ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

HOÀNG MINH THỦY

NGHIÊN CỨU CÁC PHƯƠNG PHÁP TRÍCH CHỌN THÔNG TIN VÀ ỨNG DỤNG TRÍCH CHỌN THÔNG TIN DU LỊCH TRONG VĂN BẢN TIẾNG VIỆT

Chuyên ngành: KHOA HỌC MÁY TÍNH
Mã số: 60 48 01 01

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

NGƯỜI HƯỚNG DẪN KHOA HỌC

GS. VŨ ĐỨC THI

Thái Nguyên – 2015

LỜI CAM ĐOAN

Tác giả Hoàng Minh Thủy xin cam kết rằng nội dung của Luận văn này chưa được nộp cho bất kỳ một chương trình cấp bằng cao học nào cũng như bất kỳ một chương trình đào tạo cấp bằng nào khác.

Ngoài ra, tác giả cũng xin cam kết Luận văn thạc sĩ này là nỗ lực riêng của cá nhân tác giả. Các kết quả, phân tích, kết luận trong Luận văn thạc sĩ này (ngoài các phần được trích dẫn) đều là kết quả làm việc của cá nhân tác giả.

Thái Nguyên, ngày 10 tháng 11 năm 2015

Tác Giả

Hoàng Minh Thủy

LỜI CẢM ƠN

Lời đầu tiên em xin gửi lời cảm ơn chân thành đến Các quý thầy cô giáo, Tổ chuyên môn Trường Đại học Công nghệ thông tin và Truyền thông - Đại học Thái Nguyên đã tận tình giảng dạy, truyền đạt những kiến thức, kinh nghiệm quý báu trong suốt thời gian em theo học tại trường. Các kiến thức, kinh nghiệm quý báu của các Quý thầy cô giáo không chỉ giúp cá nhân em hoàn thiện hệ thống kiến thức trong học tập mà còn giúp em ứng dụng các kiến thức đó trong công tác hiện tại tại đơn vị.

Đặc biệt, em xin chân thành cảm ơn thầy giáo GS. Vũ Đức Thi đã rất nhiệt tình và tâm huyết trong việc định hướng và giúp đỡ em hoàn thành luận văn này.

Ngoài ra, em cũng xin chân thành cảm ơn Ban lãnh đạo và cán bộ viên chức Trường Đại học Lâm nghiệp đã tạo điều kiện cung cấp những ý kiến quý báu và những kiến thức thực tiễn cho em thực hiện luận văn tốt nghiệp này.

Em cũng xin được bày tỏ tình cảm với gia đình, đồng nghiệp, bạn bè đã tạo điều kiện để cá nhân em có thể dành thời gian cho khóa học. Xin chân thành cảm ơn những người bạn lớp cao học CK13, trong 2 năm qua đã luôn luôn động viên, khích lệ và hỗ trợ em trong quá trình học tập.

Trong quá trình thực hiện Luận văn mặc dù đã cố gắng hết mình, song chắc chắn luận văn của em vẫn còn nhiều thiếu sót. Em rất mong nhận được sự chỉ bảo vào đóng góp tận tình của các thầy cô để luận văn của em được hoàn thiện hơn.

Thái Nguyên, ngày 10 tháng 11 năm 2015

Tác Giả

Hoàng Minh Thủy

MỤC LỤC

DANH MỤC CÁC BẢNG

DANH MỤC CÁC HÌNH

MỞ ĐẦU

1.1. Sự cần thiết lựa chọn đề tài

Trích chọn thông tin (IE - Information Extraction) là một lĩnh vực nghiên cứu quan trọng trong khai phá dữ liệu văn bản [3, 4]. Trích chọn thông tin là quá trình thu thập thông tin từ các nguồn dữ liệu theo nhiều định dạng khác nhau, không đồng nhất, thậm chí không có định dạng cụ thể, sau đó chuyển thành một dạng đồng nhất. Dữ liệu sau khi trích chọn được lưu vào cơ sở dữ liệu để xử lý hay được sử dụng cho những hệ thống khai phá dữ liệu. Từ dữ liệu, thông tin được trích chọn ra có thể sử dụng các kỹ thuật phân tích, khai phá để khám phá ra các mẫu thông tin có ích, tiềm ẩn trong dữ liệu.

Ngày nay, cùng với sự phát triển của công nghệ thông tin, Tin học đã dần được ứng dụng rộng rãi trong nhiều lĩnh vực như kinh tế, du lịch, thương mại, y tế, ngân hàng và mang lại nhiều lợi ích to lớn. Nền kinh tế không ngừng phát triển, đời sống văn hoá - xã hội ngày càng được nâng cao thì du lịch đã trở thành một nhu cầu không thể thiếu trong cuộc sống của người dân, trên các trang web du lịch là hàng loạt thông tin về các tour du lịch trong nước và ngoài nước. Tuy nhiên lượng thông tin về các tour du lịch trên Internet là vô cùng lớn, gây khó khăn cho người có nhu cầu du lịch trong việc lựa chọn địa điểm tham quan, lựa chọn công ty cung cấp dịch vụ,…. Do vậy, một bài toán đặt ra là cần phải xây dựng một hệ thống tìm kiếm và tư vấn du lịch, giúp người dùng có thể lựa chọn được những tour du lịch phù hợp nhất với yêu cầu đề ra. Để có một hệ thống tìm kiếm và tư vấn tốt thì trước tiên ta phải xây dựng được tập dữ liệu có độ chính xác cao. Cùng với nó là bài toán con trích chọn thông tin du lịch trong văn bản tiếng Việt.

Để có thể tiến đến tìm hiểu được những vấn đề trên, em lựa chọn đề tài "***Nghiên cứu các phương pháp trích chọn thông tin và ứng dụng trích chọn thông tin du lịch trong văn bản Tiếng Việt***" làm luận văn tốt nghiệp Thạc sĩ của mình.

1.2. Mục tiêu đề tài

Tìm hiểu các phương phương pháp trích chọn thông tin và xây dựng mô hình giải quyết bài toán trích chọn thông tin về các tour du lịch từ các trang thông tin điện tử tiếng Việt trên Internet.

1.3. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu của đề tài là các phương pháp tiếp cận giải quyết bài toán trích chọn thông tin trong văn bản tiếng Việt và các trang thông tin điện tử tiếng Việt trên mạng Internet về lĩnh vực du lịch.

Phạm vi nghiên cứu của đề tài là bài toán trích chọn thông tin về các tour du lịch trên một số trang thôn tin điện tử tiếng Việt (website) trên mạng Internet.

1.4. Phương pháp nghiên cứu

Phương pháp nghiên cứu của đề tài là nghiên cứu lý thuyết và nghiên cứu thực nghiệm.

Về nghiên cứu lý thuyết, đề tài đã tổng hợp các kết quả nghiên cứu về các phương pháp trích chọn thông tin từ văn bản tiếng Việt phục vụ phân tích, thống kê, báo cáo, ra quyết định. Về nghiên cứu thực nghiệm, đề tài xây dựng và cài đặt, thử nghiệm mô hình trích chọn thông tin du lịch từ một số trang web về du lịch bằng tiếng Việt trên mạng Internet.

1.5. Cấu trúc của luận văn

Cấu trúc luận văn gồm: mở đầu, bốn chương chính, kết luận và tài liệu tham khảo.

Phần mở đầu: Lý do chọn đề tài và bố cục luận văn

Chương 1: Giới thiệu tổng quan bài toán trích chọn thông tin và một số lĩnh vực nghiên cứu liên quan.

Chương 2: Trình bày một số phương pháp trích chọn thông tin. Trên cơ sở tìm hiểu, luận văn sẽ sử dụng một số phương pháp tiếp cận để giải quyết bài toán trích chọn thông tin du lịch trong văn bản tiếng Việt.

Chương 3: Đưa ra mô hình trích chọn thông tin du lịch trong văn bản tiếng Việt.

Chương 4: Cài đặt, thử nghiệm mô hình trích chọn thông tin du lịch trên một số trang web du lịch bằng tiếng Việt trên mạng Internet.

Phần kết luận: Tóm tắt các kết quả đạt được và hướng phát triển tiếp của đề tài.

Chương 1
TỔNG QUAN VỀ TRÍCH CHỌN THÔNG TIN VÀ BÀI TOÁN TRÍCH CHỌN THÔNG TIN DU LỊCH

Chương này giới thiệu tổng quan về trích chọn thông tin và bài toán trích chọn thông tin du lịch trong văn bản tiếng Việt.

1.1. Tổng quan về trích chọn thông tin

Trích chọn thông tin là một lĩnh vực quan trọng trong khai phá dữ liệu văn bản, nó được định nghĩa như sau: Trích chọn thông tin (IE – Information Extraction) [3, 4] là quá trình lấy thông tin từ các nguồn ở những định dạng không đồng nhất thậm chí không có định dạng cụ thể khi nó ở dạng văn bản diễn đạt bằng ngôn ngữ tự nhiên, sau đó chuyển thành một dạng đồng nhất. Dữ liệu sau khi trích chọn được sử dụng, trình bày trực tiếp cho người dùng, lưu vào cơ sở dữ liệu để xử lý sau đó hay sử dụng cho những hệ thống tìm kiếm thông tin như một dữ liệu đã qua bước tiền xử lý.

Từ dữ liệu, thông tin được trích chọn ra ta có thể sử dụng các kỹ thuật phân tích, khai thác dữ liệu (Data Mining) để khám phá ra các mẫu thông tin hữu ích. Chẳng hạn việc cấu trúc lại các mẫu tin quảng cáo, mẫu tin bán hàng trên internet có thể giúp hỗ trợ tư vấn, định hướng người dùng khi mua sắm. Việc trích chọn và cấu trúc lại các mẫu tin tìm người, tìm việc sẽ giúp cho quá trình phân tích thông tin nghề nghiệp, xu hướng công việc, … hỗ trợ cho người tìm việc, cũng như nhà tuyển dụng.

Trích chọn thông tin không đòi hỏi hệ thống phải đọc hiểu nội dung của tài liệu văn bản, nhưng hệ thống phải có khả năng phân tích tài liệu và tìm kiếm các thông tin liên quan mà hệ thống mong muốn được tìm thấy. Các kỹ thuật trích chọn thông tin có thể áp dụng cho bất kỳ tập tài liệu nào mà chúng ta cần rút ra những thông tin chính, cần thiết cũng như các sự kiện liên quan. Các kho dữ liệu văn bản về một lĩnh vực trên internet là ví dụ điển hình,

thông tin trên đó có thể tồn tại ở nhiều nơi khác nhau, dưới nhiều định dạng khác nhau. Sẽ rất hữu ích cho các khảo sát ứng dụng nếu như các thông tin thuộc các lĩnh vực liên quan được trích chọn, tích hợp lại thành một hình thức thống nhất và biểu diễn một cách có cấu trúc. Khi đó thông tin trên internet sẽ được chuyển vào một cơ sở dữ liệu có cấu trúc phục vụ cho các ứng dụng phân tích và khai thác khác nhau.

Các nghiên cứu liên quan đến trích chọn thông tin văn bản tập trung vào:

1) Trích chọn từ khóa (Keyphrase Extraction): Tìm kiếm các thuật ngữ chính có liên quan, thể hiện ngữ nghĩa, nội dung, chủ đề của tài liệu hay một tập các tài liệu.

2) Trích chọn thực thể có tên (Named Entity Recognition): Việc trích chọn ra các thực thể có tên tập trung vào các phương pháp nhận diện các đối tượng, thực thể như: tên người, tên công ty, tên tổ chức, một địa danh, nơi chốn.

3) Trích chọn quan hệ (Relationship Extraction): Cần xác định mối quan hệ giữa các thực thể đã nhận biết từ tài liệu. Chẳng hạn xác định nơi chốn cho một tổ chức, công ty hay nơi làm việc của một người nào đó. [2, 3].

1.1.1. Bài toán trích chọn thực thể

Con người, thời gian, địa điểm… là những đối tượng cơ bản trong một văn bản. Mục đích chính của bài toán trích chọn thực thể là xác định ra các đối tượng này từ đó giúp cho người đọc trong việc hiểu rõ văn bản.

Bài toán trích chọn thực thể là bài toán đơn giản nhất trong các bài toán trích chọn thông tin, tuy vậy nó lại là bước cơ bản nhất nên được thực hiện trước khi giải các bài toán phức tạp hơn trong lĩnh vực này. Rõ ràng là để có thể xác định được các mối quan hệ giữa các thực thể ta phải xác định được đâu là các thực thể tham gia vào mối quan hệ đó.

Bài toán trích chọn thực thể trong văn bản là tìm câu trả lời cho các câu hỏi: ai ?, bao giờ ?, ở đâu ?,... [19].

Bảng 1.1. Bảng phân loại thực thể

Tên nhãn	Ý nghĩa
PER	Tên người
ORG	Tên tổ chức
LOC	Tên địa danh
NUM	Số
PCT	Phần trăm
CUR	Tiền tệ
TIME	Ngày tháng, thời gian
MISC	Những loại thực thể khác ngoài 7 loại trên
O	Không phải thực thể

Ý nghĩa của bài toán trích chọn thực thể

Một hệ thống trích chọn thực thể tốt có thể được ứng dụng trong nhiều lĩnh vực khác nhau, cụ thể có thể được sử dụng để:

1) Hỗ trợ web ngữ nghĩa. Web ngữ nghĩa là các trang Web có thể biểu diễn dữ liệu "thông minh" (có khả năng kết hợp, phân lớp và khả năng suy diễn trên dữ liệu đó). Sự thành công của các Web ngữ nghĩa phụ thuộc vào các ontology cũng như sự phát triển của các trang Web được chú giải bởi các siêu dữ liệu tuân theo các ontology này. Mặc dù lợi ích mà các ontology đem lại là rất lớn nhưng việc xây dựng chúng một cách tự động lại hết sức khó khăn. Vì lý do này, các công cụ trích chọn thông tin tự động từ các trang web để "làm đầy" các ontology như hệ thống trích chọn thực thể là hết sức cần thiết.

2) Xây dựng các máy tìm kiếm hướng thực thể. Người dùng có thể tìm thấy các trang Web nói về "Clinton" là một địa danh ở Bắc Carolina một cách nhanh chóng mà không phải duyệt qua hàng trăm trang Web nói về tổng thống Bill Clinton.

3) Trích chọn thực thể có thể được xem như là bước tiền xử lý làm đơn giản hóa các bài toán như dịch máy, tóm tắt văn bản. ..

4) Như đã đề cập ở trên, một hệ thống trích chọn thực thể có thể đóng vai trò là một thành phần cơ bản cho các bài toán trích chọn thông tin phức tạp hơn.

5) Trước khi đọc một tài liệu, người dùng có thể đọc lướt qua các tên người, tên địa danh, tên công ty được đề cập đến trong đó.

6) Tự động đánh chỉ số cho các sách. Trong các sách, tài liệu phần lớn các chỉ mục là các loại thực thể.[2, 3]

1.1.2. Bài toán trích chọn quan hệ

Các nghiên cứu về trích chọn thực thể, cũng như quan hệ đã được tổ chức MUC (Message Understanding Conferences) và ACE (Automatic Content Extration) đầu tư và thúc đẩy phát triển. Trích chọn quan hệ bắt đầu được quan tâm từ hội thảo MUC lần thứ 7 năm 1998, từ đó ngày càng được chú ý đến. Trích chọn quan hệ là việc xác định mối quan hệ ngữ nghĩa giữa các thực thể trong văn bản hay trong một câu. Chẳng hạn xác định nơi chốn cho một tổ chức, công ty hay nơi làm việc của một người nào đó. Ví dụ từ một đoạn văn bản: "James Gosling vào làm việc cho Sun Microsystems từ năm 1984 nằm tại Silicon Valley " ta có thể nhận diện được các thực thể, loại thực thể và quan hệ giữa chúng như sau:

1) CON NGƯỜI làm việc TỔ CHỨC: nhận diện được hai thực thể là "James Gosling" và "Sun Microsystems". Mối quan hệ giữa hai thực thể này là "làm việc".

2) TỔ CHỨC nằm tại NƠI CHỐN: nhận diện được hai thực thể là "Sun Microsystems" và "Silicon Valley"; mối quan hệ giữa hai thực thể này là "nằm tại" [14].

Ứng dụng

Trích chọn quan hệ được ứng dụng trong nhiều lĩnh vực khác nhau. Lĩnh vực đầu tiên phải nhắc tới là việc xây dựng cơ sở tri thức mà điển hình là xây dựng Ontology – phần nhân của Web ngữ nghĩa. Trong khi những lợi ích mà Web ngữ nghĩa đem lại là rất lớn thì việc xây dựng các ontology một cách thủ công lại hết sức khó khăn. Giải pháp cho vấn đề này chính là kĩ thuật trích chọn thông tin nói chung và trích chọn quan hệ nói riêng để tự động hóa một phần quá trình xây dựng các ontology.

Trích chọn quan hệ cũng được sử dụng nhiều trong các hệ thống hỏi đáp. Một số hệ thống hỏi đáp đã được xây dựng dựa vào việc trích chọn tự động các từ, khái niệm và mối quan hệ. Ngoài ra, trích chọn quan hệ còn có ứng dụng trong các lĩnh vực xử lý ảnh như phát hiện ảnh qua đoạn văn bản (text-to-image generation). Trích chọn quan hệ cũng là một công cụ đắc lực trong lĩnh vực công nghệ sinh học như tìm quan hệ bệnh tật - Genes, ảnh hưởng qua lại giữa protein-protein (Protein-Protein interaction)…[1, 12].

1.1.3. Bài toán trích chọn cụm từ khóa

Cụm từ khóa được xem là thành phần chính hay một dạng siêu dữ liệu (Meta Data) thể hiện nội dung của tài liệu văn bản [18]. Mục đích của hầu hết các nghiên cứu trích chọn cụm từ khóa là nhằm tìm kiếm các đặc trưng tốt để mã hóa văn bản [8, 17, 18] ứng dụng trong các hệ thống phân loại, gom cụm, tóm tắt và tìm kiếm văn bản. Tùy vào đặc trưng của từng ngôn ngữ sẽ có những phương pháp khác nhau để tìm kiếm các cụm từ khóa. Hầu hết các phương pháp đều dựa trên các kỹ thuật truyền thống được dùng trong xử lý ngôn ngữ tự nhiên như tiền xử lý văn bản, tách đoạn, tách câu, tách từ, phân tích cú pháp, phân tích ngữ nghĩa, thống kê và học máy [18].

Ứng dụng

1) Các kho dữ liệu văn bản lớn như các thư viện số phát triển rất nhanh, điều đó dẫn đến gia tăng giá trị thông tin tóm tắt.

2) Hỗ trợ người dùng nhận biết về nội dung của tài liệu và kho tài liệu.

