Kỷ yếu Hội nghị Quốc gia lần thứ VIII về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR); Hà Nội, ngày 9-10/7/2015
DOI: 10.15625/vap.2015.000218

VNMATHSEARCH - HỆ THỐNG TÌM KIẾM CÁC TÀI LIỆU TOÁN HỌC BẰNG TIẾNG VIỆT

Cao Xuân Tuấn[1], Võ Trung Hùng[2], Nguyễn Mạnh Hùng[3], Nguyễn Thị Thu Hà[4]

[1]Bộ Giáo dục và Đào tạo
[2] Khoa CNTT, Trường Đại Bách khoa, Đại học Đà Nẵng
[3]Học viện Công nghệ Bưu chính Viễn thông
[4]Khoa CNTT, Trường Đại học Điện Lực

cxtuan@moet.edu.vn, vthung@dut.udn.vn, nmhung@yahoo.com, hantt@epu.edu.vn

TÓM TẮT - *Bài báo này giới thiệu kết quả nghiên cứu xây dựng một hệ thống phục vụ tìm kiếm các tài liệu toán học viết bằng tiếng Việt. Hệ thống bao gồm 2 phần mềm chính đó là tạo chỉ mục và tìm kiếm. Chúng tôi đã đề xuất 2 mô hình tổng quát cho 2 phần mềm này. Với phần tạo chỉ mục, đầu vào là các tập tin dưới định dạng PDF hoặc XHTML và đầu ra là tập tin chỉ mục. Với phần tìm kiếm, người sử dụng có thể gõ vào truy vấn bằng từ khóa hoặc công thức bất kỳ và hệ thống trả về các tài liệu có chứa từ khóa hoặc công thức đó. Để xây dựng hệ thống, chúng tôi đã đề xuất các giải pháp để chuyển đổi định dạng công thức toán học, chuẩn hóa công thức toán học trong MathML, phân tích cú pháp và tạo chỉ mục, tích hợp công cụ gõ công thức toán học vào khung tìm kiếm, xếp hạng kết quả tìm kiếm,… Chúng tôi đã xây dựng và thử nghiệm hệ thống này với hơn 5000 tài liệu toán học viết bằng tiếng Việt, kết quả tìm kiếm cơ bản đáp ứng nhu cầu người dùng cả về độ chính xác lẫn tốc độ tìm kiếm.*

Từ khóa - *tìm kiếm, chỉ mục, xếp hạng, toán học.*

I. GIỚI THIỆU

Cùng với sự phổ biến và phát triển nhanh chóng của CNTT và mạng Internet, thông tin được chia sẻ và nhu cầu tìm kiếm trên mạng Internet ngày càng phong phú đa dạng hơn. Cũng như các lĩnh vực khác, ngày càng có nhiều người chia sẻ các tài liệu toán học và tìm kiếm thông tin để giải quyết các vấn đề thông qua 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 trên mạng Internet. Các máy tìm kiếm phổ biến hiện nay như Google Search, Yahoo Search, Live Search của Microsofts chưa cho phép cung cấp và nhận diện được các công thức theo cách tự nhiên, 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 [5][6].

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), nhưng những kết quả tìm thấy chỉ giới hạn trên những tài liệu điện tử lưu trữ trên máy chủ SpringerLink,... [3] Đặc biệt, hiện nay chưa có hệ thống nào tìm kiếm chuyên dụng cho các tài liệu toán học dành cho tiếng Việt. Vì vậy, việc nghiên cứu phát triển một công cụ tìm kiếm dựa trên các công thức toán học là cần thiết và có ý nghĩa thực tiễn cao [1].

Trong bài báo này, chúng tôi giới thiệu kết quả nghiên cứu và triển khai thử nghiệm của chúng tôi trên hệ thống VNMathSearch. Hệ thống này nhằm hỗ trợ tìm kiếm các tài liệu toán học (có thể tìm kiếm trực tiếp qua các công thức hoặc các từ khóa tiếng Việt) nhằm thúc đẩy việc học tập, nghiên cứu và ứng dụng khoa học tự nhiên tại Việt Nam. Bài báo được tổ chức thành 4 phần chính. Phần 2 trình bày kết quả nghiên cứu tổng quan về các văn bản toán học, phương thức biểu diễn công thức toán học trong tài liệu và Website và một số kết quả nghiên cứu liên quan. Phần tiếp theo mô tả ứng dụng, xây dựng mô hình tổng quát và giới thiệu giải pháp lưu trữ công thức toán học trên văn bản, giải pháp tạo chỉ mục cho các tài liệu toán học và giải pháp tìm kiếm công thức toán học cũng như tích hợp công cụ hỗ trợ người dùng trong quá trình tìm kiếm. Phần cuối trình bày việc triển khai xây dựng công cụ tìm kiếm công thức toán học trên văn bản và thử nghiệm đánh giá những kết quả đã đạt được.

II. MỘT SỐ NGHIÊN CỨU LIÊN QUAN

1. Đặc tả công thức toán trên tài liệu

Công thức toán học trên tài liệu có thể được đặc tả bằng nhiều ngôn ngữ khác nhau được gọi là ngôn ngữ đánh dấu toán học. Các ngôn ngữ đánh dấu toán học phổ biến nhất hiện nay là TeX/LaTeX [8], MathML [13], OMDoc [10] và OpenMath [11]. 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.

