BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG

TRƯƠNG TIẾN DƯỠNG

NGHIÊN CỨU ỨNG DỤNG PHÂN LỚP DỮ LIỆU TRONG QUẢN LÝ KHÁCH HÀNG TRÊN MẠNG

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

TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT

Đà Nẵng - Năm 2012

Công trình được hoàn thành tại
ĐẠI HỌC ĐÀ NẴNG

Người hướng dẫn khoa học: **PGS.TS. PHAN HUY KHÁNH**

Phản biện 1: **TS. NGUYỄN TRẦN QUỐC VINH**

Phản biện 1: **PGS.TS. LÊ MẠNH THẠNH**

Luận văn được bảo vệ tại Hội đồng chấm Luận văn tốt nghiệp thạc sĩ kỹ thuật họp tại Đại học Đà Nẵng vào ngày 03 tháng 03 năm 2012

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

Trong kinh doanh yếu tố khách hàng quyết định đến sự thành bại của doanh nghiệp, khi thông tin đang trở thành yếu tố quyết định trong kinh doanh thì vấn đề tìm ra các thông tin hữu ích trong các CSDL khổng lồ ngày càng trở thành mục tiêu quan trọng của các doanh nghiệp. Vì vậy một trong những giải pháp hữu hiệu nhất nhằm khắc phục các vấn đề nêu trên là tiến hành triển khai xây dựng một hệ thống khai phá dữ liệu (KPDL), khai thác quản lý nguồn khách hàng nói trên. Đó là một hệ thống được thiết kế giúp cho lãnh đạo doanh nghiệp nắm bắt được nguồn thông tin khách hàng hữu ích và các tri thức chiết xuất được từ CSDL trên sẽ là một nguồn tài liệu hỗ trợ cho lãnh đạo xây dựng chiến lược kinh doanh. Chính vì những lý do nêu trên, tôi quyết định chọn đề tài "**Nghiên cứu ứng dụng kỹ thuật phân lớp dữ liệu trong quản lý khách hàng trên mạng**".

2. Mục đích nghiên cứu

Nghiên cứu phương pháp phân lớp dữ liệu trong KPDL, các thuật toán liên quan đến quy nạp cây quyết định, tìm hiểu các ngôn ngữ mã lệnh siêu tìm kiếm Regurlation Expressions,...

3. Đối tượng và phạm vi nghiên cứu

❖ Đối tượng nghiên cứu

Tìm hiểu các website TMĐT bán hàng trực tuyến với số lượng truy cập và giao dịch lớn phong phú, đa dạng có thể gây khó khăn trong công tác quản lý nguồn khách hàng.

❖ **Phạm vi nghiên cứu**

Ứng dụng các thuật toán của kỹ thuật phân lớp dữ liệu để xây dựng phục vụ công việc khai thác nguồn khách hàng.

4. Phương pháp nghiên cứu

Dựa trên thực trạng các website TMĐT hiện có để xây dựng ứng dụng quản lý khách hàng.

5. Ý nghĩa khoa học và thực tiễn

❖ **Ý nghĩa khoa học**

Đề xuất giải pháp ứng dụng kỹ thuật phân lớp dữ liệu vào trong khai thác quản lý nguồn khách hàng trên mạng.

❖ **Ý nghĩa thực tiễn**

Sản phẩm là hệ thống hỗ trợ đắc lực, kịp thời và có độ hiệu quả cao cho các doanh nghiệp thu thập được thông tin và đưa ra các chính sách phù hợp trong hoạt động kinh doanh của đơn vị.

6. Cấu trúc của luận văn

Nội dụng chính của luận văn này được chia thành ba chương với nội dung như sau:

Chương 1. Tổng quan về khai phá dữ liệu

Chương 2. Giải pháp phân lớp dữ liệu bằng kỹ thuật quy nạp cây quyết định.

Chương 3. Xây dựng hệ thống và thử nghiệm.

CHƯƠNG 1. TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU

1.1. Giới thiệu về khai phá dữ liệu

1.1.1. Khái niệm về khai phá dữ liệu

Khai phá dữ liệu (Data Mining) là quá trình khảo sát và phân tích một khối lượng lớn các dữ liệu được lưu trữ trong các CSDL, kho dữ liệu,…để từ đó trích xuất ra các thông tin quan trọng, có giá trị tiềm ẩn bên trong [6][10].

1.1.2. Những lợi thế và thách thức của khai phá dữ liệu

1.1.2.1. Lợi thế

KPDL là một lĩnh vực liên quan tới nhiều ngành học khác như: hệ cơ sở dữ liệu, thống kê xác suất, trực quan hoá… Thêm vào đó KPDL còn có thể áp dụng các kỹ thuật như mạng nơron, lý thuyết tập thô, tập mờ, biểu diễn tri thức…

1.1.2.2. Thách thức

Những hạn chế của các thuật toán: Hầu hết các thuật toán đều khá là tổng quát, nó sinh ra nhiều luật. Mặc dù các luật sinh ra đa số đều hữu ích nhưng ta vẫn phải đo độ đáng quan tâm của các mẫu nên vẫn cần sự can thiệp của các chuyên gia nghiệp vụ.

