

MỘT MÔI TRƯỜNG THỐNG NHẤT ĐỂ BIỂU DIỄN, LƯU TRỮ, SOẠN THẢO VÀ XỬ LÝ CÁC CÔNG THỨC TOÁN HỌC

A UNIFIED ENVIRONMENT FOR REPRESENTATION, STORAGE, EDITING AND HANDLING OF MATHEMATICAL FORMULA

Cao Xuân Tuấn², Nguyễn Mạnh Hùng¹, Võ Trung Hùng²

¹Học viện Công nghệ Bưu chính Viễn thông; mhnghuyen@ptit.edu.vn

²Đại học Đà Nẵng; cxtuan@moet.edu.vn; vthung@dut.udn.vn

Tóm tắt - Trong bài báo này, chúng tôi trình bày kết quả nghiên cứu về việc đề xuất và xây dựng một môi trường thống nhất để biểu diễn, lưu trữ, soạn thảo và xử lý các tài liệu có chứa công thức toán học. Chúng tôi đề xuất một không gian thống nhất để biểu diễn và lưu trữ các công thức toán học trên nền tảng của tiêu chuẩn MathML, đây là một tiêu chuẩn được W3C khuyến cáo sử dụng cho tài liệu có chứa các công thức toán. Đặc biệt, môi trường này cho phép người sử dụng dễ dàng thực hiện các chức năng tìm kiếm, sao chép và chuyển đổi các công thức toán học giữa các phần mềm soạn thảo khác nhau như WinWord, LaTeX,... Môi trường này có nhiều ưu điểm so với các phần mềm soạn thảo hiện nay và tạo tiền đề để xây dựng các hệ thống tìm kiếm tài liệu thông qua các câu truy vấn có chứa công thức toán học.

Từ khóa - công thức toán học; biểu diễn dữ liệu; tiêu chuẩn lưu trữ; soạn thảo văn bản.

1. Đặt vấn đề

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 hầu hết các 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. Vì vậy, việc khai thác hiệu quả các tài liệu nói chung và các tài liệu khoa học nói riêng trên Internet có ý nghĩa quan trọng trong phát triển khoa học và kinh tế vì nó góp phần đáng kể vào việc nâng cao chất lượng học tập và nghiên cứu. Theo số liệu thống kê, trong quá trình học tập và nghiên cứu con người đã chi phí một lượng lớn thời gian cho việc tìm kiếm, phân tích và tổng hợp các tài liệu hiện có. Các công cụ tìm kiếm phổ biến hiện nay như Google, Yahoo, Bing... chỉ hỗ trợ mạnh việc tìm kiếm các dữ liệu dưới dạng văn bản và hình ảnh, công việc tìm kiếm các dữ liệu đặc biệt như các công thức toán học thì các công cụ này hầu như chưa hỗ trợ.

Hiện nay, đã có nhiều công cụ cho phép soạn thảo và quản lý các công thức toán học, nhưng việc tìm kiếm còn gặp nhiều khó khăn. Để tìm kiếm một công thức toán học, chúng ta cần có một cơ chế thống nhất để mô tả, lưu trữ và tìm kiếm theo ngữ nghĩa tương ứng với công thức đó. Tương tự, đã xuất hiện một số công cụ hỗ trợ việc biểu diễn các công thức toán học trên môi trường Web, tuy nhiên các công cụ này chưa xác định được chuẩn mô hình và cách biểu diễn chung. Do sự đa dạng về cách biểu diễn công thức toán học trong các tài liệu khoa học, dẫn đến khó khăn trong việc diễn giải công thức cần tìm kiếm đối với người dùng và so sánh sự tương đồng giữa chúng.

Trong bài báo này, chúng tôi đề xuất mô hình phù hợp,

Abstract - In this paper, we present our research on proposing and building a unified environment for representation, storage, editing and handling of documents containing mathematical formulas. We have proposed a unified space to represent and store the mathematical formula based on the standard MathML. This is the standard recommended by W3C to use for documents that contain mathematical formulas. In particular, this environment enables users easily to perform functions such as searching, copying and converting the mathematical formula between different editing software programs such as WinWord, LaTeX, ... This environment has many advantages compared with current editing software, and creates prerequisites for building document search system via the query containing mathematical formulas.

Key words - mathematical formula; data representation; storage standards; text editing.

phát triển các bộ công cụ để soạn thảo công thức, soạn thảo chủ thích và tìm kiếm các công thức toán học cũng như các giải pháp lưu trữ, quản lý và khai thác dữ liệu có chứa công thức toán học trong các tài liệu khoa học, đặc biệt là trên môi trường Web. Kết quả chúng tôi đã phát triển thành công một hệ soạn thảo văn bản hỗ trợ đầy đủ các chức năng liên quan đến công thức toán học gồm: soạn thảo, sao chép, tìm kiếm và chuyển đổi qua lại với các phần mềm soạn thảo khác.

