BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG

TRẦN THỊ BÍCH LIỄU

NGHIÊN CỨU VÀ PHÁT TRIỂN CÔNG CỤ TÌM KIẾM CÔNG THỨC TOÁN HỌC TRÊN VĂN BẢN

Chuyênngành: KHOA HỌC MÁY TÍNH
Mãsố: 60.48.01.01

TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT

ĐàNẵng - Năm 2015

Công trình được hoàn thành tại

ĐẠI HỌC ĐÀ NẴNG

Người hướng dẫn khoa học: **PGS.TS. Võ Trung Hùng**

Phản biện 1: TS. Nguyễn Tấn Khôi

Phản biện 2: TS. Nguyễn Quang Thanh

Luận văn đã được bảo vệ trước Hội đồng chấm Luận văn tốt nghiệp Thạc sĩ Kĩ thuật họp tại Đại học Đà Nẵng vào ngày 18 tháng 7 năm 2015

Có thể tìm hiểu luận văn tại:

- Trung tâm Thông tin-Học liệu, Đại học Đà Nẵng
- Thư viện trường Đại học Bách Khoa, Đại học Đà Nẵng

MỞ ĐẦU

1. Lý do chọn đề tài

Ngày nay khoa học công nghệ có ảnh hưởng rất sâu rộng đến mọi lĩnh vực của đời sống xã hội ở mọi quốc gia. Thông tin điện tử đã trở thành một nguồn tài nguyên quý giá phục vụ cho hoạt động quản lý và mọi nhu cầu của đời sống xã hội. Hầu hết những nội dung, tri thức mà con người khám phá tra đều đã tồn tại dưới dạng văn bản điện tử, từ các nội dung khoa học tự nhiên - xã hội, đến cả các tin tức, chia sẻ hằng ngày của con người trên khắp thế giới.

Cũng như các lĩnh vực khác, ngày càng có nhiều người chia sẻ các nội dung toán học và tìm kiếm giải pháp cho các vấn đề của họ trên mạng Internet. Tuy nhiên, một vấn đề đặt ra là làm sao có thể tìm kiếm được các nội dung toán học cần thiết trong một kho tài liệu khổng lồ như vậy? Các bộ máy tìm kiếm văn bản bình thường không nhận diện được các kí hiệu, cấu trúc đặc biệt, do đó việc tìm kiếm thường không trả về kết quả khớp với yêu cầu người dùng. Chính vì vậy cần có một bộ máy tìm kiếm công thức toán học chuyên dụng cho phép tìm kiếm các công thức toán học trên các tài liệu và website được chia sẻ trên mạng Internet. Hiện nay trên thế giới đã phát triển một số công cụ tìm kiếm công thức toán học cho phép tìm theo nội dung hiển thị của công thức hoặc theo ngữ nghĩa của nó tuy nhiên phạm vi ứng dụng của các công cụ này còn bó hẹp, chẳng hạn như EgoMath cho phép tìm kiếm công thức toán học trên Wikipedia.org, Website LatexSearch có hỗ trợ tìm kiếm các công thức toán học được soạn thảo bằng ngôn ngữ đánh dấu LaTeX, đây là bản quyền của MPS Technologies (*Mathematical Programming System - Hệ thống lập trình toán học*), nhưng những kết quả tìm thấy chỉ giới hạn

trên những tài liệu điện tử trên máy chủ SpringerLink, vv.

Lĩnh vực nghiên cứu phát triển công cụ tìm kiếm công thức toán học có thể được xem là mới mẻ và cần một đầu tư nghiên cứu chuyên sâu để tìm ra một giải pháp thích hợp nhất. Xuất phát từ nhu cầu thực tế đó, chúng tôi đã chọn đề tài "***Nghiên cứu và phát triển công cụ tìm kiếm công thức toán học trên văn bản***". Việc nghiên cứu phát triển công cụ này nhằm hỗ trợ, thúc đẩy việc học tập, nghiên cứu và ứng dụng khoa học tư nhiên tại Việt Nam và trên thế giới.

2. Mục tiêu và nhiệm vụ

Mục tiêu của nghiên cứu này là nhằm cung cấp một website tìm kiếm công thức toán học trên văn bản. Website này cho phép người dùng nhập công thức toán học cần tìm kiếm và hệ thống sẽ hiển thị danh mục các tài liệu chứa công thức toán học đó

3. Đối tượng và phạm vi nghiên cứu

3.1. Đối tượng nghiên cứu

Đối tượng nghiên cứu trong luận văn này là các tiêu chuẩn đặc tả công thức toán học trên văn bản sử dụng ngôn ngữ đánh dấu LaTeX và MathML, phương pháp tạo chỉ mục cho các tài liệu cần tìm kiếm, một số ứng dụng hỗ trợ tạo chỉ mục phổ biến hiện nay, phương pháp tìm kiếm công thức toán học trên văn bản, các công cụ tìm kiếm công thức toán học sẵn có và phương pháp tích hợp công cụ gõ công thức toán học vào công cụ tìm kiếm.

3.2. Phạm vi nghiên cứu

Thực hiện nghiên cứu phát triển ứng dụng tìm kiếm trên kho dữ liệu chứa 50 tài liệu toán học định dạng PDF hoặc XHTML. Ứng dụng tìm kiếm và cơ sở dữ liệu được triển khai

trên máy đơn chạy trên localhost.

4. Phương pháp nghiên cứu

4.1. Nghiên cứu lý thuyết

Nghiên cứu các tài liệu liên quan đến tìm kiếm công thức toán học trên văn bản và thử nghiệm các công cụ tìm kiếm công thức toán học sẵn có.

4.2. Nghiên cứu thực nghiệm

Dựa trên lý thuyết đã nghiên cứu, tiến hành xây dựng một công cụ tìm kiếm công thức toán học trên văn bản sử dụng ngôn ngữ lập trình Java;

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

5.1. Về mặt lý thuyết

Luận văn chỉ ra một hướng đi mới trong việc nghiên cứu xây dựng một công cụ tìm kiếm công thức toán học hiệu quả trên văn bản, cũng như tạo tiền đề để xây dựng các công cụ hoàn chỉnh hơn trong tương lai.

