BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG

NGÔ VĂN KHOA

NGHIÊN CỨU KỸ THUẬT PHÂN TÍCH VÀ TRÍCH RÚT THUỘC TÍNH TÀI LIỆU PHỤC VỤ CHO CÁC BÀI TOÁN TÌM KIẾM

Chuyên ngành: Khoa học máy tính
Mã số: 60.48.01

TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT

Đà Nẵng - Năm 2013

Công trình được hoàn thành tại

ĐẠI HỌC ĐÀ NẴNG

Người hướng dẫn khoa học: TS. HUỲNH CÔNG PHÁP

Phản biện 1: **PGS.TS. VÕ TRUNG HÙNG**

Phản biện 2: **PGS.TS. TRƯƠNG CÔNG TUẤN**

Luận văn được bảo vệ trước Hội đồng chấm Luận văn tốt nghiệp thạc sĩ Kỹ thuật họp tại Đại học Đà Nẵng vào ngày 18 tháng 5 năm 2013.

Có thể tìm hiểu luận văn tại:

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

MỞ ĐẦU

1. Lý do chọn đề tài

Trên môi trường Internet ngày nay, số lượng thông tin từ các Web Site là vô cùng lớn và vẫn đang còn gia tăng nhanh chóng theo từng ngày. Với hiện trạng đó, tìm kiếm thông tin là một nhu cầu không thể thiếu cho những người sử dụng Internet. Ngày nay, loại thông tin mà người sử dụng muốn tìm kiếm cũng đã trở nên phong phú, nó không còn đơn thuần là tìm kiếm nội dung văn bản trên một trang Web trên Internet, thay vào đó còn nhiều loại khác như: hình ảnh, tập tin âm thanh, tập tin video, tài liệu dưới dạng những tập tin được soạn thảo bằng các trình soạn thảo …

Tuy nhiên, hiện nay các máy tìm kiếm là công cụ tìm kiếm hỗ trợ rất tốt cho người sử dụng để truy vấn thông tin. Với các máy tìm kiếm khá phổ biến hiện nay như Google, Yahoo, thì khi nhận một truy vấn từ người dùng, các máy tìm kiếm này thường trả về một danh sách lớn các kết quả tìm kiếm. Các kết quả tìm kiếm này thường không chính xác, các kết quả tìm kiếm thường theo danh sách các từ khóa mà người dùng truy vấn. Thêm vào đó, đối với các truy vấn "nhập nhằng", có nhiều chủ đề liên quan thì người dùng rất khó khăn và tốn nhiều thời gian xem xét các tiêu đề và đoạn tóm lược của tài liệu để tìm ra kết quả mong muốn. Ví dụ, người truy vấn muốn tìm địa danh Đà Nẵng nhưng kết quả trả về các bài viết có chứa từ khóa Đà Nẵng như FPT Đà Nẵng hay báo Đà Nẵng điện tử, du lịch Đà Nẵng, khách sạn Đà Nẵng…

Nguyên nhân cho ra kết quả như trên là do tệp chỉ mục danh sách các từ được xây dựng với mỗi từ gắn vào một từ khóa mà thôi. Từ

hiện trạng đó để nâng cao tính chính xác cho kết quả tìm kiếm, chúng tôi đề xuất xây dựng tệp chỉ mục ngữ nghĩa tốt hơn, mỗi từ khóa gắn với một từ và có các thuộc tính mô tả từ đó, để phục vụ tốt cho bài toán kiếm. Do đó chúng tôi quyết định chọn đề tài "**Nghiên cứu kỹ thuật phân tích và trích rút thuộc tính tài liệu phục vụ cho các bài toán tìm kiếm".**

Trong luận văn này, chúng tôi mong muốn sử dụng các kĩ thuật đã được nghiên cứu, sử dụng kĩ thuật phân tích và trích rút thuộc tính tài liệu cùng với một số phương pháp xử lý ngôn ngữ tự nhiên để xây dựng tệp chỉ mục ngữ nghĩa để áp dụng vào máy tìm kiếm, tìm ra tập tất cả đối tượng thỏa mãn yêu cầu do người dùng đặt ra.

2. Mục tiêu và nhiệm vụ

- Tìm hiểu các kỹ thuật phân tích và trích rút thuộc tính của tài liệu.
- Xây dựng lại tệp chỉ mục ngữ nghĩa phục vụ tốt cho kết quả tìm kiếm.
- Để đạt được mục đích trên, nhiệm vụ chính của đề tài là:

Tìm hiểu tìm kiếm khai phá dữ liệu.

Ứng dụng phân tích và trích rút thuộc tính tài liệu trong các bài toán tìm kiếm.

Biểu diễn kết quả tìm kiếm.

3. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu của đề tài gồm:

- Các tài liệu văn bản.
- Hệ thống tìm kiếm.
- Tệp chỉ mục.

Phạm vi nghiên cứu như sau:

- Tài liệu HTML, file.Doc, file PDF.
- Tệp chỉ mục.
- Ngôn ngữ tiếng việt.

4. Phương pháp nghiên cứu

Phương pháp phân tích: Thu thập, phân tích dữ liệu và đánh giá độ liên quan của bảng dữ liệu.

Phương pháp thực nghiệm: Thực hiện việc cài đặt, thử nghiệm phương pháp trích rút thuộc tính tài liệu. Đánh giá kết quả đạt được theo bảng đánh giá đã xây dựng.

