Indexing Based on Topic Modeling and MATHML for Building Vietnamese Technical Document Retrieval Effectively

Tuan Cao Xuan[1], Linh Bui Khanh[2], Hung Vo Trung[3], Ha Nguyen Thi Thu[2(✉)], and Tinh Dao Thanh[4]

[1] Vietnam Ministry of Education and Training, Hanoi, Vietnam
cxtuan@moet.edu.vn
[2] Information Technology Faculty, Electric Power University, Hanoi, Vietnam
{linhbk, hantt}@epu.edu.vn
[3] Danang University, Da Nang, Vietnam
vthung@dut.udn.vn, nmhung@yahoo.com
[4] Information Technology Faculty, Le Quy Don Technical University, Hanoi, Vietnam
tinhdt@mta.edu.vn

Abstract. The grow of data on the Internet has brought to people many information and it also opened some important problem in Information retrieval…Along with it, some search engines have developed for user's purpose. User can retrieve information by content, keyword or anything what they need. However, data on the Internet is too huge, the results feedback is often millions or hundreds millions for each query. Therefore, with the narrow field, we will meet a difficult to find related information, especially technical information that contain formulas. In this paper, we present a method for building Vietnamese technical text based on topic modeling and MathML for indexing. System has built and tested with over 500 Vietnamese technical text shown that, this system satisfied users' requires in accuracy and speed.

Keywords: Mathml · Topic modeling · Vietnamese technical text · Search engine · Information retrieval

1 Introduction

Big data is a very widespread concept in life today when data sources on the Internet become popular. The huge amounts of data share online every day brings convenience to seek information consistent with user's purpose, but also difficulties to find financing specialty information, especially is technical documents contain multiple formulations, special notation: π, Ω, $\frac{-b\pm\sqrt{b^2-4ac}}{2a}$, $(x+a)^n = \sum_{k=0}^{n}\binom{n}{k}x^k a^{n-k}$ [3, 8]. These popular search engines Google Search, Yahoo Search, Live Search,… not allow type and identify the formula naturally, so search results usually are not matched with user's requires. Therefore, should have a search engine for searching mathematical formula that have shared on the Internet [7, 8].

P.C. Vinh and V. Alagar (Eds.): ICCASA 2015, LNICST 165, pp. 322–332, 2016.
DOI: 10.1007/978-3-319-29236-6_31

There are many search engines that can search formulas developed. Egomath allow searches mathematical formulas on Wikipedia.org [18], LatexSearch support the mathematical formula that using LaTeX markup language [15]. But these search engines didn't support formula typing frame so the user feel very difficult when they want to find documents that contain flexibility formula and they can't search for particular language or specific major.

There isn't any similar system for Vietnamese technical document because it is seem difficult to Vietnamese. In this paper, we present our research on building, developing and result of experimental with Vietnamese technical text retrieval by using topic modeling and MathML to indexing formula solution, It has a frame for typing formula based on WIRIS open source [19] and topic modeling is the way to optimize retrieval technical text, easy to search, matched with user's query better than other general search engines.

The rest of the paper structured as follows: Sect. 2 introduce some related works, our method in the Sect. 3, the results and experimental in Sect. 4 and Sect. 5 is conclusion.

2 Related Works

For advantage some function of popular search engines: Google, Yahoo search, Live search in searching mathematical documents, some search engines have been built. MathWebSearch is a mathematical engine based on expression semantic. Mathematic Expressions are stored by data tree. Each node on these tree called substitution that corresponds to a function. A mathematic expression can be represented from root along path on the tree. With this, there are many mathematic expression represented on a tree, and the searching become easier. MathWebSearch can process and index expressions encoded with content MathML or OpenMath [6, 9, 11].

LeActiveMath index OMDoc documents, in this, mathematical formulas processed by OpenMath. User can search text or formulas via this system. With each document, LeActiveMath index title, content and formulas. Like some other search engines, documents will be rank by similarity score of queries. LeActiveMath is developed based on Lucene, documents stored in database [20] (Fig. 1).