1.1.3. Những nhu cầu về khai phá dữ liệu trong kinh doanh

Phân loại khách hàng để từ đó phân định thị trường, thị phần. Tăng sức cạnh tranh, làm thế nào để giữ được khách hàng cũ và thu hút được thêm nhiều khách hàng mới. Phân tích rủi ro trước khi ra các quyết định quan trọng trong chiến lược hoạt động sản xuất kinh doanh. Ra các báo cáo giàu thông tin …

Tất cả các nhu cầu xã hội trên đòi hỏi cần phải có một phương thức, công cụ nào đó hỗ trợ bên cạnh các chuyên gia kinh tế. Và KPDL là một chìa khoá hỗ trợ giải quyết vấn đề nêu trên.

1.1.4. Khai phá dữ liệu trong một số lĩnh vực quan trọng khác

1.2. Các phương pháp chính trong khai phá dữ liệu

1.2.1. Phân loại

Phân loại là tổ chức dữ liệu trong các lớp cho trước, còn được gọi là học có quan sát. Phân loại sử dụng các nhãn lớp cho trước để sắp xếp các đối tượng. Trong đó có một tập huấn luyện gồm các đối tượng đã được kết hợp với các nhãn đã biết. Một số thuật toán dùng trong bài toán phân loại như: cây quyết định, mạng nơron, Naive Bayes.

1.2.2. Phân cụm

Phân cụm là kỹ thuật KPDL tương tự như phân loại dữ liệu. Tuy nhiên, sự phân nhóm dữ liệu là quá trình học không được giám sát.

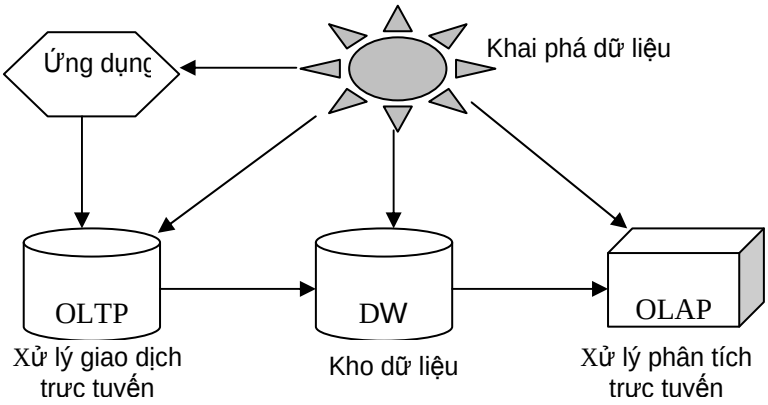
1.2.3. Luật kết hợp

1.2.4. Hồi quy

1.2.5. Phân tích chuỗi

1.3. Các bước xây dựng một giải pháp về khai phá dữ liệu

1.3.1. Mô hình luồng dữ liệu



Hình 1.1 Mô hình luồng dữ liệu

1.3.2. Vòng đời của một hệ thống khai phá dữ liệu

Bước 1: Xác định mục tiêu bài toán.

Bước 2: Thu thập dữ liệu.

Bước 3: Làm sạch dữ liệu và chuyển đổi dữ liệu.

Bước 4: Xây dựng mô hình.

Bước 5: Đánh giá mô hình hay đánh giá mẫu.

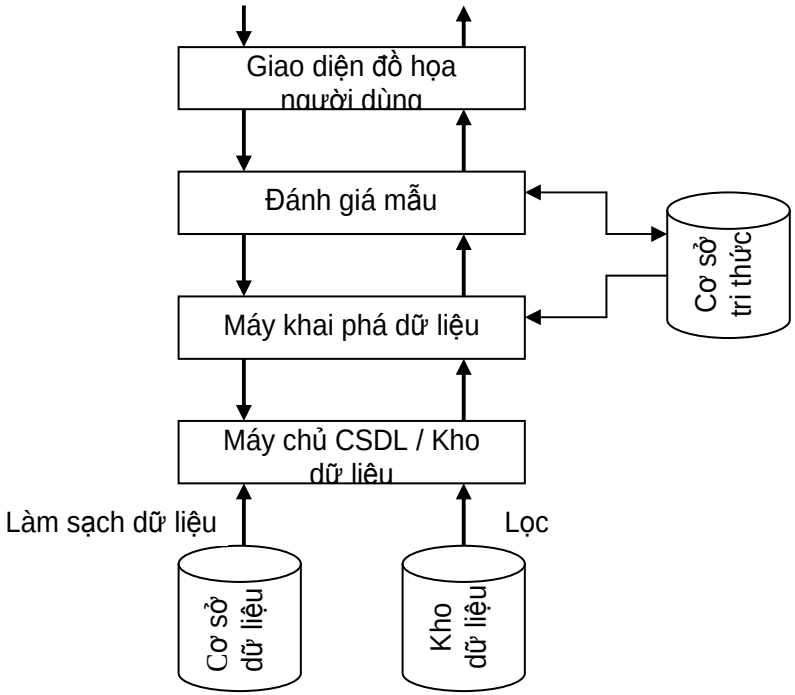
Bước 6: Báo cáo.

Bước 7: Dự đoán.

Bước 8: Tích hợp vào ứng dụng.

Bước 9: Quản lý mô hình.

