác là tận dụng các định dạng văn bản (mã hóa thông tin vào khoảng cách giữa các từ hay các dòng văn bản).

Có thể sử dụng phương pháp Steganography trong tài liệu bằng cách đơn giản là thêm khoảng trắng và tab cho đến cuối dòng tài liệu. Loại steganography này là vô cùng hiệu quả bởi vì việc sử dụng khoảng trắng và tab sẽ không thể bị mắt thường phát hiện, ít nhất là đối với tất cả các trình biên tập văn bản tài liệu.

Kỹ thuật giấu tin đang được áp dụng cho nhiều loại đối tượng chứ không riêng gì dữ liệu đa phương tiện như ảnh, video, audio. Gần đây đã có một số nghiên cứu giấu tin trong cơ sở dữ liệu quân hệ, các gói IP truyền trên mạng, chắc chắn sau này còn phát triển tiếp cho các môi trường dữ liệu số khác.

1.3.5. Độ an toàn của một hệ thống giấu tin

Việc phá vỡ một hệ thống giấu tin thông thường gồm ba phần: phát hiện, giải tin và huỷ thông tin đã giấu. Một hệ thống giấu tin mật được gọi là thực sự an toàn khi kẻ tấn công không phát hiện được sự tồn tại của thông tin giấu trong một đối tượng chứa. Trong khi phát triển một hệ giấu tin mật, người ta phải luôn luôn cho rằng kẻ tấn công có năng lực tính toán và sẵn sàng làm đủ mọi cách để phá vỡ tính an toàn của hệ thống. Nếu kẻ tấn công không thể chắc chắn một đối tượng có được giấu tin hay không thì theo lý thuyết, hệ thống đó là an toàn.

1.4. Đặc điểm giấu tin

Hiện nay giấu thông tin còn tương đối mới và đang có xu hướng phát triển rất nhanh.

Một kỹ thuật giấu tin được đánh giá dựa trên một số đặc điểm sau:

- Tính ẩn của thông tin.
- Tính bảo mật và an toàn của thông tin.
- Số lượng thông tin được giấu.
- Chất lượng vật chứa thông tin sau khi nhúng.

1.4.1. Tính ẩn của thông tin

Khái niệm này dựa trên đặc điểm của hệ thống thị giác của con người. Thông tin nhúng là không được phát hiện nếu dựa vào các giác quan của người bình thường. Hiểu biết về hệ thống giác quan con người sẽ góp một phần không nhỏ trong việc cải tiến và nâng cấp các thuật toán về giấu tin. Để nhận biết một thông tin não người phải xử lý các thông tin thu nhận như vị trí không gian, đường nét, màu sắc của thông tin.

1.4.2. Tính bảo mật và an toàn của thông tin

Nếu ai đó có thể dễ dàng phát hiện nơi bạn đã giấu thông tin của bạn và tìm thấy tin nhắn của bạn, nó đánh bại mục đích của việc sử dụng steganography. Vì vậy, các thuật toán được sử dụng phải đủ mạnh mẽ rằng ngay cả khi ai đó biết làm thế nào các công trình kỹ thuật họ không thể dễ dàng tìm ra rằng bạn đã ẩn dữ liệu trong một file nhất định.

Thuật toán nhúng tin được coi là có tính bảo mật nếu thông tin được nhúng không bị tìm ra khi bị tấn công một cách có chủ đích trên cơ sở những hiểu biết đầy đủ về thuật toán nhúng tin và có bộ giải mã (trừ khóa bí mật).

1.4.3. Số lượng thông tin được giấu

Lượng thông tin giấu so với kích thước của môi trường là một vấn đề cần quan tâm trong một thuật toán giấu tin. Đây là một trong hai yêu cầu cơ bản của giấu tin mật. Ví dụ rõ ràng là có thể chỉ giấu một bit thông tin vào mỗi ảnh mà không cần lo lắng về độ nhiễu của ảnh nhưng như vậy sẽ rất kém hiệu quả khi mà không cần lo lắng về độ nhiễu của ảnh nhưng như vậy sẽ rất kém hiệu quả khi mà thông tin cần giấu có kích thước bằng Kb. Các thuật toán đều cố gắng đạt được mục đích làm thế nào giấu được nhiều thông tin nhất mà không gây ra nhiễu đáng kể.

1.4.4. Chất lượng của phương tiện chứa sau khi nhúng

Sau khi giấu thông tin bên trong, phương tiện chứa phải đảm bảo được yêu cầu không bị biến đổi để có thể bị phát hiện dễ dàng so với trước khi nhúng. Ví dụ trong việc ẩn thông tin trong ảnh yêu cầu này khá đơn giản đối với ảnh màu hoặc ảnh xám bởi mỗi một pixel ảnh (picture element) được biểu diễn bởi nhiều bit, nhiều giá trị và khi thay đổi một giá trị nhỏ nào đó thì chất lượng ảnh không thay đổi, thông tin giấu khó bị phát hiện. Đối với ảnh đen trắng thì việc giấu thông tin phức tạp hơn nhiều, vì mỗi pixel ảnh đen trắng chỉ gồm một trong hai giá trị hoặc trắng hoặc đen, và nếu ta biến đổi một bit từ đen thành trắng thì rất dễ bị phát hiện.

Vì phương pháp giấu tin trong môi trường nhúng dựa trên việc điều chỉnh các giá trị của các bit theo một qui tắc nào đó và khi giải mã sẽ theo các giá trị đó để tìm được thông tin giấu, cho nên nếu một phép biến đổi nào đó trong tệp chứa tin làm thay đổi giá trị của các bit thì sẽ làm cho thông tin giấu bị sai lệch. Chính đặc điểm này mà giấu thông tin trong tệp có tác dụng nhận thực và phát hiện sai lệch thông tin. Trên đây là những tính chất và đặc điểm cơ bản chung của giấu tin trong ảnh. Riêng đối với ứng dụng giấu tin mật (steganography) thì các tính chất ẩn, lượng thông tin giấu và độ an toàn là ba tính chất quan trọng nhất.

Chương 2: GIẤU TIN TRONG TỆP VĂN BẢN

2.1. Tệp văn bản

2.1.1. Khái niệm tệp văn bản

Theo nghĩa rộng, văn bản được hiểu là vật mang tin được ghi bằng ký hiệu hay bằng ngôn ngữ, nghĩa là bất cứ phương tiện nào dùng để ghi nhận và truyền đạt thông tin từ chủ thể này đến chủ thể khác. Theo cách hiểu này, bia đá, hoành phi, câu đối ở đền, chùa; chúc thư, văn khế, thư tịch cổ; tác phẩm văn học hoặc khoa học kỹ thuật; công căn, giấy tờ khẩu hiệu, băng ghi âm, bản vẽ… ở cơ quan đều được gọi là văn bản. Khái niệm này được sử dụng một cách phổ biến trong giới nghiên cứu về văn bản học, ngôn ngữ học, sử học ở nước ta từ trước tới nay.

Theo nghĩa hẹp, văn bản được hiểu là các tài liệu, giấy tờ, hồ sơ được hình thành trong quá trình hoạt động của các cơ quan nhà nước, các tổ chức xã hội, các tổ chức kinh tế. Theo nghĩa này, các loại giấy tờ dùng để quản lý và điều hành các hoạt động của cơ quan, tổ chức như chỉ thị, thông tư, nghị quyết, quyết định, đề án công tác, báo cáo… đều được gọi là văn bản. Ngày nay, khái niệm được dùng một cách rộng rãi trong hoạt động của các cơ quan, tổ chức. Khái niệm văn bản dùng trong tài liệu này cũng được hiểu theo nghĩa hẹp nói trên.

Tập tin (viết tắt cho tập thông tin; còn được gọi là tệp, tệp tin, file) là một tập hợp của thông tin được đặt tên. Thông thường thì các tập tin này chứa trong các thiết bị lưu trữ như đĩa cứng, đĩa mềm, CD, DVD cũng như là các loại chip điện tử dùng kĩ thuật flash có thể thấy trong các ổ nhớ có giao diện USB. Nói cách khác, tập tin là một dãy các bit có tên và được chứa trong các thiết bị lưu trữ dữ liệu kỹ thuật số. Tệp văn bản là tệp tin chứa các tài liệu, giấy tờ, hồ sơ văn bản.

2.1.2. Đặc điểm của tệp văn bản

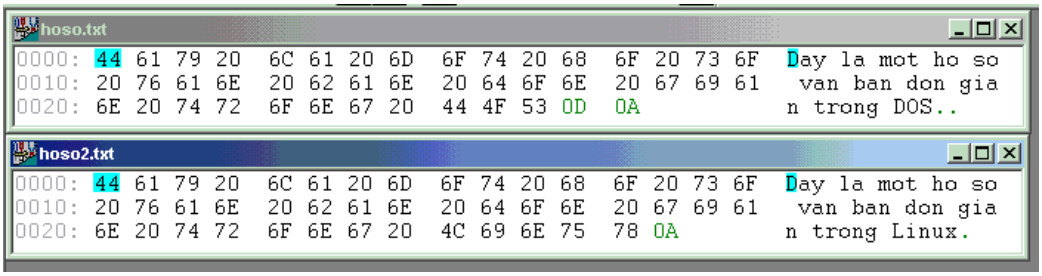
Một tệp tin văn bản có đầy đủ các điểm cơ bản như các tệp tin khác như:

Một tập tin luôn luôn kết thúc bằng 1 ký tự đặc biệt (hay dấu kết thúc) có mã ASCII là 255 ở hệ thập phân. Ký tự này thường được ký hiệu là EOF (từ chữ End of File).

Một tập tin có thể không chứa một thông tin nào ngoại trừ tên và dấu kết thúc. Tuy nhiên, điều này không hề mâu thuẫn với định nghĩa vì bản thân tên của tập tin cũng đã chứa thông tin. Những tập tin này gọi là tập tin rỗng hay tập tin trống.

Độ dài (kích thước) của tập tin có thể chỉ phụ thuộc vào khả năng của máy tính, khả năng của hệ điều hành cũng như vào phần mềm ứng dụng dùng nó. Đơn vị nhỏ nhất dùng để đo độ dài của tập tin là byte. Độ dài của tập tin không bao gồm độ dài của tên tập tin và dấu kết thúc.

Tệp văn bản là tệp tin được cấu thành bởi các ký tự theo chuẩn ASCII.Điểm đặc biệt là dữ liệu của tệp tin được lưu trữ thành các dòng, mỗi dòng được kết thúc bằng ký tự xuống dòng (new line), ký hiệu '\n', ký tự này là sự kết hợp của 2 ký tự CR (Carriage Return – Về đầu dòng, mã ASCII là 13) và LF (Line Feed – Xuống dòng, mã ASCII là 10). Mỗi tệp tin được kết thúc bởi ký tự EOF (End of File) có mã ASCII là 26 (xác định bởi tổ hợp phím Ctrl + Z). Tệp tin văn bản có thể truy suất theo kiểu tuần tự.



