

BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG

LÊ MINH TRÍ

XÂY DỰNG HỆ THỐNG
THU NHẬN VÀ XỬ LÝ THÔNG TIN ĐIỆN VĂN
PHỤC VỤ CÔNG TÁC KIỂM TRA, THỐNG KÊ
TẠI TRUNG TÂM QUẢN LÝ BAY MIỀN TRUNG

Chuyên ngành : **KHOA HỌC MÁY TÍNH**
 Mã số : 60.48.01

TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT

Đà Nẵng – Năm 2011

Công trình được hoàn thành tại
ĐẠI HỌC ĐÀ NẴNG

Người hướng dẫn khoa học: **PGS.TS. Võ Trung Hùng**

Phản biện 1:.....

Phản biện 2:.....

Luận văn sẽ được bảo vệ trước Hội đồng chấm Luận
 văn tốt nghiệp Thạc sĩ Khoa học Máy tính họp tại Đại học
 Đà Nẵng vào ngày.....tháng....năm 2011

Có thể tìm hiểu luận văn tại:

- Trung tâm Thông tin - Học liệu, Đại học Đà Nẵng
- Trung tâm Học liệu, Đại học Đà Nẵng

MỞ ĐẦU

1. Lý do chọn đề tài

Hiện nay, các thông tin trong lĩnh vực hàng không đều được trao đổi trên một hệ thống, hệ thống đó được gọi là AMSS (Automatic Message Switching System) hay còn gọi là hệ thống chuyển tiếp điện văn tự động. Các thông tin được trao đổi bao gồm thông tin về kế hoạch bay, thời tiết, thông tin về các chuyến bay...và được chuyển đi dưới định dạng text. Hệ thống này được phát triển và đóng gói để cài đặt cho các trung tâm điều hành bay các khu vực và các trung tâm. Tuy nhiên, người dùng không thể can thiệp vào mã nguồn hệ thống để sửa đổi, bổ sung những tính năng mà họ cần. Vì vậy việc thống kê báo cáo về một số thông tin như: các chuyến bay đi, đến, quá cảnh qua khu vực miền Trung... điều được thống kê bằng thủ công, dựa vào các tập tin điện văn đã được in ra trong ngày để tập trung lại.

Vấn đề đặt ra là làm thế nào để có thể trích lọc các thông tin có trong điện văn từ hệ thống AMSS và tổng hợp lại để có thể giúp cho những kiểm soát viên không lưu có thể dễ dàng thống kê, theo dõi theo một thời gian nào đó mà ta muốn.

Với những lý do như trên và là người hiện đang công tác tại Trung tâm Quản lý bay miền Trung, tôi chọn đề tài “ ***Xây dựng hệ thống thu nhận và xử lý thông tin điện văn phục vụ công tác kiểm tra, thống kê tại Trung tâm Quản lý bay miền Trung***”.

2. Mục đích nghiên cứu

Hệ thống được xây dựng với mục đích trích các thông tin từ các tập tin điện văn, để đưa ra các thông tin cần thiết. Từ đó, ta sẽ tổng hợp tất cả các thông tin đó lại với nhau, và đưa ra những báo cáo về các thông tin mà người dùng yêu cầu (thời tiết một ngày nào đó, tổng số chuyến bay đi và đến trong một tháng...).

3. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu của đề tài gồm những tập tin điện văn trong hệ thống AMSS, kỹ thuật trích lọc thông tin. Phạm vi nghiên cứu của đề tài là hệ thống chuyển tiếp điện văn tự động AMSS, nhu cầu tra cứu thông tin lập báo cáo của kiểm soát viên không lưu thuộc Trung tâm Quản lý bay miền Trung.

4. Phương pháp nghiên cứu

Phương pháp được sử dụng trong đề tài là dựa vào các tập tin trong hệ thống AMSS, kỹ thuật trích lọc thông tin, khai phá dữ liệu, tra cứu thông tin. Từ đó tạo mô-dun đọc dữ liệu điện văn, trích lọc các thông tin chứa trong các điện văn từ hệ thống AMSS để xây dựng phần mềm cho phép người sử dụng có thể tra cứu thông tin cần thiết trên đó.

5. Ý nghĩa khoa học và thực tiễn của đề tài

Sau quá trình thực hiện đề tài, tôi đã nghiên cứu và ứng dụng thành công hệ thống thu nhận và xử lý thông tin điện văn. Hệ thống được sử dụng tại Trung tâm Quản lý bay miền Trung

6. Cấu trúc của luận văn

Luận văn được tổ chức thành 3 chương như sau:

Chương 1: Nghiên cứu tổng quan: chương này trình bày khái niệm về mạng viễn thông cố định hàng không, hệ thống chuyển tiếp điện văn tự động AMSS. Cơ sở lý thuyết về kỹ thuật trích lọc thông tin, khai phá dữ liệu.

Chương 2: Giải pháp đề xuất: chương này mô tả các yêu cầu đối với hệ thống, kiến trúc tổng thể và mô hình hoạt động của hệ thống. Giải pháp để triển khai hệ thống.

Chương 3: Phát triển ứng dụng: chương này trình bày công cụ để phát triển hệ thống, xây dựng các mô-dun trích lọc thông tin, xây dựng giao diện hiển thị thông tin và cho phép người sử dụng tra cứu thông tin. Quá trình cài đặt, thử nghiệm ứng dụng và đánh giá kết quả.

CHƯƠNG 1:NGHIÊN CỨU TỔNG QUAN

Trong chương này, tôi tập trung trình bày về khái niệm về hệ thống thông tin điện văn trong lĩnh vực hàng không, hệ thống chuyển tiếp điện văn tự động AMSS, kỹ thuật trích lọc thông tin, khai phá dữ liệu. Những nội dung trong chương này là cơ sở để thực hiện các chương tiếp theo.

1.1. Hệ thống thông tin điện văn trong lĩnh vực hàng không

Trung tâm Quản lý bay dân dụng Việt Nam là đơn vị có nhiệm vụ quản lý mạng AFTN quốc gia, đảm bảo duy trì thông tin liên lạc điện văn phục vụ công tác điều hành bay một cách liên tục và an toàn.

