BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG

BÙI THỊ THANH THỦY

NGHIÊN CỨU VÀ XÂY DỰNG HỆ THỐNG TÌM KIẾM CÔNG THỨC KHOA HỌC

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.TSKH. TRẦN QUỐC CHIẾN**

Phản biện 2: **TS. HOÀNG THỊ LAN GIAO**

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

Trong những năm gần đây, cùng với sự phát triển mạnh mẽ của mạng Internet và công nghệ Web là sự bùng nổ thông tin số. Số lượng người sử dụng và lượng thông tin sản sinh ra trên mạng Internet gia tăng rất nhanh và chúng ta có thể tìm thấy mọi thông tin cần thiết khi có nhu cầu. Đặc biệt, lượng thông tin liên quan đến khoa học, phục vụ học tập, nghiên cứu cũng gia tăng nhanh chóng và phong phú về lĩnh vực.

Việc khai thác hiệu quả các tài liệu khoa học trên Web có ý nghĩa quan trọng trong khoa học và kinh tế vì nó góp phần đáng kể vào việc cải thiện quá trình học tập và nghiên cứu. Nhu cầu tìm kiếm các công thức toán học trên môi trường web là rất lớn nhưng hiện nay chưa có hệ thống nào đáp ứng, kể cả các nhà cung cấp dịch vụ nổi tiếng như Google, Yahoo, Microsoft,… Việc nghiên cứu các giải pháp để hỗ trợ soạn thảo, lưu trữ và tìm kiếm các công thức toán học trên môi trường web là rất cần thiết .

Vì vậy tôi chọn đề tài " **Nghiên cứu và xây dựng hệ thống tìm kiếm công thức khoa học"**

Ý tưởng của luận văn là đề xuất một mô hình phù hợp với các tiêu chuẩn hiện hành và cho phép chúng ta có thể mô hình hóa, lưu trữ và tìm kiếm thuận lợi các công thức toán học, hóa học… trên các tài liệu, tài liệu khoa học, web.

2. Mục tiêu của đề tài

Tìm kiếm công thức khoa học vẫn còn rất hàn lâm. Nó phải đối mặt với nhiều vấn đề phức tạp của toán học vì thiếu các tiêu

chuẩn định nghĩa của một công cụ tìm kiếm khoa học đó như thế nào và bao gồm những tính năng gì.

Để tạo ra một công cụ tìm kiếm khoa học những vấn đề mới sau đây cần phải được giải quyết:

- Trích rút nội dung toán học, hóa học … từ các tài liệu khoa học.
- Phân loại văn bản bình thường và ký hiệu khoa học.
- Lập chỉ mục của văn bản khoa học và thiết kế thuật toán xếp hạng.
- Thiết kế ngôn ngữ truy vấn.

Mục đích để tài là đề xuất một phương pháp tiếp cận mới giải quyết gần như tất cả các vấn đề cơ bản ở trên và nhấn mạnh vào khả năng sử dụng thực tế. .

Điểm khác biệt giữa các công cụ tìm kiếm khoa học trước đây và công cụ tìm kiếm đề xuất trong luận văn đó là kỹ thuật chỉ mục tài liệu khoa học và thuật toán xếp hạng. Đề xuất một hệ thống tìm kiếm được thiết kế với kỹ thuật chỉ mục mới, lạ.

Các vấn đề được trình bày trong luận văn:

- Vấn đề kết nối tìm kiếm khoa học, trình bày các giải pháp hiện tại và lý do tại sao các giải pháp đó không phù hợp với bộ sưu tập tài liệu lớn.
- Thiết kế hệ thống tìm kiếm khoa học.
- Đề xuất kỹ thuật chỉ mục các ký hiệu khoa học và thuật toán xếp hạng.
- Đề xuất ngôn ngữ truy vấn khoa học.
- Tổng kết dựa trên kết quả đánh giá.

3. Đối tượng nghiên cứu

Đối tượng nghiên cứu khi thực hiện đề tài này là các công cụ hỗ trợ soạn thảo, các tiêu chuẩn lưu trữ, phương pháp hiển thị và tìm kiếm công thức khoa học.

4. Phương pháp nghiên cứu

Tìm hiểu lý thuyết về soạn thảo, lưu trữ và tìm kiếm công thức khoa học trên văn bản. Nghiên cứu các công cụ tìm kiếm khoa học hiện có. Tìm hiểu chuẩn MathML để đặc tả các công thức toán học, hóa học...

Dựa trên lý thuyết đã nghiên cứu, tiến hành xây dựng một ứng dụng soạn thảo công thức, lưu trữ, hỗ trợ tìm kiếm công thức khoa học trên web, tài liệu khoa học ở tất cả các định dạng.

5. Ý nghĩa khoa học và thực tiễn

Việc khai thác hiệu quả các tài liệu khoa học trên Web có ý nghĩa quan trọng trong khoa học và kinh tế vì nó góp phần đáng kể vào việc cải thiện quá trình học tập và nghiên cứu. Kết quả của đề tài sẽ góp phần quan trọng trong việc xử lý các công thức khoa học trong các hệ thống hỗ trợ dạy và học qua mạng internet, diễn đàn khoa học…

6. Cấu trúc luận văn

Bố cục của luận văn được tổ chức thành 3 chương.

