基于LaTex的Web数学公式提取方法研究

陈立辉 苏 伟 蔡 川 陈晓云
(兰州大学信息科学与工程学院 兰州 730000)

摘 要 数学论坛、Wiki等社会性网站对数学教育的影响日益增长,数学公式广泛存在这些网站中,如何对这些网站中的数学公式进行搜索,对学习和科研非常重要。数学公式提取是索引系统的前提和基础,文中主要研究LaTex格式的数学公式的提取方法,结合BNF表述方式,提出自动分析提取包含LaTex公式特征的方法。依据公式包含的特征,提出提取和过滤LaTex数学公式的方法规则。通过实验发现,该方法的查全率达到75%,查准率达到99%。

关键词 数学公式,LaTex,查准率,查全率,主题爬虫,搜索引擎
中图法分类号 TP311 **文献标识码** A

Research of Extraction Method of Web Mathematical Formula Based on LaTex

CHEN Li-hui SU Wei CAI Chuan CHEN Xiao-yun
(School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, China)

Abstract The influence of Wiki, mathematics forum and other social networking sites on the mathematics education field is growing. Mathematical formulas exist widely in these websites. How to search the mathematical formulas of these websites is very important for study and research. The extraction of mathematical formulas is the premise and foundation of the indexing system. This paper mainly studied the format of the LaTex mathematical formulas, and presented the automatic analysis extraction method of Web mathematical formulas based on LaTex through the BNF paradigm. According to features the formulas contain, the paper proposed the method of extraction and filtration of LaTex mathematical formula. The experiment discovers that the recall rate reaches 75% and the precision rate comes to 99% using this method.

Keywords Mathematical formula, LaTex, Precision, Recall, Topic crawler, Search engine

1 引言

随着计算机的不断普及,出现了互联网的高速发展和Web信息的爆炸式增长。用户如何从海量信息中查找所需要的信息,是搜索引擎解决的关键问题。目前,文本搜索引擎已较为成熟,从本地的文件搜索到互联网上信息的搜索,文本搜索已被用户广泛应用。

数学是科学的工具,数学信息广泛存在于各种文献资料中,在教育教学和科学研究中起着非常重要的作用,而数学是借助数学公式来进行描述和表现的。网络与人们的关系越来越紧密,数学论坛、Wiki[1,2]等社会性网站在社会、经济、文化和教育方面都具有重大的影响和重要的意义。在Web中,数学论坛、Wiki等社会性网站包含着大量的数学公式,如何提取这些公式则显得尤为重要。Web中数学公式描述格式包括图片、MathML、LaTex、OpenMath、Infix等。调查发现,当前的数学论坛、Wiki等网站中的数学公式大部分是LaTex格式。

LaTex[3]是一种基于Tex的排版系统,它是由美国计算机学家Leslie Lamport在20世纪80年代初期开发的。利用这种格式,即使使用者没有排版和程序设计的知识也可以充分发挥由TeX所提供的强大功能。LaTex强大而优美的公式排版和渲染具有简洁、成熟和通用等优点。LaTex在渲染成图片之前本身是文本形式,不仅可以把数学公式应用在LaTex文档排版中,也能够将LaTex应用于互联网页面数学公式显示。LaTex数学公式中的数学结构主要包括分数、方根、指数与指标、域以及常用的数学修饰符等。

现今流行的搜索引擎如Baidu、Google等尚不支持数学公式方面的搜索,而具有数学搜索[4-8]能力的搜索引擎有MathWebSearch、EgoMath、MathSearch、MathDex、LeActiveMath、DLMF Search等。

MathWebSearch[9]是一个基于Content MathML的数学搜索引擎,系统中网络爬虫采集的数据存放在MySQL数据库中,主要支持MathML、OpenMath等XML形式的数学公式的搜索。EgoMath是一个基于Egothor v2[10]全文本搜索引擎的可识别数学内容的搜索引擎,它支持Presentation MathML和Content MathML表示方式,对于PDF文件,用

到稿日期:2013-08-23 返修日期:2014-01-06 本文受国家自然科学基金项目(61003139,60903102),教育部—英特尔信息技术专项科研基金(MOE-INTEL-11-03),中央高校基本科研业务费专项资金(lzujbky-2013-39,lzujbky-2013-188,lzujbky-2013-187)资助。
陈立辉(1987—),男,硕士生,主要研究方向为数学搜索、符号计算,E-mail:chenlh11@lzu.edu.cn;苏 伟(1977—),男,副教授,主要研究方向为语义搜索引擎、数学知识工程与管理、教育信息化技术、信息无障碍技术,E-mail:suwei@lzu.edu.cn(通信作者);蔡 川(1979—),女,博士生,主要研究方向为网格计算、计算机数学;陈晓云(1954—),女,教授,主要研究方向为数据挖掘、数据仓库、Web挖掘、气象信息处理。

Infty 转换器转换为 Presentation MathML 标记的数学文件。MathDex[11]是一个以 Apache Lucene 为基础引擎的数学公式引擎，能够支持不同编码格式的数学公式无语义搜索，MathDex 首先将所有检索到的文件转换成 XHTML+MathML 格式文件，根据结构及与查询项的语法相似程度进行排序。LeActive[12]是基于 Apache Lucene 的，该系统针对的数据源是 OMDoc[13]编码的具有语义的数学文件，索引阶段依赖于特殊的 OMDoc 格式，其包含语义信息以及其他的元数据。DLMF Search[14]是为美国国家标准与技术研究所的数学公式数字图书馆而建立的一个检索系统，美国国家数学公式数字图书馆中 80%以上的用户手册或网页中包含数学公式并且数学公式的存储格式以 TeX/LaTex 为主。因此，DLMF Search 的目标公式格式主要针对于 TeX/LaTex 格式。MathSearch[15]是一个基于公式的数学语义搜索引擎，能够实现对互联网上含有数学公式、符号的网页和文档等相关内容进行搜索；MathSearch 采用 Lucene 为原型系统，提供基于 Presentation MathML 和 Content MathML 的搜索，主要由 MathCrawler(网络爬虫器)、MathIndexer(索引器)和 MathUI(搜索器)3 部分组成。