Egomath has been developed by Charles University in Prague. It can search mathematical formulas that wrote in LaTeX or MathML and simple document, the search results display along with quotations that contain matches the queries. From the search interface, users can enter queries through two data fields. A field to enter the plain text and the rest to enter mathematical formulas [18].

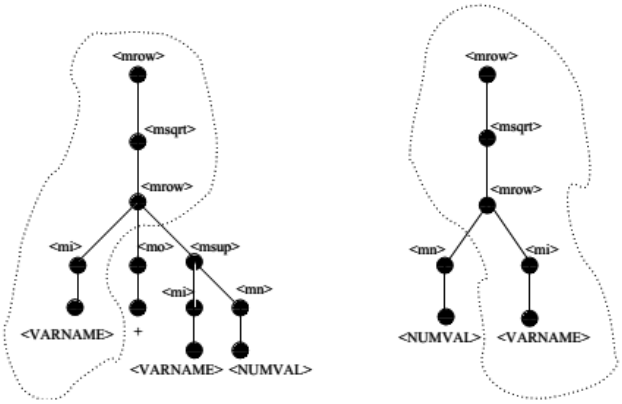

Fig. 1. Representation of $\sqrt{x+x^2}$ and $\sqrt{2x}$

3 Indexing Method Effectively for Building Vietnamese Technical Document Retrieval System

3.1 Topic Modeling

Vietnamese is a single syllable language, one of the Asian languages have single word and compound of word. These words in Vietnamese no distinction based on white spaces, so when mine Vietnamese text, the traditional methods commonly used word processing tool kit to solve the problem of Vietnamese as: text summary, text extraction, information retrieval, text classification… With this approach, always need so much times for processing, and effective is not high in Fig. 2.

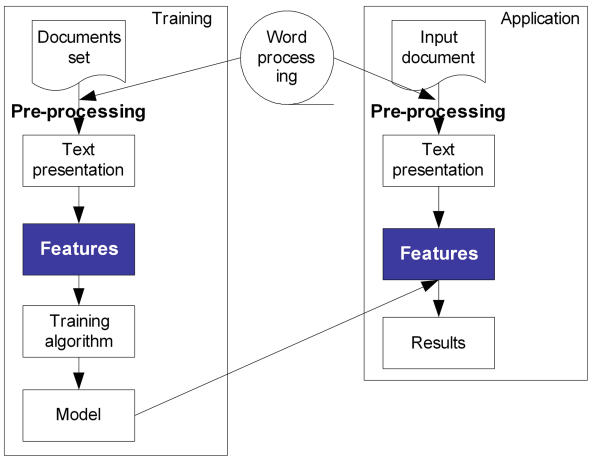

Fig. 2. Vietnamese text mining system.

For text classification problem, large of features are extracted from text and then use one of machine learning methods: SVM, Naïve Bayes, Decision tree, K-nearest neighbor… to classify or identify what category it belong is. With this approach, need high cost and time to process with a large amount of texts and features (Fig. 3).

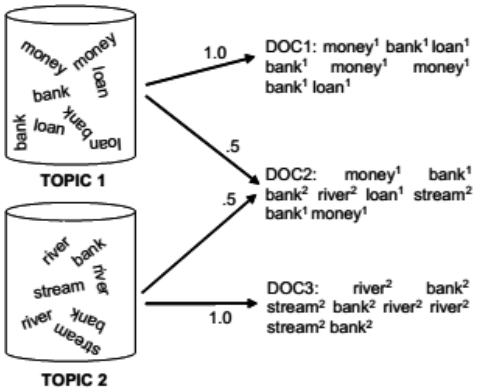

Fig. 3. Topic modeling in text classification.

Topic modeling developed by Blei [1], there are many approach for text mining have been proposed, and proven it's really effective in text mining field by reducing features in text. It helps text mining systems become faster and more accuracy [2, 5, 10].

In this paper, we use the topic modeling to identify Vietnamese technical documents automatic. With topic modeling, very easy to recognize Vietnamese texts in the large database with multiple languages. We can reduce time for processing, identify Vietnamese technical documents more correctly, and improve retrieval process by reduced large number of features.