Nội dung bài báo được tổ chức thành 5 mục chính. Mục 1 giới thiệu lý do nghiên cứu và thông tin chung của bài báo; mục 2 trình bày một số kết quả nghiên cứu liên quan; mục 3 trình bày giải pháp đề xuất bao gồm mô hình tổng quát, lưu trữ, soạn thảo và các giải pháp để sao chép, tìm kiếm, chuyển đổi,...; mục 4 trình bày kết quả thử nghiệm; mục 5 trình bày kết quả đánh giá; và cuối cùng là kết luận để tổng kết nội dung đạt được, ý nghĩa của nghiên cứu và hướng phát triển.

2. Một số nghiên cứu liên quan

Các nghiên cứu liên quan đến các tài liệu chứa công thức toán học tập trung vào 3 hướng chính: 1) Các tiêu chuẩn để lưu trữ công thức toán học trong các văn bản điện tử; 2) Soạn thảo và hiển thị các công thức toán học trên các văn bản; 3) Tìm kiếm công thức toán học. Các nghiên cứu trên chủ yếu được thực hiện bởi các nhà khoa học ở nước ngoài, ở Việt Nam các nghiên cứu về vấn đề này chưa nhiều và kết quả còn khá khiêm tốn.

Việc nghiên cứu các tiêu chuẩn để có thể lưu trữ và trao đổi các văn bản có chứa công thức toán học đã được các nhà khoa học, các công ty nước ngoài quan tâm từ rất sớm như: tiêu chuẩn TeX/LaTeX (đề xuất bởi Donald Knuth từ năm 1969 và đến năm 1991 đã có nhiều phiên bản ra đời và hỗ trợ nhiều ngôn ngữ khác nhau) [1], [2]; tiêu chuẩn MathML

(phát triển và phổ biến rộng rãi bởi W3C - World Wide Web Consortium do Tim Berners-Lee làm Chủ tịch) [3].

Bên cạnh các tiêu chuẩn lưu trữ, việc soạn thảo và xử lý các công thức toán học được nhiều nhà khoa học quan tâm. Hàng loạt các giải pháp đã được đề xuất và cho ra đời các công cụ soạn thảo, xử lý công thức toán như: MathMagic (phát triển bởi InfoLogic, Inc) [3], GtKMathView (phát triển bởi Luca Padovani, Đại học Torino, Ý) [4],...

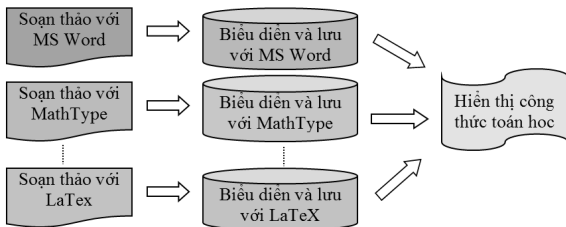
Ngoài ra, việc nghiên cứu các công cụ cho phép có thể lưu trữ, lập chỉ mục, tìm kiếm các công thức toán đã và đang được nhiều nhà khoa học, tổ chức đầu tư nghiên cứu như: WolframAlpha (tự động xử lý và tính toán giá trị các biểu thức toán học) [4], [5], Symbolab (tìm kiếm tài liệu có chứa công thức toán cho trước) [6],...

Cho đến nay, đã có một số nghiên cứu nhằm đưa tiếng Việt vào TeX và một sản phẩm tiêu biểu cho hướng này là VnTeX. VnTeX hỗ trợ tiếng Việt cho LaTeX và TeX thuần thông qua các gói macro và phong chữ. VnTeX thường được đóng gói kèm trong các bản phân phối TeX như teTeX, TeXLive... [8]. Ngoài ra, một cách soạn thảo công thức toán và tiếng Việt trong TeX có thể được vẽ bằng các lệnh của TeX. Hiện đã có một số nghiên cứu liên quan đến xử lý văn bản tiếng Việt nhưng không nhiều và chưa chú trọng đến xử lý các công thức toán học trên văn bản [9].

3. Giải pháp đề xuất

3.1. Mô hình tổng quát

Từ nhu cầu thực tiễn của việc soạn thảo, tìm kiếm công thức toán học trên văn bản và những khó khăn do sự khác nhau về chuẩn định dạng nên không thể sao chép trực tiếp công thức giữa hai trình soạn thảo, chúng tôi đề xuất xây dựng một môi trường soạn thảo có thể giải quyết được các vấn đề trên.



