

学术文献相似性检测比对资源应用分析与建设策略探究

——基于万方检测系统的实证分析

蒋勇青, 刘芳, 于洋

(中国科学技术信息研究所, 北京 100038)

摘要: 比对资源库建设是学术文献相似性检测系统建设中的重要环节。本文通过重点研究实际检测统计数据并分析引用或非典型性引用甚至可能的抄袭文献的特征, 为检测系统比对资源建设提供可行策略与发展思路。从2016年向万方数据论文相似性检测系统送检的学术文献中随机抽样650万篇, 对抽样送检文献检测结果中共计845 889篇被引用或非典型性引用甚至可能的抄袭文献从文献类型、被相似频率、发表时间等维度进行数据分析。调查发现, 比对数据中的学位论文引用或非典型性引用甚至可能的抄袭次数最多, 发表或公开时间在2—10年的比对数据引用或非典型性引用甚至可能的抄袭次数最多。此次研究为将来学术文献相似性检测系统的持续建设与优化提供全新的思路与方法借鉴。

关键词: 文献检测; 比对资源建设; 学术不端

中图分类号: G25

DOI: 10.3772/j.issn.1673-2286.2017.12.006

近年来学术不端行为层出不穷, 最初的学术不端典型个体事件逐步蔓延为社会性“群体事件”。学术不端的防治已成为国内外学术界与教育领域共同关注的焦点, 大众呼吁在保障自由学术环境的前提下, 全面构建学术规范体系以维持良好的学术秩序^[1]。为有效贯彻落实在学术及科研领域的学术不端审查工作, 利用数字化与智能化等全新技术手段推进学术文献相似性检测工具的全面建设与持续发展, 对提高我国整体学术与科技发展水平, 努力实现科技强国的宏伟目标有重大且深远的意义^[2]。

1 学术文献相似性检测相关研究进展及意义

近年来, 政府相关管理机构为有效防治学术不端行为, 陆续颁布相关政策及办法。如科技部颁布的《国家科技计划实施中科研不端行为处理办法(试行)》^[3]与教育部颁布的《教育部关于严肃处理高等学校学术

不端行为的通知》^[4], 旨在建立常态化的学术监督机制, 营造风清气正、求真务实的学术氛围。然而, 数字化与智能化等信息技术的迅猛发展为学术不端行为提供了更多方便, 也使得学术监督机制中学术不端行为的甄别判定环节增加难度。在此背景下, 学术文献相似性检测系统应运而生, 成为遏制学术不端行为的实用和有效工具。

国内学界针对学术文献相似性检测的发展与应用展开了多方面的探索研究, 综合来看, 主要包括四个方向: (1) 学术文献相似性检测工具的使用意义, 如胡政平的《学术不端文献检测与期刊质量的提升》^[5]等; (2) 学术文献相似性检测的场景性应用与对比分析, 如张旻浩等的《国内外学术不端文献检测系统平台的比较研究》^[2]、荣曼等的《学术不端文献检测系统在医学论文编辑出版过程中的合理应用》^[6]等; (3) 学术文献相似性检测结果的分析与使用, 如孔琪颖等的《正确看待“科技期刊学术不端文献检测系统”检测结果》^[7]等; (4) 学术文献相似性检测算法与技术研究, 如王晓

笛等的《学术文献抄袭检测研究进展》^[8]与《基于语义角色标注的文献相似度检测研究》^[9]。

综上所述,目前相关研究主要集中于如何有效甄别与判定学术不端行为以及对检测技术的深入研究方面。调研发现,检测工具使用者有效尝试了如何科学合理地使用检测结果数据,改变检测工具发展初期仅采用使用者相似比数值的单一指标情况。然而,现有研究少见对引用或非典型性引用甚至可能的抄袭的来源文献分布研究,对于检测系统建设中的重要环节——比对资源建设更是鲜有陈述。因此,本文通过实证分析,将会得到:(1)比对资源不同会给检测结果带来较大变化;(2)引用或非典型性引用甚至可能的抄袭的来源文献能很大程度反映其需求规律与资源利用情况。本文将重点对检测工具比对资源建设策略与规划进行拓展与完善,助力学术文献相似性检测市场专业化进程的持续发展。