Hình 2.1. Cấu trúc bit của tập tin ACSII

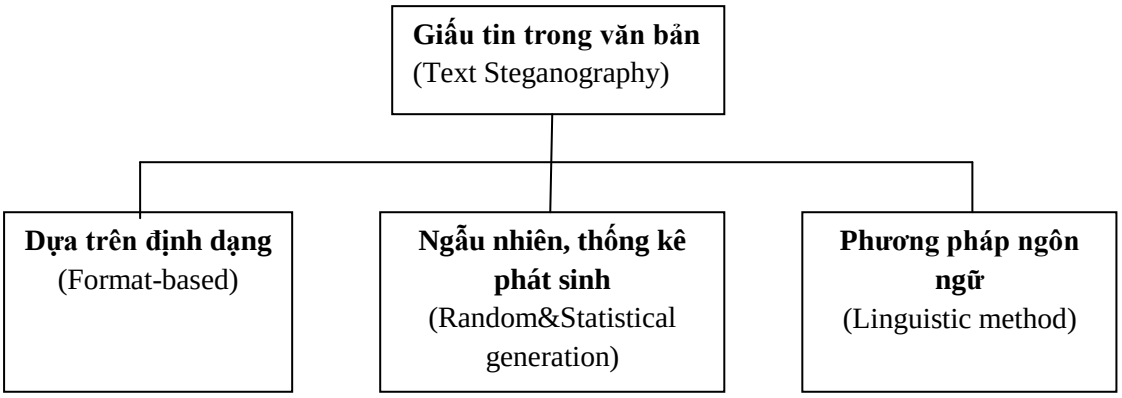
Một tên tệp tin nói chung gồm hai phần, chúng được ngăn cách với nhau bằng một dấu chấm. Phía bên trái dấu chấm là tên riêng, còn phía bên phải là phần mở rộng (extension). Phần mở rộng này cho thấy mục đích sử dụng của tệp tin. Các tệp văn bản thường có phần mở rộng là: txt, doc, wtf, pdf, .v.v.v

Tập tin văn bản dùng cho các mục tiêu khác nhau cũng sẽ có các định dạng khác nhau. Ngoài sự ràng buộc về định dạng của hệ điều hành, các tập tin dùng trong các ứng dụng hay các phần mềm khác nhau cũng sẽ khác nhau và sự khác nhau này tùy thuộc vào kiến trúc của các ứng dụng sử dụng các tập tin đó.

2.2. Giấu tin trong tệp văn bản (TextSteganography)

2.2.1. Các phân lớp của kỹ thuật giấu tin trong tệp văn bản

Kỹ thuật giấu tin trong tệp văn bản, được phân thành ba phần cơ bản: **Dựa vào định dạng** (Format-based). **Ngẫu nhiên&Thống kê phát sinh** (Random&Statistical generation). **Phương pháp ngôn ngữ** (Linguistic method).



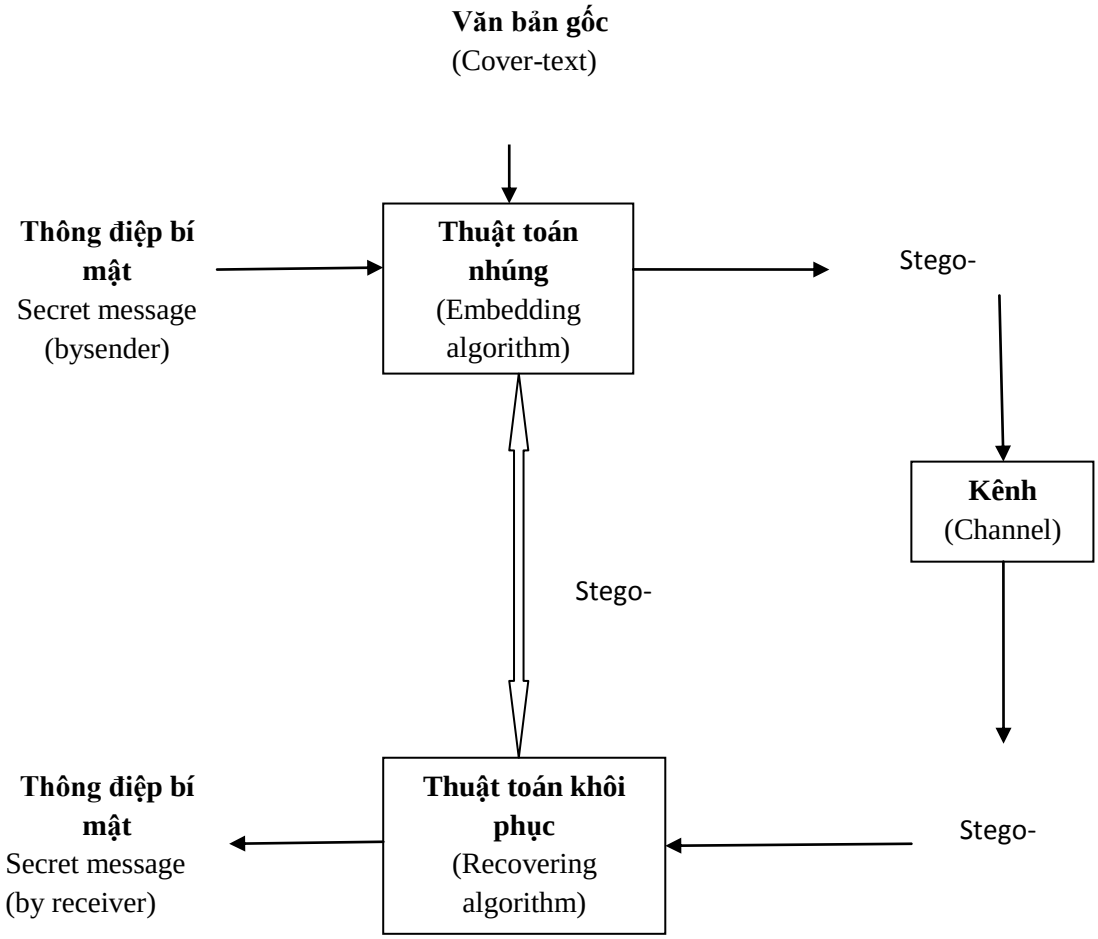
Hình 2.2. Ba loại cơ bản của Text Steganography

Phương pháp dựa vào định dạng sử dụng định dạng vật lý của văn bản như một nơi để che giấu thông tin. Nói chung, phương pháp này thay đổi văn bản hiện tại để ẩn thông tin. Chèn vào khoảng trống, cố ý để lỗi chính tả trải khắp văn bản, thay đổi kích thước các phông chữ là một số trong nhiều phương pháp dựa trên định dạng được sử dụng. Tuy nhiên, Bennett đã tuyên bố rằng các phương pháp dựa vào định dạng cố đánh lừa đôi mắt của con người, nhưng nó không thể lừa hệ thống máy tính một khi đã được sử dụng.

Ngẫu nhiên & thống kê phát sinh là tạo ra văn bản gốc theo đặc tính thống kê. Phương pháp này dựa trên các chuỗi ký tự vàcác từ chuỗi. Ẩn thông tin trong chuỗi ký tự là nhúng thông tin được xuất hiện theo thứ tự ngẫu nhiên của các ký tự. trình tự này phải xuất hiện ngẫu nhiên với bất cứ ai chặn thông điệp. Cách tiếp cận thứ hai sinh ra các ký tự là lấy thống kê thuộc tính của độ dài từ và tần suất bức thư để tạo "chữ" (không có giá trị từ vựng) sẽ xuất hiện để có cùng một thống kê thuộc tính từ thực tế trong một ngôn ngữ nhất định. Cất giấu thông tin trong phạm vi trình tự từ, các từ trong từ điển thực tế có thể được sử dụng để mã hóa một hoặc nhiều bit thông tin trên mỗi từ bằng cách ánh xạ giữa các đơn vị từ vựng và trình tự bit, hoặc các từ mà nó có thể mã hóa ẩn thông tin.

Loại cuối cùng là **phương pháp ngôn ngữ** xem xét các thuộc tính của ngôn ngữ được tạo ra và văn bản sửa đổi, thường xuyên sử dụng cấu trúc ngôn ngữ như là một nơi cho các thông điệp ẩn. Trong thực tế, dữ liệu có thể được ẩn giấu bên trong cấu trúc cú pháp của chính nó.

2.2.2. Cơ chế cơ bản của Text Steganography



Hình 2.3. Cơ chế cơ bản của Text Steganography

Theo như hình 2.3. Thứ nhất, một thông điệp bí mật (hoặc một dữ liệu nhúng) sẽ được giấu trong một văn bản gốc (cover-text) bằng cách áp dụng một thuật toán nhúng để tạo một văn bản giấu tin (stego –text). Các stego-text sau đó sẽ được truyền một kênh truyền thông, ví dụ như Internet hoặc thiết bị điện thoại di động để đến người nhận. Để khôi phục bí mật mà được gửi bởi người gửi, người nhận cần phải sử dụng một thuật toán khôi phụcđược biểu hiện bằng tham số bởi một khóa giấu tin (stego-key) để tách thông điệp ẩn. Một stego-key được dùng để điều khiển các tiến trình ẩn nhằm hạn chế phát hiện và phục hồi dữ liệu nhúng.

2.3. Kỹ thuật giấu tin trong tệp văn bản

Che giấu thông tin trong khoảng trống có vẻ là tiềm năng khi người ta khó có thể biết về sự tồn tại của các bit ẩn. Bender và cộng sự đã chỉ ra rằng một khoảng trống dịch là "0", trong khi hai khoảng trống được hiểu như là "1". Chương trình nhúng đã được áp dụng trong khoảng trống xuất hiện giữa các từ. Hạn chế lớn nhất của phương pháp Bender này là nó đòi hỏi rất nhiều khoảng trống để mã hóa vài bit. Ví dụ, một ký tự tương đương với 8 bit, và nó đòi hỏi khoảng 8 khoảng trống để mã hóa một ký tự. Như vậy, vấn đề này có thể được giải quyết nếu chúng ta nén các ký tự trong thông điệp bí mật từ 8 bit về 3 bit hoặc ít hơn trong phương pháp được đề xuất. Bằng cách kết hợp với đoạn văn, ẩn các bit bí mật có thể có hiệu quả bằng cách sử dụng hầu hết các khoảng trắng trong một tài liệu văn bản. Trong phương pháp nhúng, hai khoảng trống mã hóa một bit trên mỗi dòng, bốn mã hóa hai, tám mã hóa 3,v.v.v, dần dần tăng số lượng thông tin chúng ta muốn mã hóa.

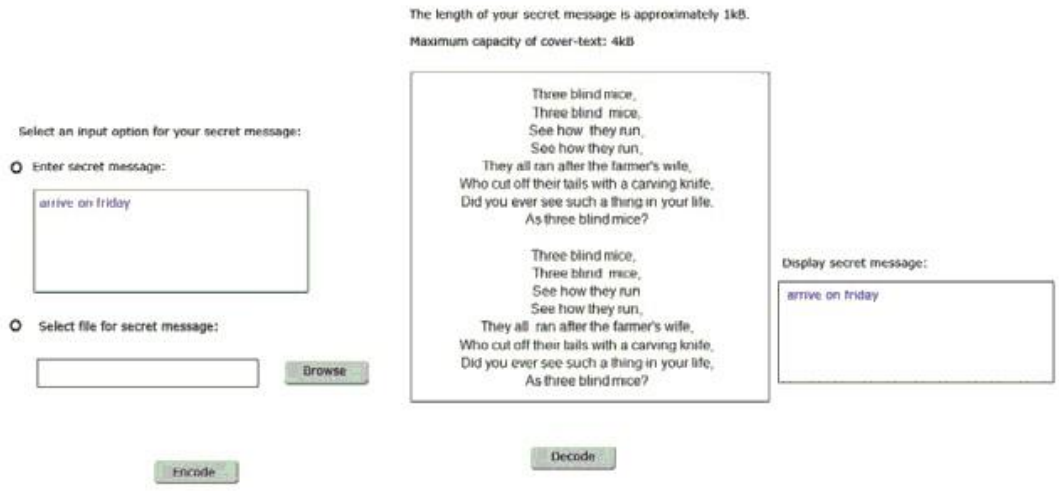
Hiện tại, thao tác với các khoảng trắng dường như có lợi và có tiềm năng trong ẩn thông tin vì khoảng trắng xuất hiện trong một tài liệu văn bản nhiều hơn sự xuất hiện của từ. Nó thậm chí còn là một lợi thế khi sẽ không có ai biết rằng một mảnh trống của tài liệu là thông tin bí mật thực sự quan trọng.

