【大数据专题】

国外大数据硕士人才培养的经验与启示
——基于大数据文本挖掘

阮 敬，刘宏晶，纪 宏
（首都经济贸易大学 统计学院，北京 100070）

摘要：利用半结构化文本数据分析方法，从国外高校 387 个大数据硕士相关项目及国内 22 个相关硕士项目人才培养方案中提取出大数据高端人才培养的七大方向，并对不同方向的培养目标、课程、学分、学制等设置及其对应的人才市场需求匹配情况等相关内容进行剖析，为改革中国大数据高端人才的供给提出一定建议。

关键词：大数据；大数据人才；人才培养；网页文本挖掘
中图分类号：C82；C41 文献标志码：A 文章编号：1007－3116(2017)09－0029－08

一、引 言

2015 年 8 月，国务院印发的《促进大数据发展行动纲要》明确了发展大数据的指导思想、发展目标和发展任务，标志着大数据正式上升为国家核心战略。同年 10 月，《中共中央关于制定国民经济和社会发展第十三个五年规划的建议》提出要“实施国家大数据战略，推进数据资源开放共享”。2017 年的政府工作报告中也专门提出“促进大数据、云计算、物联网广泛应用”，这是自 2014 年首次进入政府工作报告以来大数据连续三年成为中国政府的聚焦。然而，中国大数据发展过程中仍旧面临着众多制约因素，其中大数据人才的稀缺则是关键。

根据现有人才市场上对大数据人才的需求状况，大数据人才是掌握数学、统计学、数据分析、机器学习和自然语言处理等多方面知识且具备大数据处理能力的复合型人才。早在 2012 年 3 月，美国政府宣布启动“大数据研究和发展计划”，将“大数据研究”上升为国家意志，赞助各级学术单位进行大数据相关研究，邀请跨学科的研究人员共同探讨大数据对教学领域的影响，并鼓励科学研究院设立研究课程，培育下一代的数据科学家与工程师。继美国率先开启大数据国家战略先河之后，欧洲、日本等国家和地区也跟进出台了相应的大数据人才战略举措。

北卡罗莱纳州立大学于 2005 年 6 月首次提出开办数据分析专业申请，2007 年 2 月正式成为美国第一个数据分析硕士研究生学位授予单位，从 2011 年、2012 年开始美国高校的大数据硕士课程大量招生。北京航空航天大学于 2013 年设立了数据科学专业硕士课程，这是中国最早开设数据科学硕士培养的高校，相比而言中国大数据高端人才培养起步较晚。有学者对比了中国和美国的统计学本科专业指导性教学纲要指出，大数据背景下中国统计类专业在教学目标、课程体系和教学内容等方面都还存在一定差距；朱建平和李秋雅认为中国大数据相关专业课程体系建设要综合考虑大数据专业人才所应具备的知识和技能，并积极借鉴发达国家对大数据专业人才培养的理念[1]。解决大数据人才短缺的问题需要借鉴国外先进经验，近几年这也引起了很多学者的关注。有学者根据国内外有关大数据的研究动向，指出大数据分析的研究需多学科联合；何海地通过深入分析美国 23 所知名大学大数据分析硕士课程网站信息，总结出了美国大学数据科学专业硕士课程设置非常重视实践[2]；阮敬和陈涛以美国 20 多所院校的著名大数据项目为例，对中国应用统计

收稿日期：2017－03－04；修复日期：2017－06－09
作者简介：阮 敬，男，广西桂林人，经济学博士，教授，研究方向：大数据分析；
刘宏晶，女，山西介休人，硕士生，研究方向：大数据分析；
纪 宏，男，北京人，教授，博士生导师，研究方向：大数据分析。

专业硕士的大数据分析人才的培养模式进行了探讨[3]；迪莉娅以国外开设数据科学硕士课程的部分高校为例，分析了国内外高校数据科学专业硕士课程设置的内容和特点，提出完善中国高校数据科学硕士课程设置的策略[4]；祝丹和陈立双基于大数据驱动下学科交叉发展的视角，探讨了统计学专业人才培养模式[5]。当前对于大数据人才的培养，其相关研究基本是比较宏观的以典型学校为代表进行的研究，目前还没有利用大数据技术全面、系统地对国内外大数据项目进行分类剖析。

2015 年 11 月，国务院学位办全国应用统计专业学位研究生教育指导委员会在上海财经大学举办的学位授权点评估总结会上，笔者代表首都经济贸易大学提出了“国外有先例，国内有需求，培养有特色”的大数据专业硕士人才的培养方式。基于此，本文针对人才培养规律，利用大数据分析技术搜集、整理和分析了来自 20 个国家及地区的 257 所高校的 387 个大数据相关项目和国内 26 所高校的 22 个大数据项目的人才培养方案，在大数据视角下为中国大数据人才的培养途径提出建议，同时也为其他学科人才培养方案的设置提供一种基于大数据分析的新思路。

