

**《2007 年国家科学图书馆青年人才领域前沿项目结题报告》**

# **相关性挖掘在科学评价中的适用性 及其应用研究**

撰写人：王俊、王建芳、李泽霞

完成时间：2009 年 9 月 20 日

修改时间：2009 年 10 月 18 日

## 摘 要

本研究引入相关性算法,如最佳子集变量算法、二元变量相关性分析等,对用于大学科技竞争力评价的科学指标体系进行优化研究。

通过可获得的完整数据,分析了2007年教育部23所高校的114个与科技相关的评价指标之间及这些评价指标与大学科技竞争力的相关性,获得了大学科技竞争力评价的最优评价模型。在此基础上,利用其他年份的高校科技数据及ESI的文献计量数据对最优评价模型的适用性与合理性进行验证,进而确定了一套较为适用的、合理的表征大学科技竞争力的指标体系。

**关键词:** 评价指标; 大学科技竞争力; 相关性算法

## Abstract

Related algorithms, such as the best subset variable algorithm, binary variable correlation analysis and so on, are introduced for the evaluation of university competitiveness to optimize the evaluation index system.

Firstly, we selected 23 universities. With complete data available, we analyzed the independence of indexes in 114 technology-related evaluation indicator, and the relevance of these indicators with the universities competitive scores from others' research. Then, with related algorithms, we put forward a optimal evaluation model for the evaluation of science and technology competitiveness of university. And then, with several years science and technology data and ESI bibliometrics indicator, we validate the applicability and legitimacy of optimal evaluation model, and indentify a set of more appropriate and reasonable indicators for the evaluation of university science and technology competitiveness.

**Keywords:** University ranking; related algorithms

# 目 录

<b>一、研究背景</b> .....	<b>1</b>
1.1 研究问题提出的背景.....	1
1.2 相关领域的研究进展.....	2
1.3 总结.....	6
<b>二、研究的主要内容和思路</b> .....	<b>8</b>
2.1 研究目的与内容.....	8
2.2 研究思路.....	8
<b>三、相关性挖掘方法在科学评价中的适用性分析及验证</b> .....	<b>10</b>
3.1 相关性挖掘方法在科学评价中的适用分析.....	10
3.2 相关性挖掘方法在科学评价中的适用验证研究过程与方法.....	12
<b>四、研究结论与成果</b> .....	<b>18</b>
4.1 主要研究结论.....	18
4.2 研究成果创新点.....	24
<b>五、存在的问题及建议</b> .....	<b>25</b>
5.1 存在的问题.....	25
5.2 今后建议.....	25
<b>参考文献</b> .....	<b>26</b>
<b>致 谢</b> .....	<b>27</b>

## 一、研究背景

### 1.1 研究问题提出的背景

在情报研究课题中，科学评价是非常重要的内容，包含了科研成果的评价、科研组织和个人的评价，以及科研活动的评价等等。科学评价方法的研究与应用，经历了同行评议、定量评价的发展历程，评价方法和指标日益综合化和系统化，采用定量与定性分析相结合，运用多指标体系开展综合化、系统化评价的趋势越来越明显。

国内外的科学评价研究已形成了比较完整的评价方法体系，其研究与应用还得到了多方面的发展，如多种评价方法被应用于同一评价目标和评价过程，相互验证和检验，或者特征与原理相同的多种评价方法被整合集成在同一评价方法与模型中。

现有的评价理论与方法主要以文献计量学理论为基础，对科学评价方法与技术的研究及应用很大程度上基于简单统计指标分析，集中于指标或指标体系的构建，在此基础上进行定性分析评价。对于评价的综合化和系统化，特别是学科发展状况评价的知识化需求的支持不足。

另一方面，以文献计量为主的评价指标的适用性、可靠性等，都引起越来越多的争议。因为简单指标很难反映事物发展的诸多方面，可靠性、可信度往往不足，综合指标会存在指标构建合理性的问题，指标体系中各指标之间又存在重复、相互影响等等问题，而指标间信息的重迭会夸大评价结果，因此很难得到广泛认同。对于多重指标的遴选问题，很大程度上依赖于专家的经验，定量的方法以主成分分析等统计方法为主，能够达到简单降维的目的，但对指标以及评价指标与评价目标之间的相关性的分析远远不够，很难确保各项指标之间的相互独立，不相互交叉，不互为因果，不相互矛盾，所以会严重影响评价结果的合理性。

目前，从企业、大学到研究机构等的竞争力评价都是情报研究的热点，从理论研究到实践分析有很多的成果，但在指标的选择和评价中仍存在很多的问题。因此，本课题将相关性挖掘的方法应用于机构竞争力分析中的

指标遴选与评价，旨在优化科学评价的指标体系，为更有效的指标构建和科学评价提供理论依据和新的解决方案。

## 1.2 相关领域的研究进展

### 1.2.1 大学竞争力评价的研究与实践进展

#### (1) 大学排名的研究

大学排名尤其是学术排名是科学评价中的一个最重要的领域。科学评价方法的研究与应用，起初以同行评议为主，后来逐步发展为定量为主、定性结合的综合评价阶段，评价方法和指标日益综合化和系统化。另一大主要特点是，运用多指标体系开展综合化、系统化评价的趋势越来越明显。

国内外的科学评价研究已形成了比较完整的评价方法体系，其研究与应用还得到了多方面的发展。而且，受经济全球化和高等教育国际化趋势的影响，出现了不少国际性的院系、学科排名。

Thomson Reuters 公司出版的 ISI Essential Science Indicators 数据库<sup>[1]</sup>给出了一个学科论文总被引次数排在世界前列的大学和其他研究机构的名单，结果得到了广泛认可并经常被大学引用作为自己学科实力的证明。

瑞士科学技术研究中心 (CEST) 在 2004 年的一项研究中，对世界大学的论文数量和质量进行了统计，给出了每所大学在自然科学、工程科学、农学与环境、生命科学、医学、社会科学、人文与艺术 7 个领域 107 个学科的论文影响力排名。

英国《泰晤士报高等教育增刊》自 2004 年开始对世界大学在人文艺术、社科、科学、技术、生命医学 5 个领域进行排名，采用的唯一指标是声誉调查得分。

德克萨斯大学信息技术中心在 2005 年完成公布了世界百强商学院排名，该排名采用的指标是在商学权威刊物上发表论文的数量。另外一些单学科的世界排名包括世界大学经济学系排名，政治学系排名等，这些排名所采用的指标也仅限于论文数量和被引用情况等基于国际论文的指标。