1.3.3. Kiến trúc của một hệ thống khai phá dữ liệu điển hình



Hình 1.2 Kiến trúc của một hệ thống khai phá dữ liệu điển hình

1.3.3.1. Phương pháp đánh giá độ chính xác của mô hình phân lớp

Trong phương pháp holdout, dữ liệu dưa ra được phân chia ngẫu nhiên thành 2 phần là: tập dữ liệu đào tạo và tập dữ liệu kiểm tra. Thông thường 2/3 dữ liệu cấp cho tập dữ liệu đào tạo, phần còn lại cho tập dữ liệu kiểm tra.

Trong phương pháp k-fold cross validation tập dữ liệu ban đầu được chia ngẫu nhiên thành k tập con (fold) có kích thước xấp xỉ nhau S1, S2, …, Sk. Quá trình học và test được thực hiện k lần. Tại lần lặp thứ i, Si là tập dữ liệu kiểm tra, các tập còn lại hợp thành tập dữ liệu đào tạo.

1.3.3.2. Vấn đề quản lý KH trên mạng và sự liên quan đến DM

KPDL giúp lãnh đạo các doanh nghiệp xác định được các KH mục tiêu, phân loại để từ đó hỗ trợ các doanh nghiệp có một chiến lược quảng cáo, tiếp thị tốt. Tổng hợp các tri thức này lãnh đạo có thể lên kế hoạch hoạt động, sản xuất, kinh doanh một cách thuận tiện hơn nhằm giảm bớt thời gian thống kê, tìm hiểu thị hiếu KH. Chẳng hạn chiến lược quảng cáo cho các đối tượng KH khác nhau…

1.3.4. So sánh giữa các kỹ thuật khai phá dữ liệu

1.3.4.1. Kỹ thuật khai phá dữ liệu mô tả

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 lớp là tương đồng còn các đối tượng thuộc các cụm khác nhau sẽ không tương đồng.

1.3.4.2. Kỹ thuật khai phá dữ liệu dự đoán

Mục tiêu của phương pháp phân lớp dữ liệu là dự đoán nhãn lớp cho các mẫu dữ liệu. Không giống như phân cụm dữ liệu, phân lớp dữ liệu là học bằng ví dụ, trong khi phân cụm dữ liệu có thể coi là một cách học bằng quan sát.

CHƯƠNG 2. GIẢI PHÁP PHÂN LỚP DỮ LIỆU TRONG QUẢN LÝ KHÁCH HÀNG TRÊN MẠNG

2.1. Bài toán phân lớp dữ liệu

2.1.1. Giới thiệu

Phân lớp là một tiến trình xử lý nhằm xếp các mẫu dữ liệu hay các đối tượng vào một trong các lớp đã được định nghĩa trước.

2.1.2. Các bước chính để giải quyết bài toán phân lớp

Phân lớp dữ liệu gồm hai bước xử lý chính:

Bước 1: Học, mục đích của bước này là xây dựng một mô hình xác định một tập các lớp dữ liệu.

Bước 2 : Kiểm tra và đánh giá, bước này sử dụng mô hình phân lớp đã được xây dựng ở **bước 1** vào việc phân lớp.

2.1.3. Các cơ sở dữ liệu phục vụ cho phân lớp dữ liệu

2.1.3.1. Cơ sở dữ liệu giao tác

CSDL giao tác là tập hợp những bản ghi giao dịch, trong đa số các trường hợp chúng là những bản ghi các dữ liệu hoạt động của doanh nghiệp, tổ chức.

2.1.3.2. Cơ sở dữ liệu đa phương tiện

KPDL web thông thường được chia thành ba phạm trù chính: Khai phá cách dùng web, khai phá cấu trúc web và khai phá nội dung web.

2.1.3.3. Cơ sở dữ liệu Hypertext

HyperText là loại dữ liệu phổ biến hiện nay, và cũng là loại dữ liệu có nhu cầu tìm kiếm và phân lớp rất lớn.

2.2. Phân lớp bằng phương pháp quy nạp cây quyết định

2.2.1. Khái niệm cây quyết định

Cây quyết định là một flow-chart giống cấu trúc cây, nút bên trong biểu thị một kiểm tra trên một thuộc tính, nhánh biểu diễn đầu ra của kiểm tra, nút lá biểu diễn nhãn lớp.

2.2.2. Đánh giá cây quyết định trong lĩnh vực khai phá dữ liệu

2.2.2.1. Sức mạnh của cây quyết định

Khả năng sinh ra các quy tắc hiểu được, khả năng thực thi trong những lĩnh vực hướng quy tắc, dễ dàng tính toán trong khi phân lớp,…

2.2.2.2. Điểm yếu của cây quyết định

Dễ xãy ra lỗi khi có quá nhiều lớp, Chi phí tính toán đắt để đào tạo

2.2.3. Xây dựng cây quyết định

Quá trình xây dựng cây quyết định gồm hai giai đoạn:

Giai đoạn thứ nhất phát triển cây quyết định bắt đầu từ gốc, đến từng nhánh và phát triển quy nạp theo cách thức chia để trị cho tới khi đạt được cây quyết định với tất cả các lá được gán nhãn lớp.

Giai đoạn thứ hai cắt, tỉa bớt các cành nhánh trên cây quyết định.

2.2.4. Thuật toán quy nạp cây quyết định

Input : những mẫu học được biểu thị bằng những thuộc tính riêng biệt, một tập các thuộc tính đặc trưng và danh sách các thuộc tính.

