一种基于段落词频统计的论文抄袭判定算法

赵俊杰[1,2],胡学钢[1]
(1. 合肥工业大学,安徽 合肥 230009;
2. 安徽财经大学,安徽 蚌埠 233061)

摘 要:解决论文抄袭的判定问题不但可以减轻审稿人员的工作负担,而且对于提高学术论文质量、净化学术领域、防止学术腐败都有很重要的意义。从抄袭的定义和法律规定出发,在分析比较国内外主要的论文抄袭判定方法基础上,提出存在的问题和改进策略,然后给出一种基于段落词频统计的论文抄袭判定算法。此算法不但可以检测出抄袭者成段抄袭的情况,而且可以检测出段落中语句顺序改变、段落内容压缩和扩充的情况,若疑似抄袭还可以将抄袭论文和被抄袭论文的相似内容输出,方便用户进一步审查。

关键词:抄袭判定;词频统计;段落相似度;中文分词
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2009)04-0231-03

A Way to Judge Plagiarism in Academic Papers Based on Word-Frequency Statistics of Paragraphs

ZHAO Jun-jie[1,2], HU Xue-gang[1]
(1. Hefei University of Technology, Hefei 230009, China;
2. Anhui University of Finance & Economics, Bengbu 233061, China)

Abstract: *Solutions to judging plagiarism in the academic papers can not only lighten article editors' work load, but also have the important significance in improving academic papers' quality, purifying academic fields and preventing academic corruption. Based on the definition of plagiarism and the law, put forwards the existent problems and strategies for improvement by analyzing the main methods to deal with plagiarism judgment both at home and abroad, and produce a way to judge plagiarism according to word-frequency statistics of paragraphs. This way can help us not only inspect whether plagiarists copy the information paragraph by paragraph, but find out the following cases, such as the change in the order of sentences, compression and expansion of the content. Also output the similar content to make it convenient for users to check when the papers are suspected as copies.*

Key words: plagiarism judgment; word-frequency statistics; similarity between paragraphs; Chinese text segment

0 引 言

近些年来论文抄袭成为困扰学术界的严重问题之一,抄袭剽窃之风在今天的学术界愈演愈烈几乎是不争的事实。"抄袭是指将他人作品或者作品的片段窃为己有。"更准确地说,抄袭是指将他人作品或者作品的片段窃为己有并公开发表。

论文中可以适当地引用他人作品的部分内容,当然要指明出处。但如果引用不合理,就涉嫌抄袭。《著作权法实施条例》第二十七条第二款规定"所引用部分不能构成引用人作品的主要部分",我国文化部1985年曾对合理引用量作了规定。该规定指出,引用非诗词类作品不得超过2500字或被引用作品的十分之一;多次引用同一部长篇非诗词类作品,总字数不得超过1万字;引用诗词类作品不得超过40行或全诗的四分之一,但古体诗词例外;凡引用一人或多人的作品,所引用的总量不得超过本人创作作品总量的十分之一。但专题评论和古体诗词除外[1]。

目前,对于英文论文抄袭的判定比较成熟,主要采用数字指纹和字符串匹配等技术,而对于中文论文抄袭的判定还不太成熟,大多数算法存在识别率低、效率不高等问题。针对这种情况,提出一种基于段落词频统计的论文抄袭判定算法,提高了识别率,并对抄袭内容进行定位输出,方便用户查看。

收稿日期:2008-07-16
基金项目:教育部社科研究基金青年项目(07JC870006);安徽财经大学教研重点项目(ACJYZD200914)
作者简介:赵俊杰(1973-),男,安徽宿州人,讲师,研究方向为数据挖掘;胡学钢,教授,博士,研究方向为数据挖掘。

1 相关工作

1.1 国内外研究现状

学术论文抄袭的形式和手段多种多样,包括直接将他人论文全盘复制,只改动题目和署名;东拼西凑,抄袭多篇论文的部分段落和语句;抄袭论文的图、表与公式等。这里只讨论文字部分的抄袭判定。

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识别技术有了较大的发展,出现了多个抄袭识别系统。1994 年,Mander 开发了用于大规模文件系统中相似文件查询的 siff 工具[2]。siff 能够查询二进制和文本文件,率先使用数字指纹技术来计算文件相似度,为抄袭论文识别技术提供了新思路。1995 年,Shivakumar 等采用相关频率模型开发了复制检测系统 SCAM[3],SCAM 借鉴了信息检索技术中的向量空间模型,采用了改进的余弦法来计算文档相似度。同期,香港理工大学的 Si 和 Leong 等人建立的 CHECK 原型采用统计关键词的方法来度量文本相似性。CHECK 系统首次把文档结构信息引入到文本相似性度量中。2002 年,Hoad 和 Zobel 综合采用了词频统计和数字指纹方法来解决衍生文档的识别问题,通过对大量 XML 数据和 Linux 文件的测试以寻找较好的抄袭识别算法。另外,悉尼大学 Wise 开发了 YAP(yet another plague)1,YAP2,YAP3 系列工具[4]。YAP1 和 YAP2 是用于程序复制检测的工具,YAP3 利用程序复制检测的方法,既检测程序复制也检测文本复制。