Hình 1. Các mô hình hiện nay

Hình 1 minh họa thực trạng này: công thức toán học được soạn thảo trong môi trường MS Word thì chỉ có thể lưu trữ trong văn bản MS Word mà không thể sao chép hay lưu trữ sang một công cụ soạn thảo hay hệ thống lưu trữ khác như MathType, LaTeX... và ngược lại.

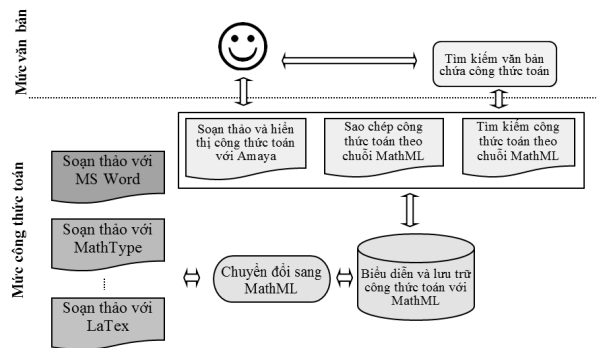
Từ thực tế này, chúng tôi đề xuất một giải pháp đồng bộ biểu diễn và lưu trữ công thức toán học như minh họa trong Hình 2. Theo đó, mô hình bao gồm hai mức xử lý: (i) mức xử lý công thức toán trong văn bản nằm ở mức thấp, và (ii) mức xử lý các văn bản chứa công thức toán, nằm ở mức cao hơn. Mức cao này sẽ được trình bày ở mục sau.

Ở mức xử lý công thức toán trong văn bản, các công thức toán học sẽ có một hệ thống soạn thảo đồng bộ chung, một hệ thống lưu trữ chung, và một hệ thống hiển thị chung. Khi đó, có hai vấn đề nảy sinh:

- Làm sao để chuyển đổi các công thức toán học từ môi trường soạn thảo truyền thống ban đầu sang môi trường soạn thảo đồng bộ chung này?

- Làm sao chuyển đổi các công thức toán học từ môi trường lưu trữ truyền thống ban đầu sang môi trường lưu trữ đồng bộ chung này?

Để giải quyết hai vấn đề này, chúng tôi đề xuất xây dựng một bộ chuyển đổi giữa cách biểu diễn các công thức toán học từ dạng ban đầu sang dạng đồng bộ chung, và ngược lại, chuyển đổi cách biểu diễn các công thức toán học từ dạng đồng bộ chung sang các dạng biểu diễn đang tồn tại trong thực tế hiện nay.



Hình 2. Mô hình đề xuất chi tiết

Như vậy, mô hình đề xuất của chúng tôi bao gồm ba thành phần được đề xuất mới:

- Một là bộ soạn thảo công thức toán học đồng bộ chung;
- Hai là hệ thống biểu diễn và lưu trữ công thức toán học đồng bộ chung;
- Ba là bộ chuyển đổi giữa các cách biểu diễn công thức toán học truyền thống đang tồn tại sang dạng biểu diễn đồng bộ chung, và ngược lại.

Trong mô hình này, có một số vấn đề cần phải xử lý liên quan đến công thức và chúng tôi đề xuất như sau.

3.2. Biểu diễn và lưu trữ công thức toán với MathML

MathML [1] 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 môi trường Web.

Đối với hiển thị trên môi trường Web, cấu trúc XML không ngắn gọn như TeX, nhưng có thể được dễ dàng sử dụng 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 MS Word, OpenOffice.org cùng với các phần mềm tính toán kỹ thuật như Maple, Mathematica, và MathCad trên các nền nhiều hệ điều hành khác nhau như Linux, Windows,...

3.3. Soạn thảo công thức toán với Amaya

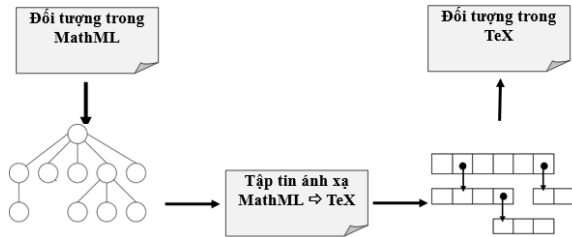
Để soạn thảo văn bản, chúng tôi đề xuất sử dụng phần mềm mã nguồn mở Amaya [10]. Amaya là phần mềm theo kiểu WYSIWYG (What You See Is What You Get), người

dùng có thể vừa soạn thảo và vừa có thể xem hiển thị kết quả trên trình duyệt. Các bộ công cụ của Amaya ở chế độ bảng trên chọn tương tự như MS Word, OpenOffice.Org Math,...