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

Sau khi thực hiện các phương pháp nghiên cứu các phương pháp trích rút thuộc tính, sẽ góp phần làm cơ sở cho việc lập chỉ mục ngữ nghĩa.

6. Bố cục của luận văn

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

Chương 1 Cơ sở lý thuyết.

Nội dung chính là tìm hiểu lý thuyết liên quan đến vấn đề nghiên cứu

- Giới thiệu xử lý ngôn ngữ tự nhiên
- Khai phá dữ liệu
- Tổng quan về hệ thống tìm kiếm thông tin

Chương 2 Các phương pháp trích rút thông tin.

Nội dung chính là tìm hiểu các phương pháp trích rút liên quan đến vấn đề nghiên cứu.

- Các phương pháp trích rút thông tin
- Đánh giá nhận xét các phương pháp

Chương 3 Thử nghiệm và đánh giá kết quả.

Trong chương này chúng tôi xây dựng chương trình

- Phát biểu bài toán
- Mô hình tổng quan.
- Ngôn ngữ thực nghiệm kết quả dự kiến.

CHƯƠNG 1.
CƠ SỞ LÝ THUYẾT

1.1. GIỚI THIỆU XỬ LÝ NGÔN NGỮ TỰ NHIÊN

1.1.1. Giới thiệu

Xử lý ngôn ngữ chính là xử lý thông tin đầu vào là "dữ liệu ngôn ngữ", tức là dữ liệu "văn bản" hay "tiếng nói". Các dữ liệu liên quan đến ngôn ngữ viết (văn bản) và nói (tiếng nói) đang dần trở nên kiểu dữ liệu chính của con người và lưu trữ dưới dạng điện tử. Đặc điểm chính của các kiểu dữ liệu này là không có cấu trúc hoặc nửa cấu trúc và chúng không thể lưu trữ trong các khuôn dạng cố định như các bảng biểu. Theo đánh giá cảu công ty Oracle, hiện có đến 80% dữ liệu không có cấu trúc trong lượng dữ liệu của loài người đang có. Với sự ra đời và phổ biến của Internet, của sách báo điện tử, của máy tính cá nhân, của viễn thông, của thiết bị âm thanh,…Người người ai cũng có thể tạo ra dữ liệu văn bản hay tiếng nói. Vấn đề là làm sao ta có thể xử lý chúng, tức là chuyển chúng từ dạng ta chưa hiểu được thành các dạng ta có thể hiểu và giải thích được, tức là ta có thể tìm ra thông tin, tri thức hữu ích cho mình.

Xử lý ngôn ngữ tự nhiên đã được ứng dụng trong thực tế để giải quyết bài toán như: nhận dạng chữ viết, tóm tắt văn bản, khai phá dữ liệu và phát hiện tri thức..

1.1.2. Khái niệm cơ bản ngôn ngữ tự nhiên

1.1.3. Khái niệm cơ bản xử lý ngôn ngữ tự nhiên

1.2. KHAI PHÁ DỮ LIỆU

Khái niệm khai phá dữ liệu (Data Mining)

Khai phá dữ liệu được định nghĩa như một quá trình chắt lọc hay khám phá tri thức từ một lượng lớn dữ liệu. Thuật ngữ Data Mining ám chỉ việc tìm một tập nhỏ có giá trị từ một lượng lớn các dữ liệu thô. Có sự phân biệt giữa khái niệm "Khai phá dữ liệu" với khái niệm "Phát hiện tri thức" (Knowledge Discovery in Databases - KDD) mà theo đó, khai phá dữ liệu chỉ là một bước trong quá trình KDD.

1.3. TỔNG QUAN HỆ THỐNG TÌM KIẾM THÔNG TIN

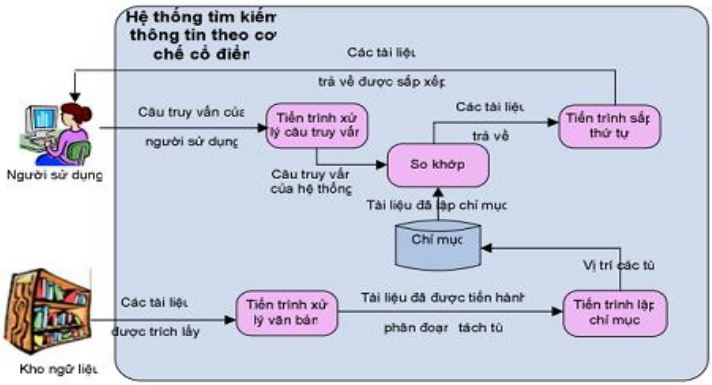
1.1.4. Giới thiệu về tìm kiếm thông tin

Tìm kiếm thông tin (Information Retrieval - IR) là tìm kiếm tài nguyên trên một tập lớn các dữ liệu phi cấu trúc được lưu trữ trên máy tính nhằm thỏa mãn nhu cầu về thông tin

1.1.5. Mục tiêu của hệ thống tìm kiếm thông tin

1.1.6. Cách thức hoạt động của một hệ thống tìm kiếm thông tin

Hình 1.1 minh họa cấu trúc, cách hoạt động cơ bản của một hệ thống tìm kiếm thông tin cổ điển.