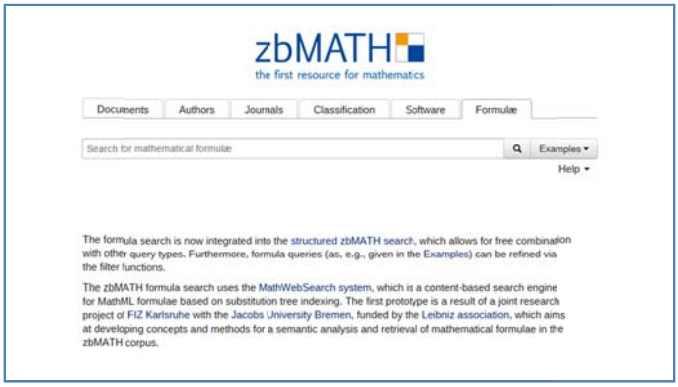
MathML (Mathematical Markup Language) là một ngôn ngữ mở rộng dựa trên XML để thể hiện ký hiệu và công thức toán học với mục đích rộng là phương thức trao đổi thông tin toán học trên máy tính (để hiển thị cũng như để tính toán) và mục đích hẹp là hiển thị tài liệu toán học trên World Wide Web. Tổ chức W3C (World Wide Web Consortium) có khuyến nghị nên sử dụng ngôn ngữ này trên mạng khi biểu diễn nội dung các công thức toán học. Đối

với hiển thị trên trang mạng, cấu trúc MathML không ngắn gọn như TeX, nhưng có thể dễ dàng phân tích bởi các trình duyệt, cho phép hiển thị ngay lập tức công thức toán học một cách đẹp mắt, đồng thời truyền tải ý nghĩa toán học cho các phần mềm tính toán. MathML được hỗ trợ bởi các phần mềm văn phòng như Microsoft Word, OpenOffice.org cùng với các phần mềm tính toán kỹ thuật như Maple, Mathematica và MathCad trên các hệ điều hành khác nhau như Linux, Windows,…

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 công thức toán học đó (Content MathML) [9].

2. Một số máy tìm kiếm dựa trên công thức toán học

MathWebSearch

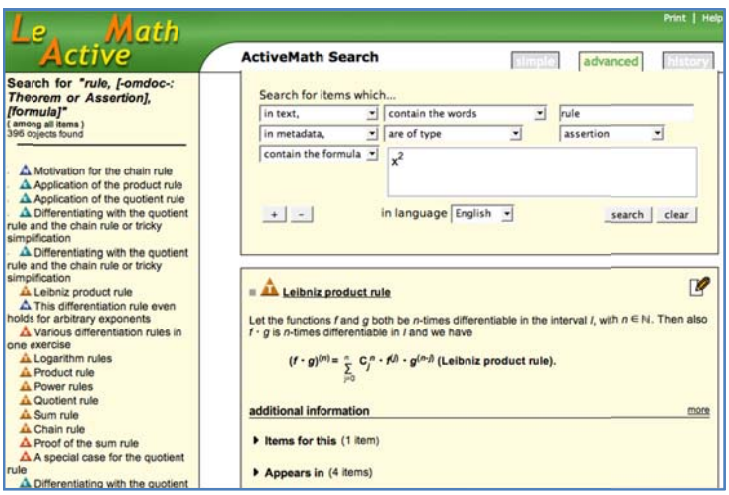


MathWebSearch là một bộ máy tìm kiếm công thức toán học dựa trên ngữ nghĩa của công thức, được phát triển tại Đại học Jacobs [2][7]. Hệ thống này tạo chỉ mục cho các công thức MathML và OpenMath, sử dụng kỹ thuật chỉ mục Substitution Tree Indexing. Công cụ tìm kiếm MathWebSearch được tối ưu cho các truy vấn nhanh và các ứng dụng tương tác. Bất kỳ dạng văn bản nào mà có chứa các công thức dưới dạng Content MathML hoặc dạng nào đó có thể dễ dàng chuyển đổi về Content MathML đều có thể được lập chỉ mục bởi MathWebSearch.

MathWebSearch có riêng bộ thu thập (Crawler) để tìm về các tài liệu có chứa Content MathML từ các kho chứa đặc biệt trên Internet, chuyển đổi các biểu thức toán học thành các chuỗi và lưu trữ nó trong cơ sở dữ liệu sử dụng MySQL. Chỉ mục sẽ được tạo trên dữ liệu này.

MathWebSearch cung cấp cả một bộ máy tìm kiếm hoàn chỉnh với giao diện trực quan và cả một API nhằm dễ dàng tích hợp vào các hệ thống sau này. Ngoài định dạng đầu vào kiểu XML và chuỗi, MathWebSearch còn cung cấp trình biên soạn công thức WIRIS nhằm hỗ trợ người dùng nhập công thức dễ dàng từ các mẫu sẵn có. Kết quả trả về từ MathWebSearch được xếp hạng dựa theo độ trùng khớp với nội dung tìm kiếm. Do đó nếu một tài liệu được xếp hạng càng cao, thì chứng tỏ số lần trùng khớp của nó với nội dung tìm kiếm càng nhiều. Hiện tại, MathWebSearch tạo chỉ mục cho hơn 1,600,000 tài liệu từ các kho chứa http://cnx.org và http://functions.wolfram.com và con số này càng ngày càng tăng. Trang chủ của MathWebSearch là: http://search.mathweb.org/.