Các văn bản gốc sẽ được tự động tạo ra theo chiều dài của thông điệp bí mật. Công suất tối đa của các bit ẩn được xác định bởi hệ thống cho dù chiều dài của thông điệp bí mật có thể được cung cấp trong 4Kb, 16Kb, 32Kb, 64Kb,128Kb hay 256Kb. 4Kb được sử dụng như là một giới hạn thấp hơn công suất trong số các tùy chọn bởi vì đầu vào tối thiểu của văn bản do người dùng đã cân nhắc (ví dụ như một thông điệp bí mật có thể chỉ chứa một vài ký tự).

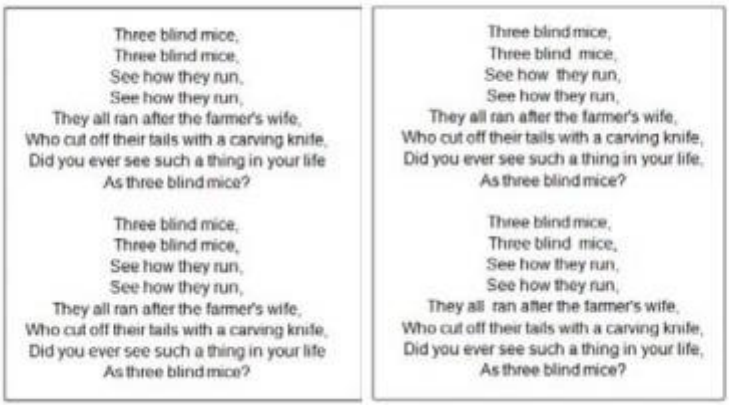
Ý nghĩa nghiên cứu của Bender là do kết hợp hai khái niệm và tạo ra một thuật toán cho dữ liệu hệ thống nhúng. Công suất của các văn bản là tùy thuộc vào độ dài của thông điệp bí mật. Dựa trên hình 2.3, một người dùng được yêu cầu cung cấp thông điệp bí mật trong phạm vi văn bản hoặc bằng lựa chọn một tập tin có chứa bí mật. Hệ thống sẽ tính toán chiều dài của thông điệp và tạo ra một văn bản chứa (cover-text) phù hợp có thể mã hóa thông điệp bí mật. Như minh họa trong hình 2.5, độ dài của thông điệp bí mật là khoảng 1KB và có thể phù hợp với các cover-text, trong đó công suất là 4kB. Sau đó, thông điệp bí mật đã được nhập và mã hóa. Cuối cùng, các văn đã giấu tin tạo ra trong hình 2.6. Khi một người dùng khác nhận được thông điệp này, các bit ẩn có thể được tách ra bằng cách sử dụng các thuật toán phục hồi.

Bảng 2.1 Văn bản chứa theo chiều dài của thông điệp

Độ dài văn bản (kB)	
Secret	Cover
<4	4
<16	16
<**32**	**32**
<64	64
<128	128
<256	356



Hình 2.4 Văn bản gốc và thông điệp ẩn



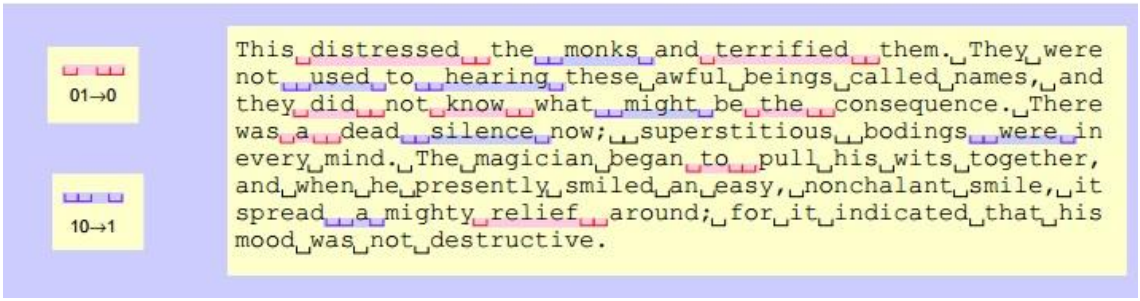
Hình 2.5 Văn bản gốc (bên trái), văn bản giấu thông điệp ẩn (bên phải)

2.3.1. Ý tưởng của kỹ thuật

Brassil và đồng sự đưa ra ý tưởng ban đầu của phương pháp giấu tin trongbài viết của mình bằng cách đề nghị giấu tin dịch chuyển trái (life-shift coding), mã hóa dịch chuyển từ (word-shift coding) và giấu tin ký tự (feature coding) để ngăn cản phổ biến bất hợp pháp của tài liệu trên mạng máy tính. Giấu tin Line-shift là phương pháp thay đổi tài liệu bằng cách chuyển vị trí của các dòng văn bản theo chiều dọc để giấu tin. Giấu tin Word-shift là phương pháp để thay đổi tài liệu bằng cách chuyển vị trí của từ theo chiều ngang trong dòng văn bản để giấu. Giấu tin ký tự hoặc giấu tin đặc trưng cụ thể là phương pháp chỉ áp dụng cho hình ảnh bitmap của tài liệu và có thể được kiểm tra để lựa chọn ký tự đặc trưng và những đặc trưng này được thay đổi, hoặc không bị thay đổi, tùy thuộc vào từ mã. Một tài liệu được đánh dấu theo cách không thể nhận ra bằng từ mã xác định đăng ký chủ sở hữu mà tài liệu được gửi đi. Nếu một bản sao tài liệu được tìm thấy thì sẽ bị nghi ngờ là bất hợp pháp, rằng bản sao có thể được giải mã và xác định chủ sở hữu đã đăng ký.

Bender và đồng sự cho là ba phương pháp giấu tin định dạng bằng văn bản khi họ thực hiện nghiên cứu về kỹ thuật ẩn dữ liệu trong văn bản. Ba phương pháp này là các phương pháp khoảng trống mở để mã hoá thông qua thao tác các khoảng trắng, cú pháp phương pháp sử dụng dấu chấm câu và ngữ nghĩa phương pháp mã hoá bằng cách sử dụng thao tác của từ đồng nghĩa. Phương pháp khoảng trống mở lấy khoảng trống giữa các câu (inter-sentence spacing), khoảng trống cuối dòng (end-of-line spaces) và khoảng trống giữa các từ (inter-word spacing) trong văn bản hợp lý. Phương pháp Inter-sentence spacing là để giấu một thông điệp nhị phân vào một văn bản bằng cách đặt một hoặc hai khoảng trống ở cuối của mỗi ký tự kết thúc. Nếu giấu "0" bằng thêm một khoảng trống duy nhất, giấu "1" bằng cách thêm hai khoảng trống. Phương pháp này hoạt động, bất lợi là không hiệu quả bởi vì nó đòi hỏi một số lượng lớn văn bản để giấu rất ít bit. Một bit trên một câu tương đương với tỉ lệ dữ liệu của khoảng một bit mỗi 160 byte, giả sử trung bình mỗi câu gồm hai dòng 80-ký tự của văn bản. Phương pháp này hoàn toàn phụ thuộc vào cấu trúc của văn bản. Phương pháp End-of-line space khai thác các khoảng trắng ở cuối mỗi dòng. Dữ liệu được giấu bằng cách định trước số lượng khoảng trống ở cuối của mỗi dòng. Ví dụ hai khoảng trống sẽ giấu được một bit, bốn khoảng trống sẽ giấu được hai bit, tám khoảng trống sẽ giấu đước ba bit. Nó hoạt động tốt hơn phương pháp inter-space bởi vì nó tăng số lượng khoảng trống có thể ẩn nhiều dữ liệu hơn.

Phương pháp thứ ba sử dụng khoảng trắng để giấu dữ liệu liên quan đến việc căn lề phải của văn bản, cũng có thể được sử dụng để giấu dữ liệu với các tệp tin văn bản. Dữ liệu được giấu bằng cách kiểm soát nơi mà các khoảng trống được đặt thêm. Một khoảng trống giữa hai từ được dịch là "0". Hai khoảng trống được dịch là "1". Không phải mọi không gian giữa hai từ có thể được sử dụng như dự liệu do khó khăn khi căn chỉnh. Bender và đồng sự sử dụng phương pháp Machester như một phương pháp để xác định khoảng trống các từ thể hiện các bit dữ liệu ẩn và đó là một phần của văn bản gốc. "01" được xem như là "0" và "10" là "1".Các chuỗi bit "00" và "11" là null..



Hình 2.6.Văn bản giấu theo phương pháp Machester

Bảng 2.2.Bảng mã theo phương pháp Machester

Mã hóa	Giá trị
01	0
10	1
00	null
11	null

2.3.2. Thuật toán giấu tin dựa vào khoảng trống của văn bản Text

Đầu vào:

-Văn bản gốc C (cover-text)

-Thông tin cần giấu M (Secret Message).

Đầu ra:

-Văn bản sau khi nhúng thông tin S (stego-text)

Các bước của thuật toán.

Bước một:

- Mã hóa thông tin M thành mã nhị phân MSG dựa vào bảng mã hóa (trong phụ lục A). Mỗi ký tự của chuỗi thông điệp tương ứng là một giá trị mã hóa trong phụ lục A. Chuyển giá trị mã hóa của mỗi ký tự từ thập phân sang nhị phân ta được một dãy nhị phân 6 bit. Ghép các dãy nhị phân lại với nhau ta được mã nhị phân MSG của thông điệp.
- Chia chuỗi ký tự văn bản gốc C thành các chuỗi ASCII $A_1A_2A_3A_4…A_n$. Mỗi chuỗi gồm K khoảng trống. Theo đó mỗi chuỗi sẽ giấu 1 bit mã nhị phân MSG. Tức là cứ K khoảng trống sẽ giấu một bit nhị phân. Vì giá trị K do người giấu tin đặt nên tăng tính bảo mật và khó giải mã thông điệp hơn.

Bước hai:

- N: Tổng số khoảng trống của văn bản gốc C.
- N/K : Là tổng số các chuỗi con được tạo ra.
- Kiểm tra điều kiện:
  - + Nếu N/K>= Length(MSG) : Thì thực hiện nhúng
  - + Nếu N/K<Length(MSG): Chương trình báo lỗi do số bit thông điệp nhị phân lớn hơn số chuỗi ACSII nên không thể nhúng hết mã nhị phân.