二、研究数据及方法

(一)研究数据

本文研究数据全部来源于互联网的公开资料。本文从 863 270 184 条与大数据硕士人才培养相关的 Google 搜索结果中选取出热度较高的 387 个相关项目的网址进行 url 解析，并利用 Python 网络爬虫爬取出网页的内容，建立了包括项目名称、项目概况、培养目标、必修课程、选修课程、学分等与人才培养相关的指标体系，根据这个指标体系分别对由所爬取数据构成的非结构化数据库运用正则表达式等文本分析技术提取并量化关键信息，再利用文本挖掘技术整理和分析课程的设置情况。为了进行国际比较，对国内 26 所高校的 22 个大数据项目也采用了上述相同思路进行分析。

为了进一步探究国外大数据人才培养途径是否符合国内的人才需求，本文还从中国最大的某招聘网站上爬取了最近一个月内发布且最低学历要求在本科及以上、全职、与大数据相关的岗位招聘信息，其信息涵盖了职位名称、工作地点、发布日期、工作经验、岗位职责及岗位要求等内容，共计 8 650 条记录。本文将对这些大数据相关职位的岗位要求与大数据项目不同方向的课程设置进行匹配度分析。

(二)研究方法

1. 正则表达式。正则表达式是一种描述字符串匹配的模式，用单个字符串来检索、匹配和替换符合某个句法规则的文本，利用正则表达式能够高效准确地对已知特征的信息进行抽取。基于正则表达式的匹配方式，首先将网页文档作为字符流来处理[6]，再根据所建立的指标体系编写出一种模式来迅速匹配到要抽取的信息，并将非结构化的数据库转化为半结构化的数据集。

2. 特征项。文本挖掘需要将特征用合适的模型组织起来合理地表示文本，最经典的将文本表示成向量、矩阵或元组的方法是向量空间模型[7]。假定文本中词与词是不相关的，将文本表示为以特征项的权重为分量的向量，每个文本表示如下：

$$d = (w_1, w_2, \cdots, w_n) \tag{1}$$

其中 d 表示文本，向量中元素 w_i 表示第 i 个特征项的权重；特征项的权重用 TF-IDF 表示，即用特征项频度和文本频度这两个参数来描述特征项表达文本内容属性的能力，其计算方法为：

$$w(t_j, d_i) = \frac{\log(\mathrm{tf}(t_j, d_i)+\alpha) \cdot \log\left(\frac{|D|}{\mathrm{df}(t_j)}\right)}{\sqrt{\sum_k \left[\log(\mathrm{tf}(t_k, d_i)+\alpha) \cdot \log\left(\frac{|D|}{\mathrm{df}(t_j)}\right)\right]^2}} \tag{2}$$

式(2)中 $D = \{d_1, d_2, \cdots, d_m\}$ 表示文本集合，集合元素 d_i 表示第 i 个文本，t_j 表示第 j 个特征项，$\mathrm{tf}(t_k, d_i)$ 表示在文本 d_i 中特征项 t_k 的频度，$\mathrm{df}(t_j)$ 表示 t_j 的文本频度，α 为平滑因子。

特征项的提取是文本聚类的前提。在对国外大数据项目进行分类时，根据项目概况的特征项利用 K-MEANS 聚类方法，将数目较多的硕士类项目分为不同方向。

3. 特征相关度和文本相似度。特征相关度和文本相似度均是基于向量空间模型来度量的。本文利用特征相关度深层次地研究不同方向的大数据项目的必修和选修课程的设置情况，两特征项的相关度为两个特征项同时出现的概率除以两个特征项分别单独出现的概率之积的对数；文本相似度是通过计算两个向量的夹角余弦值来评估的，余弦值越大表明相似性越强；在向量空间模型中如有 $D_1 = D(w_{11}, w_{12}, \cdots, w_{1n})$、$D_2 = D(w_{21}, w_{22}, \cdots, w_{2n})$，则这两个文档的相似性为：

$$\mathrm{sim}(D_1, D_2) = \cos\theta = \frac{\sum_{k=1}^{n} w_{1k} * w_{2k}}{\sum_{k=1}^{n} w_{1k}^2 * \sum_{k=2}^{n} w_{2k}^2} \tag{3}$$

在本文的大数据人才需求研究中，将各职位招聘要求与大数据不同方向项目的课程内容基于此原理分别进行文本相似度计算，最后选取出匹配度最高的大数据方向。

三、国外高校大数据硕士人才培养状况

（一）大数据人才培养项目的分布及培养方向

国外高校的大数据项目主要分为学位和证书两大类，学位项目包含硕士学位和博士学位。在本文的分析中主要以硕士项目为主，博士和证书类项目的分析结果作为对比参考。

将硕士类项目根据其概况进行文本聚类，对聚类效果进行多重比较后最终确定为数据科学、应用统计、商业分析、商务智能、健康医疗、信息系统以及MBA（大数据方向）7个方向。在对学分进行比较时考虑到欧洲国家使用不同的学分体系，因此对于学分要求的讨论按照欧洲类国家和非欧洲类国家来划分更有意义，故在此将英国、法国、德国、奥地利、瑞典、意大利、立陶宛、荷兰、芬兰、爱尔兰、西班牙和丹麦归为欧洲类，其他国家归为非欧洲类。