3.4. Chuyển đổi công thức toán giữa MathML và các dạng khác

Việc chuyển đổi định dạng giữa MathML và các định dạng khác cho công thức toán học không gặp nhiều khó khăn khi triển khai trong thực tế. Vấn đề là tìm ra sự tương đương giữa MathML và chuẩn biểu diễn tương ứng trong TeX, LATEX, MS Word,... để tạo ánh xạ khi chuyển đổi.

Ví dụ, khi chuyển đổi từ MathML sang TeX thì mô hình tổng quát như sau:



Hình 3. Mô hình chuyển đổi công thức

Một đối tượng trong MathML được biểu diễn theo một cấu trúc được qui định trong mô hình đối tượng tài liệu (DOM) qui định bởi W3C [3].

Để thực hiện việc chuyển đổi giữa MathML và TeX, ta cần có một tập tin lưu trữ thông tin ánh xạ mỗi phần tử của MathML sang một phần tử tương ứng trong TeX. Tập tin ánh xạ là một thành phần có vai trò rất quan trọng trong quá trình chuyển đổi. Nó mô tả sự tương ứng một – một giữa mỗi phần tử của MathML và TeX. Tập tin ánh xạ chứa một XML - form và bao gồm cả mẫu để biểu diễn cả MathML và TeX.

Cú pháp chung để mô tả sự tương ứng của các đối tượng trong MathML và TeX được viết như sau:

```
<pat:template>
  <pat:TeX op="\[TeX macro]"
    parameters=" TEX expression"/>
  <pat:mml op="mml-element"
    mode="math|TeX|spec">
    . . .
    [MathML expression]
    . . .
  </pat:mml>
</pat:template>
```

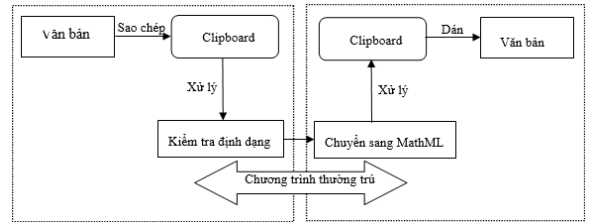
3.5. Sao chép công thức toán trong MathML

Amaya là một trình soạn thảo đồng thời là trình duyệt Web [10]. Vì vậy, mọi dữ liệu được tạo ra phải tuân thủ theo định dạng của một trang XHTML. Amaya có khả năng tự sinh thẻ khi sao chép dữ liệu dạng chuỗi từ ứng dụng khác vào trình duyệt, nhưng không thể tự sinh ra các thẻ của SVG khi sao chép dữ liệu hình ảnh từ ứng dụng khác sang.

Vì vậy, chúng tôi đề xuất xây dựng một chương trình thường trú và tích hợp vào Amaya để cho phép sao chép công thức được soạn thảo từ một ứng dụng khác sang trình duyệt Amaya.

Ý tưởng đề xuất là tiến hành sao chép công thức dưới dạng Text vào Clipboard, rồi sau đó xử lý trên Clipboard để chuyển sang ngôn ngữ đánh dấu MathML và cuối cùng chúng ta có thể dán vào Amaya.

Cách hoạt động của chương trình này như sau:



Hình 4. Mô hình sao chép công thức

Về sao chép công thức, chương trình thường trú Math Clipboard Converter sẽ nhận dạng bộ mã hóa đang dùng và tự động chuyển về chuẩn MathML. Ví dụ minh họa dưới đây cho thấy chương trình đang thông báo đã nhận dạng được chuỗi dữ liệu của công thức và thực hiện chuyển đổi dữ liệu sang MathML. Ở sơ đồ trên, khung bên trái là ngôn ngữ đánh dấu công thức toán học trong văn bản nguồn và khung bên phải là đoạn mã phát sinh tự động khi thực hiện lệnh dán từ Clipboard và hiển thị công thức tương ứng với mã lệnh.

3.6. Tìm kiếm công thức toán trong tài liệu

Để thực hiện chức năng tìm kiếm công thức toán học, chúng tôi đề xuất sử dụng phương pháp đối sánh mẫu (Pattern Matching). Bài toán yêu cầu ta tìm ra một hoặc nhiều vị trí xuất hiện của mẫu trên một văn bản. Trong đó mẫu và văn bản là các chuỗi có độ dài m và n ($m \leq n$). Việc đối sánh mẫu diễn ra với nhiều lần thử trên các đoạn khác nhau của văn bản. Trong đó cửa sổ là một chuỗi M ký tự liên tiếp trên văn bản. Mỗi lần thử chương trình sẽ kiểm tra sự giống nhau giữa mẫu với cửa sổ hiện thời. Tùy theo kết quả kiểm tra cửa sổ sẽ được dịch đi sang phải trên văn bản cho lần thử tiếp theo.

