基于大数据分析的大学生创新创业主要影响因素探究

谢晓默，林敏

(福州大学数学与计算机科学学院，福建福州，350116)

[摘要] 在收集海量创新创业数据基础上，利用大数据分析手段，从大学生创新创业数据获取层、大学生创新创业数据分析层、大学生创新创业数据应用层等三个层次，搭建大学生创新创业的大数据分析模型，尝试用定量分析的方法探究大学生创新创业的主要影响因素的具体占比，更有针对性指导大学生创新创业实践，为推动大众创新、万众创业提供参考依据。

[关键词] 大数据分析；大学生；创新创业；影响因素

[中图分类号] G641 [文献标识码] A [文章编号] 1674-893X(2018)06-0049-05

一、前言

当下大数据被广泛运用在社会各个领域，悄然改变着人们的生产方式和生活方式。哪些主客观因素影响大学生创新创业，成为当前高校创业教育的重要课题。综观国内外研究情况，笔者发现国内关于创新创业影响因素的研究起步较晚，研究成果不是很多，研究内容更多体现在微观层面；传统研究方法多是通过问卷调查的形式，普遍存在分析方法单一，主观性较强，效度、信度欠佳等不足[1]。

鉴于此，本研究试图利用大数据分析具备海量的数据来源、高效的分析速率、准确的结果判断等特点[2]，搭建大学生创新创业的大数据分析模型，对当前在校大学生创新创业影响因素展开实证研究。

二、大学生创新创业大数据分析模型构建

随着信息技术的高速发展，以微博、微信、门户网站等为代表的互联网新媒体为大学生创新创业核心影响因素的分析带来可能性[3]。从互联网丰富的大学生创新创业数据中提取影响因素，尤其探究对大学生创新创业影响的主要因素具有十分重大的现实意义。为此，本文从海量异构创新创业数据入手，构建大学生创新创业的大数据分析模型，通过对大学生创新创业数据的采集、存储、分析，探究大学生创新创业的影响因素。

该模型如图 1 所示，分为三个层次，包括大学生创新创业数据获取层、大学生创新创业数据分析层、大学生创新创业数据应用层。具体介绍如下：

(一) 大学生创新创业数据获取层

主要包括采集清理和存储两个部分。

(1) 数据采集清理。数据的采集是大学生创新创业大数据分析首先需要解决的基础性工作。网络数据潜在分布广、海量庞杂、多源异构，与此同时，网络中 90%的数据存在于深网(例如微博、微信、电子期刊等)中，常规采集手段的覆盖率无法满足创新创业大数据分析的需求。

针对互联网数据特点以及常规采集手段存在的以上问题，本文构建了一款基于 THRIFT 通信框架的分布式创新创业数据采集方法。首先，针对数据泛在分布于互联网及社交媒体的问题，构建基于 THRIFT 通信框架的分布式架构，同时通过嵌入创新创业相关主题和种子 URL 定制、采集参数配置等模块，实现可定制采集；其次，针对深网数据，本文采用模拟用户行为以及模拟登录来爬取相应信息；然后，针对数据动态增长的问题，本文采用基于 BLOOM 过滤器的判重方法，实现增量采集，使得日均采集量提升至单机的 10 倍以上；最后，针对海量庞杂和多源异构问题，本文建立了基于网页文本结构的统一抽取框架，框架针对现有互联网

[收稿日期] 2018-02-08；[修回日期] 2018-12-11

[基金项目] 福州大学教育管理研究专项课题研究成果“大数据在高校创新创业教育中的理论探讨与实践”(16SKZ30)

[作者简介] 谢晓默(1962—)，男，福建古田人，福州大学副研究员，主要研究方向：思想政治教育；林敏(1990—)，女，福建福清人，福州大学讲师，主要研究方向：思想政治理论与实践，联系邮箱：352914127@qq.com