上述的几种数学搜索中，MathSearch、LeActiveMath 针对的是 MathML 数学公式；MathWebSearch、EgoMath、MathDex、DLMF Search 能够对 LaTex 数学公式进行检索，对其中的 LaTex 数学公式采用手工下载和格式转换工具转换索引需要形式。然而这些数学搜索基本都是针对特定数据源的，对于未知的包含 LaTex 数学公式网页，进行公式获取存在一定困难。因此，本文在 MathSearch 中 MathCrawler 的基础上，对网页中 LaTex 数学公式的识别和提取进行研究，提出基于特征和过滤规则的公式识别和提取的方法。

2 相关工作

数学搜索 MathSearch 课题组于 2008 年起做了相关研究工作，文献[16]研究数学公式查询系统，并构造出数学查询语言 MQL；文献[17]研究搜索过程中数学公式的索引系统，并提出了基于语义和结构的索引方式；文献[18]研究数学搜索中的爬虫系统，对 Web 中的常见数学公式格式进行调研，并提出 MathML 格式的数学公式的识别与提取方法。本文主要研究 Web 中 LaTex 格式的数学公式识别提取方法，属于爬虫系统 MathCrawler 研究的一部分。在前期公式识别与提取研究中，对数学公式的提取方法进行了相关调研，为本文识别与提取 LaTex 格式的数学公式提供了参考。

目前，多数数学搜索引擎主要选取固定数据源作为爬虫系统的目标，其中嵌入数学公式特征固定，在应对未知数学源时，难以完成数学公式的识别与提取。为了提取未知数据源中的公式，本文提出自动分析获取网页嵌入 LaTex 数学公式特征方法。

在 LaTex 格式数学公式识别方法中，主要通过公式的组成元素和结构的特征来识别。LaTex 是一种面向排版布局的语言，而纯粹的 LaTex 数学符号组合会出现一些非公式组合，要在符号组合基础上增加表达式含义，才能表示为数学公式，因而，本文结合 LaTex 格式数学公式的结构信息和数学公式本身的语义信息来识别 LaTex 数学公式。

在分析 LaTex 数学公式嵌入特征过程中，需要使用 LaTex 数学公式的组成符号来识别公式，由于其出现的概率会影响特征识别效率和准确性，在特征识别过程中需要将其考虑在内。

3 LaTex 数学公式识别与提取

爬虫[19-22]作为搜索引擎的一部分，主要作用是抓取和解析互联网中的网页。在 MathSearch 中，爬虫 MathCrawler 作为其重要组成部分，能够自动提取含有数学公式的网页。但为了能够对网页中 LaTex 数学公式进行提取，本文对 MathCrawler 进行了改进。改进后的 MathCrawler 主要包含两部分内容：网页读取模块和网页内容分析模块。图 1 显示了 MathCrawler 的具体工作流程。

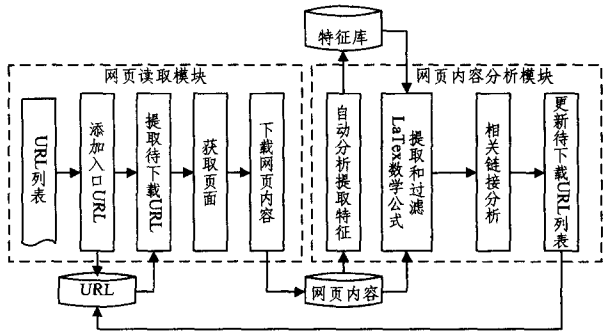

图 1 MathCrawler 结构图

为了能够精确提取出网页中的 LaTex 数学公式，本文提出了基于特征和过滤规则的公式识别和提取方法。为了使表述清晰，本文定义了以下集合符号(见表 1)。

表 1 集合符号及含义

集合	表示含义
U_{entry}	爬虫入口 URL 集合
U_{wait}	爬虫待下载 URL 集合
$U_{contain}$	网页中包含链接 URL 集合
U_{wiki}	网页中有关 Wiki 的链接 URL 集合
$D_{document}$	爬虫下载网页文档集合
T_{text}	提取特征库集合
C_{latex}	LaTex 数学符号集合
G_{filter}	过滤规则集合
G_{URL}	正则表达式 URL 过滤集合

爬虫的具体工作流程如下：

(1)将部分网站的首页添加到集合 U_{entry} 中作为爬虫入口 URL，并将其添加到集合 U_{wait} 中；

(2)从集合 U_{wait} 中提取 URL，通过 Web 获取网页，添加到集合 $D_{document}$ 中；

(3)对集合 $D_{document}$ 中的元素文档，分析 LaTex 数学公式提取特征，将特征添加到特征集合 T_{text} 中；

(4)依据特征集合 T_{text} 和网页内容集合 $D_{document}$，进行 LaTex 数学公式提取和过滤；

(5)对集合 $D_{document}$ 中的元素文档，分析页面内容，获取页面中集合 $U_{contain}$，通过相关链接分析得到集合 U_{wiki}，将集合 U_{wiki} 中的内容更新到集合 U_{wait}；