在国内,2001 年,西安交通大学宋擒豹等人提出了 CDSDG(copying detection system of digital goods)系统[5],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。2007 年,金博等人还从论文的篇章结构相似度出发提出了基于篇章结构相似度的复制检测算法[6]。它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,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。但此算法只是针对书写格式规范的学术论文抄袭情况的判定。

1.2 中文分词

英文书写时,单词之间用空格隔开,词间界限泾渭分明;而中文是字的序列,词与词之间没有间隔标记,而词又是中文中最小的能够独立运行的语法单位,所以必须经过"分词"处理后,计算机才能进行下一步的分析,因此对中文的自动分词,是中文信息处理的基础和前提。

现有的分词方法主要有以下三类[7]。

1.2.1 基于字符串匹配的分词方法

这种方法又称为机械分词方法,它是按照一定的策略将待分析的汉字串与一个"充分大的"机器词典中的词条进行匹配,若在词典中找到某个字符串,则匹配成功(识别出一个词)。按照扫描方向的不同,串匹配分词方向的不同,串匹配方式又包括正向匹配、逆向匹配和双向匹配等。

1.2.2 基于理解的分词方法

这种分词方法是通过让计算机模拟人对句子的理解,达到识别词的效果。其基本思想就是在分词的同时进行句法、语义分析,利用句法信息和语义信息来处理歧义现象。它通常包括三个部分:分词子系统、句法语义子系统、总控部分。

1.2.3 基于统计的分词方法

在上下文中,相邻的字同时出现的次数越多,就越有可能构成一个词。可以对语料中相邻共现的各个字的组合的频度进行统计,计算它们的互现信息。互现信息体现了汉字之间结合关系的紧密程度。当紧密程度高于某一个阈值时,便可认为次字组可能构成一个次。这种方法时空开销较大,识别精度不高。所以实际应用的统计分词系统都是将串匹配和串统计结合起来使用。

现在常用的分词软件有中科院的 ICTCLAS 和天津海量科技公司的分词系统。其中中科院的汉语词法分析系统 ICTCLAS 包括中文分词、词性表注、命名实体识别、新词识别、同时支持用户词典等动能。文中所涉及到的算法就是采用中科院的 ICTCLAS 进行分词预处理的[8]。

2 基于段落词频统计的论文抄袭判定算法

2.1 基本思想

前面提到的复制检测或者抄袭判定系统大多只能处理全文抄袭或大部分抄袭等情况,而对于个别段落的抄袭,尤其是从多篇文章进行段落摘抄的情况,容易疏漏。另外,在基于段落相似度比较判定时,由于很多情况下抄袭者也不是整段一字不动的抄袭,有的是调整语序,有的摘抄段落的一部分,还有的是对某些段落进行合并、扩充或者压缩等,所以判定时很容易漏查或误查。

针对上述问题,提出的算法就是基于段落的词频统计和比较来判定是否存在抄袭。其中涉及到一些细节问题包括:

在统计词频时考虑到效率问题,去除了停用词等,

而且使用散列表提高检索速度。另外,由于有的段落很短,且对于判定抄袭结果影响不大,所以在比较时可以忽略。

由于论文判定抄袭的标准一般是非法引用他人文章的十分之一或者引用部分占到自己文章的十分之一都算为抄袭,所以在判定时必须把这两种情况都考虑在内。当然在具体实现时,考虑到偶然情况,全文相似度也可以设定稍大一些,例如 15%;且对于每一段比较时还要另外设定一个阈值,这个值要比全文相似度再稍大一些,例如 20%。

2.2　基于段落词频统计的论文抄袭判定算法

基于段落词频统计的论文抄袭判定具体算法:

步骤 1:对查询论文进行分词,然后将各词分别放入以段落为单位建立的若干数组和 Hashtable 中(使用 Hashtable 主要是提高检索速度,其中键代表词,值代表词频),在放入过程中去除无关紧要的词,如叹词、虚词、停用词等。

步骤 2:.设定一个阈值(0.2),将待查论文的每个数组与查询论文的每个数组进行比较,若有重复则登记匹配的个数。每轮循环比较结束,若匹配词频总数超过阈值则疑似段落抄袭,记录相似度最大的那个,否则认为没有抄袭。直至所有段落比较完毕。

