

基于 Hadoop 高职院校毕业生就业监测预警的决策支持系统构建

杨华¹ 李骏¹ 何文静¹ 雍玲燕² 张鹏¹

1. 南充职业技术学院 四川南充 637000

2. 南充市白塔中学校 四川南充 637000

摘要: 以大数据、云计算为代表的新兴信息技术的兴起与发展,给社会带来了深刻变革,也为高职院校改革提供了有力的数据支撑,通过大数据分析可以及早发现社会人才知识需求、发展趋势,也可以通过监测高职院校毕业生就业情况分析出高职院校在人才培养过程中存在的问题,促使高职院校不断改革,完善专业设置、教学计划,从而借助大数据实现高职院校的持续健康发展。因此,设计一套科学的基于大数据的高职院校就业监测和决策支持系统,可以为高职院校专业设置、人才培养目标设定、教学计划制定、就业引导、合理配置大学生资源提供可靠支持。

关键字: 高职院校, 毕业生, 就业监测, 预警, 决策支持系统

DOI:10.69979/3029-2735.24.3.016

1. 引言

近年来,随着互联网技术和大数据的迅猛发展,各行业对数据的需求日益增长。高职院校作为培养技术技能人才的重要基地,其毕业生的就业情况不仅关乎学生个人的发展,也影响着学校的声誉和社会的稳定。然而,毕业生就业市场瞬息万变,如何及时掌握毕业生的就业动态并进行有效预警,已成为高职院校亟待解决的问题。

传统的就业监测方法通常依赖于人工统计和定期调查,效率低下且难以实时反映就业市场的变化。而基于 Hadoop 的大数据技术,采集高职院校毕业生就业相关信息等数据,面向大学生就业领域重点问题,研究建立大数据融合的分析方法和模型库,开发大学生就业相关分析与决策支持系统,全面动态反映高职学生市场供需状况,以期通过系统应用,实现对高职毕业生就业状况动态监测与分析评估,形成系列分析报告,有效支撑高职院校专业设置、人才培养目标设定、教学计划制定、就业引导、合理配置大学生资源和毕业生就业政策制定等。为此,本文提出基于 Hadoop 的高职院校毕业生就业监测预警决策支持系统的构建方案,通过大数据技术实现对就业信息的实时监测和预警。

2. Hadoop 技术概述

2.1 概述

Hadoop 是一个由 Apache 软件基金会开发的分布式计算框架,最初设计用于处理大规模数据集的存储和分析。其核心组件包括 Hadoop 分布式文件系统(HDFS)和 Hadoop YARN(Yet Another Resource Negotiator)。HDFS 通过将大规模数据集分布式存储在多个节点上,提供高容错性和高可靠性;而 YARN 则负责集群中的资源调度和任务管理,支持多种计算框架如 MapReduce、Spark 等并行运行。其可扩展性,通过添加更多计算节点来扩展处理能力,满足大规模数据处理需求;容错性、高性能即并行计算方式,高效处理大规模数据;低成本高效益使得该技术在系统设计部署上具有很大优势。

基于 Hadoop 的高职院校毕业生就业监测预警决策支持系统,通过集成多个技术组件,实现就业数据的全面采集、高效处理与智能分析,为教育机构、政府部门及毕业生提供决策支持。

2.2 关键技术

基于 Hadoop 高职院校毕业生就业监测预警的决策支持系统其关键技术如下:

Hadoop 生态系统: 利用 HDFS 进行数据存储,使用 MapReduce/Spark 进行数据处理。

数据采集框架: 使用 Apache Flume 或 Kafka 进行数据流的实时采集。

数据分析算法: 应用机器学习算法对就业数据进行

深入分析。

3. 系统设计

3.1 系统需求分析

在设计系统时，我们主要从功能和性能需求上考虑：

功能上必须具备：数据采集，即实时采集毕业生的就业信息，包括岗位、薪资、行业等；数据存储，利用 Hadoop HDFS 存储大规模就业数据；数据处理与分析，使用 MapReduce 或 Spark 对数据进行处理，分析毕业生就业趋势；预警机制，根据设定的预警规则，及时向相关部门发出预警信息；可视化展示，通过数据可视化工具展示分析结果，便于决策者理解数据这些功能。

在性能需求上要求系统：高并发处理能力，能够处理大量并发数据请求；数据存储能力，支持存储 TB 级别的就业数据；实时性：系统能够实时更新数据并及时反馈分析结果。

3.2 系统架构及功能模块设计

系统架构及功能模块主要分为以下几个部分（如图所示）：



数据源层：包括毕业生就业信息采集平台、第三方就业数据接口等，如高职院校的就业信息数据库、企业招聘数据、社交媒体就业信息等；数据采集工具可以使用 Apache Sqoop、Flume 等工具，实现结构化数据和非结构化数据的采集。