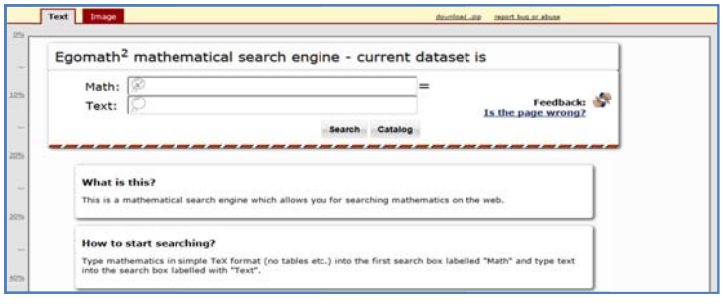
LeActiveMath



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. Người dùng có thể tìm kiếm đồng thời văn bản và công thức toán học trong ứng dụng này. Với mỗi tài liệu, LeActiveMath thực hiện lập chỉ mục cho trường tiêu đề, nội dung văn bản và công thức toán học. Tương tự như các công cụ tìm kiếm khác, các tài liệu tìm thấy cũng được sắp xếp giảm dần theo độ trùng khớp của tài liệu so với câu truy vấn. LeActiveMath được phát triển dựa trên Lucene, nó chỉ lập chỉ mục cho các tài liệu được sử dụng nội bộ trong môi trường học tập LeActiveMath. Trang chủ của LeActiveMath là http://www.leactivemath.org/.

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, kết quả tìm thấy được hiển thị cùng với đoạn trích dẫn chứa các nội dung trùng khớp với câu truy vấn, những phần trùng khớp này sẽ được làm nổi bật (highlight) nhằm giúp người dùng dễ dàng đối

chiếu và lựa chọn [4]. Từ giao diện tìm kiếm, người dùng có thể nhập câu truy vấn thông qua hai trường dữ liệu. Một trường để nhập cho các văn bản đơn giản và trường còn lại để nhập công thức toán học. EgoMath có thể xử lý được văn bản và các công thức toán học viết bằng LaTeX hoặc MathML. Trang chủ của EgoMath tại http://egomath.projekty.ms.mff.cuni.cz/.

III. GIẢI PHÁP ĐỀ XUẤT

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 cần đáp ứng một số yêu cầu như sau:

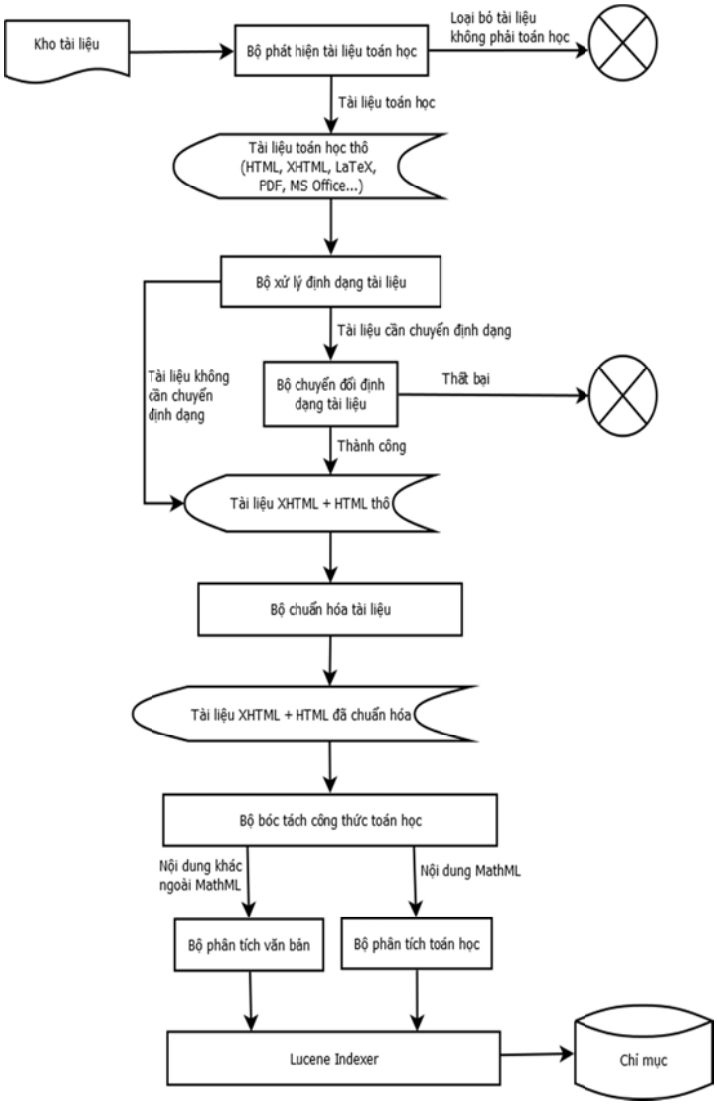
- Ứng dụng cho phép tìm kiếm được tài liệu ở các định dạng PDF và XHTML.
- Cho phép người dùng nhập công thức toán học một cách trực quan từ khung tìm kiếm.
- Cho phép tìm kiếm tài liệu toán học dựa trên nội dung tìm kiếm chứa đồng thời văn bản và công thức. Chẳng hạn người dùng có thể nhập "Pythagoras 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.