步骤 3:统计所有疑似抄袭段落匹配的词频数总和占抄袭论文与待查论文总词频数的比例,若任何一个超过指定的阈值(0.15)则认为论文存在抄袭。

步骤 4:若存在抄袭则对可能抄袭的段落进行逐句比较,输出相似的语句(包括原文章和抄袭的文章),目的是方便用户查询对比;若无抄袭则输出:"无抄袭现象"的结论。

3　实验结果分析

笔者设计了一篇抄袭的文章,其中包括整段抄袭(两篇)、调整某段落语句顺序(一篇)、摘抄段落的一部分(两篇)和对某些段落进行扩充和压缩(两篇)等多种情况,目的是检验系统对各种抄袭情况判定的效果。另外通过关键字在论文库中 1000 多篇论文的摘要或全文中进行检索,选取了 132 篇同类文章。针对不同阈值进行了三组实验,表 1 是在不同阈值下其查准率、查全率和 F1 值的情况。

表 1　不同阈值下的查准率、查全率和 F1 值

	阈值	查准率	查全率	F1 值
实验 1	0.1	0.83	1	0.91
实验 2	0.15	0.86	1	0.92
实验 3	0.2	0.86	0.86	0.86

图 1 是其中抄袭论文和某一篇待查论文的分析比较情况:

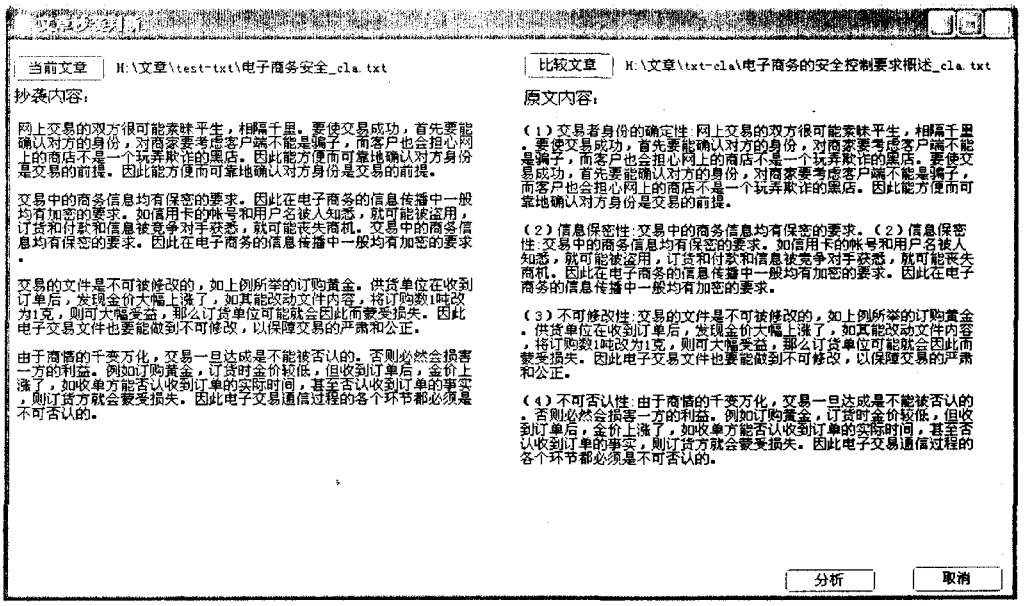

图 1　论文抄袭判定结果

要提高查准率和查全率,对于段落相似度的域值设定很重要,通过实验我们发现当阈值过大时,容易漏查;而当域值过小时,容易误查,当然这两种情况下查准率都不高。一般法律规定是引用超过 10% 则认为是不合理引用或者抄袭,由于按词频统计有一定误差,所以可以放大到 15% 左右。

另外,笔者是以自然段为单位进行比较的,但有些段落内容非常少,如小标题等。由于词语太少,矩阵太稀疏,容易造成误判,所以在扫描论文时,词语数太少的段落是忽略不记的。由于这部分内容所占比例很小,所以最后统计时对判定结果影响不大。

4　结束语

基于段落词频统计的论文抄袭判定方法的特点是:以段落为单位可以防止抄袭者将论文的段落顺序打乱,尤其是从多篇文章进行段落摘抄的情况,只要总

(下转第 238 页)

题之一，在研究和分析了当前流行的多种静态和动态的均衡算法[8]之后，提出了一种基于数据挖掘思想的负载均衡方法。该方法利用对访问历史记录的数据挖掘结果，使服务器有能力预测即将到来的访问并做出预处理。文中对数据挖掘的方法和负载均衡的策略分别进行了描述和分析，并在此基础上进行测试和实验，取得了较为理想的效果。