Chương 1: Trình bày nghiên cứu tổng quan về công cụ tìm kiếm khoa học.

Chương 2: Được dành để mô tả ứng dụng, xây dựng mô hình tổng quát, đề xuất giải pháp xây dựng môi trường soạn thảo công thức, tiêu chuẩn lưu trữ và tìm kiếm công thức khoa học.

Chương 3: Trình bày triển khai ứng dụng .

CHƯƠNG 1
NGHIÊN CỨU TỔNG QUAN

Công thức toán, công thức hóa, công thức vật lý…gọi chung là công thức Khoa học. Và các công thức khoa học này đều được biểu diễn dưới dạng toán học. Cho nên trong luận văn này, tôi tập trung đi sâu vào phân tích công thức toán học.

1.1. SOẠN THẢO CÔNG THỨC TOÁN HỌC

Với mỗi trình soạn thảo văn bản có một chuẩn lưu trữ khác nhau vì vậy sẽ gặp rất nhiều khó khăn trong việc hợp nhất các văn bản được tạo ra từ nhiều ứng dụng khác nhau.

1.1.1. Định dạng lưu trữ

a. LaTex

LaTex định nghĩa một chế độ đặc biệt để soạn thảo công thức toán học. Các công thức này có thể được đưa vào ngay trong môi trường văn bản ta có thể tách rời chúng khỏi các đoạn văn bản. Phần nội dung toán học trong đoạn văn có thể được soạn thảo ở giữa dấu \ và \ hay $ và $ hay \ begin{math} và \end{math}.

b. HTML

HTML là một ngôn ngữ đánh dấu được thiết kế để tạo nên các trang web, nghĩa là các mẫu thông tin được trình bày trên Word Wide Web.

c. MathML

MathML là một ứng dụng của XML để thể hiện ký hiệu và công thức toán học với mục đích rộng là phương cách trao đổi thông tin toán học trên máy tính (để hiển thị cũng như để tính toán) và mục đích hẹp là hiển thị tài liệu toán học trên World Wide Web. Nhóm

toán học của W3C đề xuất mọi người nên dần sử dụng ngôn ngữ này trên mạng.

MathML được thiết kế để không chỉ hiển thị tốt công thức toán học mà còn chứa ý nghĩa hiểu nội dung toán học.

1.1.2. Biểu diễn soạn thảo

Ở đây, chúng ta sẽ tìm hiểu ở bốn loại văn bản thường sử dụng hiện nay.

- Biểu diễn công thức toán học trên Microsoft Word
- Biểu diễn công thức toán học trên Website
- Biểu diễn công thức toán học trên OpenOffice.Org
- Biểu diễn công thức toán học trên MathType

a. Biểu diễn công thức toán học trên Microsoft Word

Ví dụ trong Microsoft word 2003, để hiển thị công thức $\sqrt{x}$, soạn thảo bằng phương trình Editor thì phải sử dụng đoạn mã sau:

```
{ EQ\r(2,x)}
```

Equation Editor (soạn thảo phương trình) là một trình soạn thảo công thức phát triển bởi Design Science. Cho phép người dùng xây dựng các phương trình toán học trong môi trường WYSIWYG, được tích hợp trong tất cả các phiên bản Microsoft Office.

b. Biểu diễn công thức toán học trên Website

Trang trực tuyến CodeCogs Equation Editor sử dụng ngôn ngữ đánh dấu Latex để soạn thảo công thức toán học. Đây là trang soạn thảo dạng WYSIWYG và hầu hết tất cả các trình duyệt có thể đọc được ví như Mozilla Firefox, Internet Explorer,…

Hiển thị hình ảnh $\sqrt{x}$ thì chúng ta phải nhập đoạn mã sau:

```
\sqrt [x]{2}\
```

c. Biểu diễn công thức toán học trên OpenOffice.Org

Đối với bộ OpenOffice.Org, việc tạo ra một công thức toán học trên trình soạn thảo là rất đơn giản thông qua OpenOffice.Org Math. Ta chỉ việc sử dụng bảng lựa chọn các ký hiệu cần chèn và công thức tương ứng với đoạn mã được sinh ra.

Để hiển thị công thức $\sqrt{x}$ thì chúng ta có đoạn mã tương ứng sau:

```
nroot{2} {x}
```

d. Biểu diễn công thức toán học trên MathType

MathType là một phần mềm thương mại của Design Science cho phép tạo các ký hiệu toán học để xử lý văn bản. Hỗ trợ các ngôn ngữ đánh dấu như Tex, Latex, Mathml,…cho nên khi sử dụng MathType, chúng ta có thể chuyển công thức khoa học sang các dạng chuẩn khác nhau.

1.2. TỔNG QUAN VỀ CÔNG CỤ TÌM KIẾM

1.2.1. Khái niệm cơ bản về hệ tìm kiếm thông tin

Bản chất của quá trình tìm kiếm thông tin dựa trên cơ chế "đối sánh" các tài liệu được lưu trữ trong hệ thống với yêu cầu tìm kiếm của người dùng để tìm ra kết quả cho phù hợp .

