HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Đồng Thanh Tú

TÌM HIỂU HỆ THỐNG DATA WAREHOUSE VÀ ỨNG DỤNG CỦA CHÚNG

Chuyên ngành: Hệ thống Thông tin

Mã số: 8.48.01.04

TÓM TẮT LUẬN VĂN THẠC SĨ

HÀ NỘI - 2021

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: Tiến sĩ Hồ Văn Canh……...

(Ghi rõ học hàm, học vị)

Phản biện 1: Tiến sĩ Nguyễn Vĩnh An……..

Phản biện 2: Tiến sĩ Trần Minh Tân…….

Luận văn được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: 10 giờ 30 ngày 30 tháng 08 năm 2021

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông.

MỞ ĐẦU

Trong thời đại cách mạng công nghiệp 4.0, khi mà những yếu tố cốt lõi là Trí tuệ nhân tạo (AI), vạn vật kết nối – Internet of Things (IoT) và dữ liệu lớn (Big Data) luôn là xu hướng hàng đầu giúp tự động hóa xử trí và trao đổi thông tin, dữ liệu. Big Data chứa trong mình rất nhiều thông tin quý giá mà nếu trích xuất thành công, nó sẽ giúp rất nhiều cho việc: kinh doanh, nghiên cứu khoa học, dự đoán các dịch bệnh sắp phát sinh... Chính vì thế, những dữ liệu này phải được thu thập, tổ chức, lưu trữ, tìm kiếm, chia sẻ theo một cách khác so với bình thường.

Data warehouse thực hiện quá trình truy cập dữ liệu từ các nguồn không đồng nhất, làm sạch, lọc và chuyển đổi dữ liệu, lưu trữ dữ liệu theo cấu trúc để dễ dàng truy cập, hiểu rõ và sử dụng. Từ nguồn dữ liệu ở khắp mọi nơi, hệ thống sẽ kiểm soát và ra quyết định cụ thể theo yêu cầu.

Kho dữ liệu là một hướng công nghệ mới được sử dụng phổ biến cho các bài toán lớn hiện nay như: y tế, bảo hiểm, ngân hàng, dân số, viễn thông.... Việc xây dựng kho dữ liệu không những giúp lưu trữ một lượng thông tin lớn hàng ngày mà còn giúp cho các nhà quản lý có thể trích rút nguồn tài nguyên một cách nhanh chóng, chính xác. Đây cũng là kiến thức rất hữu ích và cần thiết để có thể khai thác ngày một hiệu quả các thành tựu tin học.

Với mục đích, đưa những tiến bộ khoa học, công nghệ vào phục vụ cho cuộc sống, học viên xin chọn đề tài nghiên cứu "*Tìm hiểu hệ thống Data warehouse và ứng dụng của chúng*".

Luận văn tập trung vào nghiên cứu tổng quan về Data warehouse và xây dựng một phần mềm ứng dụng nhằm tìm kiếm các thông tin liên quan đến Bảo hiểm y tế ngành Công an được lưu trong cơ sở dữ liệu của Hệ thống thông tin giám định bảo hiểm y tế.

Nội dung của luận văn bao gồm 03 chương:

Chương 1: TỔNG QUAN VỀ DATA WAREHOUSE

Chương 2: XÂY DỰNG THUẬT TOÁN GIẤU THÔNG TIN MẬT TRONG CƠ SỞ DỮ LIỆU DATA WAREHOUSE

Chương 3: ĐỀ XUẤT THUẬT TOÁN GIẤU THÔNG TIN MẬT VÀ ỨNG DỤNG TRONG NGÀNH Y TẾ

Cuối cùng là phần kết luận và các tài liệu tham khảo

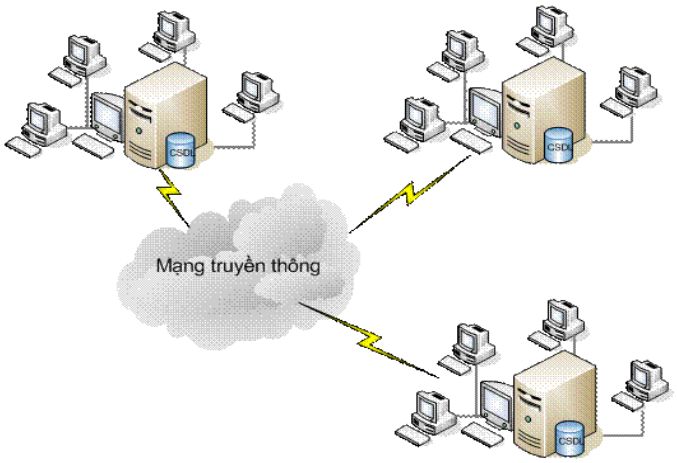
Chương 1 - TỔNG QUAN VỀ DATA WAREHOUSE

1.1. Khái niệm CSDL phân tán, kho dữ liệu

1.1.1. Mô hình CSDL phân tán

Cơ sở dữ liệu là một trong những nội dung rất được quan tâm trong quá trình xây dựng các hệ thống thông tin, đặc biệt là hệ thông thông tin quản lý.

Cơ sở dữ liệu phân tán (*Distributed DataBase – DDB*) là một tập hợp nhiều cơ sở dữ liệu có liên đới logic và được phân bố trên một mạng máy tính. Với khái niệm này, có 02 thuật ngữ quan trọng trong các định nghĩa là "liên đới logic" và "phân bố trên một mạng máy tính".



Hình 1.1. Sơ đồ cơ sở dữ liệu phân tán

Hệ quản trị cơ sở dữ liệu phân tán (*Distributed Database Management System D – DBMS*) được định nghĩa là một hệ thống phần mềm cho phép quản lý các hệ cơ sở dữ liệu phân tán và làm cho sự phân tán trở nên "trong suốt" đối với người sử dụng.

Hệ cơ sở dữ liệu phân tán (*Distributed DataBase System – DDBS*) được xây dựng dựa trên hai công nghệ cơ bản là cơ sở dữ liệu và mạng máy tính. Một hệ cơ sở dữ liệu phân tán không phải là một "tập hợp các tập tin" được lữu trữ riêng rẽ tại mỗi nút của một mạng máy tính. Để tạo ra một hệ cơ sở dữ liệu phân tán các tập tin không chỉ có liên đới logic mà chúng còn phải có cấu trúc và được truy xuất qua một giao diện chung. Có 02 hệ cơ sở dữ liệu phân tán:

- Hệ cơ sở dữ liệu phân tán không thuần nhất
- Hệ cơ sở dữ liệu phân tán thuần nhất

1.1.2. Định nghĩa kho dữ liệu (Data warehouses)

Kho dữ liệu (Data warehouse) là tuyển tập các cơ sở dữ liệu tích hợp, hướng chủ đề, được thiết kế để hỗ trợ cho chức năng trợ giúp quyết định. Kho dữ liệu thường rất lớn tới hàng trăm Girabyte hay thậm chí hàng Terabyte. Kho dữ liệu được xây dựng để tiện lợi cho việc truy cập theo nhiều nguồn, nhiều kiểu dữ liệu khác nhau sao cho có thể kết hợp được cả những ứng dụng của các công nghệ hiện đại và kế thừa được từ những hệ thống đã có sẵn từ trước.