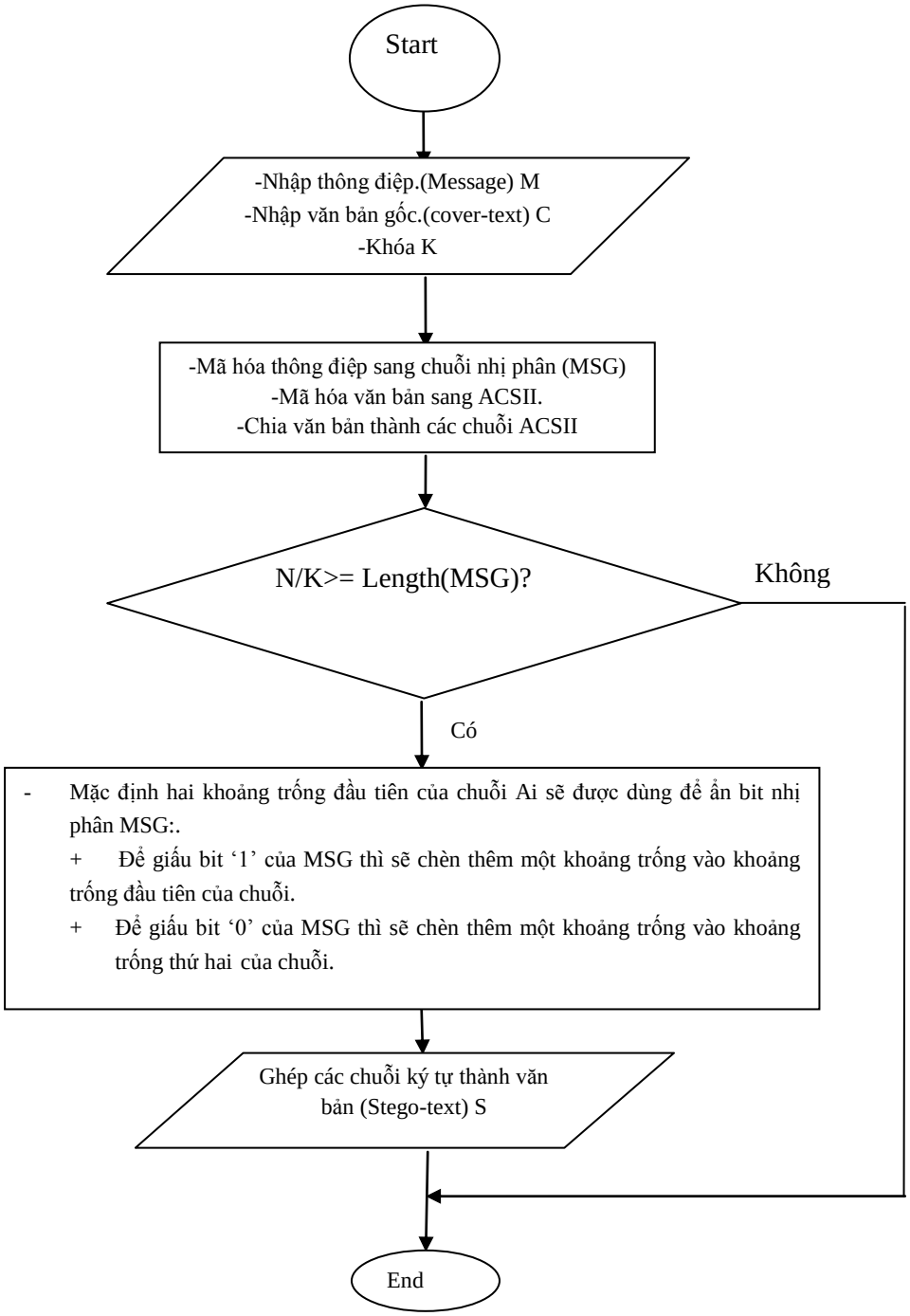
Bước ba:

Mỗi chuỗi A_i giấu được 1 bit thông điệp. Theo lần luợt các công đoạn sau:

- Mã hóa bit thông điệp (Cần 2 khoảng trống để ẩn một bit mã hóa. Do vậy số khoảng trống mỗi chuỗi K>=2). Mặc định hai khoảng trống đầu tiên của chuỗi A_i sẽ được dùng để ẩn bit.
  - + Để giấu bit ‘1’ của MSG thì sẽ chèn thêm một khoảng trống vào khoảng trống đầu tiên của chuỗi.
  - + Để giấu bit ‘0’ của MSG thì sẽ chèn thêm một khoảng trống vào khoảng trống thứ hai của chuỗi.

Quá trình giấu dừng lại cho đến khi giấu hết các bit thông điệp (MSG) vào các chuỗi. Ghép lại các chuỗi ký tự ta được văn bản sau khi nhúng S.

Hình 2.7 Lưu đồ thuật toán nhúng thông điệp



Hình 2.7. Lưu đồ giải thuật nhúng thông điệp

2.3.3. Thuật toán tách tin

Đầu vào:

- Văn bản đã nhúng (Stego-text) S

Đầu ra:

- Thông điệp được nhúng M.

Các bước của thuật toán.

Bước một:

- Chuyển văn bản S sang mã ACSII.
- Từ mã ACSII tách mã nhị phân của văn bản bằng theo nguyên tắc:
  + Một khoảng trống giữa hai từ được dịch là "0".
  + Hai khoảng trống giữa hai từ được dịch là "1".

Bước hai:

Từ mã nhị phân của văn bản ta mã nhị phân này thành các chuỗi nhị phân con, mỗi chuỗi gồm K bit.

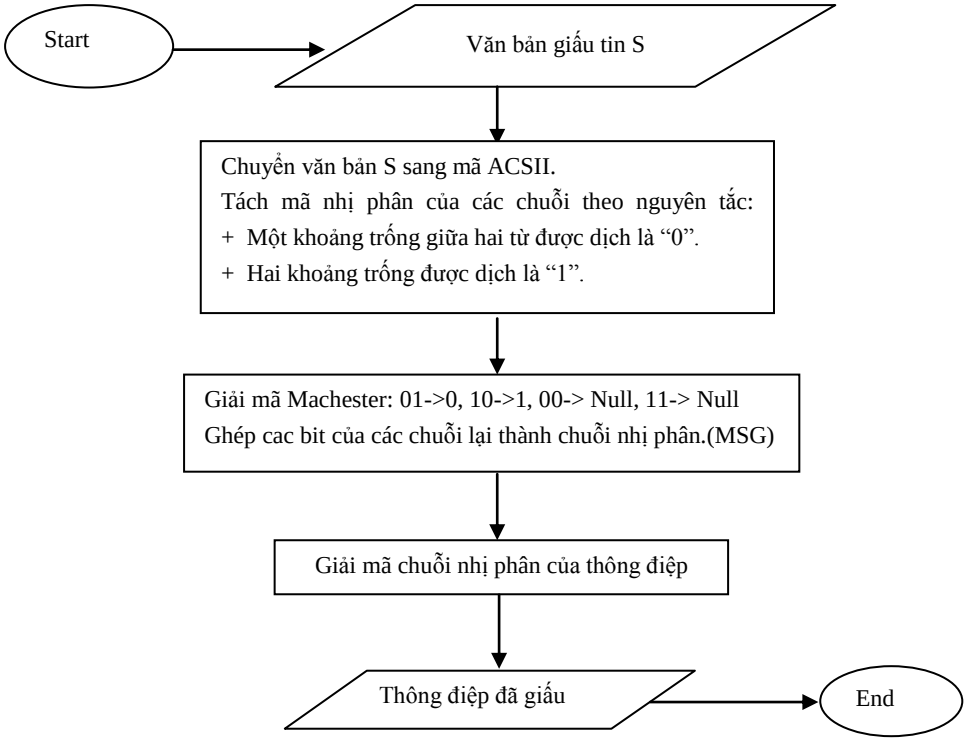
- Xét hai bit đầu tiên của các chuỗi nhị phân
- Theo bảng mã hóa Machester 01->0, 10->1, 00-> Null, 11-> Null

Với mỗi chuỗi ta được 1 bit thông điệp đã giấu, ghép các bit này lại ta được một chuỗi nhị phân MSG của thông điệp.

Bước ba:

- Dựa vào bảng mã hóa nhị phân (Phụ lục A). Giải mã (chia MSG thành từng chuỗi con 6 bit, chuyển chuỗi bit này từ nhị phân sang thập phân ta được một giá trị mã hóa. Ứng với một giá trị mã hóa trong bảng phụ lục A là một ký tự) ta có được thông điệp M đã giấu trong tệp văn bản S.

Hình 2.8 Lưu đồ giải thuật tách tin



Hình 2.8 Lưu đồ giải thuật tách tin

2.3.4. Ví dụ minh họa

Thông tin cần giấu: DH DL HP

Khóa : K=3

Tệp Văn bản che giấu có nội dung:

On a recent Saturday, Johan Dijkland, a 23-year-old student in Emmen, Netherlands, opened a free messaging app called Line on his iPhone. Then he tapped on a virtual sticker of a sleepy panda with a "good night" speech bubble and pressed send to a friend. With that action, Mr. Dijkland's text joined the tens of billions of messages that are processed every day from a fast-growing crowd of mobile messaging apps.These messaging apps—with funny names like WhatsApp, WeChat and KakaoTalk—have become an indispensable form of communication for hundreds of millions of people world-wide.They are also rankling technology giants from Silicon Valley to Seoul. That is because when users like Mr. Dijkland send messages using Line, his mobile carrier Vodafone Group PLC and iPhone maker Apple Inc. don't directly profit from the interaction.

- Hỏi sau khi giấu tin được văn bản giấu như thế nào?

Bước một.

-Mã hóa thông tin.

Bảng 2.2 Giá trị thông điệp theo ASCII

stt	Ký tự	ASCII
1	D	68
2	H	72
3	Khoảng trắng	32
4	D	68
5	L	76
6	Khoảng trắng	32
7	H	72
8	P	80

Để mã hóa ký tự thông điệp 8 ký tự mỗi ký tự cần 8 bits. Nhưng có thể dùng thuật toán nén dữ liệu chỉ dùng 6 bits cho mỗi ký tự thay vì 8 bits. Bởi vì trong bảng mã ACSII chỉ có 128 ký tự chuẩn (có thể hiển thị). Bằng cách tạo ra một bảng giá trị có thể ánh xạ vào giá trị của ACSII (theo bảo phụ lục A).

Bảng 2.3 Mã hóa thông điệp

stt	Ký tự	ASCII	Bảng mã	Mã hóa
1	D	68	20	010100
2	H	72	23	010111
3	Khoảng trắng	32	16	010000
4	D	68	20	010100
5	L	76	34	100010
6	Khoảng trắng	32	16	010000
7	H	72	23	010111
8	P	80	24	100110

- Ta được dữ liệu mã hóa nhị phân MSG là: 010100 010111 010000 010011 100010 010000 010111 100110.
- Chia chuỗi ký tự văn bản gốc thành các chuỗi ASCII $A_1A_2A_3A_4…A_n$. Mỗi chuỗi gồm K khoảng trống. Theo đó ta sử dụng vòng lặp và phép gán sao cho cứ 3 khoảng trống là tạo thành một chuỗi. Theo ví dụ ta có.

  A_1=”*79 110 32 97 32 114 101 99 101 110 116 32*”

  A_2= *“83 97 116 117 114 100 97 121 44 32 74 111 104 97 110 32 68 105 106 107 108 97 110 100 44 32 “*

  A_3=”*97 32 50 51 45 121 101 97 114 45 111 108 100 32 115 116 117 100 101 110 116 32*”

  ……

Bước hai:

- Tính độ dài thông điệp: length(MSG)=48
- Văn bản bao gồm: 188 khoảng trống.
- K=3.

188/3>48

Vậy văn bản hoàn toàn có thể giấu thông thông điệp. Chuyển sang nhúng.

Bước 3. Nhúng dữ liệu

Mỗi chuỗi A_i giấu được 1 bit thông điệp. Theo lần lượt các công đoạn sau:

- Mã hóa bit thông điệp theo (Cần 2 khoảng trống để ẩn 1 bit. Do vậy số khoảng trống mỗi chuỗi K>=2). Mặc định hai khoảng trống đầu tiên của chuỗi A_i sẽ được dùng để ẩn bit.
  + Để giấu bit '1' của MSG thì sẽ chèn thêm một khoảng trống vào khoảng trống đầu tiên của chuỗi.
  + Để giấu bit '0' của MSG thì sẽ chèn thêm một khoảng trống vào khoảng trống thứ hai của chuỗi.

Quá trình giấu dừng lại cho đến khi giấu hết các bit thông điệp (MSG) vào các chuỗi. Ghép lại các chuỗi ký tự ta được văn bản sau khi nhúng S.

Theo ví dụ ta giấu bit đầu tiên của MSG là '0' thì sẽ chèn thêm khoảng trống (32) vào khoảng trống thứ hai của chuỗi. A_1 sau khi nhúng là:

A_1="*79 110 32 97 32* ***32*** *114 101 99 101 110 116 32*"

Tiếp theo giấu bit tiếp theo của MSG là '1' bằng cách chèn thêm một khoảng trống vào khoảng trống đầu tiên của A_2 . A_2 sau khi nhúng là:

A_2= *"83 97 116 117 114 100 97 121 44 32* ***32*** *74 111 104 97 110 32 68 105 106 107 108 97 110 100 44 32 "*

Lặp quá trình giấu tin cho đến khi giấu hết các bit thông điệp (MSG) vào các chuỗi. Ghép lại các chuỗi ký tự ta được văn bản sau khi nhúng. Ta có văn bản đã giấu tin.