AFTN là loại dịch vụ chuyển bản tin loại store-and-forward (là loại chuyển điện văn mà các node trung gian sẽ nhận đầy đủ một bản tin trước khi chuyển tiếp đến node khác. Loại chuyển điện văn này sẽ yêu cầu nhiều bộ nhớ hơn tại các node trung gian) để chuyển các bản tin ký tự. Các bản tin AFTN được chuyển theo từng bước nhảy theo những tuyến được cấu hình trước để chuyển tới địa chỉ trong thời gian ngắn nhất.

1.1.1. Mạng viễn thông cố định hàng không AFTN

1.1.2. Điện văn hàng không

Điện văn hàng không là các bản tin được chuyển đi trên mạng viễn thông cố định hàng không. Các bản tin này chứa tất cả các thông tin liên quan đến lĩnh vực hàng không như:

- Điện văn kế hoạch bay: bao gồm thông tin của một chuyến bay sắp khởi hành hoặc sắp hạ cánh.
- Điện văn khí tượng: bao gồm các thông tin về dự báo thời tiết sân bay đến, dự báo khu vực và đường bay.

Ngoài ra còn những loại điện văn như: điện văn dịch vụ không báo hàng không, điện văn hàng chính hàng không, các điện văn sự vụ.

1.2. Hệ thống chuyển tiếp điện văn tự động AMSS

1.2.1. Giới thiệu

Hệ thống AMSS (Automatic Messages Switch System) hay còn gọi là hệ thống chuyển tiếp điện văn tự động. Hệ thống có chức năng thu nhận và xử lý điện văn trong mạng AFTN. Một hệ thống AMSS thuần túy thực hiện đúng chức năng chuyển tiếp điện văn AFTN. Đồng thời, đảm bảo cho hệ thống AMSS có giao tiếp mở cho phép kết nối mềm dẻo với các hệ thống xử lý dữ liệu khác có liên quan như: xử lý số liệu bay, số liệu thời tiết, sự vụ hàng không...

Hệ thống có khả năng hoạt động ở chế độ dự phòng nóng. Hai máy chủ hoạt động song song. Khi máy chính có sự cố, hệ thống sẽ tự động chuyển sang máy dự phòng và cảnh báo để kỹ thuật viên xử lý. Hệ thống được thiết kế đáp ứng được lưu lượng 30,000 điện văn một ngày. Có khả năng kết nối và quản lý 40 kênh truyền AFTN.

Có khả năng kiểm soát tình trạng hoạt động của tất cả các kênh và thiết bị trong hệ thống. Đảm bảo khả năng kiểm soát mạch UP/DOWN.

Có khả năng sử dụng hai loại Modem là tự động kết nối và quay số (leased line , dialup). Tốc độ kênh truyền đến các đầu cuối có thể thiết lập từ 50 bauds đến 230 kbps theo từng mức đối với các đầu cuối sử dụng cổng RS-232, 10/100 Mbs đối với đầu cuối trong mạng LAN. Có các tệp nhật ký ghi nhận các sự kiện trong quá trình hoạt động của tất cả các vị trí trong hệ thống. Các sự kiện được ghi vào tệp Log.

Hệ thống có khả năng định cấu hình trực tuyến với mức độ tham số hoá cao cho các tác vụ mà không phải Reset lại hệ thống.

Hệ thống có cơ chế quản lý người dùng và an ninh để đảm bảo độ ổn định và an toàn của toàn cho hệ thống.

1.2.2. Những chức năng chính của hệ thống AMSS

1.2.2.1. Xử lý điện văn

1.2.2.2. Lưu trữ điện văn

1.2.2.3. Truy xuất điện văn

1.2.2.4. Thống kê

1.3. Kỹ thuật trích lọc thông tin (Information Extraction)

1.3.1. Giới thiệu về trích lọc thông tin

Trích lọc thông tin là kỹ thuật được sử dụng sau khi thông tin qua tác vụ thu thập đã được lấy về, việc tiếp theo là phải lấy ra được những thông tin cần thiết và chỉ là những thông tin mà mình cần một cách tự động không cần tới sự thao tác của người sử dụng. Hầu hết các thuật toán trích lọc thông tin hiện nay đều dựa vào các công cụ khác nhau trên nền kỹ thuật “Wrapper”. Wrapper có thể được hiểu là những hàm để tách thông tin từ các tài nguyên. Các hàm này được viết dựa trên các luật (quy luật) đã được đúc rút ra sau khi khảo sát các bản văn chứa thông tin cần lấy. Các Wrapper có thể xây dựng dựa trên rất nhiều quy luật khác nhau và tùy thuộc vào mục đích của người sử dụng.

1.3.2. Các dạng trích lọc thông tin

Việc trích lọc nội dung văn bản phụ thuộc vào định dạng tập tin mà người dùng định nghĩa cho phép nhận diện. Các tập tin này phải được khai báo trước và có cách đọc cũng như cách trích lọc khác nhau. Sau đây là các bước trích lọc nội dung văn bản của từng định dạng tập tin.

1.3.2.1. Đối với tập tin HTML

1.3.2.1. Đối với tập tin PPT

1.4. Khai phá dữ liệu (Data Mining)

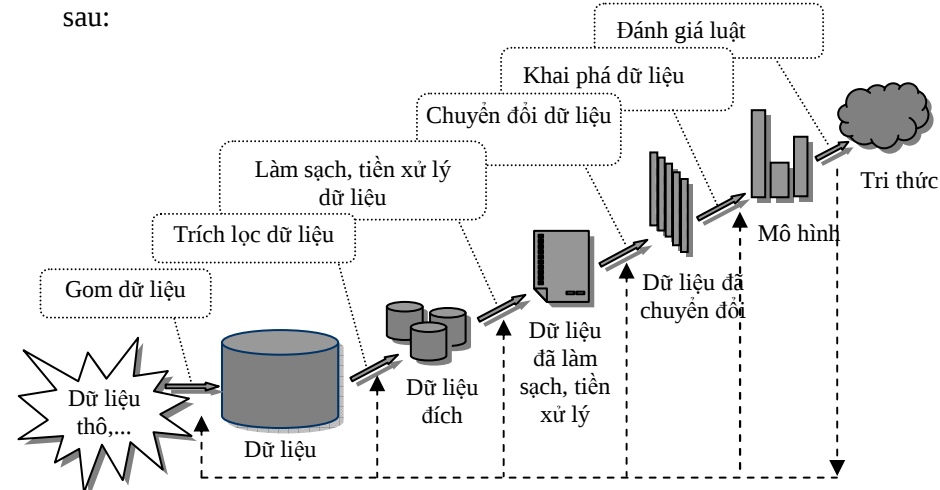
1.4.1. Giới thiệu về khai phá dữ liệu

Khai phá dữ liệu là một khái niệm ra đời vào những năm cuối của thập kỷ 1980. Nó là quá trình khám phá thông tin ẩn được tìm thấy trong các cơ sở dữ liệu và có thể xem như là một bước trong quá trình khám phá tri thức. Khai phá dữ liệu là giai đoạn quan trọng nhất trong