本文研究的国外高校大数据项目来自多个国家和涵盖多个方向。首先，对项目所属地区、类型和上课形式进行统计，其结果见图1所示。图1中图形的面积与分类的数目成正比，可以看出商业分析硕士项目最多，反映了目前市场对大数据分析人才需求大；其次，是大数据证书类，证书所花费的时间和耗用的金钱较少，在资源有限的情况下想要提高某一领域的技能时证书项目是一个有吸引力的选择，很受在职人士的青睐；再次，以研究为主要目的的博士类项目相对较少，而其他6个方向的数目相当。

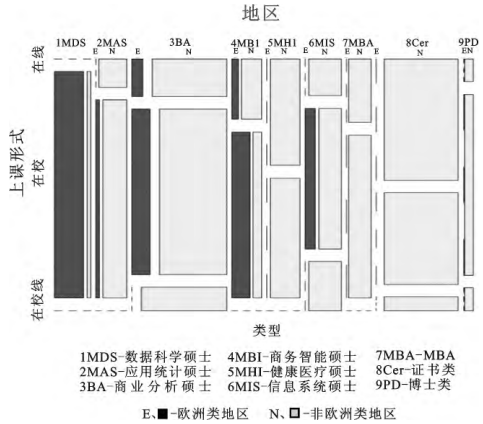
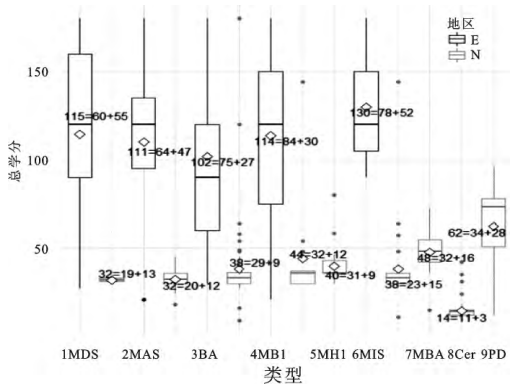


图1　国外高校大数据项目分布及培养方向图

以美国为代表的非欧洲类国家的大数据项目要远远多于以英国为代表的欧洲类国家，虽各类型占比不同但整体上以非欧洲类国家的项目为主，其中数据科学和商务智能硕士欧洲类国家开设更多，因欧洲类国家更注重培养学生管理、监督和评估商务问题的能力，期望将其培养成为新一代的数据科学家。两大类地区上课形式的比例相近，证书类项目以网络形式为主，学位类以在校教学为主，且非欧洲类地区部分高校开设的商业分析硕士、信息系统硕士、证书类和博士类项目还可根据自身情况自主选择上课形式。

（二）学分要求

不同培养方向的大数据项目学分要求不同，大数据项目的学分设置情况如图2所示。从图2可以看出欧洲类国家的总学分要比非欧洲类国家要多，这与学校每门课程学分设置情况有关。欧洲国家绝大部分学校的单课学分集中于0.5、1、1.5、2、3，非欧洲国家学校单课学分多为5、7.5、10、15、20 ECTS（欧洲制学分）；欧洲类国家的大数据项目类型以硕士学位的数据科学、应用统计、商业分析、商务智能和信息系统为主，就总学分平均水平而言，商业分析最少，信息系统最多；非欧洲类国家的大数据项目种类多样，证书类项目平均总学分最少，博士类最多，学位类居中，学位中MBA规定的总学分要比其他方向的学位类大数据项目多。



注：图中文本标签等式的3个数字依次表示该类项目平均总学分、平均必修学分和平均选修学分。

图2　国外高校大数据各类项目学分分布图

总学分包括必修学分和选修学分，两类地区相同方向的必修学分和选修学分的比例相当。证书类的大数据项目所需总学分最少但其必修学分比重最大，而博士类项目的选修学分所占比重最大。这是由不同层次项目的培养目标决定的：证书类项目是在短时间内提升学生某一方面的应用专长，课程亦是集中于相对固定内容设置的；博士类项目主要是培养学生发现其所研究领域未解决问题的新理论和新方法，倾向于理论研究。此外，博士阶段不同研究方向的课程内容也不尽相同，学生可以根据兴趣自主选择课程以便于更好地学习和研究。

(三)先修课程及数据分析工具

国外高校的大数据项目的招收群体主要有两类:一是想加深对某领域的理解,期望在该领域的未来就业市场中获得竞争优势的学生;二是想在现有职位上有所突破或想重新定位职业的专业人士。申请项目的最基本要求是拥有学士学位,所以在申请项目之前需要有一定的基础,各方向在这方面的要求不同,具体结果见表1。