Cấu trúc của một kho dữ liệu bao gồm ba tầng: tầng đáy, tầng giữa, tầng trên cùng.

Hình 1.2. Sơ đồ chung kho dữ liệu

Hình 1.2 mô tả kiến trúc cơ bản của Data warehouse, dựa trên nguyên tắc là xây dựng một kho dữ liệu thống nhất từ nhiều nguồn dữ liệu khác nhau để phục vụ truy vấn.

Công nghệ kho dữ liệu (*Data warehouse Technology*) là tập các phương pháp, kỹ thuật và các công cụ có thể kết hợp, hỗ trợ nhau để cung cấp thông tin cho người sử dụng trên cơ sở tích hợp từ nhiều nguồn dữ liệu, nhiều môi trường khác nhau.

1.2. Dữ liệu Data warehouse

Data warehouse là tập hợp dữ liệu tương đối ổn định, không hay thay đổi, cập nhật theo thời gian, được tích hợp theo hướng chủ đề nhằm hỗ trợ quá trình đưa ra quyết định về mặt quản lý.

1.2.1. Các đặc trưng của kho dữ liệu

Kho dữ liệu là một tập hợp dữ liệu hướng chủ đề, toàn vẹn, không bị rò rỉ mất mát và có giá trị lịch sử. Cụ thể các tính chất đó như sau:

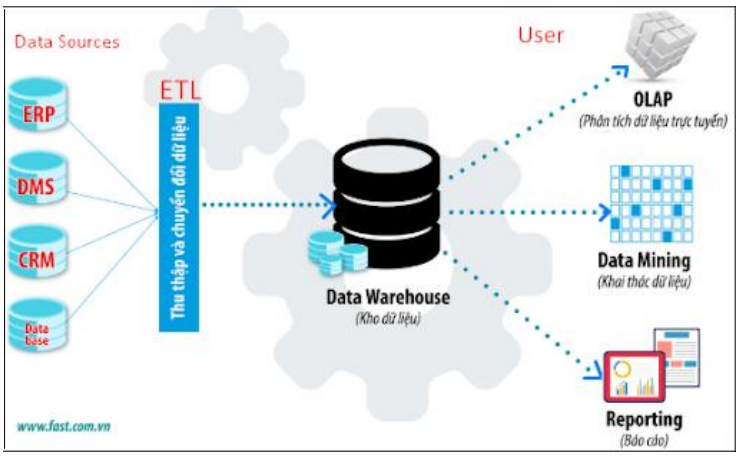
- Tính hướng chủ đề (*Subject – oriented*)
- Tính toàn vẹn (*Integrated*)

- Tính bất biến (*Nonvolatile*)

- Giá trị lịch sử (*Time – varying*)

1.2.2. Kiến trúc hệ thống Data warehouse

Khác với cơ sở dữ liệu giao dịch thông thường, Data warehouse được bổ sung thêm bộ công cụ kết xuất, chuyển đổi và tích hợp dữ liệu (*Extraction Transformation Loading – ETL*) bộ phân tích dữ liệu trực tuyến OLAP và các công cụ quản trị các tiến trình thu thập dữ liệu. Đặc biệt, Data warehouse được tổ chức nâng cao theo các chủ đề Data Mart



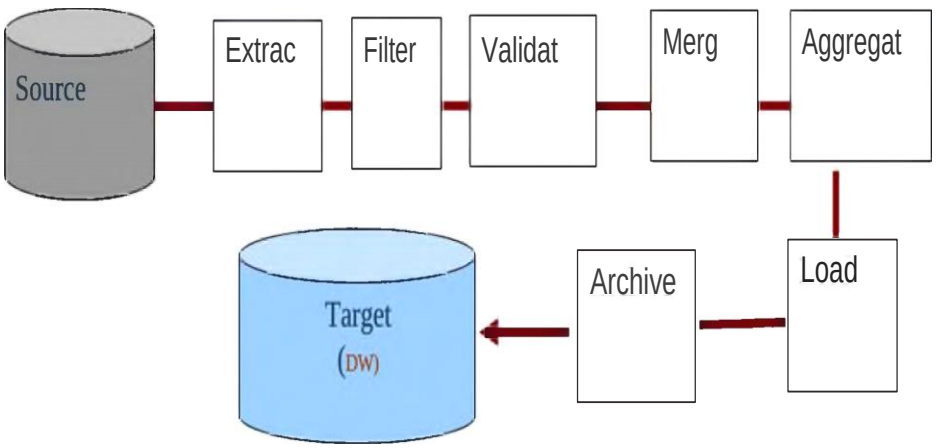
Hình 1.3. Thành phần cơ bản của Dataware house

➢ Tầng xử lý dữ liệu (*Extraction Transformation Loading – ETL*) là tầng thấp nhất, ẩn đi với người dùng cuối.

➢ Tầng kho dữ liệu (*Data warehouse*) đứng ở trung tâm một hệ thống kho dữ liệu làm nhiệm vụ lưu trữ dữ liệu bao quanh tất cả các hoạt động.

➢ Tầng khai thác dữ liệu (*User*) chứa các công cụ cho người dùng cuối khai thác, sử dụng các dữ liệu trong kho dữ liệu. Một số công cụ chính: Báo cáo OLAP (*OLAP tool*); Báo cáo tĩnh (*reporting tool*);Bộ công cụ khai phá dữ liệu (*data mining*)

1.3.3. Quy trình xây dựng kho dữ liệu



Hình 1.4. Các bước tạo lập kho dữ liệu

Hầu hết mọi hệ thống/dự án công nghệ thông tin, chúng ta đều xem xét dựa trên 3 khía cạnh: cấu trúc (*Structure*), dữ liệu (*Data*), tiến trình (*Process*). Vì vậy, khi xây dựng một hệ thống Data warehouse chúng ta cũng phải cân nhắc 3 khía cạnh này.