在国内的大学排名研究中，以上海交通大学的世界大学学术排名、武汉大学邱均平老师的大学竞争力排名，以及武书连等人进行“中国大学评价”的影响力最大。

上海交通大学在世界大学学术排名的基础上，进行了世界大学学科领域排名。其确定的学科水平的主要标志是卓越人才和高水平的科学研究，因此从这两个方面选择指标来评价。卓越的人才，即拥有并培养了一批在本学科领域享有很高声誉的学术大师；高水平的科学研究，即产出了一批创新性的、对学科发展意义重大的科研成果。为此，构建了世界大学学科领域排名的指标体系，包括：“获诺贝尔奖和菲尔兹奖的校友的折合数”、“获诺贝尔奖和菲尔兹奖的教师的折合数”、“各学科被引用次数很高的教师数”、“被科学引文索引和社会科学引文索引收录的论文数”和“高质量论文比例”等指标。此外，考虑到工程学科需要大量投入的事实，在工科排名中增加了“年度科研经费”指标，替代了获奖校友和获奖教师指标。<sup>[2]</sup>其中“获诺贝尔奖和菲尔兹奖的校友的折合数”是指对不同年代的获奖校友赋予不同的权重，每回推十年权重递减 10%，如 1991 年以后毕业的获奖校友的权重为 100%，1981-1990 年的权重为 90%，1901-1910 年的权重为 10%。最后计算 1901 年以来的获奖折合数，且如果一个校友在一所学校获得两个或以上学位，只计算最近的一次；“获诺贝尔奖和菲尔兹奖的教师的折合数”的计算方法是对不同年代的获奖者赋予不同的权重，每回推十年权重递减 10%，如 2001 年以后获奖者的权重为 100%，1991-2000 年的权重为 90%，1981-1990 年的权重为 80%，1911-1920 年的权重为 10%。最后计算 1911 年以来的获奖折合数，目的是为了更客观地反映一所大学的学术表现。<sup>[3]</sup>

武汉大学的大学竞争力排名以各高等学校出人才、出成果的数量、质量和水平为基本标准，采取综合评价与分层次、分类型、分学科、分专业的专项评价相结合的基本原则，从有关政府部门的统计数据资料、国内外有关数据库等多方面指标，对大学的办学资源、教学水平、科学研究与学校声誉等多方面进行评价。<sup>[4]</sup>2009 年武汉大学的中国大学及学科专业评价的做法有三个新的变化，一是在中国重点大学综合竞争力评价指标体系中，首次引入大学网络影响力指标替代社会声誉指标，占学校声誉指标的 50%。并明确提出了“中国一流大学”的概念、标准和排名结果，根据集中与离散分布规律，在学科专业评价结果的表示方面引入了星级表示方法。如此的改进使评价方法与评价结果更趋于科学、合理。<sup>[5]</sup>

## (2) 评价指标的分析研究

对于评价指标的选择，也是机构竞争力评价的一个主要研究内容。因此可供评价的数据往往有很多，但找出那些更能反映竞争实力，并用最省时、省力的方法评价出相对可靠的结果的指标却是值得研究的内容。另外，对于同样的基础数据，采用不同的权重确定方法，所得到的大学排名也大不相同。因此权重确定问题也是一个很重要的问题，尽量采取客观的赋权方法也是学者们追求的目标。

目前大学排名所采用的确定权重的方法都是主观赋权法，大体上是两类：专家咨询法和层次分析法。在武书连的评价中，使用的就是专家咨询法，但是遭到了许多非议，如顾海兵认为其“粗暴地把高校教员的人力总投入分为人才培养投入与科学研究投入”、“粗暴地把理科师生与文科师生分为 8:2 之数量比”，认为他这样设立权重是没有道理的。<sup>[6]</sup>

层次分析法 (AHP) 是一种定性与定量分析相结合的多因素决策分析方法。这种方法将决策者的经验判断给予数量化，在目标因素结构复杂且缺乏必要数据的情况下使用更为方便，因而在实践中得到广泛应用。AHP 法可对原始观测数直接加权运算进行综合排序，未削弱原始信息量，使评价指标逻辑判断量化且保持判断思维全过程的一致性。但构造各层指标的权重判断矩阵时，一般采用的是分级定量赋值，这可能会造成同一类中同一指标是另一指标的 5 倍、7 倍，甚至是 9 倍，从而影响了权重的合理性。

因此，专家咨询法及层次分析法都比较简便易行，但其存在的主要缺点是主观性太大，由主观赋权法确定出的权重系数真实与否，很大程度上取决于专家的知识结构、工作经验及其偏好，权重的确定存在不少主观随意性，这将严重影响评价结果的客观性。

为此，在评价指标体系赋权上，有学者进行了主观和客观相结合的组合赋权法，同时，在客观赋权法中，考虑到了指标相关性的问题，把指标相关性检验引入进来，使得指标的设置更加规范和合理。如上改进的指标体系的赋权方法，是针对大学评价中专家繁多、一些指标缺乏样本数据、指标体系中含有大量定性指标等问题提出来的。以上方法综合考虑了这些问题，是一种较好的解决方案。

### 1.2.2 相关性挖掘及其在评价指标分析中的应用

在科学评价之外，许多领域都涉及多指标综合评价的问题，如综合国力的比较、上市公司经营业绩的综合评价、投资方案的优选等。综合评价的方法尽管很多，但基本思路大体相同。在综合评价中，如何消除指标间的信息重叠、合理选择评价指标对于综合评价至关重要。这便是相关性挖掘方法可以解决的问题。

综合评价有两个基本原则，一个是全面性原则，一个是不相关性原则，即在设计评价指标时，应尽量多地选择指标，以覆盖评价对象的各个方面，选择的指标之间应互不相关。显然，指标选的太少，即使指标之间不相关，也不能反映评价对象的全貌；选的太多，指标之间势必存在相关性，这不但增加工作量，同时也会产生重复评价，影响综合评价的效果。因此，如何消除指标间的信息重叠、合理选择评价指标对于综合评价至关重要。

相关性挖掘是指对多重指标之间，以及指标与其所确定的评价目标的相关性的分析评价过程。最常用的评价指标相关性分析是用简单的线性相关性测度。另外，还有许多统计或数据挖掘的算法用于多重指标的约简，如主成分分析、粗糙集理论、神经网络等等。这些方法在指标优化方面发挥了很大的作用，但它们或多或少地存在一些不足，如主成分分析法通过对原指标的线性变换获得新的指标，揭示的信息点意义不明确；粗糙集理论应用上、下近似和等价关系等一系列概念，对指标(属性)进行约简，但是，在指标较多的情况下，其核(core)往往是不存在的，因而可能会导致算法复杂度的问题；神经网络由于本身固有的不足，对本问题不是很适合，且也会导致指标含义不明确。

随着各种评价指标体系综合化的要求，相关性研究在自然科学、社会科学领域都得到了广泛的应用。如医学领域内病症与指标的相关性分析，化学领域化合物结构与性质/活性相关性研究等等。在系统工程领域，对指标相关性的研究更加深入，如在指标相关性的消除方法上，提出了许多途径，如限制指标数量、分离重叠源、修正指标权重、主成分分析及因子分析等等<sup>[7]</sup>。还有研究人员提出依据评价指标相关矩阵，对指标集合进行等价类划分，然后利用随机向量线性相关性，对指标等价类加以优化求解，使得指标等价类之间的相关性最小，进而得到