表1 国外大数据项目先修课程及数据分析工具表

项目类型	先修课程	数据分析工具
数据科学	数学	SAS Python
应用统计	数学 统计学	SAS R
商业分析	统计学 经济学 数学 计算机	SAS R Python SQL Tableau
商务智能	计算机	SAS R Java
健康医疗	数学 统计学 生物学	SAS Java
信息系统	计算机	C++ Java
MBA	统计学 计算机	SAS R Python
博士类	数学 统计学	SAS
证书类	统计学 计算机	SAS

从表1中可以看出:国外高校大数据项目的选修课程集中于数学、统计学和计算机学科。首先,统计学最多,因大数据分析建模的核心离不开统计学,具有统计背景的学生对于大数据的学习和建模有很大优势;其次,是数学和计算机,大数据分析人才需要的是全能型人才,不仅需要有扎实的理论基础,还要求有一定的编程经验;再次,还有部分类型的大数据项目有特殊要求,如商业分析还要求掌握经济学的基本知识,健康医疗则需要有一定的生物学基础。

计算机是大数据处理技术的重要工具,掌握至少一种编程语言是大数据高端人才不可缺少的一种能力,这也是市场和行业的基本需求。由表1提炼出来的结果表明:SAS是最受国外高校青睐的编程语言或工具,R、Python和Java也使用得较多;除上述这几种软件外,信息系统类的项目对计算机的要求较高,还要求掌握C++等脚本语言;商业分析类项目所使用的大数据软件比较广泛,除大数据分析常用的SAS、R和Python外,还注重培养学生利用SQL对数据库进行处理以及采用Tableau对分析结果进行可视化的能力。

(四)课程设置

课程设置是人才培养的核心。不同层次学校的课程对教育及社会的影响程度不同,为了提高课程的分析质量,本文对照USNEWS公布的2016年全球大学综合排名前750名单,对所考察院校设置了权重,其中对哈佛大学、麻省理工学院、斯坦福大学、芝加哥大学、约翰斯·霍普金斯大学、伦敦帝国学院、康奈尔大学、伦敦大学学院、西北大学和波士顿大学这10所高校的大数据项目赋予5的权重;对以纽约大学和南加州大学为代表的15所学校赋予4的权重;对利兹大学及罗切斯特大学等20所大学赋予3的权重;对比萨大学和乔治华盛顿大学等39所学校赋予2的权重;其他学校权重为1。

1.课程概况。课程内容是基于培养目标而规划的,不同学位类型和方向的课程设置之间既有联系又有区别。图3是在90%稀疏矩阵且词条相关度不低于0.35的基础上绘制的关于所有大数据项目全部课程的无向网络结构图。图3中的课程词条为节点,圆或文字越大说明该词条的度越大,即与更多词条有相关性;课程也主要是围绕这些内容设置的,边越粗代表两节点间的权重越大,两关键词的相关度也越大;图3表明,大数据项目的课程主要是围绕统计、数据库、安全、计算机、管理、信息、分析、方法、数据、金融和会计等内容展开的。以应用统计硕士项目为例,该项目旨在传授统计思想和数据分析先进知识及技能,加强学生利用复杂数据进行有效分析、推断和预测的能力。作为如今国内最热门的大数据方向,其课程以数据、分析、统计、时间序列为核心,侧重培养学生利用统计思维对数据进行分析的能力,这与其培养目标一致(见图4)。

图3 大数据项目全部课程词条网络结构图

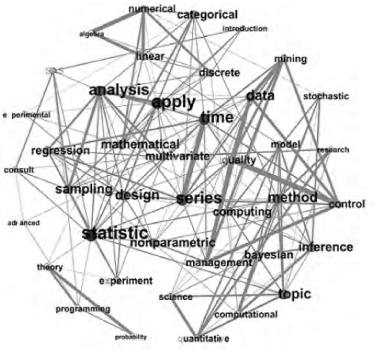


图4 应用统计项目全部课程词条网络结构图

2. 课程内容。在网络结构图的基础上，分别提取必修和选修课程的核心词来进一步探讨课程内容。应用统计硕士类项目必修和选修课程核心词条关联情况分别见图 5 图 6。图 5 图 6 中五边形内的单词是核心词条，指超过一半的该类项目的课程名称中出现了该词；椭圆里的是在一定特征相关度要求下选取的与核心词条关联性较大的一些关键词；两者连线上的数字代表词条相关性的大小，相关性越大说明国外高校开设此课程的数量越多，对连线上的节点进行组合即为项目设置的主要课程内容。从图 5 和图 6 可以看出，应用统计硕士类必修课程主要是围绕统计、分析、应用、方法、数据、概率和模型等基础内容来设置的，强调大数据分析的理论与方法；选修课程则是在必修课程的基础上增加了针对不同数据类型的建模方法，同时也强调提升学生在统计、分析和建模这三大方面的应用能力。