2 研究对象与数据来源

万方数据论文相似性检测系统是国内较早服务于高校及各类学术机构的文献检测工具。2010年将自主开发的“基于滑动窗口的低频匹配特征算法”应用于全新的万方数据论文相似性检测系统。该系统历经7年的稳健发展,在算法方面进行基于语义实现的查全查准的优化升级,在数据方面持续汇聚海量学术期刊、硕博学位论文、国家专利等全文比对数据,万方检测服务始终坚持客观、公正、精准、全面的原则,秉承关注学术环境、维护学术诚信、恪守学术道德、尊重学术成果的理念,提供多版本、多维度的检测报告,检测结果精准翔实,为科研管理机构、教育领域、出版发行领域、学术个体等用户提供各类学术科研成果的相似性检测服务。

本次研究的对象是来源于万方数据论文相似性检测工具的2016年部分抽样检测结果中列出的相似文献及其相关元数据信息,对检测过程中同一篇文献的多次检测结果进行合并,保障数据分析结果尽可能接近实际情况,使研究结果具有更高的可借鉴性。

3 研究设计

3.1 数据对象选择

在万方论文相似性检测系统2016年全部送检学

术文献中随机抽样650万篇,在检测结果中统计得出被相似文献845 889篇,并从文献类型、被相似频率、发表时间等维度进行分析与总结。

3.2 研究问题设计

本文围绕三个问题展开讨论:(1)分析2016年论文相似性检测数据的文献相似总体情况;(2)基于相似频次的文献分布情况,如相似文献发表时间、类型等;(3)基于论文检测统计结果对未来学术论文相似性检测比对数据选择的借鉴意义。

3.3 分析指标确定

针对万方数据论文相似性检测工具2016年送检论文的部分抽样检测结果数据,本文以相似文献为单元进行多维度数据分析,包括相似文献的类型(期刊文献、学位论文等)、相似文献发表或公开时间、相似文献的被相似次数以及相似文献是否被送检文献列为参考文献等。以下为本次研究中涉及的研究概念解释。

(1)相似文献:送检论文被检测出的相似文本内容的来源学术文献;

(2)被相似次数:全部送检论文检测结果中某篇相似文献出现的总次数(单篇送检论文中多次检测出同一篇相似文献,按出现一次计);

(3)单篇平均被相似次数:按不同维度划分相似文献后,某类相似文献的被相似次数总数除以相似文献数量所得的平均值。

4 数据统计与分析

4.1 基于相似次数的相似文献总体情况

本文随机抽样选择的送检文献共有845 889篇相似文献。如表1所示,被相似次数为0—9次的文献有806 949篇,占比95.40%。本文将相似文献被相似次数在10次以内的行为定义为非典型性引用和可能的抄袭,不列入本文研究重点。被相似次数大于等于10次的相似文献共38 940篇,按被相似次数呈线性分布且逐渐递减的趋势,经初步分析该分布规律较合理,可针对此次的研究数据作进一步分析。

表 1 相似次数统计表

被相似次数	文献数量/篇	占 比	被相似次数	文献数量/篇	占 比
0—9	806 949	95.396 6%	160—169	21	0.002 5%
10—19	26 256	3.104 0%	170—179	11	0.001 3%
20—29	6 319	0.747 0%	180—189	13	0.001 5%
30—39	2 544	0.300 7%	190—199	6	0.000 7%
40—49	1 256	0.148 5%	200—209	5	0.000 6%
50—59	675	0.079 8%	210—219	4	0.000 5%
60—69	525	0.062 1%	220—229	3	0.000 4%
70—79	316	0.037 4%	230—239	3	0.000 4%
80—89	252	0.029 8%	240—249	4	0.000 5%
90—99	183	0.021 6%	250—259	39	0.004 6%
100—109	148	0.017 5%	260—269	11	0.001 3%
110—119	93	0.011 0%	270—279	0	0.000 0%
120—129	73	0.008 6%	280—289	2	0.000 2%
130—139	49	0.005 8%	290—299	1	0.000 1%
140—149	36	0.004 3%	≥300	30	0.003 5%
150—159	62	0.007 3%	-	-	-

