

• 理论探索 •

基于粗糙集的本科毕业论文质量控制影响因素分析

张 昕

(天津师范大学管理学院, 天津 300387)

〔摘要〕目前高校层出不穷的学术造假现象使得学术道德下降和科学精神匮乏, 本文以高校经管类本科毕业论文作为研究对象, 针对论文质量管理中的事前控制阶段, 以论文抄袭检测系统的检测结果作为基础数据, 运用统计分析和粗糙集进行数据挖掘, 寻找论文质量事前控制的外部 and 内部影响因素, 并提出了如何通过控制这些因素来提高经管类本科毕业论文质量, 杜绝抄袭等学术不端行为的发生。

〔关键词〕论文质量; 抄袭检测; 影响因素; 粗糙集

DOI: 10.3969/j.issn.1008-0821.2012.01.008

〔中图分类号〕G642.477 〔文献标识码〕A 〔文章编号〕1008-0821(2012)01-0032-06

Analysis on Impact Factors of the Quality Control of Graduation Thesis of the Graduate Based on Rough Set

Zhang Xin

(School of Management, Tianjin Normal University, Tianjin 300387, China)

〔Abstract〕Academic fraud is currently the endless stream of the phenomenon of university academic and moral decline and make the lack of scientific spirit. In this paper, as a college administered by the undergraduate thesis study, paper quality management for the control period prior to the detection of plagiarism detection system results as the basic data, using statistical analysis and rough set data mining, looking for quality pre-control study External and internal factors, and proposes how to control these factors to improve the quality of undergraduate thesis, put an end to plagiarism and other academic misconduct occurred.

〔Key words〕thesis quality; plagiarism testing; impact factors; rough set

1 问题提出与解决思路

近年来, 我国高校学术造假现象越来越严重, 这种行为不但违背了学术道德和科学精神, 而且使得整个学术环境变得浮躁和急功近利。这种现象的泛滥和其后果的严重性使得整个学术界从上到下都在努力地进行监管和治理。面对科研创新的需求, 我们需要培养大批具有学术道德和科学精神的科研人员, 而这种培养应从本科学习阶段就要开始, 对于学术造假的监管也要从本科毕业论文做起, 所以越来越多的高校开始应用各种论文抄袭检测系统, 对本科毕业论文进行检测。但这种监管属于事后控制, 为了进一步提高毕业论文质量, 杜绝抄袭、拷贝等学术不端现象

的发生, 学校更应加强事前和事中控制, 才能对帮助学生建立良好的学术道德、严谨的科学精神起到真正的作用。

本文针对上述提出的问题, 用统计分析与粗糙集, 对所调研学院的毕业本科论文检测结果的数据进行分析与挖掘, 通过专业间的对比分析找到影响高校本科毕业论文质量事前控制的外部因素, 通过论文相关信息分析找到影响高校本科毕业论文质量事前控制的内部因素。希望这些因素为各高校本科毕业论文质量事前控制工作提供其管理的关键点。

2 调查样本的选取

本研究选取了某管理学院所设 5 个专业的本科毕业论

收稿日期: 2011-10-12

基金项目: 天津市高等学校人文社会科学研究项目“基于制造业服务化的都市工业园发展模式”(项目编号: 20082102); 天津师范大学博士基金项目“基于服务型制造的都市工业园发展模式研究”(项目编号: 52X09021)。

作者简介: 张昕(1978-), 女, 讲师, 博士, 研究方向: 制造业服务化。

文学术不端检测结果数据,共选取样本420个,其中专业1样本107个,专业2样本92个,专业3样本113个,专业4样本83个,专业5样本25个。

专业1是所调研学院特色+品牌专业,设有2个班;专业2是由所调研学院的专业1和专业5各分出一个方向而重新组合的专业,设有2个班;专业3是所调研学院最传统的专业,设有2个班;专业4是所调研学院高考录取分数最高的专业,设有1个班,此外该专业有辅修生;专业5是所调研学院学科建设最强的一个专业,设有1个班。