Hình 1.1: Mô hình một hệ tìm kiếm thông tin

CHƯƠNG 2.
CÁC PHƯƠNG PHÁP TRÍCH RÚT THÔNG TIN

2.1. GIỚI THIỆU

Như chúng ta đã biết trích rút thông tin là một lĩnh vực nghiên cứu chuyên sâu thuộc lĩnh vực xử lý ngôn ngữ tự nhiên. Vì vậy các bài toán cũng như phương pháp trích rút thông tin đều có nguồn gốc, và tương tự các phương pháp kỹ thuật được sử dụng trong xử lý ngôn ngữ tự nhiên.

Trong phần này chúng tôi sẽ trình bày tóm tắt khảo sát về các bài toán liên quan đến trích rút thông tin từ văn bản (từ khóa, cụm từ khóa, thực thể có tên, quan hệ giữa các thực thể,…) cũng như các phương pháp tiếp cận và các phương pháp trích rút mối quan hệ ngữ nghĩa. Mục đích của việc trích rút thông tin là để tìm ra các thuộc tính của thông tin để xây dựng lại tệp chỉ mục trong tìm kiếm.

2.2. CÁC PHƯƠNG PHÁP TRÍCH RÚT THÔNG TIN

2.2.1. Trích rút cụm từ khóa(Keyphrase Extraction)

2.2.2. Nhận diện thực thể có tên

2.2.3. Nhận diện mối quan hệ

2.2.4. Trích rút metadata

2.2.5. Khái quát trích rút mối quan hệ ngữ nghĩa

2.2.6. Các phương pháp trích rút mẫu quan hệ ngữ nghĩa

2.3. ĐÁNH GIÁ NHẬN XÉT CÁC PHƯƠNG PHÁP

CHƯƠNG 3.
XÂY DỰNG CHƯƠNG TRÌNH

3.1. PHÁT BIỂU BÀI TOÁN

Trong thời đại công nghệ thông tin hiên hiện nay nhu cần tìm kiếm thông tin trên Internet là vấn đề cần thiết đối với người dùng, số lượng thông tin từ các Website là vô cùng lớn và vẫn đang còn gia tăng nhanh chóng theo từng ngày. Với hiện trạng đó, tìm kiếm thông tin là một nhu cầu không thể thiếu cho những người sử dụng Internet. Ngày nay, loại thông tin mà người sử dụng muốn tìm kiếm cũng đã trở nên phong phú, nó không còn đơn thuần là tìm kiếm nội dung văn bản trên một trang Web trên Internet.

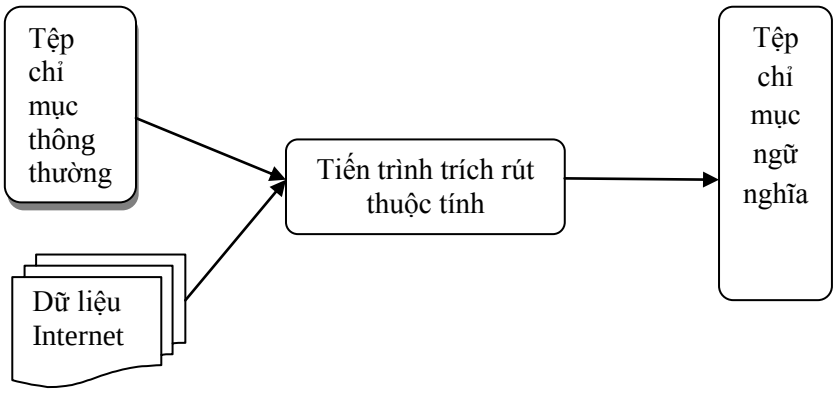
Các máy tìm kiếm là công cụ tìm kiếm hỗ trợ rất tốt cho người sử dụng. Với các máy tìm kiếm khá phổ biến như Google, Yahoo, thì khi nhận một truy vấn từ người dùng, các máy tìm kiếm này thường trả về một danh sách lớn các kết quả tìm kiếm. Các kết quả tìm kiếm này thường không chính xác, đối tượng cần tìm kiếm thì không đưa lên trang đầu tiên. Từ những vấn đề đó người tìm kiếm muốn có một công cụ tìm kiếm phải thỏa mãn hai tiêu chí: chính xác và nhanh chóng. Đây là một "thách thức" đối với tất cả các nhà phát triển khi muốn phát triển một hệ thống tìm kiếm ngày nay. Bởi vì, lượng thông tin trên Internet là vô cùng lớn, không có một máy chủ nào có thể chứa toàn bộ tất cả thông tin đó trong nó, nên các nhà phát triển phải chia lượng thông tin này thành nhiều phần để lưu trử tại các máy chủ đặt ở những nơi khác nhau. Ngoài ra, cũng do lượng thông tin lớn như vậy, nên việc tìm kiếm trên đó cũng đòi hỏi thời gian rất lơn nếu như chúng không có được tệp chỉ mục tốt. Để đáp ứng được hai tiêu chí đó, bài toán được giải quyết mà chúng tôi đưa ra là ***trích***

rút thuộc tính tài liệu sau đó xây dựng ***tệp chỉ mục ngữ nghĩa*** để áp dụng vào trong máy tìm kiếm.

Trong tệp chỉ mục ngữ nghĩa có rất nhiều loại từ, thuộc nhiều lĩnh vực khác nhau, với mỗi từ khóa thì nó sẽ có nhiều quan hệ ngữ nghĩa mô tả cho từ khóa đó. Nếu chúng tôi xây dựng tệp chỉ mục ngữ nghĩa với nhiều lĩnh vực như vậy sẽ tốn rất nhiều thời gian và không khả thi, Do đó chúng tôi giới hạn phạm vi bài toán nghiên cứu là trích rút thuộc tính tài liệu về địa danh và xây dựng tệp chỉ mục ngữ nghĩa theo địa danh trong ngôn ngữ tiếng việt.