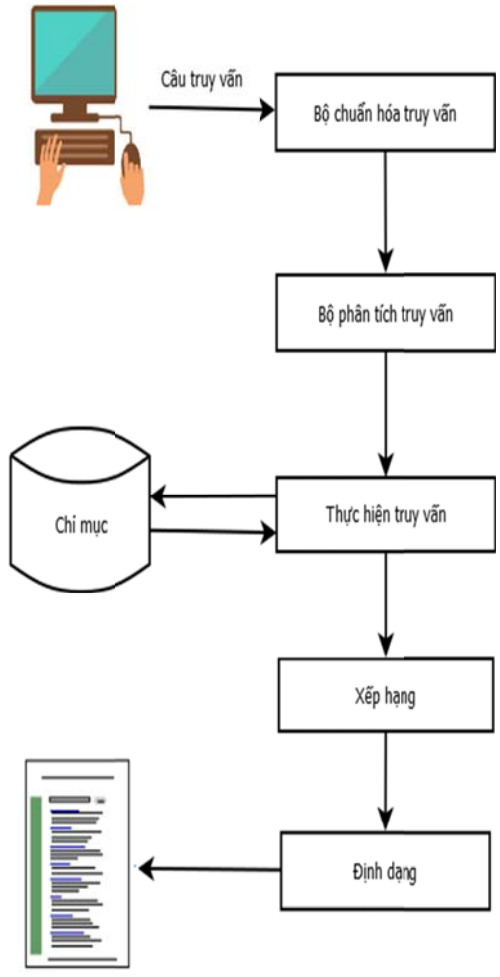
2. Mô hình tổng quát

Khi xây dựng hệ thống tìm kiếm, sau khi có kho dữ liệu chúng ta trải qua 2 quá trình chính đó là tạo chỉ mục cho các tài liệu và tìm kiếm khi có yêu cầu truy vấn của người dùng.

Mô hình tổng quá của quá trình tạo chỉ mục và tìm kiếm như sau:



Hình 1. Mô hình quá trình tạo chỉ mục



Hình 2. Mô hình quá trình tìm kiếm

3. Một số giải pháp xử lý

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 và XHTML. Để tạo chỉ mục trên tập tài liệu này, chúng tôi sẽ chuyển đổi chúng về một định dạng thống nhất là XHTML+MathML.

PDF là tài liệu đã được biên dịch từ mã nguồn TeX, LaTeX, Doc hoặc là kết quả chuyển đổi từ tập tin DVI hoặc PS, do đó không thể thực hiện tìm kiếm trực tiếp trên loại tập tin này. Để chuyển đổi PDF thành mã nguồn XHTML+MathML nhằm thuận tiện cho việc lập chỉ mục và tìm kiếm, chúng tôi đề xuất sử dụng InftyReader.

InftyReader là ứng dụng OCR (Optical Character Recognition - Nhận dạng ký tự quang học) có khả năng nhận dạng các tài liệu khoa học có chứa các công thức toán học. Kết quả nhận dạng có thể được xuất ra dưới nhiều định dạng khác nhau, chẳng hạn như: LaTeX, MathML, XHTML, HRTeX, IML và liệu Microsoft Word. InftyReader được phát triển tại thư viện Masakazu Suzuki, khoa Toán học sau đại học của trường đại học Kyushu.

Giải pháp chuẩn hóa công thức toán học

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. Quá trình này giúp cho việc tìm kiếm được chính xác hơn. Chuẩn hóa MathML là bước tối ưu hóa các công thức toán học bằng MathML nhằm loại bỏ các thẻ, các thuộc tính không cần thiết.

Để chuẩn hóa các công thức, chúng tôi thực hiện quá trình chuẩn hóa gồm các bước sau: Loại bỏ các thành phần và các thuộc tính không cần thiết; loại bỏ các thực thể ẩn (thực thể ẩn là những thực thể không hiển thị trên trình duyệt khi hiển thị công thức mà nó chỉ có tác dụng làm rõ ý nghĩa của công thức đó).

Những thuộc tính bị loại bỏ và các thực thể ẩn này chỉ có tác dụng trong việc giải thích phần ngữ nghĩa của công thức, mà không có tác dụng trong việc lập chỉ mục và tìm kiếm. Do đó những thành phần này được loại bỏ để tối ưu hóa hiệu suất của bộ máy tìm kiếm.

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.