1.2.2. Mô hình của hệ tìm kiếm thông tin

Các thành phần cơ bản bao gồm:

- Đầu vào: gồm các câu truy vấn và các tài liệu
- Đầu ra: Tập hợp các đoạn trích hay tài liệu được hệ thống đánh giá phù hợp với yêu cầu truy vấn của người dùng :
- Bộ xử lý:

Bao gồm các thành phần cơ bản :

Bộ biểu diễn câu truy vấn : Biểu diễn câu truy vấn người dùng.

Bộ biểu diễn tài liệu : Biểu diễn thông tin về tài liệu, lập chỉ mục phục vụ cho quá trình tìm kiếm.

Bộ đối sánh : Đối sánh giữa câu truy vấn và thông tin tài liệu xem tài liệu có phù hợp không

1.2.3. Các thành phần của hệ tìm kiếm thông tin

- Thu thập thông tin web (web crawler)
- Lập chỉ mục cho các tài liệu thu thập được
- Tìm kiếm

1.3 .THỰC TRẠNG TÌM KIẾM CÔNG THỨC KHOA HỌC

Hàng này, chúng ta thường sử dụng chức năng tìm kiếm tài liệu trong công việc nhưng thực tế đó chỉ là tìm kiếm văn bản dưới dạng chuỗi. Vậy tìm kiếm văn bản dưới dạng công thức thì sao? Đây là một vấn đề mà chúng ta ít quan tâm nhưng rất quan trọng.

Công thức được định dạng như một tập tin hình ảnh. Vì lý do đó, chúng ta không thể thực hiện quá trình tìm kiếm giống như tìm kiếm chuỗi trong văn bản thông thường mà chúng ta hay sử dụng.

1.3.1. CiteSeer

Citeseer là một hệ thống thu thập thông tin dữ liệu chỉ mục các bài báo khoa học được thu thập về từ trên web dựa trên nội dung và phần trích dẫn của bài báo, sau đó tổ chức thông tin dữ liệu chỉ mục thu thập được thành cơ sở dữ liệu, cho phép người dùng tìm kiếm thông tin trên dữ liệu chỉ mục này.

Một số điểm chưa được:

- CiteSeer không thể đánh dấu dữ liệu chỉ mục với các tạp chí chưa có các bản điện tử.
- CiteSeer chưa xử lý được những dữ liệu như sự nhập nhằng giữa 2 tác giả cùng một tên.

1.3.2. Google Scholar

Google Scholar là một công cụ chuyên tìm kiếm tài liệu nghiên cứu và học thuật, bao gồm các bài báo khoa học, bài tóm tắt khoa học, bài nghiên cứu sơ bộ, bài tóm tắt, báo cáo kỹ thuật, luận án, sách v.v…

Ưu điểm của Google Scholar:

- Khả năng lọc thông tin từ những nguồn đáng tin cậy.
- Tìm kiếm thông tin từ những nguồn web mở lẫn web thương mại. Do đó, người dùng có thể tìm được toàn văn của tài liệu nghiên cứu từ những nguồn miễn phí trên mạng Internet hoặc tìm được thông tin thư mục (biểu ghi) của các nguồn cơ sở dữ liệu trả tiền.
- Cung cấp công cụ hỗ trợ việc đánh giá tính đáng tin cậy của nguồn tài liệu tìm được.
- Cho phép mở rộng phạm vi tìm kiếm trên kết quả tìm**.**

1.3.3. LeActiveMath

ActiveMath là một trang web thông minh - Môi trường học toán học. Nội dung ngữ nghĩa của các tài liệu toán học được mã hóa ở OMDoc.

Công cụ tìm kiếm nhận biết toán học chuyên dụng này phù hợp cho người học toán và tìm kiếm toán học ở mức độ đơn giản. Khả năng sử dụng cho các tìm kiếm phức tạp hơn là vấn đề khó.

1.3.4. MathDex

Các tính năng chính :

- Hỗ trợ cả tìm kiếm công thứctoán học và tìm kiếm văn bản đơn giản.
- Hỗ trợ phong phú các định dạng đầu vào.

Nhược điểm:

- Giao diện người dùng không hỗ trợ đồ họa.
- Các thuật toán lập chỉ mục chỉ sử dụng phần đặc biệt của ngữ nghĩa có sẵn và tần số xuất hiện công thức.

1.3.5. MathWebSearch

MathWebSearch tìm công thức dựa trên quan đểm ngữ nghĩa.

Hiện nay MathWebSearch có thể chỉ mục tài liệu trong MathML Nội dung và OpenMathformat. Việc hỗ trợ MathML Trình bày bị giới hạn và do đó cũng hạn chế tính sử dụng của nó. Lý do là Trình bày không bao gồm dữ liệu ngữ nghĩa của các ký hiệu toán học.