On a recent Saturday, Johan Dijkland, a 23-year-old student in Emmen, Netherlands, opened a free messaging app called Line on his iPhone. Then he tapped on a virtual sticker of a sleepy panda with a "good night" speech bubble and pressed send to a friend. With that action, Mr. Dijkland's text joined the tens of billions of messages that are processed every day from mmen, Netherlands, opened a free messaging app called Line on his iPhone. Then he tapped on a virtual sticker of a sleepy panda with a "good night" speech bubble and pressed send to a friend. With that action, Mr. Dijkland's text joined the tens of billions of messages that are processed every day from a fast-growing crowd of mobile messaging apps.These messaging apps—with funny names like WhatsApp, WeChat and KakaoTalk—have become an indispensable form of communication for hundreds of millions of people world-wide.They are also rankling technology giants from Silicon Valley to Seoul. That is because when users like Mr. Dijkland send messages using Line, his mobile carrier Vodafone Group PLC and iPhone maker Apple Inc. don't directly profit from the interaction.

Tách tin chúng ta làm như sau:

Bước 1:Chuyển các ký tự văn bản sang mã thập phân ASCII .

Từ mã thập phân ASCII của văn bản ta tách mã nhị phân văn bản bằng cách cứ 1 khoảng trống (giá trị thập phân là 32) được dịch là '0', hai khoảng trống được dịch là '1'. Ta được dãy nhị phân là:

010 100 010 100 010 010 010 100 010 100 010 100 100 100 010 100 ..

79 110 32 97 32 32 114 101 99 101 110 116 32 83 97 116 117 114 100 97 121 44 32 74 111 104 97 110 32 68 105 106 107 108 97 110 100 44 32 32 97 32 50 51 45 121 101 97 114 45 111 108 100 32 115 116 117 100 101 110 116 32 105 110 32 32 69 109 109 101 110 44 32 78 101 116 104 101 114 108 97 110 100 115 44 32 111 112 101 110 101 100 32 32 97 32 102 114 101 101 32 109 101 115 115 97 103 105 110 103 32 97 112 112 32 99 97 108 108 101 100 32 76 105 110 101 32 32 111 110 32 104 105 115 32 105 80 104 111 110 101 46 32 84 104 101 110 32 104 101 32 116 97 112 112 101 100 32 111 110 32 32 97 32 118 105 114 116 117 97 108 32 115 116 105 99 107 101 114 32 111 102 32 32 97 32 115 108 101 101 112 121 32 112 97 110 100 97 32 119 105 116 104 32 32 97 32 34 103 111 111 100 32 110 105 103 104 116 34 32 32 115 112 101 101 99 104 32 32 98 117 98 98 108 101 32 97 110 100 32 112 114 101 115 115 101 100 32 115 101 110 100 32 116 111 32 32 97 32 102 114 105 101 110 100 46 32 87 105 116 104 32 116 104 97 116 32 97 99 116 105 111 110 44 32 77 114 46 32 68 105 106 107 108 97 110 100 39 115 32 32 116 101 120 116 32 106 111 105 110 101 100 32 116 104 101 32 32 116 101 110 115 32 111 102 32 98 105 108 108 105 111 110 115 32 111 102 32 109 101 115 115 97 103 101 115 32 32 116 104 97 116 32 97 114 101 32 32 112 114 111 99 101 115 115 101 100 32 101 118 101 114 121 32 100 97 121 32 102 114 111 109 32 109 101 110 44 32 78 101 116 104 101 114 108 97 110 100 115 44 32 111 112 101 110 101 100 32 97 32 102 114 101 101 32 109 101 115 115 97 103 105 110 103 32 97 112 112 32 99 97 108 108 101 100 32 76 105 110 101 32 111 110 32 104 105 115 32 105 80 104 111 110 101 46 32 84 104 101 110 32 104 101 32 116 97 112 112 101 100 32 111 110 32 97 32 118 105 114 116 117 97 108 32 115 116 105 99 107 101 114 32 111 102 32 97 32 115 108 101 101 112 121 32 112 97 110 100 97 32 119 105 116 104 32 97 32 34 103 111 111 100 32 110 105 103 104 116 34 32 115 112 101 101 99 104 32 98 117 98 98 108 101 32 97 110 100 32 112 114 101 115 115 101 100 32 115 101 110 100 32 116 111 32 97 32 102 114 105 101 110 100 46 32 87 105 116 104 32 116 104 97 116 32 97 99 116 105 111 110 44 32 77 114 46 32 68 105 106 107 108 97 110 100 39 115 32 32 116 101 120 116 32 106 111 105 110 101 100 32 116 104 101 32 116 101 110 115 32 111 102 32 98 105 108 108 105 111 110 115 32 111 102 32 109 101 115 115 97 103 101 115 32 116 104 97 116 32 97 114 101 32 112 114 111 99 101 115 115 101 100 32 101 118 101 114 121 32 100 97 121 32 102 114 111 109 32 97 32 102 97 115 116 45 103 114 111 119 105 110 103 32 99 114 111 119 100 32 111 102 32 109 111 98 105 108 101 32 109 101 115 115 97 103 105 110 103 32 97 112 112 115 46 84 104 101 32 109 101 115 115 97 103 105 110 103 32 97 112 112 115 151 119 105 116 104 32 102 117 110 110 121 32 110 97 109 101 115

……

Bước 2:

Ta chia chuỗi nhị phân của văn bản thành các chuỗi nhị phân con. Mỗi chuỗi con gồm K bit. Theo ví dụ K=3 vậy ta có các chuỗi con gồm 3 bit: "000", "100", "000", "100", "000", "000", "000", "100", "000", "100", "100", "100", "000", "100", "000", "000", "000",…….

Do bit dấu tin được nhúng vào hai khoảng trống đầu tiên của các chuỗi, nên ta xét hai bit đầu tiên của chuỗi. Từ hai bit đầu tiên của chuỗi nhị phân ta giải mã Machester được bit ‘0’ và ‘1’

010 100 010 100 010 010 010 100 010 100 010 100 100 100 010 100 ..

0 1 0 1 0 0 0 1 0 1 0 1 1 1 0 1…

Ghép các bit lại ta được chuỗi nhị phân thông điệp MSG.

010100 010111 010000 010011 100010 010000 010111 100110

Bước 3:

Theo như bảng mã hóa ký tự thông điệp 2.3 cứ 6 bit sẽ được nhóm lại thành một ký tự. Ta có thông điệp.

‘DH DL HP’

Chương 3 :CÀI ĐẶT VÀ THỬ NGHIỆM

3.1. Môi trường cài đặt

- Ngôn ngữ cài đặt: Ngôn ngữ lập trình Matlab phiên bản 7.0.

- Môi trường soạn thảo: Matlab phiên bản 7.0.

- Môi trường chạy chương trình: Môi trường giao diện Matlab 6.5.

- Cấu hình tối thiểu để cài đặt Matlab phiên bản 7.0.

+ Bộ vi xử lý Pentium hoặc Pentium Pro.

+ Windows 95 hoặc NT.

+ Dung lượng ổ cứng từ 25Mb cho tới hơn 1Gb.

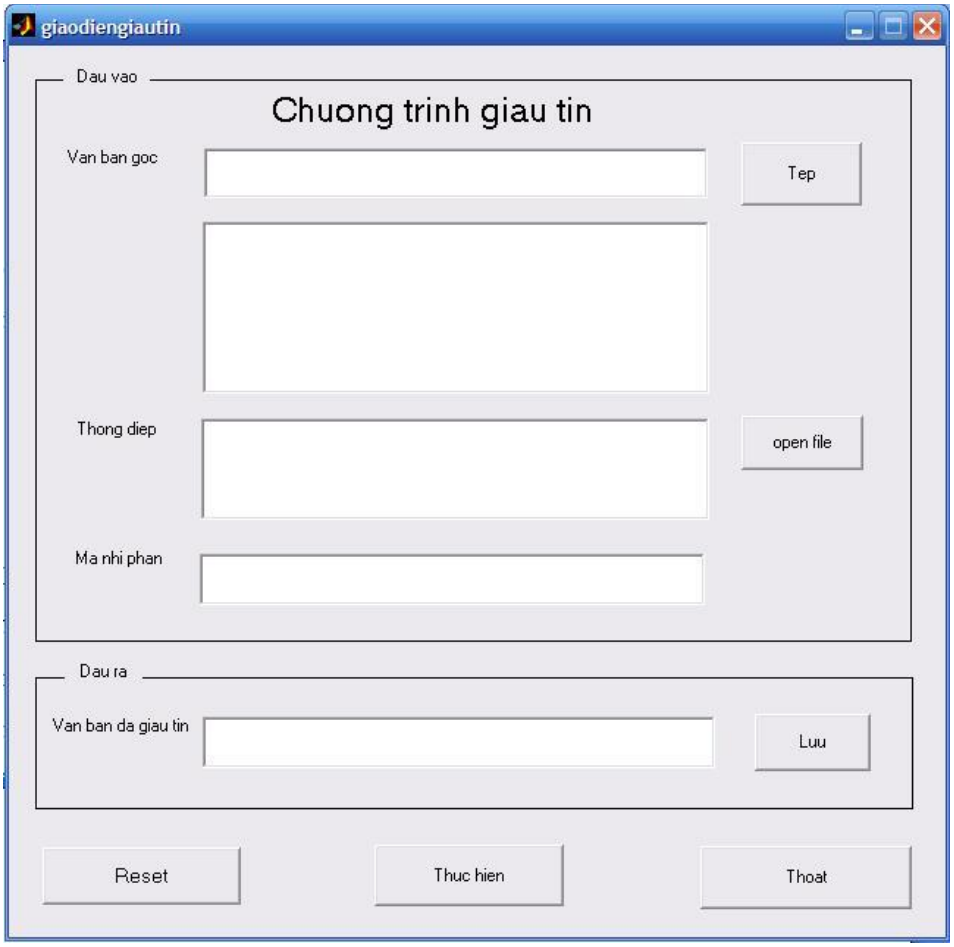
+ Bộ nhớ Ram tối thiểu 128Mb.

3.2. Giao diện Chương trình

3.2.1. Giao diện chính

Hình 3.1. Giao diện chính của chương trình

3.2.2. Giao diện giấu tin



Hình 3.2. Giao diện giấu tin

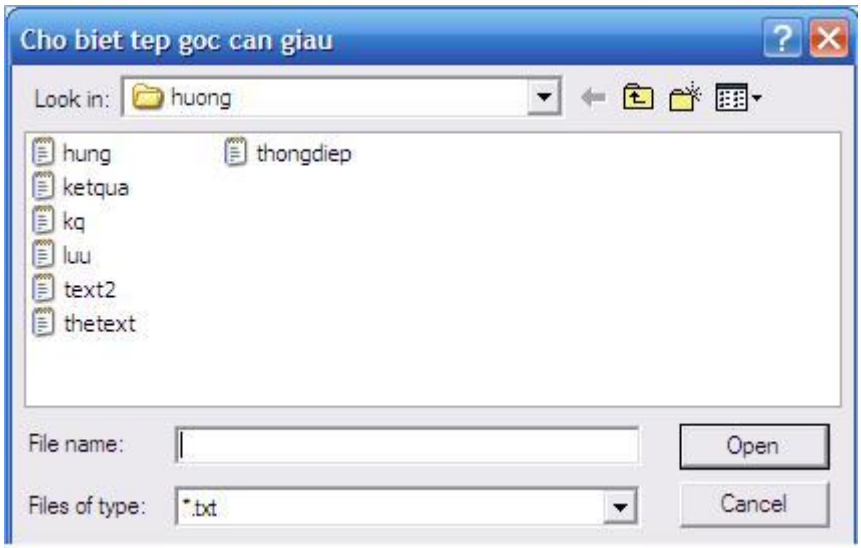
Đầu vào:

- Văn bản gốc.
- Chuỗi thông điệp cần giấu.