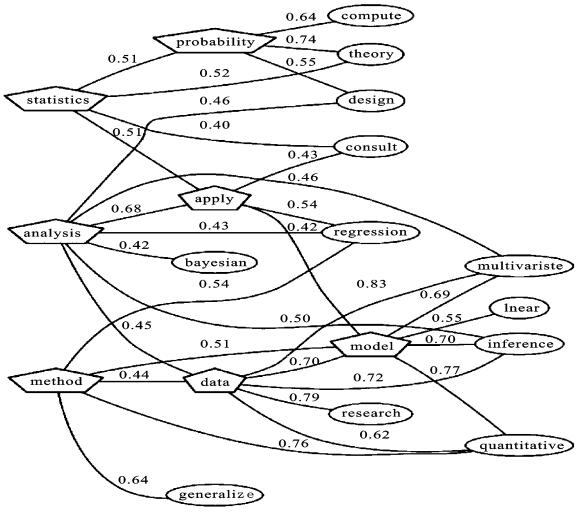


图 5　应用统计类项目必修课程核心词条相关度图

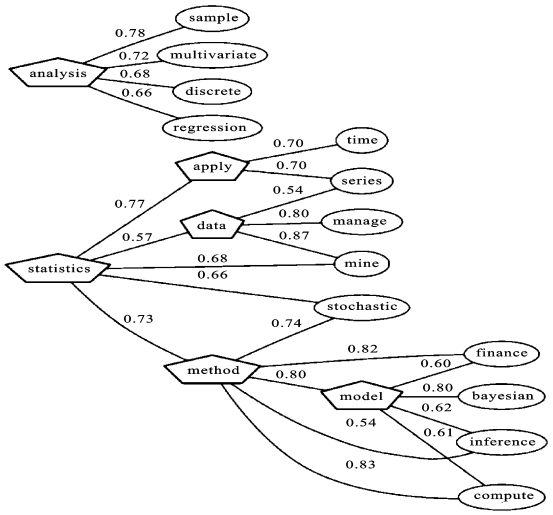


图 6　应用统计类项目选修课程核心词条相关度图

其他类型项目课程分析同应用统计硕士项目，其文本分析结果见表 2，在此不再赘述。

只有明确不同类型人才培养的核心才能更好地设置课程，对部分必修和选修课程核心词统计结果如图 7 所示。图 7 表明必修和选修课程都是以数据和分析为主要内容，不论是哪个方向的大数据项目都要求能够对该方向的数据进行分析；其次国外高校还注重管理和处理实际业务能力的培养，最大的差异是必修课程倾向于对数据挖掘知识的培养，选修课程更侧重对整体系统的设计和应用。

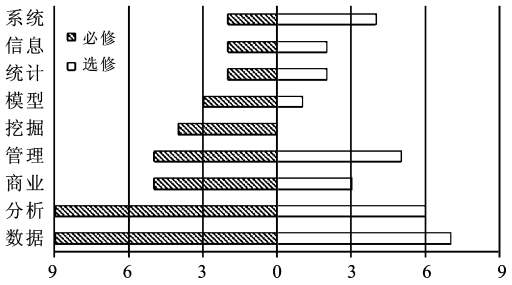


图 7　国外高校大数据各类项目必修和选修课程核心词条

四、国内高校大数据高端人才培养状况

(一)国内大数据机构布局

在“国家大数据战略”宏观背景下，近年来北京大学、中国人民大学、首都经济贸易大学、复旦大学、北京航空航天大学、贵州大学等高校和一些机构在大数据研究方面纷纷开展布局。据笔者统计，截至 2017 年 4 月有 26 所院校的 22 个大数据硕士项目已经开始实施或已提上日程，有 35 所院校的数据科学和大数据技术本科专业已经获教育部批准设立，但部分院校尚未开始招生，其中北京市各类大数据机构数量高达 12 个，北京市作为全国文化教育的中心占有大量的教育资源，在大数据高端人才培养方面同样表现突出；其次是以贵州为代表的西南地区对大数据的发展尤为重视，每年 5 月在贵州举行的大数据产业峰会更是引领着大数据的潮流。虽然中国近年来设立了许多大数据学院和研究中心，但目前大数据的覆盖面仍不广泛，那些即使已经设立了大数据机构的高校多数还处在课程筹备阶段，较少有全面开始招生。总体而言，目前中国在大数据教育方面的发展还处在一个初始阶段。