Tóm lại, Data warehouse là một kho dữ liệu được lưu trữ bằng hệ thống điện tử của một tổ chức, nó được thiết kế để hỗ trợ việc phân tích, tìm kiếm và lập báo cáo. Data warehouse cung cấp các lợi ích sau tới những người dùng:

- Dữ liệu được tổ chức để tạo thuận lợi cho các truy vấn phân tích chứ không phải cho việc xử lý các giao dịch.

- Sự khác biệt về cấu trúc dữ liệu được lưu trữ trên nhiều nguồn dữ liệu không đồng nhất sẽ được giải quyết.

- Những quy tắc thống nhất sẽ được áp dụng khi hợp nhất dữ liệu từ các hệ thống không đồng nhất sang Data warehouse.

- Tính bảo mật và hiệu suất có thể được cải thiện mà không cần phải thực hiện bất kỳ sửa đổi nào trên hệ thống dữ liệu gốc.

Để ứng dụng Data warehouse vào trong lĩnh vực y tế của ngành Công an, chúng ta cần bổ sung thêm phần an toàn – bảo mật thông tin cho hệ thống. Do đặc thù của ngành y tế trong lực lượng vũ trang nói chung và trong ngành y tế Công an nói riêng. Chương 2 sẽ giải quyết vấn đề này.

Chương 2 - XÂY DỰNG THUẬT TOÁN GIẤU THÔNG TIN MẬT TRONG CƠ SỞ DỮ LIỆU DATA WAREHOUSE

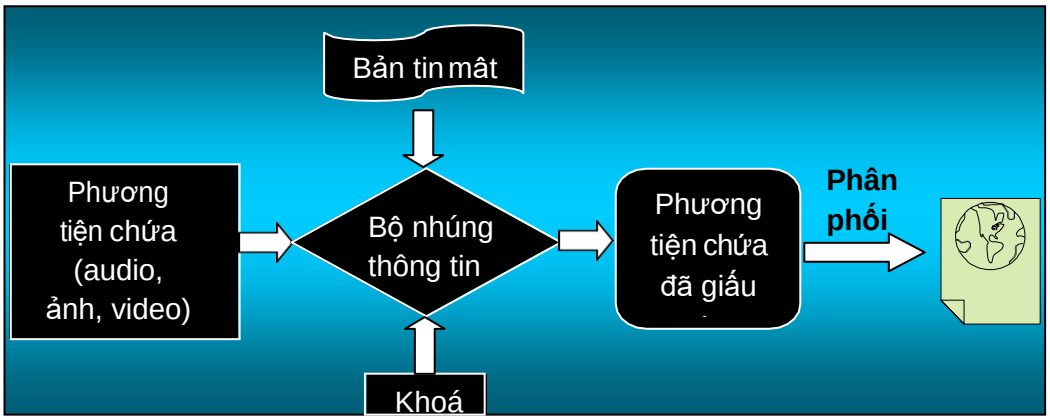
Ta hình dung Data warehouse như là một rừng con của một rừng lớn, gồm các cây cần quan tâm nào đó trong một rừng. Mỗi "cây" trong Data warehouse có những đặc tính chung nào đó nhưng chúng có những đặc điểm riêng biệt cần được giữ bí mật. Như vậy, mỗi cây trong Data warehouse được coi như là một thông báo (message) trong đó có những dữ liệu hoàn toàn được công khai nhưng cũng có những dữ liệu có tính riêng tư cần phải giữ bí mật. Vấn đề đặt ra, làm sao tất cả đều được công khai mà vẫn giữ được bí mật riêng tư?

Có thể có một số phương pháp khác nhau để giải quyết bài toán đó, trong phạm vi nghiên cứu của luận văn, học viên đề xuất một phương pháp khá đơn giản đó là: giấu dữ liệu cần giữ bí mật vào trong ảnh số. Như vậy mỗi cây trong Data warehouse là một thông điệp chứa ảnh số. Từ đó, hình thành một hệ thống Data warehouse gồm các ảnh số cùng các quan hệ giữa chúng.

2.1. Tổng quan về giấu tin

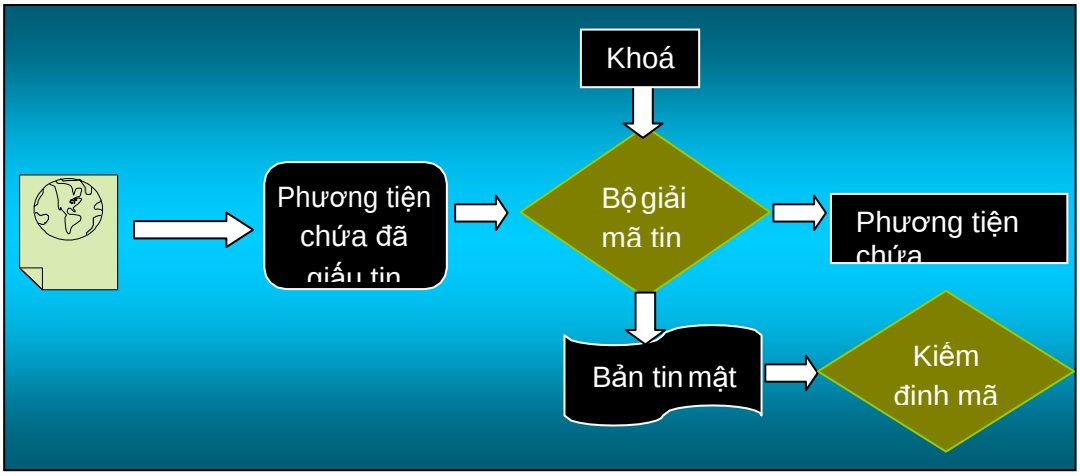
2.1.1. Khái niệm giấu tin

"Giấu tin" là một kỹ thuật nhúng (*Embedding*) một lượng thông tin số nào đó vào trong một đối tượng dữ liệu số khác. Kỹ thuật giấu tin nhằm hai mục đích: Một là, bảo mật cho dữ liệu được đem giấu. Hai là, bảo vệ cho chính đối tượng mang tin giấu. Giấu thông tin vào phương tiện chứa và tách lấy thông tin là hai quá trình trái ngược nhau và có thể mô tả qua sơ đồ khối của hệ thống.



Hình 2.1. Sơ đồ khối quá trình giấu tin

Tách thông tin từ các phương tiện chứa diễn ra theo quy trình ngược lại. với đầu ra là các thông tin đã được giấu vào phương tiện chứa.