(6)如此往复。

3.1 自动分析提取包含 LaTex 数学公式特征

随着数学论坛、Wiki 等社会性网站的影响越来越广泛，为了便于搜索引擎的搜索，识别和提取数学公式显得尤为重要。通过研究此类网页中包含的数学公式，发现存在大量

LaTex格式数学公式，并且，嵌入的LaTex数学公式存在一定特征[23]，如在Wiki中公式

$$\alpha > \int \frac{\beta}{\Sigma\gamma}$$

其源码表示为：

〈img class="tex" alt=" \alpha\geq\int\frac{\beta}{\sum\gamma}" src="//upload. wikimedia. org/math/9/b/9/9b953d65e80b05b46e4dfecce5639cbd. png"〉

可以看出Wiki中LaTex数学公式被存放在元素标记〈img〉中，其源码包含在特征属性alt内容中。在其它数学网站中也具有其他类似特征。

目前，Web中在线处理LaTex数学公式的方式主要包括以下几种：mimetex服务、WordPress. com的LaTex服务器、mathtex服务、LatexRender、JSMath/MathJax。公式在Web中的源码嵌入方式虽有不同，却具有一定特征，表2列出了同一公式的不同嵌入方式。

表2　不同处理方式下的源码表示

处理方式	Web中源码
mimetex服务	〈img src="http://tex. 72pines. com/latex. php? latex=\alpha\geq\int\frac{\beta}{\sum\gamma}" alt=""〉
WordPress. com的LaTex服务器	〈img src="http://l. wordpress. com/latex. php? latex=\alpha\geq\int\frac{\beta}{\sum\gamma}" alt=""〉
mathtex服务	〈img src="http://www. forkosh. dreamhost. com/mathtex. cgi? \alpha\geq\int\frac{\beta}{\sum\gamma}"alt="" border=0 align=middle〉
LatexRender	〈img src="d21848cdd835abcb491be1f151e9b6c6. gif" alt="\alpha\geq\int\frac{\beta}{\sum\gamma}" align="ABSMIDDLE" border="0" height="17" hspace="0" vspace="0" width="22"〉
JSMath/MathJax	〈p class="math-type-block"〉\(\alpha\geq\int\frac{\beta}{\sum\gamma}\)〈/p〉 〈p class="math-type-block"〉$\alpha\geq\int\frac{\beta}{\sum\gamma}$〈/p〉 〈p class="math-type-block"〉\[\alpha\geq\int\frac{\beta}{\sum\gamma}\]〈/p〉

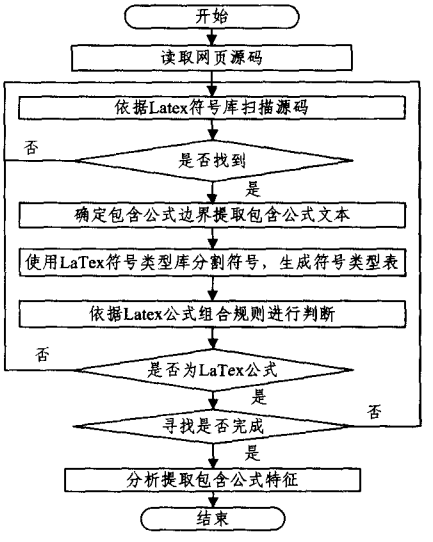

图2　自动分析提取公式特征流程图

网页中LaTex数学公式源码的嵌入方式多种多样，而对特定的一个网站而言，其嵌入LaTex数学公式的方式是固定的，可通过人工的方式分析出嵌入特征。但对大量的网站而言，人工要实现特征识别，效率低下而且不现实，因此一种自动数学公式特征识别方法则显得尤为重要。本文提出了自动分析提取包含LaTex数学公式特征方法，其流程见图2。该方法主要分5步：第1步，在网页内容中查找LaTex符号；第2步，确定包含LaTex公式文本边界，提取包含公式文本；第3步，使用LaTex符号类型库对公式文本分词；第4步，判别是否为LaTex公式；第5步，通过完整的LaTex公式，找出公式的嵌入特征。

下面对该方法进行详细介绍：

第1步，查找LaTex数学符号。在查找LaTex公式过程中，需要使用LaTex数学符号作为查找依据，而这些符号查找的先后顺序直接影响了查找效率。影响LaTex数学公式查找效率的主要因素包含两个方面：

(1)LaTex数学符号在网页公式中出现的频率(Term Frequency)；

(2)相同符号能被识别为LaTex数学符号的正确率(Term Accurate Rate)。

针对这两方面的影响因素，本文提出一种基于频率和正确率的LaTex数学符号影响因子计算方法(TF-TAR)，计算公式如下：

$$I(t)=TF(t)\times TAR(t,s)$$

其中，$I(t)$表示在一个文档集合D中LaTex数学符号t的影响因子；$TF(t)=\sqrt[a]{\frac{n}{N}}$，表示LaTex数学符号$t$在文档集合$D$中所有LaTex数学符号中出现的频率，$N$表示在文档集合$D$中出现的所有LaTex数学符号的次数，开$a$次方是为了平衡两个比率，在本文中$a$值采用4；$TAR(t,s)=\frac{n}{s}$表示符号$t$在网页公式中能被识别为LaTex数学符号的正确率，n表示符号t视为LaTex数学符号的个数，s表示符号t出现的次数。在网页公式中能被识别为LaTex数学符号的正确率越高，即$TAR(t,s)$值越大，影响因子越大；在网页中LaTex数学符号t在所有LaTex数学符号中出现的频率越高，即$TF(t)$值越大，影响因子越大。为了能够更好地查找LaTex数学公式，采用影响因子大于0.1的符号作为查找符号。

第2步　确定包含公式边界。网页在解析嵌入其中的LaTeX格式数学公式时，需要提取公式源码，而公式源码存在的位置会存在固定特征方便网页提取转换。经过分析发现，嵌入方式分两种：一种是嵌入到某个特殊属性，作为其属性值，如：属性名称＝“LaTex公式源码”；另一种是嵌入到某个特殊标记之中，作为文本内容，如：〈标记名称〉LaTex数学公式〈/标记名称〉。为了确定公式边界，在查找到LaTex数学符号时，继续向前后扫描，找到表示属性值内容或标记之间的文本内容，对该内容从左到右进行LaTex数学符号类型识别，生成符号类型表。

