IT 行业薪酬变化数据分析

万嘉杰 [1]

网物 1801 201892296

摘 要

结合 Hadoop 平台的高扩展性、高性能、与低成本的优点，设计基于 Hadoop 招聘数据分析的框架。对近 200 万条数据分词、去重、去噪、提取特征，构造特征矩阵与文本矩阵，利用奇异值分解法对文本矩阵降维，按相似度分类，对分类结果进行关联规则挖掘与数据统计分析。结果表明 Hadoop 平台数据分析效率明显提高，具有较高的加速比。实验结果（IT 行业）呈现目前就业岗位、薪资、所需技能、工作地点的关联规则与统计结果分析，为行业的发展与就业提供一定的数据参考与支撑。

关键词：

数据挖掘；数理统计；Hadoop…

1 引言

随着互联网的迅速发展，大量的人才招聘信息分布在互联网上，形成了大量的异构非结构化网络数据。对这些数据进行有效的分析，对行业的发展具有重要意义与环境的导向作用。非结构化数据在数据处理中具有独特的特点在分析阶段难以确定，大量数据分析能力不足，性能不足。文献[1]中，三个招聘平台的数据接近 8 万

对计算机行业招聘数据进行聚类分析，并对各行业招聘数据进行统计该头寸的市场需求，并计算与该头寸相关的其他维度感兴趣的相关系数。文献[2]将四家招聘网站的数据进行了划分。

其次，利用二维隐马尔可夫模型对招聘信息进行分割在招聘信息中，职位、企业名称、企业类型等关键词的书写。文献[3]通过数据预处理，分析了爬网的 50 万条数据采用奇异值分解（SVD）方法求解降低了文本矩阵的维数，采用聚类算法挖掘行业信息。文献[4] Hadoop 平台用于分析网络舆情数据。文献[5]利用 Hadoop 平台分析葡萄酒数据信息。文献[6]基于 Hadoop 平台对商业银行的数据进行了分析。Hadoop 软件技术已逐渐成为一种比较完整的分析技术。基于 Hadoop，提出了并行计算能力弱的问题。针对平台招聘数据分析研究，为近 2000 万条计算机行业招聘数据进行分析。

2 相关技术

2.1 HDFS 分布式文件系统

HDFS 文件系统 HDFS 文件系统采用主从结构，由主节点和从节点组成部分数据节点。主节点主要负责文件系统中的数据 meta 的存储管理包括存储地址的选择和空值的命名以及每个节点的访问权限和数据块之间的关系。数据节点负责特定数据块的存储和管理。包括具体的根据创建的块，读写数据并将信息反馈给主节点。什么时候？当要存储的数据文件很大时，HDFS 会对文件数据进行分割它是一个独立的数据块，由主节点引导并发送到数据中心每个数据节点中每个数据块的存储信息是相同的保存在主节点中。主节点负责调用执行数据节点，统计节点数数据库节点不定期地将更新后的数据反馈给主节点。

2.2 MapReduce 编程模型

MapReduce 的实现过程分为两个阶段，map 阶段和 Reduce 阶段[7]，map 是映射阶段，Reduce 是还原阶段。首先，主节点输入文件，执行分割操作，然后执行映射。该操作将文件解析为<key，value>格式并存储中间数据。存取节点的快取空间定期写入本机磁碟，并分成 R 个 regions，每个 region 对应一个 reduce 作业，执行 reduce 分区数据可以在操作前进行排序和合并。所有数据都是正确的。从底层文件系统，执行过程生成的临时数据存储在在当前节点的文件系统中，执行结果最终存储在底层文件系统中在分布式文件系统中。

3 分析系统的设计与实现

3.1 分析平台系统架构

将大数据技术与数据挖掘技术应用到招聘数据的分析中，实现了基于 Hadoop 平台的招聘数据分析平台，如图 1 所示，分析平台包括数据采集、文本处理、分析与展示四大模块。