3) Ứng dụng trong truy vấn thông tin, mô tả những tài liệu trả về từ kết quả truy vấn. Định hướng tìm kiếm cho người dùng.

4) Nền tảng cho chỉ mục tìm kiếm.

5) Là đặc trưng dùng trong kỹ thuật phân loại, gom cụm tài liệu [5, 10].

1.2. Bài toán trích chọn thông tin du lịch

Bài toán "Trích chọn thông tin du lịch" là một phần của bài toán trích chọn thông tin, trong đó ta sử dụng các phương pháp trích chọn trên miền dữ liệu du lịch. Mục tiêu chính của bài toán trích chọn thông tin du lịch trong văn bản tiếng Việt là trích ra các thông tin đặc trưng về một tour du lịch có trong bài viết, chuyển những thông tin đó về dạng có cấu trúc để làm dữ liệu cho việc xây dựng một hệ thống tìm kiếm và tư vấn du lịch. Hệ thống tư vấn du lịch là hệ thống hỗ trợ người dùng lựa chọn các dịch vụ du lịch phù hợp nhất với bản thân. Đồng thời, hệ thống còn có khả năng đưa ra các giải pháp đề nghị tương ứng với yêu cầu đã cho. Hệ thống tư vấn du lịch tương tự như các chuyên gia du lịch, hiểu rõ các vấn đề chuyên môn nhằm tư vấn cho khách hàng chọn lựa dịch vụ. Khi sử dụng hệ thống một người khách du lịch có thể nhập vào số tiền dành cho việc du lịch và những địa danh muốn đến, hệ thống sẽ tìm kiếm đưa ra tất cả những tour du lịch đáp ứng được yêu cầu và hỗ trợ tư vấn cho du khách về những tour phù hợp nhất. Trong phạm vi luận văn, tác giả sẽ tập trung vào mục tiêu *trích chọn ra các thông tin đặc trưng về một tour du lịch* từ các trang thông tin điện tử tiếng Việt (Website) trên Internet.

Chỉ khi xây dựng được một tập dữ liệu chính xác, đầy đủ thì mới có thể hình thành nên một hệ thống tư vấn hiệu quả.

1.3. Ý nghĩa của bài toán trích chọn thông tin du lịch

1.3.1. Ý nghĩa khoa học

Đây là một hướng trong khai phá dữ liệu văn bản nói chung và thông tin nói riêng, nó đang được nghiên cứu và ứng dụng rộng rãi....

1.3.2. Ý nghĩa thực tế

Bài toán trích chọn thông tin du lịch có ý nghĩa rất lớn trong thực tế, hầu hết mọi người khi muốn đi du lịch sẽ tìm hiểu thông tin trên Internet, nhưng các bài giới thiệu về một tour du lịch, hay một địa danh rất dài, thậm chí không có thông tin cần thiết, mục đích khi tìm hiểu về một tour du lịch là đi đâu, bao giờ xuất phát, đi trong thời gian bao lâu, khởi hành vào thời điểm nào và quan trọng nhất là giá thành là bao nhiêu, bài toán trên sẽ đáp ứng được việc trích ra đầy đủ các thông tin mà người dùng cần biết về một tour du lịch. Từ những thông tin đó, người dùng có thể quyết định có lựa chọn tour du lịch đó hay không một cách nhanh chóng.

1.4. Ứng dụng của bài toán trích chọn thông tin du lịch

1.4.1. Hệ thống tìm kiếm và tư vấn du lịch

Hệ thống tìm kiếm và tư vấn du lịch là hệ thống đưa ra tất cả các tour du lịch phù hợp với yêu cầu và hỗ trợ người dùng lựa chọn các tour du lịch phù hợp nhất. Đồng thời, hệ thống còn có khả năng đưa ra các giải pháp đề nghị tương ứng với yêu cầu đã cho. Ví dụ khi một du khách cần chọn một tour du lịch, những thông tin mà người đó quan tâm đến là: thông tin về tour đó (giá cả từ các công ty du lịch khác nhau, đi trong bao lâu, di chuyển bằng phương tiện gì, ở tại khách sạn thế nào,…), thông tin về các công ty cung cấp

dịch vụ (chế độ khuyến mãi, chất lượng dịch vụ,. ..), v.v. Họ phải tốn nhiều thời gian để tìm kiếm và tổng hợp thông tin để có thể quyết định chọn tour. Hệ thống tìm kiếm và tư vấn dịch vụ sẽ giúp trích chọn, tổng hợp các thông tin theo các yêu cầu và đưa ra những tour phù hợp nhất.

1.4.2. Bài toán dự đoán xu hướng du lịch

Từ việc đưa ra được thông tin về các tour du lịch của từng website, ta có thể thống kê được số tour đến từng địa điểm du lịch, từ đó có thể dự đoán được những thông tin sau: địa điểm du lịch nào đang được coi là thu hút với du khách, địa điểm du lịch nào đang vắng du khách, công ty cung cấp dịch vụ này có các tour du lịch thế mạnh là gì, là các tour trong nước hay nước ngoài, công ty này có liên kết mạnh với địa điểm du lịch cụ thể nào hay không?

Ví dụ sau khi trích chọn thông tin về các tour du lịch, ta thống kê thấy trong 100 tour thì có 80 tour đi đến các địa danh liên quan đến biển, thì ta có thể kết luận du lịch Biển đang là tour hot nhất trong thời điểm này.

Ví dụ tiếp theo là trong một website du lịch, nếu ta thống kê được các tour du lịch miền bắc có tần số xuất hiện nhiều hơn hẳn so với các tour du lịch tới các vùng miền khác thì ta có thể dự đoán điểm mạnh của công ty du lịch này là các tour miền bắc và lựa chọn các tour du lịch trong miền bắc của công ty này sẽ được cung cấp các dịch vụ tốt hơn so với các tour tới các vùng miền khác.

1.5. Kết luận chương

Trong chương 1, luận văn đã trình bày khái niệm và những nghiên cứu cơ bản của bài toán trích chọn thông tin, đồng thời giới thiệu về bài toán trích chọn thông tin du lịch, ý nghĩa và ứng dụng của bài toán trong khoa học và thực tế. Trong chương tiếp theo, luận văn sẽ trình bày một số phương pháp tiếp cận giải quyết bài toán trích chọn thông tin.

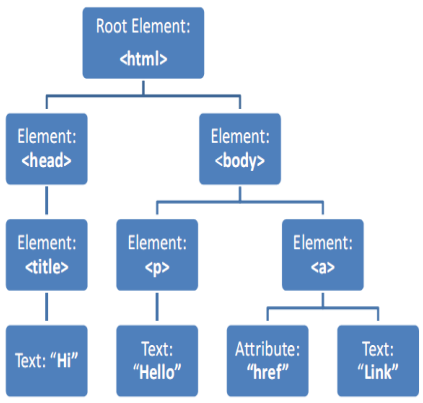
Chương 2
MỘT SỐ PHƯƠNG PHÁP TRÍCH CHỌN THÔNG TIN

Có nhiều phương pháp cũng như giải thuật được sử dụng để giải quyết bài toán trích chọn thông tin. Chương 2 sẽ giới thiệu một số phương pháp trích chọn thông tin đó là phương pháp dựa trên luật, phương pháp phân tích mã HTML thành cây DOM, phương pháp trích chọn thông tin dựa trên học máy và phương pháp kết hợp giữa phân tích mã HTML và luật. Trong phần cuối, luận văn sẽ phân tích về ưu điểm, nhược điểm của các phương pháp trên, từ đó lựa chọn ra phương pháp phù hợp cho bài toán ở chương 3.

2.1. Trích chọn thông tin dựa vào cây DOM

2.1.1. Khái niệm cây DOM

Theo W3C thì DOM (Document Object Model) là một giao diện lập trình ứng dụng (API) cho các văn bản HTML hợp lệ và các văn bản XML có cấu trúc chặt chẽ. Nó định nghĩa cấu trúc logic của các văn bản và cách thức một văn bản được truy cập và thao tác [20]. Dưới đây là một đoạn mã html đơn giản được biểu diễn dưới dạng cây DOM như sau:

<html> <head> <title>Hi</title> </head> <body> Hello Link </body> </html>	Dạng biểu diễn cây DOM của mã HTML 

Hình 2.1. Mô hình biểu diễn cây DOM

2.1.2. Xây dựng cây DOM

Xây dựng cây DOM từ những trang Web đầu vào là một bước cần thiết trong nhiều giải thuật trích chọn thông tin [20]. Hai phương pháp cơ bản để xây dựng cây DOM.

1) Sử dụng các thẻ riêng biệt

Hầu hết các thẻ HTML làm việc trong một cặp. Mỗi cặp chứa một thẻ mở <> và một thẻ đóng </>. Bên trong mỗi cặp thẻ có thể có những cặp thẻ khác, kết quả là cấu trúc trở nên chồng chéo. Xây dựng một cây DOM từ một trang Web bằng cách sử dụng mã HTML của nó là một vấn đề cần thiết. Trong một cây DOM, mỗi cặp thẻ là một node, những cặp thẻ ẩn bên trong được gọi là node con của node hiện tại. Có hai nhiệm vụ cần tiến hành đó là:

- *Làm sạch mã HTML*: một vài thẻ không cần thẻ đóng (như <li>, <hr>,<p>) mặc dù chúng có thẻ đóng. Bởi vậy một thẻ đóng nên được chèn vào để tất cả các thẻ trở thành trạng thái cân bằng. Các thẻ được định dạng không tốt cũng cần phải được sửa chữa. Một thẻ sai thường là một thẻ đóng, đó là thẻ cắt ngang các khối ẩn bên trong. Ví dụ: <tr> … <td> … </tr> … </td>, sẽ rất khó để sửa lỗi trường hợp này nếu tồn tại sự chồng chéo đa cấp. Có một vài phần mềm mã nguồn mở để làm sạch mã HTML, một số những phần mềm thông dụng như: JTidy, NekoHTML, HTMLCleaner.
- *Xây dựng cây*: Chúng ta có thể đi theo các khối con của các thẻ HTML để xây dựng được cây DOM.

2) Sử dụng các thẻ và các hộp ảo (visual cue)

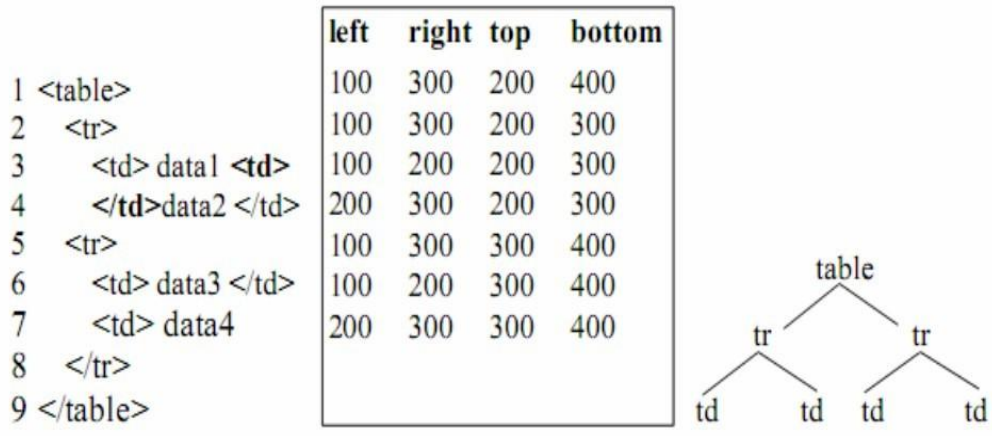
Thay vì phân tích mã HTML để sửa lỗi, có thể sử dụng sự biểu diễn hoặc các thông tin ảo (ví dụ như: địa chỉ trên màn hình mà các thẻ được biểu diễn) để suy luận mối quan hệ có cấu trúc của các thẻ và có thể xây dựng được

cây DOM. Phương thức xây dựng có thể phân tích mã HTML thành cây DOM, miễn là trình duyệt có thể hiển thị được đoạn mã đó một cách chính xác.

Trong một trình duyệt web, mỗi phần tử HTML (chứa đựng một thẻ mở, các thuộc tính tùy chọn, nội dung HTML được nhúng tùy ý và một thẻ đóng, thẻ này có thể thiếu) được biểu diễn như một hình chữ nhật. Thông tin ảo này có thể lấy được sau khi mã HTML được biểu diễn trên trình duyệt. Một cây DOM sau đó có thể được xây dựng dựa vào các thông tin ảo này. Các bước xử lý như sau:

- Tìm 4 đường biên của hình chữ nhật ứng với mỗi phần tử HTML thông qua việc công cụ trình diễn của trình duyệt, ví dụ: Google chrome.
- Theo sự tuần tự của các thẻ mở và kiểm tra xem một hình chữ nhật có nằm trong một hình chữ nhật khác không, để xây dựng cây DOM.

Ví dụ minh họa về sử dụng visual cue:



Hình 2.2. Minh họa sử dụng visual cue

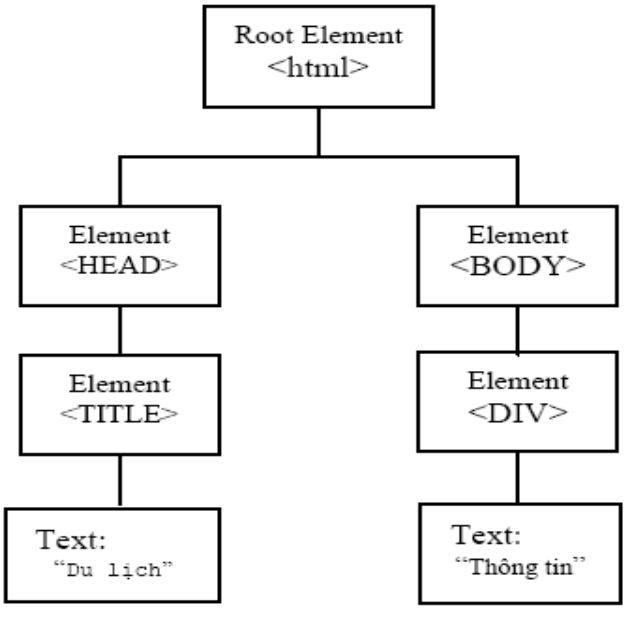
2.1.3. Sử dụng cây DOM để trích chọn thông tin

Để trích chọn được thông tin cần thiết ở một node của cây DOM, chúng ta cần chỉ rõ đường đi từ gốc của cây đến node cần trích chọn thông tin. Đường đi này gọi là một Xpath [21] hay mẫu trích chọn.

Muốn trích chọn thông tin dựa vào cây DOM thì trước hết phải xây dựng cây DOM cho mã HTML của trang web.

Các mẫu trích chọn có thể được hiểu là đường dẫn từ gốc của cây DOM đến node chứa nội dung cần trích chọn.

Ví dụ đây là cây DOM của một đoạn mã HTML chứa thông tin về một tour du lịch, gồm tên tour (title) và thông tin chi tiết về tour (div). Bài toán đặt ra là sử dụng cây DOM này trích chọn các thông tin về tên tour và thông tin chi tiết về tour. Mẫu trích chọn được xây dựng sau:



Hình 2.3. Minh họa cây DOM dùng trong mẫu trích chọn

Mẫu trích chọn tên tour: HTML → HEAD → TITLE → TEXT

Mẫu trích chọn thông tin chi tiết: HTML → BODY → DIV → TEXT

2.2. Trích chọn thông tin dựa trên tập luật

Trích chọn thông tin dựa trên tập luật hay còn được gọi là phương pháp trích chọn thông tin dựa trên tri thức (knowledge - driven). Phương pháp này dựa trên kiến thức chuyên gia (thường là do chuyên gia về ngôn ngữ và

chuyên gia miền dữ liệu tạo ra tập luật). Do vậy nó đòi hỏi người xây dựng phải hiểu dữ liệu mới có thể tạo ra được tập luật đầy đủ.

2.2.1. Hình thức và biểu diễn của luật

Một luật cơ bản có dạng: "Mẫu theo ngữ cảnh → hành động". Một mẫu theo ngữ cảnh bao gồm một hoặc nhiều mẫu được gán nhãn chứa đặc tính đa dạng của thực thể và bối cảnh thực thể xuất hiện trong văn bản. Một mẫu gán nhãn được xác định bằng biểu thức chính quy dựa vào đặc trưng của thẻ trong văn bản và nhãn tùy chọn. Các đặc trưng có thể chỉ là từ hoặc đoạn hoặc cả tài liệu trong đó có các từ xuất hiện.

Phần hành động của các luật được sử dụng để biểu thị việc gán nhãn: gán nhãn thực thể cho một chuỗi các thẻ, chèn vào dấu hiệu bắt đầu hoặc kết thúc một thực thể, hoặc gán nhiều thẻ thực thể [16].

2.2.2. Đặc trưng của từ tố (token)

Một từ tố trong câu thường là sự kết hợp của tập các đặc trưng thu được thông qua một hoặc nhiều các tiêu chí sau:

1) Chuỗi biểu diễn cho từ tố.
2) Các quy tắc ngữ pháp như: Quy định về viết hoa, viết thường, kết hợp giữa văn bản, số, ký hiệu đặc biệt, dấu cách, dấu chấm câu, …
3) Từ loại của từ tố.
4) Danh sách từ điển chứa từ tố.
5) Chú thích kèm theo các bước xử lý trước đó.

2.2.3. Tập luật xác định thực thể đơn

Tập luật xác định một thực thể đơn đầy đủ bao gồm ba loại mẫu như sau:

1) Mẫu tùy chọn ghi lại bối cảnh trước khi bắt đầu của thực thể.
2) Một mẫu so khớp các từ tố trong các thực thể.
3) Một mẫu tùy chọn để ghi lại bối cảnh sau khi kết thúc của thực thể.

Ví dụ: Thực thể tên người có dạng "Dr. Yair Weiss", thực thể tên người trong các văn bản thường xuất hiện sau chức danh, giữa chức danh và tên người là dấu ".", tên người thường bắt đầu bằng kí tự in hoa. Như vậy để xác định một thực thể tên người ta có luật như sau: Đầu tiên ta xây dựng một từ điển chức danh (có chứa các chức danh như: "Prof ", "Dr", "Mr", "Mrs", "Miss").

Sau đó so sánh các kí tự trước dấu chấm với từ điển chức danh, nếu thấy xuất hiện trong từ điển thì hai từ viết hoa sau dấu chấm sẽ là thực thể tên người.

({Dictionary – Lookup = Titles}{String = "." }{Orthography type = capitalized word}{2}) → Tên người.

Trong đó mỗi phần trong dấu ngoặc {} là một điều kiện và số theo sau cùng sẽ chỉ ra số lần lặp lại của thẻ. Ví dụ số 2 ở trên nghĩa là có hai từ viết hoa.