5.2. Về mặc thực tiễn

Luận văn cung cấp một công cụ tìm kiếm công thức toán học trên văn bản, giúp học sinh, sinh viên, giáo viên… tiết kiệm thời gian, công sức cũng như đạt được hiệu quả cao hơn trong công tác học tập và nghiên cứu các môn khoa học tự nhiên.

6. Bố cục của luận văn

Ngoài phần mở đầu và kết luận, luận văn gồm có ba chương với các nội dung chính như sau:

Chương 1: Nghiên cứu tổng quan

Trong chương này, chúng tôi sẽ trình bày những tìm hiểu sơ bộ về thực trạng tìm kiếm công thức toán học tại Vệt Nam và trên thế giới. Đồng thời, chúng tôi cũng tìm hiểu về cách đặc tả công thức toán học

trên tài liệu và website sử dụng hai ngôn ngữ đánh dấu là LaTeX và MathML. Ngoài những nội dung trên, chúng tôi cũng sẽ tìm hiểu về phương pháp tạo chỉ mục và một số ứng dụng hỗ trợ tạo chỉ mục phổ biến hiện nay, trong đó nổi bật là ứng dụng mã nguồn mở Apache Lucene. Ngoài ra, chúng tôi sẽ nghiên cứu một số công cụ tìm kiếm công thức toán học sẵn có như MathWebSearch, Nutch, EgoMath.

Chương 2: Giải pháp đề xuất

Trong chương này, chúng tôi sẽ đề xuất được mô hình tổng quát của hệ thống và đề xuất các giải pháp cụ thể nhằm giải quyết các yêu cầu bài toán đặt ra, bao gồm: giải pháp sử dụng InftyReader để chuyển đổi định dạng tập tin PDF sang XHTML+MathML, giải pháp chuẩn hóa công thức toán học, giải pháp tích hợp bộ công cụ WIRIS nhằm hỗ trợ nhập công thức vào khung tìm kiếm.

Chương 3: Thực nghiệm và đánh giá kết quả

Trong chương này, chúng tôi sẽ tiến hành xây dựng hệ thống tìm kiếm công thức toán học trên văn bản, ứng dụng mã nguồn mở Lucene, và tiến hành thử nghiệm, đánh giá những kết quả đạt được.

CHƯƠNG 1
NGHIÊN CỨU TỔNG QUAN

1.1. TỔNG QUAN VỀ TÌM KIẾM CÔNG THỨC TOÁN HỌC TRÊN CÁC VĂN BẢN

1.1.1. Khái niệm văn bản toán học

Trong phạm vi luận văn này, văn bản toán học được dùng để chỉ các văn bản có chứa các công thức toán học. Nó không đơn thuần chỉ là các tài liệu về lĩnh vực toán học, mà còn có thể là các tài liệu vật lí, hóa học, sinh học.

1.1.2. Thực trạng tìm kiếm công thức toán học trên văn bản hiện nay

a. Trên thế giới

- Đối với những người không chuyên, chẳng hạn như muốn tìm kiếm công thức $\sqrt{x}$, họ sẽ chuyển công thức này thành "The square root of x" và tìm bằng cụm từ này các bộ máy tìm kiếm văn bản phổ biến như Bing, Google, vv.

- Hướng tìm kiếm thứ hai là tìm kiếm theo tên của công thức.

- Hướng tìm kiếm thứ ba sử dụng các bộ máy tìm kiếm chuyên dụng cho công thức toán học chẳng hạn như MathWebSearch.

b. Tại Việt Nam

Việc tìm kiếm tài liệu toán học tại Việt Nam có thể nói khó khăn hơn so với mặt bằng chung trên thế giới.

1.2. ĐẶC TẢ CÔNG THỨC TOÁN HỌC TRÊN MÁY TÍNH

1.2.1. Tổng quan về đặc tả công thức toán học trên tài liệu

Công thức toán học trên tài liệu được có thể đặc tả bằng nhiều ngôn ngữ khác nhau, gọi là ngôn ngữ đánh dấu toán học. Có nhiều loại ngôn ngữ đánh dấu toán học hiện nay, phổ biến nhất là 4 loại chính:

- TeX/LaTeX [16]
- MathML [17]
- OMDoc [18]
- OpenMath [19]

Trong đó, TeX/LaTeX có cú pháp gần gũi với ngôn ngữ tự nhiên, trong khi MathML, OpenMath và OMDoc lại tối ưu hóa cho việc giao tiếp giữa các máy tính với nhau.

1.2.2. Đặc tả công thức toán học bằng ngôn ngữ LaTeX

TeX là một hệ thống sắp chữ được viết bởi Donald E.Knuth ở Đại học Stanford vào năm 1977. TeX được xem là cách tốt nhất để gõ công thức toán học phức tạp nhằm phục vụ nhu cầu soạn thảo các tài liệu toán học với chất lượng bản in cao. Bắt đầu từ năm 1980, Leslie Lamport bắt đầu tạo ra hệ thống soạn thảo văn bản ngày nay gọi là LaTeX dựa trên định dạng của TeX. Việc tạo ra tập tin PDF từ tập nguồn LaTeX được thực hiện dễ dàng nhờ vào các công cụ chuyển đổi.