3 本科毕业论文质量事前控制的外部影响因素分析

3.1 各专业论文查重率检测结果的统计分析

根据该学院所在学校规定,查重率小于等于30%的论文通过检测,查重率大于30%的论文属于不同程度的抄袭行为,需对论文进行修改后方可参加答辩。所以本研究以查重率30%为界限,对各专业毕业论文检测结果进行了汇总,如表1所示。

表1 各专业毕业论文查重率检测结果汇总

专业	查重率	人数	专业	查重率	人数
1	≤30%	58	4	≤30%	56
	>30%	49		>30%	27
2	≤30%	67	5	≤30%	21
	>30%	25		>30%	4
3	≤30%	76			
	>30%	37			

根据上表所汇总的数据,结合各专业选取的样本量大小,可得到各专业查重率超过30%在其所有被检测学生中所占的比例,如图1所示。

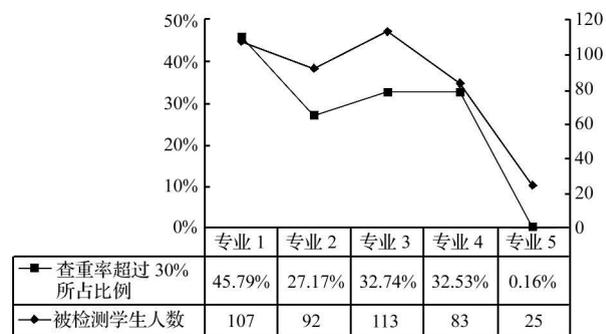


图1 各专业查重率超过30%的毕业论文所占比例

从图1可知,专业1查重率超过30%的学生占所有检测学生的比例最高,其次是专业3和专业4,而专业5是该比例最小的。

这个结果与各专业基本情况产生了很大的反差,出乎意料的是:一是作为所调研学院特色和品牌的专业1,其

中具有不同程度抄袭行为的论文占到了将近50%;二是作为所调研学院高考录取分数最高的专业4,这种情况也达到了将近1/3。但这个结果也有意料之中的发现,就是作为所调研学院学科建设较强的专业5,虽然其参与检测的人数相对较少,但具有抄袭行为的论文仅占到总数的0.16%,并且由此专业分出一个方向而组成的专业2,也相对情况好些。

3.2 外部影响因素与事前控制的要点分析

从上述分析,结合本研究调研时了解到各专业的具体情况,可找到导致上述情况产生的一些外部影响因素,而这些因素就是在毕业论文质量管理事前控制中需要把握的关键点。

3.2.1 专业的学生人数

查重结果从差到好进行排序,排在第一和第二位的分别是专业1和专业3,这两个专业都是所调研学院人数偏多的专业,而排在最后一位的是专业5,是所调研学院人数最少的一个专业。原因是学生人数偏多的专业,毕业论文指导教师所带的学生人数较多,导致毕业论文质量出现问题。所以,学校及学院教务部门每年需对毕业生较多的专业进行严格监控与管理,如对这些专业的毕业论文全部进行检测,以此来督促这些专业的指导教师对毕业论文的严格要求与监督、学生对毕业论文的充分重视和认真态度。

3.2.2 学科建设程度

从专业5的良好表现以及专业2的较好表现,可以看出,专业学科建设加强的同时会促进学术风气建设的不断提高,使得指导教师对学生论文要求越来越严格,学生受到教师的耳熏目染和身传言教,也会提高其对学术原创性的重视。

3.2.3 辅修专业的毕业论文

作为高考录取分数最多的专业4,其生源应该是5个专业中最好的,但查重结果却不尽人意。本文在寻找原因的过程中发现,该专业共选取了83个样本(包括辅修专业的学生22个),而辅修专业学生查重率超过30%的人数达到13人,占到59.09%,其抄袭行为比任何一个专业都严重。这是因为这些学生与本专业学生相比专业知识相对薄弱。而专业4本专业学生的毕业论文参加检测的有61篇,其中14篇有抄袭行为,占到22.95%,比专业2的情况还要稍好一些,仅差于专业5。所以,对于有辅修学生的专业,教务部门需要重视辅修专业的毕业论文监控工作,同样也可对辅修专业的学生毕业论文进行全部检测。