Hình 2.2. Sơ đồ khối quá trình giải mã

2.1.2. Kỹ thuật giấu tin trong ảnh

Ngày nay, giấu tin trong ảnh chiếm tỉ lệ lớn nhất hệ thống giấu tin trong đa phương tiện. Giấu tin trong ảnh được thực hiện bằng cách thay thế một vài thông tin ít quan trọng nhất của các điểm ảnh gốc, sao cho chất lượng ảnh ít bị ảnh hưởng nhất có thể. Kỹ thuật giấu tin trong ảnh bao gồm hai quá trình đó là: quá trình giấu tin vào ảnh và quá trình tách tin từ ảnh giấu tin.

Hiện nay, đã có nhiều thuật toán giấu thông tin vào ảnh số ([5], [9], [10]). Trong phạm vi đề tài luận văn học viên trình bày một thuật toán do thầy giáo hướng dẫn cùng học viên phát triển dựa trên công cụ toán học về lý thuyết trường Galoi và mã Hamming sửa sai, sau đây là những nội dung cơ sở:

2.2. Cơ sở toán học xây dựng thuật toán

Trước hết ta ký hiệu: GF(P) là trường Galois có cấp là số nguyên tố P còn GF(P)[x] là không gian véctơ các đa thức với các hệ số trên trường GF(P). Khái niệm đa thức nguyên thủy và ứng dụng:

2.2.1. Định nghĩa 1

Đa thức f(x) có cấp m trong trường GF(p) được gọi là đa thức bất khả qui (*irredueible polynomial*) nếu f(x) không thể được phân tích thành tích của các đa thức có cấp nhỏ hơn m (và > 1) trong trường GF(q).

2.2.2. Định nghĩa 2

Đa thức bất khả qui p(x) cấp m được gọi là đa thức nguyên thủy (primitive polynomial) trong trường GF(P) nếu số nguyên dương nhỏ nhất n mà $x^n - 1$ chia hết cho P(x) phải thỏa mãn $n = P^m - 1$

Trong phạm vi đề tài luận văn này, học viên chỉ xét các đa thức trong trường GF(2):

- Định lý 1. Có tất cả Ø $(2^m - 1)/m$ đa thức nguyên thủy cấp m trong trường GF(2). Trong đó, Ø (.) là hàm phi-ơle.
- Định lý 2. Tập các nghiệm $\{\alpha_i\}$ của đa thức nguyên thủy P(x) cấp m trong trường GF(2) sẽ có cấp $2^m - 1$.

Chứng minh (xem [7])

Bây giờ ta giả sử α là một nghiệm của đa thức nguyên thủy P(x) có cấp m là:

$P(x) = x^m + a_{m-1}x^{m-1} + \ldots + a_1x + a_0$ (chú ý a_0 luôn khác không) với $a_i \in GF(2)$

Khi đó $P(\alpha) = \alpha^m + a_{m-1}\alpha^{m-1} + \ldots + a_1\alpha + a_0 = 0$

Từ đây, ta suy ra: $\alpha^m = -a_0 - a_1\alpha - \ldots a_{m-1}\alpha^{m-1}$ [8]

Do $a_i \in GF(2) = \{0,1\}$ với mọi i = 0, 1, …, m-1. Trong trường GF(2) thì a + b = a – b nên ta có thể viết biểu thức cho ở [8] là: $\alpha^m = a_0 + a_1\alpha + \ldots + a_{m-1}\alpha^{m-1}$ [9]

Do định lý 2, ta suy ra: các lũy thừa của α có cấp lớn hơn hoặc bằng m có thể được biểu diễn dưới dạng đa thức có cấp nhỏ hơn m. Vì α có cấp $2^m - 1$ nên các lũy thừa khác nhau của α phải có $2^m - 1$ các biểu diễn đa thức phân biệt khác không dưới dạng: $P(x) = b_0 + b_1\alpha + \ldots + b_{m-1}\alpha^{m-1}$ với $b_i \in GF(2)$, i = 0, 1, …, m-1. Như vậy, tập hợp tất cả các nghiệm của đa thức nguyên thủy cấp m cùng với véctơ 0 có thể được xem như là một không gian véctơ trên trường GF(2).

Bây giờ ta lấy m = 6. Khi đó không gian véctơ $GF(2^6)$ trên trường GF(2) chúng gồm 64 phần tử (kể cả véctơ 0). Mỗi phần từ là một véctơ gồm 6 thành phần nhị phân. Để ứng dụng cho thuật toán giấu tin mật, ta chọn một đa thức nguyên thủy cấp 6 trong trường GF(2).

Dễ thấy rằng $P(x) = x^6 + x + 1$ là một đa thức nguyên thủy trong trường GF(2).

Vì véctơ (0, 0, 0, 0, 0, 0) không phải là một nghiệm của đa thức $P(x) = x^6 + x + 1$, nên ta sẽ có $2^6 - 1$ nghiệm của P(x).

Tất cả 63 nghiệm của $x^6 + x + 1$ được lập như sau: Lấy α là một nghiệm tùy ý của P(x), ta có $\alpha^6 + \alpha + 1 = 0$.

Do đó: $\alpha^6 = \alpha + 1$; $\alpha^7 = \alpha^2 + \alpha$; $\alpha^8 = \alpha^3 + \alpha^2$; $\alpha^9 = \alpha^4 + \alpha^3$; $\alpha^{10} = \alpha^5 + \alpha^4$; $\alpha^{11} = \alpha^6 + \alpha^5$; $\alpha^{12} = \alpha^7 + \alpha^6$; v.v.

Trong không gian véctơ $GF(2^6 - 1)$ có 6 chiều nên nó có cơ sở trực chuẩn gồm 6 véctơ cực đại độc lập tuyến tính. 6 véctơ của cơ sở này được ký hiệu là:

S = {100000, 010000, 001000, 000100, 000010, 000001}

Tất cả véctơ đó trong $GF(2^6 - 1)$ trừ véc tơ 0 = (000000) được thiết lập bởi bảng H_{6x63} như sau:

$$H = \begin{pmatrix} 1000001000011000101001111010001110010010110111011001101010111111 \\ 0100001100010100111101000111001001011011101100110101011111110000 \\ 0010000110001010011110100011100100101101110110011010101111111000 \\ 0001000011000101001111010001110010010110111011001101010111111100 \\ 0000100001100010100111101000111001001011011101100110101011111110 \\ 0000010000110001010011110100011100100101101110110011010101111111 \end{pmatrix}^T$$

=> $H = [h_{.1}, h_{.2}, \ldots, h_{.63}]$ với $h_{ij} \in \{0,1\}$ và $h_{.j} = (h_{1j}, h_{2j}, h_{3j}, \ldots, h_{6j})$ $j = \overline{1,63}$

Kết quả chính đạt được ở chương 2 là tiền đề để xây dựng một thuật toán giấu tin mật có độ an toàn cao, cho phép cải tiến để có thể giấu được mọi thông điệp mật. Thuật toán được đề xuất có thể giấu thông điệp vào trong ảnh Bmp, mầu đa cấp xám.