3.2. MÔ HÌNH TỔNG QUAN

3.2.1. Mô hình tổng quan trích rút và xây dựng tệp chỉ mục ngữ nghĩa



Hình 3.1 Mô hình tổng quan trích rút và xây dựng tệp chỉ mục ngữ nghĩa

Trong mô hình tổng quan này, đầu vào là của dữ liệu là tệp chỉ mục thông thường và tài liệu trên Internet sau khi đưa vào tiến trình trích rút thuộc tính tài liệu, thì ta xây dựng được tệp chỉ mục ngữ nghĩa.

3.2.2. Xây dựng tệp chỉ mục ngữ nghĩa

3.2.3. Nhận xét tệp chỉ mục thông thường

Trong các phương pháp xây dựng tệp chỉ mục thì tệp chỉ mục thông thường sẽ chia từ khóa theo từ có dạng như sau:

Với mỗi từ khóa sẽ gắn vào đó là các URL1, URL2, URL3… các URL này sẽ liên kết với các từ khóa đó.

Ví dụ:

Với từ khóa Đà Nẵng thì tệp chỉ mục thông thường có dạng:

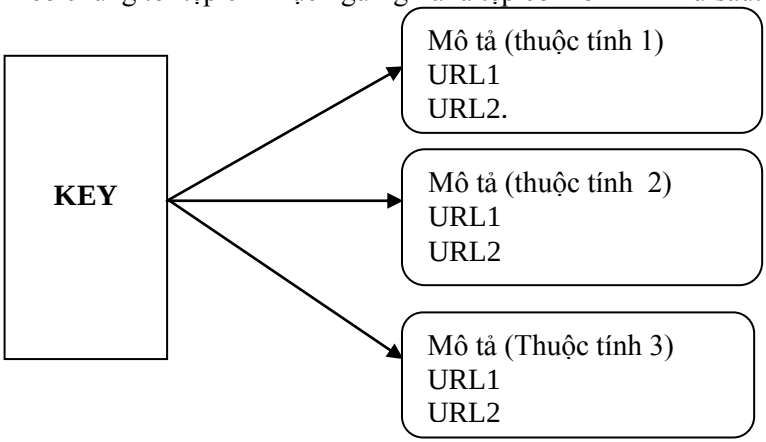
KEY	URL
Đà Nẵng	URL1,URL2,URL3..
Đà	URL1,URL2,URL3..
Nẵng	URL1,URL2,URL3…

Với từ Đà Nẵng thì có những URL chứa cả 2 từ Đà Nẵng, cũng có thể có những URL chỉ chứa một từ Đà hoặc một từ Nẵng. Ngoài ra từ Đà có những URL nói về công ty Sông Đà hoặc Thủy điện Sông Đà, với tệp chỉ mục hiện tại như vậy khi tìm kiếm sẽ hiện lên, ảnh hưởng đến kết quả tìm kiếm. Từ những nhược điểm đó chúng tôi đề xuất một công cụ ***xây dựng tệp chỉ mục ngữ nghĩa.***

3.2.4. Tệp chỉ mục ngữ nghĩa

a. Tổng quan.

Theo chúng tôi tệp chỉ mục ngữ nghĩa là tệp có mô hình như sau:



Hình 3.2 Mô hình tệp chỉ mục ngữ nghĩa

Việc lập chỉ mục theo hướng ngữ nghĩa là trích rút các thuộc tnh có trong nội dung văn bản để làm chỉ mục biểu diễn cho nội dung tài liệu. Việc trích rút thuộc tính có thể được thực hiện theo nhiều phương pháp mà một trong những phương pháp đó. Nhiều công trình lập chỉ mục theo ngữ nghĩa đã tìm các giải pháp sao cho không cần so khớp tài liệu. Từ đó việc lập chỉ mục theo hướng ngữ nghĩa chia ra 2 hướng tiếp cận lớn :

b. Nhóm các từ thuộc lĩnh vực ngữ nghĩa.

c. Kế thừa các ontology đã có.

3.2.5. Các bước xây dựng tệp chỉ mục ngữ nghĩa theo địa danh

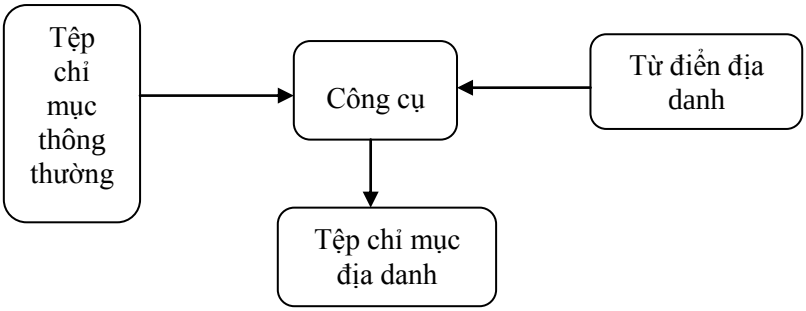
a. Bước 1: Xác định từ địa danh

Đầu vào của dữ liệu: tệp chỉ mục thông thường, các tài liệu trên Internet

Đầu ra của dữ liệu: tệp chỉ mục địa danh.