Ví dụ thực thể "Year" là các số xuất hiện sau giới từ "by" và "in". Như vậy, luật phát hiện ra thực thể "Year" như sau:

({String="by" | String="in"})({Orthography type = Number}):y → Year =: y.

Có hai mẫu được sử dụng trong luật này: mẫu đầu tiên để ghi lại ngữ cảnh xuất hiện của các thực thể "Year" là sau các giới từ "in", "on" và mẫu thứ hai ghi lại tính chất của thực thể "Year" là các con số.

Ví dụ thực thể "Timetour" của một tour du lịch có dạng "Thời gian: 6 ngày". Thực thể "Timetour" là các số xuất hiện sau các cụm từ "Thời gian:" hoặc "Thời lượng:". Như vậy, luật phát hiện ra thực thể "Timetour" như sau:

({String="Thời gian:" | String="Thời lượng:"})({Orthography type = Number})({String="Ngày" | String="Đêm"}) →TimeTour.

Có ba mẫu được sử dụng trong luật này: mẫu đầu tiên để ghi lại ngữ cảnh xuất hiện của các thực thể "Timetour" là sau các cụm từ "Thời gian", "Thời lượng", mẫu thứ hai ghi lại tính chất của thực thể "Timetour" là các

con số và mẫu thứ ba ghi lại dấu hiệu kết thúc của thực thể "Timetour" là cụm từ "Ngày" hoặc "Đêm".

2.2.4. Các luật đánh dấu biên của thực thể

Một số luật có dạng biểu thức chính quy với nhiều slot (ô, khe), mỗi slot đại diện cho một thực thể khác nhau sao cho luật này có thể đoán nhận được nhiều thực thể cùng một lúc. Những luật này rất hiệu quả khi dữ liệu được tổ chức dưới dạng bản ghi. Ví dụ, hệ thống dựa trên luật WHISK [15] sử dụng các luật này để khai thác các hồ sơ có cấu trúc như hồ sơ y tế, các bản ghi bảo trì thiết bị, và phân loại quảng cáo. Các luật này được viết lại từ những luật trong [15], để trích chọn hai thực thể, số lượng phòng ngủ và giá phòng từ một quảng cáo cho thuê căn hộ.

({Orthography type = Digit}):Bedrooms ({String ="BR"})({}*)

({String ="$"})({Orthography type = Number}):Price→Number of Bedrooms =:Bedroom, Rent =: Price [16].

2.2.5. Các luật xác định nhiều thực thể

Một số luật có dạng biểu thức chính quy với nhiều slot (ô, khe), mỗi slot đại diện cho một thực thể khác nhau sao cho luật này có thể đoán nhận được nhiều thực thể cùng một lúc. Những luật này rất hiệu quả khi dữ liệu được tổ chức dưới dạng bản ghi. Ví dụ, hệ thống dựa trên luật WHISK [15] sử dụng các luật này để khai thác các hồ sơ có cấu trúc như hồ sơ y tế, các bản ghi bảo trì thiết bị, và phân loại quảng cáo. Các luật này được viết lại từ những luật trong [15], để trích chọn hai thực thể, số lượng phòng ngủ và giá phòng từ một quảng cáo cho thuê căn hộ.

({Orthography type = Digit}):Bedrooms ({String ="BR"})({}*)

({String ="$"})({Orthography type = Number}):Price→Number of Bedrooms =:Bedroom, Rent =: Price [16]

2.2.6. Đánh giá phương pháp tiếp cận dựa trên luật

Ưu điểm: Thích hợp với hệ thống làm việc một cách thủ công, phụ thuộc nhiều vào kỹ năng và kinh nghiệm của người viết ra luật. Dựa vào trực giác, quan sát. Hiệu quả đạt được tốt hơn.

Nhược điểm: Phụ thuộc rất nhiều vào nguồn tài nguyên ngôn ngữ như bộ từ điển phù hợp, khả năng của người viết luật. Nếu một nhân tố nào bị mất, hệ thống có thể trở lên không còn chắc chắn. Việc phát triển có thể sẽ tốn nhiều thời gian, Khó điều chỉnh khi có sự thay đổi [11].

2.3. Trích chọn thông tin dựa trên học máy

Trích chọn thông tin dựa trên học máy còn được gọi là phương pháp tiếp cận dựa trên dữ liệu (data-driven). Hướng tiếp cận này không đòi hỏi người xây dựng phải thành thạo về ngôn ngữ, lĩnh vực nghiên cứu như các chuyên gia. Nhưng lại đòi hỏi một lượng lớn dữ liệu để xây dựng tập huấn luyện tốt vả đủ lớn dùng cho bộ phân lớp tối ưu. Phương pháp này thường dựa trên mô hình xác suất (probabilistic models), lý thuyết thông tin (information theory), và đại số tuyến tính (linear algebra). Một bộ đoán nhận sẽ thực hiện việc gán cho kho dữ liệu văn bản các nhãn phù hợp với từng lớp. Sau khi có tập dữ liệu huấn luyện phù hợp đã được gán nhãn, thuật toán huấn luyện được sử dụng, hệ thống sẽ sử dụng kết quả trả về từ thuật toán huấn luyện để phục vụ cho quá trình phân tích văn bản mới.

Ngoài ra, ta còn có thể sử dụng bộ quan hệ huấn luyện để tương tác với người dùng trong suốt quá trình xử lý. Người sử dụng được phép chỉ ra liệu rằng các giả thuyết của hệ thống về văn bản có đúng không, nếu không đúng, hệ thống sẽ thay đổi các quy tắc của chính nó để điều tiết thông tin mới [4, 14].

Ưu điểm: Nhấn mạnh đến việc tạo dữ liệu huấn luyện, cách tiếp cận này không cần có sự tham gia của các chuyên gia về ngôn ngữ và chuyên gia

miền. Ưu điểm tiếp theo của phương pháp là các mô hình sau khi huấn luyện có thể sử dụng với các miền dữ liệu khác nhau.

Nhược điểm: Thứ nhất, trích chọn thông tin dựa trên học máy cần một lượng dữ liệu lớn để huấn luyện mô hình. Trong một số trường hợp, việc gán nhãn dữ liệu tốn thời gian và chi phí. Thứ hai, trong các bài toán trích chọn, phương pháp tiếp cận dựa trên dữ liệu không giải quyết được các vấn đề có liên quan đến ngữ nghĩa. Thứ ba, do phương pháp tiếp cận dựa trên dữ liệu được xây dựng trên các mô hình xác suất thống kê, do đó trong một số trường hợp nếu quá trình làm dữ liệu huấn luyện không tốt dẫn đến kết quả của quá trình trích chọn không cao. Thứ tư, khi dữ liệu có sự thay đổi → có thể cần phải gán nhãn lại cho cả tập dữ liệu huấn luyện.

Thực tế cho thấy, việc thu thập tập dữ liệu huấn luyện với chất lượng tốt có khi rất tốn kém, chúng ta cần phải tốn nhiều thời gian cho việc chọn mẫu, gán nhãn và để có kết quả tốt cần rất nhiều dữ liệu. [15, 22].

2.4. Phương pháp kết hợp giữa phân tích mã HTML và luật

Sử dụng phương pháp kết hợp giữa phân tích mã HTML và dùng luật sẽ khắc phục được một số nhược điểm khi sử dụng riêng lẻ từng loại: Nếu chỉ sử dụng riêng phương pháp trích chọn thông tin dựa trên luật (rule - based), ta sẽ mất thời gian cho công việc tiền xử lý dữ liệu như: loại bỏ thẻ html, tách câu, tách từ, loại bỏ từ dừng … và có thể độ chính xác không cao do sự nhập nhằng về ngôn ngữ. Còn nếu chỉ sử dụng riêng phương pháp trích chọn thông tin dựa và cây DOM bằng đường đi XPATH, do các website không tuân thủ theo một quy cách chung, dẫn đến cùng một website nhưng trong những trang web khác nhau lại có cách bố trí khác nhau. Ví dụ: Trên website Dulichmienbac.com, có bài viết thì thông tin cần trích chọn đặt tại thẻ , có bài viết lại đặt tại thẻ . Do vậy, sau khi phân tích mã HTML xong dựa vào luật để nhận biết đâu là thông tin cần trích chọn.

Sau khi phân tích dữ liệu và đánh giá ưu điểm, nhược điểm và độ phù hợp của phương pháp, tác giả quyết định sử dụng phương pháp trích chọn thông tin dựa trên việc phân tích mã HTML và sử dụng luật, do phương pháp này có những đặc điểm sau: Thứ nhất, sử dụng phương pháp trên ta sẽ không mất công xây dựng tập huấn luyện như với phương pháp trích chọn dựa trên học máy (với miền dữ liệu du lịch, việc xây dựng tập dữ liệu rất tốn thời gian và công sức do có nhiều từ đồng nghĩa, đoạn văn có ý nghĩa nhập nhằng, cấu trúc dữ liệu không nhất quán); Thứ hai, do dữ liệu cho bài toán rất nhiều và có nhiều bài viết không liên quan, nếu dùng tất cả các bài viết thì sẽ dẫn đến mất nhiều thời gian và độ chính xác là không cao. Dẫn đến sử dụng luật để loại bỏ bớt dữ liệu dư thừa trước khi đi vào trích chọn; Thứ ba, sử dụng phương pháp phân tích mã HTML, sẽ giảm được thời gian cho việc tiền xử lý dữ liệu do thao tác ngay trên các thẻ HTML.

2.5. Kết luận chương

Chương 2 giới thiệu tổng quan về các phương pháp tiếp cận cơ bản để giải quyết bài toán trích chọn thông tin. Phương pháp tiếp cận dựa trên luật (rule – based), phương pháp tiếp cận dựa trên học máy, phương pháp tiếp cận dựa trên phân tích mã HTML thành cây DOM và phương pháp kết hợp phân tích mã HTML và sử dụng luật. Có thể thấy, mỗi phương pháp đều có những ưu điểm và nhược điểm. Sau khi đánh giá ưu điểm và mức độ phù hợp của các phương pháp với đặc điểm của miền dữ liệu du lịch, luận văn lựa chọn phương pháp kết hợp giữa phân tích mã HTML và luật. Trong chương tiếp theo, luận văn sẽ trình bày chi tiết bài toán trích chọn thông tin du lịch trong văn bản tiếng Việt và mô hình giải quyết bài toán.

Chương 3

BÀI TOÁN TRÍCH CHỌN TOUR DU LỊCH TRÊN MỘT SỐ TRANG THÔNG TIN ĐIỆN TỬ TIẾNG VIỆT

Trong chương này, luận văn sẽ tập trung làm rõ bài toán trích chọn thông tin trên một số trang thông tin điện tử tiếng Việt, phân tích ưu nhược điểm của các phương pháp đã được trình bày ở chương 2 và mục đích khi xây dựng mô hình là tạo ra một tập dữ liệu mẫu đầy đủ, không mất thời gian trong việc tiền xử lý dữ liệu nên trong chương này tác giả lựa chọn giải pháp trích chọn thông tin dựa trên phương pháp kết hợp giữa phân tích mã HTML và luật để xây dựng mô hình chi tiết cho bài toán trích chọn thông tin trên một số trang thông tin điện tử tiếng Việt.

3.1. Bài toán trích chọn thông tin du lịch trên một số trang thông tin điện tử tiếng Việt

3.1.1. Phát biểu bài toán

Sau quá trình khảo sát, nghiên cứu các bài viết về lĩnh vực du lịch từ các trang thông tin điện tử tiếng Việt trên Internet, tác giả thấy rằng thông tin về một tour du lịch sẽ bao gồm hai thành phần là tên tour và thông tin chi tiết về tour, trong thông tin chi tiết về tour thì tùy thuộc vào từng website mà ta có số lượng thuộc tính khác nhau, thông thường thì trong phần này sẽ có ba thuộc tính cơ bản là mã tour, thời gian và giá tour. Ngoài ra còn có thể có thêm các thuộc tính khác như: Phương tiện, lịch trình, điểm khởi hành, ngày khởi hành, điểm kết thúc. Như vậy, ta thấy rằng các thông tin cơ bản về một tour du lịch sẽ bao gồm hai thành phần là tên tour, thông tin chi tiết về tour.

Mục tiêu của bài toán trích chọn thông tin du lịch trong văn bản tiếng Việt là trích ra các thông tin cơ bản về các tour du lịch từ các văn bản tiếng Việt không có cấu trúc.

Đầu vào: Bài viết tiếng Việt về lĩnh vực du lịch.

Đầu ra: Thông tin về các tour du lịch trong bài viết đó.

Thông tin về một tour du lịch được định nghĩa là một bộ E gồm hai thành phần đó là: *Tên tour, thông tin chi tiết về tour*. Một cách hình thức E được định nghĩa như sau:

E = <tên tour, thông tin chi tiết về tour>

Trong đó:

1) *Tên tour*: Là tên của một chuyến du lịch được đề cập trong bài viết

2) *Thông tin chi tiết về tour*: Là tập các thuộc tính. Trên các website khác nhau thì có các thuộc tính khác nhau.

Ví dụ: E = {“Tour Du Lịch Hà Nội – Nha Trang 4 Ngày 3 Đêm”, “Thông tin tour Mã tour: NTH”, “Thời gian: 4 ngày 3 đêm”, “Điểm xuất phát Hà Nội”, “Điểm thăm quan: Nha Trang – VINPEARL LAND – Hòn Mun, Hòn Miếu...”, “Giá tour 3.350.000 VNĐ”}. Qua các thông tin trên ta có thể hiểu rằng: có một tour tham quan các điểm Nha Trang – VINPEARL LAND – Hòn Mun, Hòn Miếu..., trong 4 ngày 3 đêm, điểm xuất phát từ Hà Nội và có giá là 3.350.000 đồng.

3.1.2. Ý tưởng giải quyết

Để giải quyết bài toán trên, luận văn sử dụng phương pháp kết hợp phân tích mã HTML và xây dựng một tập luật để trích chọn ra các thông tin du lịch trong bài viết. Chi tiết phương pháp này sẽ được trình bày trong phần tiếp theo của luận văn.

3.2. Phương pháp giải quyết bài toán

Trong Chương 2 luận văn đã trình bày một số phương pháp cơ bản để trích chọn thông tin. Các phương pháp đó bao gồm: Trích chọn thông tin dựa trên phân tích mã HTML thành cây DOM, sử dụng luật (tri thức), trích chọn

thông tin dựa trên học máy (dữ liệu) và phương pháp kết hợp giữa phân tích mã HTML và luật. Phần này sẽ tiếp tục phát triển ý tưởng kết hợp phân tích mã HTML và luật cho bài toán trích chọn thông tin du lịch trong văn bản tiếng Việt.

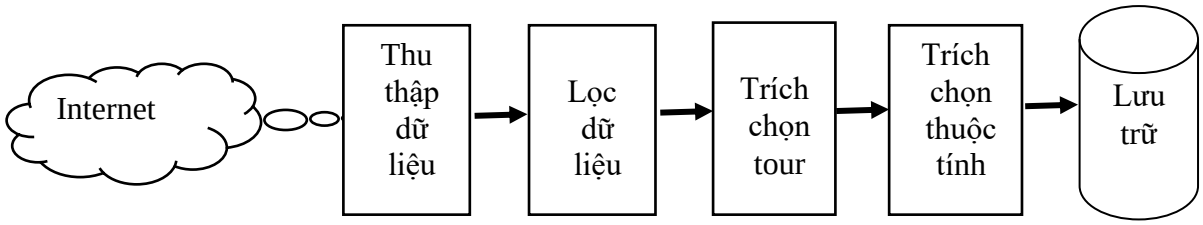
Khi thu thập dữ liệu từ Internet, ta gặp phải một vấn đề như sau: mặc dù đã lựa chọn các bài viết trên các website du lịch, nhưng không phải tất cả các bài viết đó đều chứa thông tin về các tour du lịch. Có thể đó là các bài viết giới thiệu về một danh lam thắng cảnh, những bài giải thích tên gọi của một địa danh hay những phong tục đặc trưng của một vùng miền … Từ đó hình thành một nhiệm vụ là phải giảm số lượng các bài viết trước khi đưa vào trích chọn các tour du lịch. Để làm được việc đó, tác giả đã xây dựng một tập luật để lọc ra những bài viết chứa thông tin về các tour du lịch. Trong các bộ trích chọn cần phải lấy được những phần văn bản chứa thông tin mà thông thường các thông tin này đều được đặt cố định trong một thẻ HTML, tuy nhiên việc ta cần làm là xác định được thông tin nằm ở đâu, bắt đầu lấy thông tin từ chỗ nào (ta dùng luật để làm việc này). Vậy để trích chọn được đúng các thông tin, tác giả sẽ sử dụng phương pháp kết hợp phân tích mã HTML và luật để thực hiện.

Luận văn lựa chọn phương pháp trên mà không dùng phương pháp tiếp cận dựa trên học máy là bởi những lý do sau:

1) Sử dụng luật sẽ không mất thời gian tạo dữ liệu huấn luyện, do vậy ít tốn công sức hơn.

2) Thông tin ở những nguồn dữ liệu khác nhau sẽ có những đặc trưng khác nhau do vậy nếu sử dụng phương pháp học máy ta sẽ gặp khó khăn trong việc xây dựng bộ dữ liệu huấn luyện phổ quát.

Cụ thể, mô hình giải quyết bài toán như sau:



Hình 3.1. Mô hình bài toán trích chọn

Mô hình bài toán trích chọn thông tin bao gồm năm thành phần:

1) *Bộ thu thập dữ liệu* (Crawler): Có chức năng lấy các bài viết từ trên Internet về để phục vụ cho mục đích trích chọn thông tin, bao gồm cả các bài viết có chứa thông tin về các tour du lịch và các bài viết khác.

2) *Bộ lọc dữ liệu*: Có chức năng lọc ra các bài viết chứa thông tin về tour du lịch cần trích chọn.

3) *Bộ trích chọn tour*: Có chức năng trích ra đoạn văn bản chứa thông tin về tour du lịch từ các bài viết chứa thông tin về tour du lịch đã được lọc.

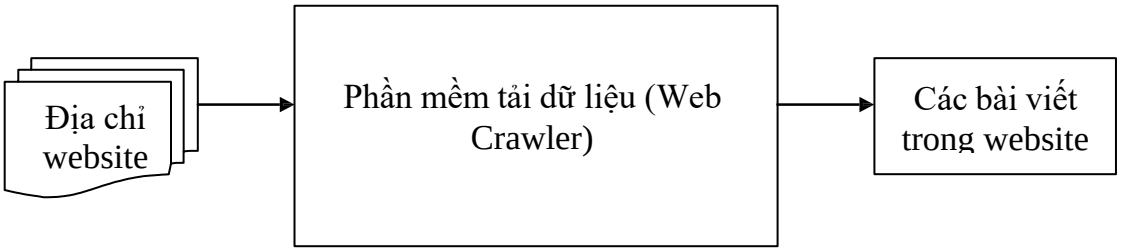
4) *Bộ trích chọn thuộc tính*: Có chức năng làm mịn dữ liệu, bỏ đi các thông tin không cần thiết và trích ra các thuộc tính cụ thể của một tour du lịch.