Chương 3 – ĐỀ XUẤT THUẬT TOÁN GIẤU TIN MẬT VÀ ỨNG DỤNG TRONG NGÀNH Y TẾ

Trong ngành y tế Công an, khi ứng dụng công nghệ thông tin đáp ứng chuyển đổi số cũng đặt ra các yêu cầu bảo mật dữ liệu trong quản lý, lưu trữ dữ liệu và trao đổi thông tin qua mạng máy tính. Đặc biệt, cần bảo đảm tính bí mật một số thông tin cá nhân, do vậy việc mã hóa các thông tin cá nhân người bệnh trong ảnh số cũng là một trong những nhu cầu thực tiễn hiện nay.

Dựa trên các tìm hiểu tại Chương 1, Chương 2 học viên xin đề xuất thuật toán giấu tin mật trong cơ sở dữ liệu Data warehouse làm tiền đề tiến tới xây dựng chương trình phần mềm tìm kiếm các thông tin liên quan đến Bảo hiểm y tế nhằm hiện thực hóa các nhu cầu quản lý, chuyển đổi số trong lĩnh vực y tế như đã phân tích ở trên.

3.1. Thuật toán giấu tin và trích chọn tin mật

3.1.1. Thuật toán Giấu tin mật (embed)

Trên cơ sở ma trận H đã được xây dựng ở chương 2, ta đề xuất thuật toán giấu tin mật như sau:

Input: Bản thông điệp M = (m_1 m_2 m_3… m_n), $m_i \in$ {a, b, c, …, z}={0, 1, 2, …, 25}; Ảnh Bitmap C; khởi điểm giấu (điểm bắt đầu đặt dữ liệu vào ảnh C)

Output: Ảnh Stego S.

Bước 1. Dùng thuật toán nén Zip để nén bản thông điệm M, ta được Zip (M) = X = (x_1, x_2, …, x_k);

Bước 2. Chuyển dãy ký tự của X thành dãy nhị phân và phần kết quả được chia thàng từng block có độ dài bằng nhau và bằng 6 (nếu khối cuối cùng không bằng 6 thì thêm vào các số 0 cho đủ 6 bit); kết quả được ký hiệu là Y = (y_1, y_2, …, y_n);

Bước 3. Trích chọn 63n các LSB của các pixel dữ liệu ảnh của C bắt đầu từ khởi điểm cho trước, ta được:

Z = (z_1, z_2, …, z_n), mỗi Z_i = (z_{i1} z_{i2} … z_{i63}) i = 1, 2, …, n. $z_{ij} \in$ {0,1};

Bước 4. Với i = 1, 2, …, n, tính:

$u_i = y_i \Theta Hz_i$ (where x^T là chuyển vị của véctơ x) và phép toán Θ là phép cộng XOR;

T

, , ,

Bước 5. Với mỗi i tìm trong ma trận H có cột $h_{.i}$ nào trùng với u_i hay không nếu không tìm thấy thì cho qua và vectơ y_i được giữ nguyên;

Bước 6. Giả sử có tồn tại một j mà $h_{.i}$ trùng với u_i^T thì ta đảo bit thứ j của véctơ z_i tại bít z_{ij} tạo thành bit $z_{ij} = z_{ij}$ Θ 1 để nhận được véctơ $z_i = (z_{i1}, \ldots, z_{ij}, z_{ij+1}, \ldots, z_{i63})$ và quay lại Bước 4;

Bước 7. Trả lại tất cả các bit $z' = (z'_1, z'_2, \ldots, z'_n)$ lần lượt đúng như vị trí đã trích chọn z (tức là thay z bằng z' vào ảnh C) ta nhận được ảnh mới S.

3.1.2. Thuật toán Trích chọn (extract)

Input: Ảnh S và khởi điểm giấu.

Output: Thông điệp mật M.

Bước 1. Trích chọn đươc $(z'_1, z'_2, \ldots, z'_n) = z'$ với $z'_i = (z'_{i1}, z'_{i2}, \ldots, z'_{i63})$ i = 1, 2, …, n; như vậy: $z = (z_1, z_2, \ldots, z_{63n})$.

Bước 2. Chia dãy z thành từng block (mỗi block có độ dài 6) ta nhận được kết quả:

$Y = (Y_1, Y_2, \ldots, Y_n)$, $Y_i = (y_{i1}, y_{i2}, y_{i3}, y_{i4}, y_{i5}, y_{i6})$ i = 1, 2, …n;

Bước 3. For i = 1, 2, …, n. Tính $X_i^T = HY_i^T$ ta được véctơ $X = (X_1, X_2, \ldots, X_n)$;

Bước 4. Giải nén Zip (unzip): M = unzip (X) là thông điệp được giấu trong ảnh S.

Chú ý: muốn khôi phục lại ảnh gốc C ta chỉ cần giấu trở lại thông điệp M đã được trích chọn bắt đầu từ khởi điểm đã cho.

3.1.3. Phạm vi ứng dụng và lý do sử dụng thuật toán

Học viên đề xuất thuật toán trên nhằm mục đích mã hóa thông điệp 6 bit vào ảnh số thẻ BHYT. Khi sử dụng thuật toán này số lượng ký tự được mã hóa đã tăng lên từ 31 ký tự (mã hóa 5 bít) thành 63 ký tự (mã hóa 6 bít) đáp ứng yêu cầu thông tin cần giấu. Từ ma trận H với 63 mã nhị phân (trang 30) ta xây dựng được bảng mã tương ứng với chữ số, chữ cái la tinh cụ thể như sau:

STT	Ký tự	Từ mã
1		000000
2	a	010000
3	b	001000
4	c	000100
5	d	000010
6	e	000001
7	f	010100
8	g	001010
9	h	000101
10	i	010110

STT	Ký tự	Từ mã
22	u	000110
23	v	000011
24	w	010101
25	x	011110
26	y	001111
27	z	011101
28	/	011010
29	0	100000
30	1	110000
31	2	111000

STT	Ký tự	Từ mã
43	E	100011
44	F	100001
45	G	101010
46	H	101011
47	I	100101
48	J	110001
49	K	110010
50	L	110111
51	M	110011
52	N	101001

11	j	001011
12	k	010001
13	l	011100
14	m	001110
15	n	000111
16	o	010111
17	p	011111
18	q	011011
19	r	011001
20	s	011000
21	t	001100