Theo như bài toán đã phát biểu, chúng ta chỉ xây dựng tệp chỉ mục ngữ nghĩa trong phạm vi các từ địa danh. Mà như chúng ta đã biết trong tệp chỉ mục thông thường và các tài liệu trên Internet, thì có rất nhiều từ khóa và mỗi từ khóa này có thể là địa danh hoặc không phải là từ nói về địa danh hay cũng có thể nói về lĩnh vực khác nữa. Như vậy chúng ta phải làm thế nào để xác định được từ khóa nào là từ nói về địa danh đây là vấn đề cần giải quyết ở bước này. Theo chúng tôi thì một trong những phương pháp được đưa ra để giải quyết vấn đề này là chúng ta có thể so khớp các địa danh trong từ điển địa danh với các tệp chỉ mục tài liệu đó, hay chúng ta cũng sử dụng phương pháp so mẫu chính xác đó là một trong các hướng đưa ra để giải quyết vấn đề.

Ta có thể khái quát mô hình xác định địa danh như sau:



Hình 3.3 Mô hình xác định từ địa danh

Phương pháp xác định từ địa danh

Chúng ta có thể thực hiện 2 phương pháp so khớp giữa các key word với từ điển hoặc sử dụng phương pháp so mẫu. Sau đây chúng tôi xin trình bày hai phương pháp so mẫu:

Phương pháp so mẫu chính xác: Cho xâu mẫu P có độ dài m(P=P1 P2…Pm-Pi là ký tự) và văn bản T độ dài n (T=T1 T2…Tn-Ti là ký tự) Tìm tất cả các vị trí xuất hiện của mẫu P trong xâu T.

Phương pháp so mẫu xấp xỉ: Tìm kiếm xấp xỉ là bài toán tìm sự xuất hiện của một mẫu trong văn bản, trong đó sự khớp giữa mẫu và xuất hiện của nó có thể chấp nhận " k" lỗi (k là một giới hạn cho trước). Có thể kể ra một vài kiểu lỗi, như những lỗi đánh máy hay lỗi chính tả trong hệ thống rút trích thông tin…vì trong các hệ thống tin học khó có thể tránh các lỗi nên vấn đề tìm kiếm xấp xỉ càng trở nên quan trọng. Ví dụ như thứ tự ghép từ khác nhau nhưng mang ngữ nghĩa giống nhau (ví dụ "toán logic" và "logic toán") hoặc do thứ tự sai song vẫn hiểu được đúng nghĩa (ví dụ "toán giải tích" và "giải tích toán",..) hoặc do lỗi đánh máy (ví dụ "sedan" viết thành " sudan")

Phương pháp được phát biểu: Cho xâu mẫu P độ dài m và văn bản T độ dài n. Từ đó xác định độ tương tự giữa hai xâu P và T

Phương pháp trên chúng ta có thể đưa về tìm xâu con chung dài nhất (hay khúc con chung dài nhất). Một xâu w là xâu con hay khúc con của xâu T nếu T= uwv (xâu u,v có thể rỗng). Xâu w là khúc con chung của hai xâu P,T nếu w đồng thời là khúc con của P,T. Khúc con chung dài nhất của hai xâu P,T.

```
Char chuoicon (char *p, char *T, int m)
     { int len,k, i, j;
        Char str[m], tam[m];
        len = 1 ;
        str='' ;
         while  ( len <= strlen(s1))
              {
                 k=strstr(p,T);
                 j = 1;
                 tam='';
                 for ( i= k; i<=strlen(p); i++)
                   if (p[j]=T[i])
                       {
                          Tam[j]=T[i];
                          j++;
                       }
                   Else
                     if (  strlen(tam) > strlen( str))
                        for (i= 1, i<=j,i++)
                          str[i]=tam[i];
               len ++;
              }
}
```

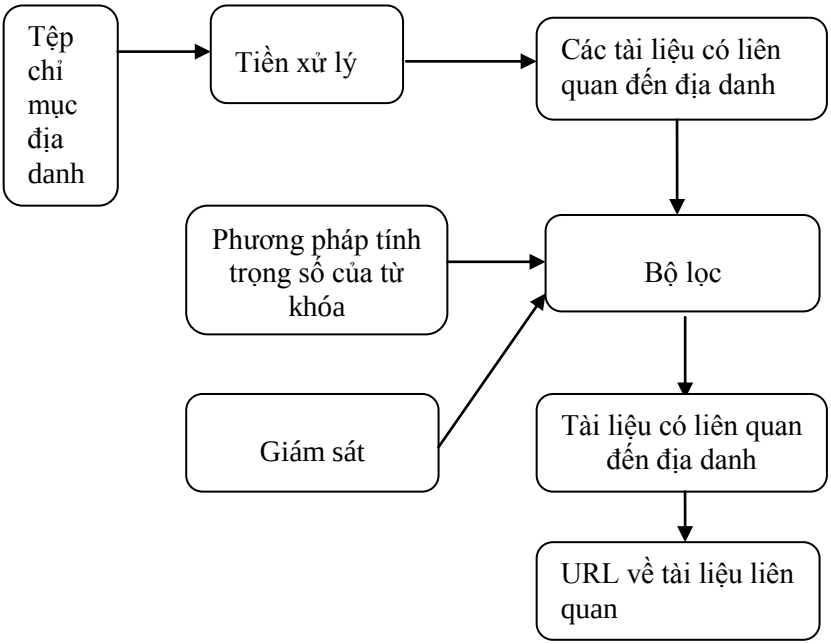
Qua các phương pháp trên chúng tôi nhận thấy rằng việc so khớp giữa key word với từ điển địa danh cho trước, từ đó chúng tôi xây dựng tệp chỉ mục địa danh. Như vậy ở đây chúng tôi sử dụng phương pháp so mẫu chính xác giữa keyword trong tệp chỉ mục với từ địa danh trong từ điển, từ đó nhận định đâu là các từ địa danh trong tệp chỉ mục. Tiếp theo chúng tôi sẽ giữ lại các từ địa danh và các URL của nó, các từ khóa và các URL không liên quan chúng tôi sẽ loại bỏ khỏi tệp chỉ mục. Như vậy sau bước này chúng tôi sẽ có được tệp chỉ mục địa danh.