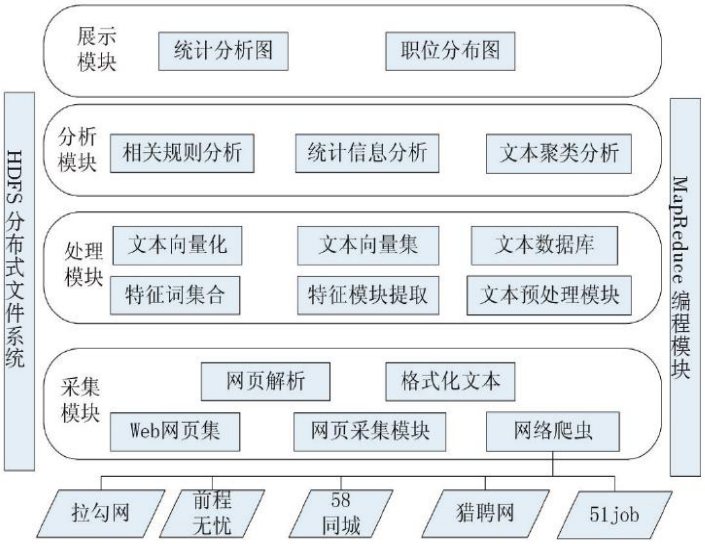

图 1 平台架构图

信息采集模块主要利用网络爬虫从各大在线招聘平台的网页上采集相关招聘信息。需要保证数据的全面性和准确性。数据采集成功后，需要对数据进行基本去噪和基本格式化，并使用统一的接口将格式化后的数据存储在数据库中。被爬网的数据大多是文本格式的数据，部分数据存在关键字字段数据为空、行数重复等问题。采集到的数据具有一定的噪声特征，对词频统计、文本聚类和相关分析都有一定的影响。为了保证数据的完整性和唯一性，补充了部分缺失的文本数据，对不同网站的相同数据进行了去重。在数据处理模块中，主要是提取特征模块，生成特征词集、文本向量化、文本向量集，最后生成文本数据库。

经过数据特征提取和稳定代价向量化，得到数据的文本数据库，并将数据发送到其他节点进行存储。数据分析模块主要采用数据挖掘技术进行分析，分别从数据统计的角度和挖掘不同维度的相关规则。

3.2 数据处理

收集的招聘数据中的一些属性需要数字化。如，工作经历的字符类型：1 年、3 年、无需，可以根据正则表达式转换为 1、3、0；薪资的字符类型：5K、8K、10K 等，可以转换为数字 5、8、10。文本中的中英文单词是分段的。正则表达式用于分割英文数据，jieba 用于分割中文数据。基于前缀字典实现了高效的词图扫描，并对扫描结果进行了分析生成由可能的新词组成的有向无环图（DAG），然后用动态规划算法寻找概率最大的路径，得到词频最大的分词组合。对于未入词典的词，采用基于汉语习语能力的 HMM 模型，汉语分词结果如图 2 所示。

```
['五险一金', '餐饮补贴', '通讯补贴', '绩效奖金', '年终奖金', '员工旅游']
['五险一金', '员工旅游', '年终奖金', '定期体检']
['五险一金', '餐饮补贴', '定期体检', '周末双休', '年终奖金', '项目奖金']
['五险一金', '员工旅游', '年终奖金', '定期体检', '股票期权', '节日福利',
['五险一金', '餐饮补贴', '通讯补贴']
['五险一金', '餐饮补贴', '年终奖金', '弹性工作', '员工旅游', '定期体检']
['周末双休', '五险一金', '通讯补贴', '专业培训', '年终奖金', '定期体检',
['五险一金', '绩效奖金', '定期体检', '弹性工作']
['五险一金', '年终奖金', '定期体检', '员工旅游', '绩效奖金', '朝九晚五']
```

图 2 中文分词

4 结果分析

4.1 关联规则挖掘

挖掘关联规则时，需要将招聘信息中的城市级别划分为一线城市、二线城市和三线城市；需要将公司规模划分为 50 人以下、50 人-100 人、100-300 人、300 人-500 人和 500 人以上；我们要把工资折算成 5K 以下、5K-8k、8k-12k、12k-20k、20k 以上的月薪；学历分为博士、硕士及以上、本科及以上、大专；工作年限分为 1 年、2 年、3-4 年、5-7 年和 10 年以上。

关联规则挖掘部分结果如表 1 所示，可知在一线城市，如果拥有两年工作经验，近 92%的企业开出的薪资在 8K-12K 之间。如硕士毕业生且有工作经验，在一线城市，有 88.75%的企业愿意开出 12K-20K 之间的薪资。大数据专业的硕士毕业生主要需求地为一线城市。在二线城市，公司规模大部分维持在 50-100 人的中小型企业，如果为本科学历，82.72%可能会拿到 5K-8K 之间的薪资。

前项	后项	实例数	支持度（%）	置信度（%）	规则支持度（%）	提升	部署能力
地点：一线城市	薪资：5k—8k	657157	21.336	90.63	19.775	1.152	1.561
地点：一线城市 经验：2 年以上	薪资：8k-12k	443018	17.842	92.03	15.719	1.120	2.123
地点：一线城市 经验：2 年以上 学历：硕士以上	薪资：12k-20k	287306	13.372	88.75	10.517	1.024	2.855
地点：二线城市	规模：50-100	320877	22.675	89.53	20.845	1.150	1.83
地点：二线城市 薪资：8k-12k	规模：50-100	232066	16.362.	86.49	14.331	1.034	2.031
学历：硕士以上 岗位：Java	一线城市	286093	13.617	89.62	11.715	1.037	1.902
薪资：5k-8k 学历：本科以上 经验：1 年以上	二线城市	343895	18.169	82.72	15.974	1.010	2.195