5) *Bộ lưu trữ*: Có chức năng lưu trữ các thông tin vừa trích chọn vào cơ sở dữ liệu phục vụ cho các mục đích thống kê, báo cáo.

3.2.1. Bộ thu thập dữ liệu

Chức năng chính của bộ thu thập dữ liệu là lấy các bài viết từ các Website về du lịch. Bộ thu thập dữ liệu sử dụng mô đun tải dữ liệu (crawler) để lấy trang web chứa thông tin du lịch từ một website cụ thể.

Mô hình làm việc của bộ thu thập dữ liệu như sau:



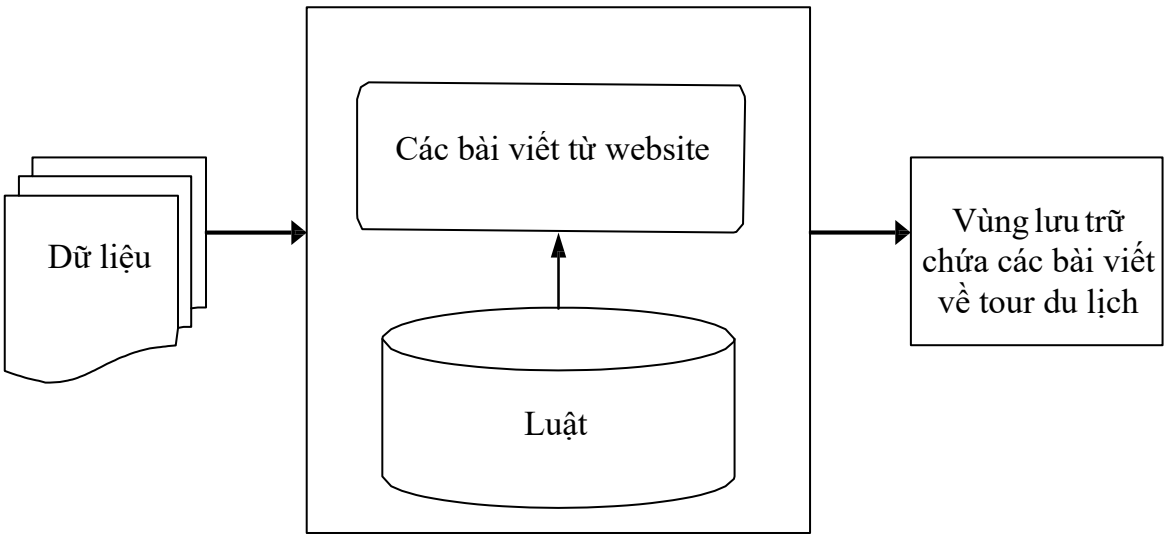
Hình3.2. Mô hình làm việc của bộ thu thập dữ liệu

3.2.2. Bộ lọc dữ liệu

Bộ lọc dữ liệu có chức năng chọn lọc các bài viết chứa các thông tin về các tour du lịch để sử dụng cho bộ trích chọn các tour du lịch [11].

Khi sử dụng mô đun tải dữ liệu để lấy dữ liệu về từ Internet (crawler) thì các bài viết của một website thường được lưu vào thư mục có tên là tên của website đó. Trong thư mục đó thì ngoài các bài viết chứa thông tin về tour du lịch, còn có các bài viết về các vấn đề khác như: bài viết giới thiệu về một địa điểm du lịch, bài viết giới thiệu về các món ăn đặc trưng của từng vùng miền hay những thông tin về nạn chặt chém du khác ở những điểm thăm quan … Khi đó bộ lọc dữ liệu sẽ làm nhiệm vụ kiểm tra những bài viết trong thư mục, chuyển những bài viết có chứa các thông tin về tour du lịch sang vùng lưu trữ để bộ trích chọn làm việc. Việc lọc thông tin trong các trang web dạng này dựa vào cấu trúc của trang web. Luận văn sử dụng thư viện JsoupParser và xây dựng một bộ lọc để thực hiện công việc này. Các bài viết được lựa chọn sẽ được ghi vào vùng lưu trữ để làm đầu vào cho bước xử lý sau.

Mô hình chung của bộ lọc dữ liệu như sau:



Hình 3.3. Mô hình làm việc của bộ lọc dữ liệu

Sau quá trình nghiên cứu dữ liệu tác giả thấy đặc điểm sau: 80% các bài viết mà tiêu đề bắt đầu bằng các từ khóa là "Tour" hoặc "Du lịch" đều là các bài viết chứa thông tin cần trích chọn, các bài viết mà tiêu đề không chứa từ khóa "Tour" hoặc "Du lịch" nếu nội dung trong thẻ div chứa một trong các từ khóa như: "Thời gian", "Giá tour", "Lịch trình", "Phương tiện", "Mã tour" thì đều là các bài viết cần trích chọn. Từ nghiên cứu trên ta xây dựng các luật cho bộ lọc dữ liệu như sau:

1) Những bài viết mà thẻ title bắt đầu bằng từ khóa "Tour" hoặc "Du lịch"

2) Những bài viết mà thẻ div chứa một trong các tiền tố "Thời gian", "Giá tour", "Lịch trình", "Phương tiện", "Mã tour", "Điểm khởi hành".

Thuật toán thực hiện cho bộ lọc dữ liệu được xây dựng như sau:

Thuật toán: Lọc các bài viết chứa thông tin về tour du lịch.

Đầu vào: Tập bài viết D dạng HTML

Đầu ra: Các bài viết chứa thông tin về tour du lịch cần trích chọn.

Phương pháp:

For each file in D

{

1. Tạo thể hiện của đối tượng HTMLDocument từ file;

2. Nội dung kiểm tra = Nội dung trong thẻ title; Nội dung trong thẻ div;

3. Dùng các luật trong tập luật để kiểm tra;

4. Nếu thỏa mãn thì chuyển file sang vùng lưu trữ các bài viết chứa thông tin về tour du lịch;

}

3.2.3. Bộ trích chọn tour

Mục tiêu của bộ trích chọn tour là dựa trên phân tích mã HTML kết hợp với các luật thích hợp để lấy đoạn văn bản chứa thông tin về tour du lịch đưa sang bộ trích chọn thuộc tính.

Một trang web du lịch không chỉ chứa các thông tin về một tour du lịch mà còn chứa các thông tin khác như các quảng cáo, các liên kết... Các thông tin du lịch trong một website tùy thuộc vào từng website khác nhau sẽ được lưu trữ trong các thẻ khác nhau. Trích chọn thông tin du lịch có trong các tài liệu HTML dạng này dựa vào kỹ thuật phân tích trích chọn thông tin từ tài liệu HTML. Việc xử lý văn bản HTML và trích chọn các phần tử trong văn bản HTML có thể thực hiện được bằng cách sử dụng biểu thức chính quy (regular expression) hoặc các công cụ phân tích tài liệu HTML còn gọi là các "HTML Parser".

Sau khi tìm hiểu một số công cụ phân tích tài liệu HTML, Giải pháp thực hiện của tác giả dựa trên phương pháp bóc tách nội dung nhờ vào phân tích mã HTML theo bộ mã nguồn JsoupParser để tạo thành cây Document Tree và các luật cụ thể được xây dựng bên dưới để lấy thông tin.

Thuật toán: Trích chọn đoạn văn bản chứa thông tin về tour du lịch

Đầu vào: Tập tài liệu D dạng HTML chứa thông tin về các tour du lịch.

Đầu ra: Văn bản T chứa thông tin về tour du lịch.

Phương pháp:

For each file in D

{

1. Tạo thể hiện của đối tượng HTMLDocument từ file.

2. Thông tin về Tour = Dùng luật để lấy đoạn văn bản chứa thông tin trong một thẻ HTML.

3. Ghi thông tin về Tour vào văn bản T.

}

Ví dụ: Tour Bắc Kinh - Thượng Hải - Hàng Châu - Tô Châu 8 ngày Du lịch Bắc Kinh - Thượng Hải - Hằng Châu - Tô Châu ngày bằng đường bay!

Bắc Kinh là thủ đô của Trung Quốc Mã Tour: TQ-086-BKTH8N Thời lượng: 8 ngày Giá: Call.

Luật dùng trong việc trích chọn tour như sau:

Những bài viết mà thẻ div, thẻ p chứa một trong các tiền tố "Thời gian", "Giá tour", "Lịch trình", "Phương tiện", "Mã tour", "Điểm khởi hành".

Sau khi đã trích chọn được các thông tin du lịch vào một văn bản dạng text, hệ thống sẽ chuyển văn bản đó sang bộ trích chọn thuộc tính để lấy ra từng thuộc tính cụ thể. Mỗi trang web sẽ có một cách trình bày riêng, do vậy số lượng thuộc tính trích chọn được cũng khác nhau.

Ví dụ: Có trang web sau khi trích chọn ta có 6 thuộc tính là: *Tên tour, điểm khởi hành, thời gian, phương tiện, điểm thăm quan, giá tour*. Còn có những trang web sau khi trích chọn ta có 5 thuộc tính là: *Tên tour, thời gian, ngày khởi hành, giá tour.*

3.2.4. Bộ trích chọn thuộc tính

Bộ trích chọn thuộc tính thực hiện hai chức năng chính như sau:

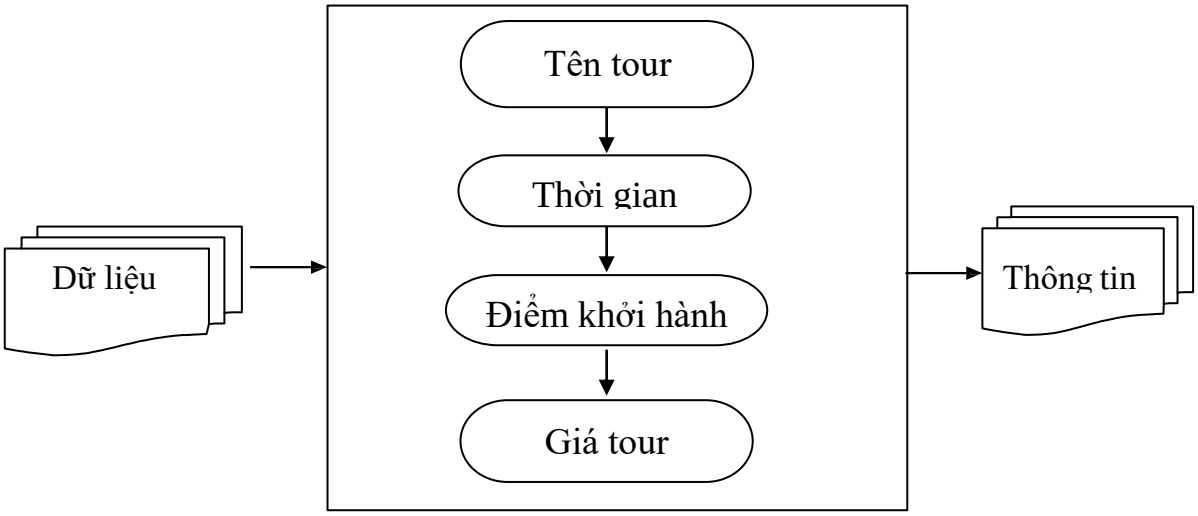
1) Làm sạch dữ liệu, loại bỏ đi các kí hiệu thừa và thông tin không cần thiết như: các thẻ HTML, thông tin quảng cáo, các đoạn giới thiệu về địa điểm du lịch …

2) Sử dụng các luật trích chọn để trích ra các thuộc tính cụ thể

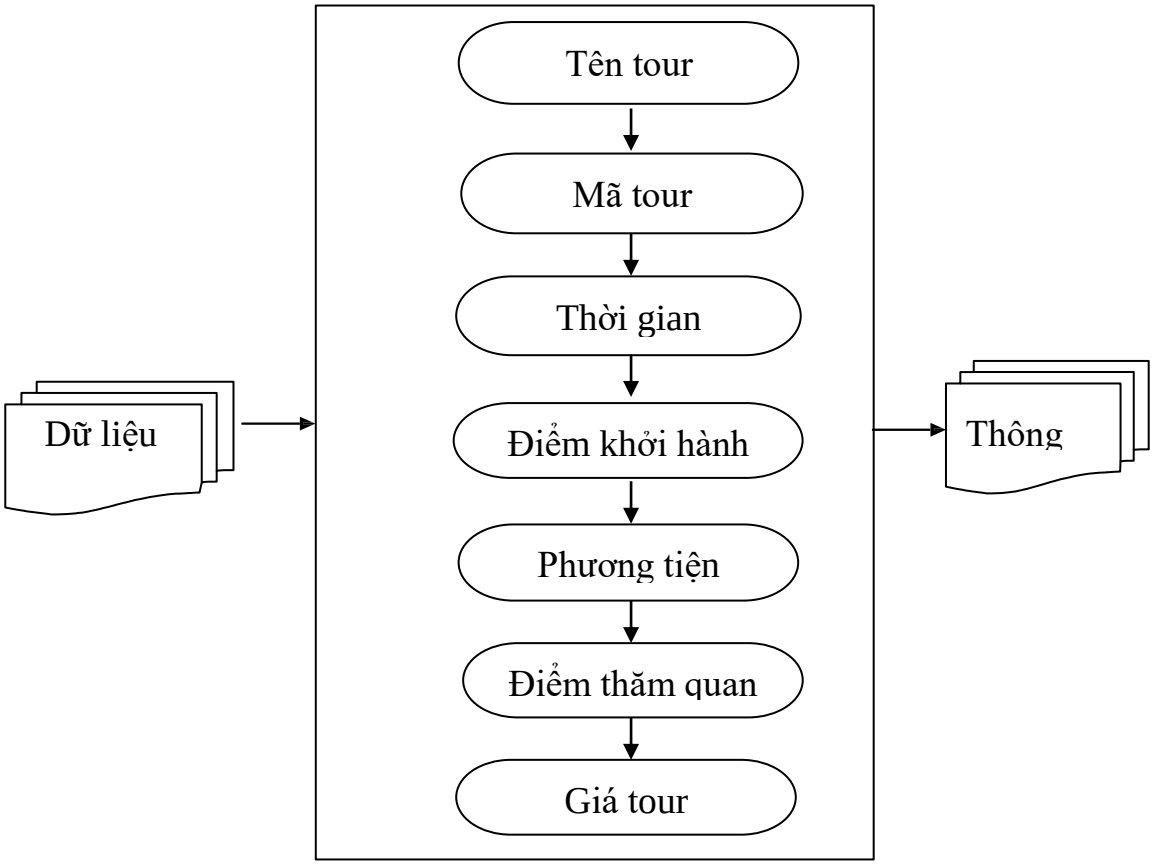
Sau khi trích chọn thuộc tính xong, các thuộc tính về tour du lịch sẽ được đưa vào một văn bản dạng text, hệ thống sẽ lưu các thông tin này vào cơ sở dữ liệu về các tour du lịch. Mỗi một tour du lịch có cấu trúc gồm tên tour, thông tin chi tiết về tour. Tùy thuộc vào từng trang web khác nhau mà thông tin chi tiết này có số lượng thông tin khác nhau (số thuộc tính khác nhau).

Luận văn tiến hành khảo sát các thông tin chi tiết của các tour du lịch ở các website http://www.dulichnamchau.vn; http://www.dulichnetviet.com.vn; http://www.dreamtravel.vn;http://www.dulichhn.com; http://dulichachau.com.vn;

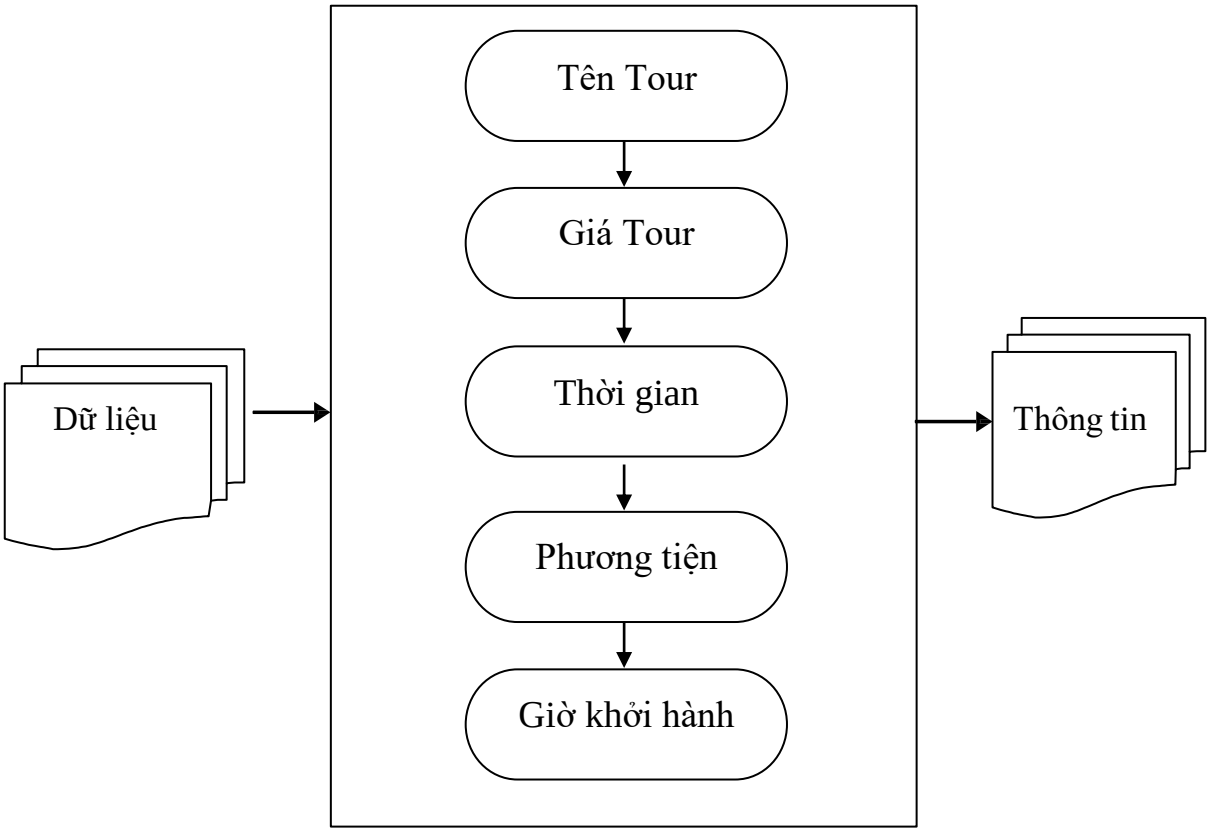
http://dulichviet.com; http://dulichmienbac.com. Kết quả khảo sát về các thông tin chi tiết của các tour du lịch như sau:



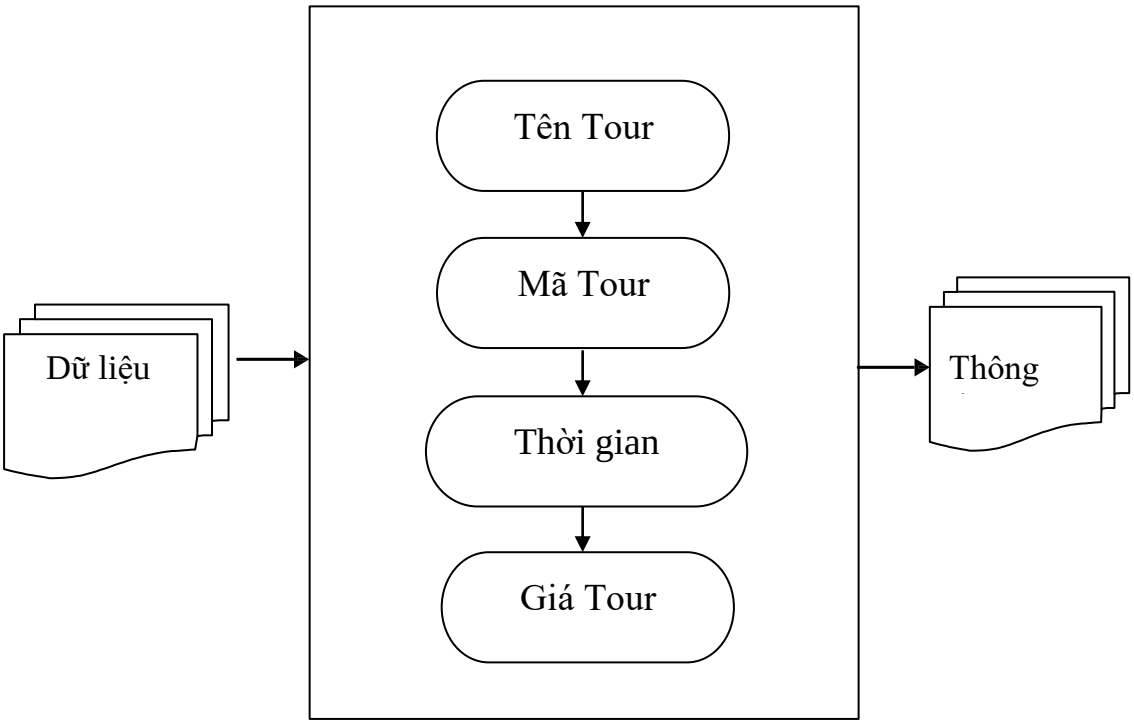
Hình 3.4. Các thông tin chi tiết về tour của website Du lịch Dấu Chân



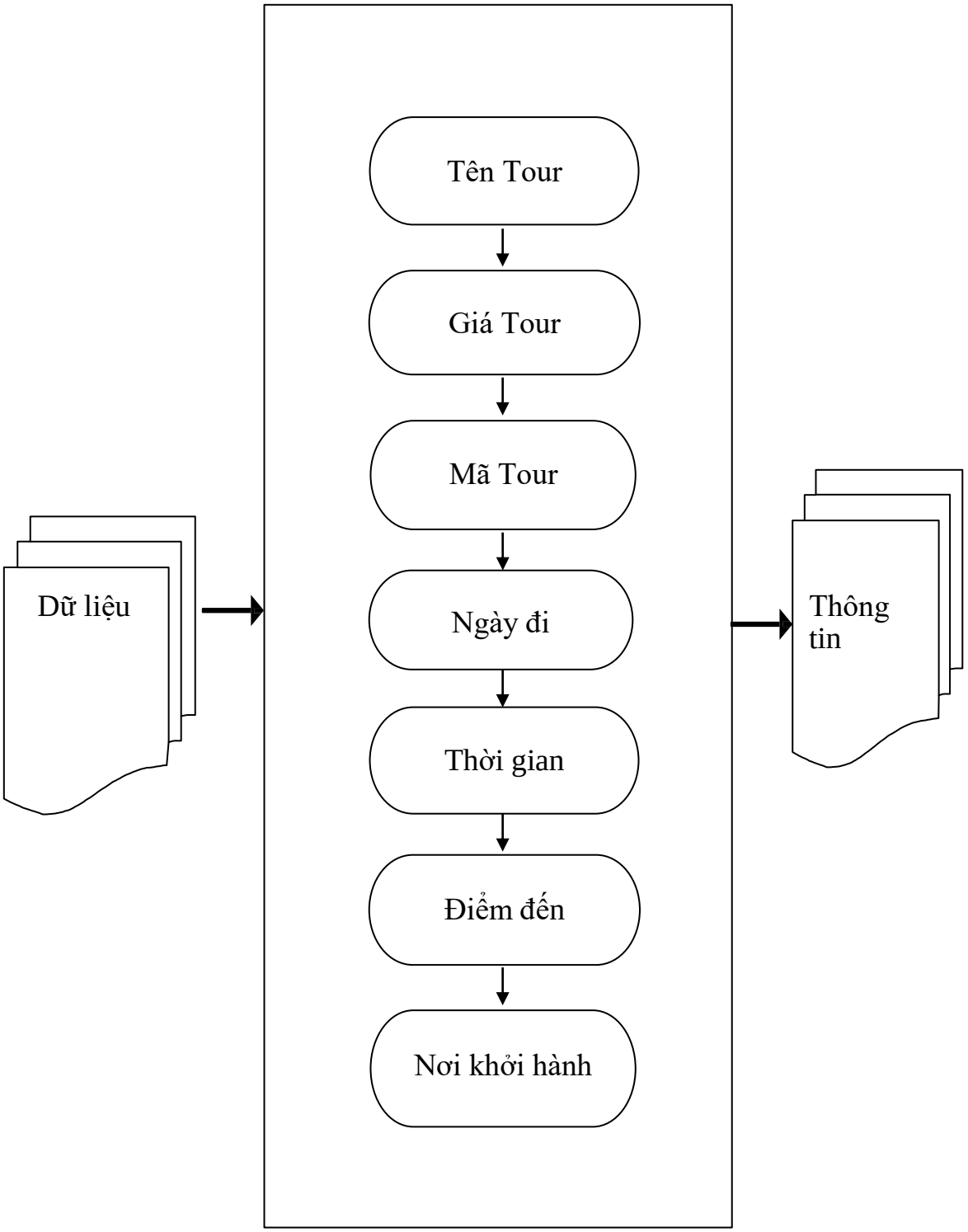
Hình 3.5. Các thông tin chi tiết về tour của website Du lịch Năm Châu



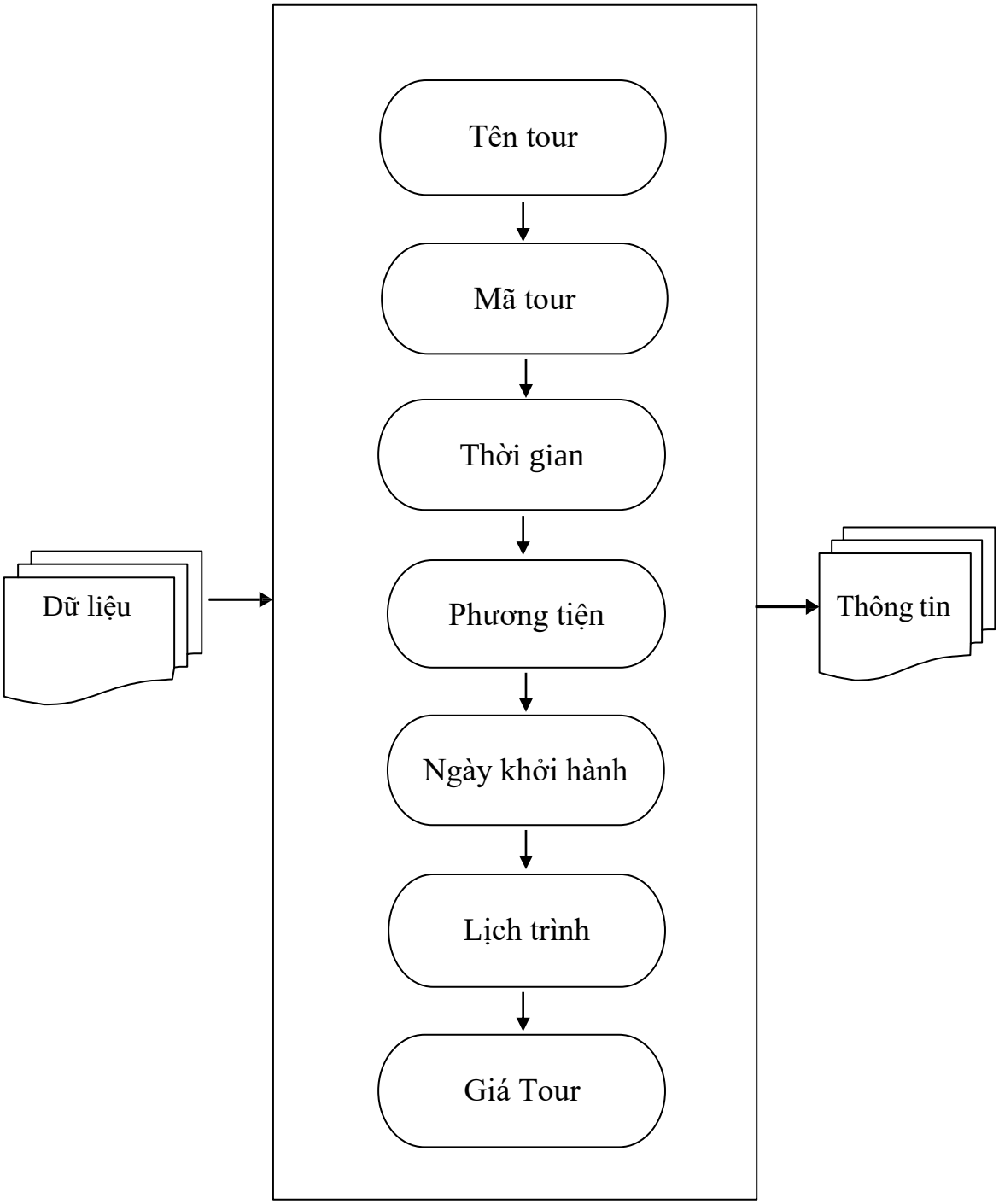
Hình 3.6. Các thông tin chi tiết về tour của website Du lịch Quốc tế Nét Việt



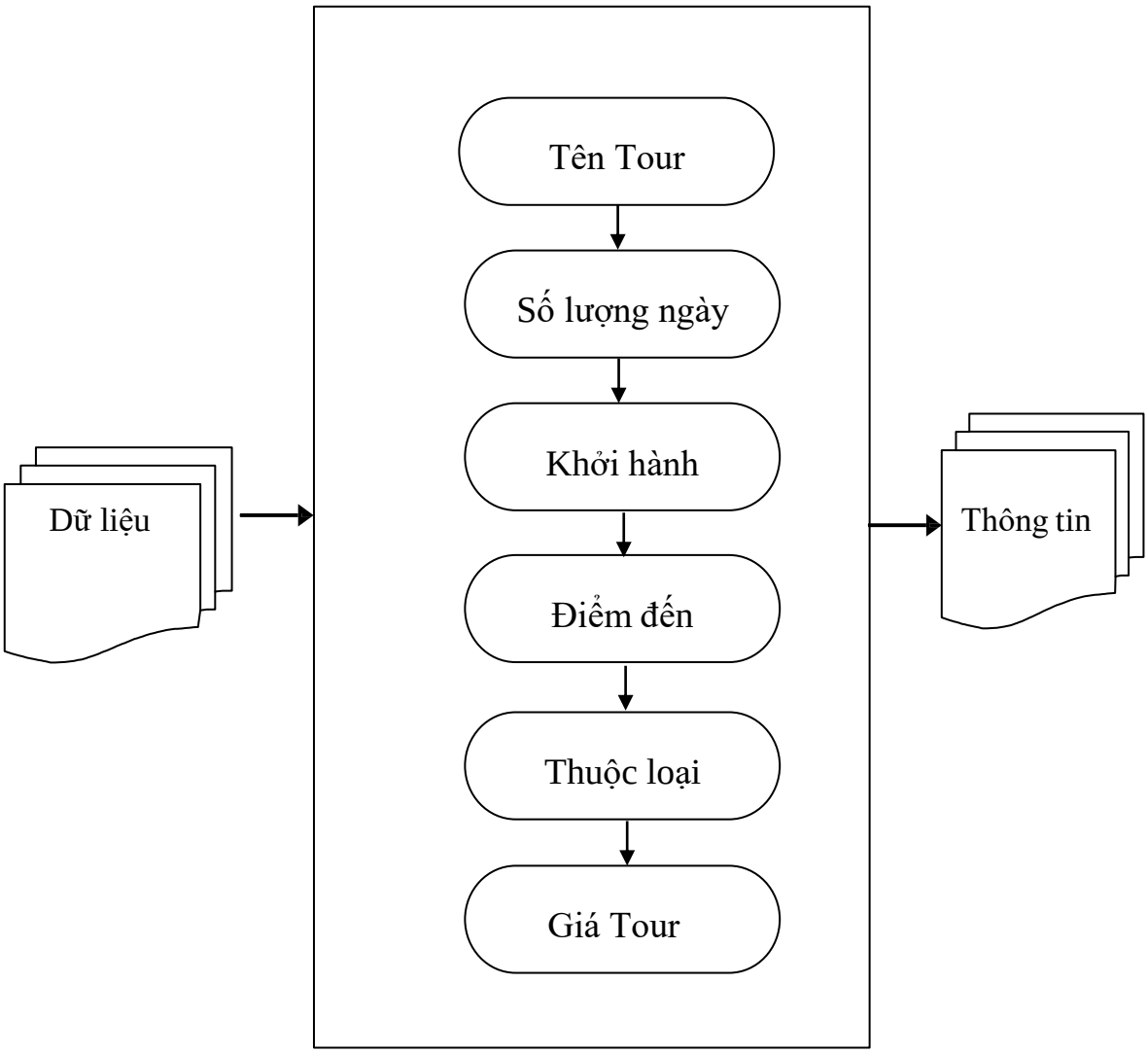
Hình 3.7. Các thông tin chi tiết về tour của website Du lịch AMI TOUR



Hình 3.8. Các thông tin chi tiết về tour của website Du lịch Giấc Mơ Việt



Hình 3.9. Các thông tin chi tiết về tour của website Du lịch Việt



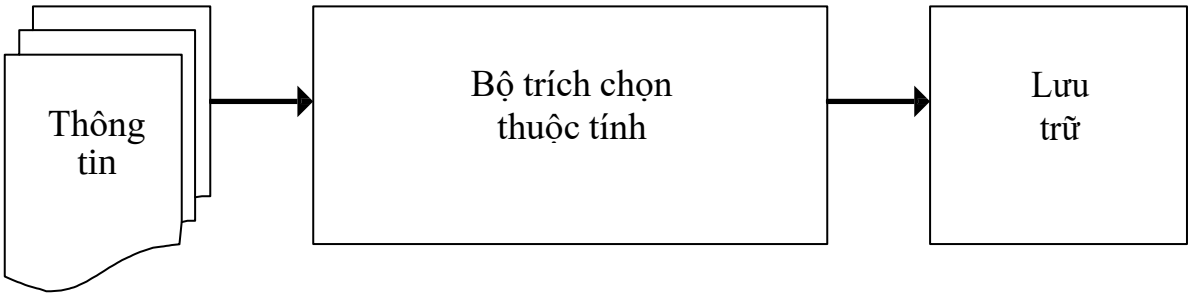
Hình 3.10. Các thông tin chi tiết về tour của website Du lịch Á Châu

Ví dụ 1: {"Tour: Hà Nội – Nha Trang – Đà Lạt – 5 Ngày 4 Đêm", "Thông tin tour: Mã tour: TN-NT4D06", "Thời gian: 5 ngày 4 đêm", "Điểm xuất phát: Hà Nội", "Phương tiện: Máy bay, ô tô", "Điểm thăm quan: Vịnh Nha Trang, Dốc Lết, Suối khoáng Tháp Bà, Khu du lịch Vinpearl Land, Chùa Long Sơn, Tháp Bà Ponaga, Đà Lạt", "Giá tour 3.990.000 VNĐ"}.

Ví dụ 2: {"Tour du lịch khám phá vẻ đẹp các tỉnh Tây Bắc", "Thông tin tour: Số lượng ngày: 6 ngày 5 đêm", "Khởi hành: Hằng ngày", "Điểm đến:

Hà Nội, Mộc Châu, Sơn La, Điện Biên, Sapa, đền Hùng, Phú Thọ, Hà Nội", "Thuộc loại: Du lịch"}

Mô hình làm việc của bộ trích chọn thuộc tính như sau:



Hình 3.11. Mô hình làm việc của bộ trích chọn thuộc tính

Để trích chọn chính xác các thuộc tính trong một tour du lịch, ta xây dựng bộ luật như sau:

3.2.4.1. Thông tin về tour

Tên tour thường ở một trong hai dạng như sau:

Dạng 1: **TÊN TOUR = <TIỀN TỐ> + <THÔNG TIN>**

Dạng 2: **TÊN TOUR** là danh sách các địa danh phân tách nhau bởi dấu " - ".

Trong đó: Tiền tố: "Du lịch", "Tour", "Tour Du lịch"

Ví dụ 1: Tour thăm quan Mỹ Tho, Bạc Liêu, Cà Mau, Sóc Trăng, Cần Thơ, 4 Ngày 3 Đêm.

Ví dụ 2: Tour Du Lịch: Đà Nẵng – Bà Nà – Hội An – Huế – Động Thiên Đường – Vũng Chùa (5N4Đ).

Ví dụ 3: Du Lịch Hà Nội – Hạ Long (2 Ngày 1 Đêm – Du Thuyền 4 Sao).