tên trình khai phá tri thức từ cơ sở dữ liệu, các tri thức này hỗ trợ trong việc ra quyết định trong khoa học và kinh doanh. Để hình dung vấn đề này, ta có thể sử dụng một ví dụ đơn giản sau: khai phá dữ liệu được ví như tìm một cây kim trong đồng cỏ khô.

1.4.2. Quá trình khai phá dữ liệu

Một quá trình khai phá dữ liệu bao gồm 6 giai đoạn chính như sau:



Hình 1.1: Quá trình phát hiện tri thức

- Gom dữ liệu (Gathering)
- Trích lọc dữ liệu
- Làm sạch, tiền xử lý và chuẩn bị trước dữ liệu
- Chuyển đổi dữ liệu
- Phát hiện và trích mẫu dữ liệu
- Đánh giá kết quả mẫu

Trên đây là 6 giai đoạn trong quá trình khai phá dữ liệu, trong đó giai đoạn 5 là giai đoạn được quan tâm nhiều nhất hay còn gọi là Data Mining

1.4.3. Các kiểu khai phá dữ liệu

1.4.3.1. Khai phá dữ liệu dự đoán

Nhiệm vụ của khai phá dữ liệu dự đoán là đưa ra các dự đoán dựa vào các suy diễn trên dữ liệu hiện thời. Nó sử dụng các biến hay các trường trong cơ sở dữ liệu để dự đoán các giá trị không biết hay các giá trị tương lai. Bao gồm các kỹ thuật: phân loại (classification), hồi quy (regression)...

a. Phân loại

Mục tiêu của phương pháp phân loại dữ liệu là dự đoán nhãn lớp cho các mẫu dữ liệu. Quá trình phân loại dữ liệu thường gồm 2 bước: xây dựng mô hình và sử dụng mô hình để phân loại dữ liệu.

Bước 1: Xây dựng mô hình dựa trên việc phân tích các mẫu dữ liệu cho trước. Mỗi mẫu thuộc về một lớp, được xác định bởi một thuộc tính gọi là thuộc tính lớp. Các mẫu dữ liệu này còn được gọi là tập dữ liệu huấn luyện. Các nhãn lớp của tập dữ liệu huấn luyện đều phải được xác định trước khi xây dựng mô hình, vì vậy phương pháp này còn được gọi là học có giám sát.

Bước 2: Sử dụng mô hình để phân loại dữ liệu. Trước hết chúng ta phải tính độ chính xác của mô hình. Nếu độ chính xác là chấp nhận được, mô hình sẽ được sử dụng để dự đoán nhãn lớp cho các mẫu dữ liệu khác trong tương lai.

Hay nói cách khác, phân loại là học một hàm ánh xạ một mục dữ liệu vào một trong số các lớp cho trước.

b. Hồi quy

Phương pháp hồi quy khác với phân loại dữ liệu ở chỗ, hồi quy dùng để dự đoán về các giá trị liên tục, còn phân loại dữ liệu thì chỉ dùng để dự đoán các giá trị rời rạc.

Hồi quy là học một hàm ánh xạ một mục dữ liệu vào một biến dự báo giá trị thực. Các ứng dụng hồi quy có nhiều, ví dụ như đánh giá xác suất một bệnh nhân sẽ chết dựa trên tập kết quả xét nghiệm chẩn đoán,

dự báo nhu cầu của người tiêu dùng đối với một sản phẩm mới dựa trên hoạt động quảng cáo tiêu dùng.

1.4.3.2. Khai phá dữ liệu mô tả

Kỹ thuật này có nhiệm vụ mô tả về các tính chất hoặc các đặc tính chung của dữ liệu trong CSDL hiện có. Bao gồm các kỹ thuật: phân cụm (clustering), phân tích luật kết hợp (association rules)...

a. Phân cụm

Mục tiêu chính của phương pháp phân cụm dữ liệu là nhóm các đối tượng tương tự nhau trong tập dữ liệu vào các cụm sao cho các đối tượng thuộc cùng một cụm là tương đồng còn các đối tượng thuộc các cụm khác nhau sẽ không tương đồng. Phân cụm dữ liệu là một ví dụ của phương pháp học không giám sát. Không giống như phân loại dữ liệu, phân cụm dữ liệu không đòi hỏi phải định nghĩa trước các mẫu dữ liệu huấn luyện. Vì thế, có thể coi phân cụm dữ liệu là một cách học bằng quan sát (learning by observation), trong khi phân loại dữ liệu là học bằng ví dụ (learning by example). Trong phương pháp này ta sẽ không thể biết kết quả các cụm thu được sẽ như thế nào khi bắt đầu quá trình. Vì vậy, thông thường cần có một chuyên gia về lĩnh vực đó để đánh giá các cụm thu được. Phân cụm dữ liệu được sử dụng nhiều trong các ứng dụng về phân đoạn thị trường, phân đoạn khách hàng, nhận dạng mẫu, phân loại trang Web... Ngoài ra phân cụm dữ liệu còn có thể được sử dụng như một bước tiền xử lý cho các thuật toán khai phá dữ liệu khác.

b. Luật kết hợp

Mục tiêu của phương pháp này là phát hiện và đưa ra các mối liên hệ giữa các giá trị dữ liệu trong CSDL. Mẫu đầu ra của giải thuật khai phá dữ liệu là tập luật kết hợp tìm được. Khai phá luật kết hợp được thực hiện qua 2 bước:

- **Bước 1:** tìm tất cả các tập mục phổ biến, một tập mục phổ biến được xác định qua tính độ hỗ trợ và thỏa mãn độ hỗ trợ cực tiểu.