3.2 Indexing and Store Database with MathML

MathML (Mathematical Markup Language) is an extension language based on XML to present symbols and mathematical formulas. MathML's purpose is to display mathematic communication methods on the computer and the World Wide Web. For display on websites, structure of MathML is not concise like TeX, but it is easy analyzed by the browser, immediate display of mathematical formulas, and transmit to calculating applications. MathML is supported by the office software like Microsoft Word, OpenOffice.org, and other calculation software like Maple, Mathematica and MathCad and on the difference operating systems like Linux, Windows [13, 16].

We can present $a + b^2$ formula can be written in MatML.

```
<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>
```

After that, it convert to

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

3.3 Methodology of Vietnamese Technical Document Retrieval

Based on topic modeling and MathML for indexing and storing Vietnamese technical document in the database, we present a solution for building Vietnamese technical document retrieval system through some steps:

- Step 1: Collect training set with n Vietnamese technical documents $D = \{d_1, d_2, \ldots, d_n\}$ automatic by classify documents based on topic modeling [5].
- Step 2: Store terms in Technical field that extract from topic modeling.
- Step 3: Identify formulas from training set by Infty Reader. And use MathML to store formulas in database by tree indexing.
- Step 4: Feature representation of each document with probabilistic.

Four steps can be present like Fig. 4 below.

4 Experimental

4.1 Functions of Vietnamese Technical Documents Retrieval System

We built a system for searching Vietnamese technical documents that contains mathematical formulas by entering formula visually on the input box. Here is some functions of system:

- Allow searching PDF, .Doc, .docx and XHTML format type.
- User can enter formula from input box.
- User can enter text from input box.
- Searching document based on content or formula. (User can enter: "Pythagoras", all documents contain formula: $a^2 + b^2 = c^2$ or content: Pythagoras will be appearance in the interface of system).
- Results are ranked by user's queries.

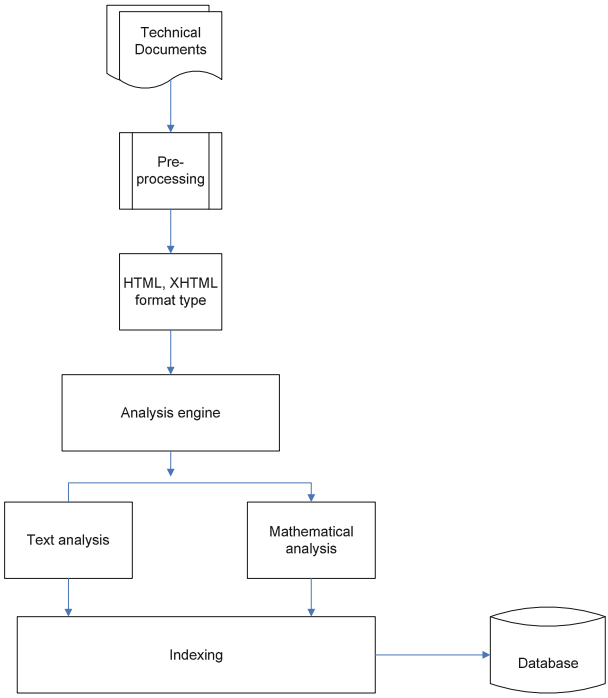

Fig. 4. Indexing and storing Vietnamese technical documents on database.

4.2 Integrate Formula Input Box

On the interface of system, user can type mathematical formulas directly on the search box by integrating mathematical formulas tool called WIRIS. WIRIS is an open source that wrote by JavaScript helps users enter and edit formulas, it is a visual formulas editor like equation tool in Microsoft word. Users select format of formula and then they edit its to complete form.

WIRIS can display all web browsers: Firefox, Internet Explorer, Chrome, Safari,… and anything of operators Windows, Linux, Mac,…It is integrated in the web application like a plugin. Responds of results are stored MathML and we use it for indexing (Fig. 5).