3.2.4.2. Thông tin về thời gian

THỜI GIAN = <TIỀN TỐ> + <ĐỊNH DẠNG> + <HẬU TỐ>

Trong đó:

Tiền tố: "Thời gian", "Thời lượng", "Số lượng ngày"

Định dạng: Bao gồm các ký tự {0, 1, 2,…, 9, "\", "/", "N"}

Hậu tố: "Ngày", "Đêm", "N", "N/Đ"

Ví dụ: Thời gian: 5N/ 4Đ, Thời lượng: 4 ngày 3 đêm, Số lượng ngày: 3 ngày 2 đêm.

3.2.4.3. Thông tin về giá tour

GIÁ TOUR = <TIỀN TỐ> + <ĐỊNH DẠNG> + <HẬU TỐ>

Trong đó:

Tiền tố: "Giá tour", "Giá", "Giá từ", "Giá khuyến mãi", "Price"

Định dạng của giá: Dạng số, bao gồm các ký tự {0, 1, 2,…,9, ",", "."}

Hậu tố: "VNĐ", "VND/ KHÁCH", "Đ", "vnđ / khách", "VND"

Ví dụ: Giá tour: 4.200.000 VNĐ, Giá từ: 8,500,000 VND.

3.2.4.4. Thông tin về điểm khởi hành

ĐIỂM KHỞI HÀNH = <TIỀN TỐ> + <ĐỊA ĐIỂM>

Trong đó:

Tiền tố: "Điểm khởi hành", "Khởi hành từ", "Từ", "Khởi hành", "Giờ khởi hành", "Bắt đầu", "Xuất phát", "Điểm xuất phát", "Nơi khởi hành"

Địa điểm: Danh từ chỉ nơi chốn

Ví dụ: Điểm khởi hành: Hà Nội, Khởi hành từ: Đà Lạt, Điểm xuất phát: Sài Gòn, Nơi khởi hành: Đà Nẵng...

3.2.4.5. Thông tin về phương tiện

PHƯƠNG TIỆN = <TIỀN TỐ> + <PHƯƠNG TIỆN DI CHUYỂN>

Trong đó:

Tiền tố: "Phương tiện", "Di chuyển bằng", "Vận chuyển"

Phương tiện di chuyển: Tên một loại phương tiện giao thông

Ví dụ: Phương tiện: Ô tô hoặc máy bay, Di chuyển bằng: Du thuyền 4 sao, Vận chuyển: Máy bay Vietnam Airlines.

3.2.4.6. Thông tin về lịch trình

LỊCH TRÌNH = <TIỀN TỐ> + <CÁC ĐỊA DANH>

Trong đó:

Tiền tố: "Lịch trình", "Điểm thăm quan", "Nơi đến", "Đến", "Điểm đến", "Điểm dừng", "Hành trình"

Các địa danh: Tên các địa danh trong hành trình du lịch

Ví dụ: Lịch trình: New York - Washington DC - Los Angeles - Las Vegas; Điểm thăm quan: Vịnh Nha Trang,Suối khoáng Tháp Bà, Khu du lịch Vinpearl Land, Chùa Long Sơn, Tháp Bà Ponaga, Đà Lạt; Điểm đến: Hà Nội, Phan Thiết, Sài Gòn, Củ Chi, Mekong.

Chương 4
THỬ NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

Trong chương này, tác giả sẽ trình bày về môi trường, công cụ cũng thử nghiệm và đánh giá kết quả. Tác giả cũng trình bày một số bàn luận về kết quả của phương pháp, cũng như hướng phát triển trong tương lai.

4.1. Bài toán thử nghiệm

Thử nghiệm hệ thống trích chọn thông tin về các tour du lịch từ các website sau:

1) http://www.dulichnamchau.vn
2) http://www.dulichnetviet.com.vn
3) http://www.dreamtravel.vn
4) http://www.dulichhn.com
5) http://dulichachau.com.vn
6) http://dulichviet.com
7) http://dulichmienbac.com

Như vậy, *đầu vào* của bài toán là 07 website về du lịch, *đầu ra* là thông tin về các tour du lịch (bao gồm các tour du lịch và các thuộc tính) được lưu trong cơ sở dữ liệu phục vụ kết xuất các báo cáo, thống kê.

4.2. Môi trường và các công cụ thử nghiệm

4.2.1. Môi trường thử nghiệm

Bảng 4.1. Cấu hình hệ thống thử nghiêm

Thành phần	Chỉ số
CPU	Core i3 – 3240 3.4 (GHz)
RAM	4GB
OS	Windows 7 Ultimate
HDD	500GB

4.2.2. Công cụ phần mềm sử dụng để thử nghiệm

1) Các công cụ có sẵn, môi trường lập trình

Bảng 4.2. Công cụ phần mềm có sẵn

TT	Tên công cụ	Chức năng	Nguồn
1	Teleport Pro	Công cụ Crawler tải dữ liệu từ các website	http://teleport-pro.en.softonic.com
2	Eclipse Standard/Kepler Release	Tạo môi trường để viết chương trình	http://eclipse.org/eclipse
3	JsoupParser	Bộ công phân tích mã HTML	http://jsoup.org/apidocs/org

2) Công cụ phần mềm được lập trình

Project "TravelFeatureExtractor" thực hiện các công việc liên quan đến trích chọn thông tin về các tour du lịch được nêu trong các bài viết về lĩnh vực du lịch, bao gồm các gói:

FilterData: Lọc ra các bài viết có chứa thông tin về các tour du lịch sang vùng lưu trữ để trích chọn thông tin về các tour du lịch.

TourExtraction: Trích ra đoạn văn bản chứa thông tin về tour du lịch.

AttributeExtraction: Trích ra các thuộc tính của các tour du lịch và lưu trữ vào cơ sở dữ liệu.

4.3. Xây dựng cơ sở dữ liệu

Để lưu trữ dữ liệu phục vụ quá trình trích chọn thông tin, chúng tôi sử dụng các bảng lưu trữ các website, lưu trữ các bài viết lấy từ các website qua công cụ crawler, lưu trữ các tour du lịch, lưu trữ các thuộc tính của các tour du lịch phục vụ kết xuất báo cáo, thống kê.

1) Bảng dữ liệu lưu trữ các website

- Tên bảng: WEBSITE
- Cấu trúc:

STT	Tên trường	Kiểu dữ liệu	Mô tả
1	ID	N(05)	Mã hiệu
2	TenGoi	TEXT	Tên website

2) Bảng dữ liệu lưu trữ các bài viết chứa thông tin về các tour du lịch sau khi qua bộ lọc dữ liệu

- Tên bảng: BAIVIET
- Cấu trúc:

STT	Tên trường	Kiểu dữ liệu	Mô tả
1	ID	N(05)	Mã hiệu
2	Website	TEXT	Tên website
3	NoiDung	TEXT	Nội dung bài viết

3) Bảng dữ liệu lưu trữ các tour du lịch được trích lọc từ các bài viết

- Tên bảng: TOUR
- Cấu trúc:

STT	Tên trường	Kiểu dữ liệu	Mô tả
1	ID	C(05)	Mã tour
2	Ten_Tour	TEXT	Tên tour du lịch
3	Mô tả	TEXT	Mô tả chi tiết
3	Website	TEXT	Tên website

4) Bảng dữ liệu lưu trữ chi tiết các thuộc tính của các tour du lịch. Như đã trình bày ở mục 3.2.4 về bộ trích chọn thuộc tính, số lượng thuộc tính của các tour du lịch ở các website là khác nhau. Do đó, bảng dữ liệu phải được thiết kế sao cho chứa tất cả các thuộc tính của các tour du lịch.

- Tên bảng: THUOCTINH_TOUR
- Cấu trúc:

STT	Tên trường	Kiểu dữ liệu	Mô tả
1	**ID**	C(05)	Mã tour
2	**Ten_Tour**	TEXT	Tên tour du lịch
3	**Mô tả**	TEXT	Mô tả chi tiết
4	**Website**	TEXT	Tên website
5	**Dia_Danh**	TEXT	Địa danh du lịch
6	**Diem_Khoi_Hanh**	TEXT	Điểm khởi hành
7	**Ngay_Khoi_Hanh**	DATE	Ngày khởi hành
8	**Thoi_Gian**	DATE	Thời gian
9	**Phuong_Tien**	TEXT	Phương tiện
10	**Diem_Tham_Quan**	TEXT	Điểm thăm quan
11	**Khach_San**	TEXT	Khách sạn
12	**Lich_Trinh**	TEXT	Lịch trình
13	**Gia_Tour**	N(20)	Giá tour

4.4. Thử nghiệm quy trình trích chọn tour du lịch

4.4.1. Thu thập dữ liệu (Web Crawler)

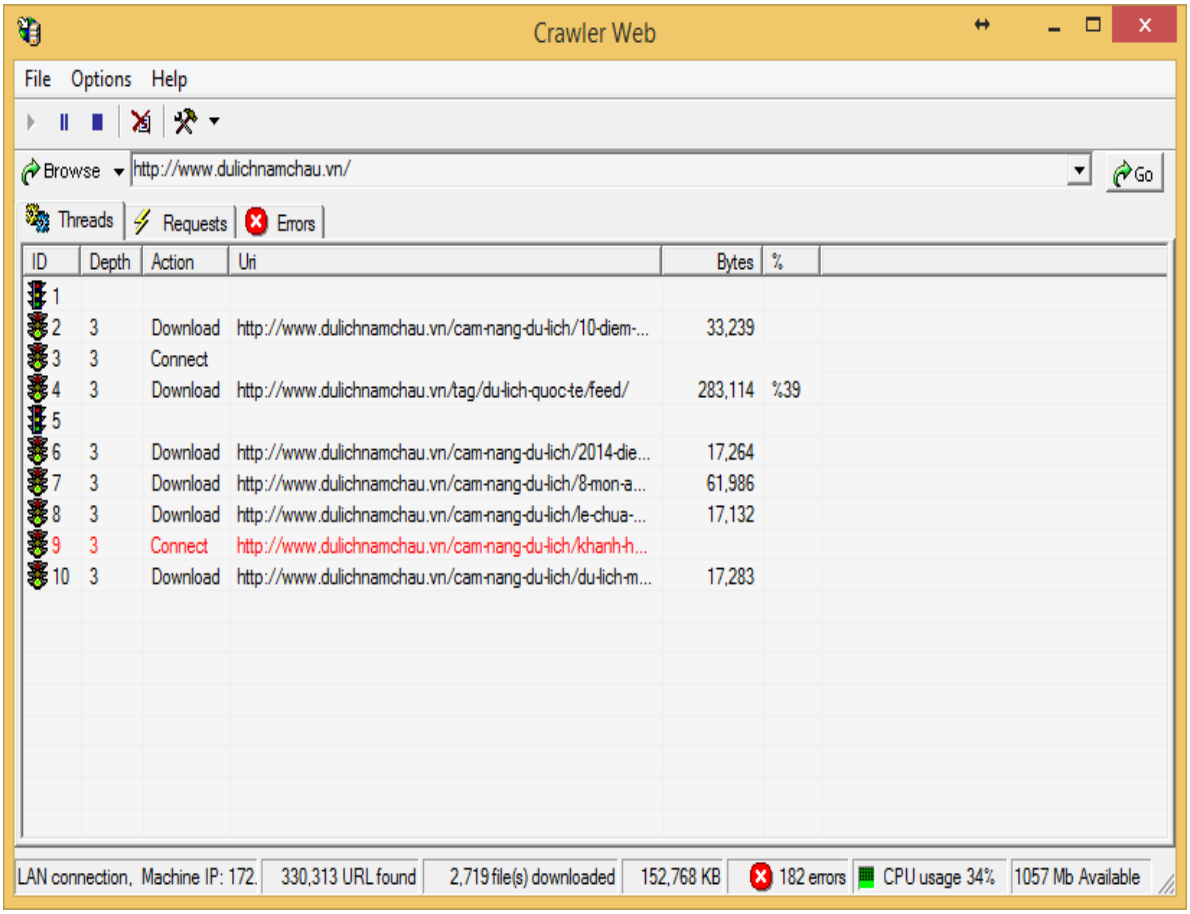
Dữ liệu trong luận văn được thu thập từ các website:

1) http://www.dulichnamchau.vn
2) http://www.dulichnetviet.com.vn
3) http://www.dreamtravel.vn
4) http://www.dulichhn.com
5) http://dulichachau.com.vn
6) http://dulichviet.com
7) http://dulichmienbac.com

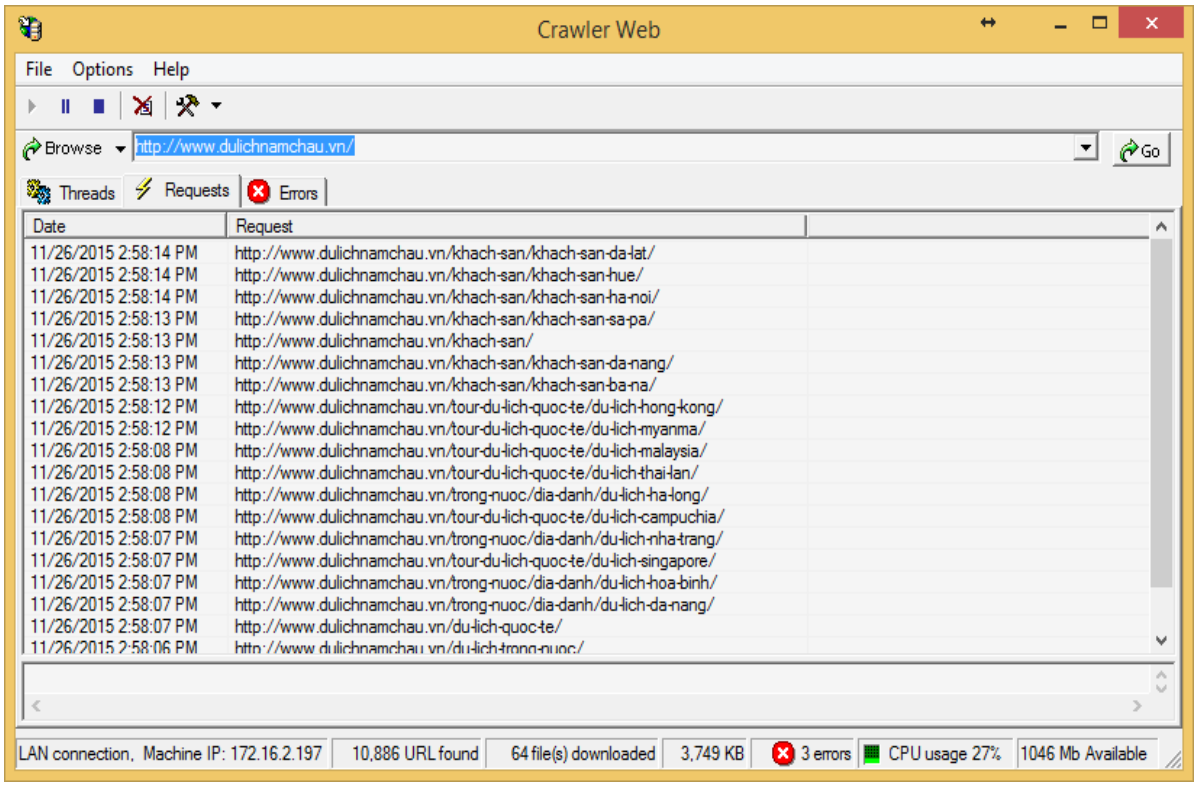
Luận văn lựa chọn những website trên là để đảm bảo tính toàn diện của dữ liệu, có những website, thông tin chi tiết về các tour du lịch chứa đầy đủ các thông tin như tên tour, thời gian, ngày khởi hành, điểm khởi hành, giá tour, lịch trình, như: dulichnamchau.vn, dreamtravel.vn, dulichviet.vn, còn có những website chỉ chứa các thông tin đặc trưng như tên tour, thời gian, điểm khởi hành, giá tour như: dulichhn.com, dulichmienbac.com.

Việc thu thập dữ liệu (web crawler) sẽ được thực hiện bằng phần mềm Teleport Pro, phần mềm này sẽ lấy về 500 bài viết từ các website trên, như vậy sau khi thu thập dữ liệu ta có 3500 bài viết từ các website du lịch ở trên.

Hình 4.1, Hình 4.2 là giao diện phần mềm crawler thu thập dữ liệu từ trang www.dulichnamchau.vn.



Hình 4.1. Thu thập dữ liệu từ trang www.dulichnamchau.vn.



Hình 4.2. Quá trình thu thập dữ liệu từ trang [www.dulichnamchau.vn.](http://www.dulichnamchau.vn)

4.4.2. Lọc dữ liệu

Dữ liệu sau khi được thu thập về dưới dạng HTML sẽ được đưa qua bộ lọc dữ liệu để lấy ra các bài viết liên quan tới lĩnh vực du lịch.

Bộ lọc dữ liệu là chức năng đầu tiên trong quá trình trích chọn thông tin du lịch, làm nhiệm vụ lọc các bài viết được lấy từ bộ thu thập dữ liệu. Như đã trình bày ở chương 3, chức năng này được thực hiện dựa trên các luật như sau:

1) Những bài viết mà thẻ title bắt đầu bằng từ khóa "Tour" hoặc "Du lịch".

2) Những bài viết mà thẻ div chứa một trong các tiền tố "Thời gian", "Giá tour", "Lịch trình", "Phương tiện", "Mã tour", "Điểm khởi hành".

Kết quả thực hiện cho thấy, đầu vào của bộ lọc dữ liệu là 3500 bài viết được thu thập từ 07 website ở trên (mỗi website lấy 500 bài viết), đầu ra là

1832 bài viết có chứa thông tin về các tour du lịch. Kết quả chi tiết trong Bảng 4.3 sau:

Bảng 4.3. Kết quả lọc các bài viết chứa thông tin về các tour du lịch

STT	Tên website	Số bài viết thu thập bởi crawler	Số bài viết chứa thông tin tour du lịch
1	Dulichnamchau	500	197
2	Dulichviet	500	351
3	Dulichachau	500	293
4	Dreamtravel	500	217
5	Dulichhn	500	226
6	Dulichmienbac	500	292
7	Dulichnetviet	500	256
	Tổng số	**3500**	**1832**

Hình 4.3 là giao diện kết quả lọc các bài viết có chứa thông tin về các tour du lịch.

