**Mining Rising Stars in Academic Graph**

Yao Lu

July, 2016

**1 已有的一些论文思路**

关于rising star这个问题，目前有三篇这样的文章。第一篇文章是提出了用学术质量作为网络的vertex value，用合作关系来建立学术网络（如果两个人一起发表一篇文章，就存在两条directed edge，edge的weight通过两个人合作的文章占各自发表的文章比重来确定），最后用pagerank框架来计算，按照一个人相邻几年的rank 增长趋势来评估一个人是不是rising star(类似于rank的上升加速度)

第二篇文章就是纯粹把第一篇文章的feature提取过程更加细化，没有算法上和特征上的创新。

第三篇文章分别计算了作者发表的文章的rank(也就是文章网络)，得到一个文章的分数，也通过文章keyword来计算一个作者和他的合作者的研究方向的差异有多大（合作信息熵的概念，因为他的假设是一个人能和不同领域人合作，就能有更好的发展前景），最后把文章网络的分数和hits算法给出的分数和他的合作信息熵相乘，作为rank分数。

前两篇文章都是用dblp数据集跑的，最后的evaluation就是比较选出的rising star在未来的引用数是不是高于平均值，每篇文章给出的解释都是说没有groundtruth数据集，所以就用引用数来评价一下。有些杰出rising star，文章会列表说明他是ieee fellow这种的，总之就是evalutation这个过程也不可信。

**2 目前已有的方法的缺陷**

**2.1 没有考虑cs大领域的划分而是直接排序**

A 采用整个领域来进行计算，但是，这样会导致一些研究人员众多的领域的rank高于小的研究领域，比如计算机视觉这个领域的热度大于系统领域。我们应该专注于某一个领域的排序而非整个cs领域的排序。

**2.2没有考虑细分领域以及跨领域合作的权重关系（community detection）**

对于一个cs的子领域，比如nlp，其下有很多的细分方向，如果直接排序，可能会出现某个细分方向的最厉害的人在其他方向也排名不高（因为不同研究方向的研究人员数量也不同），所以在rank时候，我们应该更多考虑nlp的graph中，cluster之间的cooperation的weight应该被弱化，而clutser内部的cooperation应该被强化。

**2.3缺少可信的evaluation metrics**

因为之前文章都是用dblp整个cs的数据集来计算，或者是用aps物理学会的论文数据集来计算，所以没有办法界定选出的rising star后来到底有没有变成star，所以只能用选出的top author的平均引用数和其他random selected author对比。

并且用同一个算法对于不同的人在3年5年7年这种不同的预测时间间隔上，计算rank list的一致性。

之前的文章给出的理由是说，For the ranking of rising stars, there is no ground truth of it currently. Therefore, we adopt the scholar’s future citation counts as the ground truth to validate whether these rising stars have achieved their expectations.

但是，对于rising star的界定，顶级会议的area chair就可以认为是star了，所以我们建立了acl系列的会议的area chair数据集来作为groundtruth进行预测。

**3 目前我们工作的进展**

之前我尝试用从学术图谱中获取的整个cs网络来进行排序，但是效果不好。主要是引用网络太复杂，存在大量噪音。

所以我只采用了nlp领域五个顶级会议(ACL,NAACL,EMNLP,EACL,COLING)的发表数据构建数据集。

此外，我建立了从2006-2016年之间的这五个会议的chair的数据集作为我们的groudtruth来验证算法。

我们通过截止到2005年的数据，来预测2000-2005年之间开始发表第一篇论文的作者未来的排名，并且和groundtruth数据集进行对比进行评估。

之前方法是 publication\_quality作为vertex initial value，cooperation frequency作为edge weight，加上pagerank进行计算

我目前简单验证了一下把publication\_quality 换成publication number in top conference这个特征，此外算法不变，初步验证了我们的数据清洗和特征选取的有效性。

实验结果如下所示，novarank是我们的方法，starrank是之前文章的方法。

|  |  |
| --- | --- |
| 找出的rising star的数目对比 | 找出的rising star的准确性对比 |
| /private/tmp/1.png | /private/tmp/1.png |

**总结**

目前我们工作是清洗数据，建立了数据集，这些已经全部完成，并且通过初步尝试，确认了rising star这个项目还是有可行性的。

目前我的想法是

1通过community detection，给同一个community内部的连接赋予更高的权重，来进行计算。(假设某个作者有10个合作者，但是其中有2个合作者不属于这个作者所在的community，其权重应该降低。)

2 如果一个作者，能和多个不同机构的属于同一个community的作者合作，那么他的未来发展更大(通常来说，多个机构合作完成的paper质量会比较高)