Nói chung, tìm kiếm các công thức toán học là một vấn đề rất khó khăn vì các ký hiệu toán học phụ thuộc vào ngữ cảnh. Ví dụ như biểu diễn hệ số nhị thức có thể có đến một loạt các ký hiệu mà phụ thuộc vào bối cảnh như $\binom{n}{k}$, C_n^k

đều có cùng một nghĩa là $\frac{n!}{k!(n-k)!}$. Trong việc tìm kiếm một công thức ở đây chỉ muốn lấy tất cả hình thức chứ không phân biệt các ký hiệu.

Để thực hiện đối sánh mẫu, chúng ta có thể dùng phương pháp đối sánh mẫu chính xác hoặc đối sánh mẫu xấp xỉ.

Bài toán đối sánh mẫu chính xác có thể mô tả như sau: Cho xâu mẫu P độ dài m ($P = P_1P_2 \dots P_m$ - P_i là ký tự) và văn bản T độ dài n ($T = T_1T_2 \dots T_n$ - T_i là ký tự). Tìm tất cả các vị trí xuất hiện của mẫu P trong xâu T.

Bài toán đối sánh mẫu xấp xỉ có thể mô tả như sau: Cho

xâu mẫu P độ dài m và văn bản T độ dài n. Xác định độ tương tự giữa hai xâu P và T. Như vậy, việc tìm sự xuất hiện của một mẫu trong văn bản, trong đó sự “khớp” giữa mẫu và sự xuất hiện của nó có thể chấp nhận “*k* lỗi” (k là một giới hạn cho trước). Có thể kể ra một vài kiểu “lỗi”, như những lỗi đánh máy hay lỗi chính tả trong hệ thống trích rút thông tin,... Vì trong các hệ thống tin học khó có thể tránh được các “lỗi” nên vấn đề tìm kiếm xấp xỉ càng trở nên quan trọng.

Bài toán trên chúng ta có thể đưa về tìm xâu con chung dài nhất (hay khúc con chung dài nhất): Một xâu w là xâu con hay khúc con (substring or factor) của xâu T nếu $T = uwv$ (xâu u, v có thể rỗng). Xâu w là khúc con chung của hai xâu P, T nếu w đồng thời là khúc con của P và T. Khúc con chung dài nhất của hai xâu P và T.

Đối với tìm kiếm công thức toán, chúng tôi đề xuất sử dụng thuật toán tìm kiếm xấp xỉ vì việc thay đổi thứ tự các thành phần trong công thức không có nhiều ý nghĩa như trong ngôn ngữ tự nhiên. Ví dụ : trong ngôn ngữ tự nhiên thì thứ tự của 2 từ “trường” và “học” sẽ tạo ra 2 cụm từ có ý nghĩa hoàn toàn khác nhau là “trường học” và “học trường”. Tuy nhiên, với công thức toán thì việc viết “ $x+y$ ” hoặc “ $y+x$ ” là có nghĩa như nhau. Mặt khác, tìm kiếm xấp xỉ cho phép tìm kiếm được các công thức có cùng ý nghĩa những cách viết khác nhau.

Thuật toán tìm kiếm xấp xỉ được mô tả như sau:

Input: Mẫu P có độ dài m; văn bản T có độ dài n.

Output:

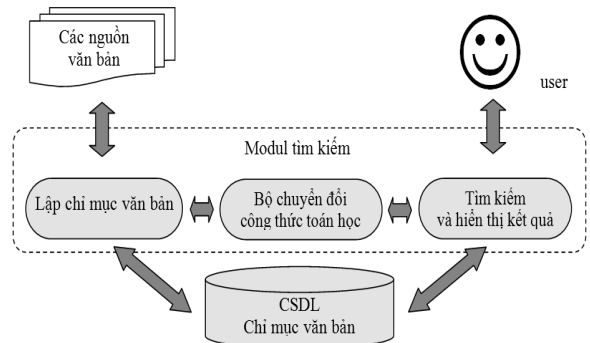
- Khúc con chung dài nhất giữa P và T
- Vị trí chuỗi mẫu Substr trong T

Thuật toán:

```
Char chuoicon (char *p, char *T, int m)
{
    int len, k, i, j;
    Char str[m], tam[m];
    len = 1;
    str = '';
    while (len <= strlen(s1))
    {
        k = strstr(p, T);
        j = 1;
        tam = '';
        for (i = k; i <= strlen(p); i++)
            if (p[j] == T[i])
            {
                Tam[j] = T[i];
                j++;
            }
        Else
            if (strlen(tam) > strlen(str))
                for (i = 1, i <= j, i++)
```

```
str[i] = tam[i];
len++;
}
}
```

3.7. Tìm kiếm tài liệu chứa công thức toán



Hình 5. Kiến trúc modul tìm kiếm văn bản chứa công thức toán

Nội dung phần này sẽ trình bày giải pháp tìm kiếm những tài liệu hoặc trang Web được tạo ra từ mô hình biểu diễn và lưu trữ đồng bộ công thức toán học trong văn bản như đã được đề xuất trong phần trước. Mô hình tìm kiếm này có hai 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 quát của quá trình tạo chỉ mục và tìm kiếm ở Hình 6.