等价类的指标综合函数。

如在评价企业竞争力水平时,有学者提出了基于信息粒度的属性约简进行指标体系构建。在属性权重的确定方面,在计算过程中将核 core 看作是约简,但核是所有约简的交集,所以核不一定是约简,因此又结合约简的定义,进一步提出了一种基于信息粒度的属性约简算法<sup>[8]</sup>。

### 1.3 总结

总结国内外主要的大学竞争力评价,有的以科研为主,但大部分侧重于综合实力评价,评价的指标各有侧重,方法大不相同,详见下表。

机构	排名类型	指标及方法
武汉大学科学评价研究中心	科研竞争力排名	科研生产力、科研影响力、科研创新力、科研发展力等;文献计量作为评估标准。
上海交通大学	学术排名	教育品质、教师品质、研究成果、机构规模。
泰晤士报高教副刊	以学术为主的综合排名	学术声誉、研究成果、教学品质及国际化。
美国新闻与世界报	综合排名	学术声誉、教师资源、学生素质、毕业率、校友捐赠等;以学术声誉调查为基础。
德国高等教育发展中心(The Center for Higher Education Development)	综合排名	学生与校友意见调查、教师专业意见调查、学校现况;学科排名结果以“分层法”的方式,根据评比结果分为顶尖组、中等组和低分组,以避免学科差异。

由表及上述分析可以看出,国外综合排名较多,且评价数据的来源和指标的确定以定性为主,受人为因素影响比较大,定量处理方法还比较模糊,没有形成比较完善的定量指标的确定和分析方法。

而以上相关性挖掘的方法比较有效地解决了评价指标合理选择的问题,在各领域的综合评价中发挥了比较重要的作用,而且研究人员也在不断引入新的方法来优化评价指标和模型,以提高评价的有效性。但如前文所述,在情报研究范围的科学评价,评价指标及其体系的构建很大程度上停留在简单的统计分析层面,综合化的评价分析指标建立往往存在盲目性,这在某种程度上影响了评价的效果。

科学评价的指标与方法问题是一个长期未解决的问题,在评价指标体系的设

计中，由于对指标之间关系的不确定，经常易导致部分指标之间存在不同程度的相关；当指标之间高度相关时，会出现多重共线性现象。简单说，共线性就是评价指标(自变量)对评价目标(因变量)的影响降低，标准误差扩大，由于评价指标之间的线性相关，导致它们所反映的信息高度重合，降低了评价指标的效度和信度。相关性挖掘方法正是提供了一种解决该问题的思路。因此，本研究尝试进行科学评价指标之间，及评价指标与评价目标间的相关性挖掘，以期在确定评价指标时既客观全面，又避免重复干扰，尽可能地保证指标的科学性、有效性、独立性和可操作性。

## 二、研究的主要内容和思路

### 2.1 研究目的与内容

本研究的目的是引入相关性挖掘的算法,研究科学评价指标体系的方法和过程。为此,选取比较热门的研究领域——大学学术评价作为案例,依据能够获取的完整数据,对大学科研机构科技竞争力评价指标体系进行研究,优选评价指标,进而与其他常用的指标进行对比,证明其适用性与可靠性。

因此,本课题的主要研究内容包括两个方面:

#### (1) 相关性挖掘算法的适用性分析与验证

本部分在文献调研的基础上,分析各种常用算法的适用范围和特征,总结现有应用方法的经验,在此基础上借鉴自然科学、社会科学等领域中相关性研究的理论和方法,引入相关性挖掘方法,对选定的大学的科技竞争力评价指标体系中的评价指标相互之间的关系进行各种相关性分析与挖掘进行数据处理和验证,以便用于机构的科技竞争力综合评价的指标优化工作中。

#### (2) 相关性挖掘算法在机构科技竞争力评价中的应用研究

选择目前情报研究领域的一个热点课题——大学竞争力评价作为案例,进一步进行算法适用性的研究,并利用应用案例进行指标的分析 and 约简,在此基础上提出大学科研竞争力的评价模型,并对其有效性进行验证。

本研究中,在文献调研的基础上,选定高校科技竞争力评价指标体系,利用相关性挖掘算法对科技竞争力与多重综合评价指标的相关性,以及指标之间的相关性进行分析,优化指标集,在此基础上对现有评价指标体系的研究进行佐证、补充与完善。

### 2.2 研究思路

本研究针对情报研究中的大学科技竞争力评价工作,引入相关性挖掘的理论和方法,对综合评价体系进行适当的技术处理,以达到消除指标间相关性的目的,进而提高评价结果的有效性。

根据可以获得的数据情况,本研究选取 23 所教育部直属高校(以下称为“大学”)作为研究对象机构(研究样本),选取与大学科技相关的信息(定量数据)

为大学科技竞争力评价指标体系的备选指标集合,选取大学科技竞争力得分为评价目标值。利用相关性算法对评价指标进行遴选,并对各类评价指标项的显著性进行分析等,以获得评价性能更科学有效的评价指标。将评价指标进行多层次的约简,直到最小的指标集合,可以更好表达竞争力。

根据所得最优指标集合,采用跨年度的指标数据,进行分析验证,进而通过一定的算法得出由最优指标构成的评价模型。

对得到的评价模型,再从不同的角度进行分析验证,证实方法过程的有效性,和所得结论的可靠性。

### 三、相关性挖掘方法在科学评价中的适用性分析及验证

#### 3.1 相关性挖掘方法在科学评价中的适用分析

相关性挖掘是指对多重指标之间，以及指标与其所确定的评价目标的相关性的分析评价过程。各种相关性挖掘方法适用于不同的应用领域，对于评价指标的选择优化发挥不同的作用，常见的方法有交叉验证、最佳子集变量算法 ( Leaps-and-bounds )、主成分分析法、因子分析法等；其他方法比如将指标集进行等价类划分，构造指标综合函数，并求解等价类指标间相关系数最小化模型，以消除指标的相关性；再如利用格兰姆-施密特正交化的方法也可以消除指标间的相关性等。

不同的算法功能与应用场景有所不同，常用的相关性挖掘算法及应用范围分析见表 3-1。

**表 3-1 相关性挖掘算法备选表**

序号	算法名称	基本思想	应用
1	一元回归分析	研究两个变量之间是否存在某种线性关系。	检验两变量相关性的最简单方法。
2	多元回归分析	研究一个变量与多个变量之间的相关性，一元回归分析的扩展。	(1) 确定几个变量之间是否存在相关关系； (2) 处理多重共线性问题； (3) 用一个或几个自变量的值预测或控制另一个因变量的取值。
3	最优子集回归	核心是回归树 (regression tree)，即按照一定算法，将测试数据形成某种树型结构，树中的结点为一数据子集。	用于选择最佳回归数学模型，也可用于变量的评估
4	二元变量相关分析	若两变量相关系数很高，则说明这两变量所表征的信息有较大的重复，可以考虑删除一个变量。	测试两变量之间的相关程度
5	偏相关分析	研究两个变量间存在相关性时，控制可能对其产生影响的变量。	
6	交叉验证	首先，抽取出一个试样，对剩下的试样进行多元回归分析，并对抽出的试样进行预测；	作为评估回归模型的一种判据