- **Bước 2:** sinh ra các luật kết hợp mạnh từ tập mục phổ biến, các luật phải thỏa mãn độ hỗ trợ cực tiểu và độ tin cậy cực tiểu.

Phương pháp này được sử dụng rất hiệu quả trong các lĩnh vực như marketing có chủ đích, phân tích quyết định, quản lý kinh doanh,...

1.4.4. Kỹ thuật phân loại trong khai phá dữ liệu

Các cơ sở dữ liệu với rất nhiều thông tin ẩn có thể được sử dụng để tạo nên các quyết định kinh doanh thông minh. Phân loại là một dạng của phân tích dữ liệu, nó dùng để trích ra các mô hình mô tả các lớp dữ liệu quan trọng hay để dự đoán các khuynh hướng dữ liệu tương lai. Phân loại dùng để dự đoán các nhãn xác thực (hay các giá trị rời rạc). Nhiều phương pháp phân loại được đề xuất bởi các nhà nghiên cứu các lĩnh vực như học máy, hệ chuyên gia, thống kê... Hầu hết các giải thuật dùng với giả thiết kích thước dữ liệu nhỏ. Các nghiên cứu khai phá cơ sở dữ liệu gần đây đã phát triển, xây dựng mở rộng các kỹ thuật phân loại có khả năng sử dụng dữ liệu thường trú trên đĩa lớn. Các kỹ thuật này thường được xem xét xử lý song song và phân tán.

1.4.4.1. Khái niệm về phân loại

1.4.4.2. Các vấn đề quan tâm của phân loại

a. Chuẩn bị dữ liệu để phân loại

Các bước tiền xử lý dữ liệu sau đây giúp cải thiện độ chính xác, hiệu suất và khả năng mở rộng của phân loại.

- **Làm sạch dữ liệu:** Đây là quá trình thuộc về tiền xử lý dữ liệu để gỡ bỏ hoặc làm giảm nhiễu và cách xử lý các giá trị khuyết. Bước này giúp làm giảm sự mập mờ khi học.
- **Phân tích sự thích hợp:** Nhiều thuộc tính trong dữ liệu có thể không thích hợp hay không cần thiết để phân loại. Vì vậy, phép phân tích sự thích hợp được thực hiện trên dữ liệu với mục đích gỡ bỏ bất kỳ những thuộc tính không thích hợp hay không cần thiết. Trong học máy, bước này gọi là trích chọn

đặc trưng. Phép phân tích này giúp phân loại hiệu quả và nâng cao khả năng mở rộng.

- **Biến đổi dữ liệu:** Dữ liệu có thể được tổng quát hoá tới các mức khái niệm cao hơn. Điều này rất hữu ích cho các thuộc tính có giá trị liên tục. Ví dụ, các giá trị số của thuộc tính *thu nhập* được tổng quát hoá sang các phạm vi rời rạc như *thấp, trung bình và cao*. Tương tự, các thuộc tính giá trị tên như *đường phố* được tổng quát hoá tới khái niệm mức cao hơn như *thành phố*. Nhờ đó các thao tác vào/ra trong quá trình học sẽ ít đi.

b. So sánh các phương pháp phân loại

1.4.4.3. Phân loại bằng cây quyết định quy nạp

Cây quyết định là cấu trúc cây có dạng biểu đồ luồng, mỗi nút trong là kiểm định trên một thuộc tính, mỗi nhánh đại diện cho một kết quả kiểm định, các nút lá đại diện cho các lớp. Nút cao nhất trên cây là nút gốc. Để phân loại một mẫu chưa biết, các giá trị thuộc tính của mẫu sẽ được kiểm định trên cây. Đường đi từ gốc tới một nút lá cho biết dự đoán lớp đối với mẫu đó. Cây quyết định có thể dễ dàng chuyển đổi thành các luật phân loại. Giải thuật nền tảng của cây quyết định quy nạp là ID3, một giải thuật cây quyết định quy nạp nổi tiếng.

a. Chiến lược cơ bản của ID3

Giải thuật cây quyết định quy nạp bao gồm các chiến lược sau:

- Cây bắt đầu là một nút đơn đại diện cho các mẫu huấn luyện.
- Nếu tất cả các mẫu cùng lớp thì nút trở thành một lá và được gắn nhãn với lớp đó.

Ngược lại, giải thuật sử dụng một phép đo Entropy để lựa chọn thuộc tính. Đây là thuộc tính sẽ phân tách tốt nhất các mẫu vào trong các lớp riêng biệt. Thuộc tính này trở thành thuộc tính "kiểm định" hay "quyết định" tại nút đó. Trong version này của giải thuật, tất cả các

thuộc tính đều là xác thực, tức là giá trị rời rạc. Các thuộc tính giá trị liên tục phải được rời rạc hóa.

Một nhánh được tạo lập cho từng giá trị đã biết của thuộc tính kiểm định và các mẫu được phân chia một cách phù hợp.

Giải thuật sử dụng cùng xử lý đệ quy để hình thành nên cây quyết định cho các mẫu tại mỗi lần phân chia.