从上述3点分析,可知作为学科建设较强的专业5查重结果比较理想的原因所在,也可找到生源最好的专业4查重结果较差的原因所在。而专业1和专业3之所以查重结果不理想,一部分原因可归结为学生人数较多导致。但

专业2学生人数也相对较多,为何没有出现专业1和专业3的情况。尤其是专业1作为品牌专业,为何会出现目前的结果。本文接下来将通过进一步的数据挖掘对上述这些疑问进行探究。

4 本科毕业论文质量事前控制的内部影响因素分析

4.1 专业1和专业3论文查重率检测结果的统计分析

专业1样本中涉及到的指导教师共11名,以101~111作为指导教师的代号。将11名教师所指导的毕业论文的检测统计如下,见表2。

根据表2所汇总的数据,可得到专业1每位教师所指导的毕业论文中查重率低于30%和超过30%各占比例,如图2所示。

表2 专业1毕业论文检测结果汇总

指导教师	查重率	人数	指导教师	查重率	人数
101	≤30%	7	107	≤30%	6
	>30%	4		>30%	2
102	≤30%	8	108	≤30%	3
	>30%	4		>30%	11
103	≤30%	2	109	≤30%	7
	>30%	6		>30%	4
104	≤30%	3	110	≤30%	6
	>30%	5		>30%	5
105	≤30%	5	111	≤30%	3
	>30%	3		>30%	5
106	≤30%	8			
	>30%	0			

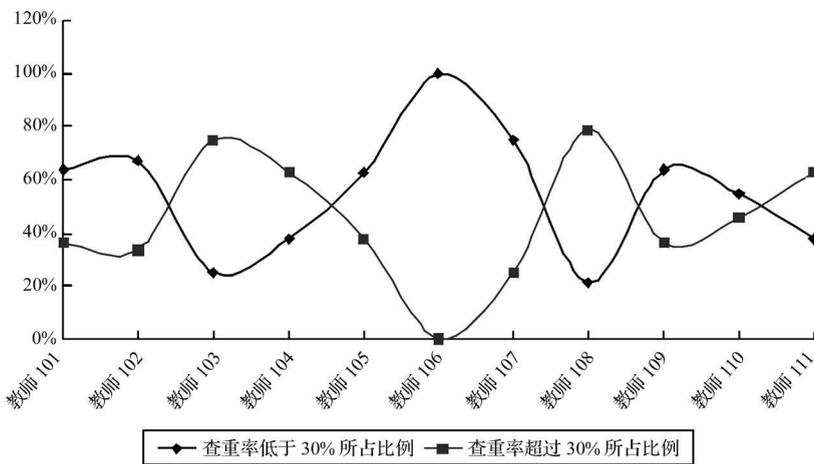


图2 专业1各教师所指导的毕业论文查重率低于30%和高于30%各占比例

从图2可知,专业1中的教师103、教师104、教师108和教师111都出现了查重率超过30%的论文所占比例高于查重率小于30%的论文所占比例,其中以教师108和教师103最为严重;其他不存在这种情况的教师中,教师106所指导论文的查重结果最好,全部为查重率低于30%,不存在论文抄袭情况;此外,教师110虽然所指导论文中查重率低于30%的比重高于查重率高于30%的比重,但仅是略高于,相差很近。

从统计数据分析,虽然能总结出上述这些结论,但导致这些情况发生的原因却不得而知,接下来本文将利用粗糙集的方法对专业1所有查重率超过30%的样本进行数据挖掘,找出导致查重率偏高的原因,得出论文质量的影响因素,将此作为提高论文质量工作的关键点进行监控。

4.2 专业1论文质量事前控制的内部影响因素分析

4.2.1 方法的选取

首先,本科毕业论文质量的影响因素非常多,且这些因素又存在着太多的不确定性因素,导致所搜集的这些数据又未必都对论文质量产生重要影响,存在一定的噪音。

其次,所搜集到的这些数据都是离散型的,且可能存在不精确甚至不完整的问题。鉴于上述考虑,本文运用粗糙集这种较新的软计算方法,对这些数据进行处理,根据这些不确定甚至不完整的信息进行推理,找出主要是哪些内部因素影响了本科毕业论文的质量,以及它们是如何影响的,影响程度又如何。