表 1　部分关联规则分析结果

4.2 统计分析

平台的统计数据如图 3-5 所示。从图 3 据我们了解，目前有近一半的 Java 岗位要求本科以上学历，34.4%的 Java 岗位要求大专以上学历，拥有硕士学位的 Java 编程岗位相对较少。从图 4 可以看出，要求 10 年以上工作经验的 Java 岗位比例仅为 0.5%，大部分岗位要求 1-4 年内工作经验，占 69.6%。应届毕业生的工作不需要工作经验。图 5 显示了不同语言编程工作的人数和工资的统计数据。从图 5 可以看出，Java 语言开发工作的需求量很大，Java 和前端的高薪工作更多。图 6 显示了不同工作要求和工资之间的关系。从图中可以看出，Java 和前端的工作需求比较大。

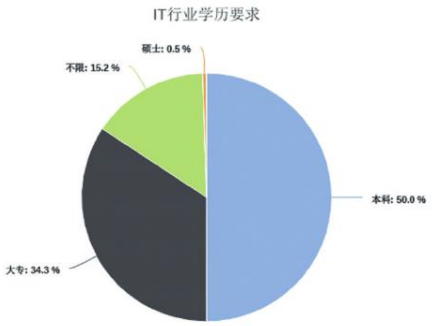

图 3 Java类岗位学历要求统计

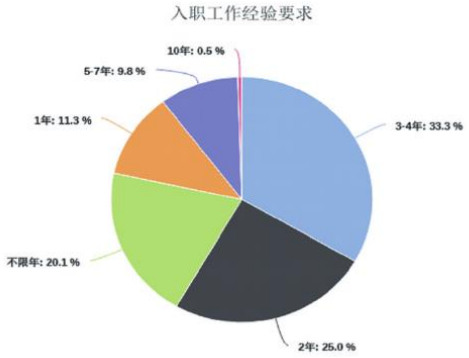

图 4 Java 类岗位入职经验统计

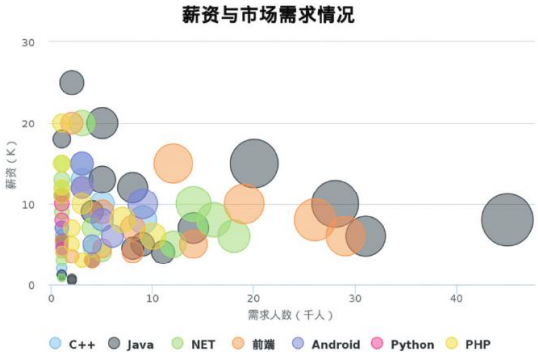

图 5 薪资与市场需求统计图

结论

本文将离线招聘数据的分析搬迁到 Hadoop 平台上，设计与实现了数据分析平台，平台包括数据采集模块，处理模块，分析模块，展示模块。利用关联规则算法对岗位，所需技能，薪资，工作经验等特征维度进行关联规则挖掘，同时利用统计分析法对就业分布，薪资，市场比例等进行分析，形成可视化统计数据。

参考文献

[1]钟晓旭，胡学钢. 基于数据挖掘的 Web 招聘信息相关性分析[J]. 安徽建筑工业学院学报（自然科学版），2010，18（04）:93-96.

[2]王静. Web 对象的信息抽取的关键技术研究[D]. 西安:西安电子科技大学，2011.

[3]张学新，贾园园，饶希，蔡黎. 海量非结构化网络招聘数据的挖掘分析[J]. 长春师范大学学报，2017，36（10）:28-36.

[4]谌志华. 基于大数据的网络舆情分析系统[J]. 现代电子技术，2017，40（24）:15-17.

[5]郝艳妮，田维丽. 基于 Hadoop 的数据挖掘算法在葡萄酒信息数据分析系统中的应用[J]. 计算机应用，2017，37（S1）:72-74+79.

[6]张登耀. 基于 Hadoop 分布式文件系统的商业银行大数据分析[J]. 山东农业大学学报（自然科版），2018，49（05）:884-888.

[7]Bendre M，Manthalkar R. Time Series Decomposition and Predictive Analytics Using MapReduce Framework[J]. Expert Systems with Applications，2019，116:108-120.

[8]邬启为. 基于向量空间的文本聚类方法与实现[D]. 北京:北京交通大学，2014.

[9]廖飞. 基于关联规则的试题生成与数据分析方案研究[D]. 广州:华南理工大学，2018.