Phân chia đệ quy này dừng khi một trong những điều kiện sau là đúng:

- Tất cả các mẫu thuộc về cùng một lớp
- Không còn thuộc tính nào để tiếp tục phân chia các mẫu. Trong trường hợp này, lựa chọn theo số đông (*majority voting*) được dùng. Lúc này nút được tạo trở thành lá với nhãn là lớp đã lựa chọn theo số đông.
- Không còn mẫu nào cho nhánh $test-attribute = a_i$. Lúc này, một lá được tạo với nhãn là lớp chiếm đa số trong các mẫu.

b. Phép đo lựa chọn thuộc tính

1.4.5. Kỹ thuật phân cụm trong khai phá dữ liệu

1.4.5.1. Khái niệm về phân cụm

Xử lý nhóm một tập các đối tượng vào trong các lớp các đối tượng giống nhau được gọi là phân cụm. Một cụm là một tập hợp các đối tượng dữ liệu giống nhau trong phạm vi cùng một cụm và không giống nhau với các đối tượng trong các cụm khác. Phép phân tích cụm là một hoạt động quan trọng. Bằng phân cụm, ta có thể nhận biết các vùng đông đúc và thưa thớt, bởi vậy tìm ra toàn bộ các mẫu phân bố và các tương quan thú vị giữa các thuộc tính dữ liệu. Trong kinh doanh, phân cụm có thể giúp cho các nhà nghiên cứu thị trường tìm ra các nhóm riêng biệt dựa trên khách hàng của họ và mô tả các nhóm khách hàng dựa trên các mẫu mua sắm.

Như là một nhánh của thống kê, phép phân tích cụm được nghiên cứu mở rộng đã nhiều năm, tập trung chính trên phép phân tích cụm dựa

trên khoảng cách. Các công cụ phân tích cụm dựa trên *k-means*, *k-medoids* và một số các phương pháp khác cũng được xây dựng trong nhiều gói phần mềm hay hệ thống phân tích thống kê như S-Plus, SPSS và SAS.

1.4.5.2. Các kiểu dữ liệu trong ghép phân cụm

1.4.5.3. Độ tương đồng và không tương đồng

1.4.5.4. Phân loại các phương pháp phân cụm chính

Hiện có một số lượng lớn các giải thuật phân cụm trong các tài liệu. Việc lựa chọn giải thuật phân cụm tùy thuộc vào kiểu dữ liệu cho sẵn, mục đích riêng và ứng dụng. Nếu như phép phân tích cụm được dùng như một công cụ mô tả hay thăm dò thì có thể thử một vài giải thuật trên cùng dữ liệu để xem dữ liệu có thể thể hiện được điều gì. Nhìn chung, các phương pháp phân cụm chính được phân thành các loại như sau.

a. Các phương pháp phân chia

b. Các phương pháp phân cấp

c. Các phương pháp dựa trên mật độ

d. Phương pháp dựa trên lưới

Một phương pháp dựa trên lưới lượng tử hoá không gian đối tượng vào trong một số hữu hạn các ô hình thành nên một cấu trúc lưới. Sau đó nó thực hiện tất cả các thao tác phân cụm trên cấu trúc lưới (tức là trên không gian đã lượng tử hoá). Thuận lợi chính của tiếp cận này là thời gian xử lý nhanh chóng của nó độc lập với số các đối tượng dữ liệu và chỉ tùy thuộc vào số lượng các ô trong mỗi chiều của không gian lượng tử.

CHƯƠNG 2: GIẢI PHÁP ĐỀ XUẤT

Đây là một trong những chương quan trọng của đề tài. Trong chương này trình bày các nội dung sau: bài toán đặt ra, mô tả hệ thống bao gồm: xác định yêu cầu đối với hệ thống, kiến trúc tổng thể và mô hình hoạt động của hệ thống; từ hiện trạng hệ thống phục vụ tra cứu

thông tin từ tập tin điện văn ta thiết kế cơ sở dữ liệu và cuối cùng là thiết kế các chức năng.

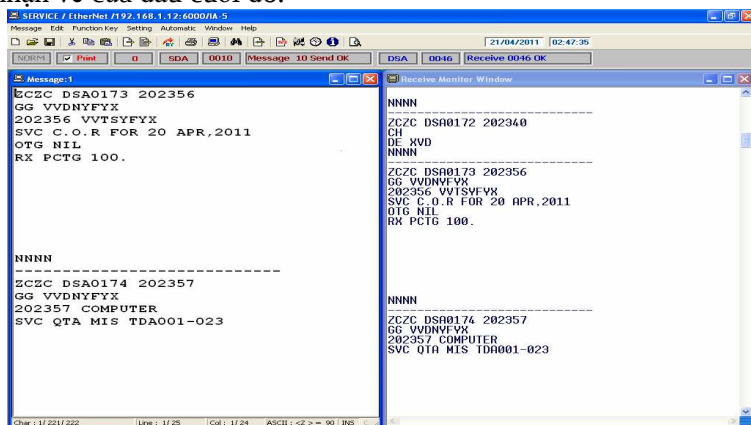
2.1. Mô tả bài toán

Hiện nay, tại Đài chỉ huy cất hạ cánh sân bay Đà Nẵng, những thông tin thuộc lĩnh vực hàng không dân dụng mà các kiểm soát viên không lưu khai thác đều từ những điện văn được chuyển đi trên hệ thống AMSS. Những điện văn này bao gồm tất cả các thông tin liên quan đến lĩnh vực hàng không dân dụng như: kế hoạch bay, điện văn thời tiết, điện văn sự vụ..... Vì thông tin điện văn chỉ hiện thị theo mẫu sẵn, trong một vài trường hợp điều này có thể dẫn đến sai sót khi kiểm soát viên không lưu đọc không chính xác các thông tin có trong điện văn.

Dưới đây là giao diện phần mềm khai thác điện văn mà các kiểm soát viên không lưu sử dụng. Giao diện này gồm 2 phần chính:

- Phần bên trái: đây là phần soạn thảo điện văn của phần mềm.
- Phần bên phải: đây là phần hiển thị các điện văn mà đầu cuối đó nhận được.

Ngoài ra, phần mềm còn hiển thị tổng số điện văn đã được gửi đi và nhận về của đầu cuối đó.



Hình 2.1: Giao diện phần mềm khai thác điện văn tại đầu cuối

Bài toán đặt ra ở đây là từ những tập tin điện văn ban đầu được lưu trữ trong hệ thống AMSS, ta sử dụng kỹ thuật bóc tách thông tin, chia nhỏ tập tin điện văn đó ra thành nhiều tập tin text nhỏ chứa thông tin mà ta cần và đưa vào cơ sở dữ liệu. Tiếp đó ta sẽ tiến hành khai phá dữ liệu trên cơ sở dữ liệu đó và đưa ra được những thông tin mà người sử dụng yêu cầu thông qua phần mềm tra cứu thông tin.