4.2.2 数据处理

我们从原有的420个样本中,只选取了属于专业1的107个样本,其中有19个样本查重率为零,没有显示出基本信息,故有效样本88各。每个样本的基本信息主要包括性别、论文字数、指导教师、选题类型和查重率。现将每个样本的这些信息都进行数字化处理,为此有如下设置:

- ①性别:男为1;女为2。
 - ②论文字数 X : $6000 < X < 10000$ 为1; $10000 < X$ 为2。
 - ③指导教师:共11名,分别用101~111代表。
 - ④选题类型:科学技术为1;生产实践为2;社会经济为3;经典为4;自拟为5;其它为6。
 - ⑤查重率 R : $R \leq 30\%$ 为1; $30\% < R < 50\%$ 为2; $50\% \leq R$ 为3。
- 由此得到表3。

表3 专业1毕业论文基本信息的数字化表示

U	A				
	a	b	c	d	e
1	2	2	101	3	1
2	1	1	101	3	1
3	2	1	101	3	1
4	2	1	101	3	1
5	2	1	101	2	1
6	2	1	101	4	2
7	2	2	101	2	2
8	2	1	101	4	2
9	1	1	101	3	3
10	2	1	102	3	1
11	2	2	102	2	1
12	2	2	102	2	1
13	2	2	102	4	2
14	2	1	102	4	2
15	1	1	102	4	3
16	1	2	102	4	3
17	2	1	103	3	1
18	2	2	103	3	1
19	2	2	103	4	2
20	2	1	103	4	2
21	2	1	103	4	2
22	2	1	103	4	3
23	2	2	103	4	3
24	2	2	103	4	3
25	2	1	104	3	1
26	2	2	104	3	1
27	2	1	104	3	1
28	2	2	104	4	2
29	2	1	104	4	2
30	2	1	104	4	3
31	2	1	104	4	3
32	1	2	104	4	3
33	2	1	105	3	1
34	2	1	105	2	1
35	1	1	105	2	1
36	2	1	105	2	2
37	2	1	105	2	3
38	2	2	105	2	3
39	2	2	106	2	1
40	2	1	106	2	1
41	2	1	106	2	1
42	1	1	106	2	1
43	1	2	106	2	1
44	2	1	106	2	1
45	2	1	107	3	1
46	2	2	107	3	1
47	2	1	107	4	2
48	2	1	107	4	2
49	2	1	108	2	1

续表3

U	A				
	a	b	c	d	e
50	2	1	108	3	1
51	2	2	108	2	1
52	2	1	108	4	2
53	2	1	108	4	2
54	2	2	108	4	2
55	2	1	108	4	2
56	2	1	108	4	2
57	2	1	108	4	2
58	2	2	108	4	3
59	2	2	108	2	3
60	2	1	108	4	3
61	2	1	108	4	3
62	2	1	108	4	3
63	1	2	109	3	1
64	2	2	109	2	1
65	2	1	109	2	1
66	2	2	109	3	1
67	1	1	109	2	1
68	2	1	109	4	2
69	2	1	109	4	2
70	2	1	109	4	3
71	1	1	109	4	3
72	1	1	110	3	1
73	2	1	110	3	1
74	2	1	110	3	1
75	2	1	110	2	1
76	2	1	110	4	2
77	2	1	110	4	2
78	2	1	110	4	2
79	2	2	110	4	3
80	2	2	110	4	3
81	2	1	111	3	1
82	1	2	111	2	1
83	1	1	111	2	1
84	2	2	111	4	2
85	2	1	111	4	2
86	2	1	111	4	3
87	2	2	111	4	3
88	2	2	111	4	3

表3中 $U = \{1, 2, \dots, 88\}$ 表示个体集, 分别为88名学生毕业论文的序号。 $A = \{a, b, c, d, e\}$ 表示属性集, 分别为属性: 性别、论文字数、指导教师、选题类型和查重率。现用 Rough 集方法对表3进行约简, 以达到挖掘数据潜力、节约时间和空间的目的。第一步, 删去重复行。第二步, 对于出现矛盾的规则, 删去查重率情况较差的, 因为这可能是学生写作过程中出现问题所导致的, 属事中控制, 不属于本文研究内容。约简后经过整理, 得到表4。