Output : một cây quyết định.

1) Khởi tạo một node N;

2) **if** tất cả các mẫu đều thuộc vào cùng một lớp C **then**

3) **return** node N, được xem là 1 node lá và đặt tên là lớp C;

4) **if** danh sách thuộc tính là rỗng **then**

5) **return** node N, là một node lá được đặt tên lớp là lớp chung nhất trong các mẫu ;

6) Chọn thuộc tính thử, là một thuộc tính trong danh sách thuộc tính mà có độ đo cao nhất;

7) Đặt tên node N với tên của thuộc tính thử;

8) Với mỗi giá trị a_i đã biết của thuộc tính thử

9) Tạo ra 1 nhánh từ node N cho điều kiện thuộc tính thử = a_i;

10) Đặt S_i là một tập các mẫu lấy trong các mẫu ban đầu với thuộc tính thử = a_i;

11) **if** S_i là rỗng **then**

12) Tạo ra một node lá trên cây quyết định, được đặt tên lớp là lớp chung nhất của hầu hết các mẫu ;

13) **else** thêm vào một node là cây kết quả của thuật toán tạo cây với tham số đầu vào

2.2.5. Rút trích luật phân lớp từ cây quyết định

Tri thức trên cây quyết định có thể được rút trích và biểu diễn thành một dạng luật phân lớp **IF - THEN**. Khi đã xây dựng được cây quyết định, ta có thể dễ dàng chuyển cây quyết định này thành một tập các luật phân lớp tương đương, một luật tương đương với một đường đi từ gốc đến node lá.

2.3. Tìm hiểu các công nghệ ứng dụng

2.3.1. Giới thiệu thuật toán cây quyết định Microsoft

Cây quyết định của Microsoft là thuật toán cây quyết định lai ghép được phát triển bởi nhóm nghiên cứu của Microsoft. Nó hỗ trợ cả hai nhiệm vụ phân loại và hồi quy.

2.3.2. Data Mining eXtensions

DMX - Data Mining eXtensions là một ngôn ngữ truy vấn khai phá dữ liệu được định nghĩa trong OLE DB dành cho khai phá dữ liệu, được kế thừa hầu hết các khái niệm quan hệ và cấu trúc của nó dựa trên ngôn ngữ truy vấn SQL.

2.3.3. Giới thiệu về Regular Expressions

Regular Expression (regex) là một chuỗi miêu tả một bộ các chuỗi khác, tập hợp các phép xử lý văn bản tìm kiếm, so khớp, cắt ghép,… theo những quy tắc cú pháp nhất định. Regex làm việc dựa trên những mẫu văn bản theo các quy tắc quy định sẵn trước.

2.3.4. Giới thiệu về lập trình tương tác Windows services

Windows services [12] cung cấp phương tiện cho application logic chạy liên tục trên máy tính, thông thường là việc cung cấp điều khiển thiết bị hoặc các dịch vụ hệ điều hành. Windows services là một ứng dụng chạy trên máy chủ hoặc máy trạm và cung cấp những chức năng mà sự diễn tiến của nó không cần sự tương tác trực tiếp của người dùng.

2.4. Khảo sát hiện trạng

2.4.1. Phân tích quy trình, hoạt động khách hàng TMĐT

Để thực hiện đăng ký thành viên hoặc đăng tin, giao dịch mua bán trên website TMĐT, khách hàng phải đăng ký xác nhận các thông tin của KH mà dường như các website thương mại điện tử đều yêu cầu đó là: email, tên khách hàng, điện thoại, địa chỉ,…

❖ **Các hình thức giao dịch trong thương mại điện tử.**

TMĐT được phân chia thành một số loại như B2B, B2C, C2C dựa trên thành phần tham gia hoạt động thương mại.

❖ **Đặc điểm của thương mại điện tử**

Tính cá nhân hoá, đáp ứng tức thời, giá cả linh hoạt, các "điệp viên thông minh"

2.4.2. Thực trạng khách hàng thương mại điện tử

Kết quả khảo sát thống kê khách hàng giao dịch từ website TMĐT http://www.raovat30s.com

Bảng 2.1 Bảng thống kê KH giao dịch TMĐT tại một thời điểm

Tên KH	Địa chỉ	Điện thoại	Email	Nhu cầu	Mô tả	Ngày cập nhật
Hải Nam	TPHCM	0972105943	tinh.hn@gmail.com	mua	máy tính	14/09/2011
Ngân	Hà Nội	0974386284	thaong@yahoo.com	mua	máy tính	14/09/2011
Tiến	Hà Nội	09761383 53	tien@gmail.com	bán	Laptop	14/09/2011
Tiến Bình	Đà Nẵng	0983552518	tnbinh@gmail.com	mua	Desktop	14/09/2011
Hà	Đà Nẵng	0982734515	hant@yahoo.com	mua	Laptop	14/09/2011
….	….	….	….	….	….	….

Bảng thống kê kết quả khảo sát số lượng KH quan tâm đến những sản phẩm, dịch vụ trong một thời điểm nhất định.