1.3.6. XML Searching

Hiệu quả của tìm kiếm XML vẫn còn là một lĩnh vực nghiên cứu học tập. XML cơ bản xử lý ngôn ngữ truy vấn XQuery và XPath. Cả hai đều rất mạnh mẽ nhưng đều không thể thực hiện cho bộ sưu tập lớn các tài liệu với các tính năng:

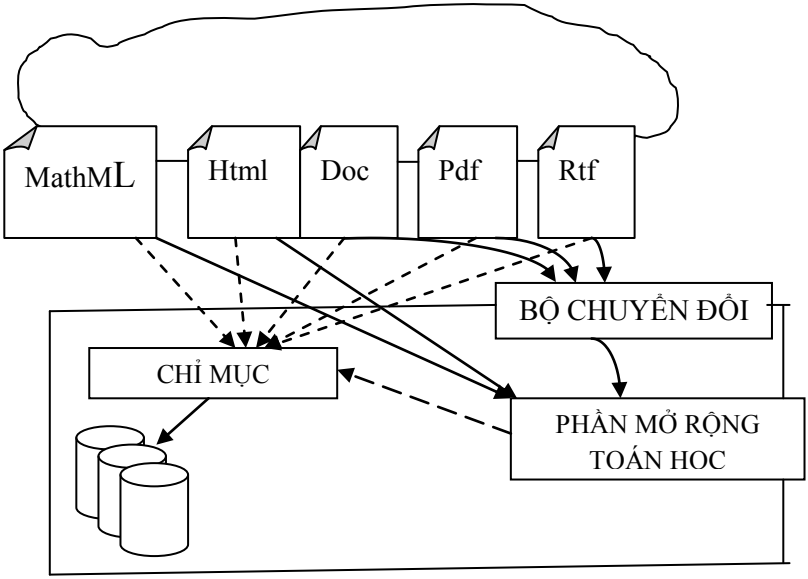
- Biến đổi phức tạp - cần thiết cho các ký hiệu toán học.
- Hỗ trợ văn bản đầy đủ.

CHƯƠNG 2
ĐỀ XUẤT GIẢI PHÁP

2.1. MÔ HÌNH TỔNG QUAN

Tóm tắt các tính năng chính:

- Nhận biết toán học.
- Có khả năng tìm kiếm các công thức đồng dạng và các công thức con.
- Hỗ trợ cả nội dung lẫn trình bày.
- Ngôn ngữ truy vấn đơn giản, có thể mở rộng nhưng rõ ràng.



Hình 2.1: Thiết kế của công cụ tìm kiếm toán học đề nghị

Giải pháp được đề xuất dựa trên một công cụ tìm kiếm toàn văn.

Một tài liệu toán học có thể được tách thành hai phần khác nhau.

- Phần đầu tiên là phần văn bản.
- Phần thứ hai là phần toán học.

Bằng cách này, tất cả những lợi thế của một công cụ tìm kiếm văn bản đầy đủ có thể được sử dụng với khả năng giới hạn tìm kiếm theo nội dung toán học. Một tìm kiếm cho một "chứng minh" của một công thức có thể cho kết quả tìm kiếm "chứng minh" từ trong đoạn văn bản và công thức trong phần toán học. Công cụ tìm kiếm đề xuất xem xét ký hiệu toán học như một số loại siêu dữ liệu rất quan trọng của tài liệu được lập chỉ mục và được sử dụng trong bảng xếp hạng.

Để khai thác tất cả các tính năng hiện tại đề xuất công cụ tìm kiếm toán học được thiết kế như là một phần mở rộng của một công cụ tìm kiếm văn bản bỏ qua tất cả các chức năng không liên quan. Bằng cách này, nền tảng lý thuyết của các giải pháp được đề xuất có thể được thực hiện trong bất kỳ công cụ tìm kiếm hiện tại. Hai thay đổi được yêu cầu là:

- Sự hỗ trợ của một ngôn ngữ truy vấn toán học ở đầu vào.
- Sử dụng các dữ liệu toán học khi xếp hạng các tài liệu

Các thiết kế của phần mở rộng toán học trong giai đoạn lập chỉ mục được minh họa trong hình 2.1. Phần mở rộng của toán học là một trung gian giữa các tài liệu toán học và phần lập chỉ mục.

2.2. XÂY DỰNG TỆP CHỈ MỤC

Phần quan trọng nhất của một công cụ tìm kiếm là lập chỉ mục và gần như tất cả các vấn đề cần phải được giải quyết đều nằm trong giai đoạn chỉ mục. Giai đoạn lập chỉ mục có thể được chia hai giai đoạn:

- Phần đầu tiên bao gồm các thuật toán lập chỉ mục

Các thuật toán lập chỉ mục của công cụ tìm kiếm toàn văn phải sử dụng một kỹ thuật nhận dạng ngôn ngữ - chịu trách nhiệm phân tích các phần văn bản và phân tích chúng một cách chính xác. Kỹ thuật tương tự cũng được sử dụng trong công cụ tìm kiếm toán học và được gọi là kỹ thuật nhận dạng công thức. Sau khi công thức được nhận dạng nó được biến đổi thành các phép biểu diễn riêng bao gồm một hoặc nhiều từ.

- Giai đoạn thứ hai xác định xếp hạng các kết quả.

Thông thường có nhiều kết quả có liên quan đều được sắp xếp bởi một thuật toán xếp hạng. Các số liệu thống kê cho thấy rằng người sử dụng công cụ tìm kiếm không nhìn vào hơn hai mươi kết quả và do đó các thuật toán xếp hạng là rất quan trọng.

