基于大数据技术的大学生热点问题预测研究

唐青　王俊凯　泮佳伟　王雨泽
（衢州学院　浙江衢州）

摘要：随着经济的日渐发展，社会逐渐向现代化、网络化、科学化转变，大数据逐渐深入人们的生活，对社会上各个行业带来了一定的冲击，尤其是教育行业大数据的冲击为各行各业带来变革，越来越多的大学生在网络上发表言论。本系统是基于大数据技术的思想，将爬虫技术与 Hadoop 框架结合并应用系统中，本项目利用 Apriori、二次指数平滑法等算法，并且进行分词和词频统计，设计一套可快速、准确地掌握和预测大学生关注的热点问题的系统。

关键词：热点预测；大数据分析；Hadoop 框架

1.引言

大学生关注的热点问题反应了大学生的思想状况以及价值观念，为了客观、全面、深入地了解当代大学生关注的热点问题，把握其思想状况及其价值观念，国内外对大学生热点问题的研究从未停歇并且高度重视。目前，对于大学生热点问题的研究大多数采用抽样调查、问卷调查等方法，这些方法存在一些不足：效率不高，研究对象有限，且易受地域差异和经济差异等因素的局限。在此背景下，提出了一种新的技术方法，即利用大数据技术分析 与预测大学生热点问题，该方法有三个方面的优点 1. 耗时少，热点不易过时；2. 研究成本低；3. 研究对象广泛，研究对象遍布全国各地。其中，大数据技术指的是从各种各样的巨型数据中快速 获得有价值信息的技术，该技术已经渗透各行各业。

2.国内外研究现状

在国外，对于社交媒体(Social media)数据挖掘与分析方面，预测问题的基本表现形式是基于用户对 UGC 内容的短期(reference date)反馈（即短期流行度 N(v, Tr)）来预测它们未来(target date)的流行度(N(v, Tr, Tt))，这里的流行度可以简单地定义为 UGC 内容的查看数(the number of views)，Szabo and Huberman(Communic. of ACM, 2010) 首先提出了简单的预测模型，该模型只有一个参数 alpha，在算出 alpha 之后，未来流行度 N(v,Tr,Tt)=alpha*N(v,Tr)。该模型简单易于实现，但是没有考 reference date 和 target date 之间流行度的变化趋势。在此基础上，Pinto, Almeida and Goncalves (Proc.of WSDM,2013)考虑了流行度变化趋势，提出了多元回归预测模型(Multivariate Linear,ML)，并且在准确度不断地在提高中。

在国内，对于学生校园行为分析数据包括结构化和非结构化的数据两大类，结构化数据可以通过“一卡通”、网络监控、教务、考勤等校园部署的信息系统进行数据采集和转换。非结构化、半结构化数据大部分来源于互联网、社区论坛等，需要利用数据采集平台，运用云化 ETL 工具、流数据处理、网络爬虫等工具进行采集，针对性的学生行为大数据可视化模型刚刚起步。

3.分布式数据爬取

本系统数据获取来源：本系统数据来源采用 Python 中的网络爬虫技术，爬虫是按照一定的规则，自动地抓取万维网信息的程序或者脚本，本质就是模拟浏览器打开网页，获取网页中我们想要的那部分数据。

HDFS 分布式文件系统和 MapReduce 分布式计算框架是 Hadoop 的核心内容，MapReduce 处于 HDFs 的上一层，负责海量数据的分布式计算，图 2 展示了 MapReduce 工作原理：首先将任务分割成若干个小任务，发送到各个计算节点，经由计算节点计算后将计算结果反馈，最后迅速整合形成答案。在整个系统中，无论是数据采集任务、数据存储任务、数据分析任务还是其他，只要把将要执行的任务转化为 MapReduce 任务，就可以实现分布式计算，从而大大提高计算效率。

4、大学生热点问题分析与预测框架设计

为有效地、快速地、精准地挖掘大学生言论潜在价值，建立了大学生热点问题研究框架，该框架基于 Hadoop 平台，主要包括四个层面，分别是数据采集层、数据处理层、分析预测层和结果展示层。

数据采集层：利用网络爬虫在各大高校贴吧、论坛和微博爬行，实现自动辨别网址，深度延伸智能化，全面采集大学生发表的言论极其个人信息。

数据处理层：建立词典库，参照词典库去掉无用的、无意义的字词、或标点符号。然后经过分类、聚类、关联分析等手段对数据进行综合性处理，尽量避免提取关键字所带来的偏差。

分析预测层：分析预测层对数据进行综合性分析与预测，重点挖掘热点与热点之间存在的关联性，并且预测热点的曲线分布。

结果展示层：结果可视化展示，通过可视化图标清晰展示数据分析结果，为数据的预测分析提供数据支撑。

5、数据处理

数据采集层采集下来的数据必须经过数据分析和处理技术，去掉一些多余的、无用的数据，才能更好地挖掘潜在

价值。

传输到数据处理层的数据首先经过数据清洗，清洗掉大量噪声、冗余和重复的数据，然后将经过清洗的数据递交给数据整合模块进行数据整合。数据整合以词典库和决策树为依据，对这些数据分门别类，不符合数据库表格模式的数据将进行二次清洗。

6、预测与结果分析

预测分析是大学生热点问题研究的重要部分，作用是利用预测算法预测某个热点的发展趋势。用于预测分析的算法较多，经过多次探讨，大学生热点的预测分析算法采用了二次指数平滑法。二次指数平滑法可以消除时间序列的偶然性变动，提高近期数据在预测中的重要程度，并且具有计算简单、样本要求量较少、适应性较强、结果较稳定等优点。

以大学生一年内提及游戏的次数为例，预测游戏这个热点的趋势走势。本次二次指数平滑初值给出：S(1)0=s(2)0=(Yl+Y2+Y3+Y4+Y5)/5.0=3884.8，通过计算选取使均方误差平方和(MSE)最小的平滑常数为最佳平滑常数，得平滑常数仪=0.52，得到了预测模型 YT=at+btT，T=1,2,3,…,9，YT 为未来月份 T 的预测值。

7、结束语

本系统可为分析大学生热点问题等关于调查分析系统的建设提供一定的参考价值，填补了有关大学生舆情预测，思想状况调查的空白，本系统通过构建互联网下的大数据分析平台，更及时地了解与预测分析大学生的热点问题，使教育机构、社会有明确目标地引导大学生成长，培养全面发展的青年人。

参考文献：

[1]方新丽.浅析数据挖掘技术在计算机审计中的应用[J].电脑知识与技术，2013.9(15)：3445-3446.

[2]张良将.基于Hadoop平台的海量数字图像数据挖掘的研究[D].上海交通大学，2013

项目基金：2020 年衢州学院大学生科技创新项目（No.Q20X034）国家大学生科技创新项目（202011488025，202011488026）。2019 年衢州学院大学生科技创新项目（Q19X009）