4.2 相似文献的时间范围分布

相似文献的时间取值以年份为基本单位,其中期刊文献的时间值为其发表年份,学位论文的时间值为学位授予年份,会议论文的时间值为会议举办年份。

相似文献的时间分布为1992—2016年,其中无1993年及1994年的文献,同时,2016年比对数据文献的更新具有明显的延时性。据统计,相似文献的数量按发表年份呈线性分布。即在全部检测出的相似文献中,文献发表或公开的年份越接近当前日期,相似文献的数量越多,表明作者在选择可能引用或可能抄袭的文献时倾向于新发表或公开的论文成果。

相似文献数量按发表时间逐年增加,2014年相似文献的数量达到峰值。由于从期刊正式发表或学位授予,到其他学者可从平台获取学术文献并引用存在一定时间差,2016年在万方数据论文相似性检测平台中送检的论文引用或非典型性引用甚至可能抄袭2014年发表或公开的学术文献最多,共有相似文献5 272篇。

根据相似文献发表或公开时间分布情况,对发表或公开时间在2008—2014年的相似文献作进一步的类型分析,如表2所示。2008—2014年相似文献中学位论文占比为77%—87%,呈逐年下降趋势;期刊论文占比为11%—22%,呈逐年上升趋势;相似文献中会议论文数量较少,并未体现出明显的分布趋势。

论文作者引用或非典型性引用甚至可能的抄袭期刊论文时,对文献新颖性要求更高,更易选择最新发表的期刊论文。引用或非典型性引用甚至可能的抄袭学位论文时,对新颖性要求相对较低,可能选择学位授予时间为8—10年的学位论文。原因可能有两点。(1)学位论文自学位授予日起至被收录到学术文献数据平台供其他学者获取所经历的时间较期刊文献更长。相较期刊论文的纸质刊发、收录及在网络平台上线的公开流程,学位论文的网络公开过程效率较低,执行流程不够成熟。(2)学位论文较期刊论文的研究更具系统性。其中,国内外研究背景、课题绪论等内容更易被引用或非典型性引用甚至可能的抄袭,且网络开放的学位论文主要为国内高校优秀硕博士毕业论文,论文学术价值整体较高。

表 2 2008—2014年发表或公开的相似文献类型分析

年 份	会 议	学 位	期 刊	相似文献数量总计
2008	20	2 687	367	3 074
2009	19	3 420	484	3 923
2010	31	3 341	618	3 990
2011	24	3 395	657	4 076
2012	17	3 629	730	4 376
2013	26	4 017	1 014	5 057
2014	18	4 093	1 161	5 272

4.3 被相似文献的类型分布

在学术论文写作过程中,可引用多种类型的学术资源,包括学术期刊文献、硕博学位论文、会议论文、科技报告文献、标准文献、专利文献、成果文献,以及部分网络资源。而学术价值高,且被学者普遍引用的主要为学术期刊文献、学位论文及会议论文三大类。

全部相似文献中,期刊论文共447 416篇,占比52.89%;学位论文共382 514篇,占比45.22%;会议论文共15 959篇,占比1.89%。被相似次数大于等于10次的相似文献中,期刊论文共7 307篇,占比18.76%;学位论文共31 407篇,占比80.66%;会议论文共226篇,占比0.58%。

基于上述结论,本文对相似文献类型分布进行更深入的数据统计,统计显示全部相似文献中期刊文献的数量为7 307篇,贡献被相似次数共139 040次,平均单篇期刊文献被相似19.02次;学位论文的数量为31 407篇,贡献被相似次数共730 034次,平均单篇学位论文被相似23.24次;会议论文的数量为226篇,贡献被相似次数共4 224次,平均单篇会议论文被相似18.69次。

期刊文献与会议论文的平均单篇被相似次数较接近,而学位论文的平均单篇被相似次数达23.24次,明显高于期刊文献与会议论文,再次证明学位论文是文献检测中质量高且被引用或非典型性引用甚至可能的抄袭最为频繁的一种文献类型,对学术文献相似性检测结果具有重要影响。