Có rất nhiều cấu trúc dữ liệu chỉ mục mà có thể được sử dụng để lưu trữ dữ liệu (chỉ số nghịch đảo, cây hậu tố, ngram, cấu trúc cây, ...).

Các cấu trúc dữ liệu phổ biến nhất là các chỉ số nghịch đảo. Cấu trúc này cho phép tìm kiếm văn bản đầy đủ. Tất cả các công cụ tìm kiếm văn bản đầy đủ đều sử dụng loại cấu trúc dữ liệu này.

Có vài yếu tố rất quan trọng phải được xem xét khi thiết kế chỉ mục:

- Các yếu tố phải hợp nhất.
- Kỹ thuật lưu trữ.
- Kích thước chỉ mục.
- Tốc độ tra cứu.
- Duy trì chỉ mục theo thời gian.
- Sai số cho phép.

2.2.1. Thuật toán chỉ mục

Những lý do để lựa chọn MathML như là định dạng được hỗ trợ chính được mô tả trong Chương 1. Nó hỗ trợ cả trình bày và nội dung MathML. Ví dụ về định dạng tài liệu MathML được thể hiện trong Bảng 2.1

MathML Nội dung	MathML Trình diễn
<apply>	<msup>
<power/>	<mfenced/>
<apply/>	<mi>a</mi>
<plus>	<mo>+</mo>
<ci>a</ci>	<mi>b</mi>
<ci>b<ci>	</mfenced>
</apply>	<mn>2</mn>
<cn>a</cn>	</msup>
</apply>	

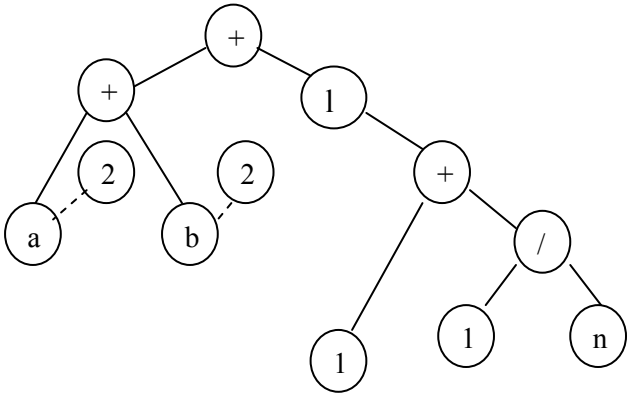
Bảng 2.1: Ví dụ đánh dấu MathML đơn giản

Định dạng nội dung sao chép cấu trúc phân cấp của một công thức toán học và do đó có thể được chuyển đổi tự nhiên vào cây. Định dạng trình bày rất tuyến tính và nó là mơ hồ (ab cũng có thể có nghĩa là a * b).

Mỗi công thức toán học bao gồm các toán hạng và các toán tử. Các toán tử là hàm số với tham số, đối số. Toán hạng được phân thành nhiều loại:

- Hằng số 0, 100, -104, ...
- Hằng số phổ biến không đổi - Π, e ...(Lưu ý: chỉ có vài ký hiệu được định nghĩa trước như là hằng số, vì cũng có nhiều ký hiệu thường được sử dụng có nghĩa là một cái gì đó khác trong lĩnh vực toán học khác).

- Toán tử phổ biến được biết đến - ∑, ∫...
- Biến không xác định –a, F, ...



Hình 2.2: Cây công thức cho $a^2 + b^2 + \ln (1 + 1 / n)$

Cây công thức – Cấu trúc dữ liệu cây có kết nối không định hướng, trong đó mỗi nút có tối đa hai con. Lá mà có con được gọi là các nút tâm. Tất cả các nút khác được gọi là các nút lá. Các nút tâm chứa các hàm số toán học (+, /, ln, ...). Các nút lá có thể chứa số, các hằng số và các biến. Mỗi nút có thể có một hoặc nhiều công thức toán con kết nối với nó giống như chỉ số hoặc số mũ. Độ sâu của tất cả các nút phải được lưu trữ. Ví dụ x + y + 10 / x sẽ có ba cấp độ, với +, + như nút gốc. Nút x, y, / như các nút con và 10, x là con của /. Độ sâu phải đáp ứng một điều kiện là tại cùng một cấp độ sâu thì có thể giao hoán.

Toán tử - Nút tâm của cây công thức.

Toán hạng của một Toán tử - con của một toán tử.

Những toán tử quan trọng nhất - Tất cả các nút với mức độ sâu thấp nhất.

Những toán hạng quan trọng nhất - Tất cả các toán hạng của các toán tử quan trọng nhất.

Công thức biểu diễn tiêu chuẩn - Kết quả của thuật toán xếp hạng được sử dụng trên cây công thức.

Từ quan điểm toán học, trong giai đoạn này một hàm Q được xây dựng.

Miền xác định là không gian của tất cả các công thức toán học (F) và miền giá trị là F^N: = $F_1 \times F_2 \times$, ..., $\times F_N$. Công thức f_1, f_2, ..., f_N là kết quả của công thức đầu vào. Hàm Q có thể được định nghĩa: Q: $F \rightarrow F^N$, Q (f) = <f_1, ..., f_N>. Hàm số Q phải đáp ứng một trong những điều kiện trong miền F^N: cho tất cả giá trị i hợp lệ, F_{i+1} là một tổng quát của F_i.