Tuần tự các bước thực hiện trong mô hình này diễn ra theo hai giai đoạn như sau:

- **Giai đoạn 1:** Thu thập văn bản và lập chỉ mục văn bản. Quá trình này thu thập các văn bản chứa công thức toán học từ nhiều nguồn khác nhau từ Internet, từ nhiều dạng văn bản khác nhau như PDF, XHTML... Sau đó đánh chỉ mục các văn bản theo đặc trưng của văn bản và lưu vào CSDL chỉ mục văn bản. Trước khi trích các đặc trưng của văn bản, các nội dung chứa công thức toán học trong văn bản được tách ra và chuyển đổi về dạng thống nhất nhờ bộ chuyển đổi công thức toán học.

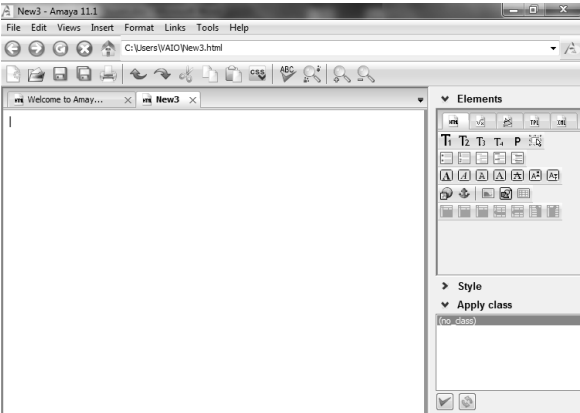
- **Giai đoạn 2:** Tìm kiếm và hiển thị kết quả. Giai đoạn này cho phép người dùng nhập từ khóa hoặc công thức toán học vào để tìm kiếm. Hệ thống sau đó sẽ tìm và tính điểm từng văn bản liên quan đến công thức muốn tìm trong CSDL chỉ mục các văn bản. Kết quả sẽ được hiển thị lên giao diện cho người dùng.

Như vậy, trong mô hình đề xuất cho chức năng tìm kiếm công thức toán học trong văn bản, có 3 mô-đun: Lập chỉ mục văn bản, chuyển đổi công thức toán học, tìm kiếm và hiển thị kết quả.

4. Thử nghiệm

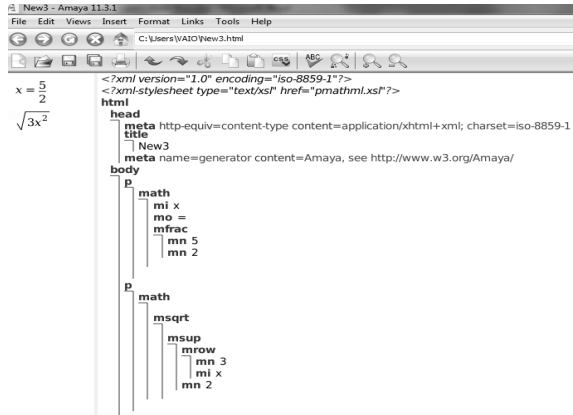
Trên cơ sở đề xuất, chúng tôi đã phát triển một môi trường soạn thảo các tài liệu có chức công thức toán học, bao gồm cả soạn thảo tài liệu Web. Môi trường được phát triển dựa trên: tiêu chuẩn lưu trữ MathML, phần mềm mã nguồn mở Amaya và các mô-đun do chúng tôi phát triển mới tích hợp vào để phục vụ soạn thảo, tìm kiếm, sao chép và chuyển đổi công thức toán với các môi trường soạn thảo khác.

- Giao diện chính của môi trường soạn thảo bao gồm:
- Thực đơn cho phép người sử dụng chọn thực hiện các thao tác xử lý.
 - Khung màn hình soạn thảo để chứa nội dung tài liệu.
 - Khung công cụ bên phải để hỗ trợ định dạng và gõ các công thức toán.