数据存储层：基于 Hadoop HDFS 进行数据存储。

数据处理层：使用 MapReduce/Spark 进行数据处理和分析，提取就业相关的关键指标，如通过编写 MapReduce 作业，对数据进行清洗和统计。例如，计算毕业生就业率的 MapReduce 作业如下：

```
public class EmploymentRateCalculator {
```

```
    public static class Map extends Mapper<LongWritable, Text, Text, IntWritable> {
        public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {
            String[] fields = value.toString().split(",");
            String status = fields[3]; // 假设状态在第四列
            context.write(new Text(status), new IntWritable(1));
        }
    }

    public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable> {
        public void reduce(Text key, Iterable<IntWritable> values, Context context) throws IOException, InterruptedException {
            int sum = 0;
            for (IntWritable val : values) {
                sum += val.get();
            }
            context.write(key, new IntWritable(sum));
        }
    }
}
```

同时应用 Apache Hive 进行 SQL 查询分析，查询监测就业状态：

```
SELECT profession,
COUNT(CASE WHEN employment_status =
'employed' THEN 1 END) AS employed_count,
COUNT(*) AS total_count,
(COUNT(CASE WHEN employment_status =
'employed' THEN 1 END) / COUNT(*)) * 100 AS
employment_rate
```

```
FROM graduate_employment
GROUP BY profession;
```

使用 Apache Pig 进行复杂的数据流处理，利用机器学习算法（如逻辑回归、决策树等）进行模型训练。

```
from pyspark.ml.classification import
LogisticRegression
# 划分训练集和测试集
train_data, test_data =
```

```
final_data.randomSplit([0.7, 0.3])
# 构建模型
lr = LogisticRegression(featuresCol='features',
labelCol='label')
model = lr.fit(train_data)
# 模型评估
test_results = model.evaluate(test_data)
print(f"Accuracy: {test_results.accuracy}")
达到就业趋势预测, 如就业率、就业薪资分布、就
业行业分布等, 实现就业趋势预测模型, 提供未来就业
市场的预测分析。
同时在该模块上我们提供预警模块: 设定就业率、
就业质量等关键指标的预警阈值, 当实际数据超过或低
于预警阈值时, 自动触发预警通知。
为了实现预警机制, 可以编写一个简单的 Java 程
序, 定期检查就业率, 若低于某个阈值则触发警报。
public void checkEmploymentRate() {
// 假设通过 JDBC 连接到 Hive 获取就业数据
Connection conn =
DriverManager.getConnection("jdbc:hive2://local
host:10000/default", "user", "password");
Statement stmt = conn.createStatement();
ResultSet rs = stmt.executeQuery("SELECT
profession, employment_rate FROM
employment_statistics");
while (rs.next()) {
double rate = rs.getDouble(2);
if (rate < 70.0) { // 假设 70%为预警阈值
System.out.println("Warning: Employment rate
for " + profession + " is below threshold: " + rate
+ "%");
}
}
rs.close();
stmt.close();
conn.close();
}
决策支持层: 生成定制化的就业报告, 为学校管理
层提供决策依据。并提供针对专业设置、教学内容、就
业指导等方面的优化建议。
```

用户界面层: 开发 Web 前端界面, 提供可视化的数
据展示和交互式查询功能; 设计移动应用, 方便学校管
理人员和就业指导老师随时访问就业数据, 为用户提供
友好的操作界面。

4. 系统部署

将模型部署为 API, 使用 Flask 或 FastAPI 等框架
进行接口暴露。

```
from flask import Flask, request, jsonify
app = Flask(__name__)
@app.route('/predict', methods=['POST'])
def predict():
data = request.get_json()
# 处理输入数据并进行预测
# 返回预测结果
return jsonify(result)
if __name__ == '__main__':
app.run(debug=True)
```

5. 结论与展望

本文基于 Hadoop 平台构建了高职院校毕业生就业
监测预警的决策支持系统, 通过实时数据采集、存储、
分析与挖掘, 为高职院校提供了科学、精准的就业指导
和决策依据。未来, 可以进一步优化系统性能, 提高数
据处理效率和分析精度; 同时, 还可以拓展系统的应用
范围, 将其应用于其他领域的监测预警和决策支持中。

参考文献

- [1]任建新. 基于 Hadoop 平台的大数据应用系统架构
的研究与实现[D]. 北京: 北京邮电大学, 2014.
 - [2]关长海. 基于高校毕业生就业状况监测的专业预警
体系建构初探. 北京教育(德育)
 - [3]金义富, 吴涛. 大数据环境下学业预警系统设计与
分析[J]. 教育信息化, 2016
 - [4]李鹏, 蔡治廷. 大数据时代的大学生就业工作探析
[J]. 黑龙江高教研究, 2015(5):
 - [5]杨锐, 夏红. 大数据时代下大学生就业数据信息的
应用研究[J]. 中国电力教育, 2014(20)
- 作者简介: 杨华 男 大学本科 副教授 主要研究方向:
计算机应用及算法、图形图像处理、教育管理等。
- 基金项目: 南充职业技术学院 2022 年度项目, 项目编
号: ZRB2202;