对三类主要相似文献的被相似次数进行标准差分析发现,会议论文被相似次数波动较小,而期刊文献与学位论文被相似次数波动较大。相似文献中期刊文献与学位论文的被相似情况可能存在较明显的最大值与最小值群分布差距,需要通过更多维度对上述两种类型的文献进行分析,如学科领域分布、文献来源机构分布、核心期刊与非核心期刊分布、硕博学位论文分布等。

因在被相似次数大于等于10次的相似文献中,学位论文占比超过80%,本文重点对相似文献中的学位论文情况作进一步深入剖析。相似文献来源单位总计547家,相似文献在10篇以下的共286家,贡献被相似次数共853次;相似文献大于等于10篇的共261家,贡献被相似次数709 726次。

按相似文献单篇被相似次数降序统计发现,中南财经政法大学共计相似文献28篇,合计被相似4 166次,单篇学位论文被相似频率约149次;西南财经大学共计相似文献86篇,合计被相似次数11 236次,单篇学位论文平均被相似次数约131次;哈尔滨理工大学共计相似文献82篇,合计被相似次数7 158次,单篇学位论文平均被相似次数87次。

对各学位授予单位的相似文献数量及相似文献合计被相似次数进行统计。如图1所示,按相似文献数量降序排列结果发现,电子科技大学、厦门大学、重庆大学的学位论文合计被相似次数与相似文献数量均分列前三位,分别有1 170篇文献被相似29 596次、860篇文献被相似23 435次、921篇文献被相似22 956次。

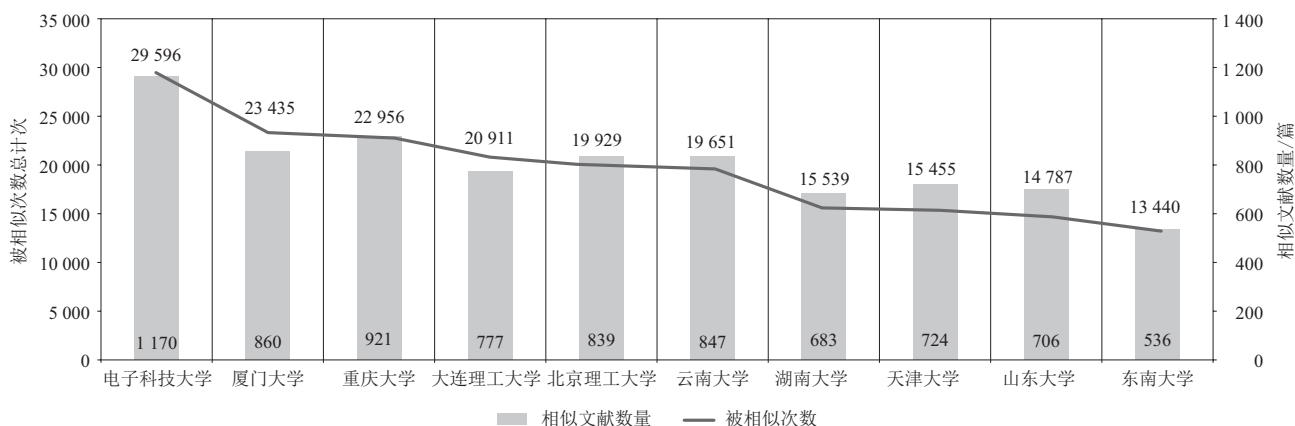


图1 基于引用或非典型性引用甚至可能的抄袭次数的学位论文相似文献来源统计

选取相似文献数量前20位的学位授予单位进行分析(见表3),发现这20家学位授予单位中,综合性大学占比较高,也不乏一定数量的理工类专业院校。同时,前20位的学位授予单位中,有11家为“双一流”建设