1.2.3. Đặc tả công thức toán học bằng MathML

a. Giới thiệu chung về MathML

MathML (*Mathematical Markup Language - Ngôn ngữ Đánh dấu Toán học*) 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. MathML cung cấp hai cách thức trình bày ngôn ngữ đánh dấu toán học, một cách thức nhằm nhấn mạnh cách trình bày của công thức (Presentation MathML) và cách thức thứ hai nhấn mạnh nội dung của toán học của công thức đó (Content MathML).

b. Presentation MathML

Presentation MathML tập trung vào mặt hiển thị của một công thức, nó có hơn 30 thẻ khác nhau. Tên của các thẻ đều bắt đầu bằng kí tự m. Một biểu thức Presentation MathML được xây dựng từ các thẻ kết hợp với nhau sử dụng các thẻ ở cấp cao hơn để nhằm điều khiển bố cục của chúng.

c. Content MathML

Content MathML tập trung vào nghĩa của công thức hơn là cách trình bày của công thức đó. Khác với Presentation MathML, các toán tử trong Content MathML được biểu diễn bằng những thẻ có ngữ nghĩa, chẳng hạn phép chia được biểu diễn bằng thẻ <devide/>, phép mũ được biểu diễn bằng thẻ <power/>. Có hơn 100 loại thẻ cho các hàm và toán tử khác nhau.

d. Kết luận

MathML được tạo ra nhằm mục đích rộng là 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. MathML trở thành một ngôn ngữ có nhiều triển vọng trong tương lai để biểu diễn công thức toán học trên website và các phần mềm.

1.3. TỔNG QUAN VỀ PHƯƠNG PHÁP TẠO CHỈ MỤC TRÊN TÀI LIỆU

1.3.1. Phương pháp tạo chỉ mục trên tài liệu

Lập chỉ mục là quá trình phân tích và xác định các từ, cụm từ thích hợp cốt lõi có khả năng đại diện cho nội dung của tài liệu. Như vậy, vấn đề đặt ra là phải rút trích ra những thông tin chính, có khả năng đại diện cho nội dung của tài liệu. Việc rút trích này chính là việc lập chỉ mục trên tài liệu.

1.3.2. Một số ứng dụng hỗ trợ tạo chỉ mục

a. Xapian

Xapian là một thư viện mã nguồn mở tìm kiếm, được phát hành dưới giấy phép GPL (*General Public License - Giấy phép công cộng*) [20]. Nó được viết bằng C++, nhưng cũng có thể được tích hợp sử dụng từ Perl, Python, PHP, Java, Tcl, C# và một số ngôn ngữ lập trình khác. Xapian có tính tương thích cao, nó có thể được dễ dàng tích hợp vào các ứng dụng khác, nó cũng hỗ trợ các toán tử tìm kiếm logic chẳng hạn như AND, OR, NOT, vv.

b. Ứng dụng mã nguồn mở Apache Lucene

Lucene là phần mềm mã nguồn mở, dùng để phân tích, đánh chỉ mục và tìm kiếm thông tin với hiệu suất cao bằng Java. Lucene không phải là một ứng dụng hoàn chỉnh mà chỉ là một thư viện, nó cung cấp các thành phần quan trọng nhất của một máy tìm kiếm đó là tạo chỉ mục và truy vấn. Mặc dù thiết kế và xây dựng ban đầu từ Java nhưng hiện nay cũng đã có một số phiên bản cho các ngôn ngữ khác : .NET, C++, Perl, vv.

1.4. MỘT SỐ CÔNG CỤ TÌM KIẾM CÔNG THỨC TOÁN HỌC SẴN CÓ

1.4.1. MathWebSearch

MathWebSearch là một bộ máy tìm kiếm công thức toán học dựa trên ngữ nghĩa của công thức, được phát triển tại Đại họcJacobs. Hệ thống này tạo chỉ mục cho các công thức MathML và OpenMath, sử dụng kỹ thuật chỉ mục Substitution Tree Indexing.

1.4.2. LeActiveMath

LeActiveMath là một ứng dụng hỗ trợ học tập có khả năng tương tác được phát triển bởi ActiveMath group. LeActiveMath thực

hiện lập chỉ mục cho các tài liệu OMDoc, trong đó các công thức toán học được mã hóa bằng OpenMath.

1.4.3. Egomath

Egomath là một công cụ tìm kiếm toán học phát triển tại Đại học Charles ở Prague. Nó có thể tìm kiếm các công thức toán học viết bằng LaTeX và văn bản đơn giản. EgoMath có thể xử lý được văn bản và các công thức toán học viết bằng LaTeX hoặc MathML. Cơ sở dữ liệu chính thức của các Egomath là Wikipedia (phiên bản tiếng Anh) từ năm 2011 trở đi.

TIỂU KẾT CHƯƠNG 1

Trong chương này, chúng tôi đã trình bày những tìm hiểu sơ bộ về thực trạng tìm kiếm công thức toán học tại Vệt Nam và trên thế giới.Đồng thời, chúng tôi cũng tìm hiểu về cách đặc tả công thức toán học trên tài liệu và website sử dụng hai ngôn ngữ đánh dấu là LaTeX và MathML. Đây là hai trong số những ngôn ngữ đánh dấu toán học phổ biến nhất hiện nay. Với những ưu điểm nổi bật của nó, hai ngôn ngữ này đang ngày được sử dụng rộng rãi trong các tài liệu khoa học và trên các trang web. Ngoài những nội dung trên, chúng tôi cũng đã tìm hiểu về phương pháp tạo chỉ mục và một số ứng dụng hỗ trợ tạo chỉ mục phổ biến hiện nay, trong đó nổi bật là ứng dụng mã nguồn mở Apache Lucene. Ngoài ra, chúng tôi sẽ nghiên cứu một số công cụ tìm kiếm công thức toán học sẵn có như MathWebSearch, Nutch, EgoMath.

CHƯƠNG 2
ĐỀ XUẤT GIẢI PHÁP

2.1. MÔ TẢ ỨNG DỤNG

Xuất phát từ nhu cầu thực tiễn cần có một công cụ để tìm kiếm công thức toán học trên văn bản, chúng tôi đề xuất xây dựng một ứng dụng tìm kiếm công thức trên một kho chứa các tài liệu toán học ở các định dạng PDF và XHTML. Từ quan điểm người dùng, ứng dụng đáp ứng một số yêu cầu như sau:

- Ứng dụng cho phép tìm kiếm được tài liệu ở các định dạng PDF và XHTML.