表4 专业1毕业论文一致决策表

U	A				
	a	b	c	d	e
2	1	1	101	3	1
5	2	1	101	2	1
3	2	1	101	3	1
6	2	1	101	4	2
7	2	2	101	2	2
1	2	2	101	3	1
15	1	1	102	4	3
16	1	2	102	4	3
10	2	1	102	3	1
14	2	1	102	4	2
11	2	2	102	2	1
13	2	2	102	4	2
17	2	1	103	3	1
20	2	1	103	4	2
18	2	2	103	3	1
19	2	2	103	4	2
32	1	2	104	4	3
25	2	1	104	3	1
29	2	1	104	4	2
26	2	2	104	3	1
28	2	2	104	4	2
35	1	1	105	2	1
34	2	1	105	2	1
33	2	1	105	3	1
38	2	2	105	2	3
42	1	1	106	2	1
43	1	2	106	2	1
40	2	1	106	2	1
39	2	2	106	2	1
45	2	1	107	3	1
47	2	1	107	4	2
46	2	2	107	3	1
49	2	1	108	2	1
50	2	1	108	3	1
52	2	1	108	4	2
51	2	2	108	2	1
54	2	2	108	4	2
67	1	1	109	2	1
71	1	1	109	4	3
63	1	2	109	3	1
65	2	1	109	2	1
68	2	1	109	4	2
64	2	2	109	2	1
66	2	2	109	3	1
72	1	1	110	3	1
75	2	1	110	2	1
73	2	1	110	3	1
76	2	1	110	4	2
79	2	2	110	4	3

续表4

U	A				
	a	b	c	d	e
83	1	1	111	2	1
82	1	2	111	2	1
81	2	1	111	3	1
85	2	1	111	4	2
84	2	2	111	4	2

4.2.3 对条件属性实行约简

采取对条件属性逐个被去掉的方法, 检查去掉该条件后是否会出现新的矛盾规则, 以此判定该条件属性是否可被约简。由此得到, 条件属性 a、b、c 的都不能被约去, 都为核。

4.2.4 决策分析

由上述分析可知, 性别、论文字数、指导教师和选题类型都对毕业论文查重率产生影响。在这些因素中, 指导教师属于主观因素, 而其他因素要么是客观存在, 要么可在毕业论文写作指导要求中给予明确规定, 相对较好在事前进行控制。所以, 当我们把指导教师这一属性约去后, 同时对产生矛盾的规则, 选择删除查重率较差的, 原因在前文已阐述。可得到表5。

表5 属性约简表

U	A			
	a	b	d	e
35	1	1	2	1
42	1	1	2	1
67	1	1	2	1
83	1	1	2	1
2	1	1	3	1
72	1	1	3	1
43	1	2	2	1
82	1	2	2	1
63	1	2	3	1
5	2	1	2	1
34	2	1	2	1
40	2	1	2	1
49	2	1	2	1
65	2	1	2	1
75	2	1	2	1
3	2	1	3	1
10	2	1	3	1
17	2	1	3	1
25	2	1	3	1
33	2	1	3	1
45	2	1	3	1
50	2	1	3	1
73	2	1	3	1
81	2	1	3	1

续表 5

U	A			
	a	b	d	e
11	2	2	2	1
39	2	2	2	1
51	2	2	2	1
64	2	2	2	1
1	2	2	3	1
18	2	2	3	1
26	2	2	3	1
46	2	2	3	1
66	2	2	3	1
6	2	1	4	2
14	2	1	4	2
20	2	1	4	2
29	2	1	4	2
47	2	1	4	2
52	2	1	4	2
68	2	1	4	2
76	2	1	4	2
85	2	1	4	2
13	2	2	4	2
19	2	2	4	2
28	2	2	4	2
54	2	2	4	2
84	2	2	4	2
15	1	1	4	3
71	1	1	4	3
16	1	2	4	3
32	1	2	4	3

