ĐẠI HỌC THÁI NGUYÊN

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

NGUYỄN CÔNG BẰNG

WEB NGỮ NGHĨA VÀ ỨNG DỤNG TRONG TRA CỨU VĂN HÓA ẨM THỰC TẠI HẢI PHÒNG

Chuyên nghành : Khoa học máy tính

Mã số : 60.48.01

LUẬN VĂN THẠC SĨ CÔNG NGHỆ THÔNG TIN

Người hướng dẫn khoa học: PGS.TS ĐOÀN VĂN BAN

Thái nguyên – Năm 2014

Mục lục

Mở đầu

1. Lý do chọn đề tài

Ngày nay khoa học và công nghệ phát triển cùng với sự bùng nổ về internet thì Word Wide Web phát triển cả về nội dung lẫn hình thức. Nó có một khối lượng thông tin khổng lồ, được tạo ra từ các tổ chức, cộng đồng và nhiều cá nhân với lý do khác nhau. Người sử dụng Web có thể dễ dàng truy cập những thông tin này bằng cách chỉ ra địa chỉ URL và theo các liên kết để tìm ra các tài nguyên liên quan khác.

Tính đơn giản của Web hiện nay đã dẫn đến một số hạn chế, việc tìm kiếm thông tin trên Web có thể trả về một lượng lớn thông tin không hợp lý và không liên quan. Tính đơn giản này đã gây ra hiện tượng thắt cổ chai, tạo khó khăn trong việc tìm kiếm, trích rút thông tin. Máy tính chỉ biết gửi và trả thông tin, chúng không thể truy xuất những nội dung cần. Nó chi hỗ trợ ở mức độ giới hạn nào đó trong việc truy xuất và xử lý thông tin. Kết quả là người sử dụng phải đảm nhiệm việc truy cập, xử lý thông tin, trích lọc thông tin phù hợp với việc tìm kiếm.

Để khắc phục các hạn chế này, khái niệm web ngữ nghĩa đã ra đời. Web ngữ nghĩa là một bước tiến vượt bậc so với kỹ thuật web trước đó dựa vào khả năng làm việc với thông tin của chúng thay vì chỉ đơn thuần là lưu trữ thông tin.

Hải Phòng là một trong 5 thành phố trực thuộc trung ương và là một đô thị loại 1 trung tâm cấp quốc gia, là thành phố lớn thứ 3 của Việt Nam,có vị trí quan trọng về kinh tế xã hội và an ninh, quốc phòng của vùng Bắc Bộ và cả nước.

Ẩm thực Hải Phòng bình dị và dân dã, không cầu kỳ nhưng đậm đà khó quên. Nơi đây nổi tiếng với các món hải sản. Các nhà hàng hải sản ở khu vực Đồ Sơn nổi tiếng với tôm cua cá mực rất tươi và giá phải chăng. Phong cách chế biến hải sản ở Hải Phòng theo phong cách dân dã, nhấn mạnh thực chất và vị tươi ngon của nguyên liệu nhiều hơn sự cầu kỳ trong gia vị và cách chế biến.

Các món ăn như bánh đa cua, bún cá, bánh mỳ cay, cơm cháy hải sản, ốc cay, nem cua bể (nem vuông), giờ đây đã quá quen thuộc và nổi tiếng. Những món ăn này

có thể được tìm thấy trên đường phố của những nơi khác như TP.Hồ Chí Minh, Hà Nội,... nhưng được thưởng thức chúng trên Thành phố Hoa phượng đỏ vẫn là lý tưởng nhất vì sự lựa chọn nguyên liệu tươi ngon cùng với những bí quyết ẩm thực riêng của người đầu bếp. Ẩm thực Hải Phòng đã từng được quảng bá sang Châu Âu tại lễ hội biển Brest 2008 (Cộng hòa Pháp) và đã gây được tiếng vang lớn.

Ngoài ra, Hải Phòng còn nổi tiếng với nhiều món ăn khác như lẩu bề bề, nộm giá, thịt san biển, sủi dìn, bánh bèo,... Một số món ăn không thể thưởng thức ở những nơi khác mà chỉ có tại Hải Phòng.

Với những lý do trên, tôi xin chọn đề tài "**Web ngữ nghĩa và ứng dụng trong tra cứu đặc trưng văn hóa ẩm thực tại Hải Phòng**"

2. Mục tiêu

Ứng dụng Semantic Web xây dựng ứng dụng tra cứu đặc trưng văn hóa ẩm thực tại Hải Phòng.

3. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu:

- Tìm hiểu về web ngữ nghĩa, phương pháp xây dựng Ontology.
- Tìm hiểu về các thông tin đặc trưng văn hóa ẩm thực tại Thành phố Hải Phòng.

Phạm vi nghiên cứu:

- Nghiên cứu xây dựng tập từ vựng cơ bản về đặc trưng văn hóa ẩm thực tại Thành phố Hải Phòng.
- Tổ chức lưu trữ dữ liệu của ứng dụng với Protégé và tính năng truy xuất dữ liệu trong Ontology.

4. Phương pháp nghiên cứu

- Tìm hiểu các vấn đề về Web ngữ nghĩa.
- Thu thập các tài liệu liên quan.
- Triển khai xây dựng ứng dụng.

5. Ý nghĩa khoa học và thực tiễn đề tài

- Xây dựng tập từ vựng về văn hóa ẩm thực ở Hải Phòng.
- Góp phần nâng cao khả năng tra cứu và chia sẻ thông tin về văn hóa ẩm thực tại thành phố Hải Phòng.

6. Dự kiến bố cục luận văn

Luận văn được chia làm 3 chương:

Chương 1: Trình bày giới thiệu tóm tắt về Web ngữ nghĩa, kiến trúc của Web ngữ nghĩa, cũng như giới thiệu RDF – nền tảng của Web ngữ nghĩa.

Chương 2: Giới thiệu các công nghệ xây dựng Web ngữ nghĩa cụ thể là đi sâu vào nghiên cứu Ontology. Đồng thời đưa ra giải pháp về ngôn ngữ và công cụ để xây dựng ứng dụng Semantic web.

Chương 3: Giới thiệu về ứng dụng, phân tích và đề xuất giải pháp xây dựng ứng dụng. Tiến hành xây dựng ontology, xử lý dữ liệu, cài đặt ứng dụng và đưa ra một số kết quả đạt được.

CHƯƠNG 1: GIỚI THIỆU VỀ WEB NGỮ NGHĨA

1.1. Cách thức tìm kiếm thông tin của bộ máy tìm kiếm (Search engine)

Search engine hay còn gọi là máy tìm kiếm là một trang Web cho phép người dùng tìm kiếm nội dung số của các trang Web trên Internet [1].

Thường kỳ, máy tìm kiếm sẽ dò quét nội dung tất cả các trang Web trên Internet và cập nhật nội dung văn bản text vào cơ sở dữ liệu khổng lồ của mình mà người dùng có thể khai thác sau đó. Để làm việc này các máy tìm kiếm thường gửi các Web crawler, web spider hay web robot (ví dụ googlebot của Google – Yahoo slurp của Yahoo) đến các trang cần đánh chỉ số. Các bọ tìm kiếm này sẽ truy cập phân tích và gửi nội dung về các máy tìm kiếm.

Máy tìm kiếm sắp xếp các trang Web dựa vào nội dung HTML của trang. Việc này khác với các thư mục Web truyền thống mà những người kiểm duyệt sắp đặt trong các mục riêng biệt với tên site và miêu tả đi kèm.

1.1.1. Một số bộ tìm kiếm thông dụng

Bộ thu thập thông tin

Cơ sở dữ liệu cuả các search engine được cập nhật hoá bởi các chương trình đặc biệt thường gọi là "robot", "spider" hay "Webcrawler". Các chương trình này sẽ tự động dò tìm và phân tích từ những trang có sẵn trong cơ sở dữ liệu để kiếm ra các liên kết (links) từ các trang và trở lại bổ xung dữ liệu cho các search engine sau khi phân tích.

Về bản chất robot chỉ là một chương trình duyệt và thu thập thông tin từ các site theo đúng giao thức web. Những trình duyệt thông thường không được xem là robot do thiếu tính chủ động, chúng chỉ duyệt web khi có sự tác động của con người.

Bộ lập chỉ mục – Index

Hệ thống lập chỉ mục hay còn gọi là hệ thống phân tích và xử lý dữ liệu, thực hiện việc phân tích, trích chọn những thông tin cần thiết (thường là các từ đơn, từ ghép, cụm từ quan trọng) từ những dữ liệu mà robot thu thập được và tổ chức thành cơ sở

dữ liệu riêng để có thể tìm kiếm trên đó một cách nhanh chóng, hiệu quả. Lập chỉ mục là giai đoạn phân tích tài liệu (document) để xác định các chỉ mục biểu diễn nội dung của tài liệu. Hệ thống chỉ mục là danh sách các từ khoá, chỉ rõ các từ khoá nào xuất hiện ở trang nào, địa chỉ nào.