		然后，放回这一试样，再抽出另一试样，如此反复 $N-1$ 次，算出每一次的标准偏差，进而得到平均标准偏差值。 若此值较小，说明所得数学模型较稳定。	
6	主成分分析	根据本征值在总加和中的比率，可以选择相应的主成分作为新的变量。	主成分分析的结果可以作为变量评估和选择的理论依据： (1) 回归分析中变量的优化、约简； (2) 回归分析中的共线性检验； (3) 多元统计分析中异常点的鉴别； (4) 适用于变量数大于样本数的情况。
7	主成分回归	首先采用主成分分析法选取重要因子，然后再采用常规回归分析进行数学模型的构造	
8	因子分析	将多个变量表示为较少的因子	简化变量、降低变量维数，对原始变量再解释及命名
9	偏最小二乘法		适用于变量数大于样本数的情况

注：以上方法中涉及的相关性分析统一采用相关系数  $R$  进行衡量。选取的目标机构即为样本，评价指标即为变量。

以上算法在其他领域都不同程度地被用作评价指标的分析和优选，以消除指标的相关性，并得到最优指标集合。科学评价的指标，同样也存在相关性的问题。如科学计量指标中，定量的经费投入与其相关的过程指标之间，本身存在相关性，但往往同时被用作评价指标。另外，用于科学评价的文献计量指标，如论文数量与引文指标之间，往往存在一定的相关性，这些问题在科学评价中往往被忽视，造成某些参数的夸大，影响了评价结果的客观性。因此，在科学评价中对指标进行相关性分析也是非常必要的。

为此，本研究选取适当的案例，并建立在已有的指标体系基础上，利用文献调研等方法，对指标进行相关性分析，特别尝试在其他领域常用的相关性算法用于科学计量指标分析，如交叉验证、最佳子集变量算法

( Leaps-and-bounds )、关联分析等方法，分别将其用于评价指标相关性分析、评价指标遴选，通过算法分析与案例验证相结合，证明所选取的算法在评价指标体系构建中的适用性。

## 3.2 相关性挖掘方法在科学评价中的适用验证研究过程与方法

### 3.2.1 研究的方法

本研究在文献调研的基础上，采用了定量方法与定性分析相结合，通过理论分析与定量计算，并结合专家咨询、定期讨论交流实施研究工作。

定性分析主要结合文献调研与定期讨论交流的方式，对研究中的相关问题进行定性分析与经验判断。

定量研究方法是本研究采用的主要方法，利用上述的相关性挖掘算法进行指标的定量评价和分析。定量分析的软件工具为 SPSS、Matlab 以及自行开发的相关性分析工具。

### 3.2.2 研究工作计划

研究工作原计划：在科学评价中应用相关性挖掘，包括相关性挖掘在机构竞争力评价与学科评价中的应用研究。

开题汇报后，根据专家评委的指导建议，对任务书中计划的研究内容做了适当的调整，重点进行相关性挖掘在机构竞争力评价中的应用研究，并根据数据的可获得情况，选取大学作为目标机构；对任务书中涉及的“相关性挖掘在学科评价中的应用研究”部分暂不做研究。

调整后的研究计划主要分以下几步：

#### ( 1 ) 文献调研

在开题前期初步文献调研的基础上，进行深入的文献调研，主要针对机构竞争力评价的国内外研究现状进行调研分析，并对广泛应用于自然科学及社会科学诸多领域的相关性挖掘算法进行调研，了解并汇总其应用的总体情况。

#### ( 2 ) 相关性挖掘算法的分析与遴选

在 ( 1 ) 的基础上选取可应用于本研究的相关性挖掘算法。

#### ( 3 ) 研究对象的遴选

根据可以获得的数据情况，选取 23 所教育部直属高校（以下称为“大学”）作为本项目的研究对象机构（研究样本），为进行相关性挖掘算法的应用研究做准备；根据可以获得的与大学科技相关的信息（定量数据）为大学科技竞争力评价指标体系的备选指标集合。

#### （4）数据采集与标准化处理

根据相关性算法对数据对象的要求，对选定的大学科技竞争力评价的相关数值型数据进行标准化处理。

#### （5）大学竞争力评价的相关性挖掘

根据不同阶段的研究内容，选择合适的相关性算法，对大学科技竞争力与评价指标的相关性及各评价指标的重要程度进行分析，进而遴选出能表达大学科技竞争力的最优评价指标集合，并在此基础上尝试构建更有效的评价指标。

#### （6）撰写结题报告

总结以上的研究工作，整理并分析研究结果，撰写结题报告。

### 3.2.3 研究工作的组织与实施

#### （1）研究团队的组成及人员参与情况

项目成员	参与情况	研究交流形式
王俊	<ul style="list-style-type: none"> <li>◇ 负责项目申请及项目研究工作的组织与实施；</li> <li>◇ 数据调研及搜集；</li> <li>◇ 数据处理、分析及分析结果的汇总；</li> <li>◇ 组织并撰写结题报告。</li> </ul>	交流讨论
王建芳	<ul style="list-style-type: none"> <li>◇ 参与项目申请；</li> <li>◇ 文献调研及资料整理；</li> <li>◇ 数据处理、分析及相关结果整理；</li> <li>◇ 撰写结题报告。</li> </ul>	互相学习 协作研究
李泽霞	<ul style="list-style-type: none"> <li>◇ 参与项目申请；</li> <li>◇ 数据初期处理（如对纸版数据的扫描、整理、汇总）；</li> <li>◇ 数据分析及相关结果整理；</li> <li>◇ 参与撰写结题报告。</li> </ul>	

#### （2）研究工作的实施

本研究选取可以获得统计资料为数据，深入研究若干数据挖掘算法，以大学科技竞争力评价为例，对选取的相关性挖掘算法在情报研究中的适用性



进行分析。在此基础上，分析评价指标的相关性、对评价指标进行遴选以及对评价指标项的显著性进行分析等，以获得评价性能更稳定的评价指标。具体执行情况分以下几步：

#### 1) 文献调研

调研了与大学科技竞争力研究、相关性评价及相关性算法理论与应用方面的文献资料，并总结撰写了文献综述。

#### 2) 相关性挖掘算法的确定

在对相关性挖掘算法理论学习与研究的基础上，确定了本研究不同研究阶段采用的相关性挖掘算法，包括：针对指标的优选采用变量优选算法、针对指标与目标值的相关性研究采用多元回归分析等方法。