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" lang="vi">
<head profile="http://gmpg.org/xfn/11">
<meta http-equiv="Content-Type" content="text/html; charset=UTF-8" />
<meta name="viewport" content="width=device-width, user-scalable=yes">
<meta http-equiv="content-language" content="vi, en">
<title>Kyoto - thành phố cổ kính đúng nghĩa, không có các tòa</title>
<meta name="description" content="Du lịch Nhật Bản Mùa Thu : TOKYO - NÚI PHÚ SỸ - KYOTO - OSAKA (Lịch khởi hành tháng 10-2015)"/>
<meta name="keywords" content="Du lịch Nhật Bản Mùa Thu  TOKYO - NÚI PHÚ SỸ - KYOTO - OSAKA (Lịch khởi hành tháng 10-2015)" />
<meta name="title" content="Kyoto - thành phố cổ kính đúng nghĩa, không có các tòa" />
<meta name="y_key" content="Du lịch Nhật Bản Mùa Thu  TOKYO - NÚI PHÚ SỸ - KYOTO - OSAKA (Lịch khởi hành tháng 10-2015)" />
<meta name="author" content="dreamtravel">
<meta name="copyright" content="Copyright DreamTravel.vn - 2008">
<meta name="email" content="sales@dreamtravel.vn">
<meta http-equiv="Content-Language" content="vn">
<meta name="Charset" content="UTF-8">
<meta name="Rating" content="General">
<meta name="Distribution" content="Global">
<meta name="Robots" content="INDEX,FOLLOW">
<meta name="Revisit-after" content="1 Day">
<meta name="google-site-verification" content="O46MfqbQRbqHm4c-PysaYDbDB1gYSUKLqQaNrbf0XVM" />
<meta name="placename" content="Vietnam">
<META NAME="Search Engines" CONTENT="www.google.com, www.google.com.vn, www.google.co.uk, www.google.it, www.google.es, www.google.com.au, www.altaVista.com, www
www.infoseek.com, www.excite.com, www.hotbot.com, www.lycos.com, www.magellan.com, www.cnet.com, www.voila.com, www.google.fr, www.yahoo.fr, www.yahoo.com, www.a
www.msn.com, www.netscape.com, www.nomade.com, www.bing.com">
<meta http-equiv="audience" content="General">
<meta name="googlebot" content="INDEX,FOLLOW" />
<!-- END META TAG -->
<!-- Link -->
<link rel="shortcut icon" href="http://www.dreamtravel.vn/themes/default/images/root/favicon.ico" >
<link rel="stylesheet" type="text/css" href="http://www.dreamtravel.vn/themes/default/css/style.css" />
<link rel="stylesheet" type="text/css" href="http://www.dreamtravel.vn/themes/default/css/mobile.css" />
<link rel="stylesheet" type="text/css" href="http://www.dreamtravel.vn/themes/default/css/back2top.css" />

```

Hình 4.3. Kết quả lọc các bài viết chứa thông tin về các tour du lịch

4.4.3. Trích chọn các tour du lịch và các thuộc tính

Sau khi đã lọc được các bài viết chứa thông tin về các tour du lịch của 7 website, chức năng trích chọn các tour du lịch có nhiệm vụ trích chọn ra các đoạn văn bản chứa các tour du lịch để lưu trữ vào bảng TOUR trong cơ sở dữ liệu. Bảng TOUR gồm các thông tin: *Id, Mã tour, tên tour, mô tả, website*. Như đã trình bày ở chương 3, luật dùng trong việc trích chọn tour như sau:

Những bài viết mà thẻ div, thẻ p chứa một trong các tiền tố: "Thời gian", "Giá tour", "Lịch trình", "Phương tiện", "Mã tour", "Điểm khởi hành".

Sau khi đã trích chọn được các tour du lịch, công việc tiếp theo là trích chọn ra các thuộc tính của các tour du lịch và lưu trữ vào bảng cơ sở dữ liệu THUOCTINH_TOUR, bao gồm các thông tin: *Mã tour, tên tour, mô tả chi tiết, tên website, điểm khởi hành, ngày khởi hành, thời gian, phương tiện, điểm thăm quan, khách sạn, lịch trình, giá tour*. Như đã trình bày, với mỗi

tour du lịch thì các thuộc tính như trên sẽ không có đầy đủ dữ liệu vì phụ thuộc vào từng website.

Để tiến hành thực nghiệm, với mỗi website ở Bảng 4.3, tác giả lấy ngẫu nhiên 50 bài viết chứa thông tin về tour du lịch được lọc để thực hiện công cụ trích chọn các tour du lịch và trích chọn các thuộc tính của tour du lịch. Kết quả trích chọn được mô tả ở Bảng 4.4 sau đây:

Bảng 4.4. Kết quả trích chọn tour du lịch và trích chọn thuộc tính

STT	Tên website	Số bài viết chứa thông tin tour	Số tour được trích chọn	Số tour có thuộc tính được trích chọn
1	Dulichnamchau	50	47	44
2	Dulichviet	50	38	33
3	Dulichachau	50	45	42
4	Dreamtravel	50	43	41
5	Dulichhn	50	46	43
6	Dulichmienbac	50	34	32
7	Dulichnetviet	50	40	34

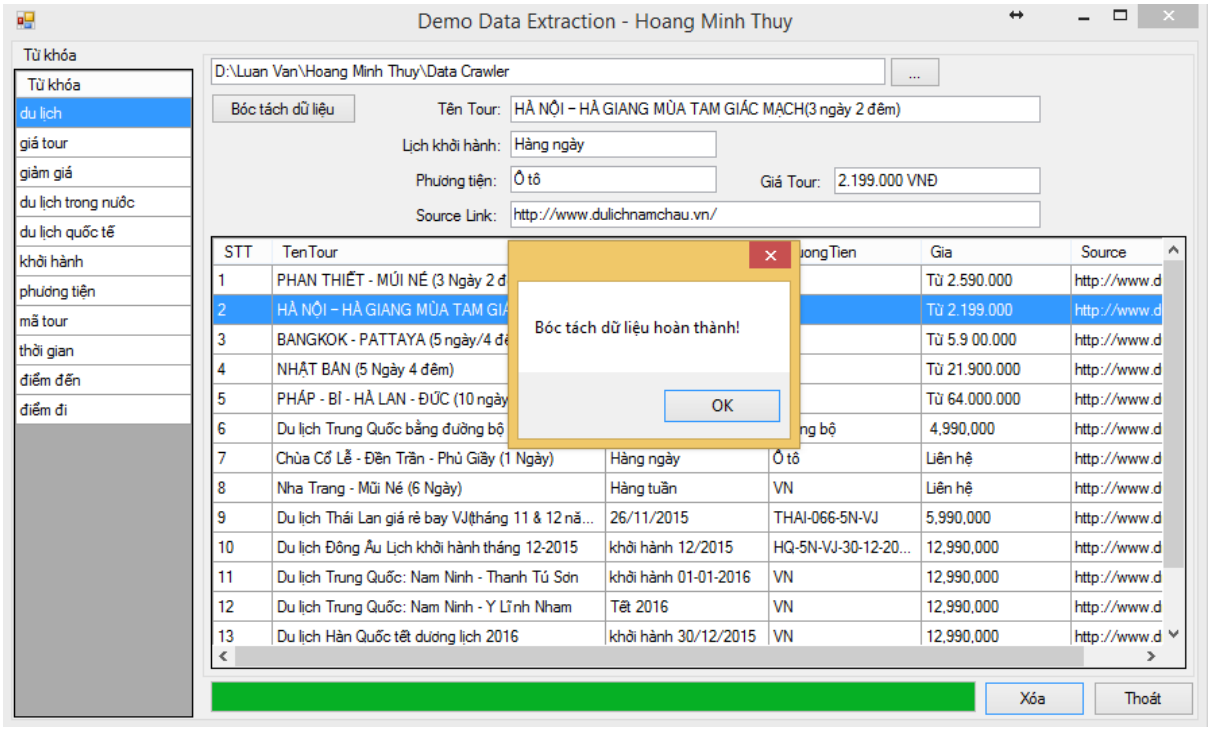
Kết quả thử nghiệm cho thấy, số tour được trích chọn nhỏ hơn số bài viết chứa thông tin về tour du lịch và số tour có thuộc tính được trích chọn nhỏ hơn số tour được trích chọn, nghĩa là một số tour được trích chọn không trích chọn được thuộc tính. Nguyên nhân này là do các lỗi, có thể là do bộ luật chưa bao hết các trường hợp, có thể do website. Vấn đề lỗi này sẽ được phân tích ở mục sau. Hơn nữa, ta thấy có sự khác nhau giữa số lượng tour được trích chọn trong mỗi website là do thiết kế của từng website. Có website thiết

kế theo kiểu List Page như website dulichnamchau, dulichviet, dulichachau, dulichmienbac, dulichnetviet có website thiết kế theo kiểu Detail Page như website dreamtravel, dulichhn.

Trong đó:

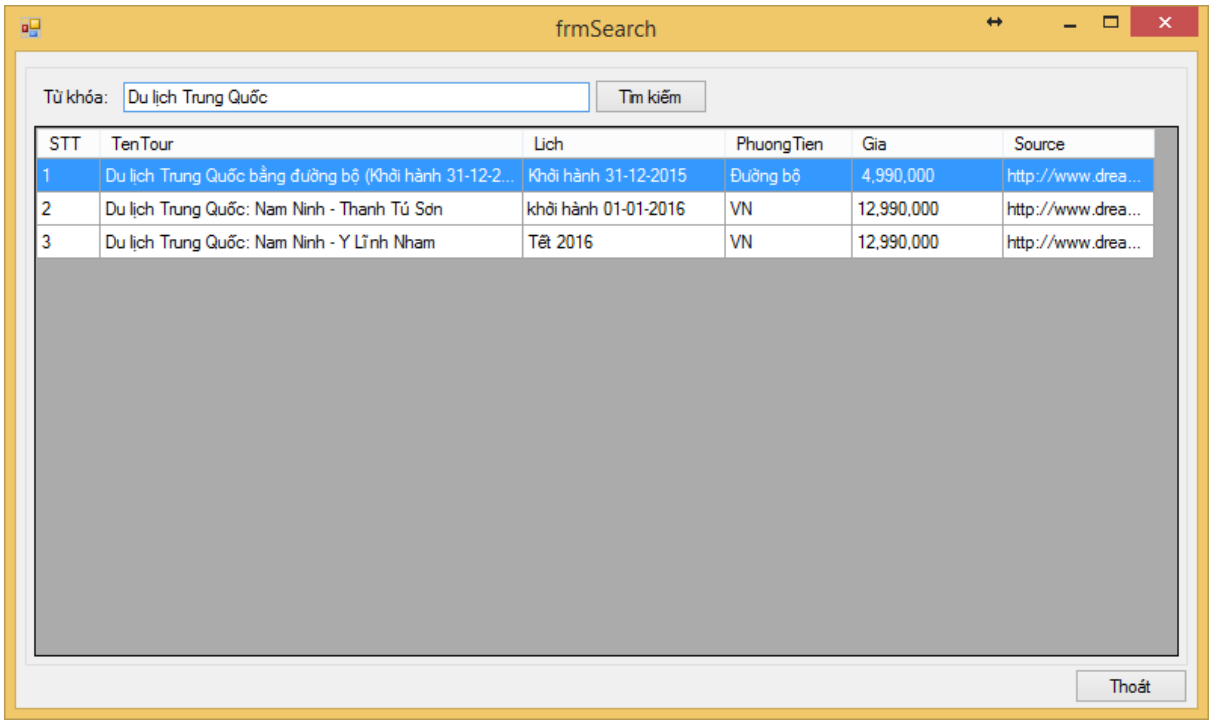
- List Page: là trang chứa một vài danh sách của các đối tượng. Có hai dạng trang list, đó là trang list bố trí theo chiều ngang hoặc chiều dọc.
- Detail Page: là trang chỉ giới thiệu một đối tượng đơn. Nó chứa đựng tất cả các thông tin về một tour du lịch như: tên tour, mã tour, thời gian, giá tour … [6].

Hình 4.4 là giao diện kết quả trích chọn tour du lịch và trích chọn thuộc tính.



Hình 4.4. Kết quả trích chọn các tour du lịch

Hình 4.5 là giao diện tra cứu các tour du lịch sau khi được trích chọn và lưu vào trong cơ sở dữ liệu.



Hình 4.5. Giao diện tra cứu tour du lịch

4.5. Phân tích lỗi

4.5.1. Phân tích lỗi của bộ lọc dữ liệu

Trong quá trình phát hiện thông tin về tour du lịch, kết quả trong Bảng 4.3 chỉ ra rằng bộ lọc dữ liệu hoạt động không tốt trong một số trường hợp. Kết quả phân tích cho thấy những trường hợp bộ lọc dữ liệu hoạt động không tốt là do luật lọc dữ liệu theo thẻ tiêu đề bắt đầu bằng "Tour" hoặc "Du lịch" nhưng trong một số trường hợp thông tin về tour lại ở dạng hình ảnh hoặc dạng lựa chọn (như Hình 4.6). Hoặc trong luật lọc theo thẻ div bằng các từ khóa như "Mã tour", "Thời gian", "Giá tour"… thì xảy ra lỗi do bài viết nói về các dịch vụ khác như cho thuê xe du lịch hay đặt vé máy bay… (như Hình 4.7).

Hình 4.6. Lỗi lọc dữ liệu khi thông tin ở dạng lựa chọn

Hình 4.7. Lỗi lọc dữ liệu khi không có thông tin về tour du lịch

4.5.2. Phân tích lỗi của quá trình trích chọn

Trong pha trích chọn thông tin thì khả năng trích chọn thông tin của trang Du Lịch AMI TOUR là thấp nhất, tác giả đã tìm hiểu nguyên nhân và thấy rằng nguyên nhân trang Du Lịch AMI TOUR cho kết quả trích chọn thấp là do có sự không đồng nhất giữa các bài viết về du lịch trên website này, dẫn đến bộ luật dùng cho website này không bao phủ được toàn bộ dữ liệu.

Ví dụ: Cùng là bài viết trên website dulichmienbac.com, nhưng có bài viết chỉ chứa thông tin là {tên tour, thời gian, giá}. Có bài viết lại chứa đầy đủ các thông tin như {tên tour, thời gian, khởi hành, giá tour, phương tiện, điện thoại, hotline, email}

Với các website khác, bộ trích chọn làm việc sai là do các bài viết bị sai chính tả nên không khớp với bộ luật mà tác giả xây dựng.

Ví dụ: Với luật xác định giá ta có:

GIÁ TOUR = <TIỀN TỐ> + <ĐỊNH DẠNG> + <HẬU TỐ>

Trong đó:

Tiền tố: “Giá tour:”, “Giá: ”, “Giá từ”, “Giá khuyến mãi”, “Price”

Định dạng của giá: Dạng số, bao gồm các ký tự {0, 1, 2,…, 9, “,”, “.”}

Hậu tố: “VNĐ”, “VND/ KHÁCH”, “Đ”, “vnđ / khách”

Nhưng ở bài viết như sau: Trọn gói: 4.200.000 VNĐ, Gia tour: 3.800.000 VNĐ, Gái: 10.450.000 VND… dẫn đến bộ trích chọn không trích ra được thuộc tính giá tour. Tương tự như vậy với các thuộc tính còn lại.

4.6. Một số ứng dụng kết quả trích chọn tour du lịch

Để có các báo cáo tổng hợp tương đối đầy đủ số liệu về 07 website về du lịch nêu trên, với mỗi website luận văn thực hiện thu thập dữ liệu (web crawler) 4000 bài viết, như vậy sau khi thu thập dữ liệu ta có 28.000 bài viết từ 07 website du lịch ở trên. Kết thúc quá trình thu thập, lọc dữ liệu, trích

chọn tour, trích chọn thuộc tính ta thu được các tour du lịch và các thuộc tính được lưu trữ trong cơ sở dữ liệu phục vụ thống kê, báo cáo.

4.6.1. Thống kê theo định danh

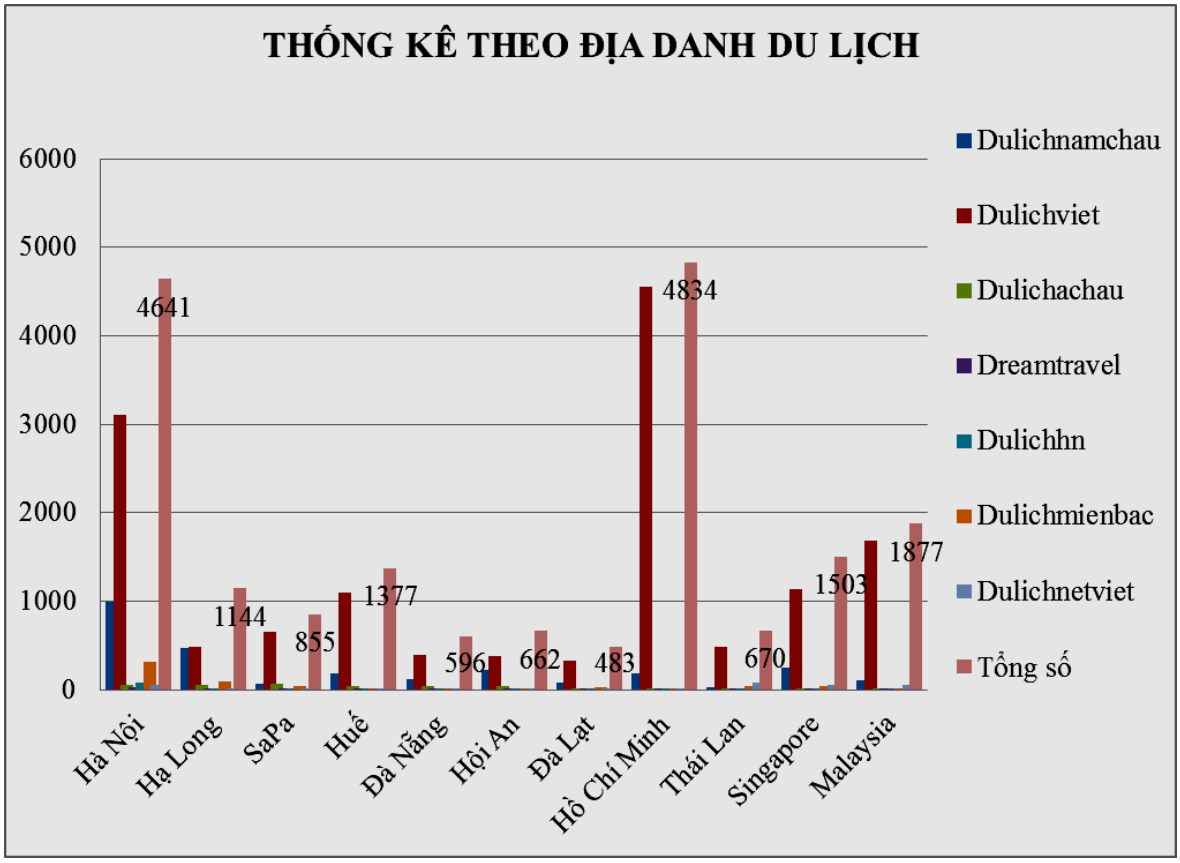
Sau quá trình trích chọn thông tin từ các website, ta có thể thống kê được số tour đến từng địa danh du lịch. Từ kết quả thống kê ta có thể có một vài nhận xét như:

- Địa danh du lịch nào đang được du khách quan tâm (thể hiện ở chỗ có nhiều tour).
- Địa danh thế mạnh của từng website.

Do số lượng địa danh du lịch rất nhiều nên tác giả chỉ lựa chọn một số địa danh điển hình.

Bảng 4.5. Bảng thống kê số tour theo địa danh du lịch

WEBSITE	Hà Nội	Hạ Long	SaPa	Huế	Đà Nẵng	Hội An	Đà Lạt	Hồ Chí Minh	Thái Lan	Singapore	Malaysia
Dulichnamchau	998	467	66	190	118	219	85	188	27	244	110
Dulichviet	3112	489	652	1104	390	375	329	4562	488	1140	1682
Dulichachau	56	48	69	41	40	37	5	21	15	12	15
Dreamtravel	24	20	10	10	12	8	13	15	12	9	11
Dulichhn	81	11	4	15	16	8	8	16	1	2	1
Dulichmienbac	312	90	38	6	5	6	25	18	45	37	9
Dulichnetviet	58	19	16	11	15	9	18	14	82	59	49
Tổng số	**4641**	**1144**	**855**	**1377**	**596**	**662**	**483**	**4834**	**670**	**1503**	**1877**



Hình 4.8. Biểu đồ thống kê số tour theo địa danh du lịch

Nhận xét