Thuật toán chính được sử dụng trong phần mở rộng toán học thực hiện các bước sau:

Đầu vào: *Công thức đầu vào – f*

Đầu ra: $f_1, f_2, \ldots f_N, f_i$ *– công thức*

$f_0 := f;$

for i = 1 to N do

$f_i := Q_i (f_{i-1});$

$f_i := xephang (f_i);$

end

Thuật toán 1: Thuật toán chính

- Ưu điểm đầu tiên là kỹ thuật này không cần dấu ngoặc đơn.
- Ưu điểm thứ hai là kỹ thuật này cho phép một loại tìm kiếm tương tự như minh họa ở ví dụ sau đây.

Công thức (a + b) - (c + d) được chuyển đổi thành ab + cd + -.

Giả định rằng khóa công thức là ab + và cd + và vì thế cơ sở dữ liệu chỉ mục có chứa 3 từ theo thứ tự này: ab +, cd +, -. Điều này

cho phép tìm kiếm các công thức con (ab+, cd+) mà không cần biết các toán tử giữa chúng.

a. Quy tắc chuyển đổi

Danh sách các quy tắc chuyển đổi:

- Ước tính giá trị ở mỗi mức

Ví dụ: 7 + a + 5 → 12 + a

- Phép xấp xỉ của hằng số

Ví dụ: 3,14 = 3

- Hủy bỏ dấu ngoặc sử dụng phép phân phối

Ví dụ: a * (b + c) → a * b + b * c

- Nhân phân phối

Ví dụ:

$$\frac{a+b}{2} * \prod \rightarrow \frac{\prod a + \prod b}{2}$$

- Chia mỗi tử số cho mẫu số riêng của nó

Ví dụ :

$$\frac{\prod a + \prod b}{2} \rightarrow \frac{\prod a}{2} + \frac{\prod b}{2}$$

- Thay thế các hằng số với biểu tượng const

Ví dụ :

$74 + a^2 + b^2$ -> const $+ a^{const} + b^{const}$

Hoặc -> const $+ a^2 + b^2$

- Thay thế các hằng số chưa biết và ẩn số bằng biểu tượng *id*

Ví dụ :

$a^2 - b^2 - c^2 + 2bc\ \mathbf{cos}\alpha \rightarrow id_1^2 - id_1^2 - id_1^2 + 2id_1id_2\mathbf{cos}id_3$

hoặc $\rightarrow id_1^2 - id_2^2 - id_3^2 + 2id_1id_2\mathbf{cos}id_3$

hoặc $\rightarrow id_1^2 - id_2^2 - id_3^2 + 2id_2id_3\mathbf{cos}id_3$

Tên biến có thể khác nhau. Tìm kiếm cho x + y cũng như a + b hoặc R + S.

- Quy tắc đơn giản hóa

- Ràng buộc và biến số của toán tử không cần xét đến.

b. Thuật toán xếp hạng

Ưu tiên các toán tử cơ bản +, -, *, /. Sau đó các thuật toán xếp hạng sử dụng thứ tự này:

1. Hằng số nhỏ nhất.

2. Hằng số với thứ tự được xác định trước (e, Π, ...), nếu đẳng thức sử dụng toán tử ưu tiên.

3. Những toán tử được biết với thứ tự được xác định (∫, ln….), nếu đẳng thức sử dụng toán tử ưu tiên.

4. Công thức con khác được xếp theo quy tắc sau:

(a) xếp theo toán tử ưu tiên , trước tiên : + x^2, thứ hai : - x^2

(b) xếp theo việc đếm toán hạng - các toán tử với các toán hạng ít nhất thì khởi đầu

(c) nếu toán tử ưu tiên và số lượng của các toán hạng bằng nhau cho hai công thức con sau đó xếp chúng đệ quy theo các quy tắc:

i. ưu tiên cho các toán tử khác nhau .

ii. nếu chính xác công thức con chứa hằng số thì nó là đầu tiên.

iii. nếu công thức con chứa nhiều hằng số chung thì nó là đầu tiên.

iv. nếu công thức con chứa nhiều toán tử biết đến thì nó là đầu tiên

v Các trường hợp khác: sử dụng so sánh chuỗi

c. Chứng minh thuật toán xếp hạng

Các thuật toán xếp hạng phải đáp ứng hai điều kiện.

- Thứ nhất, hai công thức toán học tương đương với các toán hạng như nhau nhưng hoán vị bằng cách sử dụng giao hoán phải có cách biểu diễn tiêu chuẩn như nhau.