Here is the interface of WIRIS

4.3 Ranking

Ranking equivalent with two searching ways: based on text query and based on visual formula. For formulas searching, we used similarity score between input formula and documents in database that contain formula indexed by tree graph. For example: when users enter a^2, the first document contain independent a^2, and from the second

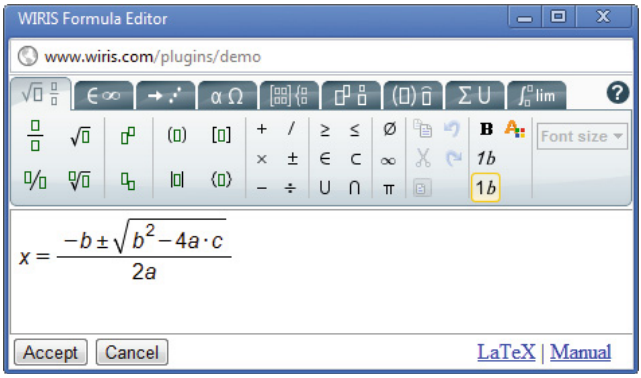

Fig. 5. Formulas input frame - WIRIS

document can contain formulas that deployment from a^2. With queries, the results rank by frequency of terms that occurred on queries and documents.

4.4 Results

Typically, a search engine includes three components: Crawl, index creation and search. Our corpus has been built from many source: Internet, library of Danang University, and corpus includes: articles, technical reports, scientific projects, e-books,… The table below is the description of the corpus (Table 1).

Index creator is a function of administrator when we developed this system. Administrator can create new indexes or delete. Figure below is the index creator what we use indexing with documents and then store it in the database of system (Fig. 6).

Table 1. Corpus

Sources	Library of Danang university, Online material, Offline material, …
Quantity	50 files collected manually from library of Danang university: articles, reports, scientific projects, e- books,… 530 files collected from Internet by Vietnamese text classification system based on topic modeling [5]
Format type	.doc, .docx, .pdf, .html, .latex
Number of formulas after indexing	694

Formulas are converted to MathML form after indexing and stored in the SQL Server Database (Fig. 7).

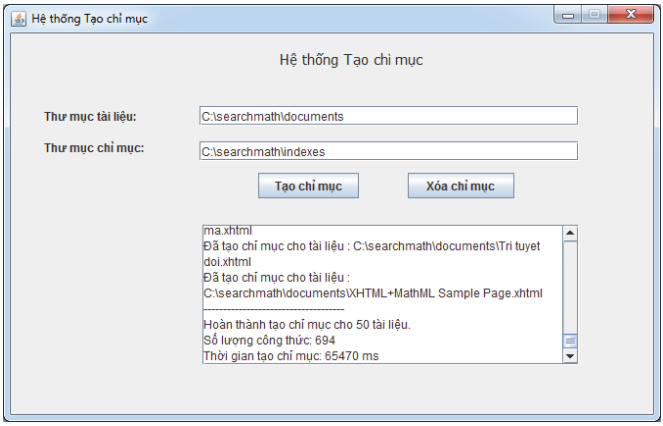

Fig. 6. Indexing creator

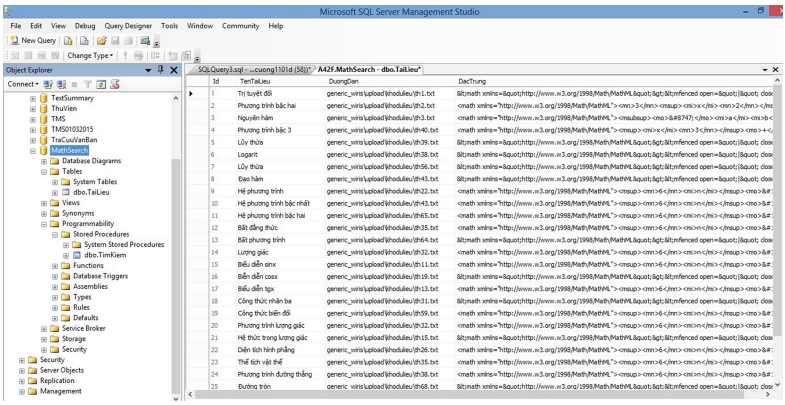

Fig. 7. Stored formulas in the database

The Vietnamese technical document retrieval engine has been built and developed by web technology. Searching interface of system includes input box that allow user enter text or formula. Visual formula frame has been integrated in it (Fig. 8).