- Hệ thống được xây dựng bao gồm 3 phần chính:

- **Sưu tập dữ liệu:** ở phần này, ta sẽ tiến hành lấy dữ liệu từ hệ thống AMSS, cụ thể là lấy tất cả các tập tin điện văn liên quan mà ta cần phân tích.
- **Xây dựng cơ sở dữ liệu:** ở phần này, ta sử dụng kỹ thuật trích lọc thông tin để lấy các thông tin trong tập tin điện văn mà ta cần. Tạo một cơ sở dữ liệu để lưu trữ các thông tin đã được trích lọc.
- **Khai phá thông tin:** tại phần này, dựa trên yêu cầu người dùng. Xây dựng giao diện tra cứu thông tin mà người dùng cần xem.

- Dưới đây là mô hình hệ thống xử lý thông tin điện văn:

tin mà chúng ta cần để đưa vào cơ sở dữ liệu.

2.2.2. Bóc tách thông tin từ tập tin điện văn

Vì điện văn được chuyển trong hệ thống AMSS theo một mẫu nhất định, vì thế ta chỉ cần quan tâm nội dung của điện văn.

2.2.2.1. Các mẫu điện văn không lưu

Điện văn không lưu là những điện văn được sử dụng để trao đổi tin tức giữa những người sử dụng qua mạng viễn thông hàng không cố định (AFTN) và di động.

- Mỗi điện văn không lưu gồm có: Mã điện văn; ký hiệu chỉ độ khẩn (ưu tiên); nhóm địa chỉ nơi gửi, nơi nhận và thời gian; các thành phần dữ liệu được tổ chức thành từng nhóm (field).
- Căn cứ theo mục đích sử dụng, điện văn không lưu được chia thành các loại điện văn với các độ khẩn tương ứng.
- Các điện văn liên quan đến việc cung cấp dịch vụ không lưu sẽ do các cơ sở cung cấp dịch vụ không lưu được chỉ định phát.
- Các điện văn liên quan đến các hoạt động và kiểm soát; điện văn thông báo tin tức về chuyến bay mà không vì mục đích cung cấp dịch vụ không lưu sẽ do tổ lái, hãng hàng không hoặc đại diện nhà khai thác phát.
- Các điện văn kế hoạch bay; sửa đổi hoặc hủy bỏ kế hoạch bay chỉ được gửi đến các cơ sở cung cấp dịch vụ có liên quan.

- a. Điện văn kế hoạch bay không lưu (FPL)
- b. Điện văn cất cánh (DEP - Departure)
- c. Điện văn hạ cánh (Arr – Arrival)
- d. Điện văn hủy bỏ kế hoạch bay
- e. Điện văn kế hoạch bay hiện hành
- f. Điện văn khí tượng Metar

2.2.2.2. Một số qui luật dùng để bóc tách thông tin từ điện văn

Với những thông tin chứa trong tập tin điện văn và theo khảo sát yêu cầu người dùng. Phần trích lọc thông tin, báo cáo chỉ tập trung vào các điện văn kế hoạch bay và điện văn thời tiết. Dữ liệu sau khi trích lọc được tổng hợp bằng việc trích lọc những thông tin mô tả và lưu vào cơ sở dữ liệu

Đối với điện văn kế hoạch bay, phần đầu mọi điện văn là như nhau, ta chỉ căn cứ vào từ khóa FPL để tiến hành tách nội dung trong đó và khi nào gặp ký tự kết thúc thì ta dừng.

Đối với điện văn khí tượng, phần đầu mọi điện văn là như nhau, ta căn cứ vào hai dạng từ khóa đó là **Metar** và **Speci** để tiến hành tách nội dung về thời tiết.

2.2.3. Quá trình khai phá dữ liệu

Từ dữ liệu đã trích lọc được từ tập tin điện văn, ta tiến hành khai phá dữ liệu đó. Quá trình khai phá dữ liệu này bao gồm các bước sau:

2.2.3.1. Khảo sát yêu cầu người dùng

Vì chương trình nhằm phục vụ cho các kiểm soát viên không lưu nên quá trình khảo sát chỉ tập trung vào nhu cầu khai thác điện văn thuộc kế hoạch bay, thời tiết tại khu vực Đà Nẵng. Một số yêu cầu chính sau khi đã tiến hành khảo sát:

- Hiện thị thông tin chứa trong một điện văn kế hoạch bay như: mã hiệu chuyến bay, tên chuyến bay, sân bay xuất phát, sân bay đến, sân bay dự bị khi cần hạ cánh khẩn cấp, phương thức bay, thời gian bay...
- Hiện thị thông tin chứa trong điện văn thời tiết như: thời tiết tại khu vực sân bay Đà Nẵng trong từng thời điểm cụ thể.
- Tổng hợp các chuyến bay đi, bay đến phục vụ cho việc thống kê sản lượng bay hàng tháng.

2.2.3.2. Xây dựng các luật truy vấn

2.3. Nhận xét

Giải pháp đã đáp ứng được yêu cầu mà bài toán đặt ra trong đề tài. Nó giúp cho việc diễn giải các thông tin có trong các điện văn kế hoạch bay, điện văn thời tiết được rõ ràng hơn so với việc phải đọc toàn bộ điện văn. Giao diện tương tác người dùng được thiết kế giúp người sử dụng dễ dàng tương tác với ứng dụng, nó hiển thị thông tin rõ ràng dễ đọc.

CHƯƠNG 3: PHÁT TRIỂN ỨNG DỤNG

Trong chương này tập trung trình bày về ngôn ngữ phát triển ứng dụng, tạo cơ sở dữ liệu và các mô-đun của ứng dụng. Quá trình cài đặt ứng dụng.

3.1. Công cụ phát triển ứng dụng

3.1.1. Lưu trữ cơ sở dữ liệu

3.1.2. Ngôn ngữ phát triển ứng dụng

Visual Studio.Net là 1 sản phẩm công nghệ .NET của Microsoft với nhiều ngôn ngữ lập trình: C++.Net, Visual Basic.Net, ASP.Net, C#, J#, giúp chúng ta có thể dễ dàng tạo ra những giao diện dễ dàng, nhanh chóng với những công cụ hỗ trợ đi kèm. Trong đề tài này tôi chọn công nghệ .NET với ngôn ngữ C# để phát triển ứng dụng.