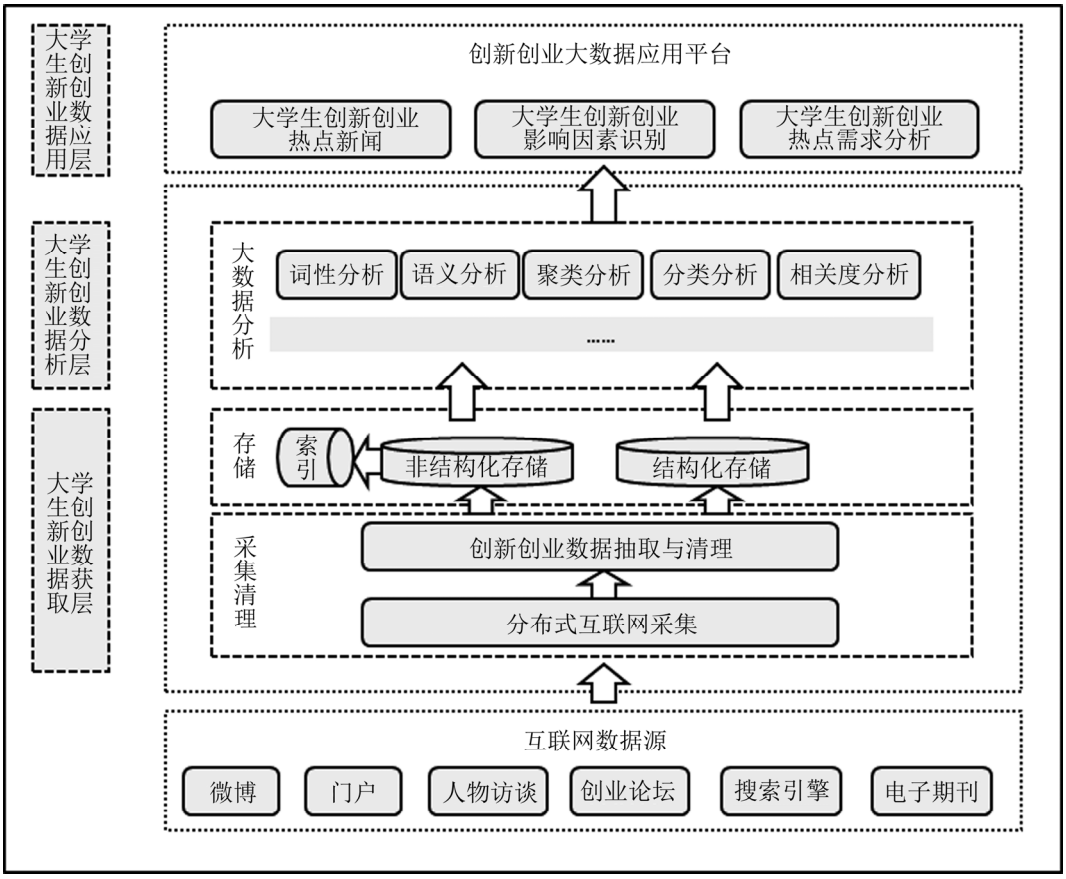

图 1 大学生创新创业大数据分析模型

文本特点，将其分为长、短文本两部分，并针对长、短文本的各自特点，采用不同的基于 DOM 树结构的文本抽取模型。

(2) 创新创业数据存储。目前，还未存在公开的针对大学生创新创业领域的标准数据构建方法。以往单一数据库只能存储结构化数据，其无法满足对非结构化数据和半结构化数据(如 HTML 等)的存储需求。因此，本文尝试创建一个创新创业数据的统一表达方式。一方面，针对结构化创新创业数据(如信息的倾向性、信息所在领域等)，本文利用关系数据库进行存储，方便后续应用层的查看；另一方面，对非结构化和半结构化数据，本文利用基于 XML 的多源异构数据表示方法对抽取出的创新创业数据进行统一表达。