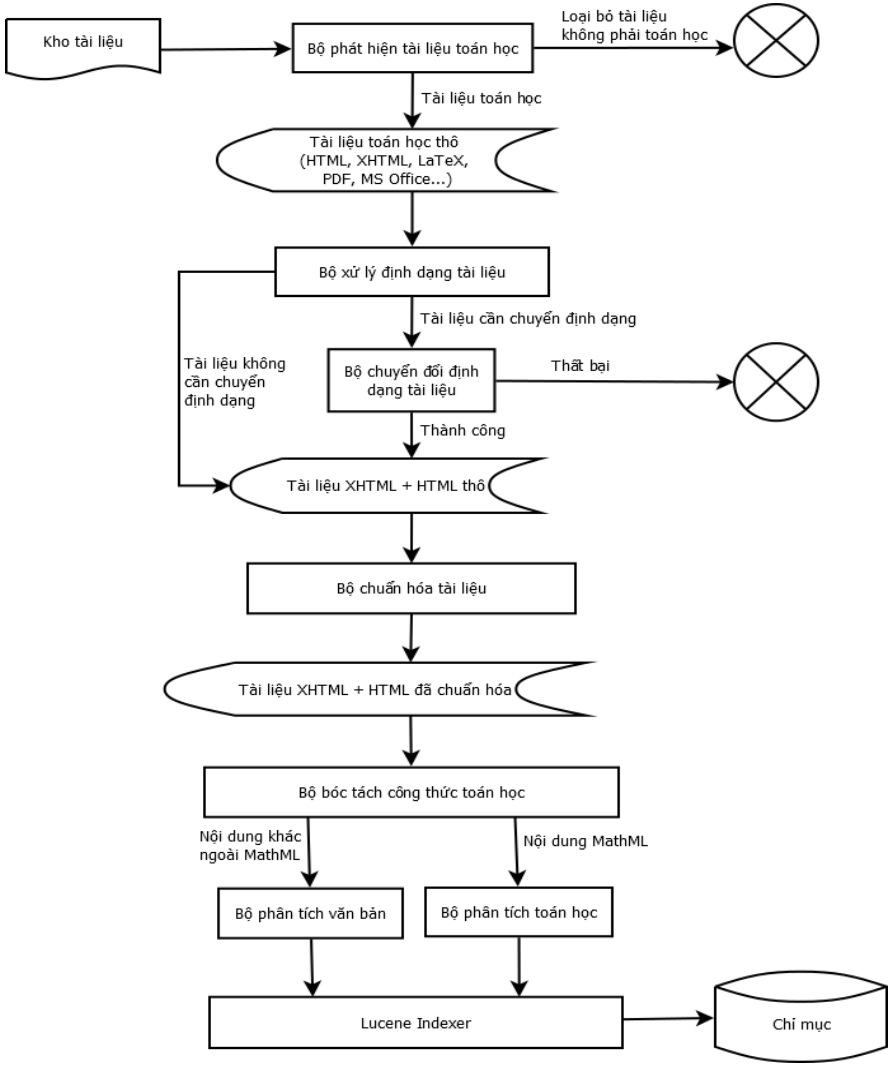
- Cho phép người dùng nhập công thức toán học một cách trực quan từ khung tìm kiếm.

- Cho phép tìm kiếm tài liệu toán học dựa trên nội dung tìm kiếm chứa đồng thời văn bản và công thức. Chẳng hạn người dùng có thể nhập "Pythagoras formula$a^2 + b^2 = c^2$" để tìm kiếm nội dung chính xác hơn.

- Ứng dụng xếp hạng kết quả trả về cho người dùng theo thứ tự giảm dần theo độ trùng khớp với câu truy vấn của người dùng.

