ĐẠI HỌC THÁI NGUYÊN

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

Phùng Tuấn Anh

TÌM KIẾM DỮ LIỆU WEB VỚI NGÔN NGỮ XML

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

TÓM TẮT LUẬN VĂN

1. Thông tin chung:

Tên đề tài: Tìm kiếm dữ liệu Web với ngôn ngữ XML

Giáo viên hướng dẫn: PGS. TS Đỗ Trung Tuấn

Học viên thực hiện: Phùng Tuấn Anh

Lớp: Cao học K9A

Cơ sở đào tạo: Trường Đại học Công nghệ Thông tin và Truyền thông – Đại học Thái Nguyên

Chuyên ngành: Khoa học máy tính

Mã số chuyên ngành: 60 48 01.

2. Mục tiêu:

- Tìm hiểu, làm rõ một số nội dung về ngôn ngữ XML và khả năng sử dụng XML trong việc tìm kiếm dữ liệu.

- Ứng dụng ngôn ngữ XML trong việc tìm kiếm dữ liệu Web phục vụ công tác Quản lý nhà nước về lĩnh vực công thương.

3. Nội dung chính:

- Nghiên cứu tổng quan về đặc tính XML và các ngôn ngữ truy vấn XML: Xparth, XQuery, XSL

- Nhu cầu thực tế về tìm kiếm dữ liệu Web hiện nay trên Internet.

- Xây dựng chương trình đề mô, ứng dụng XML để tìm kiếm thông tin trên Website cho người sử dụng.

4. Kết quả đạt được:

Luận văn đã tìm hiểu những vấn đề liên quan đến dữ liệu và cơ sở dữ liệu, đặc biệt ngôn ngữ XML và cơ sở dữ liệu XML. Trong phần đầu luận văn, một số nhu cầu về sử dụng dữ liệu XML được trình bày.

Trong chương hai, luận văn trình bày một số khía cạnh của ngôn ngữ XML, đặc biệt những cú pháp và tìm kiếm dữ liệu XML. Việc này có ý nghĩa quan trong trong hệ thống các văn bản Internet đã và đang trở nên thông dụng.

Ngoài những tìm hiểu về hệ thống XML, luận văn đã thử nghiệm với một vài công việc thực tế tại Sở Công Thương tỉnh Thái Nguyên.

Đã xây dựng chương trình đề mô, ứng dụng XML để tìm kiếm thông tin trên Website cho người sử dụng.

MỤC LỤC

DANH MỤC CÁC TỪ VIẾT TẮT

4th Dimension, ANTs Data Server, Dataphor, Daffodil database, FileMaker Pro, Informix, InterBase, Matisse, Microsoft Access, Mimer SQL, NonStop SQL, Sand Analytic Server , SmallSQL, Sybase ASA, Watcom SQL, Sybase, Sybase IQ, Teradata, ThinkSQL, VistaDB	Các hệ quản trị cơ sở dữ liệu thương mại
ASCII	Bộ mã
B2B	Business to business
B2C	Business to customer
Cloudscape, Firebird, HSQLDB, Ingres, MaxDB, MonetDB, PostgreSQL, SQLite, tdbengine.	Các hệ quản trị cơ sở dữ liệu mã mở
CNTT	Công nghệ Thông tin
DB2	Hệ quản trị cơ sở dữ liệu của IBM
DBMS	Hệ quản trị cơ sở dữ liệu
DTD	Document Type Definition
ER	Mô hình thực thể quan hệ
GML	Generalized Markup Language
HTML	Hyper Text Markup Language
ICT	Công nghệ Thông tin và Truyền thông
IP	Giao thức IP
ISO	International Standard Organisation
MVS (Multiple Virtual Storage)	Lưu trữ ảo
NXD	Native XML database (Cơ sở dữ liệu XML nguyên gốc)
RDBMS	Hệ quản trị cơ sở dữ liệu quan hệ
RFC	Request for Comments
SGML	Standard Generalized Markup Language
SQL	Structured Query Language
SQL SERVER	Hệ quản trị cơ sở dữ liệu SQL SERVER
UCS	Universal Character System
URI	Uniform Resource Identifier
WML	Wireless Markup Language
XML	Ngôn ngữ đánh dấu mở rộng, XML

DANH MỤC CÁC HÌNH VẼ

LỜI CẢM ƠN

Tôi xin chân thành cảm ơn Ban giám hiệu, các Thầy Cô trường Đại học Công nghệ thông tin và Truyền thông Thái Nguyên, đặc biệt là các thầy cô trong và ngoài trường đã tận tình giảng dạy, trang bị cho tôi những kiến thức cần thiết trong suốt những năm học tập tại trường.