- Và thứ hai, hai công thức không tương đương hoặc tương đương nhưng với các toán hạng khác nhau phải có cách biểu diễn tiêu chuẩn khác nhau.

d. Mô tả chính thức

Một mô tả chính thức của thuật toán:

Đầu vào : Đầu vào công thức – f, tập hợp các quy tắc biến đổi $T_1,\ldots,T_7$

Đầu ra: tập hợp các công thức f_1, f_2, f_3, f_4, f_5.

f_0: = f;

Q_1 : = T_1,T_2;

Q_2 : = T_3,T_4, T_5;

Q_3 : = T_6;

Q_4 : = T_7;

f_1: = xephang(f);

for i = 1 to 4 do

  f_{i+1} := xephang (Q, (f_i));

end

Thuật toán 2: Thuật toán tổng quát

2.2.2. Xếp hạng kết quả

Vì công cụ tìm kiếm toán học tạo ra các từ cũng được xếp hạng. Nếu hai tài liệu đều chứa từ khóa giống nhau nhưng cách biểu diễn khác nhau, $f_i = f_j$; i#j; i<j, thì tài liệu f_i sẽ được xếp hạng cao hơn.

Gọi R là xếp hạng của 1 từ, thì điều kiện xếp hạng của những từ công thức có thể được viết như sau: $R(f_1) \geq R(f_2) \geq \ldots \geq R(f_N)$. Nền tảng toán học của phương pháp này đó chính là "tìm kiếm tương tự".

2.3. THU THẬP, NHẬN BIẾT CÔNG THỨC

Nhận biết công thức nhằm phân loại văn bản đơn giản và ký hiệu khoa học.

Mỗi tài liệu được xử lý bằng công cụ tìm kiếm riêng của mình trong các định dạng tài liệu mà công cụ tìm kiếm văn bản đầy đủ hỗ trợ. Nếu một tài liệu có chứa ký hiệu toán học, nó được chuyển đổi sang một định dạng được hỗ trợ (MathML) và sau đó giao cho phần mở rộng toán học được trích rút chỉ có công thức toán học từ nội dung. Nội dung được đánh dấu là toán học.

2.4. TRÍCH RÚT NỘI DUNG TOÁN HỌC

Trích rút và phân loại nội dung của một tài liệu toán học bao gồm cả chuyển đổi tài liệu sang định dạng được hỗ trợ và phân tích cú pháp định dạng.

Sau khi phân tích một số định dạng (OpenMath, MathML, LATEX, TEX, XML, pdf, ps, HTML) có thể chứa nội dung toán học hoặc trình bày của toán học. Và do rất thích hợp cho chỉ mục toán học, MathML đã được lựa chọn như là định dạng chính.

Lý do là:

- Định dạng có thể mã hóa toán học.
- Nó có thể được nhúng trong HTML.
- Định dạng dễ chuyển đổi.

Và cuối cùng nó được mở rộng bởi Tập đoàn W3C

2.5. TÌM KIẾM

Giai đoạn cuối của một công cụ tìm kiếm là tương tác người sử dụng. Người sử dụng phải xây dựng một truy vấn tìm kiếm trong đó quy định cụ thể nhu cầu tìm kiếm.

Vài kết quả đầu tiên (khoảng hai mươi) là quan trọng nhất và kết quả được sắp xếp theo một thuật toán xếp hạng. Giai đoạn tìm kiếm phải thực hiện các bước sau:

- Phân tích cú pháp truy vấn. (*)
- Ánh xạ toán tử truy vấn để hỗ trợ cấu trúc nội bộ.(**)
- Tìm kiếm tất cả các từ / cụm từ trong truy vấn.
- Đánh giá tính logic của các truy vấn và thu thập các tài liệu phù hợp.
- Sắp xếp chúng theo xếp hạng và phù hợp với truy vấn.
- Trả về danh sách kết quả.

2.5.1. Ngôn ngữ truy vấn

Một trong những ngôn ngữ lập trình được sử dụng nhiều nhất là LATEX. Để cho phép người dùng trực quan, ngôn ngữ truy vấn tìm kiếm được đề xuất trong nghiên cứu này tương tự như LATEX.

2.5.2. Thực hiện truy vấn

Truy vấn	Ý nghĩa
a^{2}+b^{2}	$a^2 + b^2$
\Pi(a+b)	$\pi(a + b)$
\int e^ {-x^{1}}	$\int e^{-x}$
\lambda+y= \ Pi*r^{2}	$\Lambda + y = \pi * x^2$

2.5.3. Phương pháp, thuật toán so khớp

Gọi miền xác định của hàm biểu diễn câu truy vấn q là Q, tập hợp các câu truy vấn có thể có; và miền giá trị của nó là R, không gian thống nhất biểu diễn thông tin. Gọi miền xác định của hàm biểu

diễn tài liệu d là D, tập hợp các tài liệu; và miền giá trị của nó là R. Miền xác định của hàm so sánh c là R× R và miền giá trị của nó là [0,1], tập các số thực từ 0 đến 1. Trong một hệ thống tìm kiếm lý tưởng:

c(q(query), d(doc)) = j(query, doc) , query Q, doc D

Với j: Q × D -> [0,1] biểu diễn việc xử lý của người dùng về mối quan hệ của thông tin trên câu truy vấn và thông tin trong tài liệu.