Chuỗi nén này được tạo ra theo quy luật sau: một cặp thẻ XML (bao gồm thẻ mở và thẻ đóng) sẽ được thay thế bằng tên của thẻ và tiếp sau đó là chuỗi các tham số của thẻ đó sẽ được đặt trong cặp dấu ngoặc. Ví dụ công thức $a + b^2$ được viết trong MathML như sau:

```
<math  xmlns="http://www.w3.org/1998/Math/MathML">
                <mrow>
                        <mi>a</mi>
                        <mo>+</mo>
                        <msup>
                        <mi>b</mi>
                        <mn>2</mn>
                </msup>
                </mrow>
</math>
```

sẽ được chuyển đổi sang chuỗi nén tuyến tính như sau:

```
math(mrow(mi(a)mo(+)msup(mi(b)mn(2))))
```

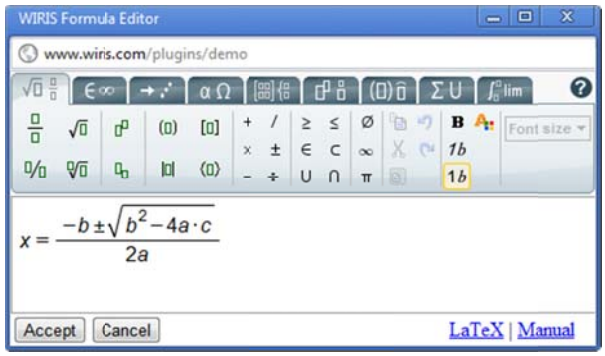
Giải pháp tích hợp công cụ gõ công thức toán học vào khung tìm kiếm

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). Trình biên soạn công thức WIRIS hoạt động tương tự bộ công cụ Equation trong Word. Người dùng chọn format của công thức cần nhập sau đó chỉnh sửa các giá trị ở trong format đó để tạo thành một công thức hoàn chỉnh.

Trình biên soạn WIRIS chạy được trên bất cứ trình duyệt (Firefox, Internet Explorer, Chrome, Safari, vv.) và bất cứ hệ điều hành nào (Windows, Linux, Mac, vv.). Nó có thể được tích hợp vào các ứng dụng Web và ứng dụng Desktop như một plugin.

Kết quả trả về của công thức được lưu trữ dưới dạng Presentation MathML, công thức này cũng có thể được chuyển đổi sang Content MathML hoặc LaTeX tùy vào nhu cầu tìm kiếm. Tuy nhiên trong phạm vi luận văn này, chúng tôi chuyển đổi công thức nhập vào thành Presentation MathML để thuận tiện cho quá trình lập chỉ mục.

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



Hình 3. Giao diện công cụ gõ công thức toán học WIRIS

Giải pháp xếp hạng kết quả tìm kiếm

Chúng tôi sử dụng thuật toán xếp hạng TF-IDF (Term Frequency - Inverse Document Frequency - Tần số mục từ - Tần số tài liệu nghịch đảo). Ý tưởng của thuật toán này là mục từ truy vấn nào xuất hiện càng nhiều trong tài liệu, tài liệu sẽ có điểm càng cao.

Thuật toán này được biểu diễn dưới công thức sau: $\mathrm{TF-IDF(t,d,D)} = \mathrm{TF(t,d)} * \mathrm{IDF(t,D)}$

Trong đó, ***t*** là query term, ***d*** là document cần được chấm điểm và ***D*** là tập hợp tất cả các tài liệu.

TF là tần suất xuất hiện của mục từ ***t*** trong tài liệu ***d*** và được tính $TF(t,d) = frequency(t,d)$

IDF là chỉ số biểu hiện cho tần suất xuất hiện của mục từ ***t*** trong toàn bộ các tài liệu. ***t*** xuất hiện càng nhiều, chỉ số càng thấp (vì xuất hiện quá nhiều đồng nghĩa với độ quan trọng rất thấp), $IDF(t,D) = \log \frac{N}{||\{d \in D: t \in d\}||}$

IV. THỰC NGHIỆM

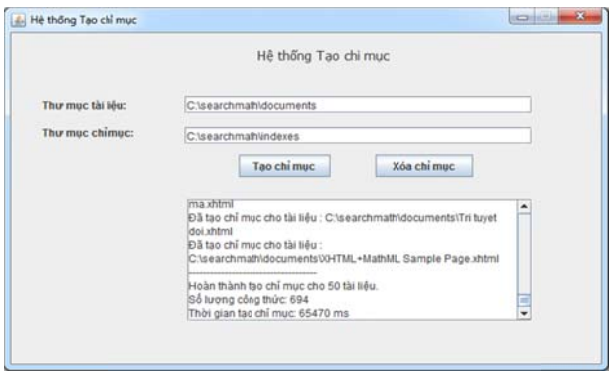
Thông thường, một hệ thống tìm kiếm gồm có 3 thành phần cơ bản gồm bộ thu thập thông tin, thành phần tạo chỉ mục và thành phần tìm kiếm. Kho dữ liệu chúng tôi xây dựng tổng hợp từ các bài báo, báo cáo, đề tài nghiên cứu khoa học, các sách điện tử về toán học tại Đại học Đà Nẵng, Giáo trình ebook và một số các tài liệu khác được thu thập trên mạng. Bảng sau mô tả về kho dữ liệu được sử dụng trong nghiên cứu này như sau:

Bảng 1. Mô tả dữ liệu thực nghiệm

Nguồn dữ liệu	Thư viện Đại học Đà nẵng
Số lượng	50 file tài liệu: giáo trình, báo cáo, bài báo khoa học,…
Định dạng	.doc, .docx, .pdf, .html, .latex
Số lượng công thức sau khi đánh chỉ mục	694

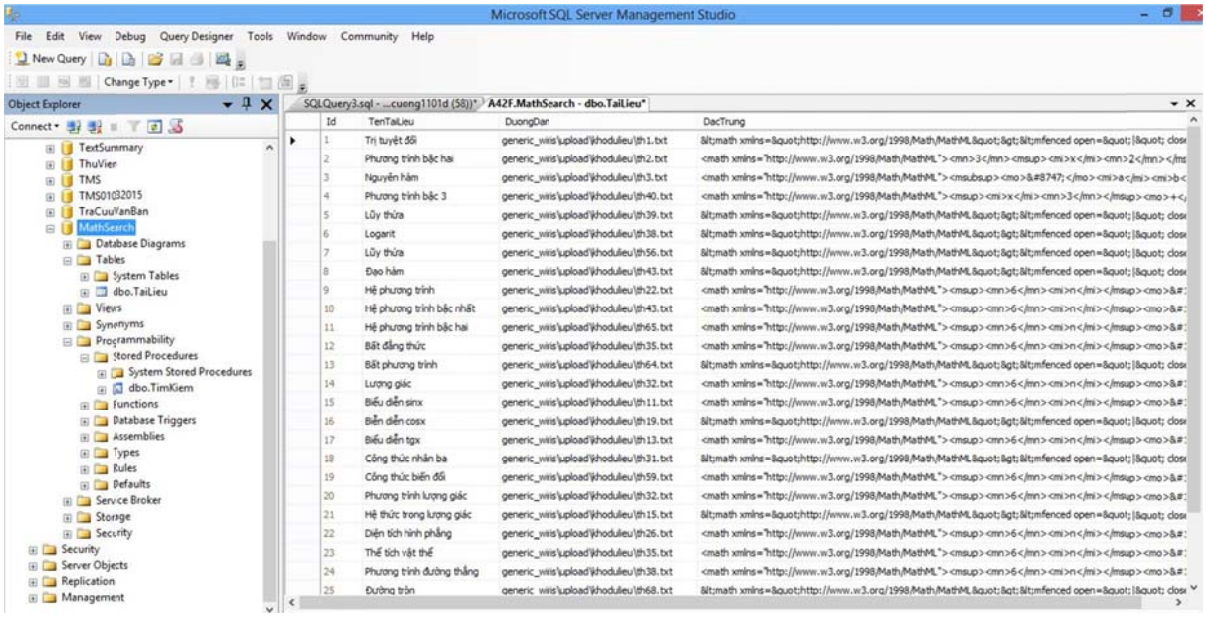
Chúng tôi phát triển hệ thống tạo chỉ mục như một chức năng dành cho người quản trị trị hệ thống. Chức năng này cho phép người quản trị chỉ định thông tin dữ liệu dùng để tạo chỉ mục, thực hiện tạo chỉ mục và xóa chỉ mục. Chương trình lập chỉ mục này được xây dựng độc lập với chương trình tìm kiếm. Người quản trị có thể chỉ định thư mục chứa tài liệu cần lập chỉ mục và thư mục chứa nội dung chỉ mục tùy ý.

Đầu vào của chương trình là thư mục chứa tất cả các tập tin cần lập chỉ mục và đầu ra là tập hợp chỉ mục trong thư mục Indexes, ngoài ra hệ thống cũng liệt kê số lượng và danh sách chi tiết các tập tin được tạo chỉ mục, số công thức được lập chỉ mục và tổng thời gian thực hiện việc lập chỉ mục.



Hình 4. Giao diện hệ thống lập chỉ mục

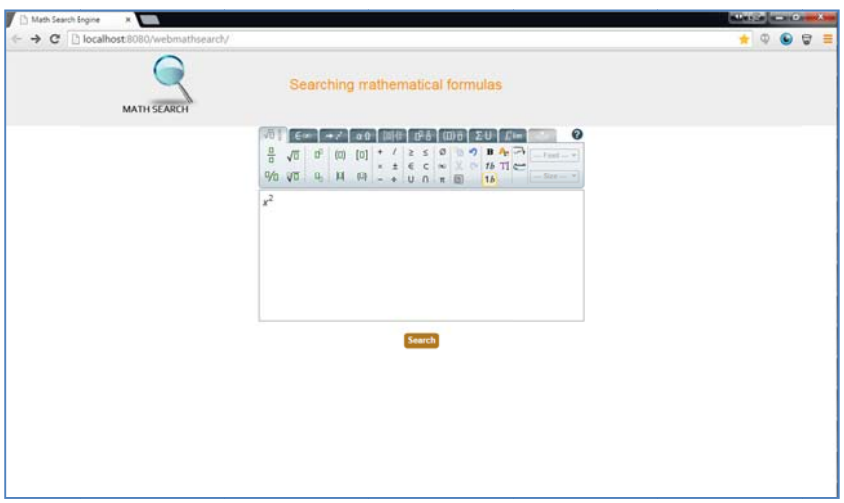
Các công thức sau khi chuyển đổi định định dạng sang MathML, được lưu trữ trong cơ sở dữ liệu SQL Server, phục vụ cho việc tìm kiếm.