Em xin chân thành cảm ơn thầy Đỗ Trung Tuấn đã tận tình quan tâm, hướng dẫn và giúp đỡ em trong thời gian qua để em có thể hoàn thành tốt luận văn của mình.

Tôi xin chân thành cảm ơn các anh chị cùng các bạn đã có những nhận xét, đóng góp ý kiến, động viên, quan tâm và giúp đỡ tôi vượt qua khó khăn.

Cuối cùng, Tôi xin gửi lòng biết ơn sâu sắc đến cha mẹ, gia đình, Lãnh đạo Sở Công Thương đã tạo mọi điều kiện về thời gian, vật chất, tinh thần động viên, khích lệ và hỗ trợ tôi trong suốt thời gian qua./.

Thái Nguyên, ngày 12 tháng 9 năm 2012

LỜI NÓI ĐẦU

Thế kỷ XXI, thế kỷ của sự bùng nổ công nghệ thông tin, các công nghệ tiên tiến phát triển như vũ bão, mang một luồng gió mới thổi vào nhận thức của mỗi người. Song song đó, thế giới đang trong xu thế toàn cầu hóa, tất cả đều mang ý nghĩa hội nhập. Lúc này, các doanh nghiệp và chính phủ không chỉ cạnh tranh với các doanh nghiệp trong một quốc gia mà còn cạnh tranh với các doanh nghiệp, chính phủ ở khắp thế giới. Vì thế, để tồn tại và phát triển, mục tiêu mà các doanh nghiệp hướng đến đầu tiên là nâng cao chất lượng phục vụ khách hàng. Khách hàng là yếu tố sống còn của bất kỳ doanh nghiệp nào trong thế kỷ XXI. Một Chính phủ muốn quốc gia mình phát triển phải xem nhân dân và doanh nghiệp là khách hàng. Doanh nghiệp nào làm cho khách hàng thỏa mãn, doanh nghiệp đó sẽ phát triển tốt, Chính phủ nào làm nhân dân hài lòng Chính phủ đó sẽ vững mạnh.

Trong bối cảnh phát triển mạnh mẽ của Internet, thương mại điện tử, Chính phủ điện tử đã ra đời và phát triển khá nhanh, điều này không phải là mới trên thế giới nhưng vẫn là rất mới đối với Việt Nam. Nhiều vấn đề đặt ra là làm thế nào để ứng dụng công nghệ thông tin vào đời sống được hữu dụng nhất.

Chính vì vậy đề tài “Tìm kiếm dữ liệu Web với ngôn ngữ XML” được đưa ra nhằm giúp cho các doanh nghiệp hay các tổ chức nhà nước, hỗ trợ cho người dùng đạt hiệu quả. Với luận văn này, tôi mong muốn sẽ xây dựng được một hệ thống chương trình phần mền "hệ thống tìm kiếm dữ liệu WEB" để phục vụ cho việc tra cứu văn bản chuyên ngành của Sở Công Thương Thái Nguyên.

Trong quá trình tìm hiểu và viết về những đặc tính của XML sẽ không tránh khỏi những sai sót và hạn chế, rất mong được sự góp ý của Hội đồng và toàn thể những ai đã đọc luận văn này của tôi, để bản luận văn của tôi được hoàn thiện hơn nữa.

Luận văn được chuẩn bị theo các chương :

- Chương 1 trình bày nhu cầu xử lý dữ liệu và dẫn đến vai trò của dữ liệu XML và thông tin trên Internet. Các hệ thống quản trị cơ sở dữ liệu được đề cập để thấy được vai trò của chúng.
- Chương 2 trình bày một số khía cạnh của XML và thách thức đối với bài toán tìm kiếm thông tin trên Web. Những kiến thức tìm hiểu là cơ sở để phát triển ứng dụng trong chương sau.
- Chương 3 trình bày những ứng dụng thử nghiệm tại địa bàn Thái Nguyên. Một số trang màn hình cho phép thể hiện kết quả thử nghiệm.

Phần cuối là kết luận, trình bày các kết quả làm được và định hướng nghiên cứu tiếp.