b. Bước 2: Thu thập tài liệu nói về từ địa danh

Sau khi thực hiện bước 1 ta có được tệp chỉ mục bình thường về địa danh với mỗi địa danh như vậy có rất nhiều URL có thể liên quan đến địa danh đó hoặc không liên quan đến địa danh, mà chỉ chứa từ khóa địa danh đó thôi.

Ví dụ : Trang Web nói nói về một công ty ở Đà Nẵng có từ khóa Đà Nẵng nhưng không thể hiện các nội dung liên quan đến địa danh của Đà Nẵng thì các URL đó không phải là cái mà chúng tôi quan tâm.

Để thực hiện được công việc như vậy, chúng tôi đề xuất công cụ thu thập tài liệu nói về địa danh có mô hình như sau:



Hình 3.4 Mô hình công cụ thu thập tài liệu nói về địa danh

Đầu vào: tệp chỉ mục ngữ nghĩa địa danh với rất nhiều URL.

Đầu ra: tệp chỉ mục địa danh với những URL có chứa tài liệu liên quan đến địa danh.

Trong bước này chúng tôi cần quan tâm đến các tài liệu và URL của tài liệu đó mà nội dung liên quan đến từ khóa về địa danh. Nên chúng tôi tiến hành thu thập tất cả các tài liệu mà liên quan đến địa danh đó thôi.

Công cụ của chúng tôi qua bước tiền xử lý trong tài liệu thuộc các URL đó, sau đó bóc tách bỏ thẻ HTML và các thẻ không quan trọng trong trang Web chúng tôi chỉ lấy văn bản thuần trong trang Web đó thôi. Sau bước tiền xử lý ta có tài liệu liên quan đến địa danh, với mỗi tài liệu đó chúng tôi lưu địa chỉ các URL trước đó cùng với các tài liệu đó và phương pháp trọng số chúng tôi tiến hành lọc lại dưới sự giám sát của con người để có được tài liệu liên qua đến địa danh. Với các tài liệu liên quan đến địa danh có một URL tương ứng.

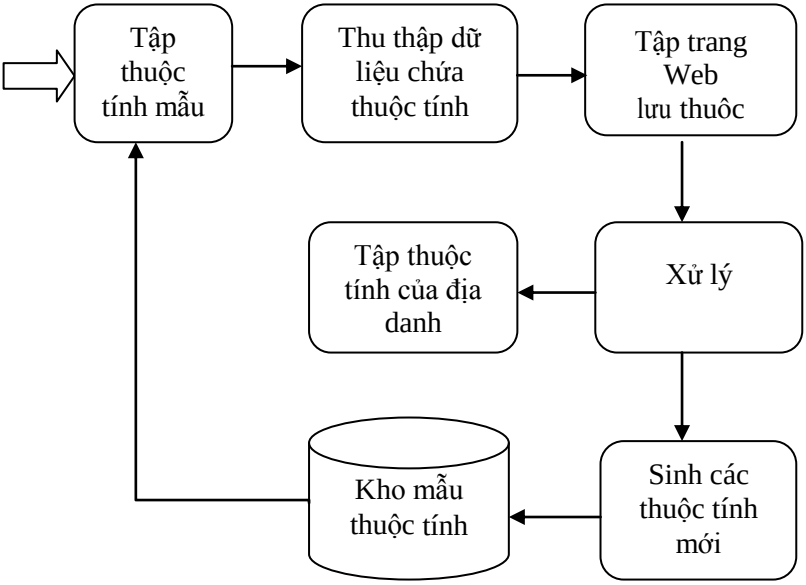
Sau bước này chúng tôi có tài liệu liên quan đến địa danh và các URL của tài liệu.

c. Bước 3: Xác định thuộc tính của địa danh

Đầu vào của dữ liêu: Địa danh cụ thể và tài liệu (URL nói về địa danh)

Đầu ra của dữ liệu: Các thuộc tính của địa danh và các danh sách URL theo các thuộc tính đó.

Chúng tôi tiến hành sử dụng phương pháp trích rút mẫu quan hệ ngữ nghĩa từ đó xây dựng công cụ của mình. Mô hình công cụ của chúng tôi như sau:



Hình 3.5 Mô hình trích rút thuộc tính theo mẫu quan hệ ngữ nghĩa

Trong mô bước này chúng sử dụng trích rút mối quan hệ ngữ nghĩa cụ thể là phương pháp Snowball và phương pháp máy tìm kiếm để trích rút các thuộc tính về địa danh.

Thu thập tài liệu

Tài liệu này đã có trong tệp chỉ mục địa danh chúng tôi đã thực hiện ở bước 1.