表 2　国外高校各类大数据项目课程内容表

编号	类型	核心词条	课程
1	应用统计	必修：统计 分析 应用 方法 数据 概率 模型	必修课程：线性模型(0.55) 统计理论(0.52) 时间序列分析(0.49) 数据建模(0.45) 应用回归分析(0.38) 统计计算(0.33) 定量分析方法(0.28)等
		选修：统计 分析 应用 数据 方法 模型	选修课程：数据挖掘(0.87) 数据分析方法与算法(0.83) 数据管理(0.80) 抽样分析(0.78) 应用统计分析(0.77) 离散数据分析(0.68) 随机过程(0.66) 贝叶斯统计(0.50) 金融统计模型(0.46)等
2	数据科学	必修：数据 分析 机器 挖掘 学习	必修课程：机器学习(0.74) 数据挖掘(0.59) 网络分析(0.47) 数据管理(0.47) 探索性数据分析(0.46) 统计推断(0.46) 数据库管理 (0.44) 数值优化(0.35) 广义回归模型(0.30) 实验设计(0.26)等
		选修：学习 数据 机器 系统 计算	选修课程：计算机系统概论(0.76) 数据可视化(0.73) 系统安全(0.57) 统计推断(0.47) 多变量分析(0.39) 生存分析(0.29) 统计质量控制(0.28)等
3	商业分析	必修：分析 数据 商业 管理 统计 模型 挖掘	必修课程：数据可视化(0.53) 商务智能(0.40) 数据挖掘(0.38) 数据分析(0.38) 商业预测(0.34) 金融管理(0.34) 数据模型与决策(0.32) 应用统计分析(0.32) 最优化理论与算法(0.31)等
		选修：分析 管理 数据 商业 市场	选修课程：文本分析(0.67) 战略管理(0.65) 时间序列分析(0.63) 电子商务(0.56) 管理会计(0.55) 市场调查与预测(0.43) 多元统计分析(0.29) 社交网络分析(0.18)等
4	商务智能	必修：数据 商业 分析 管理 智能 挖掘	必修课程：商务智能(0.74) 数据分析(0.70) 人工智能(0.66) 数据管理(0.49) 文本挖掘(0.43) 神经网络(0.40) 贝叶斯网络(0.38) 统计建模(0.33) 企业管理(0.26)等
		选修：管理 商业 分析	选修课程：决策支持系统(0.94) 算法分析(0.94) 战略营销管理(0.92) 商业分析(0.89) 信息管理(0.79) 企业风险管理(0.78) 高级数据库(0.77) 金融工程(0.53)等
5	健康医疗	必修：卫生 信息 系统 管理 分析 看护 数据 研究 保健	必修课程：医疗信息学(0.56) 数据挖掘(0.54) 流行病分析(0.45) 信息系统设计(0.45) 生物统计(0.41) 卫生管理信息系统(0.36) 数据库设计与开发(0.22) 探索性数据分析(0.18)等
		选修：卫生 管理 信息 保健	选修课程：医疗战略规划(0.65) 统计推断(0.63) 系统分析设计(0.61) 临床决策支持系统(0.60) 数据安全管理(0.52) 数据可视化(0.52) 试验设计(0.50) 管理决策(0.44)等
6	信息系统	必修：系统 数据 管理 信息 分析 商业 设计	必修课程：数据挖掘(0.74) 商务智能(0.57) 数据库设计(0.53) 数据分析与应用(0.44) 定量数据分析方法(0.42) 财务管理(0.39) 信息系统管理与战略(0.21)等
		选修：管理 信息 数据 系统	选修课程：信息安全(0.65) 系统分析与设计(0.61) 数据管理(0.52) 数据可视化(0.52) 业务流程建模(0.50) 操作系统原理 (0.47) 企业决策管理(0.44) 图像识别(0.33)等
7	MBA	必修：管理 商业 分析 数据 金融 战略 会计	必修课程：数据挖掘(0.80) 管理会计(0.75) 金融市场学(0.73) 市场营销战略(0.71) 商业经济(0.61) 金融管理(0.61) 经济发展战略(0.58) 业务分析(0.51)等
		选修：管理 数据 分析 商业	选修课程：信息管理学(0.92) 商务智能(0.87) 研发管理(0.85) 数据库系统(0.84) 预测分析(0.78) 经济学分析方法(0.75) 金融政策理论(0.50) 时间序列分析(0.45)等
8	博士类	必修：分析 数据 模型 线性	必修课程：决策模型(0.83) 最优化理论与算法 (0.83) 多元统计分析(0.49) 统计计算(0.48) 数据研究(0.41) 高级回归分析(0.35) 分布式并行计算(0.30)等
		选修：系统 计算 数据 分析	选修课程：计算机系统(0.98) 高级数据库系统(0.85) 人工智能(0.81) 模式识别(0.81) 自然语言处理(0.77) 知识发现与数据挖掘(0.65) 动态回归与决策系统(0.52)等
9	证书类	必修：数据 分析 商业 统计 挖掘	必修课程：数据挖掘(0.59) 应用回归分析 (0.55) 统计计算(0.53) 网络分析(0.37) 商务智能(0.36) 应用统计(0.36) 多元统计(0.24)等
		选修：分析 数据 统计 系统	选修课程：信息系统(0.85) 商业分析(0.75) 数据评估(0.74) 数据库设计与开发(0.70) 时间序列分析(0.58) 机器学习(0.42)等

