# 学术论文抄袭检测方法研究综述

赵俊杰 ①

(安徽财经大学 成人教育学院,安徽 蚌埠 233061)

**摘 要**:从学术论文抄袭的现象和危害出发,分析归纳了学术论文抄袭的主要类型及形式。接着从论点抄袭、文本抄袭、算法与程序代码抄袭和图片与公式抄袭等几个方面,综述了近阶段所采用的主要抄袭检测方法。最后概述了解决论文抄袭检测问题的重要意义,并对如何防止学术论文抄袭提出建议。

**关键词**:抄袭检测;文本相似度;词频统计;数字指纹;图像匹配

**中图分类号**: G256.22 **文献标识码**: A **文章编号**: 1674-117X(2010)01-0157-03

## Detective Ways against Academic Plagiarism

ZHAO Junjie

(College of Adult Education, Anhui University of Finance and Economics, Bengbu Anhui 233061, China)

**Abstract:** The main types of academic plagiarism and the forms are analyzed and summarized starting from mentioning the phenomenon to the harm done. Recently adopted main detective ways against plagiarism are summed up from the aspects of theme, text, calculation, program code, picture and formula plagiarisms. Finally the important significance of plagiarism detection problem solving is mentioned and suggestions as to how to prevent academic plagiarism are put forward.

**Key words:** plagiarism detection; similarity of texts; word-frequency statistics; digital finger printing, picture matching

自 20世纪 90年代学术界提出反对学术腐败以来,被揭露出来的学术腐败事件最多的是学术造假,其中又以学术著作和论文的抄袭为最。抄袭行为不仅侵害了作者的权益,而且严重破坏了学术发展的生态环境,损害了学术共同体的尊严,还影响到我国科研水平和科技竞争力的提高,损害了国家和公众的利益。论文抄袭的类型主要分为两种情况:一是论点抄袭,这种情况是从质的角度来考虑,主要是看是否引用他人作品作为自己作品的主要部分或实质部分。例如抄袭他人的创意、主要的观点以及核心思想、分析论证方法等;二是内容抄袭,主要是从量,有时也结合质的角度来考虑,例如抄袭论文的文字、图片、表格、数据、模型与公式等具体内容。对于不同的学术论文抄袭形式其检测方法也必然不同,下面根据不同的论文抄袭形式介绍其常用的判定方法。

### 一、论点抄袭的判定方法

抄袭他人论文的核心思想、观点或创意及分析与论证方法,有可能不是整篇整段地抄袭,抄袭的数量也可能不超过 1/10,因此不能简单以抄袭的量加以衡量。这种抄袭一般难以直接判定,论点抄袭一般比较隐蔽,难以直接检测出来,可行的方法是先借助某种模式识别方法,在怀疑抄袭论文与相似论文之间进行比较,如果相似度超过一定的域值,则给出可能抄袭的初步判定。由于可能会出现误判,所以还需要进一步进行人工判定。

① **收稿日期**: 2009-07-05

**基金项目**:教育部社科研究基金青年项目"文本挖掘技术在论文抄袭判定中的应用研究"(07JC870006);安徽财经大学教研重点项目"学生论文抄袭的检测防范研究"(ACJYZD200914)

**作者简介**:赵俊杰(1973- ),男,安徽宿州人,安徽财经大学讲师,硕士,主要从事数据挖掘与情报检索研究。

晋耀红等人提出了基于语境框架的文本相似度计算。[1]语境框架是一个三维的语义描述,它把文本内容抽象成领域(静态范畴)、情景(动态描述)、背景(褒贬、参照等)三个侧面。在语境框架的基础上,计算文本的相似度。算法从概念层面入手,充分考虑了文本的领域和对象的语义角色对相似度的影响,重点针对文本中的歧义、多义、概念组合现象,以及语言中的褒贬倾向,实现文本间语义相似程度的量化。算法应用到文本过滤系统中,用以比较用户过滤要求和待过滤文本之间的相似度。

另外,还可以从论文的篇章结构相似度出发进行检测。例如金博等人提出了基于篇章结构相似度的复制检测算法。[2]此算法是在学术论文理解的基础上,针对学术论文的特有结构,对学术论文进行篇章结构分析。文章的篇章结构用数据库表可以表示为编号、全文特征值、发表时间、标题、作者、单位、摘要、关键词集合、中图分类号、段落集合、参考文献集合等。其中全文特征值是对某篇论文的全文进行Hash处理得到的整数值。接着再通过数字指纹和词频统计等方法计算出学术论文之间的相似度,从而找出抄袭的现象。不过此算法只针对书写格式规范的学术论文的抄袭现象。