#### 3) 目标机构的遴选

考虑研究样本数据的可获得性，从而确定采用教育部直属的 23 所大学为目标机构。

**表 4-1 教育部直属 23 所大学**

序号	大学名称	序号	大学名称
01	北京大学	13	东南大学
02	清华大学	14	浙江大学
03	北京科技大学	15	山东大学
04	南开大学	16	武汉大学
05	天津大学	17	华中科技大学
06	大连理工大学	18	湖南大学
07	吉林大学	19	中南大学
08	复旦大学	20	中山大学
09	同济大学	21	重庆大学
10	上海交通大学	22	四川大学
11	华东理工大学	23	西安交通大学
12	南京大学		

数据来源：《高等学校科技统计资料汇编》，中华人民共和国教育部科学技术司编

#### 4) 评价指标备选集合及相关数据的选取与处理

根据可以获得的数据，选取 2007 年的 23 所大学的科技信息定量数据为评价指标的备选集合（说明：该部分工作开始于 2008 年 1 月，当时能够获得的最新数据为 2007 年的数据），全部备选指标集合详见表 4-2；根据相关性挖

据算法的要求，研究中对相关的数值型数据进行了适当的标准化处理。

**表 4-2 大学科技竞争力评价指标备选集合 ( 114 个指标 )**

指标类别及对应指标	指标类别及对应指标
<b>01 科技人力 ( 15 个 )</b>	<b>04 研究与发展项目 ( 20 个 )</b>
X1.1 科技活动人员总数	<研发项目合计>
X1.2 科技活动中科学家和工程师人数	X4.1 研发项目总数
X1.3 科技活动中辅助人员数	X4.2 研发项目当年总投入人员 ( 人年 )
X1.4 研发人员总数	X4.3 各项研究总在读研究生数
X1.5 研发中科学家和工程师数	X4.4 各项研究当年总拨入经费
X1.6 研发中辅助人员数	X4.5 各项研究当年总支出经费
X1.7 研发全时人员合计 ( 人年 )	<基础研究>
X1.8 研发全时科学家和工程师 ( 人年 )	X4.6 基础研究项目数
X1.9 研发全时辅助人员数	X4.7 基础研究当年投入人员( 人年 )
X1.10 R&D 成果应用及科技服务人员总数	X4.8 基础研究在读研究生
X1.11 应用及服务中科学家和工程师数	X4.9 基础研究当年拨入经费
X1.12 应用及服务中辅助人员数	X4.10 基础研究当年支出经费
X1.13 应用及服务全时合计 ( 人年 )	<应用研究>
X1.14 应用及服务全时科学家与工程师人数	X4.11 应用研究项目数
X1.15 应用及服务全时辅助人员数	X4.12 应用研究当年投入人员( 人年 )
<b>02 教学与科研人员职务 ( 8 个 )</b>	X4.13 应用研究在读研究生
X2.1 教授人数	X4.14 应用研究当年拨入经费
X2.2 副教授人数	X4.15 应用研究当年支出经费
X2.3 讲师人数	<试验发展>
X2.4 助教人数	X4.16 试验发展项目数
X2.5 教师系列其他人数	X4.17 试验发展当年投入人数
X2.6 高级技术职务人数	X4.18 试验发展在读研究生
X2.7 中级技术职务人数	X4.19 试验发展当年拨入经费
X2.8 初级技术职务人数	X4.20 试验发展当年支出经费
<b>03 科技经费收支 ( 7 个 )</b>	<b>05 R&amp;D 成果应用及科技服务项目 ( 15 个 )</b>
X3.1 政府资金拨入	<成果应用及科技服务项目合计>
X3.2 企事业单位委托投入	X5.1 成果应用及科技服务项目数
X3.3 其他经费投入	X5.2 成果应用及服务当年总投入人数
X3.4 劳务费支出	

<p>X3.5 业务费支出 X3.6 转拨外单位支出 X3.7 其他支出</p>	<p>X5.3 应用及服务总在读研究生 X5.4 应用及服务当年总拨入经费 X5.5 应用及服务总当年支出经费</p> <p>&lt;研发成果应用&gt; X5.6 研发成果应用项目数 X5.7 研发成果应用当年投入人数</p> <p>X5.8 研发成果应用在读研究生 X5.9 研发成果应用当年支出经费</p> <p>X5.10 研发成果应用当年拨入经费 &lt;科技服务&gt; X5.11 科技服务项目数 X5.12 科技服务当年投入人数 X5.13 科技服务在读研究生 X5.14 科技服务当年拨入经费 X5.15 科技服务当年支出经费</p>
<p><b>06 国际科技交流 ( 6 个 )</b></p> <p>X6.1 合作研究派遣人数 X6.2 合作研究接受人数 X6.3 出席国际会议人数 X6.4 国际会议交流论文数 X6.5 国际会议特邀报告数 X6.6 主办国际会议次数</p>	<p><b>07 科技成果获奖 ( 5 个 )</b></p> <p>X7.1 国家自然科学奖 X7.2 国家发明奖 X7.3 国家科技进步奖 X7.4 国务院各部门科技进步奖 X7.5 省、市、自治区、科技进步奖</p>
<p><b>08 科技成果 ( 23 个 )</b></p> <p>&lt;出版科技著作、学术论文&gt; X8.1 出版著作数量 X8.2 出版著作千字 X8.3 出版学术论文总篇数 X8.4 国外论文数</p> <p>&lt;不同项目来源的国家级验收项目&gt; X8.5 国家级验收项目数 X8.6 与外单位合作项目数 X8.7 973 计划 X8.8 科技攻关计划 X8.9 863 计划 X8.10 自然科学基金</p>	<p><b>09 技术转让 ( 15 )</b></p> <p>&lt;合同数&gt; X9.1 各类合同总数 X9.2 国有企业合同数 X9.3 外资企业合同数 X9.4 民营企业合同数 X9.5 其他类型合同数</p> <p>&lt;合同金额&gt; X9.6 各类型合同总金额 X9.7 国有企业合同金额 X9.8 外资企业合同金额 X9.9 民营企业合同金额 X9.10 其他类型合同金额</p>

<p>X8.11 其他项目</p> <p>&lt;专利情况&gt;</p> <p>X8.12 专利申请总数</p> <p>X8.13 申请发明专利数</p> <p>X8.14 申请实用新型专利数</p> <p>X8.15 申请外观设计专利数</p> <p>X8.16 专利授权总数</p> <p>X8.17 发明专利授权数</p> <p>X8.18 实用新型专利授权数</p> <p>X8.19 外观设计授权数</p> <p>X8.20 专利出售合同数</p> <p>X8.21 专利出售总金额</p> <p>X8.22 专利出售当年实际收入</p> <p>X8.23 其他知识产权件数</p>	<p>&lt;当年实际收入&gt;</p> <p>X9.11 各类合同当年实际总收入</p> <p>X9.12 国有企业合作收入</p> <p>X9.13 外资企业合作收入</p> <p>X9.14 民营企业合作收入</p> <p>X9.15 其他类型合同收入</p>
---	--