高校,其中前10位中有9家为“双一流”建设高校,可见“双一流”建设高校产出的学术成果在国内学术环境中对学术研究有重要影响。

表 3 学位论文相似文献来源统计

学位授予单位	相似文献数量/篇	被相似次数总计次	单篇学位论文平均被相似次数	单篇学位论文平均被相似次数标准差
电子科技大学	1 170	29 596	25.295 73	33.09
厦门大学	860	23 435	27.250 00	23.78
重庆大学	921	22 956	24.925 08	25.42
大连理工大学	777	20 911	26.912 48	26.59
北京邮电大学	839	19 929	23.753 28	20.68
云南大学	847	19 651	23.200 71	19.84
湖南大学	683	15 539	22.751 10	16.79
天津大学	724	15 455	21.346 69	19.63
山东大学	706	14 787	20.944 76	19.09
东南大学	536	13 440	25.074 63	30.11
东北大学	580	11 551	19.915 52	14.45
西南财经大学	86	11 236	130.651 2	366.32
北京交通大学	528	11 002	20.837 12	20.87
西安电子科技大学	473	9 874	20.875 26	16.18
苏州大学	416	9 341	22.454 33	21.20
北京工业大学	409	9 339	22.833 74	19.00
上海交通大学	444	9 034	20.346 85	21.78
南开大学	418	9 010	21.555 02	18.62
西南交通大学	338	8 456	25.017 75	30.77
东华大学	382	8 362	21.890 05	19.68

5 启示与建议

本次研究发现, 比对资源类型与发表时间是学术文献检测的重要影响因素, 针对资源类型及发表时间进行比对资源建设策略的制定对学术文献相似性检测系统的专业化与精准化发展具有积极作用与实际意义。

(1) 重点关注高新颖性比对资源建设。研究发现, 新颖性较高的学术文献资源在论文检测中具有更高贡献度, 然而最新发表的学术文献并非贡献度最高, 研究显示学术论作者虽更易选择新颖性高的文献, 但也体现出较明显的“延迟引用现象”。

董建军对期刊文献引用分布规律的研究发现国内期刊文献的引用次数达到高峰所需时间较长^[10], 一般在2—6年呈现逐步上升的趋势, 此现象与本研究通过论文检测相似文献统计的时间分布情况十分吻合。同时, 此规律也符合王名扬等提出的在部分学科中明显存在“延迟引用现象”的论证^[11]。因此, 可初步认为, 发表或公开时间2—10年的学术文献具有较明显的被引用或抄袭可能, 是学术不端行为检测工具重点需要关注的比对数据。

因此, 作为学术不端行为预防的有力武器, 在选择

论文相似性检测工具进行检测时, 不仅要着眼于两年以内的新论文或学术成果, 还要重点检查发表或公开时间在2—10年的学术文献。研究结果还反映出论文检测工具的数据建设不需要过多投入在旧文献上, 否则既会给论文检测带来不必要的干扰, 也会因大量数据的不断累积导致论文检测工具性能的持续下降。

(2) 全面丰富论文检测比对数据, 确保高质量文献的稳定供应。研究发现论文作者在不同类型学术文献中, 更倾向于引用或非典型性引用甚至可能的抄袭硕士学位论文。同时, 论文作者引用或非典型性引用甚至可能的抄袭的相似文献类型日趋丰富。随着检测比对数据的不断扩容, 论文检测工具可检测出的学术不端行为范围也随之扩大。除本文中重点分析的期刊文献、学位论文及会议论文外, 论文作者也出现了引用或可能的抄袭网络文献、专利文献等其他文献类型的情况。

由此可见, 在选择论文相似性检测工具的比对数据类型时, 可以遵循“全面建设、重点保障”的原则。一方面, 不断丰富比对数据类型, 全方位预防学术不端行为; 另一方面, 重点收录引用或非典型性引用甚至可能的抄袭出现频率较高的文献类型, 对其进行重灾区重点防治。其中, 以学位论文为重点关注对象, 通过对不

同学位层次及不同类型学位授予单位的差异性分析,进行合理的学位论文比对数据建设。建议加强硕博学位论文收集范围,重点收录“双一流”建设高校及“双一流”建设学科的学位论文成果。

(3) 基于比对资源特征调优文献相似性检测算法路线。本文的结果不仅为未来学术文献相似性检测系统的比对数据建设提供重要决策依据,也为相似性检测算法的专业化与精准化发展提供参考。数据显示学位论文具有最多的被引用或非典型性引用甚至可能的抄袭次数,对学术文献相似性检测的结果有重要影响。相较其他类型文献,学位论文篇幅较长,结构清晰,通过此次研究可初步推断,学位论文的摘要、绪论、研究方法设计等部分在实际检测结果中的贡献值也存在差异,同质化检测处理将无法进一步提升检测结果精度。针对学位论文被引用或非典型性引用甚至可能的抄袭情况相对较多的问题,后续将专门针对学位论文进行结构化分析,采用多层次多策略的方法,建立文本结构化识别模型与特征向量计算模型,为不同文本部分选择不同的相似性计算算法与条件权值,以有效提升检测结果精准度。通过上述检测数据能进一步分析论文作者在引用或非典型性引用甚至可能的抄袭学位论文时的内容偏好,为学术不端行为预防与处理提供更多决策依据。另外,相似文献的时间属性也可以考虑被赋予一定权值纳入检测结果,减少误识的相似结果,提高算法的精准度。