Phân loại dữ liệu chứa thuộc tính

Dựa theo tập thuộc tính mẫu, chúng tôi sẽ sử dụng phương pháp so khớp để tìm ra, phân loại các tài liệu chứa thuộc tính đó. Tuy nhiên để biết chính xác các tài liệu có chứa thuộc tính đó có phải đang mô tả cho thuộc tính đó không, chúng tôi sẽ sử dụng cơ chế duyệt lại các tài liệu đó. Như vậy ở bước này để đảm bảo độ chính

xác khi phân loại tài liệu theo thuộc tính, chúng tôi kết hợp việc học máy, có giám sát của con người.

Xử lý

Ở bước này, với các tài liệu đã xác định được thuộc tính cụ thể thì chúng tôi tiến hành sắp xếp lại theo kiểu: 1 thuộc tính – các URL tài liệu liên quan. Song song với việc đó, với những tài liệu liên quan đến địa danh, nhưng chưa được xếp vào thuộc tính mẫu nào, chúng tôi sẽ tiến hành trích rút để tìm ra thuộc tính mới, bổ sung cho tập thuộc tính mẫu. Chúng tôi tiến hành như sau tách câu trên tập dữ liệu thu được và giữ lại những câu chứa cả hai thành phần. Tách từ trong tiếng Việt, loại bỏ từ dừng cho tập câu này. Áp dụng phương pháp sinh tự động tập thực thể để mở rộng tập thực thể từ những thực thể ban đầu cho từng mối quan hệ đã được xác định trước các nhãn thực thể. Phương pháp này được trình bày ở phần tiếp theo.

Gán nhãn tổng quát

Dựa vào tập thực thể mở rộng, tiến hành tìm và xác định nhãn cho các thực thể có chứa trong tập câu thu được ở bước trên.

Sau khi các thực thể được gán nhãn, xác định các thành phần trái, thành phần phải, thành phần giữa cho các thực thể có chứa trong tập thuộc tính dựa vào tập câu thu được.

Biểu diễn các thành phần trái, thành phần phải và thành phần giữa dưới dạng các vector, ta thu được một tập các mẫu thô.

Phân cụm mẫu.

Tiến hành so khớp các thành phần trái, thành phần phải và thành phần giữa cho các mẫu thô để loại bỏ các mẫu thô trùng.

Dựa theo phương pháp Snowball, xác định các mẫu quan hệ được thực hiện bằng việc phân cụm mẫu thô. Mỗi cụm đại diện bởi

một mẫu và quá trình phân cụm mẫu được thực hiện như sau: Với những mẫu thô mới được sinh ra, tiến hành tính độ tương đồng với các mẫu đại diện theo công thức sau:

Match(mẫu1,mẫu2=(prefix1.prefix2)+(suffix1.suffix2)+(middle1.middle2)

Sinh thuộc tính mới

Những mẫu tổng quát đã thu được sẽ làm đầu vào cho vào máy tìm kiếm để tìm ra tập các câu có chứa các mẫu đó.

Nhận dạng các thực thể có chứa trong tập câu dựa vào tập các thực thể mở rộng.

Kiểm tra độ tin cậy của các thuộc tính mới được sinh ra. Những thuộc tính vượt qua được giá trị ngưỡng thì giữ chúng lại. - Sau đó quay lại bước 1, sử dụng tập thuộc tính mới thu được cùng với tập thuộc tính ban đầu đưa vào máy tìm kiếm để tiến hành sinh tập thuộc tính mới. Vòng lặp sẽ được dừng khi số lượng thuộc tính mới hoặc mẫu mới không còn được tiếp tục sinh ra.

Sau khi thực hiện việc trích rút mẫu thì ta có được tập thuộc tính của từ địa danh và chuyển sang bước 4.

d. Bước 4: Xây dựng mô tả từ địa danh

Sau khi thực hiện các thao tác ở bước 3 thì chúng tôi có được thuộc tính về địa danh như chúng ta đã biết với mỗi địa danh thì có rất nhiều thuộc tính mô tả về địa danh đó, nhưng mỗi thuộc tính lại liên kết với một lớp này mô tả làm giàu thông tin cho thuộc tính đó, mỗi lớp này giống như một ontology chứa các thông tin mô tả các thuộc tính đó.

Sau bước 3 chúng tôi có được tập thuộc tính địa danh, với tập thuộc tính đại danh đó chúng tôi sẽ sử dụng các thuộc tính địa danh kết hợp URL liên quan thuộc tính, tiếp tục xây dựng tệp chỉ mục mô tả địa danh

theo cụ thể với mỗi địa danh chúng tôi sẽ đưa các URL liên quan đến thuộc tính địa danh đó vào.

Chúng tôi cũng có thể mô tả tệp chỉ mục ngữ nghĩa dưới dạng cây chỉ mục cây chỉ mục có dạng XML cụ thể về địa danh của từ đó được mô tả tổng quan như sau:

```
<Địa danh>
        <Thuộc tính 1>
                <URL1>...</URL3>
                <URL2>...</URL3>
                <URL3>...</URL3>
        </thuộc tính 1>
        <Thuộc tính 2>
                <URL1>...</URL1>
                <URL2>...</URL2>
                <URL3>...</URL3>
        </thuộc tính 2>
</Địa danh>
```

Sau khi mô tả từ địa danh đó chúng tôi tiến hành lập chỉ mục ngữ nghĩa cho địa danh đó, tệp chỉ mục ngữ nghĩa địa danh này được trình bày trong mục 3.2.4.