## 二、内容抄袭检测方法

### (一)文本抄袭的检测方法

文本抄袭包括中文、英文和数据的抄袭,现在所采用的检测方法主要有两种:数字指纹法和词频统计法。数字指纹是通过某种选取策略对论文中的有些特征进行HASH计算而生成的,这些HASH函数可以为论文的每一特征语句或段落产生惟一整数值,通过比较指纹来计算论文间的相似程度。词频统计是采用空间模型(VSM)来表示,在模型中,论文空间被看做由一组独立词条所组成的向量空间,每个论文表示为一个特征向量进行相似度计算,常采用的计算公式包括点积法和余弦法等。[3]

在国外,自从1991年用于查询重复基金申请书的WordCheck软件应用以后,自然语言文本的抄袭检测技术有了较大的发展,出现了多个抄袭检测系统,如siff工具、复制检测系统SCAM、SE方法和Winnowing算法等。但由于英文论文和中文论文的语法和格式等有很大差别,所以检测方法也有很大区别,一般不能直接套用。

在国内,2001年西安交通大学宋擒豹等人提出了CDSDG系统,[4]这是为了解决数字商品非法复制和扩散问题而开发的一个基于注册的复制监测原型系统。此系统通过对数字正文的多层次、多粒度表示来构建基于统计的重叠度度量算法,取得了较好的效果。

金博、史彦军等提出的利用知网的知识结构及其知识描述语言的语法进行相似度计算的方法。[5]在词语的相似度计算中,利用知网义原树状结构及知网知识的网状知识特点,计算全面可靠;通过对实词集合的相似度计算来更有效地计算句子相似度;再将基于知网的语义理解相似度计算推广到段落及文本范围,使相似度计算更具实用价值。

霍华、冯博琴提出的基于压缩稀疏矩阵矢量相乘的文本相似度计算方法,[6]能够减少计算和存储空间的开销。该方法仅对非零元素存储和表示,然后用压缩稀疏矩阵矢量相乘的方法计算文本和查询的相似度,可通过给定相似度阈值来判定一个文本是否和查询相似。

余刚、裴仰军等提出的基于词汇语义计算的文本相似度研究。[7]采用了基于知网的词汇语义计算方法来计算两篇文章向量的相关性,并用最大匹配算法来获得这两篇文章的相似度,通过该计算过程达到揭示文本所涉及概念的目的。

化柏林开发了一个基于句子匹配的文章自写度测试系统。[8]句子是组成文章的重要单位,也是表明作者行文观点的最小单位,所以对于任意一篇稿子,利用句子匹配分析可以得到文章的自写度(自写不一定为创新,但相同可能为抄袭或引用)。对每一个句子都有匹配度,审核人员可以一目了然地看清有哪些句子是抄的,哪些句子是参考别人的,哪些句子是自己写的。

此外还有麻会东、刘国华等人提出了基于提取关键词的中文文档复制检测方法,[9]王涛,樊孝忠等人提出了基于复杂特征集的剽窃检测算法等,[10]都有一定的特色和检测效果。

笔者也提出了一种基于基于分类思想的论文抄袭判定系统(CBTPJS),[11]可以在分类结果的基础上进行比较精确的抄袭判定并输出抄袭段落中的具体抄袭内容。其主要思路是从分类出发,先进行全篇相似度计算,经过初步筛选,然后对筛选结果再进行精确比较,即进行段落相似度计算,最后如果判定是抄袭则输出具体抄袭的内容。

另外,中国知网推出的科技期刊学术不端文献监测系统、社科期刊学术不端文献监测系统和学位论文学术不端文献监测系统,从2009年也开始投入

使用,其主要采用的是数字指纹技术。

(二)算法与程序代码的抄袭判定方法

对于程序代码的抄袭,有的是直接复制或稍加改动,例如修改变量的名称,修改输入、输出语句的格式等。有的改动较大,例如抄袭者采用另一种程序设计语言进行实现而不做说明,其实算法是相同的,这属于算法的抄袭。算法的表示形式有很多种,包括程序流程图、N - S图、过程设计语言等,对于某种算法用另一种形式进行描述,或者用另一种语言进行实现,这实际是抄袭了他人的核心思想。