基于2016年在万方论文相似性检测系统中送检的学术文献检测部分抽样结果,本次研究主要针对论文检测比对资源的资源类型及资源发表时间两项影响因素进行分析与探究,提出在上述因素影响下论文检测比对资源建设策略应重点关注的方向。未来研究还将

融入对相似文献的学科分布、来源分布的深入分析统计,并进一步扩大样本容量。

参考文献

- [1] 万苏春.学术不端行为及其治理研究——以2010—2016年国内学界文献研究为综述对象[J].宜春学院学报,2017,39(4):111-116.
- [2] 张旻浩,高国龙,钱俊龙.国内外学术不端文献检测系统平台的比较研究[J].中国科技期刊研究,2011,22(4):514-521.
- [3] 国家科技部.第11号令《国家科技计划实施中科研不端行为处理办法(试行)》.2007年1月1日实行[EB/OL].(2007-08-13)[2017-04-26].<http://scitech.people.com.cn/GB/25509/39796/41759/6108064.html>.
- [4] 中央政府门户网站.教育部关于严肃处理高等学校学术不端行为的通知[EB/OL].(2009-03-21)[2017-07-01].http://www.gov.cn/gzdt/2009-03/21/content_1264527.htm.
- [5] 胡政平.学术不端文献检测与期刊质量的提升[J].出版发行研究,2012(11):87-89.
- [6] 荣曼,董海原.学术不端文献检测系统在医学论文编辑出版过程中的合理应用[J].中国科技期刊研究,2017,28(3):215-219.
- [7] 孔琪颖,蔡斐,张利平等.正确看待“科技期刊学术不端文献检测系统”检测结果[J].编辑学报,2009,21(6):544-546.
- [8] 王晓笛,王效岳,白如江.学术文献抄袭检测研究进展[J].图书情报工作,2013,57(8):141-148.
- [9] 王晓笛,祝娜,白如江,等.基于语义角色标注的文献相似度检测研究[J].图书情报工作,2014,58(12):130-135.
- [10] 董建军.科技期刊文献引用分布规律的探讨[J].中国科技期刊研究,2013,24(4):688-693.
- [11] 王名扬,于光,于达仁.文献的“延迟引用现象”分析[J].情报杂志,2011,30(5):60-62,84.

作者简介

蒋勇青,男,1963年生,高级工程师,研究方向:信息资源建设与服务、知识服务、科研管理、企业管理, E-mail: jiang@wanfangdata.com.cn.
刘芳,女,1992年生,硕士,研究方向:知识服务、项目管理。
于洋,女,1983年生,硕士,工程师,研究方向:数字编辑。

The Exploration of Resource Construction Strategy in Academic Literature Similarity Detection System: Empirical Analysis Based on Wanfang Detection

JIANG YongQing, LIU Fang, YU Yang
(Institute of Science and Technology Information of China, Beijing 100038, China)

Abstract: The data base construction is an important part of the academic literature similarity detection system development. This paper gathered the information of 845 889 similarity articles, acquired from 6.5 million random sampling journal papers, dissertations and other type literatures submitted in the Wanfangdata's Similarity Detection system in 2016. The paper summarized and evaluated the statistical data in the aspects of literature types, plagiarism numbers and published time. Results showed that, the dissertations had the highest cited or plagiarized number among all the literature types, at the same time, the papers published in 2 to 10 years had the most cited or plagiarized times. The research provided new ideas and methods for the further development of the academic literature similarity detection system.

Keywords: Literature Detection; Comparison Resource Construction; Academic Misconduct

(收稿日期: 2017-12-13)