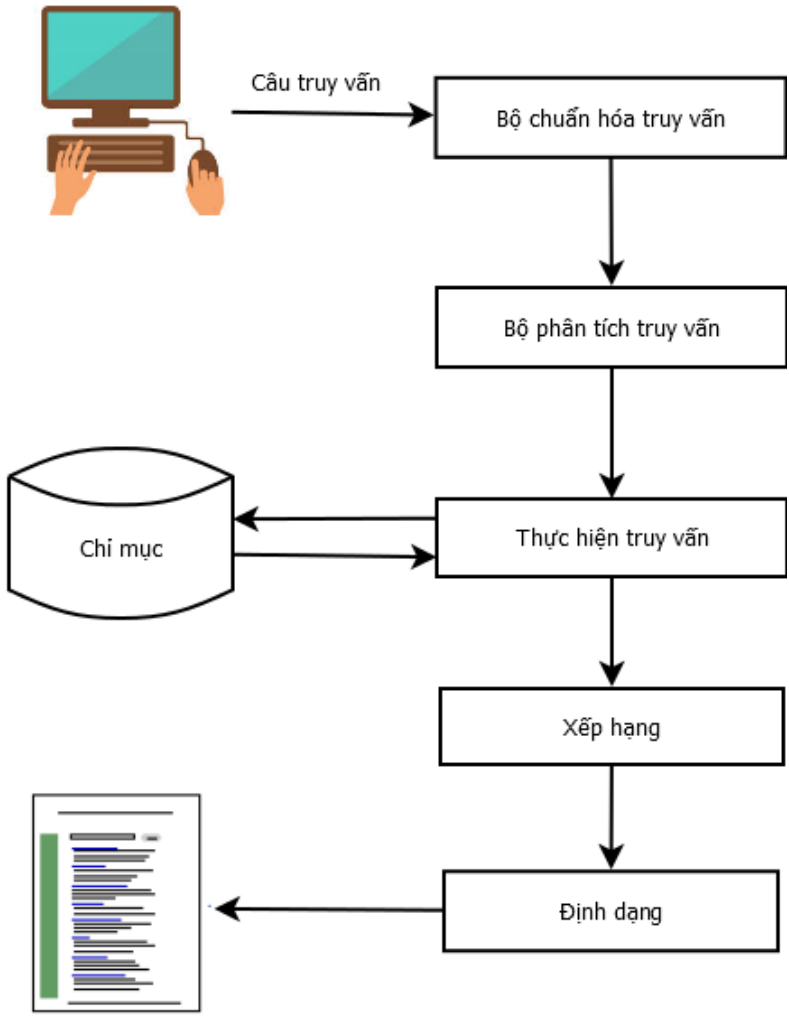
2.2. MÔ HÌNH TỔNG QUÁT

2.2.1. Mô hình tạo chỉ mục



Hình 2.1. Mô hình giải pháp lập chỉ mục

2.2.2. Mô hình tìm kiếm



Hình 2.2. Mô hình giải pháp tìm kiếm

2.3. ĐỀ XUẤT GIẢI PHÁP

2.3.1. Giải pháp chuyển đổi định dạng công thức toán học

Hệ thống của chúng tôi cho phép tìm kiếm trên các định dạng tài liệu PDF vàXHTML. Để tạo chỉ mục trên tập tài liệu này, chúng tôi sẽ chuyển đổi chúng về một định dạng thống nhất là XHTML+MathML sử dụng phần mềm InftyReader.

2.3.2. Giải pháp chuẩn hóa công thức toán học

a. Khái niệm chuẩn hóa

Chuẩn hóa là bước chuyển đổi các công thức toán học MathML có định dạng khác nhau (nhưng ý nghĩa giống nhau) về một định dạng chung.

*Ví dụ:*Loại bỏ các thuộc tính không cần thiết trong thẻ <mfrac>

```
    <mfrac linethickness=”2”
bevelled=”true”>
        <mi> a </mi>
        <mi> b </mi>
    </mfrac>
```

Thuộc tính `linethickness=”2” bevelled=”true”` chỉ có tác dụng hỗ trợ định dạng khi hiển thị công thức trên trình duyệt. Việc bỏ đi những thuộc tính này hoàn toàn không làm ảnh hưởng đến ý nghĩa của công thức. Do đó ta có thể tối ưu hóa công thức này thành:

```
    <mfrac>
      <mi> a </mi>
      <mi> b </mi>
    </mfrac>
```

b. Các bước trong quá trình chuẩn hóa

- *Loại bỏ các thành phần và các thuộc tính không cần thiết*

Có nhiều thành phần trong MathML được sử dụng trong Presentation MathML giúp ích rất nhỏ hoặc hầu như không có đóng góp gì vào việc lập chỉ mục và tìm kiếm công thức toán học đó. Chẳng hạn như các thẻ quy định giao diện hiển thị của công thức như <mspace>, <mpadded>, <mphantom>, <maligngroup>, <malignmark>, vv. Do đó cần loại bỏ những cặp thẻ này nhằm tối ưu hóa tốc độ tìm kiếm, cũng như trả về kết quả tìm kiếm chính xác hơn.

- *Loại bỏ các thực thể ẩn*

Thực thể ẩn là những thực thể không hiển thị trên trình duyệt khi công thức. Nó chỉ có ý nghĩa làm rõ ý nghĩa của công thức đó. Chẳng hạn công thức $ax^2 + bx + c$ sử dụng toán tử nhân vô hình *⁢* nhằm biểu diễn phép nhân vô hình giữa a và x^2 trong biểu thức ax^2, và giữa *b* và *x* trong biểu thức *bx*.

Có ba thực thể vô hình cần loại bỏ:

Bảng 2.1. Các thực thể vô hình trong Presentation MathML

Tên thực thể	**Mã nguồn**	**Ví dụ**
U+2062 INVISIBLE TIMES: Phép nhân vô hình	⁢	*xy*
U+2063 INVISIBLE SEPARATOR: Dấu cách vô hình	⁣	m_{12}
U+2064 INVISIBLE PLUS: Phép cộng vô hình	&InvisiblePlus;	$4\frac{1}{2}$

Ví dụ 1: Biểu diễn dấu nhân vô hình giữa *x* và *y* trong biểu thức *xy* (nhằm tránh hiểu nhầm *xy* là tên một biến)

```
<mrow>
   <mi>x</mi>
      <mo>&InvisibleTimes;</mo>
   <mi>y</mi>
</mrow>
```

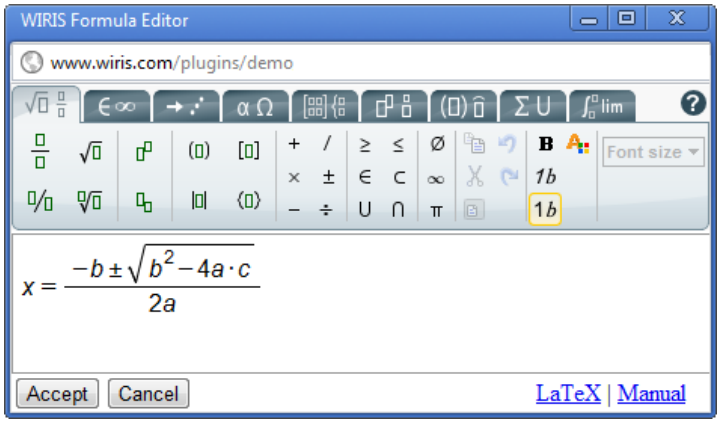
2.3.3. Giải pháp phân tích cú pháp và tạo chỉ mục

Đầu tiên nội dung tài liệu sẽ được phân tách thành nội dung văn bản và nội dung toán học. Các nội dung văn bản được lập chỉ mục theo cách thông thường. Còn các công thức toán học sau khi đã hoàn thành bước chuẩn hóa sẽ được chuyển đổi thành một chuỗi nén (chuỗi nén là chuỗi không có xuống dòng, không có khoảng trống trong chuỗi) mà có thể được lập chỉ mục như một chuỗi văn bản bình thường.

