一种基于二叉树的数学公式匹配算法

秦玉平[1] 唐亚伟[2] 伦淑娴[1] 王秀坤[3]
(渤海大学工学院 锦州 121000)[1] (渤海大学信息科学与技术学院 锦州 121000)[2]
(大连理工大学计算机科学与技术学院 大连 116024)[3]

摘 要 提出了一种基于二叉树结构的LaTeX格式数学公式匹配算法。首先根据数学公式的LaTeX格式生成其二叉树表示,并对树形结构作归一化处理,然后先序遍历二叉树得到公式元素序列,并对序列中的变量名称作归一化处理。对于待匹配的两个数学公式,根据两个公式元素序列对应位相同的公式元素数计算两个公式的相似度。实验结果表明,该算法实现了数学公式的准确匹配,是一种较实用的算法。

关键词 数学公式,二叉树,归一化,相似度
中图法分类号 TP181 **文献标识码** A

Mathematical Formula Matching Algorithm Based on Binary Tree

QIN Yu-ping[1] TANG Ya-wei[2] LUN Shu-xian[1] WANG Xiu-kun[3]
(College of Engineering, Bohai University, Jinzhou 121000, China)[1]
(College of Information Science and Technology, Bohai University, Jinzhou 121000, China)[2]
(School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China)[3]

Abstract A mathematical formula matching algorithm based on binary tree was proposed. Firstly, generating the binary tree form of a mathematical formula by its LaTeX form, normalizing the binary tree structure, and then pre-order traversing the binary tree to get the formula elememn sequence, normalizing the variable names. For two mathematical formulas to be matched, the similarity is computed by the number of the equal formula element at corresponding position. The experimental results show that the algorithm realizes the accurately recognition of mathematical formula, so it is a more practical algorithm.

Keywords Mathematical formula, Binary tree, Normalization, Similarity

1 引言

为了保护知识产权,防止学术论文抄袭,论文查重检测技术已成为信息检索领域的一个研究热点,并取得了一些研究成果。针对英文学术论文的检测主要有数字指纹法和词频统计方法[1-3],针对中文学术论文的检测主要有篇章结构相似度法、段落结构相似度法和句子相似度法[4-6],但这些方法都只适用于纯文本内容的检测。目前,对数学公式的识别已取得了一些研究成果[7-10],但对数学公式的抄袭检测尚处于探索阶段。一篇学术论文,尤其是自然科学学术论文,其思想、观点或创意等核心内容往往主要体现在公式中,因此,在学术论文抄袭检测中,对数学公式的匹配检测具有十分重要的意义。

LATEX软件被广泛地用于制作科技论文、书籍、档案、学位论文、手稿、私人信件以及各种复杂的符号公式等。另外,其他格式的文档可转化为LaTeX格式[11],为此,本文提出了一种基于二叉树结构的LaTeX格式数学公式匹配算法。首先在LaTeX格式文档中提取用字符串表示的数学公式,通过词法和结构分析,生成其二叉树表示,并对二叉树的树形结构作归一化处理,然后先序遍历二叉树,得到公式元素的先序遍历序列,并对变量名作归一化处理,最后根据序列中对应位相同的公式元素数计算两个公式的相似度[12,13]。

本文第2节介绍了由数学公式的LaTeX格式生成其二叉树表示的方法;第3节详细阐述了基于二叉树的数学公式匹配算法;第4节给出了对数学公式进行不同修改的实验结果;最后得出结论。

2 数学公式的二叉树表示

2.1 二叉树构造

在LaTeX文档中,根据标记提取用字符串形式表示的数学公式。由于LaTeX格式的数学公式具有较强的结构特征,因此,可将一个复杂的数学公式分解为多个子式,再将每个子式分解成多个更小的子式[7],如此反复,直到不能分解为止。分解后得到的最小子式称为公式元素。

对于三目运算符,如求和运算"Σ",它与上、下、右都有联

到稿日期:2012-07-23 返修日期:2012-10-19 本文受国家自然科学基金(60974071),辽宁省自然科学基金(201202003),辽宁省教育厅重点实验室项目(LS2010180)资助。

秦玉平(1965-),男,博士,教授,主要研究领域为机器学习,E-mail:jzqinyuping@gmail.com;唐亚伟(1988-),硕士生,主要研究领域为机器学习;伦淑娴(1972-),女,博士,教授,主要研究领域为模式识别;王秀坤(1945-),女,教授,博士生导师,主要研究领域为数据库系统。