2.5.4. Hiển thị kết quả

Giai đoạn tìm kiếm của hệ tìm kiếm toán học phải hiển thị ít nhất một tóm tắt nhỏ của văn bản được trích từ tài liệu

2.6. ĐÁNH GIÁ

Có nhiều cách đo lường khác nhau cho việc đánh giá mức độ xử lý trả về kết quả của một hệ thống tìm kiếm thông tin. Các cách đo lường đều đòi hỏi một tập tài liệu và một câu truy vấn trên tập tài liệu đó, giả sử rằng mỗi tài liệu có thể liên quan hoặc không liên quan đến câu truy vấn.

Đánh giá dựa trên các thông số:

- Độ chính xác
- Độ bao phủ
- Kết quả sai

CHƯƠNG 3
TRIỂN KHAI ỨNG DỤNG

3.1. THIẾT KẾ HỆ THỐNG

3.1.1. Đặc tả hệ thống:

Hệ thống gồm 3 modul

Module Web Robot. Chức năng chính là thu thập dữ liệu từ các trang Web trên các Website.

Module ConvertFile. Công việc chính của module này là thống nhất lại toàn bộ dữ liệu trước khi chuyển sang định dạng MathML thực hiện việc lập Chỉ mục (Index). Việc thống nhất này bao gồm:

- Chuyển toàn bộ các định dạng (PDF, DOC, EXCEL...) sang dạng MathML.
- Chuyển mã của trang (TCVN3, VNI, Unicode...) sang bảng mã Windows-CP1258 (Unicode tổ hợp).
- Lưu dữ liệu vào trong CSDL để index.

Modul Truy vấn

3.1.2 Thiết kế các chức năng của hệ thống.

a. Module Web Robot.

Gồm các thành phần sau:

- CollectEngine
- DomainCollector.
- DataCollector.

b. Module Convert File

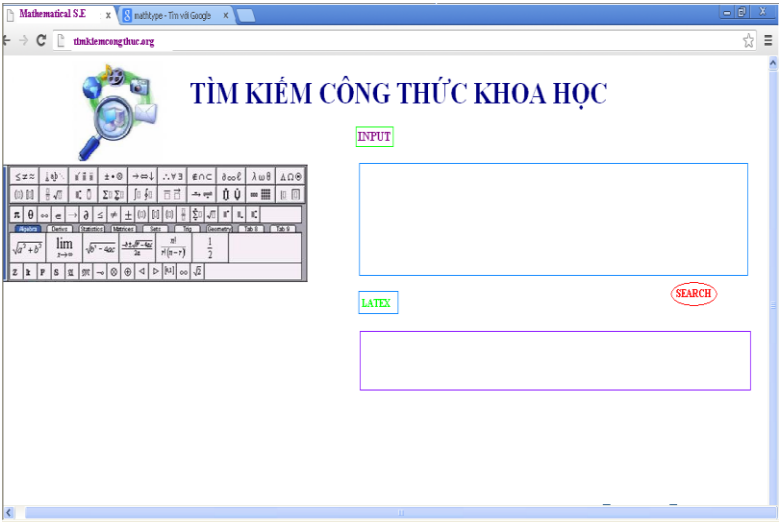
- Chuyển định dạng
- Chuyển bảng mã
- Ghép nối công đoạn chuyển định dạng và chuyển mã

3.2. XÂY DỰNG MÔI TRƯỜNG SOẠN THẢO

3.2.1. Giao diện Module Web Robot

3.2.2. Giao diện Module Convert File

3.2.3. Giao diện Module truy vấn



KẾT LUẬN

Nhu cầu tìm kiếm các tài liệu khoa học, những luận văn, đồ án chuyên sâu…là rất lớn. Vì vậy việc nghiên cứu tạo ra những hệ thống tìm kiếm chuyên biệt tài liệu khoa học hoàn thiện, hỗ trợ tốt cho ngôn ngữ tiếng Việt là rất cần thiết. Điều này có ý nghĩa cao cả về mặt khoa học lẫn thực tiễn.

Về mặt lý thuyết, trong quá trình thực hiện đề tài tôi đã nghiên cứu môi trường và công cụ soạn thảo công thức toán học đang sử dụng hiện nay. Từ đó phân tích thực trạng và khó khăn đối với người sử dụng. Đề xuất thiết kế chỉ mục và thuật toán xếp hạng cho tìm kiếm công thức khoa học. Nghiên cứu tìm hiểu ngôn ngữ đánh dấu MathML. Từ đó đã đem lại những kết quả khởi đầu cho việc xây dựng một hệ thống tìm kiếm công thức khoa học Việt Nam trong tương lai.

Về mặt thực tiễn, hệ thống còn trong giai đoạn chưa hoàn thiện. Nhưng bước đầu đã xây dựng được các module tìm kiếm. Là cơ sở để phát triển hệ thống tìm kiếm công thức khoa học mong đợi.

Vì thời gian có hạn nên trong luận văn, tệp chỉ mục được xây dựng còn nhỏ gọn, chưa đầy đủ nên khi tiến hành tìm kiếm còn bị giới hạn.

Dựa vào kết quả này, tôi sẽ nghiên cứu thêm để hoàn chỉnh hệ thống. Hướng nghiên cứu tiếp theo là xây dựng một kho dữ liệu lớn để phục vụ bài toán tìm kiếm công thức khoa học, tìm kiếm nội dung công thức và tìm kiếm chuyên sâu đến các tài liệu khoa học có công thức liên quan.