第3步　对公式文本分词。LaTex数学公式是由LaTex数学符号和普通字符组成。普通字符由单一字符或字符串组成，其中单一字符是指单个数字或字母，字符串以数字或字母开头，以字母结尾。

本文依据LaTex数学公式的结构特征以及公式本身的符号结合特征，对LaTex数学符号进行了以下分类(见表3)。

表 3　符号类型

类型	包含符号
二元符号	二元关系符、二元运算符、箭头符号、AMS 二元关系符、AMS 二元运算符、AMS 箭头、AMS 否定关系符和箭头、特殊运算符
一元符号	函数、数字字母宏、数学模式重音符、数学修饰符
分组符号	{…}、定界符
忽略符号	\!、\,、\left、\right、\quad、\qquad、\;、空白、\空白、\color、\pagecolor
分数	\frac、\cfrac、\tfrac、\dfrac
根号	\sqrt
求和/求导	\sum 、\int、\iint、\iiint、\iiiint、\oint、\prod、\coprod、\bigcup、\bigcap、\bigsqcup、\biguplus、\bigwedge、\bigvee
上标/指数符号	_、^
堆积符号	\stackrel、\atop、\choose
域符号	\begin 、\end
特殊符号	杂项符号
参数变量	单个字母或数字、字符串、希腊字母

依据表 3 中定义的符号类型，可生成单词类型表(见表 4)。

表 4　单词类型表

单词 1	类型 1
单词 2	类型 2
…	…

第 4 步　判别是否为 LaTex 公式。LaTex 公式是由表 3 中的类型符号组成的，而为了能够清晰地表述组合 LaTex 公式过程，判别是否为 LaTex 公式，本文先对表 3 中的符号类型提出了 BNF 范式描述(见表 5)。

表 5　LaTex 数学公式的符号单词类型 BNF 范式

数字	⟨d⟩::=0|1|2|3|4|5|6|7|8|9
字母	⟨l⟩::=a|b|c|d|e|f|g|h|i|j|k|l|m|n|o|p|q|r|s|t|u|v|w|x|y|z|A|B|C|D|E|F|G|H|I|J|K|L|M|N|O|P|Q|R|S|T|U|V|W|X|Y|Z
分数符号	⟨f_1⟩::="\frac"|"\cfrac"|"tfrac"|"cfrac"
上标/指数符号	⟨s_1⟩::="_" ⟨s_2⟩::="^"
根式符号	⟨q_1⟩::="\sqrt"
求和/求积分符号	⟨h1⟩::="\sum"|"\int"|"\iint"|"\iiint"|"\iiiint"|"\prod"|"\oint"|"\coprod"|"\bigcup"|"\bigcap"|"\bigsqcup"|"\biguplus"|"\bigwedge"|"\bigvee"
分组符号	⟨g_1⟩::="{" ⟨k_1⟩::="}" ⟨g_2⟩::="[" ⟨k_2⟩::="]" ⟨g_3⟩::="(" ⟨k_3⟩::=")" ⟨g_4⟩::="\{" ⟨k_4⟩::="\}" ⟨g_5⟩::=" \langle" ⟨k_5⟩::=" \rangle" ⟨g_6⟩::=" \lfloor" ⟨k_6⟩::=" \rfloor" ⟨g_7⟩::=" \lceil" ⟨k_7⟩::=" \rceil" ⟨g_8⟩::=" \llcorner" ⟨k_8⟩::=" \lrcorner" ⟨g_9⟩::=" \ulcorner" ⟨k_9⟩::=" \urcorner"
域符号	⟨t_1⟩::="\begin" ⟨t_2⟩::="\end" ⟨t_3⟩::="vmatrix"|"Vmatrix"|"bmatrix"|"Bmatrix"|"pmatrix"|"vmatrix"|"smallmatrix"|"cases"|"align"|"alignat"|"array"
堆积符号	⟨n_1⟩::="\stackrel" ⟨n_2⟩::="\atop" ⟨n_3⟩::="\choose"
一元符号	⟨x_1⟩::=⟨一元符号⟩
二元符号	⟨y_1⟩::=⟨二元符号⟩
参数变量	⟨v⟩::=⟨d⟩|⟨l⟩|⟨v⟩⟨l⟩|⟨d⟩⟨l⟩

在第 3 步中生成了单词类型表，依据 LaTex 公式本身的特征来判别表中数据能否组成 LaTex 数学公式，判别要素包括：首先，LaTex 公式由表 3 中类型符号组成；其次，LaTex 公式具有一定的数学结构；再者，若仅仅考虑前两者，会出现许多无意义的数学形式，如 LaTex 数学结构“+-+”等等，这些结构不符合公式的通用形式，LaTex 公式还应具有数学含义。因此，本文鉴于上述 3 点提出了 LaTex 数学公式组合规则的 BNF 范式描述(见表 6)，通过判断是否满足表达式组合规则，来确定完整 LaTex 公式。