2.3.4. Giải pháp tích hợp công cụ gõ công thức toán học

Trên giao diện ứng dụng, người dùng có thể gõ công thức toán học trực tiếp vào khung tìm kiếm nhờ tích hợp một bộ công cụ gõ công thức toán học gọi là WIRIS. WIRIS là tập hợp các công cụ JavaScript giúp người dùng nhập và chỉnh sửa công thức toán học, trong đó có trình biên soạn WIRIS là một trình biên soạn trực quan, hay còn gọi là WYSIWYG (*What You See Is What You Get*).Kết quả trả về của công thức được lưu trữ dưới dạng Presentation MathML.

Dưới đây là giao diện của công cụ gõ công thức toán học WIRIS:



Hình 2.6. Giao diện công cụ gõ công thức toán học WIRIS

TIỂU KẾT CHƯƠNG 2

Trong chương này, chúng tôi đã đề xuất được mô hình tổng quát của hệ thống, trong đó hệ thống gồm hai thành phần chính là: thành phần tạo chỉ mục và thành phần tìm kiếm. Hai thành phần này sử dụng chức năng chuẩn hóa toán học như một bước đệm để chuẩn hóa tài liệu trước khi tạo chỉ mục hoặc chuẩn hóa câu truy vấn trước khi tìm kiếm. Đồng thời, trong chương này, chúng tôi cũng đã đề xuất được các giải pháp cụ thể nhằm giải quyết các yêu cầu bài toán đặt ra, bao gồm: giải pháp sử dụng InftyReader để chuyển đổi định dạng tập tin PDF sang XHTML+MathML, giải pháp chuẩn hóa công thức toán học, giải pháp tích hợp bộ công cụ WIRIS nhằm hỗ trợ nhập công thức vào khung tìm kiếm và giải pháp xếp hạng kết quả tìm kiếm theo mô hình xếp hạng động. Những mô hình và giải pháp đề xuất trong chương này là cơ sở để chúng tôi triển khai xây dựng ứng dụng ở chương 3.

CHƯƠNG 3
TRIỂN KHAI ỨNG DỤNG

3.1. MÔ HÌNH HỆ THỐNG

3.1.1. Đặc tả chức năng

Một hệ thống tìm kiếm thông thường phải đầy đủ 3 thành phần cơ bản: bộ thu thập thông tin, thành phần tạo chỉ mục và thành phần tìm kiếm. Tuy nhiên vì thời gian có hạn, nên chúng tôi chỉ tập trung xây dựng thành phần tạo chỉ mục và thành phần tìm kiếm, sử dụng một kho dữ liệu có sẵn trên máy tính cá nhân thay vì xây dựng thêm bộ thu thập thông tin để thu thập tài liệu trên Internet. Thành phần tạo chỉ mục và thành phần tìm kiếm sử dụng chung một hệ thống con, là thành phần chuẩn hóa toán học:

- Thành phần chuẩn hóa toán học

Thành phần này có chức năng chuẩn hóa tài liệu XHTML và chuẩn hóa câu truy vấn dạng MathMLnhư loại bỏ các thẻ, các thuộc tính không cần thiết, vv.

- Thành phần tạo chỉ mục

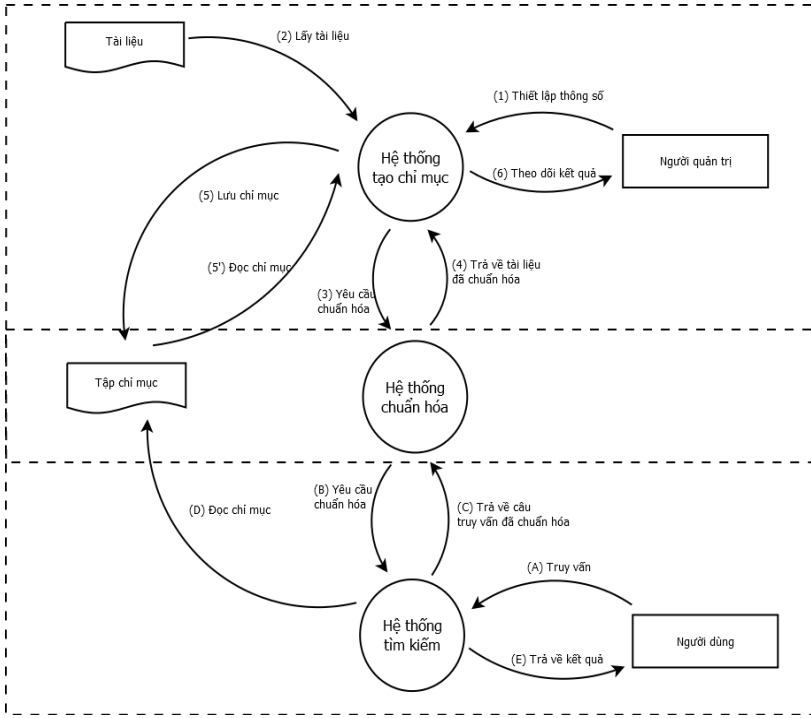
Thành phần này thuộc phần quản trị hệ thống, bao gồm các chức năng chính như chỉ định dữ liệu lập chỉ mục, thực hiện phân tích tài liệu, tạo chỉ mục và lưu trữ xuống tập chỉ mục, vv.

- Thành phần tìm kiếm

Thành phần tìm kiếm thuộc giao diện người dùng cuối, bao gồm các chức năng chính như: nhận thông tin truy vấn, biên dịch và tìm kiếm, trình bày kết quả và liên kết đến tài liệu gốc.