系，通过增加一个“link”运算符，将其与右面的子式结合起来。

从左向右遍历增加了“link”运算符后的字符串，生成了公式元素的优先级列表。依据结构特征和优先级列表可生成数学公式的二叉树表示。二叉树中结点的数据结构如表 1 所列。

表 1　二叉树结点的数据结构

成员	数据类型	含义
公式元素	字符串	公式元素
元素类别	字符串	OPS(可交换运算符),OPU(不可交换运算符),VAR(变量),CON(常量)
优先级	整型数	运算符优先级(值越大,优先级越高,常量和变量的优先级为最大机器数)
结合方式	字符串	LR(左右),UD(上下),SG(独立)
结点高度	整型数	以该结点为根的二叉树高度
结构码	字符串	以该结点为根的树结构(二叉树结点的结构码为:“左子树结构码”+“结点高度”+“右子树结构码”)

由数学公式的公式元素串表示生成其二叉树表示的方法是：先建立根结点，根结点的公式元素是公式元素串中优先级最低的公式元素，由于公式串中位于根结点公式元素之前的公式元素在根结点的左子树，位于根结点公式元素之后的公式元素在根结点的右子树，因此递归建立根结点的左右子树。

每个结点的公式元素类别和结合方式根据公式元素确定。每个结点的高度根据式(1)计算。每个结点的结构码在对树形结构作归一化处理后生成。

$$H=H_r>H_l?\ H_r:H_l+1 \tag{1}$$

式中，H_l 是左子树高度，H_r 是右子树高度。

例如，数学公式$(\sum_{i=1}^{10}a^i+x\times y\times z)\times(x\times y+y\times z)$的 LaTeX 格式为“(\sum_{i=1}^{10}a^i+x \times y \times z)\times(x \times y+y\times z)”，其对应二叉树表示如图 1 所示。

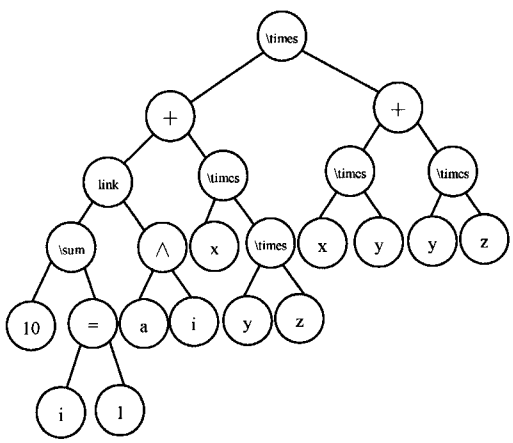

图 1　数学公式的二叉树表示

2.2　归一化处理

由于一些运算符的操作数具有可交换属性，即交换前后数学公式的含义不变，但它们对应的二叉树树形结构可能不同，因此，必须对二叉树树形结构作归一化处理。树形结构归一化处理的方法是先序遍历二叉树，若当前结点公式元素类别为 OPS 且其左子树的高度大于右子树的高度，则交换其左右子树。例如，对图 1 作归一化处理后的二叉树如图 2 所示。

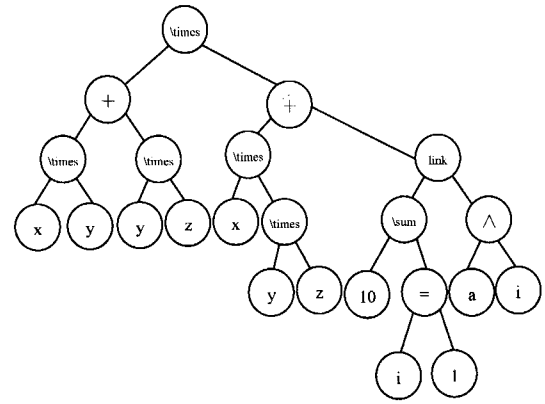

图 2　树形结构归一化后的二叉树

对树形结构作归一化处理后，通过后序遍历二叉树生成各结点的结构码。其中，叶子结点的结构码为 1，非叶子结点的结构码为“左子树根结点结构码”+“结点高度”+“右子树根结点结构码”。

另外，数学公式中的变量名与公式含义无关。对于一个树形确定的二叉树，为使按给定的遍历方式遍历后得到的公式元素序列唯一，必须对序列中的变量名作归一化处理。对变量名作归一化处理的方法是按从左到右的顺序，用固定的变量名序列依次替换遍历序列中公式元素类别为 VAR 的公式元素。