注：相关经费单位为千元。

#### 5) 大学科技竞争力与评价指标的相关性挖掘

主要实施情况分为三个阶段：(a) 根据 2007 年 23 所大学的科技相关信息数据与调研获得的大学科技竞争力得分 ( 设为  $Y_{obs}$  )<sup>[9]</sup>，通过相关性挖掘，得出大学科技竞争力得分  $Y_{obs}$  与遴选的最优评价指标集合的最优线性方程 ( 即大学科技竞争力评价的最优指标模型 )；(b) 再利用遴选出的最优评价指标与可获得的其他年份的数据，代入 (a) 的最优线性方程进行验证，通过计算得出大学竞争力得分的计算值 ( 设为  $Y_{cal}$  )，与调研获得的其他年份的 23 所大学的科技竞争力得分  $Y_{obs}$ <sup>[10]</sup> 进行线性拟合，以验证最优线性方程的适用性；(c) 选取 SCI 论文数、被引频次及篇均被引次数 3 个文献计量指标数据<sup>[11]</sup>与 (b) 中得到的大学科技竞争力得分的计算值  $Y_{cal}$  进行相关性分析，从而验证最优线性方程的合理性，进而确定了一套较为适用的、合理的表征大学科技竞争力的指标体系。

#### (6) 撰写结题报告

总结以上的研究情况，整理并分析研究结果，撰写结题报告。

## 四、研究结论与成果

本研究主要从定量评价指标的角度，在力求比较全面反映科研实力的基础上，通过相关性挖掘优选评价大学科技竞争力的较为科学、客观的指标。

### 4.1 主要研究结论

本项目的研究结论主要从以下三个研究阶段获得，主要研究步骤及相关数据显示如下。

#### 4.1.1 大学科技竞争力评价的最优指标模型的确定

##### (1) 数据的表示及预处理

为了符合相关性挖掘的要求，需要先将样本数据（即 2007 年 23 所大学的与科技相关的 9 大类 114 个指标）表示为如下形式：

$$X_{m,n}' = (X_{1,1}, X_{1,2}, \dots, X_{2,1}, X_{2,2}, \dots, X_{3,1}, X_{3,2}, \dots, X_{m,n}) \quad (\text{公式 5.1})$$

其中， $m$  表示指标所属类别的编号，取值范围为[1,9]； $n$  表示所属指标类的指标编号，由于每类的指标数不同， $m$  取值不同  $n$  取不同整数值。

为了减小不同指标数据之间在数量级上的差别，需要进行数据的预处理，即对相关指标数据进行标准化处理，如下所示标准化公式：

$$X_{std} = X_{mn} / Arg(X_{mn}) \quad (\text{公式 5.2})$$

其中， $X_{std}$  为标准化数据， $Arg(X_{mn})$  为 23 个样本的各指标数据的平均值。

##### (2) 大学科技竞争力得分与各类评价指标的相关性分析

为了确定每个指标与大学科技竞争力得分  $Y_{obs}$  的相关性，及每类指标对大学科技竞争力得分  $Y_{obs}$  的重要程度，进行了如下分析：

- ◇ 每个指标与大学科技竞争力得分做二元变量分析，获得评价指标与大学科技竞争力得分  $Y_{obs}$  的相关性情况，得到了与大学科技竞争力得分  $Y_{obs}$  的相关系数  $R$  大于 0.5 的评价指标集合（如表 5-1 所示）。这 26 个指标与大学科技竞争力得分  $Y_{obs}$  的多元回归分析的结果显示，该指标集合与大学科技竞争力的相关性较好，回归的相关系数  $R$  达到 0.90。

表 5-1 相关系数高于 0.5 的评价指标集合

序号	指标名称	相关系数 $R$
01	X4.6 基础研究项目数	0.8438
02	X3.1 政府资金拨入	0.8213
03	X6.5 国际会议特邀报告数	0.7634
04	X8.4 国外论文数	0.7038
05	X4.1 研发项目总数	0.6657
06	X4.7 基础研究当年投入人数	0.6397
07	X8.1 出版著作数量	0.6382
08	X2.1 教授	0.6202
09	X3.7 其他支出	0.6040
10	X4.9 基础研究当年拨入经费	0.5860
11	X6.3 出席国际会议人数	0.5836
12	X8.7 973 计划	0.5812
13	X8.13 申请发明专利数	0.5716
14	X6.4 国际会议交流论文数	0.5621
15	X1.5 研发中科学家和工程师数	0.5600
16	X1.8 研发全时科学家和工程师人数	0.5599
17	X1.7 研发全时人员合计	0.5520
18	X1.4 研发人员总数	0.5519
19	X4.2 研发项目当年总投入人数	0.5519
20	X9.14 民营企业合同收入	0.5483
21	X6.2 合作研究接受人数	0.5444
22	X9.13 外资企业合同收入	0.5399
23	X9.8 外资企业合同金额	0.5338
24	X8.17 发明专利授权数	0.5334
25	X8.10 自然科学基金	0.5119
26	X8.12 专利申请总数	0.5095

注：表中评价指标按相关系数由高到低排列。

- ◇ 对各个评价指标类与大学科技竞争力得分  $Y_{obs}$  进行多元回归分析，分析各个评价指标类对大学科技竞争力得分  $Y_{obs}$  的重要程度（如表 5-2 所示）。结果显示，9 类评价指标中对大学科技竞争力的影响最大的是科技成果类指标，而科技成果获奖类指标对大学科技竞争力的影响力为最小。

表 5-2 各个指标类对大学科技竞争力得分的重要程度

指标类别名称	回归显著性
08 科技成果	0.9750
06 国际科技交流	0.8760
03 科技经费收支	0.8560

09 技术转让	0.7440
02 教学与科研人员职务	0.7280
01 科技人力	0.6960
05 R&D 成果应用及科技服务项目	0.6730
04 研究与发展项目	0.6320
07 科技成果获奖	0.5880

### (3) 各类评价指标间的相关性分析

在不考虑大学科技竞争力得分影响的情况下，考察各类评价指标的相关性。通过各类指标内部的两两指标之间的相关性分析，获得投入-产出指标相关性分析结果。

表 5-3 投入-产出指标相关性比较

投入指标 \ 产出指标	出版学术论文数	国外论文数	973 计划	863 计划	自然科学基金	发明专利数	合同总数
政府资金拨入	0.6523	0.7680	0.6854	0.4010	0.7242	0.8137	0.3893
教授	0.7654	0.6392	0.3964	0.3006	0.4154	0.5463	0.4676
当年拨入经费	0.6510	0.4876	0.3351	0.4615	0.3809	0.7853	0.4880
基础研究当年拨入经费	0.3891	0.4267	0.2072	0.4454	0.2604	0.5661	0.5456
科技活动中科学家和工程师人数	0.5375	0.3452	0.1669	0.2198	0.1854	0.3064	0.2026
科技活动人员总数	0.5365	0.3368	0.1885	0.2489	0.1980	0.3144	0.2032
应用研究当年拨入经费	0.4998	0.3069	0.1864	0.3434	0.2609	0.6082	0.3998