表 6　LaTex 数学公式组合规则 BNF 范式

分数	⟨f⟩::=⟨f_1⟩⟨g_1⟩⟨b⟩⟨k_1⟩⟨g_1⟩⟨b⟩⟨k_1⟩
上标/指数	⟨s⟩::=⟨b⟩⟨s_1⟩⟨b⟩|⟨b⟩⟨s_2⟩⟨b⟩
根式	⟨q⟩::=⟨q_1⟩⟨g_1⟩⟨b⟩⟨k_1⟩|⟨q_1⟩⟨g_2⟩⟨b⟩⟨k_2⟩⟨g_1⟩⟨b⟩⟨k_1⟩
求和/求积分	⟨h⟩::=⟨h_1⟩⟨s_1⟩⟨b⟩⟨s_2⟩⟨b⟩|⟨h_1⟩⟨s_2⟩⟨b⟩⟨s_1⟩⟨b⟩
一元运算	⟨x⟩::=⟨x_1⟩⟨v⟩|⟨x_1⟩⟨b⟩
二元运算	⟨y⟩::=⟨b⟩⟨y_1⟩⟨b⟩
域	⟨t⟩::=⟨t_1⟩⟨g_1⟩⟨t_3⟩⟨k_1⟩⟨b⟩⟨t_2⟩⟨g_1⟩⟨t_3⟩⟨k_1⟩
堆积	⟨n⟩::=⟨n_1⟩⟨g_1⟩⟨b⟩⟨k_1⟩⟨g_1⟩⟨b⟩⟨k_1⟩|⟨g_1⟩⟨b⟩⟨n_2⟩⟨b⟩⟨k_1⟩|⟨g_1⟩⟨b⟩⟨n_3⟩⟨b⟩⟨k_1⟩
表达式	⟨b⟩::=⟨v⟩|⟨f⟩|⟨s⟩|⟨q⟩|⟨h⟩|⟨x⟩|⟨y⟩|⟨g_1⟩⟨b⟩⟨k_1⟩|⟨g_2⟩⟨b⟩⟨k_2⟩|⟨g_3⟩⟨b⟩⟨k_3⟩|⟨g_4⟩⟨b⟩⟨k_4⟩|⟨g_5⟩⟨b⟩⟨k_5⟩|⟨g_6⟩⟨b⟩⟨k_6⟩|⟨g_7⟩⟨b⟩⟨k_7⟩|⟨g_8⟩⟨b⟩⟨k_8⟩|⟨g_9⟩⟨b⟩⟨k_9⟩|⟨b⟩⟨b⟩

第 5 步　确定公式嵌入特征。三元组(r,p,q)表示包含公式特征，其中 r 表示属性或文本，p 表示属性值或文本中 LaTex 公式之前的固定标识部分，q 表示属性值或文本中 LaTex 公式之后的固定标识部分，p、q 可为空。在网页中嵌入 LaTex 数学公式方式有两种：属性值嵌入和文本内容嵌入。

(1)对于属性值而言，网站需要解析 LaTex 公式，其属性名称是固定的。为了能够找出特征元素，则需要对第 2 步找到的完整公式继续进行前后扫描，找到特征元素中的 r、p、q 值，通过比较该值，若发现多次相同，则该特征元素即可作为此类网站的提取特征，并添加到提取特征集合中，作为此网站的提取特征，然后提取网页 LaTex 公式。

(2)对于文本内容而言，网站需要解析 LaTex 公式，其标记名称并非固定，一方面，在文本中，完整公式前后无其他内容，那么网站解析 LaTex 公式的特征就必须依据于包含此文本的标记名称，三元组(r,p,q)的值分别为(标记名称，空，空)；另一方面，在文本中，完整公式前后包含其他内容，那么标记名称往往不固定，无法作为提取特征，网站要解析 LaTex 公式，则该完整公式前后必须包含成对的特征，如 $ $、\(\)、\[\]等，通过多次比较提取出的内容，找到公式前后的相同部分，即可作为提取三元组(r,p,q)，其值为(空，开始字符串，结束字符串)，并添加到提取特征集合中，作为此网站的提取特征，然后提取网页的 LaTex 公式。

3.2　提取和过滤 LaTex 数学公式

在 MathCrawler 中网页内容分析模块主要实现获取网页中的数学公式功能。本小节主要介绍提取和过滤 LaTex 公式的主要过程。首先爬虫获取网站页面集合 $D_{document}$，再通过进一步对 $D_{document}$ 中元素页面进行 DOM 解析，提取和过滤出 LaTex 数学公式。图 3 显示了解析页面中提取和过滤 LaTex 公式的流程。

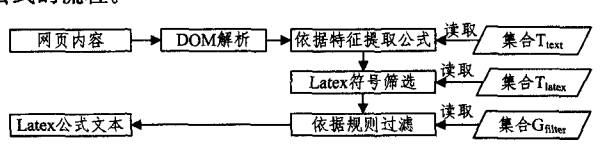

图 3　提取和过滤 LaTex 数学公式流程图

3.2.1 采用提取特征提取公式

集合 T_{text} 中包含了网站 LaTex 数学公式的嵌入特征。运用 DOM 解析，找出满足集合 T_{text} 的节点，提取该节点的包含内容，从而达到提取公式的目的。

3.2.2 采用 LaTex 数学符号筛选公式

通过特征提取出的内容中存在大量噪音。对于 LaTex 数学公式，它是由 LaTex 数学符号集合 C_{latex} 组成的。因此可以使用集合 C_{latex} 对提取出的内容进行筛选，从而增加提取 LaTex 数学公式的精度。

3.2.3 采用过滤规则过滤提取内容

在 LaTex 数学符号筛选出的内容中包含了许多噪音数据(见表 7)。

表 7 噪音类型及示例

编号	噪音类型	示例
1	图片后缀	.jpg、.png、.jpeg、.gif 等
2	网址链接	link =等
3	非 LaTex 数学符号 ASCII 码	Chhim-cháu、纳戈尔诺-卡拉巴赫等
4	说明文字	Page move-protected 等
5	其他	人名、词典未登录词等

对表 7 中的第 5 种噪音而言，由于存在不确定性，无法采用确定的规则去过滤，因此过滤规则集合 G_{filter} 主要针对第 1—4 种噪音，包含的规则元素如下：