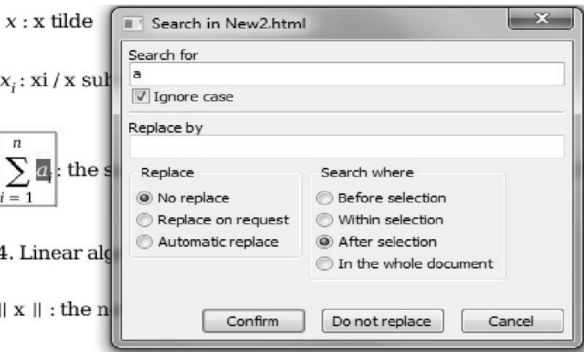


Hình 6. Giao diện chính màn hình soạn thảo

Trong môi trường soạn thảo này, người dùng có thể xem công thức toán dưới dạng trực quan hoặc mã MathML.



Hình 7. Biểu diễn công thức toán



Hình 8. Ví dụ về tìm kiếm công thức trong tài liệu

Người sử dụng cũng dễ dàng tìm kiếm các công thức toán:

Nhìn chung, môi trường do chúng tôi phát triển đã cơ bản đáp ứng các yêu cầu cho người sử dụng khi cần soạn thảo bất kỳ một tài liệu nào có chứa công thức toán học và hỗ trợ các thao tác xử lý khác.

5. Đánh giá

5.1. Dữ liệu đánh giá

Bộ dữ liệu đánh giá được tổng hợp và xây dựng từ nguồn gồm 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 1 mô tả về kho dữ liệu được sử dụng trong nghiên cứu này như sau:

Bảng 1. Mô tả dữ liệu thực nghiệm

Tiêu chí	Mô tả
Nguồn dữ liệu	Thư viện Đại học Đà Nẵng
Số lượng	580 file tài liệu: giáo trình, báo cáo, bài báo khoa học,...
Định dạng	.doc, .docx, .pdf, .html, .LaTeX

Dữ liệu bao gồm 580 văn bản từ nhiều lĩnh vực khác nhau (Bảng 2): toán, vật lí, công nghệ thông tin, điện tử và tự động hóa...

Bảng 2. Bảng phân loại dữ liệu

Kiểu tài liệu	Số lượng
Toán	210
Vật lí	17
Công nghệ thông tin	140
Điện, điện tử và tự động hóa	152
Các ngành khác	61
Tổng	580

Các tài liệu này khi thử nghiệm trên Semantic Web đã được chuyển đổi và lưu trữ trên Ontology với các lớp là các kiểu tài liệu và bổ sung mô tả tóm tắt về tài liệu, mô tả về các công thức toán có chứa trong tài liệu.

5.2. Phương pháp

Kịch bản đánh giá

Chúng tôi thực hiện đánh giá với hai kịch bản như sau:

- Kịch bản 1: Nhập truy vấn dưới dạng nội dung văn bản.

- Kịch bản 2: Nhập truy vấn dưới dạng công thức: công thức được gõ trực tiếp từ công cụ WIRIS trên hệ thống.

Tham số đầu ra

Chúng tôi quan sát các tham số đầu ra như sau:

- Số mẫu đúng trả về (TP - true positive): số lượng văn bản có chứa công thức truy vấn xuất hiện trong kết quả tìm kiếm.

- Số mẫu sai trả về (FP - false positive): số lượng văn bản KHÔNG chứa công thức truy vấn nhưng vẫn xuất hiện trong kết quả tìm kiếm.

- Số mẫu đúng không trả về (FN - false negative): số lượng văn bản có chứa công thức truy vấn, nhưng lại KHÔNG xuất hiện trong kết quả tìm kiếm.

- Số mẫu sai không trả về (TN - true negative): số lượng các văn bản KHÔNG chứa công thức truy vấn và KHÔNG

xuất hiện trong kết quả tìm kiếm.

- Khi đó, độ chính xác (Precision) được xác định bởi công thức:

$$precision = \frac{TP}{TP + FP}$$

- Độ triệu hồi (Recall) được xác định bởi công thức:

$$Recall = \frac{TP}{TP + FN}$$

Theo đó, nếu độ chính xác và độ triệu hồi càng cao thì mô hình đang được đánh giá sẽ được coi là càng tốt.

5.3. Kết quả

Kết quả đánh giá được thể hiện ở Bảng 3. Kết quả cho thấy kịch bản 1 nhập công thức truy vấn theo dạng nội dung văn bản cho kết quả thấp hơn kịch bản 2 nhập công thức truy vấn trực tiếp (84% so với 96%).