32	3	111100
33	4	111110
34	5	111111
35	6	101000
36	7	101100
37	8	101110
38	9	101111
39	A	100100
40	B	100110
41	C	100111
42	D	100010

53	O	110100
54	P	111101
55	Q	111011
56	R	101101
57	S	111001
58	T	110101
59	U	111010
60	V	010011
61	W	001001
62	Y	010010
63	Z	001101

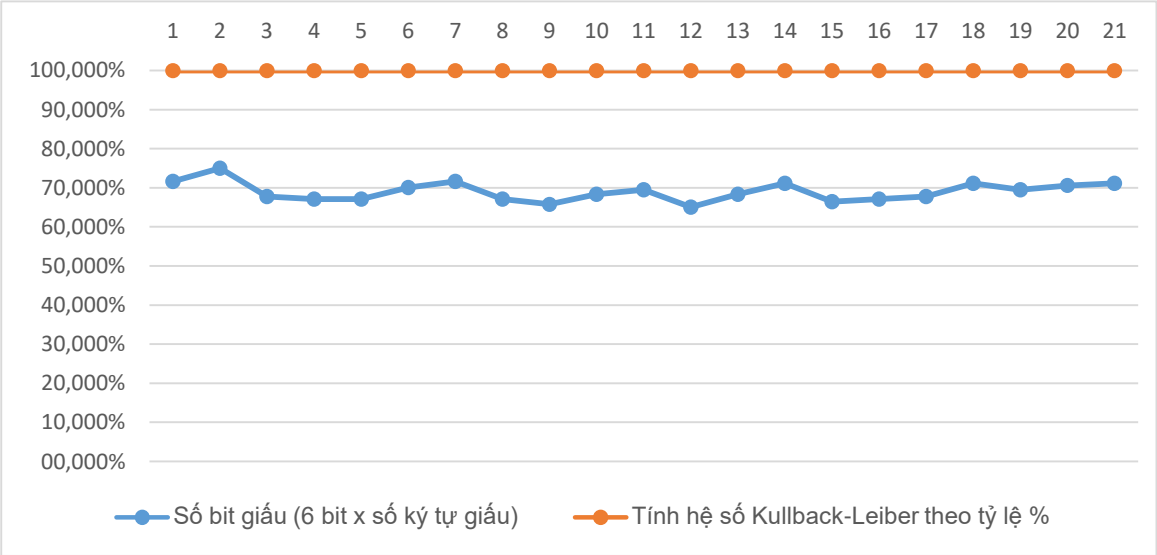
3.1.4. Thử nghiệm và đánh giá thuật toán

3.1.4.1. Thử nghiệm

Để so sánh hiệu quả giữa thuật toán được đề xuất ở trên và thuật toán đã được công bố trong [9], học viên sử dụng máy tính cấu hình CPU Intel core i5-6200U, 2.3Ghz 8Gb RAM. Thuật toán được mô phỏng thực nghiệm bởi ngôn ngữ Python.

Hình 3.1. Ứng dụng mã hóa thông điệp 6 bits vào ảnh

Kết quả thực nghiệm trên 21 ảnh chân dung, kích thước 4x6, có độ phân giải khoảng 300dpi, định dạng BMP được thể hiện trong bảng 3.1 dưới đây:



Hình 3.2. Biểu đồ K-L theo số Bit giấu tin

3.1.4.2. Đánh giá kết quả đạt được

Hiệu quả của thuật toán giấu tin mật được so sánh với một số thuật toán khác đã được công bố dựa trên 03 tiêu chí:

- Tỷ lệ thông tin giấu.
- Khả năng khó có thể phát hiện thông tin ẩn trong ảnh
- Tốc độ tính toán của thuật toán.

Để đánh giá mức độ an toàn của thuật toán vừa được trình bày ở trên, người ta thường sử dụng hai tham số: Sai số bình phương trung bình – MSE (*mean square error*) và phương pháp đề xuất với hệ số tỷ lệ tín hiệu/tín hiệu tạp PSNR (*Peak Signal to Noise Ratio*).

Thông thường, nếu PSNR ≥ 37 dB thì hệ thống mắt người gần như không phân biệt được giữa ảnh gốc và ảnh khôi phục. PSNR càng cao thì chất lượng ảnh khôi phục càng tốt [5]. Khi hai hình ảnh giống hệt nhau, MSE sẽ bằng 0 và PSNR đi đến vô hạn.

3.2. Ứng dụng trong ngành y tế

Chương trình phần mềm tìm kiếm các thông tin liên quan đến BHYT thiết lập theo mô hình Kho dữ liệu Data Warehouse. Chương trình được xây dựng xuất phát từ yêu cầu thực tế: tổng hợp số liệu để xây dựng báo cáo theo yêu cầu của Lãnh đạo cấp trên, cũng như phục vụ công tác chỉ đạo, điều hành của Lãnh đạo cấp trên dựa trên thiết kế theo định hướng lấy đối tượng tham gia BHYT làm trung tâm, và ứng dụng thông tin chính xác.

3.2.1. Phân tích yêu cầu chức năng của ứng dụng

Việc triển khai hệ thống Data wearhouse nhằm xây dựng một kho dữ liệu tập trung về BHYT của ngành Công an là công cụ hỗ trợ ra quyết định với công nghệ hiện đại, tiên tiến và thông minh sẽ mang đến nhiều lợi ích hơn. Hệ thống này sẽ góp phần cải cách thủ tục hành chính, và ứng dụng công nghệ thông tin trong quản lý các hoạt động nghiệp vụ của BHYT ngành Công an.

Hệ thống Data warehouse sẽ được vận hành với hai cấu phần chính: xây dựng một kho dữ liệu của BHYT ngành Công an; và các công cụ hỗ trợ là các ứng dụng công nghệ thông tin để cung cấp thông tin dữ liệu mang tính chính xác, toàn vẹn và duy nhất cho các cấp quản lý của BHYT ngành Công an, cũng như mang lại nhiều cách thức khai thác dữ liệu chuyên nghiệp, linh hoạt, đa dạng hơn.

3.2.2. Phân tích hệ thống

Yếu tố quan trọng nhất trong việc xây dựng hệ thống Data warehouse là bắt đầu từ đâu? Phân tích dữ liệu như thế nào? Đây cũng là "nút thắt" cản trở khi quyết định xây dựng hệ thống tương tự. Nguồn dữ liệu phân tán, được lưu trữ theo nhiều cách thức, tại nhiều đơn vị khác nhau, kể cả khi được mã hóa để lưu trữ điện tử cũng không có một chuẩn thống nhất là thực trạng chung đang diễn ra.