规则 1 对于第 1、2 种噪音，它们都有明显的特征，图片均包含固定后缀，而链接包含固定字段，直接通过其包含特征达到过滤目的。

规则 2 第 3 种噪音由于 LaTex 数学公式中不包含 ASCII 码大于 126 的字符，因此可通过字符 ASCII 码值来判断是否为 LaTex 数学公式，部分判断代码见表 8。

表 8 判断为第 3 种噪音代码

```
private boolean isASCIIOfStr(String str) {
  for (int i=0;i < str. length();i++) {
    int acs=str. charAt(i);
    if (acs > 126) {
      return false;}
    }
return true;}
```

规则 3 对于第 4 种噪音主要为英文、中文、其他语言的说明文字。其中的中文和其他语言的说明可通过规则 2 过滤掉，余下的英文说明文字，绝大多数包含空格或连字符(即“-”)。因此过滤英文说明文字，则需要对其进行简单分词，对分出的词匹配单词库的单词。

在规则 3 中，对分出的词匹配的方法有两种：(1)用单词库去匹配分出的词。这种方法可以使用，但有缺点，对于特征属性中的说明文字而言，往往文字比较短小，因此这种方法会影响效率。(2)用分出的词匹配单词库。这种方法也存在缺点，会出现匹配错误，如提取出的公式为“$a+b$”，通过空格分词之后，出现 3 个词“a”、“+”、“b”，用其匹配单词库，发现词“a”会被类似“able”的单词匹配到，从而公式“$a+b$”被过滤掉，但它是 LaTex 数学公式，类似例子有很多。面对这种问题，可以采用两种方法，一种通过计算二者的哈希值进行匹配；另一种通过对原有的单词库做简单处理，即在匹配过程中在二者基础上添加一些特殊标记(如“@”)来达到匹配目的。图 4 显示了过滤流程。

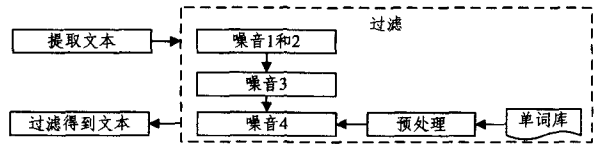

图 4 过滤工作流程

4 实验结果与分析

评价数学公式提取方法性能的主要指标为查准率和查全率。查准率是指检索到的相关文档与检索到的全部文档的比率；查全率是指检索到的相关文档与所有满足条件的文档数目的比例。

本文选取了 4 个网站进行实验分析。实验在操作系统为 Ubuntu 的 PC 上进行，手动设置爬虫爬取范围为 Wiki、Spaces、Aoshoo、Nist 网页，以网站首页作为第一次提取集合 U_{entry} 元素，以 LaTex 数学公式符号表作为第二次提取集合 T_{text} 元素，以过滤规则 1—3 作为集合 G_{filter} 元素(其中规则 3 中使用牛津词典单词过滤集)。实验基于 Nutch 完成爬取网页任务，通过获取不同基数页面计算各个网站中数学公式的查全率和查准率。图 5 中 Wiki、Spaces、Aoshoo、Nist 分别显示 4 个网站的查全率，图 6 中 Wiki、Spaces、Aoshoo、Nist 分别显示 4 个网站的查准率(Wiki、Spaces、Aoshoo、Nist 分别代表网站 http://www. wikipedia. org/、http://www. aoshoo. com/、http://spaces. ac. cn/、http://www. nist. gov/)。