Đầu ra:

- Văn bản đã giấu tin.

Ban đầu ta nhấn “Tep” để chọn văn bản gốc. Ta có hộp thoại. Ta chọn file văn bản sẽ dung để giấu tin.



Hình 3.3. Hộp chọn văn bản

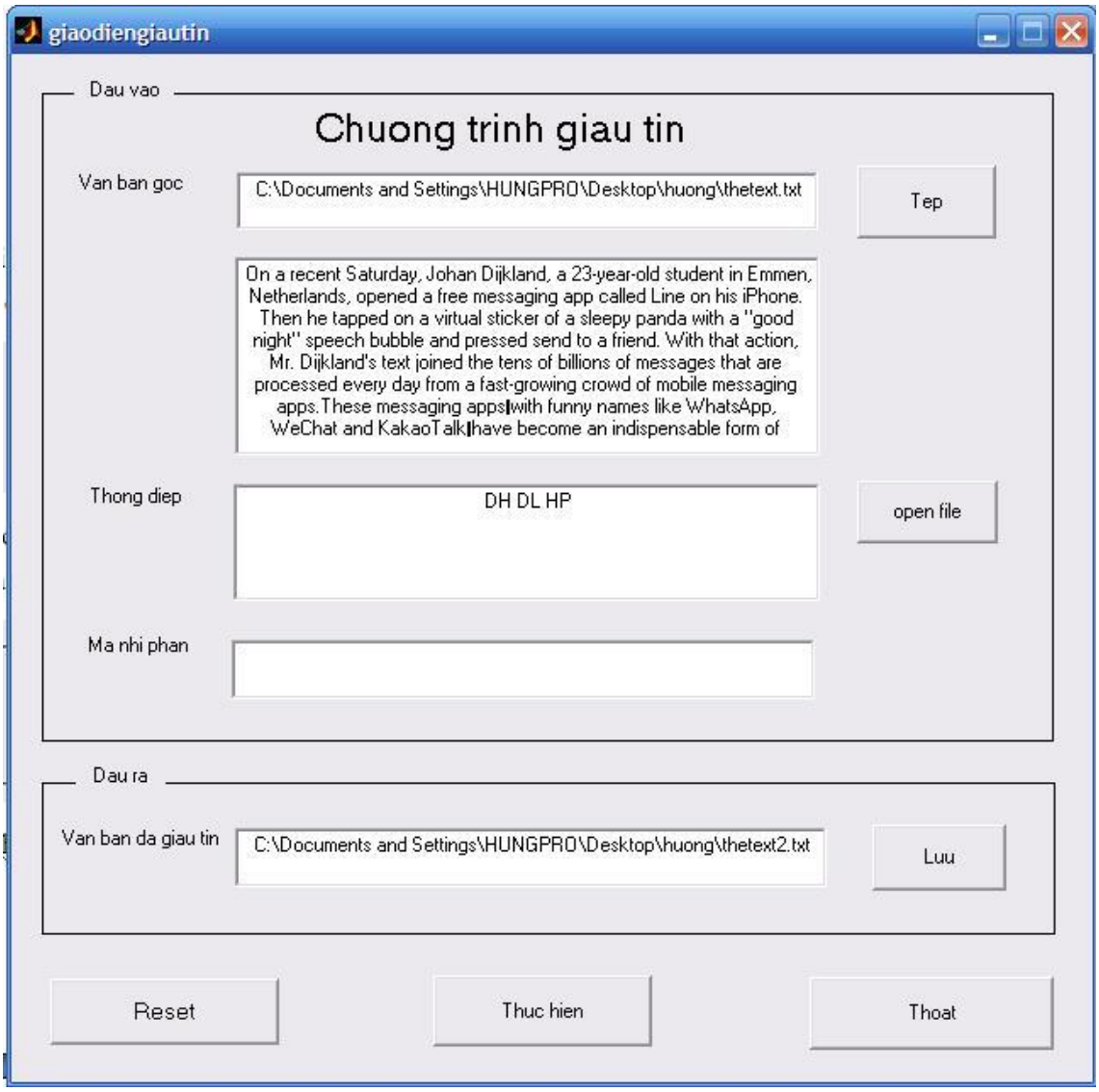
Tiếp theo nhập thông điệp cần giấu vào văn bản. Hoặc ta sẽ nhập trực tiếp thông điệp từ bàn phím vào ô "Thong diep".

Hình 3.4. Nhập thông điệp

Hoặc nhấn "open file" để chọn file chứa thông điệp. Giống như khi ta chọn file văn bản ở trên.

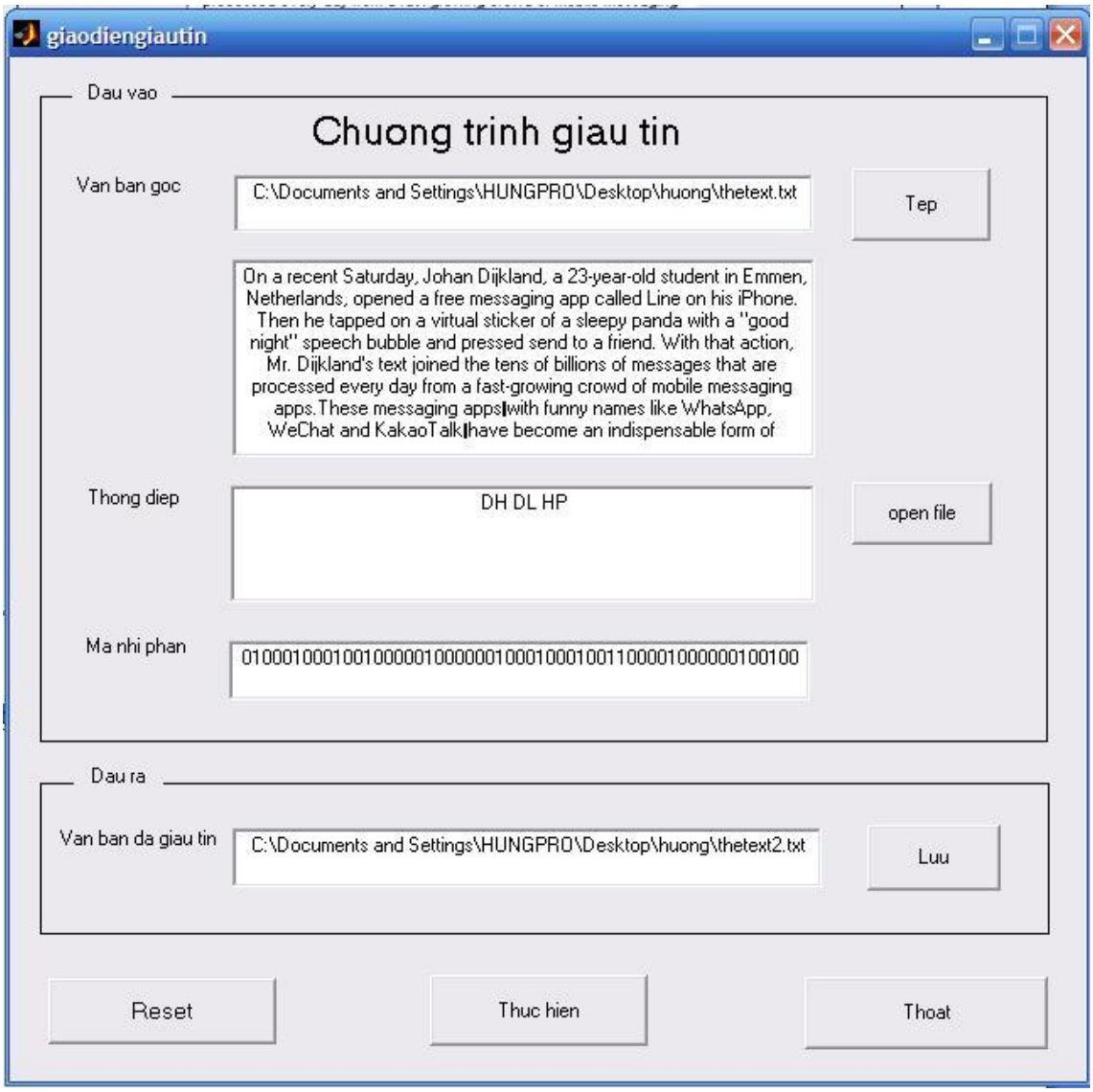
Hình 3.5. Lưu văn bản

Nhấn "Luu" để Lưu văn bạn đã giấu tin.



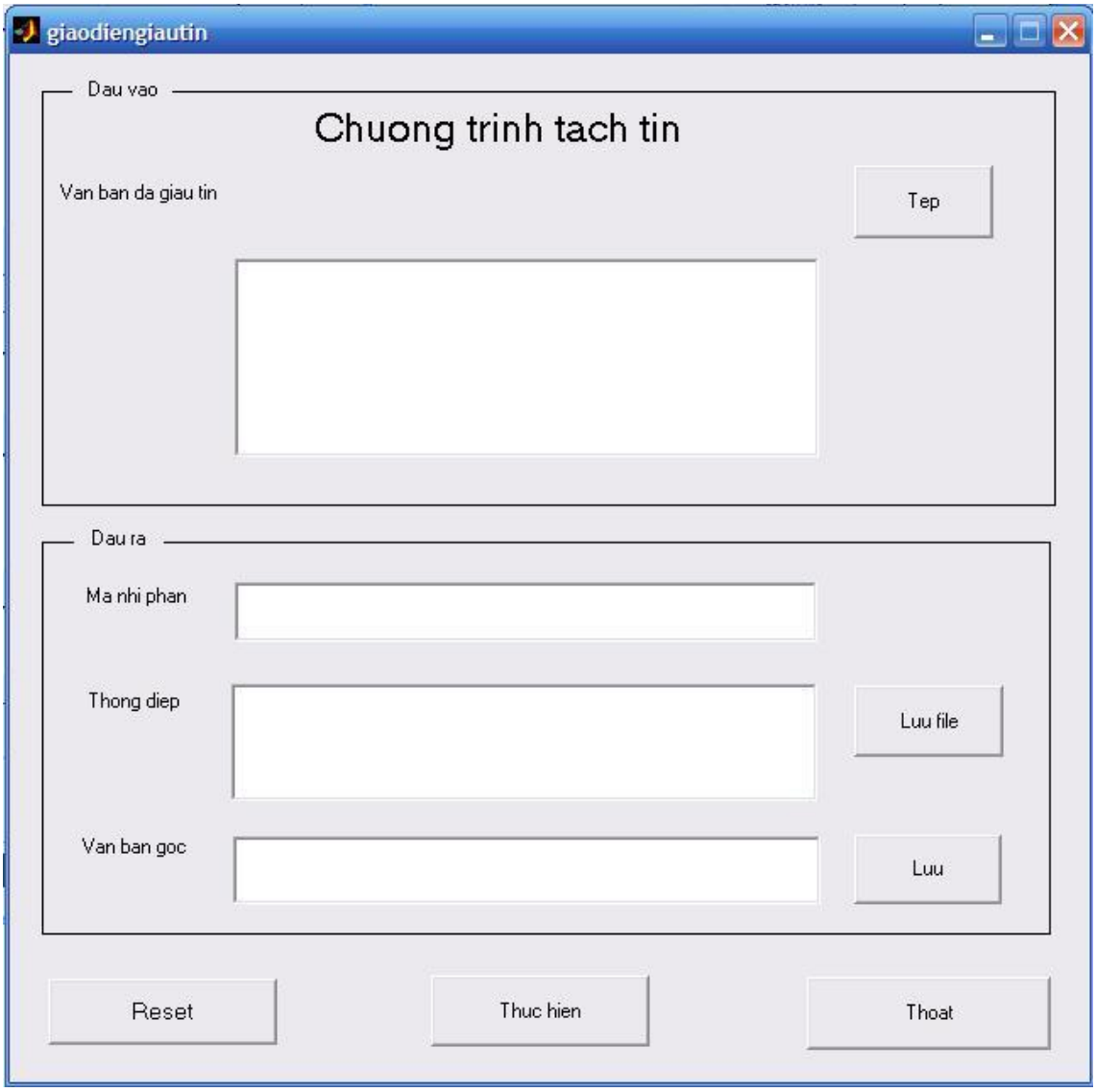
Hình 3.6. Chương trình sau khi nhập đầy đủ

Nhấn "Thuc hien" để tiến hành giấu tin. Ta có kết quả như hình 3.6.