- Theo Hình 4.8 ta thấy 3 địa danh trong nước có số lượng tour nhiều nhất là: Thành phố Hồ Chí Minh (tổng số tour 4834), Hà Nội (4641 tour), Huế (1377 tour) đây cũng là các địa danh được mạng thông tin du lịch điện tử quốc tế Touropia (touropia.com) bình chọn là những địa điểm không thể bỏ qua khi tới Việt Nam. 2 địa điểm quốc tế có lượng tour nhiều nhất là Malaysia (1877 tour) và Singapore (1503 tour) cũng là các địa điểm được Huffingtonpost xếp vào danh sách những điểm đến ở Đông Nam Á "có thể thay đổi cuộc sống của bạn".
- Bảng 4.5 cho thấy các website Du Lịch Việt, Du lịch Năm Châu có số lượng tour nhiều hơn hẳn các website du lịch khác qua đó ta có thể nhận

định rằng đây là các website có uy tín, khi lựa chọn các tour du lịch trên các website này có thể được cung cấp dịch vụ tốt hơn.

- Qua Hình 4.8 ta biết được các địa điểm thế mạnh của các website. Ví dụ như trên trang Du Lịch Việt, số tour đến thành phố Hồ Chí Minh là 4562 tour lớn gấp 326 lần so với trang Du Lịch Nét Việt, khi muốn đến địa điểm Hồ Chí Minh thì chọn tour của trang Du Lịch Việt sẽ có giá hợp lý và các dịch vụ sẽ tốt hơn.

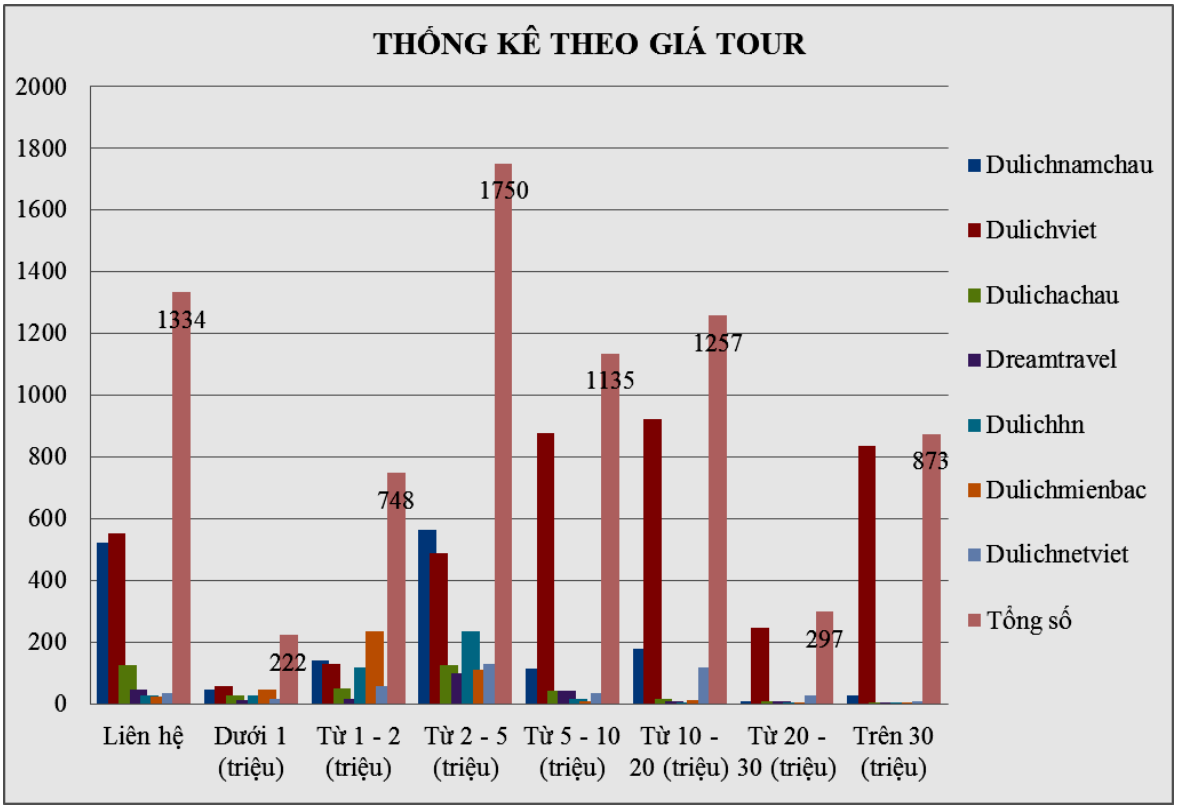
4.6.2. Thống kê theo giá tour

Sau khi trích chọn thông tin về tour, ta có thể thống kê được số lượng tour với từng mức giá cụ thể. Từ kết quả đó có thể có một vài nhận xét như:

- Giá tiền phổ biến của một tour thuộc từng website là bao nhiêu?
- Nên lựa chọn các tour thuộc website nào cho phù hợp với số tiền mình có?

Bảng 4.6. Bảng thống kê số tour theo giá

Tên website	Liên hệ	Dưới 1 (triệu)	Từ 1 - 2 (triệu)	Từ 2 - 5 (triệu)	Từ 5 - 10 (triệu)	Từ 10 - 20 (triệu)	Từ 20 - 30 (triệu)	Trên 30 (triệu)
Dulichnamchau	522	44	141	565	115	178	9	25
Dulichviet	553	55	129	488	878	924	246	835
Dulichachau	126	25	49	123	40	17	8	1
Dreamtravel	46	11	17	99	43	7	6	4
Dulichhn	28	25	119	236	15	3	1	1
Dulichmienbac	24	46	236	110	8	10	1	1
Dulichnetviet	35	16	57	129	36	118	26	6
Tổng số	**1334**	**222**	**748**	**1750**	**1135**	**1257**	**297**	**873**



Hình 4.9. Biểu đồ thống kê số tour theo giá tiền

Nhận xét

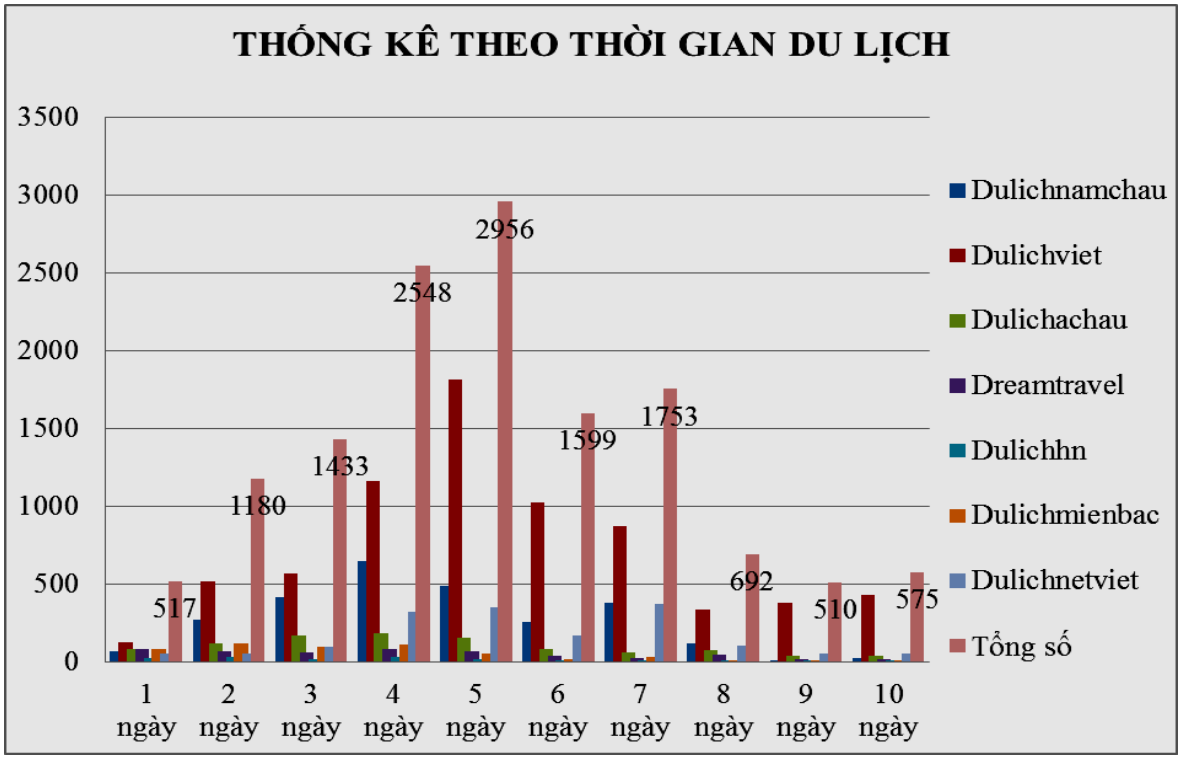
- Hình 4.9, ta thấy mức giá có nhiều tour nhất là từ 2 - 5 triệu (1750 tour).
- Bảng 4.6 cho ta biết số tour trong các mức giá của từng website. Từ kết quả của Bảng 4.6 ta có thể biết được mỗi website có thế mạnh là các tour ở mức bao nhiêu. Ví dụ với website Du Lịch AMI TOUR thì các tour ở mức từ 1- 2 triệu là các tour thế mạnh. Nếu ta có số tiền trong khoảng từ 1- 2 triệu thì nên chọn các tour của trang này.

4.6.3. Thống kê theo thời gian

Từ kết quả của quá trình trích chọn thông tin, ta thống kê được thời gian du lịch phổ biến là bao lâu. Qua đó có thể có những nhận định như: Số lượng tour du lịch trong từng khoảng thời gian là bao nhiêu? Thời gian phổ biến nhất của một tour là bao nhiêu ngày?

Bảng 4.7. Bảng thống kê số tour theo thời gian du lịch

Tên website	1 ngày	2 ngày	3 ngày	4 ngày	5 ngày	6 ngày	7 ngày	8 ngày	9 ngày	10 ngày
Dulichnamchau	67	272	416	646	489	258	379	122	9	23
Dulichviet	123	518	565	1162	1816	1023	871	338	378	432
Dulichachau	83	118	169	185	158	80	63	73	43	43
Dreamtravel	85	69	61	82	71	42	26	46	15	16
Dulichhn	22	33	21	34	17	11	11	7	5	3
Dulichmienbac	85	118	101	114	53	17	33	1	9	1
Dulichnetviet	52	52	100	325	352	168	370	105	51	57
Tổng số	**517**	**1180**	**1433**	**2548**	**2956**	**1599**	**1753**	**692**	**510**	**575**



Hình 4.10. Biểu đồ thống kê số tour theo thời gian

Nhận xét

- Từ Hình 4.10 ta thấy khoảng thời gian có nhiều tour nhất là 5 ngày (2956 tour) và 4 ngày (2548 tour).
- Bảng 4.7 cho ta biết số tour trong các khoảng thời gian của từng website. Từ kết quả của Bảng 4.7 ta có thể biết được mỗi website có thế mạnh là các tour trong khoảng thời gian nào. Ví dụ với website Du Lịch Năm Châu thì các tour trong khoảng 3 ngày (416 tour), 4 ngày (646 tour), 5 ngày (489 tour) là các tour thế mạnh.

4.7. Kết luận chương

Chương 4 trình bày kết quả thử nghiệm mô hình trích chọn thông tin về các tour du lịch trên 07 website về du lịch được chọn. Bao gồm các công việc sau:

1) Sử dụng công cụ (web crawler) thu thập các bài viết chứa các thông tin về các tour du lịch từ 07 website.

2) Lọc ra các bài viết chứa các thông tin về các tour du lịch.

3) Trích chọn các tour du lịch từ các bài viết theo tập luật được định nghĩa trước.

4) Trích chọn các thuộc tính của các tour du lịch theo tập luật được định nghĩa trước.

5) Lưu kết quả trích chọn vào cơ sở dữ liệu

6) Lập một số báo cáo, thống kê phục vụ công tác quản lý.

KẾT LUẬN

1. Những kết quả chính của luận văn

Luận văn đã đạt được mục tiêu đề ra ban đầu:

1) Tìm hiểu tổng quan về các phương pháp trích chọn thông tin, tìm hiểu bài toán trích chọn thông tin về các tour du lịch từ các website tiếng Việt, đưa ra phương pháp, mô hình giải quyết bài toán.

2) Thử nghiệm mô hình trích chọn thông tin về các tour du lịch trên 07 website về du lịch, lập một số báo cáo, thống kê phục vụ công tác quản lý, điều hành.

2. Một số hạn chế

Luận văn vẫn còn một số hạn chế như sau:

1) Không tự động trích chọn thông tin khi đưa vào một bài viết thuộc website mới.

2) Tập luật được xây dựng thủ công, do đó khó bao phủ tới toàn bộ miền dữ liệu. Điều này dẫn tới tập luật có thể bỏ sót những dữ liệu có liên quan tới miền dữ liệu.

3) Kết quả của bộ lọc dữ liệu chưa cao, còn bỏ qua nhiều bài viết chứa thông tin du lịch.

3. Định hướng tương lai

Định hướng nghiên cứu trong thời gian tới của luận văn là tiếp tục hoàn thiện và phát triển mô hình trích chọn thông tin du lịch trong văn bản tiếng Việt, tập trung vào các phương pháp trích chọn tự động, từ các thông tin trích chọn được xây dựng được hệ thống tư vấn du lịch và dự đoán xu hướng du lịch. Do hạn chế về thời gian và kiến thức cùng những khó khăn trong quá trình thu thập và tiền xử lý dữ liệu nên luận văn chưa sử dụng các phương pháp tự động. Vì vậy, nghiên cứu tiếp theo cũng sẽ tập trung vào việc sử dụng các phương pháp tự động trong trích chọn và phát triển ứng dụng.

TÀI LIỆU THAM KHẢO

Tài liệu tiếng Việt

[1] Hà Quang Thụy, Phan Xuân Hiếu, Đoàn Sơn, Nguyễn Trí Thành, Nguyễn Thu Trang, Nguyễn Cẩm Tú (2009). *Giáo trình khai phá dữ liệu Web,* Nhà xuất bản giáo dục Việt Nam.

Tài liệu tiếng Anh

[2] Alexander Yates. Information Extraction from the Web: Techniques and Applications. Phd thesis, University of Washington, 2007.

[3] Adam Berger. The Improved Iterative Scaling Algorithm: A gentle Introduction. School of Computer Science, Carnegie Mellon University

[4] A. McCallum, D. Freitag, and F. Pereia. Maximum entropy markov models for information extraction and segmentation. In Proc. Interational Conference on Machine Learning, 2000.

[5] A. Rauber, D. Merkl, and M. Dittenbach: The Growing Hierarchical Self-Organizing Map: Exploratory Analysis of High-Dimensional Data in: IEEE Transactions on Neural Networks, Vol. 13, No 6, pp. 1331-1341, IEEE, November 2002.

[6] Bing Liu, Web Data Mining Exploring Hyperlinks, Contents, and Usage Data, http://www.cs.uic.edu/~liub/WebMiningBook.html, December, 2006.

[7] F. Ciravegna, "Adaptive information extraction from text by rule induction and generalisation," in *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI2001)*, 2001.

[8] I. Muslea, S. Minton, and C. A. Knoblock, "A hierarchical approach to wrapper induction," in *Proceedings of the Third International Conference on Autonomous Agents*, Seattle, WA, 1999.

[9] J. Blitzer, R. McDonald, and F. Pereira, "Domain adaptation with structural correspondence learning," in *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP)*, 2006.

[10] Michael Dittenbach, Andreas Rauber, Dieter Merkl, Uncovering Hierarchical Struture in Data Using the Growing Hierarchical Self-Organizing Map, Institute of Software Technology, Vienna University of Technology, Vienna Austria, 24 July 2002.

[11] Minh-Tien Nguyen and Tri-Thanh Nguyen. “Extraction of Disease Events for a Real-time Monitoring System”, SoICT'2013, Da Nang, Vietnam, Decem- ber 5-6, 2013.

[12] M. E. Calif and R. J. Mooney, "Relational learning of pattern-match rules for information extraction," in *Proceedings of the Sixteenth Nation Conference on Artificial Intelligence (AAAI-99)*, pp. 328-334, July 1999.

[13] N. Kushmerick, "Wrapper induction for information extraction," PhD thesis, University of Washington, 1997.

[14] Scott Miller, Heidi Fox, et al. A Novel use of statistical parsing to extract information from Text, In 6th Applied Natural Language Processing Conference, 2000.

[15] S. Soderland, "Learning information extraction rules for semi-structured and free text," *Machine Learning*, vol. 34, 1999.

[16] Sunita Sarawagi (2008). Information Extraction.

[17] Teuvo Kohonen, et al. Self-Organizing Maps, Third edition, Springer, 2002.

[18] Yi-fang Brook Wu, Quanzhi Li. Document keyphrases as subject metadata: incorporating document key concepts in search results. Inf Retrieval -Springer. 2008.

[19] Zhou GuoDong, Su Jian, et al. Exploring Various Knowledge in Relation Extraction. Proceedings of the 43rd Annual Meeting of ACL, pages 427 - 434, Association for computational linguitics, 2005.

[20] *http://www.w3.org/DOM/*

[21] *http://www.w3.org/TR/xpath*

[22] *http://www.dcs.bbk.ac.uk/~ptw/teaching/ssd/toc.html*