(二)国内大数据硕士项目主要研究方向

国内大数据项目的文本挖掘结果显示，当前中国的大数据教研机构主要是以统计学科、计算机学科和以业务需求为依托，这三类教研机构的大数据高端人才培养途径不同。

1. 以统计学科为依托，致力于基础性数据挖掘、分析和建模。以北京大学、中国人民大学、中央财经大学、首都经济贸易大学和中国科学院大学联合共建的“大数据分析硕士培养协同创新平台”最为成熟。该平台旨在建成一个政产学研有机融合的协同创新平台，自 2014 年始现已成功开展了三届大数据方向的应用统计专业硕士的人才培养，至今已培养高端大数据人才 190 名。该培养方案中有 6 门必修课采用五校联合授课的形式，必修课程的重点内容为统计学和计算机学科的交叉部分，侧重于培养从大数据到价值的实践能力，选修课程则由各校分别开设。

2. 以计算机学科为依托，致力于工学设计、计算原理和数据存储及处理等。目前，北京航空航天大学、上海交通大学、中国科学院大学、西安电子科技大学、武汉大学、复旦大学、中国人民大学、山东大学等高校的软件学院都有与此相关的大数据工程硕士专业，这类专业课程体系一般包括公共基础课程和专业课程，专业课程涵盖大数据的基础设施、集成、存储、建模、管理、分析、挖掘以及安全等多方面。

3. 以业务需求为依托，致力于提供商业问题的

解决方案。典型代表有中央财经大学商学院开设的金融与大数据营销工商管理硕士专业，旨在适应“互联网＋”及大数据融合发展趋势，满足国家经济建设和社会发展对金融、市场、大数据的跨界复合型、应用创新型人才的需求；西南交通大学的金融大数据研究院针对本校学生开设了大数据方向专业，目前已经开设了四门研究生课程，考试成绩优良者将颁发西南交通大学数据科学证书；中国人民大学信息资源管理学院开设的两年制的大数据分析与应用方向的信息资源管理专业，采取理论与实践相结合、课堂讲授与自学相结合的方式，满足大专以上学历、勇于探索和勤于学习要求的即可申请；北京化工大学经济管理学院为适应中国现代工程事业发展对工程管理人才的迫切需求，开设了大数据技术与应用方向的工程管理硕士专业。

五、大数据高端人才市场需求与人才培养匹配度

如今，中国对大数据高端人才的需求越来越紧迫，而国内开设大数据项目的高校不多，以业务需求为依托的大数据项目则更少，因此亟需借鉴国外的人才培养经验来解决大数据人才短缺的问题。那么，国外的这种大数据高端人才培养途径是否符合中国的大数据人才市场需求呢？本文将根据当前大数据岗位的招聘信息来进行人才培养匹配程度的实证分析。

本文从中国最大招聘网站上爬取了符合条件的8 650条招聘信息，其结果显示招聘人数共计28 333人。考虑到公司发布的招聘信息会在一段时间内撤回，因此实际上市场对大数据人才的需求要更大。通过对大数据岗位工作地点分布情况统计发现，中国大数据岗位和大数据相关机构的分布基本是一致的，反映了市场的需求推动了大数据人才的培养。此外，大数据岗位的招聘信息中对于学历的要求不是很高，本科占较大比重，一方面是由于对相关工作经验提出了较高要求，另一方面也说明大数据高端专业人才的严重稀缺。

根据岗位职责并利用文本聚类方法，大数据相关职位大致可以分为系统、分析和管理三大类，具体又可细分为18小类，其中分析类职位的招聘人数占52.88％，系统类占27.46％、管理类占19.66％，大数据高端分析人才缺口最大。利用文本相似度原理对每类岗位的要求和7个大数据方向的课程内容进行对比，最终选取的相似性最大的大数据研究方向结果见表3所示。

表3　国内大数据岗位要求与国外大数据项目匹配表

类型	职位名称（占比）	岗位职责	项目匹配（匹配度）	职位名称（占比）	岗位职责	项目匹配（匹配度）
系统类	数据软件工程师（5.30％）	负责以Hadoop和Java为主的代码开发和软件实施	信息系统（0.59）	研发工程师（5.66％）	进行系统、算法等的设计、开发、优化	信息系统（0.63）
	大数据架构师（2.61％）	负责大数据应用架构设计以及应用产品规划	商务智能（0.67）	数据营运及维护（13.89％）	负责数据库日常监控、运营、维护、性能优化	信息系统（0.56）
分析类	数据统计分析师（1.33％）	对数据进行整理和分析并指导业务发展	应用统计（0.81）	数据挖掘分析师（5.68％）	对大量数据进行挖掘并发现潜在关系	数据科学（0.79）
	业务数据分析师（31.23％）	对业务数据分析处理并做出业务评估和预测	商业分析（0.62）	数据科学家（0.26％）	设计大数据挖掘模型的开发、评估、部署、优化	数据科学（0.77）
	机器学习（1.23％）	负责机器学习算法的研究及平台的搭建	数据科学（0.62）	人工智能（0.46％）	研发智能领域语义分析算法，解决智能搜索与知识发现的问题	商务智能（0.22）
	大数据工程师（6.26％）	设计实施大数据平台并制定系统技术架构	信息系统（0.30）	数据可视化（0.32％）	将数据以图形形式展示并探究数据间的关系	商业分析（0.33）
	数据采集（0.84％）	通过多渠道搜集所需数据	应用统计（0.45）	数据处理（2.13％）	对原始数据库进行预处理和分类	应用统计（0.63）
	数据建模（1.09％）	利用算法和工具对数据进行模型设计和开发	应用统计（0.72）	其他数据专员（2.05％）	包括数据推广专员、激励专员、营销专员等	商业分析（0.37）
管理类	数据顾问（1.19％）	负责数据技术咨询业务并维护客户关系	数据科学（0.54）	数据产品经理（18.47％）	负责产品推进过程中的协调沟通并带领团队开展相关业务	MBA（0.87）