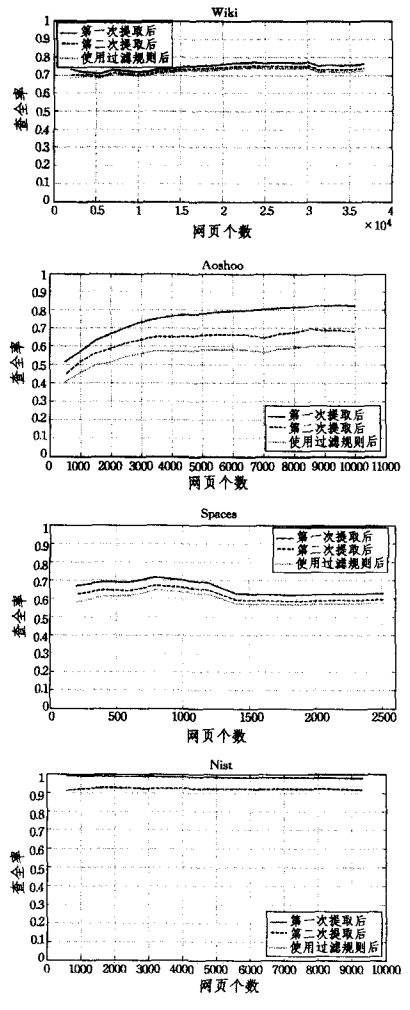

图 5 公式查全率

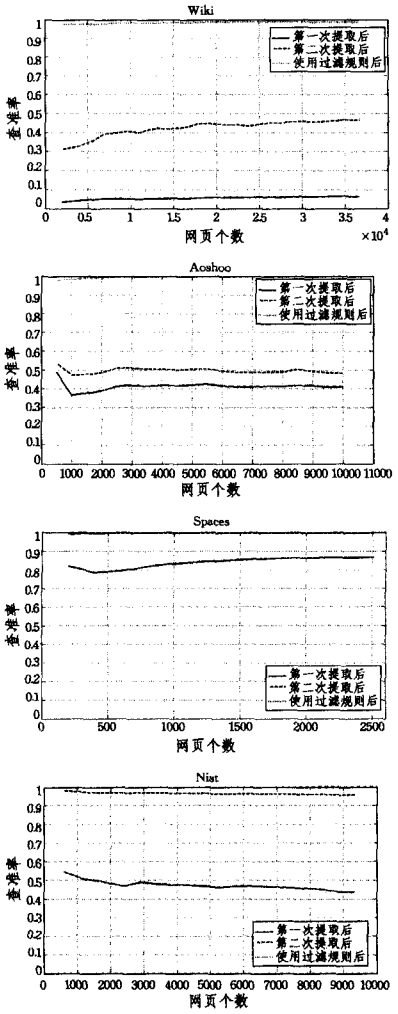

图 6　公式查准率

通过实验数据，计算出爬虫对数学公式的平均查全率和平均查准率(见表 9)。经过实验分析，影响因素为一些其他类型数学公式的存在，主要包括化学式子(如 $2H_2+O_2\rightarrow 2H_2O$)、简单二维公式(如 a^2b^3)、图片公式。查全率方面，被遗漏的公式基本为单变量(如 x)、隐含乘(如 xy)等；查准率方面，影响查准率的主要因素是人名(如 Kramp-Karrenbauer)、未登录词(如 E-maila)、编号或日期(如 N60-90)、隐式代码(如网页中显示的\sin)等。在表 9 中，网站 Spaces、Aoshoo 的查全率偏低，其原因主要是，在数学论坛中由于讨论需要，会出现大量图片公式被添加到版面中，同时还存在一部分 Infix(中缀)数学公式。

表 9　实验数据均值

项目	Wiki	Spaces	Aoshoo	Nist	所有网站均值
查全率	0.9536	0.5931	0.5559	0.9211	0.7559
查准率	0.9844	0.9969	0.9952	0.9968	0.9933

表 10　数据统计

项目	Wiki	Aoshoo	Spaces	Nist
每 1000 个网页中具有公式的网页个数	32.81	656.76	281.45	172.36
平均每个网页中包含的公式数	0.673	1.644	3.982	7.310
平均每个具有公式的网页中包含的公式数	22.08	2.503	14.15	42.41
平均每个具有公式的网页中包含 LaTex 公式数	16.49	2.150	10.06	41.57

另一方面，为了能够更清楚地反映出这些网站中数学公式特征，本文在实验中对公式和网页做了统计，通过统计计算，可以得出如表 10 所列的数据。

结束语　本文对网页中 LaTex 数学公式进行了研究，并提出基于 LaTex 的 Web 数学公式提取方法，通过获取网站网页，自动分析包含 LaTex 公式的特征，运用特征提取和过滤的方法获得 LaTex 数学公式。在此基础上，通过实验验证该方法。实验结果产生了少量噪音，并且在这些网站中还存在一些 Infix 中缀、图片数学公式，还有待改进。实验中对网页包含数学公式的数量特征进行了统计分析，便于以后进行与此类网站分析相关的工作。而一些专门的数学网站和论坛中公式也存在特征，这些也是下一步需要改进和完善的。

参 考 文 献

[1] 赵飞，周涛，张良，等. 维基百科研究综述[J]. 电子科技大学学报，2010，39(3)：322

[2] Krebs M, Ludwig M, Müller W. Learning Mathematics using a Wiki[J]. Procedia-Social and Behavioral Sciences, 2010, 2(2): 1469-1476

[3] Lamport L. LATEX: User's Guide & Reference Manual [M]. Addison-Wesley Publishing Company Inc, 1994

[4] 聂俊，陈天莹，符红光. 基于 Latex 的互联网数学公式搜索引擎[J]. 计算机应用，2010(12)：312-315

[5] 赵琳. 基于知识本体的数学公式语义检索方法与技术研究[D]. 天津：南开大学，2011

[6] Samarasinghe S H, Hui S C. Mathematical document retrieval for problem solving [C] // 2009 International Conference on Computer Engineering and Technology. 2009, 1: 583-587

[7] Misutka J, Galambos L. Mathematical extension of full text search engine indexer[C]//ICTTA. Damascus, April 2008: 1-6

[8] Shatnawi M, Youssef A. Equivalence detection using parse-tree normalization for math search [C] // 2nd International Conference on Digital Information Management, 2007 (ICDIM'07). IEEE, 2007, 2: 643-648

[9] Kohlhase M, Sucan I. A Search Engine for Mathematical Formulae[C] // 8th International Conference on Artificial Intelligence and Symbolic Computation (AISC 2006). 2006: 241-253

[10] Mišutka J, Galamboš L. System description: EgoMath2 as a tool for mathematical searching on wikipedia. org [M] // Intelligent Computer Mathematics. Springer Berlin Heidelberg, 2011: 307-309

[11] Miner R, Munavalli R. An approach to mathematical search through query formulation and data normalization [M] // Towards Mechanized Mathematical Assistants, MKM 2007. 2007: 342-355

[12] Libbrecht P, Melis E. Methods to access and retrieve mathematical content in activemath[C] // Proceedings of the Second International Conference on Mathematical Software. 2006: 331-342

[13] Kohlhase M. OMDoc-An Open Markup Format for Mathematical Documents [version 1. 2]: Foreword by Alan Bundy [M]. Springer, 2006

[14] Youssef A. Roles of math search in mathematics [C] // Proceedings of the 5th International Conference on Mathematical Knowledge Management. Springer Berlin Heidelberg, 2006: 2-16