Bảng 2.2 Bảng thống kê lượng KH quan tâm đến sản phẩm

Tên KH	Địa chỉ	Điện thoại	Email	Nhu cầu	Mô tả	Ngày thống kê	SL xem
Hải Nam	TPHCM	0972105943	tinh.hn@gmail.com	mua	máy tính	14/09/2011	1053
Thảo Ngân	Hà Nội	0974386284	thaong@yahoo.com	mua	máy tính	14/09/2011	1153
Tiến	Hà Nội	097613 3 53	tien@gmail.com	bán	laptop	14/09/2011	953
Tiến Bình	Đà Nẵng	0983552518	tienbinh@gmail.com	mua	desktop	14/09/2011	753
Hà	Đà Nẵng	0982734515	hant@yahoo.com	mua	laptop	14/09/2011	1250
…	…	…	…	…	…	…	…

Hàng ngày có rất nhiều thông tin được cập nhật trên các website TMĐT này bao gồm cả thư từ, các tệp văn bản, các cơ sở dữ liệu, các bản tính, các hình ảnh, các biểu mẫu,... Nên rất khó khăn

cho doanh nghiệp khi muốn tìm kiếm, xử lý khai thác nguồn thông tin của khách hàng, mất rất nhiều thời gian và dễ bỏ sót.

2.4.3. Nhu cầu quản lý khách hàng

Trên thực tế hiện có rất nhiều website TMĐT đang hoạt động với số lượng giao dịch của KH rất lớn. Tuy nhiên doanh nghiệp chưa có giải pháp để quản lý nguồn khách hàng này sao cho có hiệu quả. Việc ứng dụng các kỹ thuật KPDL nhằm tìm kiếm, khai thác tự động sẽ giúp cho các doanh nghiệp luôn có nguồn KH mua bán dồi dào mà không cần phải bỏ nhiều công sức và nguồn nhân lực.

2.4.4. Giải pháp xây dựng và kịch bản hệ thống

❖ **Giải pháp xây dựng hệ thống**

Xây dựng chương trình có bộ lập lịch để tự động chạy trên máy tính như một services của hệ điều hành windows.

❖ **Kịch bản sử dụng hệ thống**

Tiến hành triển khai cho máy học với tập dữ liệu huấn luyện được xây dựng bằng các mã lệnh và trích lọc từ nguồn dữ liệu web. Sau quá trình học, so khớp được hệ thống sẽ trả về kết quả dưới dạng bảng với các trường tương ứng. Phần thứ nhất liên quan đến việc thực hiện giải thuật học mẫu. Phần thứ hai chỉ đơn giản là phần áp dụng của các dữ liệu đã tìm ra.

2.4.5. Triển khai ứng dụng học quy nạp cây quyết định

2.4.5.1. Xây dựng các mẫu học

Ứng dụng các mã lệnh siêu tìm kiếm để xây dựng các mẫu trong đề tài, như xây dựng một số mẫu sau:

2.4.5.2. Thuật toán quy nạp cây quyết định dựa vào dữ liệu học

Input : những mẫu học được biểu thị bằng những thuộc tính riêng biệt, một tập các thuộc tính đặc trưng.

Output : một cây quyết định.

CHƯƠNG 3 XÂY DỰNG HỆ THỐNG VÀ THỬ NGHIỆM

3.1. Giới thiệu bài toán

3.1.1. Tính chất

Thông qua website TMĐT http://www.raovat30s.com/, Phân tích các Weblog để khám phá ra các mẫu truy cập của người dùng trong trang Web.

3.1.2. Mục tiêu

Dựa vào dữ liệu giao dịch thu thập được, hệ thống sẽ khai thác, trích rút được các thông tin cần thiết của KH.

3.1.3. Yêu cầu

Đầu vào: Cập nhật danh sách các website TMĐT, đọc nội dung html của các URL.

Đầu ra: Bộ dữ liệu phân lớp, chứa đựng thông tin email, điện thoại, tên, địa chỉ và nhu cầu của khách hàng,…

3.2. Giải pháp kỹ thuật

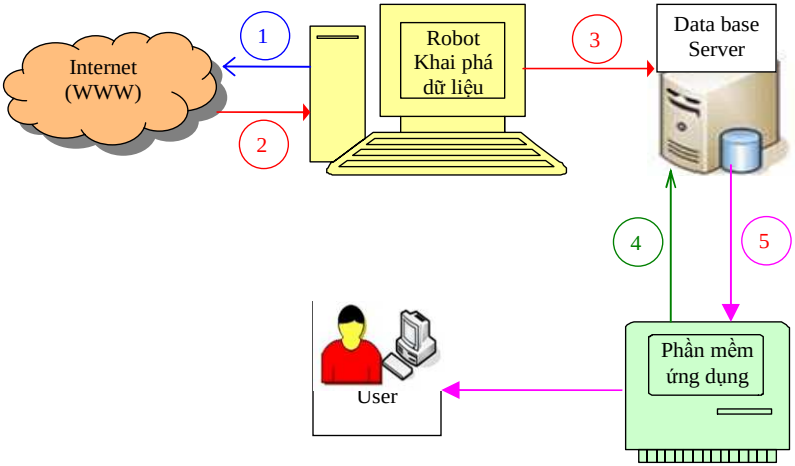
3.2.1. Tổng quan

Các trang TMĐT khi diễn ra các hoạt động giao dịch rao vặt, mua, bán hàng, đăng ký thành viên,…thường thể hiện các thông tin có tính cấu trúc như: email, điện thoại, tên KH, nhu cầu, địa chỉ,…