Hình 5. Cơ sở dữ liệu hệ thống.

Chương trình tìm kiếm phục vụ người sử dụng. Đây là gói ứng dụng Web cho phép người dùng thực hiện tìm kiếm từ xa và lấy kết quả trả về. Chức năng chính của thành phần này là thực hiện tìm kiếm theo yêu cầu của người dùng, trả về kết quả dạng liên kết để người dùng tham chiếu.

Hệ thống tìm kiếm được xây dựng như một trang Web và cài đặt trên máy chủ tìm kiếm. Giao diện tìm kiếm bao gồm một khung hỗ trợ nhập công thức toán học và một nút Search:

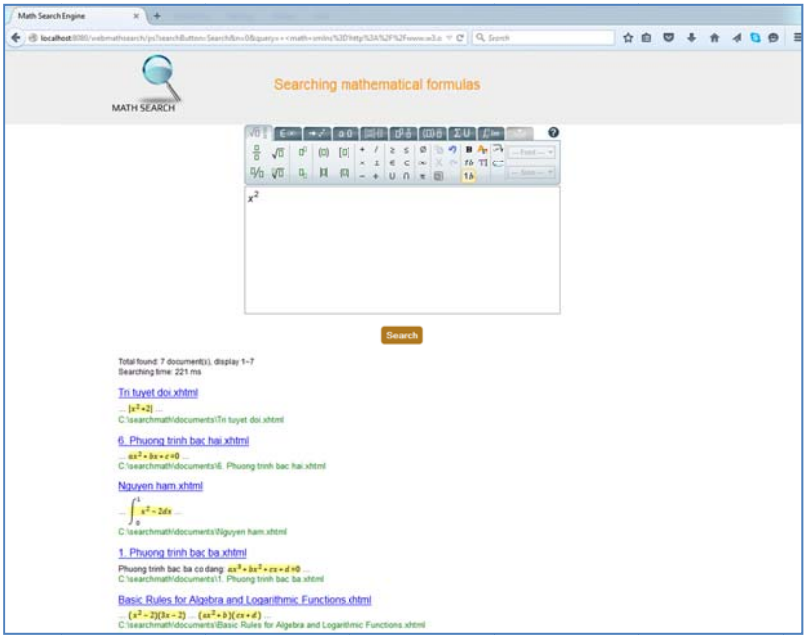


Hình 6. Giao diện ứng dụng tìm kiếm

Sau khi người dùng nhập công thức toán học và nhấn nút Search, hệ thống sẽ thực hiện tìm kiếm các tài liệu liên quan đến câu truy vấn của người dùng tại thư mục chỉ mục và trả về danh sách các tài liệu liên quan cho người dùng. 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. Mỗi tài liệu được sẽ được hiển thị lên giao diện web với các thông tin như sau:

- Tên tài liệu tìm thấy.
- Trích dẫn một phần tài liệu có chứa công thức được tìm thấy. Phần công thức trùng khớp với công thức trong câu truy vấn sẽ được làm nổi bật (highlight) để người dùng dễ dàng đối chiếu và lựa chọn.
- Đường dẫn tới tài liệu được tìm thấy.

Ngoài những thông tin trên, người dùng còn có thể xem số lượng tài liệu được tìm thấy ứng với câu truy vấn này, cũng như thời gian thực hiện truy vấn (tính bằng đơn vị millisecond).



Hình 7. Giao diện hiển thị kết quả tìm kiếm

Hiện nay, các hệ thống tra cứu tài liệu toán học bằng tiếng Việt chưa có. Do vậy, rất khó khăn để so sánh kết quả nghiên cứu của chúng tôi với các phương pháp khác. Trong bài báo này, chúng tôi đã thực hiện đánh giá kết quả của hệ thống xây dựng bằng phương pháp dùng độ đo chính xác (Precision) được mô tả theo công thức sau:

$$\text{Precision} = \frac{A \cap B}{B}$$

Trong đó: A là tập tài liệu liên quan tới nội dung tra cứu và B là tập tài liệu tìm được.

Chúng tôi thử nghiệm với tập ngữ liệu gồm 80 tài liệu toán học tiếng Việt, thực nghiệm được tiến hành và đánh giá theo 02 phương thức truy vấn: truy vấn theo công thức và truy vấn theo nội dung. Truy vấn theo công thức được gõ trực tiếp từ công cụ WIRIS trên hệ thống và truy vấn theo nội dung dựa trên câu truy vấn nhập vào.

Kết quả thực nghiệm được thể hiện ở bảng 2 dưới đây.

Bảng 2. Kết quả truy vấn

Truy vấn	Precision
Truy vấn theo công thức	0.87
Truy vấn theo nội dung	0.76

V. KẾT LUẬN

Các công cụ tìm kiếm tiện ích trên mạng cho phép người sử dụng dễ dàng tìm kiếm những tài liệu liên quan tới mục đích của họ, tuy nhiên khi số lượng thông tin quá nhiều, các kết quả trả về tới hàng trăm triệu văn bản tương ứng với mỗi câu truy vấn sẽ khó khăn khi tra cứu những tài liệu ở lĩnh vực hẹp.