### (4) 确定大学科技竞争力评价优选指标集合

运用最优子集变量回归法对(2)中相关系数  $R$  高于 0.5 的 26 个评价指标进行再次指标选择，基于指标数目尽量少且相关系数相对合理的原则，遴选出 X4.6 基础研究项目数、X8.4 国外论文数和 X8.10 自然科学基金数 3 个评价指标，这三个评价指标与大学科技竞争力得分的多元回归分析结果显示，它们与大学科技竞争力的相关系数  $R$  约为 0.90，显示其相关性与 26 个指标的集合的相关性相当。由此，获得一套指标数量较少、且能够较好地表征大学科技竞争力的评价指标。

### (5) 确定大学科技竞争力评价的最优指标模型

通过对 2007 年 23 所大学的科技相关的 114 个指标之间的相关性及其与大学科技竞争力得分的相关性进行分析，从而得出大学科技竞争力得分与遴选

的最优评价指标集合的线性方程 ( 如下公式所示 ), 即大学科技竞争力评价的最优指标模型 :

$$F(x) = a + \sum_{i=1}^{n-1} b_{n-1} X_{n-1} = 1.0061 + 0.1559X_0 + 0.1098X_1 - 0.0830X_2 \quad (\text{公式 5.3})$$

#### 4.1.2 大学科技竞争力评价指标模型适用性的验证

利用遴选出的最优评价指标的可获得的其他年份的数据 ( 包括 2006 年及 2008 年 ), 代入公式 5.3 的最优线性方程进行验证, 由此得到各年的大学科技竞争力得分的计算值  $Y_{cal}$ , 分别如表 5-4、表 5-5 所示; 并与调研获得的其他年份的 23 所大学的科技竞争力得分  $Y_{obs}$  进行线性拟合, 相关系数  $R$  均能够达到 0.90, 显示最优线性方程能够适用于其他年份的大学科技竞争力评价指标数据。<sup>[10]</sup>

表 5-4 2006 年 23 所大学的科技竞争力得分的计算值

目标机构	大学科技竞争力得分观察值 $Y_{obs}$	大学科技竞争力得分计算值 $Y_{cal}$
北京大学	124.49	410.27
清华大学	154.68	428.23
北京科技大学	14.54	86.85
南开大学	36.46	160.96
天津大学	39.56	176.37
大连理工大学	23.54	186.41
吉林大学	50.39	181.08
复旦大学	74.1	135.62
同济大学	24.07	88.96
上海交通大学	88.25	455.60
华东理工大学	17.11	92.72
南京大学	66.73	239.23
东南大学	26.58	142.56
浙江大学	114.72	587.84
山东大学	39.07	146.07
武汉大学	58.16	282.48
华中科技大学	60.02	175.21
湖南大学	13.13	60.88
中南大学	32.04	99.66
中山大学	55.6	269.01
重庆大学	16.77	84.43
四川大学	51.86	383.45
西安交通大学	47.59	162.29



表 5-5 2008 年 23 所大学的科技竞争力得分的计算值

目标机构	大学科技竞争力 得分观察值 $Y_{obs}$	大学科技竞争力 得分计算值 $Y_{cal}$
北京大学	122.25	929.18
清华大学	160.88	924.61
北京科技大学	14.31	311.50
南开大学	43.33	262.08
天津大学	38.03	666.88
大连理工大学	29.13	417.99
吉林大学	49.73	419.80
复旦大学	72.80	740.12
同济大学	24.92	626.65
上海交通大学	84.42	1086.44
华东理工大学	15.37	324.64
南京大学	76.39	360.00
东南大学	28.88	410.85
浙江大学	118.14	1894.42
山东大学	40.95	436.85
武汉大学	50.97	754.90
华中科技大学	81.12	964.44
湖南大学	17.69	235.80
中南大学	30.11	510.75
中山大学	51.98	759.98
重庆大学	18.87	377.59
四川大学	44.15	827.05
西安交通大学	49.38	456.18

#### 4.1.3 大学科技竞争力评价指标模型合理性的验证

为了进一步验证最优线性方程的合理性，本研究选取了 ESI 的 SCI 论文数、被引频次及篇均被引次数 3 个文献计量指标数据（如表 5-6）<sup>[11]</sup>与（5.1.2）中得到的大学科技竞争力得分的计算值  $Y_{obs}$  进行线性拟合，得出这三个文献计量学指标与大学科技竞争力得分计算值  $Y_{obs}$  的相关系数  $R$  能够达到 0.90，显示了最优线性方程的合理性。

表 5-6 23 所大学文献计量指标数据

学校名称	文献计量指标	2003 年	2006 年	2007 年	2008 年
北京大学	SCI 论文数	7,963	11,925	13,078	14,365
	被引频次	17,498	42,629	50,598	62,668
	篇均被引次数	2.2	3.57	3.87	4.36

清华大学	SCI 论文数	9,036	15,189	16,221	14,146
	被引频次	12,496	34,134	41,464	49,697
	篇均被引次数	1.38	2.25	2.56	3.51
北京科技大学	SCI 论文数	1,382	2,005	2,213	2,551
	被引频次	1,472	2,310	2,535	3,375
	篇均被引次数	1.07	1.15	1.15	1.32
南开大学	SCI 论文数	3,090	4,794	5,407	6,155
	被引频次	5,519	12,692	16,172	22,124
	篇均被引次数	1.79	2.65	2.99	3.59
天津大学	SCI 论文数	1,660	4,148	4,825	5,428
	被引频次	1,721	6,206	8,285	11,036
	篇均被引次数	1.04	1.5	1.72	2.03
大连理工大学	SCI 论文数	1,734	3,798	4,738	5,713
	被引频次	2,279	7,839	10,939	14,530
	篇均被引次数	1.31	2.06	2.31	2.54
吉林大学	SCI 论文数	3,259	5,371	6,272	7,095
	被引频次	6,454	13,184	17,576	21,710
	篇均被引次数	1.98	2.45	2.8	3.06
复旦大学	SCI 论文数	4,374	7,545	8,645	9,778
	被引频次	9,565	22,772	28,544	37,465
	篇均被引次数	2.19	3.02	3.3	3.83
同济大学	SCI 论文数	1,225	2,201	2,627	3,202
	被引频次	1,724	4,145	5,842	7,378
	篇均被引次数	1.41	1.88	2.22	2.3
上海交通大学	SCI 论文数	3,499	8,741	10,749	12,990
	被引频次	4,514	15,860	22,308	33,922
	篇均被引次数	1.29	1.81	2.08	2.61
华东理工大学	SCI 论文数	1,080	2,049	2,490	3,107
	被引频次	1,515	4,466	6,422	9,301
	篇均被引次数	1.4	2.18	2.58	2.99
南京大学	SCI 论文数	6,179	8,629	9,335	9,877
	被引频次	12,232	23,666	29,069	35,772
	篇均被引次数	1.98	2.74	3.11	3.62
东南大学	SCI 论文数	546	1,135	1,397	1,712
	被引频次	573	2,112	3,230	4,567
	篇均被引次数	1.05	1.86	2.31	2.67
浙江大学	SCI 论文数	6,169	12,583	14,887	17,289
	被引频次	10,015	28,387	38,206	50,988
	篇均被引次数	1.62	2.26	2.57	2.95
山东大学	SCI 论文数	3,353	5,666	6,378	7,355
	被引频次	4,862	11,774	14,766	18,694
	篇均被引次数	1.45	2.08	2.32	2.54
武汉大学	SCI 论文数	2,405	4,713	5,578	6,341
	被引频次	3,994	11,139	15,863	21,444
	篇均被引次数	1.66	2.36	2.84	3.38
华中科技大学	SCI 论文数	2,273	4,921	5,871	7,120
	被引频次	2,707	8,285	10,949	14,557
	篇均被引次数	1.19	1.68	1.86	2.04