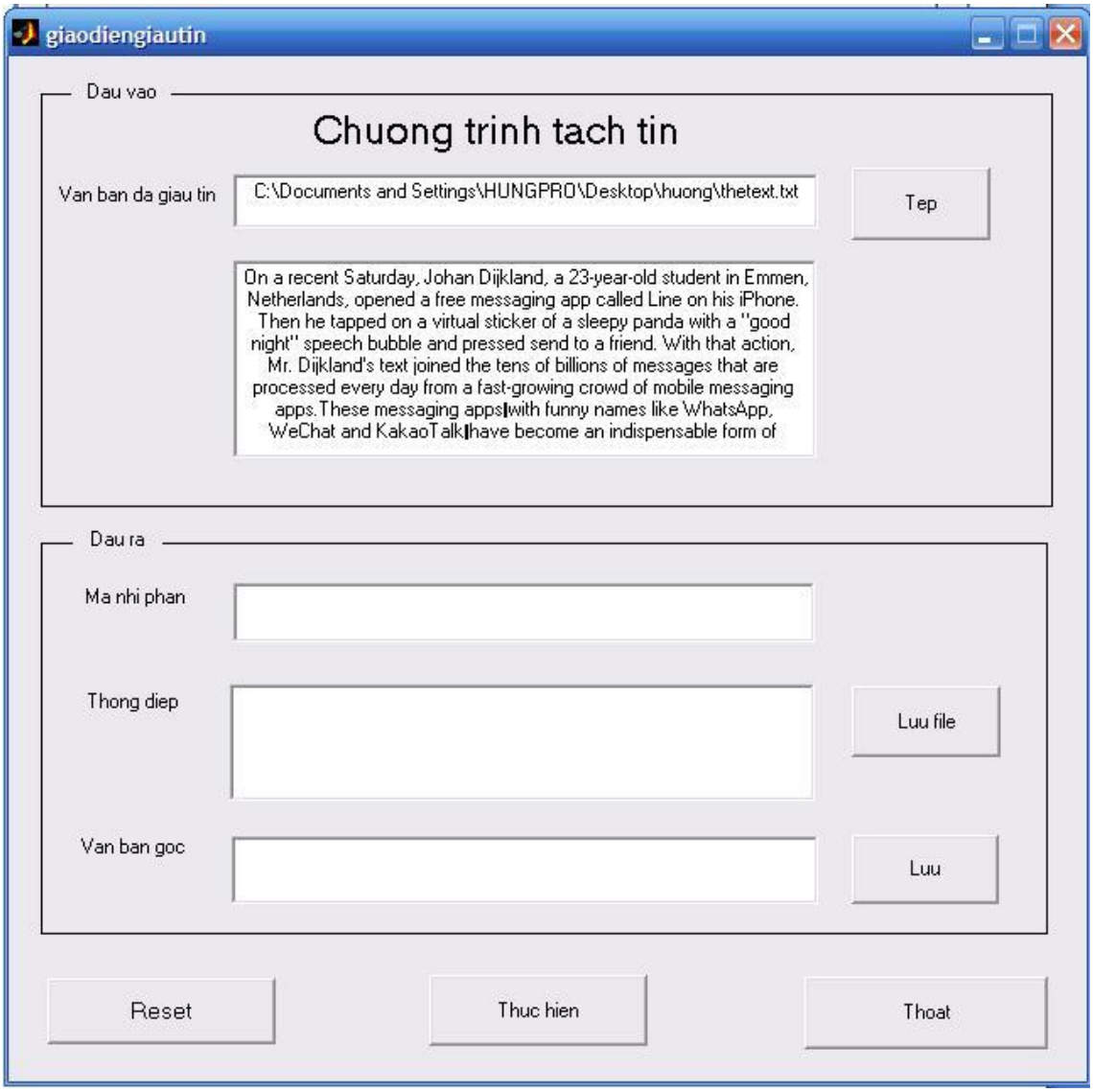
Hình 3.7. Giao diện Chương trình giấu tin thành công.

3.2.2. Giao diện tách tin



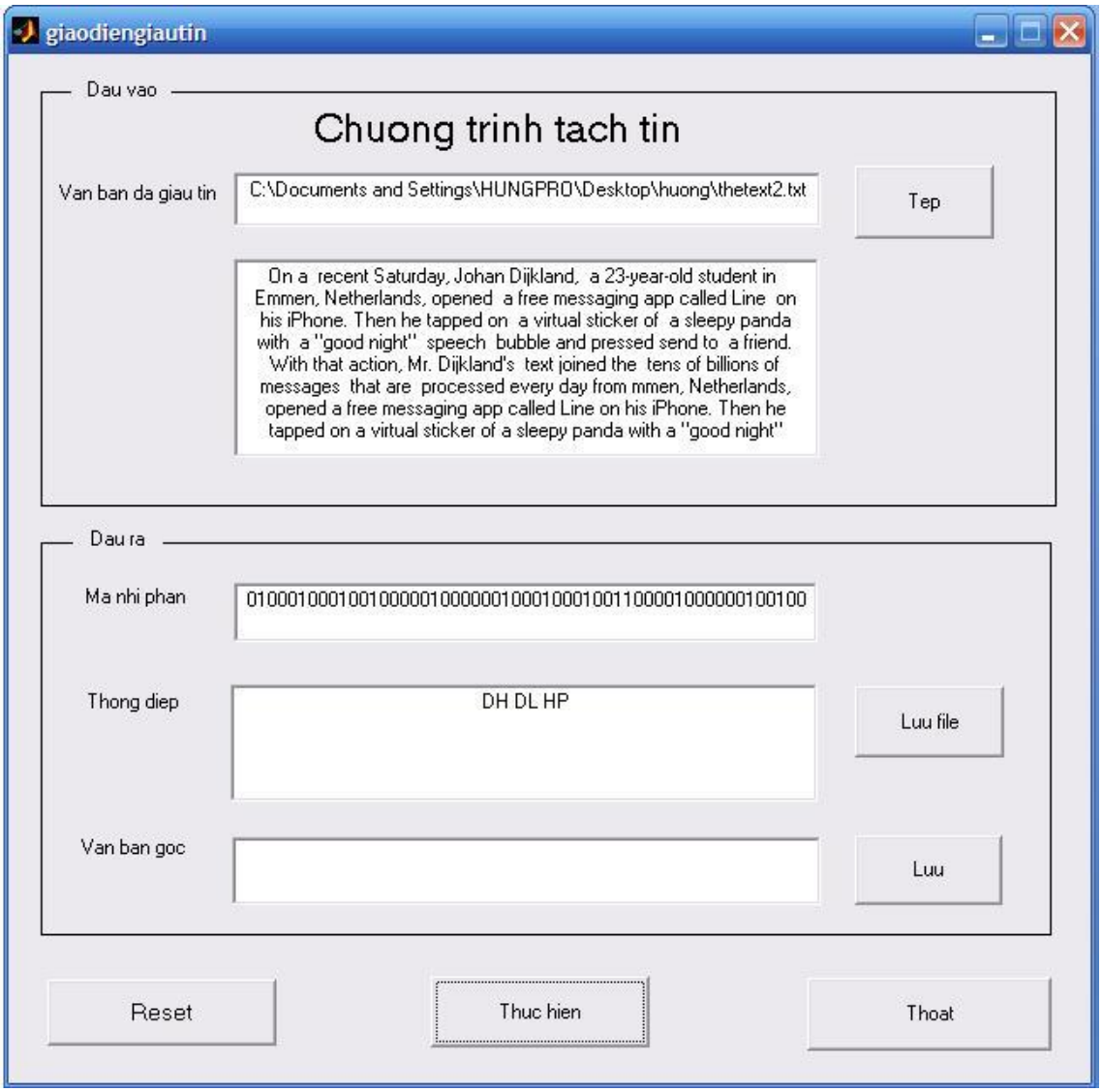
Hình 3.8. Giao diện tách tin

Nhấn "Tep" để chọn file văn bản đã chứa thông điệp.



Hình 3.9. Giao diện sau khi chọn văn bản

Sau đó nhấn thực hiện để tách thông điệp.

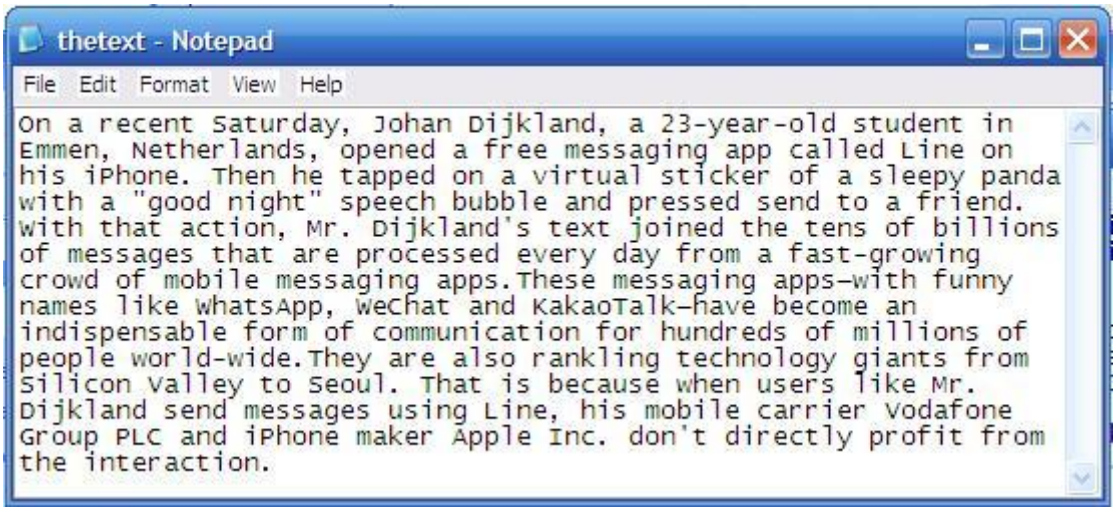


Hình 3.10. Tách thông điệp.

3.3. Cài đăt và nhận xét

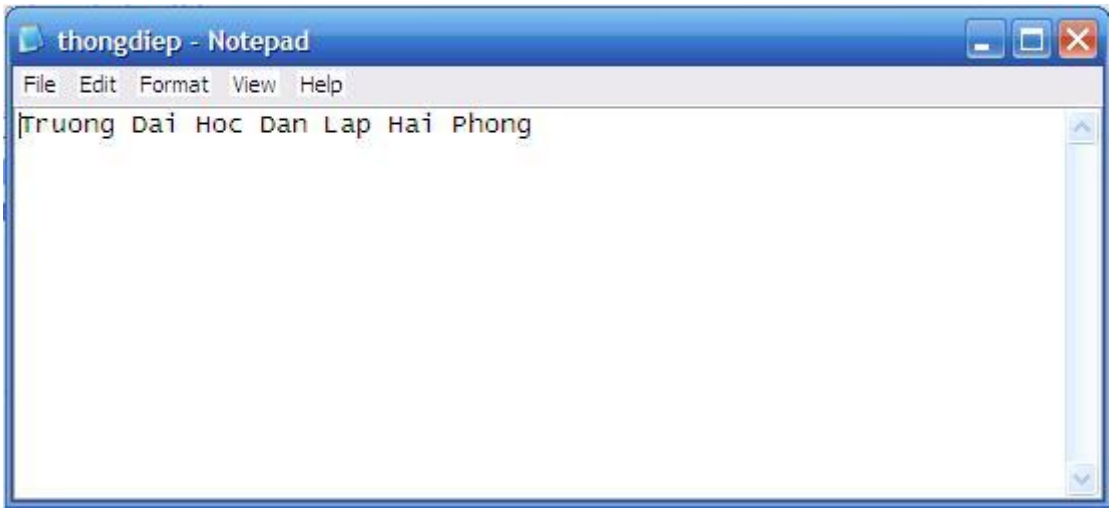
3.3.1. Cài đặt

- Cho một tệp văn bản có kích cỡ 836 byte.