After users enter mathematical formulas or any query and then click Search button, the system will find related documents. Results displayed and ranked on the interface:

- Name of document.
- Quote part document that contains formula or terms that user entered on the input box.
- Query execution time.

Vietnamese technical document retrieval system hasn't built before. Therefore, it is difficult to compare our method with others. In this paper, we have to evaluate the results that generated by system by Precision and Recall (Fig. 9).

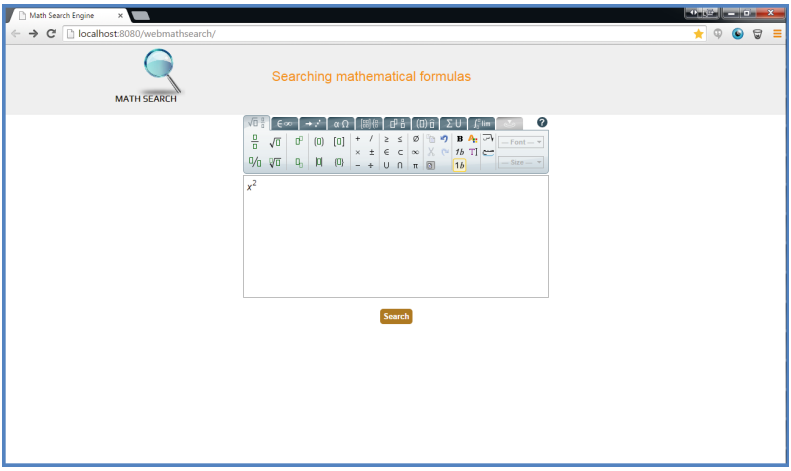

Fig. 8. Searching interface

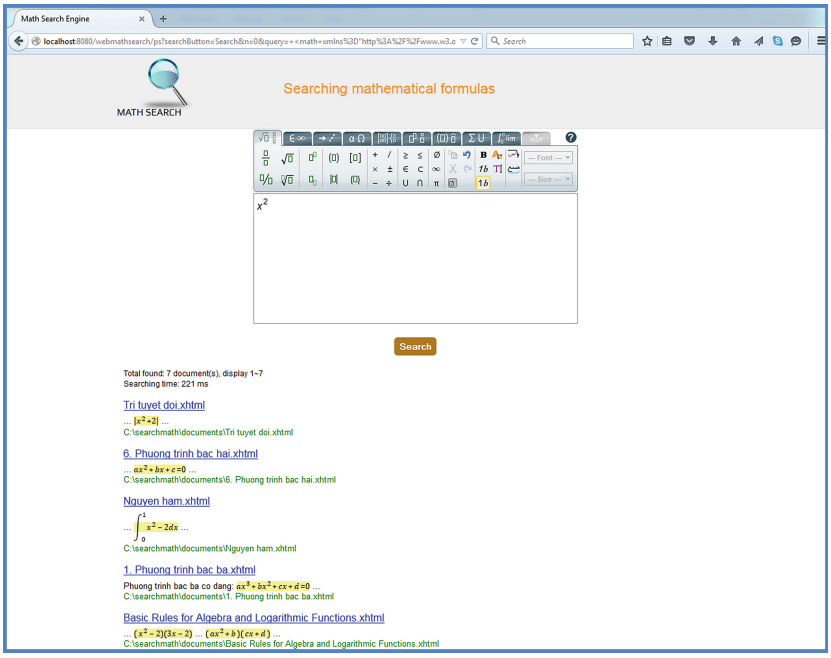

Fig. 9. Results interface

We tested based on 580 Vietnamese technical documents

- Mathematical major: 210 documents.
- Physic major: 17 documents.
- Information technology major: 140 documents.
- Electric engineering, electronic and automation major: 152 documents.
- Others: 61 documents.

On the other documents have $\sim$70 % documents isn't contain in its. Experimental were performed two searching methods: by query and by formula. Formulas entered from WIRIS that integrated on system.