(二) 大学生创新创业数据分析层

主要包括创新创业数据的词性分析、语义分析，聚类分析、分类分析以及相关度分析。

(1) 创新创业数据词性分析、语义分析。针对互联网创新创业数据难以统一表达的问题，该部分主要采用 WORD2VEC 技术对创新创业数据进行词性分析、语义分析，将其映射到统一的特征空间中，该部分分析的主要目的是从海量的创新创业数据中挖掘出影响大学生创新创业的因素。

(2) 创新创业影响因素聚类分析。该部分分析主要采用 SINGLEPASS 聚类方法对 2.2.1 的数据聚集出大学生创新创业的主要影响因素。

(3) 创新创业影响因素分类分析。基于 2.2.2 挖掘出的影响因素对大学生创新创业的影响程度存在较大差异，因此该部分采用基于互信息的特征选择方法，对创新创业的主要影响因素进行特征选择，由此将创新创业的影响因素按其影响程度大小进行有序排列。

(三) 大学生创新创业数据应用层

该层主要功能是对分析层的结果进行可视化展示。应用场景包括大学生创新创业热点需求分析、大学生创新创业项目跟踪、大学生创新创业影响因素分析等。大学生创新创业数据应用层涵盖的领域广，内容丰富，前景可观。考虑到本文重点研

究大学生创新创业影响因素分析，故而针对大学生创新创业的其他应用方向暂不做展开。

三、基于大数据分析的大学生创新创业主要影响因素

从本文构建的平台出发，通过采集存储互联网中海量的创新创业信息，利用大数据分析技术，探究影响大学生创新创业的影响因素，根据影响因素的大小进行排序。

(一) 创新创业数据的采集与抽取

底层数据的好坏关系到大数据分析质量的高低，这要求采集的互联网数据源覆盖广，实时性高，数据量大。

为此，在数据源的选取上，本文利用互联网分布式采集系统，从搜索引擎、门户网站、微博、微信、论坛、电子报纸、电子期刊等媒介中采集信息。其中，搜索引擎涵盖当下主流引擎“百度搜索”“搜狗搜索”等；门户网站采集涵盖主流大门户“新浪网”“凤凰网”，各创新创业相关门户网站如“中青在线-创家”以及各类名人或商界访谈门户网站如“极客网访谈”等；微博数据来源于时下热门社交网络平台“新浪微博”；贴吧采集目标为主流贴吧提供商“百度”“天涯”和“猫扑”等；电子期刊采集范围为近五年来各期刊会议所发表的与创新或创业因素相关的论文。数据来源基本达到上述要求。具体如表 1 所示。

表 1　数据源部分列表

类别	数据源
搜索引擎	百度搜索、新浪搜索、搜狗搜索、360 搜索、环球搜索……
门户网站	新浪网、凤凰网、极客访谈、大学生在线、腾讯网大学生创业……
微博	新浪微博、腾讯微博
贴吧	清华、北大等知名学校百度贴吧、天涯社区、猫扑社区……
电子期刊	知网数据库、万方数据库、维普数据库……

为了让多源异构信息结构化成大数据分析方法所能利用的信息，针对门户网站、搜索引擎等长文本网页内容，采用基于 DOM 树的文本密度算法进行信息抽取；针对贴吧、微博等短文本内容，文本采用基于 DOM 树层次特征的多记录网页抽取算法进行网页源码的文字识别，基于以上两项技术，多源异构网页信息的识别率高于 90%，能够保证网页关键信息不遗漏。最后，在数据分类上，以长文本、微博、贴吧、微信、期刊论文为分类依据，方便接下去的大数据分析进行有针对性的因素识别。综上，本文对采集到的数据进行了统计，结果如图 2 所示。