Bảng 3. Kết quả đánh giá

Kịch bản	Precision	Recall
Truy vấn theo nội dung	0.84	0.23
Truy vấn theo công thức	0.96	0.35

Mặc dù hai kịch bản đều cho kết quả độ chính xác cao, nhưng lại cho độ triệu hồi thấp. Có nghĩa là còn khá nhiều văn bản có chứa công thức được truy vấn, nhưng không xuất hiện trong kết quả tìm kiếm. Đây có thể coi là một điểm hạn chế còn tồn tại của mô hình đề xuất. Hạn chế này có thể xuất phát từ các yếu tố kỹ thuật sau:

- Bộ chuyển đổi công thức toán học: Có thể bộ chuyển đổi từ công thức truy vấn do người dùng nhập vào thành tập các mục từ khóa tìm kiếm còn một số hạn chế, khiến bộ từ khóa sinh ra không giúp tìm kiếm được triệt để các công thức toán học được lưu trong các văn bản.

- Thuật toán đối sánh xấp xỉ mẫu: Có thể thuật toán đối sánh xấp xỉ chưa phát hiện được đầy đủ các chuỗi con của tập từ khóa xuất hiện trong chuỗi văn bản tìm kiếm, khiến cho hiệu quả phát hiện được công thức toán học trong văn bản chưa cao.

Để khắc phục được những hạn chế này, chúng tôi sẽ tiếp tục xem xét cải tiến hiệu quả các bước chuyển đổi công thức toán học về dạng đồng bộ và bước đối sánh xấp xỉ trong tìm kiếm chuỗi biểu diễn công thức toán học trong văn bản. Đây là những hướng mở rộng đầy tiềm năng trong tương lai.

6. Kết luận

Bài báo này đã đề xuất được mô hình biểu diễn và lưu trữ đồng bộ các công thức toán học. Trong đó, MathML được đề xuất sử dụng như là tiêu chuẩn trung tâm phục vụ việc lưu trữ và xử lý các công thức toán học. Môi trường soạn thảo được đề xuất xây dựng dựa trên mã nguồn mở Amaya, có hỗ trợ soạn thảo công thức toán học theo chuẩn MathML. Mô hình này hỗ trợ việc chuyển đổi, sao chép công thức toán học giữa các dạng khác nhau vào một môi trường soạn thảo và lưu trữ thống nhất sử dụng tiêu chuẩn MathML.

Trên cơ sở đề xuất, chúng tôi đã tiến hành phát triển công cụ thử nghiệm dựa trên mã nguồn mở Amaya và về cơ bản công cụ đã đáp ứng các yêu cầu của người dùng khi soạn thảo và thực hiện các thao tác xử lý.

Trong thời gian đến, chúng tôi sẽ tiếp tục thực hiện các nghiên cứu để tạo các chú giải (mô tả ngữ nghĩa) cho các công thức toán, để người dùng có thể tìm kiếm bằng cách gõ trực tiếp công thức (hoặc một phần công thức) cần tìm hoặc gõ vào mô tả công thức và hệ thống tìm công thức thông qua tìm mô tả ngữ nghĩa của nó (dưới dạng văn bản).

TÀI LIỆU THAM KHẢO

- [1] R. Miner, *The importance of mathml to mathematics communication*, Notices of the AMS, vol. 52(5):532–538, 2005.
- [2] L. Lamport, *LaTeX: A document preparation system: User's guide and reference*, Addison-Wesley, ISBN 0-201-52983-1, 1994.
- [3] David Carlisle, Patrick Ion, Robert Miner, *Mathematical Markup Language (MathML) Version 3.0*, W3C Recommendation, 2014.
- [4] Z39.59-1998, *AAP Math DTD, Standard for Electronic Manuscript Preparation and Markup*, Washington, DC: Association of American Publishers, 1998.
- [5] R.Dillet, *Wolfram Alpha Makes CrunchBase Data Computable Just In Time For Disrupt SF*, TechCrunch Published, 2012.
- [6] G. Nass, *GES 2014: Symbolab takes the hassle out of the equation*, MindCet Published, 2015.
- [7] M. Kohlhase, B.A. Matican, *MathWebSearch 0.5 – Scaling*, in Intelligent Computer, Conferences on Intelligent Computer, Bremen, Germany, 2012.
- [8] H. T. Thanh, *Micro-typographic extensions to the TEX typesetting system*, TUGBoat, Volume 21, No. 4, 2000.
- [9] Le T.N., Vo T.H., Cao X.T., Hoang T.M.L., “Mathis - Hệ thống hỗ trợ tạo chú thích và tìm kiếm tài liệu khoa học”, *Tạp chí Khoa học và Công nghệ*, Đại học Đà Nẵng, pp. 15-20, 2010.
- [10] V. Quint, *An introduction to Amaya*, *Journal World Wide Web Journal - Special issue: scripting languages: automating the Web*, Volume 2 Issue 2, pp 39-46, O'Reilly & Associates, Inc. Sebastopol, CA, USA, 1997.

(BBT nhận bài: 23/01/2017, hoàn tất thủ tục phản biện: 08/02/2017)