Results of experimental are display on the Table 2.

Table 2. Results of experimental.

Retrieval results by content of text		Retrieval results by formulas	
Precision	Recall	Precision	Recall
0.84	0.233	0.96	0.35

5 Conclusion

The convenient search engine on the Internet allows users find their purpose's relatedly documents very easily. However, when the amount of information is too much, the results returned to hundreds of millions of documents with each query, it become difficult to find documents in a narrow field.

In this paper, we presented our research and solutions for Vietnamese technical documents retrieval. It can help for Scientists, technicians search technical documents contain formulas through enter formulas from visualization input box and system display related documents contain formulas that user entered.

We have carried out to build system and evaluated results of the system based on precision and recall measure. Its results shown that, our method and solution is really effectively and high accuracy with each queries. In the future, we will develop our system to online and receive any feedback from users to improve Vietnamese technical document retrieval system.

References

1. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
2. Vulic, I., De Smet, W., Moens, M.F.: Cross language information retrieval models based on latent topic models traned with document aligned comparable corpora. Inf. Retrieval **16**(3), 331–368. Springer (2013)

3. Mišutka, J., Galamboš, L.: Extending full text search engine for mathematical content. Charles University in Prague, Ke Karlovu 3, 121 16 Prague, Czech Republic (2008)
4. Lau, J.H., Newman, D., Karimi, S., Baldwin, T.: Best topic word selection for topic labelling. In: Coling 2010: Posters, pp. 605–613 (2010)
5. Thu, H.N.T., Thanh, T.D., Hai, T.N., Ngoc, V.H.: Building Vietnamese topic modeling based on core terms and applying in text classification. In: Proceedings of the Fifth IEEE International Conference on Communication Systems and Network Technologies, pp. 1284–1288 (2015). doi: 10.1109/CSNT.2015.22
6. Kohlhase, M., Prodescu, C.: MathWebSearch: low-latency unification-based search. Center for Advanced Systems Engineering, Jacobs University Bremen, Germany, NTCIR-10 (2013)
7. Růžička, M.: Maths information retrieval for digital libraries. Technical report, Brno University (2013)
8. Adeel, M., Cheung, H.S., Khiyal, S.H.: Math go! Prototype of a content based mathematical formula search engine. J. Appl. Theor. Inf. Technol. JATIT **4**(10), 1002 (2008)
9. Kohlhase, M.: An open markup format for mathematical documents. Technical report, Computer Science, International University Bremen (2009)
10. Moens, M.-F., Vulić, I.: Monolingual and cross-lingual probabilistic topic models and their applications in information retrieval. In: Serdyukov, P., Braslavski, P., Kuznetsov, S.O., Kamps, J., Rüger, S., Agichtein, E., Segalovich, I., Yilmaz, E. (eds.) ECIR 2013. LNCS, vol. 7814, pp. 874–877. Springer, Heidelberg (2013)
11. Caprotti, O., Cohen, A.M., Cuypers, H., Sterk, H.: 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. Sojka, P., Líška, M.: Indexing and searching mathematics in digital libraries. Masaryk University, Faculty of Informatics, Botanická 68a, 602 00 Brno, Czech Republic (2011)
13. Ion, P.D.F.: MathML: a key to math on the web. Mathematical Reviews, P.O. Box 8604, Ann Arbor, MI 48107, USA (1999)
14. Anca, S., Kohlhase, M.: MaTeSearch, a combined math and text search engine. Jacobs University (2007)
15. Oetiker, T., Partl, H., Hyna, I., Schlegl, E.: The not so short introduction to LATEX. Version 5.04 (2014)
16. Trung Hung, V., Tuan, C.X.: MathML for the management of mathematical formula in text editor. Int. J. Eng. Res. Technol. **4**(05) (2015)
17. Trung Hung, V., Tuan, C.X.: VM-SEMWEB: a semantic web for vietnamese mathematical documents. Int. J. Eng. Res. Technol. **4**(05) (2015)
18. https://en.wikipedia.org/wiki/Egomath
19. http://www.wiris.com/
20. http://www.leactivemath.org/