Chương trình phần mềm tìm kiếm các thông tin liên quan đến BHYT là phần mềm ứng dụng tập trung sẽ được cài đặt, lưu trữ tại Cục Chính sách- BCA và được triển khai tại các đơn vị. Hệ thống được thiết kế hỗ trợ việc thống kê, phân tích dữ liệu, thiết lập các báo cáo, hỗ trợ ra quyết định của Lãnh đạo Ngành. Trong đó, hệ thống Data warehouse sẽ được tích hợp với các hệ thống nguồn như sau:

1. Phần mềm Thu và quản lý thẻ BHYT.
2. Phần mềm Xét duyệt chính sách.
3. Phần mềm Giám định BHYT.
4. Phần mềm Kế toán tập trung.
5. Phần mềm Cấp mã số và quản lý thẻ BHYT.
6. Phần mềm Quản lý nhân sự.
7. Phần mềm Giao dịch điện tử.
8. Phần mềm Quản lý định danh và chia sẻ dữ liệu.

Từ những phân tích nêu trên, chương trình phần mềm tìm kiếm các thông tin liên quan đến BHYT cần được thiết kế dựa trên yêu cầu của 02 nhóm người sử dụng hệ thống Data warehouse.

3.2.2.1. Nhóm quản trị hệ thống Data warehouse

Là tập hợp những người sử dụng có chung đặc điểm, vai trò, tính chất và quyền hạn trong công tác quản trị, vận hành hệ thống. Trên hệ thống Data warehouse có các nhóm quản trị như: Quản trị ứng dụng cấp Bộ; Quản trị ứng dụng cấp Tỉnh.

3.2.2.2. Nhóm khai thác các chỉ tiêu báo cáo

Là tập hợp những người sử dụng có chung đặc điểm, vai trò, tính chất và quyền hạn thuộc đối tượng được lãnh đạo đơn vị giao nhiệm vụ sử dụng hệ thống Data warehouse để sử dụng hệ thống khai thác các chỉ tiêu, báo cáo.

3.2.3. Giao diện của hệ thống

3.2.3.1. Giao diện để truy cập hệ thống

Quyền truy cập hệ thông Data warehouse là quyền truy cập gắn với địa chỉ hộp thư điện tử của người sử dụng được cấp quyền đăng nhập và truy cập vào hệ thống Data warehouse, thực hiện các nhiệm vụ được giao.

3.2.3.2. Giao diện để truy cập báo cáo

Giao diện để truy cập báo cáo là công cụ dành cho người dùng khai thác, phân tích và phát triển các báo cáo phức tạp, nhiều trang, nhiều truy vấn với nhiều loại cơ sở dữ liệu. Người dùng có thể tạo bất kỳ báo cáo nào theo các yêu cầu quản lý như: tổng số thẻ bảo hiểm, tổng chi hàng tháng, chi một lần hay các hoạt động báo cáo cân đối kế toán…

3.2.4. Đánh giá hệ thống

3.2.4.1. Tính hiệu quả

Việc áp dụng Data warehouse sẽ giải quyết được vấn đề tích hợp, chia sẻ dữ liệu mà không phụ thuộc vào hệ điều hành và nền tảng công nghệ. Bên cạnh đó, việc triển khai kho dữ liệu là nền tảng tốt cho BHYT Công an, giúp lãnh đạo Y tế ngành Công an quản lý, điều hành và đưa ra quyết định một cách hiệu quả, chính xác và kịp thời.

Hiệu quả mà hệ thống tổng hợp và phân tích dữ liệu tập trung mang lại sẽ phát huy hơn nữa khi được tăng cường kết nối với các hệ thống báo cáo phục vụ cho công tác thống kê, tổng hợp, phân tích, đánh giá... Nhằm đưa ra các báo cáo phân tích các cấp, hỗ trợ công tác quản lý chặt chẽ, kịp thời, tổ chức thực hiện trở nên thiết thực và sâu sát hơn với các đối tượng tham gia BHYT của ngành Công an.

3.2.4.2. Các ưu và nhược điểm

Cần sớm triển khai xây dựng kho dữ liệu bởi vì nguồn dữ liệu của BHYT Công an ngày càng lớn, đa dạng và đã bộc lộ nhiều bất cập trong việc lưu trữ, quản lý, chia sẻ, khai thác sử dụng, nếu không bắt tay ngay vào việc xây dựng kho dữ liệu, thì Y tế Công an không những không hạn chế được bất cập, mà còn làm bất cập tăng lên đến mức sẽ không thể kiểm soát được.

Tuy nhiên, quá trình xây kho dữ liệu sẽ gặp nhiều khó khăn và thách thức như: có rất nhiều loại dữ liệu trong hệ thống; logic nghiệp vụ phức tạp; nguồn nhân lực chưa đáp ứng được yêu cầu… Để giải quyết vấn đề này, học viên đề xuất giải pháp xây dựng kho dữ liệu đầu vào theo chủ đề từ đó từng bước xây dựng kho dữ liệu đầu vào tập trung.

1. Việc xây dựng kho dữ liệu theo từng giai đoạn, trước tiên sẽ xây dựng kho dữ liệu đầu vào theo hướng kho dữ liệu chủ đề (data mark), sau đó sẽ tích hợp các kho dữ liệu chủ đề thành kho dữ liệu tập trung. Xây dựng kho dữ liệu của BHYT ngành Công an sẽ dựa trên nền tảng công nghệ thông tin sẵn có của BHYT Công an nhân dân, đó là công nghệ khách/chủ. Hệ quản trị CSDL thích hợp nhất với công nghệ khách/chủ là Microsoft SQL server phiên bản 2018. Chuyển đổi dữ liệu sẽ sử dụng công cụ có sẵn là SQL Server Integration Services của Microsoft SQL server.

2. Vần đề an ninh, an toàn mạng và bảo mật dữ liệu (đã được đề cập ở chương II) là vấn đề quan trọng, cần được đầu tư đồng bộ với công nghệ khách/chủ, hệ quản trị CSDL.

3. Hơn nữa, xây dựng kho dữ liệu của BHYT ngành Công an là công việc rất lớn và mới của ngành, do đó, cần có sự hỗ trợ kỹ thuật của chuyên gia có kinh nghiệm xây dựng kho dữ liệu.

