会议总结

2020-8-15

**【报告总结】**

**(彭志远)可视化**

在上次可视化图的基础上，开发出新的特点：

1. 更新了用户筛选渗透框和时间拖动条样式；
2. 可以选择特定的User(s)，以不同颜色进行标识，通过拖动时间轴展示其职业变迁轨迹；
3. 对Echart数据承受性进行调研，可支持5000点的顺畅时态切换。

**(刘志威)标签分类**

1. 将标签整体分为Basic和Hybrid两大类。Basic label 可从简历文本中直接匹配到， Hybrid label由Basic进一步扩展得到。某些标签既属于Basic，也属于Hybrid类型。
2. 在Basic标签匹配中，除了关键字匹配以外，增加正则匹配的方式。
3. 在Hybrid 标签匹配中，利用多种集合逻辑关系综合匹配。
4. 简历匹配成功的召回率为0.88，未成功匹配的原因：预处理逗号；无法识别的私营企业名称等。

**（李可）简历分词**

1. 分析了混合模型在经过OCTree加入之后召回率下降的两个主要原因：proportion 和 unconditional compress
2. 针对召回率下降的原因，做出修改，其中c和d对于效果提升最大。
3. 增加一倍训练集（300条简历）
4. 增大词频的multiplier（从2到15）
5. 前序遍历OCTree,对单字节点进行Split.
6. 压缩节点时加入exclude suffix进行过滤
7. 经过上述调整，准召率以及F1 score均有大幅度提升，从40%的水平稳定到70-75%的水平，分词结果接近期望结果。
8. 但仍存在着叶节点不压缩导致职位被分成两个节点的问题，日后考虑通过词性标注进行区分，对compress增加过滤条件。

**（孔德旭）先锋数据**

1. 介绍了图结构中的割集和shortcuts.

2. 分析了添加shortcuts以达到最优平均路径长度的最优化方法和贪心算法

3. 使用Path Screening Method在现有合作图上找到了最优的shortcut，分别是来自于龙岗区和坪山区的党支部，他们的地理距离是很近的。

【讨论】

1. 数据展示应和用途有关，要有意义，有目的，有导向，搞清楚期望从可视化中看到什么样的数据。譬如单个用户和群组用户可视化的区别、意义是什么，我们想要从群组可视化中得到什么结论？
2. 对于职业网络的可视化，一方面可以通过一些特殊的视图获得新的信息，例如考虑某个部门的晋升途径，人员从哪里来，到哪里去；另一方面可以改变展示方式，以用户为单位而非以时间为单位。
3. Demo时不一定要有完整的算法实现，可以人工筛选出具有代表性的数据， 更好的讲故事，让听众理解motivation和研究目标。
4. 在先锋项目中，如何去量化合作效率的提升？

**【未来工作计划】**

1. （彭志远）可视化：
   1. 快速原型开发： 从不同的视角对数据进行可视化并作相应研究。
2. （刘志威）基于知识的标签学习系统
   1. 补充更多基于规则的简历标签分类器的判断规则(关键词、正则、混合逻辑)；
   2. 设计简历标签与人员标签结合的数据结构；
   3. 设计满足多标签Query的标签查询结构。
3. （李可）分词系统
   1. 对模糊分词“深圳市政府”进行定义，确保consistency。
   2. 对整个offline的分词流程进行优化
4. （孔德旭）先锋数据-社交网络分析
   1. 需要明确研究shortcut问题的意义。考虑通过引入战疫先锋上更多的数据，来找到衡量添加shortcut的效果的指标，如参与度等，可在小程序上推荐合作任务。
   2. 从图的动态生成的视角，找到一个特定的合作产生的时间节点，以及其之后对图产生的影响。