当然，多个城市之间的数据访问和同步将会得到实施，在具体的操作过程中还需要对目前的策略进行补充和修改。

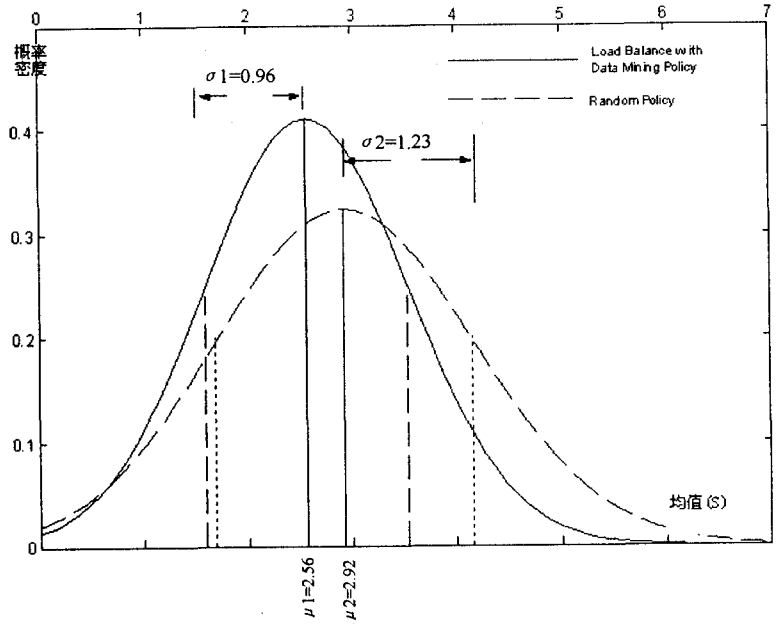

图 4 基于本地的多线程并发访问比较

参考文献：

[1] 李丙锋，祝永志，魏榕晖. 异构 Beowulf 系统负载均衡技术的研究与实现.[J] 计算机技术与发展，2008，18(7)：60－62.

[2] 陈国良. 并行算法实践[M]. 北京：高等教育出版社，2004.

[3] Group W. A high performance portable implementation of the message passing interface[J]. Journal of Parallel Computing, 1996, 22 (6): 789－828.

[4] 奎 因. 并行程序设计 C、MPI 与 OpenMP [M]. 北京：清华大学出版社，2005.

[5] 王 刚，王本年. 基于 FNN 与 GA 相融合的数据挖掘方法研究[J]. 计算机技术与发展，2008，18(2)：119－125.

[6] Ishigami H, Fukuda T, Shibata T, et al. Structure optimization of Fuzzy Neural Network by Genetic Algorithm[J]. Fuzzy Sets and Systems, 1995, 71(3): 257－264.

[7] Takagi T, Sugeno M. Fuzzy identification of systems and its applications to modeling and control[J]. IEEE Transactions on Systems, Man and Cybernetics, 1985, 15(1): 116－132.

[8] 李 旸，高正光，李启炎. 基于神经网络与遗传算法的数据挖掘体系结构[J]. 计算机工程，2004，30(6)：155－156.

(上接第 233 页)

量超过一定阈值，也能够检测出；同时，基于段落词频统计可以检测到将段落语句次序打乱重新组合和对段落进行扩充或压缩的情况。由于最后做出的结论有一定误差，还需要人工进一步判定，所以分别在两个窗口输出抄袭论文和待查论文疑似抄袭的段落，使用户不必再从整篇论文中查找、定位抄袭内容，从而方便用户进一步查看与判定。

对于中文学术论文的抄袭识别问题，相对于英文论文抄袭识别来说，由于需要额外考虑汉语的词切分、词法及语法特点，因此，难度较大。对于文中提出的算法和文中提到的其他算法都存在一定的误判，而且效率还需要进一步提高。另外，对于论点抄袭更是难以判定，一般需要借助于人工智能进行语意分析和判断。因此，对于论文抄袭问题还需要进一步研究，还不能完全替代人工判定。

致谢：文中的中文分词程序采用了中国科学院计算所软件研究室的汉语词法分析系统 ICTCLAS。

参考文献：

[1] 金 帛. 剽窃、抄袭他人的作品是一种严重的侵权行为——兼谈对剽窃、抄袭行为的认定[J]. 晋图学刊，2001 (4)：77－78.

[2] 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.

[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，2005，45(1)：50－57.

[4] 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)：1753－1760.

[5] 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，2001，38(1)：121－125.

[6] 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，2007，47(1)：125－130.

[7] 王小捷，常宝宝. 自然语言处理技术基础[M]. 北京：北京邮电大学出版社，2002.

[8] Zhang Hua ping. HHMM－based Chinese lexicalanalyzer ICTCLAS[C]// Second SIGHAN Workshop Affiliated with 41st ACL. Sapporo: [s. n.], 2003: 63－70.