湖南大学	SCI 论文数	874	1,949	2,439	2,981
	被引频次	1,381	4,568	6,946	9,619
	篇均被引次数	1.58	2.34	2.85	3.23
中南大学	SCI 论文数	488	1,588	1,961	2,177
	被引频次	313	2,448	3,665	5,490
	篇均被引次数	0.64	1.54	1.87	2.52
中山大学	SCI 论文数	2,869	4,683	5,650	6,711
	被引频次	6,453	14,031	18,675	25,260
	篇均被引次数	2.25	3	3.31	3.76
重庆大学	SCI 论文数	483	1,262	1,538	1,949
	被引频次	474	1,731	2,432	3,400
	篇均被引次数	0.98	1.37	1.58	1.74
四川大学	SCI 论文数	2,131	4,333	5,477	6,896
	被引频次	2,447	8,461	12,098	16,585
	篇均被引次数	1.15	1.95	2.21	2.41
西安交通大学	SCI 论文数	1,964	4,103	4,868	5,736
	被引频次	2,378	6,496	7,847	10,704
	篇均被引次数	1.21	1.58	1.61	1.87

#### 4.1.4 对情报研究工作的意义

本研究对情报研究工作的意义主要有两个方面：

##### (1) 实践中：评价指标的优选

本研究依据能够获取的完整数据，将相关性算法（如，最佳子集变量算法、二元变量相关性分析等）应用于大学科技竞争力评价指标体系的研究中，研究结果显示备选指标集合中遴选的评价指标能够较好地表征大学科技竞争力。这对于从定量的角度构建具有科学性、适用性的评价指标体系是一种大胆的尝试。

##### (2) 方法上：尝试相关性挖掘方法在情报研究中的应用

在日常情报研究工作中，采用相关性挖掘辅助情报研究，对定性研究确定的指标体系的科学性进行评价，并进行指标的优选。

#### 4.2 研究成果创新点

国内外关于大学竞争力评价的研究中，关于评价数据的来源和指标的确定大多以定性研究为主。

本研究的创新点在于，采用相关性分析算法对评价指标数据进行定量统计与分析，分析了评价指标之间及这些评价指标与大学科技竞争力的相关性，获得了大学科技竞争力评价的最优评价模型。并利用逐年的高校科技数据及 ESI

的文献计量数据对最优评价模型的适用性与合理性进行验证，进而确定了一套较为适用的、合理的表征大学科技竞争力的指标体系。

## 五、存在的问题及建议

### 5.1 存在的问题

结合本研究工作的具体情况，总结本研究存在的问题主要有两个：

#### (1) 数据获取问题

在科学评价中，合理的数据来源是重要的基础，我们的研究也是为了选择合理的数据用于评价。但由于种种原因，最合适的表征特征往往不能得到相应的数据，在我们的研究中，希望尽量获取全面的数据，在大量数据分析的基础上遴选优选指标。但在我们获取的研究样本数据中，存在部分指标数据某些年缺失的情况，影响了对最优评价指标模型的更广泛验证。

#### (2) 相关性挖掘算法的利用

相关性挖掘算法的利用方面，除了本研究中涉及的算法，还有其他算法或许在这里也是有意义的，因此，在未来的研究中，需要进一步尝试新的算法和理论。另一方面，在本研究中验证适用的算法，还需要更多的实验数据和场合进行进一步的深入验证，并在其他情报研究课题中进一步验证其适用性。

基于以上算法适用性及其在各方面案例中的应用分析，进一步探索相关性挖掘在评价指标体系的评估、优化以及其他类情报研究中的更深入应用。

### 5.2 今后建议

希望再进一步深入研究其他相关性挖掘方法的应用，并在大学竞争力评价中，充分考虑指标优化的问题，提高指标确定的科学性。尝试将定量分析的思想引入科研机构竞争力评价研究中，并开展相关性研究思想的适用性在其他情报研究课题中的实证研究。

## 参考文献

- [1] <http://esi.isiknowledge.com/home.cgi>
- [2] 刘念才,程莹等.我国名牌大学学科领域离世界一流有多远-从世界大学学科领域排名说起  
[J].高等教育研究,2007,28(10):1-8
- [3] Liu NC, Cheng Y. Academic Ranking of World Universities – Methodologies and Problems,[2008-12-15]. <http://www.arwu.org/rank/file/ARWU-M&P.pdf>
- [4] 邱均平, 赵蓉英等. 中国大学评价报告(2006-2007)-中国大学及学科专业评价的理念与实践. 科技进步与对策, 2006(7)
- [5] 邱均平, 罗力等. 大学评价的改进与完善. 高教发展与评估, 2009, 25(3):19-28
- [6] 王琴.我国大学评价方法研究[J].评价与管理, 2008,12(4):57-61
- [7] 徐祥发, 肖人彬. 评价指标相关性的消除方法研究. 系统工程理论与实践. 2002(11)
- [8] 刘杰. 基于信息粒度的属性约简改进算法在企业竞争力指标体系构建中的应用. 中国管理信息化. 2008,11(15): 98-100
- [9] 中国重点大学科技竞争力排行榜,  
<http://edu.sina.com.cn/gaokao/2007-05-21/104883494.html>
- [10] <http://www.people.com.cn/GB/kejiao/39/20030212/921947.html>,  
<http://info.edu.hc360.com/2006/01/11102188494-2.shtml>,  
<http://learning.sohu.com/20080107/n254507326.shtml>
- [11] ISI web of knowledge 的 Essential Science Indicators,2009

---

## 致 谢

感谢国家科学图书馆青年人才领域前沿项目对本研究的资助。

感谢总分馆领导及专家评委对本研究从立项申请到研究组织实施所给予的大力支持与有益指导，感谢业务处及情报研究部各位老师对本项目各方面工作的监督与帮助。

感谢项目组成员王建芳、李泽霞在本研究工作进行过程中付出的努力。