从表 5 可发现如下规律：(1) 从查重率为 3 的几个样本可看出，男生比女生更容易出现严重抄袭的情况；(2) 从样本可看出，论文字数超过 1 万字的论文更容易出现查重率较差的情况；(3) 从样本可看出，查重率为 1 的都是选题类型为生产实践和社会经济的论文；(4) 从样本可看出，查重率为 2 和 3 的都是选题类型为经典的论文。

此外，对于指导教师这一内部影响因素，如前文所述，专业 1 中的教师 103、教师 104、教师 108 和教师 111 所指导论文的查重率结果相对不好，其中以教师 108 和教师 103 最为严重，而教师 106 所指导论文的查重结果不存在抄袭问题。根据表 3 所反映的信息，结合指导教师的具体情况，我们可发现：接近退休年龄的指导教师所带的毕业论文更容易出现问题；指导学生偏多的教师也较容易出现论文质量问题。

4.3 专业 1 论文质量事前控制的要点分析

从上述分析，可发现对某个专业而言，其毕业论文质量的内部影响因素：学生性别、论文字数、选题类型、指导教师。只有控制好这些关键点才能做好毕业论文质量管理的事前控制工作。而针对这些内部影响因素如何进行控

制，本文建议如下：

(1) 在论文指导中需明确指出：论文选题应更多侧重于真题真做的生产实践和反映现实问题的社会经济方面的论文，尽量避开经典题目；论文字数要求控制在 6 000~10 000 之间，不易过长。

(2) 对指导教师：①根据学校和专业具体情况，明确指出指导学生人数的上限；②可根据指导教师的具体情况，如年龄、专业等，适当调整指导学生的数量。

(3) 在事中管理控制中，要加强对男性学生论文质量的监控，加强对年龄偏大和所带人数偏多的教师论文指导工作的监督。

5 结 论

本文基于统计分析和粗糙集，对所调研学院的本科毕业论文查重结果进行了数据挖掘，从而找到了本科毕业论文质量的外部 and 内部影响因素。首先，外部影响因素主要是 3 个：专业的学生人数、学科建设程度、辅修专业的学生人数；其影响方式是：专业的学生人数较多将导致查重结果相对较差，辅修专业的学生人数偏多会导致本专业论文查重率变差，学科建设较好的专业论文查重率相对较好。其次，内部影响因素主要是 4 个：学生性别、论文字数、选题类型、指导教师；其影响方式是：男生比女生的论文相对查重率结果更差；论文字数偏多的相对查重率结果更差；选题类型为经典的论文查重结果更差；不存在抄袭行为的论文基本选题类型都为生产实践和社会经济；接近退休年龄的和指导人数偏多的教师所指导的毕业论文查重率结果更差。

本文所研究内容希望对具体负责本科毕业论文质量管理工作的部门有所启示和帮助，但本文仅选取了一所学校进行调研，必然导致所研究问题具有一些特殊性，但这种研究思路和方法可以应用到其他学校的毕业论文质量管理工作中。

参 考 文 献

- [1] 黄海蓉. 提升毕业论文质量的关键过程控制点技术研究 [J]. 中国校外教育: 理论, 2009, (3): 58.
- [2] 姚志红. 从毕业论文选题看提高学生毕业论文质量的对策 [J]. 兰州教育学院学报, 2010, 26 (6): 214- 216.
- [3] 张祖庆, 姜雅莉. 本科毕业论文质量实证分析 [J]. 科技信息, 2010, 35: 19- 20.
- [4] 陆玉梅. 经管类本科毕业论文质量的调查与分析 [J]. 江苏技术师范学院学报, 2010, 16 (10): 105- 108.
- [5] 吴平. 新建本科院校毕业论文 (设计) 管理和教学环节对应性研究 [J]. 赤峰学院学报: 汉文哲学社会科学版, 2010, 31 (12): 194- 196.
- [6] 邢泽晶, 马维军, 张旭. 高校毕业论文全过程管理与系统实现探析 [J]. 黑龙江教育学院学报, 2010, 29 (12): 22- 23.
- [7] 刘清. Rough 集及 Rough 推理 [M]. 北京: 科学出版社, 2001, 8: 40- 95.