3　数学公式匹配算法

设 E_1 和 E_2 是两个 LaTeX 格式的数学公式，其中，E_1 是源公式，E_2 是目标公式，其具体匹配算法描述如下：

步骤 1　生成 E_1 的二叉树表示并对树形结构作归一化处理，得二叉树 T_1。先序遍历二叉树 T_1，得公式元素序列，对序列中的变量名作归一化处理，得归一化后的公式元素序列 L_1；

步骤 2　生成 E_2 的二叉树表示并对树形结构作归一化处理，得二叉树 T_2。先序遍历二叉树 T_2，得公式元素序列，对序列中的变量名作归一化处理，得归一化的公式元素序列 L_2，若 T_1 和 T_2 根结点的结构码不相同，则根据式(2)计算 E_1 和 E_2 的相似度，转步骤 5，否则转步骤 3；

$$\mathrm{sim}(E_1,E_2)=\mathrm{sim}(L_1,L_2)=\frac{2*matnum(L_1,L_2)}{strlen(L_1)+strlen(L_2)} \tag{2}$$

式中，$matnum(s,t)$为公式元素序列 s 和 t 中对应位相同的公式元素个数，$strlen(s)$为公式元素序列 s 中的公式元素个数；

步骤 3　若 T_2 中不存在公式元素类别为 OPS 且左、右子树高度相同的结点，则根据式(2)计算 E_1 和 E_2 的相似度，转步骤 5，否则转步骤 4；

步骤 4　对 T_2 中公式元素类别为 OPS 且左、右子树高度相同的结点(设有 n 个)，交换这类结点的左右子树，可得到 2^n-1 棵树形结构与 T_2 相同的二叉树序列 $T_2^1,T_2^2,\cdots,T_2^{2^n-1}$，分别先序遍历这 2^n-1 棵二叉树，得到相应的公式元素序列，分别对每个公式元素序列中的变量名作归一化处理，得到 2^n-1 个归一化后的公式元素序列 $L_2^1,L_2^2,\cdots,L_2^{2^n-1}$，先根据式(2)计算 L_1 与 $L_2^i(i=1,2,\cdots,2^n-1)$的相似度，再根据式(3)计算 E_1 和 E_2 的相似度，转步骤 5；

(下转第 278 页)

Euclidean skeleton in 2D and 3D[J]. Image and Vision Computing, 2007, 25(10): 1543-1556

[5] Beristain A, Grana M. Pruning algorithm for voronoi skeletons [J]. Electronics Letters, 2010, 46(1): 39-41

[6] Couprie M, Bertrand G. New characterizations of simple points in 2D, 3D and 4D discrete spaces[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(4): 637-648

[7] Chen Y, Drechsler K, Zhao W, et al. A Thinning-based Liver Vessel Skeletonization Method[C]// International Conference on Internet Computing and Information Services. 2011: 152-155

[8] Palagyi K, Kuba A. Directional 3D Thinning Using 8-Subiterations[C]// Proceedings of Discrete Geometry for Computer Imagery. 1999, 1568: 325-336

[9] 陈磊，王胜军，郑全录，等. 基于CT图像的三维拓扑细化算法及其在心脏CAD中的应用[J]. 计算机应用，2007，6(27)：406-410

[10] 滕奇志，康瑕，唐棠，等. 基于升序复核的并行三维图像骨架化算法[J]. 光学精密工程，2009，10(17)：2528-2534

[11] Friman O, Hindennach M, Khnel C, et al. Multiple hypothesis template tracking of small 3D vessel structures[J]. Medical Image Analysis, 2010, 14(2): 160-171

[12] 朱应礼，徐益明，崔秋梅. MSCT肺血管成像对肺动脉栓塞的诊断价值[J]. 医学影像学杂志，2008，18(6)：597-599

[13] Lin K S, Tsai C L, Sofka M, et al. Vascular tree construction with anatomical realism for retinal images[C]// Proceedings of the 9th IEEE International Conference on BioInformatics and BioEngineering. 2009: 313-318

(上接第 252 页)

$$\mathrm{sim}(E_1,E_2)=\max(\mathrm{sim}(L_1,L_2),\mathrm{sim}(L_1,L_2^1),\mathrm{sim}(L_1,L_2^2),\cdots,\mathrm{sim}(L_1,L_2^{2^n-1})) \quad (3)$$