Bộ tìm kiếm thông tin – Search Engine

Search engine là cụm từ dùng chỉ toàn bộ hệ thống bao gồm bộ thu thập thông tin, bộ lập chỉ mục & bộ tìm kiếm thông tin. Các bộ này hoạt động liên tục từ lúc khởi động hệ thống, chúng phụ thuộc lẫn nhau về mặt dữ liệu nhưng độc lập với nhau về mặt hoạt động.

Search engine tương tác với user thông qua giao diện web, có nhiệm vụ tiếp nhận và trả về những tài liệu thoả yêu cầu của user.

Bộ Query Engine

Bộ công cụ truy vấn có nhiệm vụ nhận và tìm kiếm các yêu cầu của người sử dụng, Bộ công cụ này sẽ dựa vào bảng chỉ mục và các kho lưu trữ. Bởi kích thước của web rất lớn, thêm nữa khi sử dụng chỉ đưa vào một hay hai từ khóa sau đó sẽ nhận được tập kết quả. Do đó phải có một modul sắp xếp kết quả theo thứ tự sao cho nó gần với nội dung đang cần tìm nhất.

Sắp xếp

Đây là một modul có chức năng sàng lọc thông tin từ hàng triệu trang tương tự nhau để sắp xếp vị trí từng trang sao cho phù hợp nhất.

1.1.2. Cách thức tìm kiếm

Tìm kiếm thông tin nói chung là giải quyết các vấn đề như: biểu diễn, lưu trữ, tổ chức và truy cập đến các mục thông tin. Việc tổ chức và biểu diễn thông tin giúp người sử dụng dễ dàng truy cập thông tin mà mình quan tâm. Nhưng để mô tả các thông tin đó không phải là điều dễ dàng. Do vậy, hệ thống tìm kiếm thông tin bao gồm quá trình cơ bản sau: Biểu diễn nội dung các tài liệu, biểu diễn yêu cầu người dùng và so sánh hai biểu diễn này.

Quy trình biểu diễn tài liệu thường gọi là quá trình chỉ số hóa. Quá trình này có thể lưu trữ thực sự các tài liệu trong hệ thống nhưng thường chỉ lưu một phần tài liệu, chẳng hạn như phần tiêu đề, phần tóm tắt. Quá trình biểu diễn yêu cầu của người

dùng gọi là quá trình truy vấn. Truy vấn biểu thị sự tương tác giữa hệ thống và người sử dụng. Việc so sánh truy vấn với tài liệu cũng được gọi là quá trình đối sánh và cho kết quả là một danh sách các tài liệu được sắp xếp theo thứ tự mức độ liên quan với truy vấn.

Rõ ràng, để mô tả thông tin yêu cầu một cách đầy đủ, người sử dụng không thể trực tiếp yêu cầu thông tin sử dụng các giao diện hiện thời của hệ thống tìm kiếm. Thay vì người sử dụng đầu tiên phải chuyển đổi thông tin yêu cầu này thành một truy vấn mà có thể được xử lý bởi hệ thống tìm kiếm (hoặc hệ thống thu hồi thông tin (Information Retrieval - IR)). Thông thường, phép chuyển đổi này tạo ra một tập hợp các từ khoá (hoặc các term chỉ số) mô tả khái quát yêu cầu của người sử dụng.

Như vậy, việc tìm kiếm các tài liệu dựa trên nội dung thực sự của văn bản mà không phụ thuộc vào các từ khoá gắn với văn bản đó. Các công cụ tìm kiếm văn bản nổi tiếng hiện nay như Google, Altavista, Yahoo,…là những hệ tìm kiếm đưa ra danh sách các văn bản theo độ quan trọng của câu hỏi đưa vào. Để xây dựng một hệ tìm kiếm văn bản có hiệu quả cao, trước hết các văn bản và truy vấn ở dạng ngôn ngữ tự nhiên phải được tiền xử lý và chuẩn hoá.

Sau đây là hai mô hình chi tiết cho bộ công cụ tìm kiếm thông tin truyền thống và bộ công cụ tìm kiếm thông tin trên mạng.