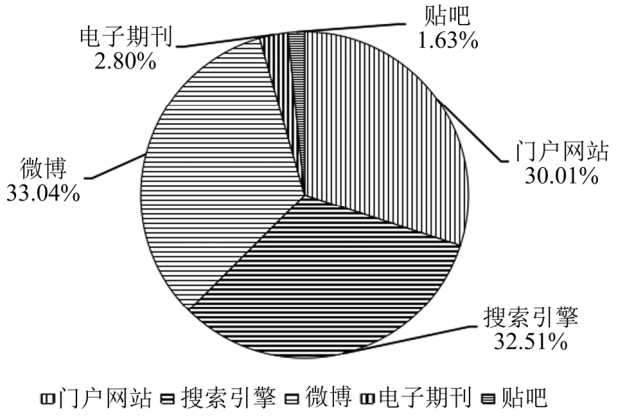

图 2　数据源饼状图

(二) 大学生创新创业主要影响因素分析

通过观察数据可知，互联网数据中蕴含大量大学生创新创业影响因素，同时，不同因素间存在一定的联系，例如“创新教育”与“创业教育”，两者通常描述一类事物。基于以上观察结果，本文提出了一种基于 WORD2VEC 的 SINGLEPASS 创新创业因素识别方法，该方法首先通过 WORD2VEC 内在向量化文本，然后使用 SINGLEPASS 聚类方法聚合同类影响因子，以更加精确化描述影响因素以及减少冗余度，最后采用线性回归的方法对各影响因素进行权重计算，并依权数大小进行排序。

1. 基于 WORD2VEC 的 SINGLEPASS 创新创业因素识别

WORD2VEC 可以把对文本内容的处理简化为向量空间中的向量运算，通过计算出向量空间上的相似度，来表示文本语义上的相似度。WORD2VEC 因其效率高、效果好，被广泛应用在语义分析之中。同时，WORD2VEC 适合于一个序列的数据，在序列局部数据间存在着很强的关联。因此，针对本文数据中各创新创业因素存在的相关性，WORD2VEC 能够较好地将文本数据进行向量化。SINGLEPASS 算法是一种流式的聚类算法，每个

数据只会参与一次样本聚类，聚类结果与数据的先后顺序有一定的依赖关系。SINGLEPASS 算法是一种增量算法，适合对流数据进行挖掘，而且算法的时间效率高。因此，针对本文增量采集的创新创业数据，SINGLEPASS 能够快速从各种创新创业因素中聚合得到相应的因素类簇。

基于以上两个方法，本文首先将采集到的创新创业影响因素相关文本数据分词，然后过滤停用词、常用词等得到候选词组集合；然后用 WORD2VEC 计算候选词组集合中每一个词组的词向量；接着利用 SINGLEPASS 聚类方法对每个词组进行聚类，计算结果如表 2 所示。其中，簇类标签由人工给出，本文首先挑选三名有标注经验并且有创新创业相关经验的人员分别对这些类簇打上

表 2 部分类簇关键词

类簇编号	影响因子类别	类内词组
1	创新创业教育	创新教育，创业教育，创业分类教育，创业教育制度，创业教育教师队伍，创业教育程度，创新创业教育投入，创新创业教育产出
2	学校教育	大学生创业意向结构自身背景影响要素，大学生创业意向结构的外在环境影响，产业结构，来自大学自身的影响因素，来自研究型大学的影响因素，学校因素，学生因素，研究影响大学生自主创业意向，创业者自身的因素，产学研
3	创新创业能力	创新能力，创业能力，创新，创新性，创业能力素质,
4	政府政策	政策扶持，政策支持，政策资金扶持，政策支持力度，政策制度，政府政策
5	自我效能感	创业自我效能感，自我效能感，创业自我效能感和创业意向，创业效能感与创业人格，个体创业自我效能感，创业自我效能
6	就业创业模式	商业模式创新，创业就业模式，创业模式，体制创新
7	社会资源	社会资源，资金资源，社会资本，资金，企业社会资本
8	资金	创新资金，创新资产，创业资金，金融创新，创新资源，金融理念创新
……		……

簇类标签，然后利用投票的方式得到簇类名称。

2. 基于线性回归的创新创业影响因素分析

线性回归分析方法是确定两种或两种以上变量间相互之间的相关关系的一种分析方法，其广泛应用于大数据分类计算、特征选择等分析领域。因此，利用线性回归的方法能够较好地满足创新创业影响因素分析的需要。