步骤 5　结束。

4　实验结果及分析

数学公式抄袭的手段主要有原样复制、修改变量名称和交换可交换运算符两边子式的位置等。为了验证算法的性能，本文实验从已公开发表的学术论文中选取 50 个含义和结构互不同的数学公式进行测试。根据抄袭手段的不同，人工对这些公式进行修改，具体的修改方式如表 2 所列。

表 2　修改方式的种类

修改种类	1	2	3	4	5
修改方式	未作修改	修改变量名称	交换可交换符号两边子式	修改变量名称和交换可交换符号两边子式	修改常量值

实验中，按表 2 的修改方式修改每个公式，并分别计算源公式与修改后的目标公式的相似度，每一种修改方式的平均相似度如表 3 所列。

表 3　各种修改方式的检测结果

修改方式	1	2	3	4	5
相似度(%)	100	100	100	100	78.41

从实验结果可以看出，未作修改的公式、修改变量名称的公式以及交换可交换运算符两边子表达式的公式，其相似度都为 100%，这是因为对源公式和目标公式的树形结构作归一化处理后，源公式和目标公式的树形结构相同，虽然与源公式树形结构形同的目标公式二叉树可能有多种，但在对所有的先序遍历序列中变量名称作归一化处理后，目标公式的遍历序列中至少存在一个与源公式遍历序列完全相同的遍历序列，使修改后的目标公式恢复原形。修改常量值时的相似度较低，这是因为修改公式中的常量后，公式的含义在一定程度上与源表达式不同，即源表达式和目标表达式是两个不完全相同的表达式。例如，$\sum_{i=1}^{10} i^2$ 和 $\sum_{i=10}^{20} i^3$ 是两个含义不相同的数学表达式。另外，算法中用结构码标识树形结构，并且字符串匹配过程是以公式元素为单位，匹配速度较快。

结束语　基于数学公式的二叉树表示，本文提出了一种 LaTeX 格式的数学公式匹配算法。该方法对未作修改的公式、修改变量名称的公式和交换可交换运算符两边子表达式的公式进行的检测都十分精确，且具有较高的检测速度，有效地解决了论文抄袭检测中的数学公式检测问题。但有些问题还有待进一步研究，如具有分配属性的运算处理、矩阵处理以及复杂的多行表达式处理等。

参 考 文 献

[1] Mander U, Baker B S. Deducing similarities in java sources from bytecode [C]// Usenix 1998 Annual Technical Conference. New Orleans: The Advanced Computing Systems Association, 1998: 179-190

[2] 史彦军，滕弘飞，金博. 抄袭论文识别研究与发展[J]. 大连理工大学学报，2005，45(1)：50-57

[3] 鲍军鹏，沈钧毅，刘晓东，等. 自然语言文档复制检测研究综述[J]. 软件学报，2003，14(10)：1753-1761

[4] 金博，史彦军. 基于篇章结构相似度的复制检测算法[J]. 大连理工大学学报，2007，47(1)：125-130

[5] 秦新国. 基于句子相似度的文档复制检测算法研究[J]. 现代图书情报技术，2007(11)：63-66

[6] Si A, Leong H V, Lau R W H. CHECK: A document plagiarism detection system[C]// Proceedings of the ACM Symposium for Applied Computing. 1997: 70-77

[7] 郭育生，黄磊，刘昌平. 基于多候选的数学公式识别系统[J]. 计算机研究与发展，2007，44(7)：1144-1150

[8] Chan K F, Yeung D Y. Mathematical Expression Recognition: A Survey [J]. International Journal on Document Analysis and Recognition, 2000, 3(1): 3-15

[9] 靳简明，江红英，王庆人. 数学公式图像处理综述[J]. 模式识别与人工智能，2005，18(4)：429-440

[10] Lee H-J, Wang J-S. Design of a Mathematical Expression Recognition System[J]. Pattern Recogniton Letters, 1997, 18(8): 289-298

[11] 靳简明，江红英，王庆人. 数学公式识别系统：MatheReader[J]. 计算机学报，2006，29(11)：2018-2026

[12] 刘宏哲，须德. 基于本体的语义相似度和相关度计算研究综述[J]. 计算机科学，2012，39(2)：8-13

[13] 林学民，王炜. 集合和字符串的相似度查询[J]. 计算机学报，2011，34(10)：1853-1862