Giải pháp tìm kiếm tài liệu toán học bằng tiếng Việt hỗ trợ cho các nhà khoa học, kỹ thuật của Việt Nam tìm kiếm những tài liệu văn bản liên quan tới các công thức bằng cách nhập dữ liệu trực quan và hiển thị những tài liệu liên quan có chứa những công thức cần tìm kiếm.

Với giải pháp đề xuất, chúng tôi đã tiến hành xây dựng hệ thống và đánh giá kết quả xây dựng bằng phương pháp sử dụng độ đo chính xác cho kết quả phù hợp với yêu cầu của người dùng. Hệ thống có một số ưu điểm nổi bật đối với các máy tìm kiếm hiện nay là đã hỗ trợ bộ gõ công thức toán học vào khung tìm kiếm, làm nổi bật (highlight) được kết quả tìm kiếm và mô-đun hóa các thành phần quản trị và thành phần tìm kiếm để dễ dàng cho việc phát triển sau này. Tốc độ lập chỉ mục và tìm kiếm khá nhanh.

Trong thời gian tới, chúng tôi tiếp tục bổ sung kho dữ liệu bằng phương pháp thu thập tự động trên Internet, tiếp tục hoàn thiện một số chức năng của hệ thống như: đa dạng hóa chức năng của bộ lập chỉ mục như cho phép xóa chỉ mục, cập nhật chỉ mục; bổ sung thêm nhiều định dạng tài liệu đầu vào khác như Word, Excel, PowerPoint,… tối ưu hóa tốc độ lập chỉ mục và tìm kiếm.

VI. TÀI LIỆU THAM KHẢO

[1] Vo Trung Hung, Cao Xuan Tuan, "VM-SEMWEB: A Semantic Web for Vietnamese Mathematical Documents", International Journal of Engineering Research & Technology, Volume. 4 - Issue. 05 , 2015.

[2] M. Kohlhase, C. Prodescu, "MathWebSearch:Low-Latency Uni_cation-based Search", Center for Advanced Systems Engineering, Jacobs University Bremen, Germany, NTCIR-10, 2013.

[3] M Růžička, "Maths Information Retrieval for Digital Libraries", Technical Report, Brno University, 2013.

[4] M. Adeel, H.S. Cheung, S.H. Khiyal, "Math go! Prototype of a content based mathematical formula search engine", Journal of Applied Theoretical and Information Technology, JATIT, 2008.

[5] J. Mišutka, L. Galamboš, "Extending Full Text Search Engine for Mathematical Content", Charles University in Prague, Ke Karlovu 3, 121 16 Prague, Czech Republic, 2008.

[6] P. Sojka, M. Líška, "Indexing and Searching Mathematics in Digital Libraries", Masaryk University, Faculty of Informatics, Botanická 68a, 602 00 Brno, Czech Republic, 2011.

[7] S. Anca, M. Kohlhase, "MaTeSearch, A combined math and text search engine", Jacobs University, 2007.

[8] T. Oetiker, H. Partl, I. Hyna, E. Schlegl, "The Not So Short Introduction to LATEX", Version 5.04, 2014.

[9] P.D.F. Ion, "MathML: A Key to Math on the Web", Mathematical Reviews, P. O. Box 8604, Ann Arbor, MI 48107, USA, 1999.

[10] M. Kohlhase, "An Open Markup Format for Mathematical Documents", Technical Report, Computer Science, International University Bremen, 2009.

[11] O. Caprotti, A.M. Cohen, H. Cuypers, H. Sterk, "OpenMath Technology for Interactive Mathematical Documents", Technical Report, Department of Mathematics and Computing Science, Eindhoven University of Technology, P.O. Box 513, NL-5600 MB Eindhoven, The Netherlands, 2002.

[12] Vo Trung Hung, Cao Xuan Tuan, "MathML for the Management of Mathematical Formula in Text Editor", International Journal of Engineering Research & Technology, Volume. 4 - Issue. 05 , 2015.

VNMATHSEARCH – A SEARCH ENGINE FOR MATHEMATICAL DOCUMENTS IN VIETNAMESE

Cao Xuan Tuan, Vo Trung Hung, Nguyen Manh Hung, Nguyen Thi Thu Ha

ABSTRACT *- This paper presents the research results to build a search engine for mathematical documents written in Vietnamese. The system consists of two main softwares that are creating the index and search. We have proposed two general models for 2 these softwares. With the index, the input is files as PDF or XHTML and the outputis an index file. With search modul, the user can type into the query by keywords or any formula and the system returns the documents that contain keywords or formulas. To build the system, we have proposed solutions to convert mathematical formulas, standardized mathematical formula in MathML, parse and index creation, integrated tool to type formulas in the search box, the search results ratings, ... We have built and tested the system with more than 5,000 mathematical documents written in Vietnamese, search results satisfy consumer demand the accuracy and speed of search.*