首先，本文对采集到的数据进行人工筛选和分类，一类为创新创业相关数据，另一类为非创新创业数据。

接着，利用 3.2.1 得到的结果，将每一个类簇当作一个特征，对所有采集到的数据进行特征向量化，本文定义每篇文档的特征向量如下：

$$d=[\delta(1),\delta(2),\delta(3),\cdots,\delta(n)]=y$$

其中，d 表示一篇文档，$\delta(i)$ 是(i=1,2,3,4…n)一个指示函数，它表示一篇文档是否包含了表 2 类簇编号中第 i 个类簇的某一个影响因素关键词，如果包含则 $\delta(i)$=1，不包含 $\delta(i)$=0。y 表示每一篇文档是否属于创新创业新闻，如果属于，则 y=1，反之，则 y=0。

经过上述步骤，所有的文档数据就用特征向量来表示，利用线性回归的方法对所有文档的特征向量进行分析，线性回归的公式如下：

$$h_\theta(x)=\theta_0+\theta_1x_1+\theta_2x_2+\cdots+\theta_nx_n$$

其中 $h_\theta(x)$ 表示文档是否是创新创业新闻数据，θ_i (i=1,2,3,4…n) 表示各特征值的相关权重，x_i (i=1,2,3,4…n)表示影响因子特征的数值，对应上述文档向量中的 $\delta(i)$ 值。对于一个线性回归的训练过程，需要一个评价函数评估回归结果的好坏，因此，有如下损失函数：

$$J(\theta)=\frac{1}{2}\sum_{i=1}^{n}(h_\theta(x_i)-y_i)^2$$

$$\min_\theta J(\theta)$$

公式中的 i 表示第 i 个文档特征向量，一共有 n 个文档，θ 为特征权重向量。本文采用梯度下降的方法调整 θ 中每个特征权重值使得 $J(\theta)$ 取得最小值，迭代过程当 $J(\theta)$ 收敛或变化幅度很小时，则停止迭代，得到每一个特征值的权重。根据线性回归的最终结果，特征因素的重要性如表 3 所示

表 3　部分类簇占比情况

影响因素类别	影响权重占比(%)	类内词组
机会	4.52	创业机会，机会创新性，机会识别效能，创业机会感知，创业机会和创业资源
自我效能感	4.31	创业自我效能感，自我效能感，创业自我效能，创业自我效能感和创业意向……
创新创业环境	2.64	创业环境, 创新创业环境，全民创业氛围，创业氛围，社会评价
师资力量	2.47	创业师资队伍，创业教育教师队伍，师资力量，师资质量，教师素质，教师因素，教学能力
人格特质	2.43	人格特质，创业大学生人格特质，好奇，独立思考，追求完美等创新特质，冒险，意志坚定……
资源	2.43	创业资源，政策资源，场地资源，管理资源，外部资源因素
年龄	2.09	年龄
文化因素	2.07	区域文化因素，文化因素，地区文化与政策，地区文化，经济文化，文化氛围，服务文化……
就业创业模式	2.07	商业模式创新，创业就业模式，创业模式，体制创新
……	……	……

(因篇幅有限，本文只列出占比排名前十的特征因素)。

由表 3 结果可知，机会、自我效能感、创新创业环境、师资力量、人格特征等对大学生创新创业都存在着影响，其影响随着占比比例的减小而相应减弱。

综上，本文的研究得出了大学生创新创业主要影响因素的具体占比，这将更有针对性地指导大学生创新创业实践。

参考文献：

[1] 丛明,寇福生,王诗白.“互联网+”背景下的研究生创新创业能力培养研究与实践[J].时代教育,2017(09):44-45.

[2] 郑石明.大数据驱动创新创业教育变革:理论与实践[J].清华大学教育研究,2016(03):65-73.

[3] 蓝荣聪,陈永福.大数据视域下大学生创新能力培养的思考[J].思想教育研究, 2014(11):70-72.

[编辑：何彩章]