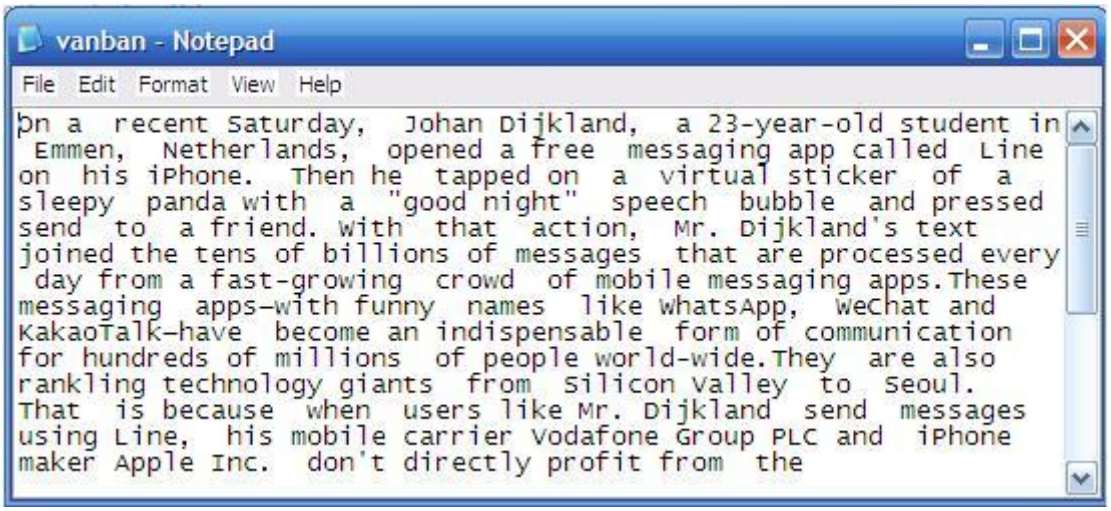
Hình 3.11. Văn bản trước khi giấu thông điệp

- Một tệp thông điệp gồm 32 ký tự



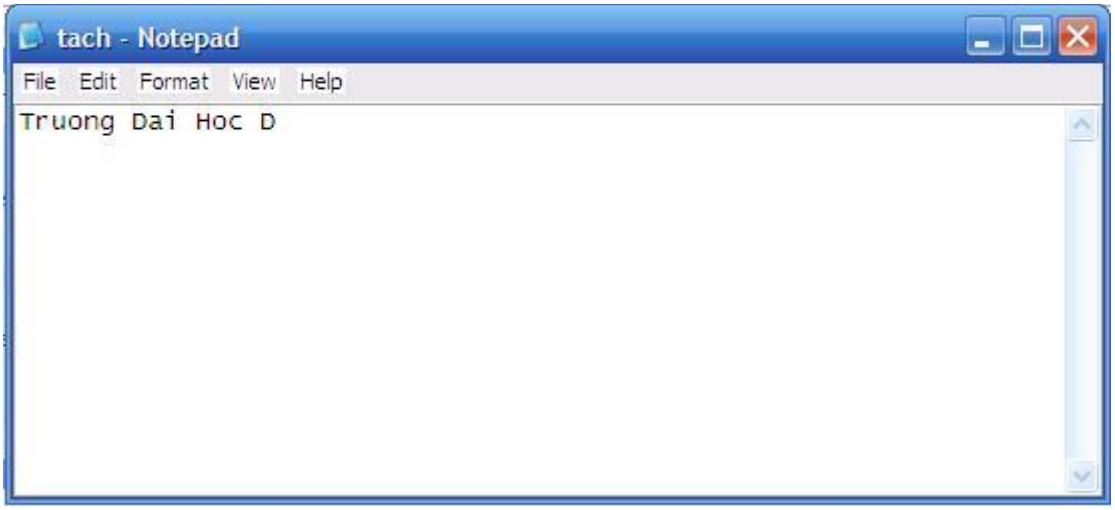
Hình 3.12. Thông điệp cần giấu

- Sau khi thực hiện giấu tin ta được tệp mới 1.597 byte



Hình 3.13. Văn bản sau khi giấu thông điệp

- Sau khi tách tin ta có thông điệp gồm 16 ký tự.



Hình 3.14. Thông điệp được tách

3.3.2. Nhận xét

- Tệp văn bản trên giấu được tối đa 16 ký tự. Muốn giấu nhiều ký tự hơn thì phải dùng tệp văn bản có độ dài lớn hơn.
- Do mỗi văn bản giới hạn số ký tự có thể giấu, nếu thông điệp dài quá thì sau khi tách sẽ nhận được một phần thông điệp bị mất một phần.

3.3.3. Ưu và nhược điểm của chương trình

Chương trình gọn nhẹ và thực hiện nhanh chóng,được viết dựa trên phương pháp sử dụng khoảng trống giữa các từ để giấu tin. Chương trình chưa thể xử lý thông điệp có dấu. Chương trình chỉ có thể thao tác trên tệp văn bản có đuôi mở rộng "txt vàkhông thể thao tác tệp có đuôi mở rộng khác như "doc".v.v.

KẾT LUẬN

Báo cáotrình bày tổng quan về giấu thông tin trong văn bản, một trong nhưng lĩnh vực còn khá mới mẻ hiện nay. Chúng được phân thành 3 loại:

- **Phương pháp dựa vào định dạng** sử dụng định dạng vật lý của văn bản như một nơi để che giấu thông tin. Chèn vào khoảng trống, cố ý để lỗi chính tả trải khắp văn bản, thay đổi kích thước các phông chữ là một số trong nhiều phương pháp dựa trên định dạng được sử dụng.
- **Ngẫu nhiên & thống kê phát sinh** là tạo ra văn bản gốc theo đặc tính thống kê. Phương pháp này dựa trên các chuỗi ký tự và các từ chuỗi. Ẩn thông tin trong chuỗi ký tự là nhúng thông tin được xuất hiện theo thứ tự ngẫu nhiên của các ký tự. Cất giấu thông tin trong phạm vi trình tự từ, các từ trong từ điển thực tế có thể được sử dụng để mã hóa một hoặc nhiều bit thông tin trên mỗi từ bằng cách ánh xạ giữa các đơn vị từ vựng và trình tự bit, hoặc các từ mà nó có thể mã hóa ẩn thông tin.
- **Phương pháp ngôn ngữ** xem xét các thuộc tính của ngôn ngữ được tạo ra và văn bản sửa đổi, thường xuyên sử dụng cấu trúc ngôn ngữ như là một nơi cho các thông điệp ẩn.

Nội dung báo cáo nói về kỹ thuật giấu tin bằng cách sử dụng khoảng trống trong văn bản. Phương pháp khoảng sử dụng khoảng trống giữa các câu (inter-sentence spacing), khoảng trống cuối dòng (end-of-line spaces) và khoảng trống giữa các từ (inter-word spacing) trong văn bản. Trong ba phương pháp trên báo cáo đi sâu vào phương pháp khoảng trống giữa các từ trong văn bản.

Hướng nghiên cứu tiếp nên tập trung hướng tới tối ưu hóa sự vững mạnh của thuật toán giải mã. Điều này là bởi vì các dữ liệu ẩn sẽ phá hủy một khi khoảng trống bị xóa bởi một số phần mềm xử lý văn bản. Bên cạnh đó để nâng cao năng lực nhúng cần sử dụng các phương pháp nén hiệu quả hơn.

Qua quá trình làm đồ án, em đã học thêm nhiều kiến thức thực tế và biết vận dụng kiến thức đã học để giải quyết một bài toán đặt ra. Tuy nhiên kết quả còn rất hạn chế, cần có sự hỗ trợ rất nhiều của giáo viên hướng dẫn. Để có khả năng làm tốt việc vận dụng lý thuyết vào thực hành và có kỹ năng nhất định, em thấy cần phải thực hành nhiều hơn nữa.

TÀI LIỆU THAM KHẢO

[1]. Nguyễn Xuân Huy, Trần Quốc Dũng,*Giáo trình giấu tin và thuỷ vân ảnh*, Trung tâm thông tin tư liệu,TTKHTN -CN 2003.

[2].Ching –Yu Yang, Chih-Hung Lin, Wu-Chih Hu, *Reversible Data Hiding By Adaptive IWT-coefficient Adjustment*,Journal of Information Hiding and Multimedia Signal Processing, ©2011 ISSN 2073 -4212, (2011).

[3].L.Y.POR,B.Delina, *Information Hiding:A New Approach in Text Steganography*, 7th WSEAS Inf. Conf. on APPLIED COMPUTER & APPLIED COMPUTATIONAL SCIENCE (ACACOS '08), Hangzhou, China, April 6-8, 2008.

[4] W. Bender, D. Gruhi, N. Morimoto, A.Lu,"Techniques for Data Hiding", IBM SYSTEMS JOURNAL, VOL 35, NOS 3&4, 1996.

[5] Roshidi Din, T. Zalizam T. Muda, Puriwat Lertkrai, Mohd Nizam Omar, Angela Amphawan, Fakhrul Anuar Aziz,*Text Steganalysis Using Evolution Algorithm Approach*, InterNetWorks Research Laboratory, School of Computing, UUM College of Arts and Sciences Universiti Utara Malaysia UUM Sintok, 06010, Kedah MALAYSIA.

[6] Indradip Banerjee, Souvik Bhattacharyya and Gautam Sanyal,*An Approach of Quantum Steganography through Special SSCE Code*, World Academy of Science, Engineering and Technology 56 2011.

PHỤ LỤC A

Bảng mã hóa ký tự

Ký tự	ACS II	Mã	Mã nhị phân	Ký tự	ACSII	Mã	Mã nhị phân
0	48	00	000000	G	71	22	010110
1	49	01	000001	H	72	23	010111
2	50	02	000010	I	73	24	011000
3	51	03	000011	&	36	25	011001
4	52	04	000100	.	46	26	011010
5	53	05	000101	]	93	27	011011
6	54	06	000110	(	40	28	011100
7	55	07	000111	<	60	29	011101
8	56	08	001000	\	92	30	011110
9	57	09	001001	^	94	31	011111
[	91	10	001010	J	74	32	100000
#	35	11	001011	K	75	33	100001
@	64	12	001100	L	76	34	100010
:	58	13	001101	M	77	35	100011
>	62	14	001110	N	78	36	100100
?	63	15	001111	O	79	37	100101
(space)	32	16	010000	P	80	38	100110
A	65	17	010001	Q	81	39	100111
B	66	18	010010	R	82	40	101000
C	67	19	010011	-	45	41	101001
D	68	20	010100	$	36	42	101010
E	69	21	010101	*	42	43	101011

F	70	44	101100	)	41	54	110110
;	59	45	101101	X	88	55	110111
‘	39	46	101110	Y	89	56	111000
`	96	47	101111	Z	90	57	111001
/	47	48	110000	<	60	58	111010
S	83	49	110001	,	44	59	111011
T	84	50	110010	%	37	60	111100
U	85	51	110011	=	61	61	111101
V	86	52	110100	“	34	62	111110
W	87	53	110101	!	33	63	111111