KẾT LUẬN

Trong toàn bộ Đề tài luận văn của mình, em đã giải quyết được 03 vấn đề cơ bản sau đây:

1. Nghiên cứu, tìm hiểu tổng quan về Data warehouse. Đây là vấn đề không mới nhưng hiện nay nó có nhiều ứng dụng trong thực tiễn, đặc biệt dối với an ninh-quốc phòng. Hiện tại, luận văn đã trình bày những nét cơ bản nhất của hệ thống Data warehouse với mục đích đưa vào ứng dụng trong Ngành Y tế - BCA.

2. Tìm hiểu và xây dựng một thuật toán giấu tin mật trong môi trường ảnh kỹ thuật số. Đây là một lĩnh vực về an toàn – bảo mật thông tin hiện nay đang phát triển mạnh trên thế giới [16]. Ở Việt Nam ta chỉ mới có ứng dụng trong Ngành Công an và chưa được phát triển nhiều. Thuật toán mà học viên xây dựng chủ yếu là sự mở rộng của thuật toán đã được công bố trong [7]. Cụ thể: học viên đã tìm hiểu và cải tiến thuật toán giấu tin 5 bit thành thuật toán giấu tin 6 bit, để tăng số lượng ký tự mã hóa (từ 31 ký tự thành 63 ký tự). Chương trình thể hiện thuật toán trên máy tính đã cho chạy thử nghiệm trên 21 mẫu (trang 35). Trong đó, có so sánh giữa 02 thuật toán và kết quả thuật toán cải tiến tốt hơn (trang 37).

3. Đề xuất phương pháp bảo vệ thông tin trong lĩnh vực y tế bằng hệ thống Data warehouse có bảo mật. Chương này chỉ mới phác thảo mục đích gợi ý cho bảo toàn thông tin trong ngành Y tế - BCA. Ứng dụng của hệ thống Data warehouse mới dừng ở việc lên ý tưởng xây dựng chương trình phần mềm tìm kiếm các thông tin liên quan đến Bảo hiểm y tế, chưa triển khai được trong thực tế.

Bên cạnh những kết quả đã đạt được, đề tài này cần được tiếp tục phát triển và hoàn thiện trong các năm tiếp theo. Do thời gian nghiên cứu có hạn và trình độ hiểu biết của bản thân còn nhiều hạn chế nên khóa luận của học viên không tránh khỏi những thiếu sót. Học viên rất mong nhận được sự góp ý quý báu của tất cả các thầy cô giáo để khóa luận của học viên được hoàn thiện hơn, góp phần đưa luận văn vào thực tiễn.

Học viên xin chân thành cảm ơn!

DANH MỤC TÀI LIỆU THAM KHẢO

Tiếng Việt

[1]. PGS.Nguyễn Kim Anh (2016). *Các hệ cơ sở dữ liệu tiên tiến*, Trường Đại học Bách Khoa Hà Nội.

[2]. TS.HồVăn Canh, TS.Nguyễn Viết Thế (2010). *Nhập môn phân tích thông tin có bảo mật,* Nhà xuất bản Thông tin và Truyền thông, trang 304 - 308.

[3]. PGS.TS.Nguyễn Đức Nghĩa (2014). *Thiết kế và Phân tích thuật toán*, Trường Đại học Bách Khoa Hà Nội.

[4]. PGS.TS.Thái Hồng Nhị, TS.Phạm Minh Việt (2004). *An toàn thông tin*, Nhà xuất bản Khoa học và kỹ thuật, 188 trang.

[5]. Nguyễn Văn Tảo (2007). *Một số thuật toán giấu tin và áp dụng giấu tin mật trong ảnh*, Tạp chí Khoa học và Công nghệ, số 4(44), Tập 2.

[6]. Hồ Thị Hương Thắm (2012). *Luận án Tiến sĩ: Nghiên cứu đề xuất phương pháp nhận dạng ảnh có chứa thông tin ẩn,* Đại học Công nghệ - Đại học Quốc Gia Hà Nội, 2012.

[7]. Lê Hải Triều (2019). *Luận án Tiến sĩ: Nghiên cứu phương pháp bảo mật thông tin giấu trong ảnh số,* Học viện Công nghệ Bưu chính Viễn thông – Hà Nội, 2019.

Tiếng Anh

[8]. Bruyn donckx, O.J.,J. Quisquater, and B. Macq, "Spatial Method for Copyright Labeling of Digital Images", *In Non-linear Signal Processing Workshop, Thessloniki, Greece*, 1995, pp.456-459.

[9]. C-C-Raos (1968): "Linear Statisties and its applications", NXB MOSCOW, 1968.

[10]. Chanfang yang, Xiangyang Leo, and Fenlin Liu: "Embedding Ratio Estimating for Each Bit plane of Image", *Zenhgzhou Information Science and Technology Information Zhengzhou,* China 2015.

[11]. Cox, I., et al, "Seceere Spread Spectrum Watermarking for Multimedia", *Technical report, NEC Research Institute,* 1995.

[12]. Depovere, G., T. Kalker, and J. – P. M. G. Linnartz, "Improved Watermark Detection Reliability Using Filtering Before Correlation", *In Proceedings of the International*

Conference on Image Proceeding, vol. 1, IEEE Signal Proceeding Society, Chicago, ILLinois, USA, oct. 1998.

[13]. Fisher, Y. (ed), Fractal Image Compression: Theory and Application, New York Springer-Verlag, 1995.

[14]. M. Warkentin, M. B. Schmidt, E. Bekkering (2008): "Steganography and Analysis", *Premier Reference Source – Intellectual Property. Protection for Multimedia Information Technology*, *chapter XIX*, pp.374-380 (2008).

[15]. Puate, J., and F. Jordan, "Using Fractal Compression Scheme to Embed a Digital Signature in to an Image", *In Proceedings of the SPIE 2015. Video Techniques and Software for Full-Service Networks,* 1996, pp.108-118.

[16]. R. Ibrahim and J.S. Kuan (2010): "Steganography Imageng Syotem (SIS), hiding Seeeret Message inside an Image", *Proceedings of the world Congress on Engineering and Computer Science 2010 San Francisco,* USA, pp.144-148.

[17]. Rosziati Ibrahim and Jeoh Suk Kuan (2011): "Steganography Algorithm to Hide Seeret Message inside an Image". *Computer Technology and Application 2* (2011).

[18]. Stefan Katzenbeisser, Fabien A. P. Petitcolas: (2000): "Information Hiding Techniques for Steganography and Digital watermarking", *Artech House Boston. London*.

[19]. T. Jahnke, J. Seitz (2008): "An Introduction in digital watermarking Applications, Porneiples and problems", *in: H. Nemati (Ed), Security and Ethics: Concepts, Methodologies, Tools and Applications. NewYork: Information Science Reference*, pp.554-569.

MỤC LỤC