通过表3内容对比可以看出：信息系统方向的项目基本满足系统类的岗位要求；分析类项目涉及的研究方向较多，主要包括应用统计、商业分析、商务智能和数据科学；MBA方向则是与管理类职位最为匹配。借鉴国外高校大数据人才培养的先进经验对中国大数据人才的培养很有意义，但是还有部分职位如人工智能作为当前关注度最高的职位之一，除对学历提出更高的要求外，由于国内外各方面技术发展程度不同等原因导致匹配度很低，反映了国内大数据人才市场需求与国外人才培养有所偏离，因此在借鉴国外对大数据专业人才培养的理念时要考虑中国国情和市场需求，综合多方面的因素

制定切实可行的培养方案。

六、总结与建议

相比于国外完善的大数据人才培养项目体系，中国的人才培养发展相对滞后，然而解决大数据高端人才短缺的问题已经刻不容缓，借鉴国外的先进经验是缩小大数据人才供需差额的一种有效途径。基于本文研究提出以下建议：

第一，明确大数据的培养目标。国外高校大数据项目的部分课程主要围绕数据、分析、商业和管理这些内容展开，其余部分课程不同类型的大数据项目间差异较大。课程内容是基于培养目标而规划的，明确培养目标才能更有针对性地设置课程。

第二，完善大数据培养方式。在国外的大数据项目中证书类数目仅次于商业分析硕士类，证书类项目学费低、时间短和效果快的优势很受在职人士的青睐，将这些有一定基础和经验的专业人士培养成为大数据高端人才，能够快速填补现在市场上大数据人才的缺口。

第三，增加大数据专业的方向。目前，中国大数据的研究方向主要是以统计和计算机为背景，而以业务为导向的项目很少，涉足其他方向的更是稀缺，因市场需求在不同大数据应用领域各具特色，全面培养大数据高端人才是大势所趋。

第四，结合市场需求，借鉴国外先进经验，有针对性地培养人才。国外高校大数据人才培养的先进经验对中国大数据人才的培养很有意义，但国内人才市场需求与国外人才培养途径有差异，国内高校在借鉴国外经验时要考虑中国的国情，各方努力进一步协调共同制定大数据人才培养方案，培养有中国特色的大数据高端人才。

第五，运用大数据分析技术制定培养方案。将大数据分析技术的思路和整套方法应用在国内高校的大数据高端人才培养途径的分析上，能够快速全面地获取先进的人才培养经验，其他学科人才培养方案的制定也可借鉴这种模式。

参考文献：

[1] 朱建平，李秋雅. 大数据对大学教学的影响[J]. 中国大学教学，2014(9).
[2] 何海地. 美国大数据专业硕士研究生教育的背景、现状、特色与启示——全美23所知名大学数据分析硕士课程网站及相关信息分析研究[J]. 图书与情报，2014(2).
[3] 阮敬，陈涛. 大数据背景下的应用统计专业硕士人才培养模式研究[J]. 统计与管理，2015(8).
[4] 迪莉娅. 高校数据科学专业硕士课程设置研究[J]. 教学研究，2014(6).
[5] 祝丹，陈立双. 大数据驱动下统计学人才培养模式研究[J]. 统计与信息论坛，2016(12).
[6] 李保利，陈玉忠，俞士汶. 信息抽取研究综述[J]. 计算机工程与应用，2003(10).
[7] Gerard Salton，Wong A，Yang C S. A Vector Space Model for Automatic Indexing[J]. Communications of the ACM，1975(18).

Experiences and Inspirations of Cultivating Big Data Talents in Foreign Countries: Based on Text Mining

RUAN Jing，LIU Hong-jing，JI Hong

(School of Statistics，Capital University of Economics and Business，Beijing 100070，China)

Abstract: This paper utilizes big data mindset and semi-structured text data analysis method to directions for big data high-end personnel training and the setting of courses，credits，school system for big data high-end talents，providing in-depth analysis on the corresponding needs for trainers to give recommendations on how to cultivate big data high-end talents in China from the perspective of supply-side reform.

Key words: big data; big data talent; talent development; web text mining

(责任编辑：郭诗梦)