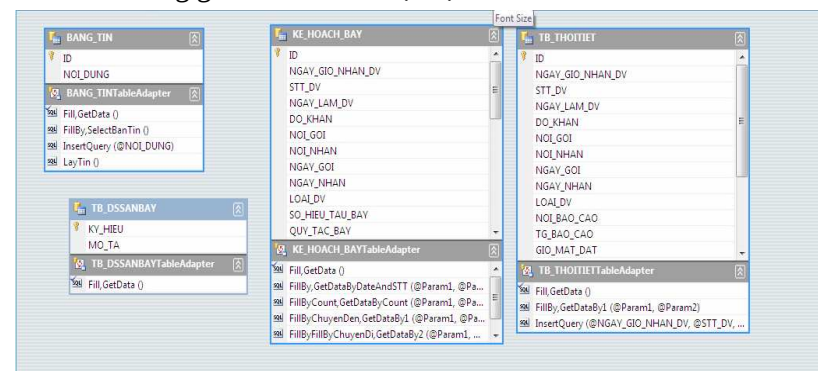
3.2. Xây dựng cơ sở dữ liệu

Dựa vào những thành phần của tập tin điện văn sau khi đã được trích lọc, ta xây dựng cơ sở dữ liệu của chương trình bao gồm các bảng ghi sau:

- Bảng ghi BANG_TIN: dùng để lưu nội dung của điện văn. Bảng này gồm 2 trường ID và NOI_DUNG
- Bảng ghi KE_HOACH_BAY: dùng để lưu nội dung điện văn kế hoạch bay sau khi đã trích lọc thông tin.
- Bảng ghi TB_THOI_TIET: dùng để lưu các thông tin liên quan đến thời tiết tại khu vực miền Trung sau khi đã trích lọc thông tin.

d. Bảng ghi TB_DSSANBAY: bảng này dùng để lưu thông tin về các sân bay có trong điện văn mà chúng ta trích lọc được. Nó phục vụ cho công tác thống kê sản lượng bay

Các bảng ghi sau khi đã được tạo hoàn tất:



Hình 3.1: Các bảng ghi của chương trình

3.3. Xây dựng các mô-đun của ứng dụng

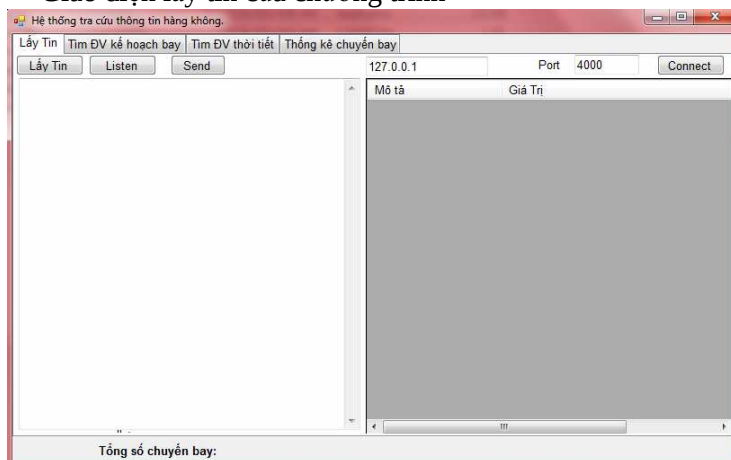
3.3.1. mô-đun trích lọc thông tin

mô-đun này được xây dựng và tích hợp thẳng vào chương trình, nó sẽ đọc dữ liệu điện văn từ bộ đệm và phân tích dữ liệu đó. Trong quá trình đọc dữ liệu, nó tách riêng từng điện văn dựa vào 2 từ khóa: bắt đầu điện văn (ZCZC) và kết thúc điện văn (NNNN). Tiếp theo nó sẽ phân tích nội dung điện văn để tìm những điện văn kế hoạch bay và thời tiết mà chúng ta cần dựa vào từ khóa FPL, Metar, Speci. Những điện văn có chứa những từ khóa đó sẽ được tiếp tục trích lọc nội dung và đưa vào cơ sở dữ liệu, những điện văn nào không chứa các từ khóa đó sẽ được loại bỏ.

3.3.2. mô-đun khai phá dữ liệu

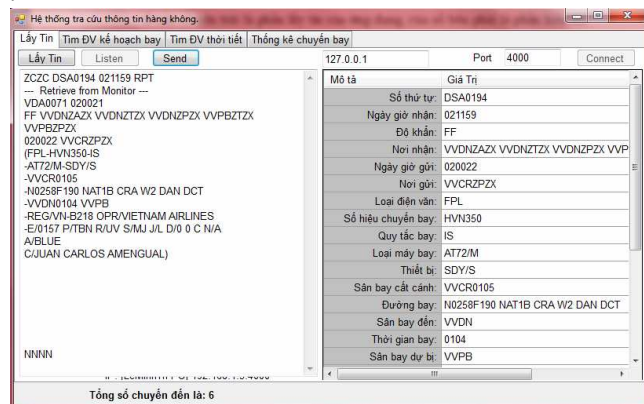
Modun khai phá dữ liệu sử dụng những luật truy vấn được trình bày ở trên để đưa ra kết quả khi người sử dụng yêu cầu. Nó được tích hợp vào giao diện người dùng. Dưới đây là những giao diện chính của chương trình:

- Giao diện lấy tin của chương trình



Hình 3.3: Giao diện chính của chương trình

Ở cửa sổ bên trái là phần lấy tin của ứng dụng, của sổ bên phải là phần hiển thị nội dung điện văn sao khi được trích lọc. Sau khi thiết lập đầy đủ các thông số cho chương trình, nó sẽ tự động chọn những tập tin điện văn cần phân tích và đưa ra kết quả ở màn hình bên phải như hình bên dưới:



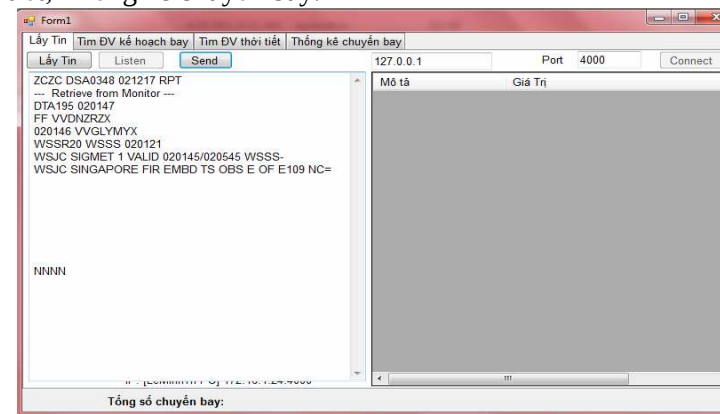
Hình 3.4: Giao diện trích lọc thông tin điện văn

3.4. Cài đặt và thử nghiệm

Chương trình có thể chạy trên các hệ điều hành Window. Vì chương trình sử dụng hệ quản trị CSDL SQL Server 2005, nên để chạy chương trình ta cần cài SQL Server 2005 và nạp lại cơ sở dữ liệu cho chương trình.

Sau khi đã nạp thành công cơ sở dữ liệu, ta chạy chương trình bằng tập tin KeHoachBay.exe ở đường dẫn sau: demo->KeHoachBay->Bin->Release. Khi chạy chương trình, giao diện tương tác với người dùng như hình dưới :

Giao diện gồm 4 tab: Lấy tin, Tìm DV kế hoạch bay, Tìm DV thời tiết, Thống kê chuyến bay.



Hình 3.8: Giao diện tương tác người dùng

Để kết nối với hệ thống AMSS, ta chọn Listen rồi chọn Connect. Chương trình sẽ tự động phân tích những điện văn kế hoạch bay và thời tiết nhận được từ hệ thống AMSS. Ngoài ra, người dùng cũng có thể chọn phần **Lấy tin** để chọn điện văn cần xem lại.

Hình dưới biểu thị điện văn kế hoạch bay đã được trích lọc bằng ứng dụng.

Hình 3.9: Giao diện trích lọc điện văn kế hoạch bay

Ngoài các chức năng trích lọc thông tin từ điện văn kế hoạch bay, thời tiết, chương trình còn có những chức năng khác như:

- Để tìm điện văn kế hoạch bay ta chọn tab Tìm ĐV kế hoạch bay, chọn ngày hoặc số thứ tự của điện văn để tìm

Hình 3.10: Giao diện tìm điện văn kế hoạch bay theo ngày

- Ngoài các chức năng tìm và trích lọc thông tin từ điện văn kế hoạch bay, điện văn thời tiết, ứng dụng còn có thể thống kê sản lượng bay bao gồm các chuyến bay đi, bay đến sân bay Đà Nẵng theo ngày hoặc tháng dựa vào các loại điện văn trên.

Dưới đây là giao diện thống kê sản lượng bay.

Hình 3.15: Giao diện thống kê sản lượng bay

3.5. Đánh giá kết quả

Trong quá trình thực hiện chương này, tôi đã áp dụng thành công các kỹ thuật trích lọc thông tin, khai phá dữ liệu vào ứng dụng. Dựa vào nhu cầu tra cứu thông tin, thống kê sản lượng bay của các kiểm soát viên không lưu đang công tác tại Đài chỉ huy cất hạ cánh Đà Nẵng để từ đó xây dựng “Hệ thống thu nhận và xử lý các thông tin điện văn” để hỗ trợ trong công tác kiểm tra, thống kê, lập báo cáo về các thông tin hàng không tại Trung tâm Quản lý bay Đà Nẵng.

Tuy nhiên, trong ứng dụng vẫn còn tồn tại những hạn chế là chưa thể đọc và phân tích từng cụm từ cụ thể có trong điện văn kế hoạch bay hay điện văn thời tiết để đưa ra chi tiết ý nghĩa của từng cụm từ đó.

KẾT LUẬN

Với mục tiêu giảm thiểu đến mức tối đa các sai sót không đáng có của kiểm soát viên không lưu nhằm phục vụ công tác điều hành bay an toàn tuyệt đối. Sau quá trình thực hiện đề tài, bước đầu tôi đã áp dụng kỹ thuật trích lọc thông tin, khai phá dữ liệu để xây dựng thành công “ Hệ thống thu nhận và xử lý thông tin điện văn”, và đã được đưa vào sử dụng. Ứng dụng đã phần nào giúp ích cho các kiểm soát viên không lưu trong việc khai thác thông tin trong mạng viễn thông cố định hàng không. Ngoài ra nó còn giúp cho việc thống kê sản lượng bay trở nên dễ dàng thuận lợi hơn.

Ứng dụng đã phần nào giúp ích cho các kiểm soát viên không lưu phần trong việc xem các điện văn kế hoạch bay hoặc thời tiết một cách dễ dàng hơn. Nó còn giúp cho việc thống kê sản lượng bay trở nên nhanh chóng và chính xác hơn so với việc thống kê bằng tay

Mặc dù đã xây dựng hệ thống thành công nhưng những gì tôi làm được vẫn chỉ là bước đầu và còn có nhiều giới hạn. Ứng dụng chưa đáp ứng được trong việc tra cứu tất cả các thông tin được chuyển tải trên mạng AFTN mà các kiểm soát viên không lưu cần. Tôi hy vọng rằng những nghiên cứu và sản phẩm của tôi sẽ được tiếp tục phát triển và thật sự có ích hơn nữa cho các kiểm soát viên không lưu đang công tác tại Đài chỉ huy cất hạ cánh sân bay Đà Nẵng.

Để ứng dụng thực sự hoàn chỉnh, có thể tiếp tục phát triển hệ thống nhằm tra cứu được tất cả các thông tin khác liên quan đến lĩnh vực hàng không. Dựa vào các thông tin có trong kế hoạch bay để đưa ra các cảnh báo cho kiểm soát viên không lưu như: hai máy bay có cùng mức bay gần nhau, hai máy bay cùng trên một đường bay nhưng tốc độ bay khác nhau, thông tin dự báo về những hiện tượng thời tiết bất thường sắp xảy ra.