[15] 刘志伟. 数学搜索引擎研究[D]. 兰州：兰州大学，2011

[16] Guo Wei, Su Wei, Lian Li, et al. MQL: A Mathematical Formula Query Language for Mathematical Search[C]// IEEE 14th International Conference on Computational Science and Engineering (CSE). IEEE, 2011: 245-250
[17] 景珂. 网络数学搜索中的数学查询语言与索引的研究[D]. 兰州: 兰州大学, 2009
[18] 崔林卫, 苏伟, 郭卫, 等. 基于 Nutch 的 Web 数学公式提取[J]. 广西师范大学学报: 自然科学版, 2011, 29(1)
[19] Srinivasan P, Menczer F, Pant G. A general evaluation framework for topical crawlers[J]. Information Retrieval, 2005, 8(3): 417-447
[20] Menczer F, Pant G, Srinivasan P. Topical web crawlers: Evaluating adaptive algorithms[J]. ACM Transactions on Internet Technology (TOIT), 2004, 4(4): 378-419
[21] 郑冬冬, 赵朋朋, 崔志明, 等. Deep Web 爬虫研究与设计[J]. 清华大学学报: 自然科学版, 2005, 45(9): 1896-1902
[22] 谭思亮. 聚焦爬行系统的设计—算法视角[D]. 成都: 中国科学院研究生院(成都计算机应用研究所), 2006
[23] Fuentes Sepúlveda J, Ferres L. Improving accessibility to mathematical formulas: the Wikipedia Math Accessor[J]. New Review of Hypermedia and Multimedia, 2012, 18(3): 183-204
[24] Abelson H, Dybvig R K, Haynes C T, et al. Revised report on the algorithmic language scheme[J]. ACM SIGPLAN Lisp Pointers, 1991, 4(3): 1-55

(上接第 112 页)

服务节点进行下载，因此密钥分量提取成功的量一直保持在较低水平，系统稳定性差。加入了信任模型的 DDTV 方案随着提取周期个数的增加，也就是提取次数的增加，节点之间的信任值计算逐渐准确，由此密钥分量提取成功的量也逐渐增加。

(3)密钥分量提取成功率

密钥分量提取成功率是整个系统的密钥分量提取成功次数在所有提取次数中所占的比例，它直观地反映了抵制恶意节点的抗攻击能力。为保证结果的准确性，每个方案的周期循环执行 20 次，采集每次循环得出的数据，并计算出平均值，结果如图 7 所示。

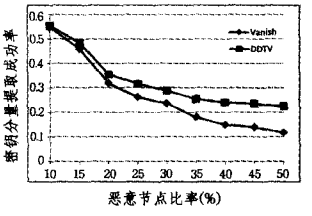

图 7 密钥分量提取成功率

由图 7 所示，当恶意节点比例特别低时，密钥分量提取成功率都接近于 58%。随着恶意节点比例的增加，在没有引入信任模型的情况下，节点缺乏抵制恶意节点的能力，表现为 Vanish 的仿真曲线迅速下降。相比 Vanish，DDTV 由于引入了信任模型，故能有效地抑制恶意节点，因而曲线的斜度更加平缓。

从以上的实验可以看出，DDTV 方案密钥的分发和重构虽然多了选择信任节点的时间，但时间开销在可接受的范围之内，而且同原来的 Vanish 方案相比，密钥分量提取成功率更高，使系统具有更高的可靠性。

结束语 本文设计了一种基于信任值的云存储数据确定性删除方案。在封装阶段，选择信任值较高的节点存放用户密钥分量。在解密阶段，从 DHT 网络中得到足够多的密钥分量后解密得到数据。在这两个阶段中均会根据节点的行为进行相应评价，综合这些评价就是该节点的信任值。引入信任值后，使用户密钥存放更安全，确保用户数据在授权时间内可用。实验结果表明，该方案是可行的，可以有效地抑制恶意节点，提高用户密钥分量提取成功率。接下来，将进一步完善我们的信任模型，使密钥分量提取成功率得到进一步提高。

参 考 文 献

[1] 武永卫, 黄小猛. 云存储[J]. 中国计算机学会通讯, 2009, 5(6): 44-52
[2] Kohno G T, Levy A, Levy H M. Vanish: Increasing data privacy with self-destructing data [C]// Proceedings of the 18th USENIX Seeurity Symposium. 2009
[3] Yue Feng-shun, Wang Guo-jun, Liu Qin. A secure self-destructing scheme for electronic data[C]// Proc of EUC2010. New York: IEEE Press, 2010: 651-658
[4] Zeng Ling-fang, Shi Zhan, Xu Sheng-jie, et al. Safevanish: An improved data self-destruction for protecting data privacy[C]// Proc of CloudCom 2010. New York: IEEE Press, 2010: 521-528
[5] 王丽娜, 任正伟, 余荣威. 一种适于云存储的数据确定性删除方法[J]. 电子学报, 2012(2): 266-273
[6] Perlman R. File System Design with Assured Delete [C]// SISW'05 Proceeding of the Third IEEE International Security in Storage Workshop. 2005: 83-88
[7] Tang Yang, Lee P P C, Lui J C S, et al. FADE: Secure overlay cloud storage with file assure ddeletion[C]// Proc of the SecureComm'10. New York: ACM Press, 2010. 380-397
[8] Stoica I, Morris R, Karger D, et al. Chord: A scalable peer-topeer lookup service for internet applications[C]// Proc of the SIGCOMM 2001. New York: ACM Press, 2001: 149-160
[9] Dabek F. A Distributed Hash Table [D]. Massachusetts: Massachusetts Institute of Technology, 2005
[10] Falkner J, Piatek M, John J, et al. Profiling a million user DHT [C]// Proc of the 7th ACM SIGCOMM Conference on Internet Measurement. New York: ACM Press, 2007: 129-134
[11] Rhea S, Godfrey B, Karp B, et al. OpenDHT: A public DHT service and its uses[C]// Proceedings of ACM SIGCOMM. 2005: 73-84
[12] Azureus[OL]. http://www.vuze.com/
[13] Shamir A. How to share a secret [J]. Communications of the ACM, 1979, 22(11): 612-613
[14] Dou W, Wang H M, Jia Y, et al. A recommendation-based Peer-to-Peer trust model[J]. Journal of Software, 2004, 15(4): 571-583
[15] The Stanford P2P sociology project[OL]. http://p2p.standford.edu/
[16] Vanish. [EB/OL]. http://vanish.cs.washington.edu/. 2011-07-29