Regular expressions của microsoft cung cấp giải pháp tìm kiếm theo cấu trúc rất mạnh và hiệu quả. Kỹ thuật này hỗ trợ mạnh mẽ cho việc xử lý chuỗi như tìm kiếm, so khớp cắt ghép…

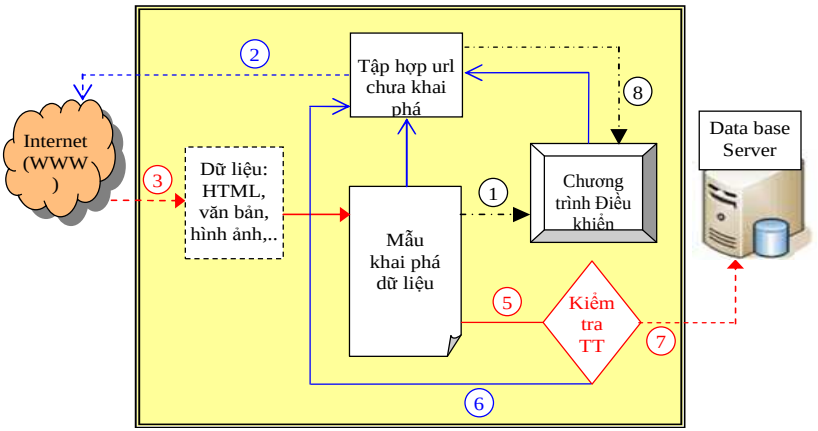
3.2.2. Mô hình giải pháp

3.2.2.1. Mô hình giải pháp tổng thể



Hình 3.1 Mô hình giải pháp tổng thể

3.2.2.2. Mô hình giải pháp Robot khai phá dữ liệu



Hình 3.2 Mô hình giải pháp Robot khai phá dữ liệu

Trong đó:

(1): Học mẫu KPDL. Các mẫu này được xây dựng theo yêu.

(2): Danh sách các url sẽ KPDL. DS thường xuyên được cập nhật.

(3): Dữ liệu trả về sau khi khai phá một url có cấu trúc.

(4): Nếu dữ liệu khai phá được từ một url không phù hợp với các mẫu thì quay lại bước **(2)**

(5): Url khai phá phù hợp với một trong số các mẫu.

(6): Nếu URL này đã tồn hoặc các thông tin khai phá được từ url này đã tồn tại trong CSDL thì quay lại bước **(2)**.

(7): Nếu kết quả khai phá từ URL phù hợp với các mẫu và chưa có trong CSDL thì đưa vào CSDL.

(8): CT Điều khiển kết thúc phiên làm việc khi tất cả các website đều được duyệt qua. Ngược lại thì tiếp tục bước **(2)**.

3.2.2.3. Mô hình giải pháp phần mềm ứng dụng

Hình 3.3 Mô hình giải pháp phần mềm ứng dụng

3.2.3. Các chức năng chính của hệ thống

❖ **Robot khai phá dữ liệu:**

Xây dựng mẫu:

Xây dựng và học mẫu các nhóm url cần khai phá.

Download một word về máy tính:

Chức năng dưới dạng mã lệnh cho phép download dữ liệu của một url về máy tính để phân tích mẩu.

Chuyển dữ liệu sang UTF-8

Chức năng này dưới dạng mã lệnh dùng để chuyển các dữ liệu dưới dạng mã ký tự sang Unicode UTF-8.

Lập danh sách url từ nhóm url:

Chức năng này dưới dạng mã lệnh dùng để phân tích chi tiết các url từ nhóm url và đưa vào danh sách để khai phá dữ liệu.

Kiểm tra sự tồn tại ở CSDL:

Chức năng này dưới dạng mã lệnh dùng để kiểm tra url đã được khai phá chưa.

Khai phá:

Chức năng này dưới dạng mã lệnh dùng để KPDL theo mẫu đã lập.

Đưa dữ liệu đã khai phá vào CSDL

Chức năng này dưới dạng mã lệnh dùng để chèn các dữ liệu đã khai phá được vào CSDL.

Đặt lịch khai phá dữ liệu:

Chức năng dùng đặt lịch để tự động KPDL theo thời gian lập trước.

Thường trú Robot khai phá như Windows service:

Chức năng cho phép cài đặt Robot KPDL chạy thường trú như một Windows service.

❖ **Phần mềm ứng dụng khai thác dữ liệu**

Phân loại thông tin:

Chức năng này cho phép phân loại các thông tin khai phá được theo các tiêu chí:

Tìm kiếm thông tin:

Tìm kiếm thông tin khai phá được qua các trường dữ liệu

Đánh giá khách hàng tiềm năng:

Đánh giá tiềm năng KH dựa vào thông tin khai phá được qua các trường dữ liệu

3.3. Xây dựng mô hình phân lớp dữ liệu trực quan

3.3.1. Thiết kế CSDL vật lý với MSSQL Server

Các bảng dữ liệu sử dụng trong chương trình

1) Thongtinkhaipha

Mục đích: Lưu các thông tin khai phá được từ website TMĐT

Bảng 3.1 Bảng dữ liệu thông tin khai phá

Trường	Kiểu dữ liệu	NULL	Mô tả
Matin	int	Không	Trường khóa
Tieude	nvarchar(100)	Có	Tiêu đề của url khai phá
Email	varchar(50)	Có	Email người đăng tin
Dienthoai	nvarchar(50)	Có	Điện thoại người đăng tin
Hoten	nvarchar(50)	Có	Họ tên người đăng tin
Diachi	nvarchar(50)	Có	Địa chỉ người đăng tin
Tinhthanh	nvarchar(20)	Có	Tỉnh thành cần mua bán,….
Nhucau	nvarchar(20)	Có	Nhu cầu mua, bán,…
Gia	nvarchar(50)	Có	Giá cả
Url	varchar(160)	Có	Url gốc khai phá về
Mota	nvarchar(160)	Có	Mô tả nội dung của url
Ngay	datetime	Có	Ngày khai phá url
Ngaylammoi	datetime	Có	Ngày cập nhật, làm mới url do người đăng tin tự cập nhật
Cophi	int	Có	=1 nếu đây là tin VIP (có phí) và =0 thì ngược lại

2) Urltuchoi

Mục đích: Lưu trữ các url đã duyệt qua nhưng không thỏa mãn

Bảng 3.2 Bảng dữ liệu URL từ chối

Trường	Kiểu dữ liệu	NULL	Mô tả
Matin	int	Không	Trường khóa
Link	nvarchar(160)	Có	Link url gốc không thỏa mãn

3) URLdaduyet

Mục đích: lưu các url đã duyệt qua và thỏa mãn các tập mẫu.

Bảng 3.3 Bảng dữ liệu URL đã duyệt

Trường	Kiểu dữ liệu	NULL	Mô tả
Matin	int	Không	Trường khóa
Link	nvarchar(160)	Có	Link url gốc không thỏa mãn
Tieude	nvarchar(100)	Có	Tiêu đề của url

4) Taphopmau

Mục đích: Lưu các tên và giá trị các mẫu

Bảng 3.4 Bảng dữ liệu tập hợp mẫu

Tên cột	Kiểu dữ liệu	NULL	Ghi chú
Khoa	Int(4)	Không	Trường khóa của bảng
Tenmau	NVarchar(50)	Có	Tên mẫu
Hammau	Nvarchar(2000)	Có	Hàm của mẫu

4) Lichkhaipha

Mục đích: Lưu thời gian các lịch để tự động khai phá thông tin

Bảng 3.5 Bảng dữ liệu lịch khai phá

Trường	Kiểu dữ liệu	NULL	Mô tả
Idkey	int	Không	Trường khóa
Tenlich	nvarchar(50)	Có	Tên lịch
Ngaybdhl	datetime	Có	Ngày lịch bắt đầu hiệu lực
Ngaykthl	datetime	Có	Ngày lịch kết thúc hiệu lực
Loop	int	Có	Lịch tự động KPDL
Looptype	nvarchar(10)	Có	Lặp lại quá trình khai phá:

6) Taikhoan:

Mục đích: Tài khoản sử dụng chương trình

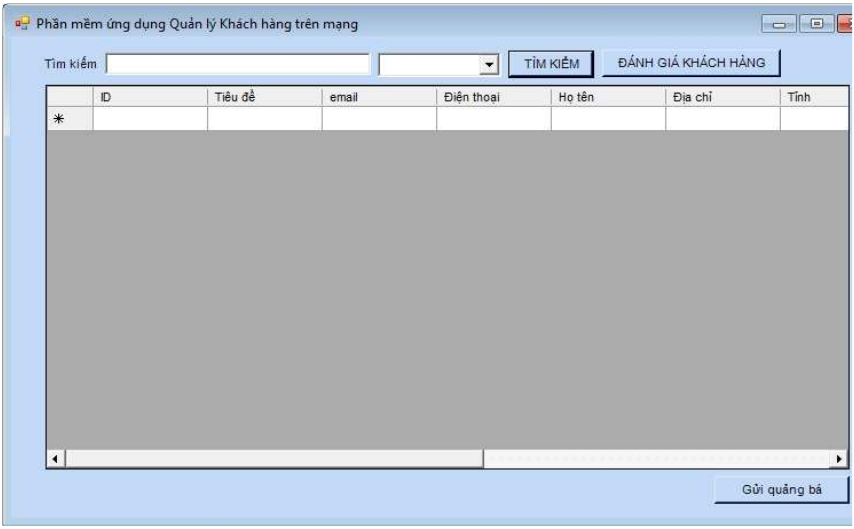
3.3.2. Giao diện chương trình

3.3.2.1. Giao diện chương trình Robot khai phá dữ liệu

Giao diện chính của hệ thống khai phá thông tin

Hình 3.4 Giao diện chương trình Robot khai phá dữ liệu

3.3.2.2. Giao diện chương trình phần mềm ứng dụng



Hình 3.5 Giao diện chương trình phần mềm ứng dụng

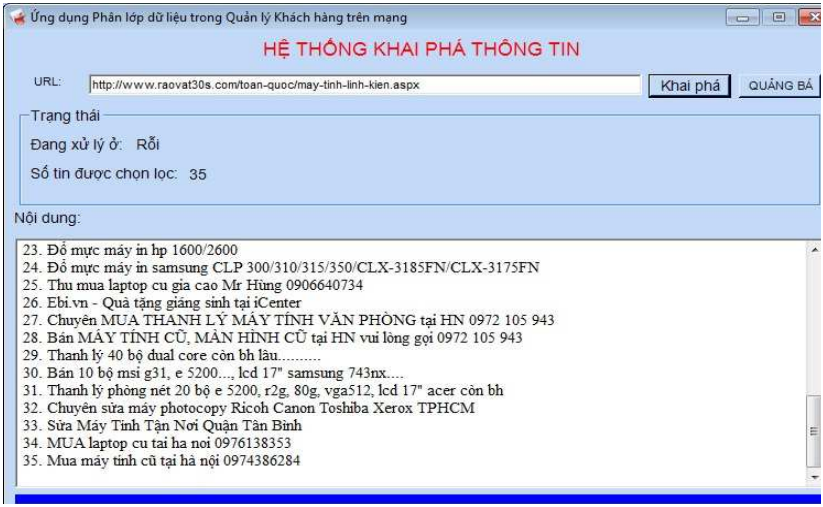
3.4. Kết quả thử nghiệm

3.4.1. Khai phá dữ liệu từ các trang thương mại điện tử

Nhập các URL thương mại vào hệ thống: Đây là chức năng cho máy học với tập dữ liệu mẫu có thể được chọn.

Hình 3.6 Chọn URL để khai phá

Hiển thị dữ liệu được huấn luyện sau khi học:



Hình 3.7 Hiển thị dữ liệu được huấn luyện

Hệ thống hoạt động theo các phiên làm việc đã được lập lịch cho trước, thông tin khai phá là những mẫu mới được cập nhật chưa tồn tại trên hệ thống.

3.4.2. Xử lý các dữ liệu thu được từ khai phá

❖ **Tìm kiếm**

Hiển thị kết quả sau khi đã nhập các thông tin có liên quan: Phần này sẽ hiển thị kết quả tương ứng với dữ liệu được người sử dụng nhập vào.

❖ **Phân loại theo nhu cầu**

Phần này sẽ hiển thị kết quả tương ứng với dữ liệu được người sử dụng chọn.

❖ **Đánh giá Khách hàng tiềm năng**

Đây là phần giúp cho người dùng, lãnh đạo có kết quả đánh giá, phân tích nguồn KH tiềm năng.

❖ **Gởi email quảng bá khách hàng**

Các nguồn thông tin mà hệ thống trích rút được như email, điện thoại sẽ giúp cho doanh nghiệp tiếp cận kịp thời và tư vấn với khách hàng thông qua hệ thống gởi mail, tin nhắn,...

3.5. Đánh giá kết quả chương trình

Hệ thống khai phá hoạt động chạy tự động theo bộ lập lịch định sẵn. Hệ thống đã xử lý và trích rút được những thông tin KH tương đối chính xác và phù hợp với yêu cầu của đề tài.

Kết quả về các thông tin của KH sau khi chương trình khai phá từ trên trang TMĐT được đưa vào CSDL theo các trường, bảng dữ liệu. Rất thuận tiện cho doanh nghiệp triển khai và khai thác các ứng dụng như theo dõi, tìm kiếm, phân loại khách hàng.

KẾT LUẬN

❖ Kết quả đạt được

Nội dung nghiên cứu trong đề tài, tác giả đã đưa ra một giải pháp từ việc phân loại dữ liệu trên các phiên giao dịch, trên TMĐT ,... rồi tiến hành khai thác xử lý chúng để chiết xuất ra các tri thức cần thiết. Các tri thức này lại được tối ưu hoá và đem vào sử dụng một cách hiệu quả trên các phiên giao dịch trong những lần tiếp theo.

Đề tài đã đi sâu vào tính ứng dụng, đưa ra cách thức xử lý thi hành các tri thức được chiết xuất một cách hiệu quả.

Về mặt lý thuyết, đã nêu được giải pháp ứng dụng kỹ thuật phân lớp dữ liệu vào bài toán quản lý khách hàng trên mạng.

Về mặt thực tiễn, có thể khẳng định đề tài đã đáp ứng được các mục tiêu đề ra, hệ thống đã khai phá được các thông tin khách hàng giao dịch trên mạng hữu ích và cần thiết, nhằm hỗ trợ doanh nghiệp có được nguồn khách hàng dồi dào và nắm bắt kịp thời các cơ hội kinh doanh. Đồng thời thông tin thu được sẽ là nguồn dữ liệu cơ sở để cho doanh nghiệp phân tích và định hướng chiến lược trong hoạt động kinh doanh của đơn vị.

❖ Hướng phát triển

Trong khuôn khổ của đề tài, chỉ tiến hành thực nghiệm trên website TMĐT http://www.raovat30s.com với các mẫu dữ liệu về máy tính và linh kiện. Tuy nhiên rất dễ dàng để phát triển trên các trang TMĐT khác và thêm các mẫu về: Điện thoại, điện máy điện tử, y tế, bất động sản,…. Hầu hết tất cả các mặt hàng kinh doanh hiện nay.

Nghiên cứu thiên về tính ứng dụng trong CSDL giao dịch, song việc nghiên cứu sẽ được tiếp tục phát triển trên các cơ sở dữ liệu khác nhằm mục đích tìm ra một quy luật ứng dụng cho các tri thức đã chiết xuất.