3.1.2. Sơ đồ luồng dữ liệu hệ thống

Sơ đồ luồng dữ liệu (DFD - Data Flow Diagram) giữa các hệ thống như sau:



Hình 3.1. Sơ đồ luồng dữ liệu giữa các hệ thống con

3.2. CÀI ĐẶT HỆ THỐNG

3.2.1. Hệ thống chuẩn hóa

Hệ thống chuẩn hóa được thiết kế bao gồm 2 module tương ứng với 2 nội dung cần chuẩn hóa: Module chuẩn hóa các thẻ và các thuộc tính và Module chuẩn hóa toán tử.Chúng tôi sử dụng đối tượng XMLStreamReader để đọc các thẻ từ mã nguồn MathML, và dùng đối tượng XMLStreamWriter để ghi kết quả.

3.2.2. Hệ thống lập chỉ mục

Sau khi đã được chuẩn hóa bằng bộ chuẩn hóa, bộ lập chỉ mục sẽ tiến hành lập chỉ mục trên tài liệu sử dụng lớp IndexWriter.

3.2.3. Hệ thống tìm kiếm

a. Tích hợp bộ gõ WIRIS vào khung tìm kiếm

Để tích hợp bộ gõ WIRIS vào khung tìm kiếm, chúng tôi tiến hành tải WIRIS plugin từ địa chỉ http://www.wiris.com/plugins/docs/editors/generic, sau đó tiến hành chỉnh sửa đường dẫn tới các thư mục tương ứng.

b. Xây dựng bộ tìm kiếm

Tiến hành tìm kiếm tài liệu sử dụng lớp IndexSearcher mặc định của Lucene.

3.3. THỬ NGHIỆM HỆ THỐNG

3.3.1. Thu thập và chuyển hóa dữ liệu

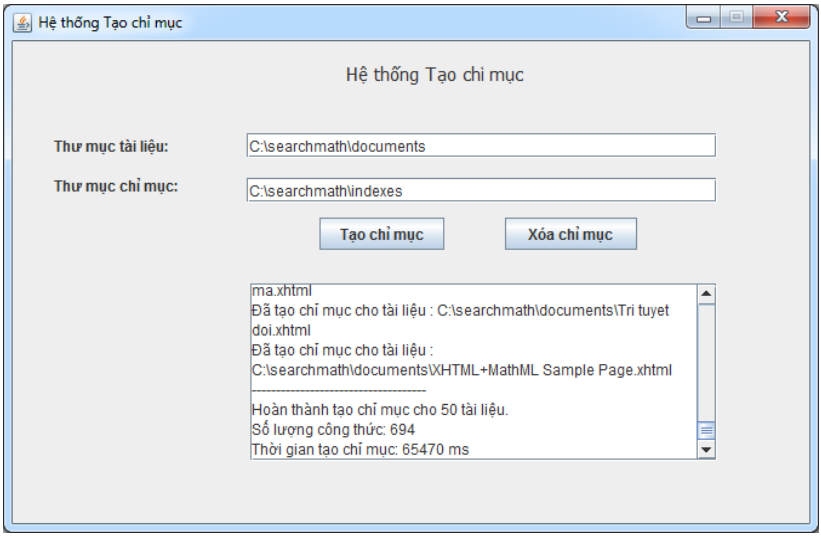
Cơ sở dữ liệu hiện tại bao gồm 50 tập tin toán học, trong đó có 10 tập tin XHTML+MathML, và có 40 tập tin toán học định dạng PDF. Những tập tin PDF này sẽ được chuyển đổi sang định dạng XHTML+MathML bằng phần mềm InftyReader.Cấu trúc thư mục dữ liệu của hệ thống bao gồm 3 thư mục con nằm trong thư mục *C:\searchmath*, trong đó:

- Thư mục *originaldocuments* chứa các tài liệu nguyên gốc định dạng PDF.
- Thư mục *documents* chứa các tài liệu XHTML+MathML.
- Thư mục *indexes* chứa chỉ mục của các tài liệu XHTML+MathML trong thư mục *documents*.

3.3.2. Thử nghiệm ứng dụng

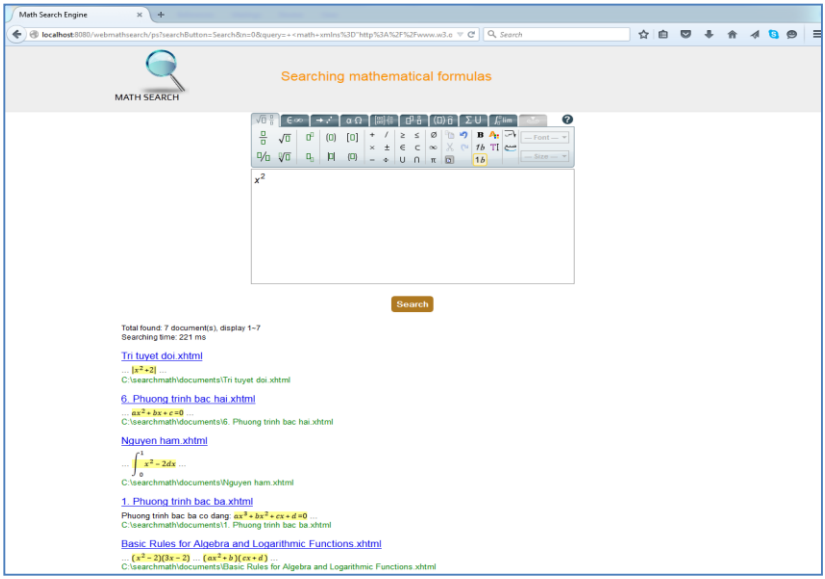
a. Hệ thống Tạo chỉ mục

Chương trình lập chỉ mục được xây dựng như một hệ thống độc lập với hệ thống tìm kiếm.Đầu ra của chương trình là tập hợp chỉ mục trong thư mục *indexes*.



Hình 3.7. Giao diện hệ thống lập chỉ mục

b. Hệ thống Tìm kiếm



Hình 3.9. Giao diện hiển thị kết quả tìm kiếm

Ứng dụng Tìm kiếm thuộc về người sử dụng. Hệ thống Tìm kiếm được xây dựng như một trang web sử dụng máy chủ localhost. Các kết quả tìm thấy sẽ được hiển thị sắp xếp giảm dần theo độ trùng khớp của tài liệu đó so với câu truy vấn.

