S. L. Wong et al. (Eds.) (2010). Proceedings of the 18th International Conference on Computers in Education. Putrajaya, Malaysia: Asia-Pacific Society for Computers in Education.

Development of IR Tool for Tree-Structured MathML-based Mathematical Descriptions

Tetsuya WATANABE[a*], Yoshinori MIYAZAKI[b]
[a]*Graduate School of Informatics, Shizuoka University, Japan*
[b]*Faculty of Informatics, Shizuoka University, Japan*
*gs09061@s.inf.shizuoka.ac.jp

Abstract: The quantity of Web contents including math has skyrocketed in recent years. Some pieces of previous research have dealt with the development of IR systems targeting MathML-based math expressions. They are still developing in terms of lack of fuzzy search functions or low hit rates. One of the authors in ICCE2008 proposed the IR tool enjoying a fuzzy search function by adopting regular expressions. The objective of this study is to propose a "tree structure" algorithm for the fuzzy search function with better precision.

Keywords: IR system for math, MathML, fuzzy search, tree structure

Introduction

With the expansion of Web contents dealing with math expressions, the needs have increased for IR(Information Retrieval) systems treating math. Some pieces of previous research have worked on the development of IR systems targeting MathML-based math expressions. It has to be admitted, however, that none of them are complete in terms of availability for composite retrieval and fuzzy search. In [1], one of the authors reported on the IR tool implementing fuzzy search. The validity and usefulness of the tool were verified, but the limitation of the tool was also revealed. In this study, it is our objective to propose a "tree structure" algorithm for the fuzzy search function with better precision.

1. Previous Research

[2], [3], and [4] are similar to our study although the systems are not found completely satisfactory in either 1. math structure is not fully considered, 2. indexing tag information is too rigid to realize fuzzy searching, or 3. only partial implementation has been made.
[5] is unique in proposing an IR system by incorporating math expressions in "extended" MathML-formats for better grasping their (mathematical) meanings. Likewise, no results of implementation or experiments are shown.

2. Retrieval with Tree Structure

(Presentation) MathML data need to be further hierarchized before retrieving process since parentheses, operators, variables and numbers are all treated as siblings. This doesn't help the system execute retrieval considering data structure. Another problem is that MathML allows degree of freedom in describing same math expressions. In order to solve these problems, MathML data is further hierarchized and for minimizing noises in retrieving. The tree structure for "z=x/(-2y+1)" is updated as Fig. 1 after processing hierarchization.

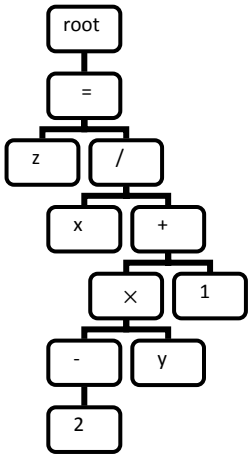
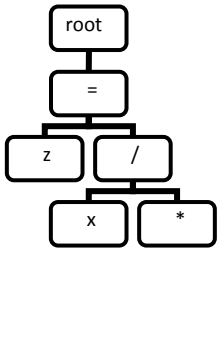
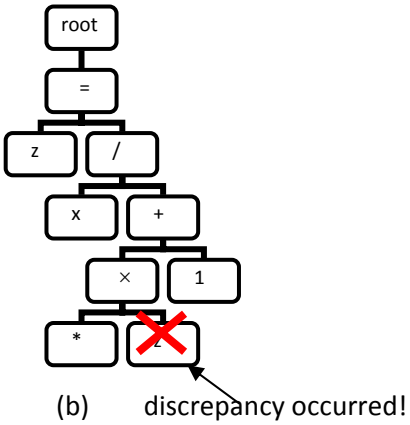


Fig. 1: Optimized Tree Structure



(a)



(b)

Fig. 2: Examples of Query Tree

A query tree is compared with target trees by the following retrieving algorithm (depicted by a pseudo-program style). Its implementation doesn't require lots of labor since operations for comparison are provided in line with the service for DOM.

<< Retrieving Algorithm>>

```
Set t1 = pointer to root node of target tree;
Do
    Matched = True; Set t = t1; Set t1 = Move(t1);
    Set q = pointer to root node of query tree;
    Do
        Set q = Move(q);
        If (Node(q) < > "*" {
        If (corresponding node of target tree doesn't exist) { Matched = False; Exit Do;}
        Else {Set t = Move (t); If (Node(q) < > Node(t)) { Matched = False; Exit Do;}}}
    While (next unvisited node of query tree exists)
While (not Matched Or next unvisited node of target tree exists)
```

Move(p) sets p to point to

1. p's child node
2. p's unvisited sibling node
3. p's parent node

, according to these priorities.

If Matched = True holds true when the outer Do loop is exited. This represents that the query math expression matched the target expression in some way. It is verified that, for instance, the query tree of Fig. 2 (a) matches the target tree (Fig. 1) while that of (b) doesn't.

3. Experiments

3.1 Outline of Experiments

IRs were attempted for 1,000 MathML data sampled from Wolfram Functions Site (http://functions.wolfram.com/), with two algorithms separately. The results of the experiments are shown in Table 1. The numbers of math expressions hit by the selected algorithm are given in the table. The numbers in parentheses are the ones hit incorrectly.

Table 1: Results of Experiments ("*" is a wildcard)

Retrieval / Algorithm	IR1	IR2	IR3	IR4	IR5
Regular Expression	9	40(9)	18	1	231(215)
Tree Structure	9	37	18	1	16
Correct numbers	9	37	18	1	16

IR1: $\cos^2(z)$, IR2: $\frac{\sqrt{*}}{*}$, IR3: $\cos^2(*)$, IR4: $\frac{\sin(*)}{\cos(*)+*}$, and IR5: $\frac{*-*}{*+*}$

4. Summary and Future Plan

In this study, IR tool targeting MathML-based math expressions was implemented. Focusing on its feature, IR algorithm applying tree structure as well as hierarchization of presentation MathML were proposed. Our future plan is the implementation of IR tool with fuzzy search function, with no restrictive conditions. The augumentation of dataset is also strongly expected for more various search. Another constraint of the current tool is its interface, especially in terms of query-inputs in math expressions.

References

[1] Y. Miyazaki, Y. Iguchi, "Development of Information Retrieval Tool for MathML-based Math Expressions", ICCE2008, pp. 419-426 (2008).
[2] T. Nakanishi, et al., "An Implementation Method of Composite Association Retrieval System for Data of Mathematical Formulas", DBSJ Letters, Vol.4, No.1, pp. 29-32 (2005).
[3] H. Hashimoto, et al., "A Survey of index formats for the search of MathML objects", IPSJ SIG Notes 2007-DBS-142-(10), Vol.2007, No.54, pp. 55-59 (2007).
[4] K. Otagiri, T. Murata, "Search of Mathematical Formulas Using MathML", 22nd Annual Conference of the Japanese Society for Artificial Intelligence, (2008).
[5] B. R. Miller, A. Youssef, "Augmenting Presentation MathML for Search", Intelligent Computer Mathematics, AISC 2008, pp. 536-542 (2008).