程序代码相似度自动度量技术的研究始于20世纪70年代,至今已比较成熟。目前的抄袭检测系统大部分使用了结构度量技术,即通过系统比较表示程序结构的字符串来检测抄袭,但表示程序结构的字符串不需要精确匹配。有的系统混合使用了结构度量技术和属性计数技术。比较有代表性的有Alex Aiken于1994年开发的MOSS系统,主要用于检测用C、C++、JAVA、PASCAL、Ada、ML、Lisp、Scheme等编写的源程序的相似性;Michael Wise于1996年开发的YAP3,不但可以检测源程序代码的抄袭,还可以检测自然语言文本间的相似性。

由于算法可能以不同形式表示,直接检测不易,因此可以考虑把算法通过某种工具,如ROSE等CASE工具,转换成统一的形式,例如伪码或PAD图等,然后再进一步比较伪码或图形。

(三)图片与公式抄袭的检测方法

图片抄袭是指对他人论文中的图形或图像直接插入到自己的论文中作为自己成果的一部分。图片不做处理,也可能稍加处理。抄袭的图片大多是重要的论据,或者是实验的结果等,这种抄袭从量上也可能不足1/10,但实际上也构成了抄袭。对于论文中的图片抄袭问题,一般借助图像匹配方法进行检测。图像匹配是指通过一定的匹配算法在两幅或多幅图像之间识别同名点。图像匹配主要包括以灰度为基础的匹配和以特征为基础的匹配。即使抄袭者对图片进行了少量修改,通过此方法也能检测出来。

论文中涉及到的模型、公式和定义等也可能被别人抄袭,这部分内容可能只占很少的篇幅,但这可能是论文的精华部分和亮点,整篇文章都是基于此模型的实验结果或者公式的演算结果进行分析和论证的,因此这也属于抄袭的一种形式。由于公式不同于普通文本,如果采用一般的文本抄袭检测方法可能判断不出公式是否被抄袭。因为抄袭者可能会更改变量名、调整表达式中常量、变量或函数等成份位置等,而且公式中还使用了很多专用数学符号。因此,对于公式的抄袭检测要采用特殊的方法。一种方法是把公式当作图片,采用前面提到的图像匹配技术,即使有所改动也能大致判断出来是否涉嫌抄袭;另一种方法是采用特定技术把两个公式中的数学符号、常量、变量、函数等分别抽取出来,然后对比,从使用的个数及顺序的相似程度上进行检测是否涉嫌抄袭。当然这两种方法都存在一定的缺陷,可能会误判,进一步人工判定还是必要的。

学术论文抄袭问题已经越来越被大家所关注,解决论文抄袭的检测问题不但对于保护知识产权、提高学术论文质量、净化学术领域、防止学术腐败都有很重要的意义,而且可以有效地防止一稿多投和减轻审稿人员的工作负担。抄袭者之所以去抄袭,一是利益驱动;二是抱有不会被发现的侥幸心理。因此,除了设法进一步提高论文抄袭检测系统的效果和效率外,还要加强科学道德教育和完善相关法律条款,从多方面入手,使得论文抄袭者无机可乘。

参考文献:

[1] 晋耀红. 基于语境框架的文本相似度计算[J]. 计算机工程与应用, 2004(16): 36 - 39.

[2] 金 博,史彦军,滕弘飞. 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(1): 125 - 130.

[3] 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(1): 50 - 57.

[4] 宋擒豹,沈钧毅. 数字商品非法复制和扩散的检测机制[J]. 计算机研究与发展, 2001(1): 121 - 125.

[5] 金 博,史彦军,滕弘飞. 基于语义理解的文本相似度算法[J]. 大连理工大学学报, 2005(2): 291 - 297.

[6] 霍 华,冯博琴. 基于压缩稀疏矩阵矢量相乘的文本相似度计算[J]. 小型微型计算机系统, 2005(6): 988 - 990.

[7] 余 刚,裴仰军,朱征宇,等. 基于词汇语义计算的文本相似度研究计算机工程与设计[J]. 2006(2): 241 - 244.

[8] 化柏林. 基于句子匹配的文章自写度测试[J]. 现代图书情报技术, 2007(11): 40 - 44.

[9] 麻会东,刘国华,李 旭,等. 基于提取关键词的中文文档复制检测研究[J]. 计算机工程与科学, 2007(10): 63 - 64, 88.

[10] 王 涛,樊孝忠,林培光,等. 基于复杂特征集的剽窃检测[J]. 北京理工大学学报, 2008(2): 129 - 133.

[11] 赵俊杰. 基于分类思想的论文抄袭判定系统的设计与实现[J]. 数字图书馆论坛, 2008(11): 73 - 75.

责任编辑:骆晓会