3.4. KẾT QUẢ VÀ ĐÁNH GIÁ

3.4.1. Kết quả thu được

a. Ứng dụng Tạo chỉ mục

Qua nhiều lần thử nghiệm với số lượng tài liệu khác nhau, kết quả thu được qua các lần thực hiện như sau:

Bảng 3.1. Thống kê kết quả lập chỉ mục

STT	Số lượng tài liệu	Số lượng công thức	Thời gian lập chỉ mục
1	10	138	5739 ms
2	20	306	8907 ms
3	50	694	17564 ms
Trung bình		~32 ms / 1 công thức	~456.84 ms / 1 tài liệu

Từ thống kê trên cho thấy tốc độ lập chỉ mục khá nhanh, khoảng nửa giây cho mỗi tài liệu và khoảng 1/31 giây cho mỗi công thức.

b. Ứng dụng Tìm kiếm

Kết quả thu được qua nhiều thử nghiệm với ứng dụng tìm kiếm như sau:

Bảng 3.2.Thống kê kết quả tìm kiếm

STT	Truy vấn	Số TL chứa công thức đang truy vấn	Số TLtìm thấy	Số TL tìm thấy chứa công thức truy vấn	Thời gian tìm kiếm
1	x^2	8	7	7	25625ms
2	$ax^2 + bx + c$	3	2	2	18635ms
3	d/dx	3	3	3	24640ms
4	sin(x)	3	3	2	15102ms
5	x+1	3	4	3	12457ms
Thời gian tìm kiếm trung bình					19291ms

Độ chính xác của hệ thống tìm kiếm có thể được tính như sau:

X = $(Sốlượngtàiliệutìmthấychứacôngthứctruyvấn)/(Sốlượngtàiliệutrong CSDLchứacôngthứctruyvấn)$

= $(7/8 + 2/3 + 3/3 + 2/2 + 3/3)/5 \ x\ 100\% = 90\%$

Từ kết quả thống kê trên cho thấy:

- Tỉ lệ kết quả chính xác của hệ thống tìm kiếm: 90%

- Thời gian trung bình tìm kiếm: 19291ms

3.4.2. Đánh giá kết quả

a. Đánh giá chung

Hệ thống tìm kiếm công thức toán học trên văn bản đã đáp ứng được những yêu cầu cơ bản của người quản trị và người sử dụng.

b. Ưu điểm và hạn chế của chương trình

➢***Ưu điểm***:

- Đáp ứng cơ bản nhu cầu sử dụng và quản trị hệ thống.

- Phát triển riêng lẻ thành phần quản trị và thành phần tìm kiếm giúp cho công tác quản lý hệ thống và tìm kiếm dễ dàng hơn.

➢ ***Hạn chế***:

- Chức năng của bộ lập chỉ mục còn hạn chế.
- Kiểu định dạng của tài liệu đầu vào chưa đa dạng.
- Chưa hiển thị được tài liệu gốc PDF trên kết quả tìm kiếm.

c. Hướng phát triển

- Đa dạng hóa chức năng của bộ lập chỉ mục, như cho phép xóa chỉ mục, cập nhật chỉ mục, vv.
- Bổ sung thêm nhiều định dạng tài liệu khác như Word, Excel, vv.
- Bổ sung bộ thu thập thông tin toán học (Math Crawler) để hệ thống có thể thu thập và tìm kiếm được các tài liệu từ Internet.

TIỂU KẾT CHƯƠNG 3

Trong chương này, chúng tôi đã ứng dụng được những chức năng cơ bản của hệ thống tìm kiếm công thức toán học trên văn bản, ứng dụng mã nguồn mở Lucene. Hệ thống có thể chuyển đổi tập hợp các tài liệu PDF thành các tài liệu XHTML, thực hiện tạo chỉ mục và tìm kiếm trên tập tài liệu XHTML này. Mặc dù hệ thống còn đơn giản tuy nhiên nó cũng đã giải quyết quyết được những nhu cầu cơ bản là lập chỉ mục cho các tài liệu toán học và tìm kiếm trên tập chỉ mục đó.

KẾT LUẬN

Qua thời gian nghiên cứu, thử nghiệm và ứng dụng, luận văn đã đạt được một số thành công nhất định trong lĩnh vực tìm kiếm công thức toán học trên văn bản.

Về mặt lý thuyết, chúng tôi đã nghiên cứu những kiến thức cơ bản liên quan đến lĩnh vực tìm kiếm công thức toán học, chẳng hạn như phương thức đặc tả công thức toán học trên văn bản và website, nghiên cứu các ứng dụng tìm kiếm công thức toán học sẵn có, tổng quan về phương pháp tạo chỉ mục và tìm kiếm công thức toán học. Từ đó, chúng tôi đã đề xuất mô hình ứng dụng tìm kiếm công thức toán học của mình, và đề xuất các giải pháp để hiện thực hóa mô hình tìm kiếm này.

Về mặt ứng dụng, chúng tôi đã xây dựng thành công công cụ lập chỉ mục và tìm kiếm công thức toán học trong văn bản trên một kho dữ liệu ở máy tính cá nhân. Tuy rằng ứng dụng chưa đủ chưa hoàn thiện để có thể đưa vào sử dụng trong thực tế, nhưng nó cũng đã làm tiền đề cho việc xây dựng các ứng dụng tương tự. Trong tương lai, ứng dụng có thể được tích hợp bộ thu thập dữ liệu để có thể lập chỉ mục và tìm kiếm được trên mạng Internet và tối ưu hóa để tăng độ chính nhằm có thể đưa dự án vào sử dụng trong thực tế, đáp ứng được nhu cầu tìm kiếm công thức toán học hiện nay.