Kết luận:

Qua 4 bước thực hiện ở bước 1 chúng tôi xác định được từ địa danh bằng phương pháp so mẫu, bước 2 chúng ta thu thập tài liệu nói về địa danh đó ở bước 3 sử dụng phương pháp trích rút mẫu quan hệ để lấy ra

các thuộc tính, bước 4 xây dựng và mô tả từ địa danh đó rồi lập tệp chỉ mục ngữ nghĩa địa danh.

3.3. NGÔN NGỮ THỰC NGHIỆM, KẾT QUẢ DỰ KIẾN

3.3.1. Ngôn ngữ XML

a. Lịch sử

XML (viết tắt từ tiếng Anh Extensible Markup Language, "Ngôn ngữ Đánh dấu Mở rộng"). Vào giữa những năm 1990, các chuyên gia SGML đã có kinh nghiệm với World Wide Web (vẫn còn khá mới vào thời đó). Họ tin tưởng rằng SGML có thể cung cấp giải pháp cho các vấn đề mà Web đang gặp phải. Jon Bosak đưa ra ý kiến W3C nên tài trợ một chương trình mang tên "SGML trên Web".

b. Đặc điểm

XML cung cấp một phương tiện dùng văn bản (text) để mô tả thông tin và áp dụng một cấu trúc kiểu cây cho thông tin đó. Tại mức căn bản, mọi thông tin đều thể hiện dưới dạng text, chen giữa là các thẻ đánh dấu (markup) với nhiệm vụ ký hiệu sự phân chia thông tin thành một cấu trúc có thứ bậc của các dữ liệu ký tự, các phần tử dùng để chứa dữ liệu, và các thuộc tính của các phần tử đó. Về mặt đó, XML tương tự với các biểu thức S (S-expression) của ngôn ngữ lập trình LISP ở chỗ chúng đều mô tả các cấu trúc cây mà trong đó mỗi nút có thể có một danh sách tính chất của riêng mình.

c. Cú pháp

Cú pháp XML cơ bản cho một phần tử là

```
<tên thuộc_tính="giá trị">nội dung</tên>
```

3.3.2. Kết quả dự kiến

Sau khi thực hiện các phương pháp trích rút thuộc tính và xây dựng tệp chỉ mục ngữ nghĩa thì chúng tôi mong muốn luận văn đạt được kết quả sự kiến cụ thể như tệp chỉ mục ngữ nghĩa về địa danh Đà Nẵng được mô tả theo ngôn ngữ XML có cấu trúc cây như sau:

< *Đà Nẵng*>
<*Dân số*>
<*URL1*>http://www.vietnamtourism.com/v_pages/country/province.asp?uid=73 </URL>
<*URL*>http://infonet.vn/Thoi-su/Den-nam-2030-dan-so-Da-Nang-se-len-den-2-trieu-nguoi/64725.info </URL>
<*URL*>http://www.danang.gov.vn/portal/page/portal/danang/chinhquyen/gioi_thieu/Dan_so... </*URL*>
</*Dân số*>
<*Vị trí địa lý*>
<*URL*>http://vi.wikipedia.org/wiki/%C4%90%C3%A0_N%E1%BA%B5ng </*URL*>
<*URL*>http://www.dulichdanang.info/gioi-thieu-du-lich-da-nang/vi-tri-dia-ly-dien-tich-tu-nhien-thanh-pho-da-nang.html</*URL*>
<*URL*>http://www.danang.gov.vn/portal/page/portal/danang/chinhquyen/gioi_thieu/Dieu_kien_tu_nhien </*URL*>
</*vị trí địa lý*>
<*Điểm du lich*>
<*URL*>http://www.dulichdanang.info/gioi-thieu-du-lich-da-nang/vi-tri-dia-ly-dien-tich-tu-nhien-thanh-pho-da-nang.html</*URL*>
<*URL*>http://www.danangxanh.com/thong-tin-du-lich/diem-tham-quan-da-nang.html</*URL*>
<*URL*>http://www.web-du-lich.com/dich-vu/type.php?iCha=10&iCat=103&module=news</*URL*>
</*điểm du lịch*>
...
< /*Đà Nẵng*>

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Trong luận văn đã tập trung nghiên cứu các phương pháp trích rút thông tin và các phương pháp trích rút mối quan hệ ngữ nghĩa, trong luận văn đã sử dụng phương pháp trích rút mối quan hệ nghĩa cụ thể là phương pháp Snowball để trích rút các thuộc tính về địa danh.

Trong thời gian không nhiều, nên cũng chưa thực nghiệm được các phương pháp trích rút để đánh giá các phương pháp trích rút đó. Tuy nhiên luận văn cũng đạt được những yêu cầu đề ra, Phân tích các vấn đề xung quanh bài toán trích chọn thuộc tính thuộc tính địa danh và xây dựng mô hình tệp chỉ mục ngữ nghĩa để áp dụng cho bài toán tìm kiếm.

Do khuôn khổ có hạn về thời gian cũng như lượng kiến thức có được và gặp khó khăn trong quá trình thu thập dữ liệu thử nghiệm nên còn một số vấn đề mà luận văn phải tiếp tục hoàn thiện và phát triển trong thời gian tới và hướng phát triển cho tương lai.

Thử nghiệm trên một dữ liệu lớn hoàn chỉnh hơn, với nhiều từ địa danh hơn, và mở rộng trích rút thuộc tính trong nhiễu lĩnh vực. Xây dựng các phương pháp hiệu quả hơn trong việc trích chọn cụm danh từ, ngữ nghĩa tiếng Việt.