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

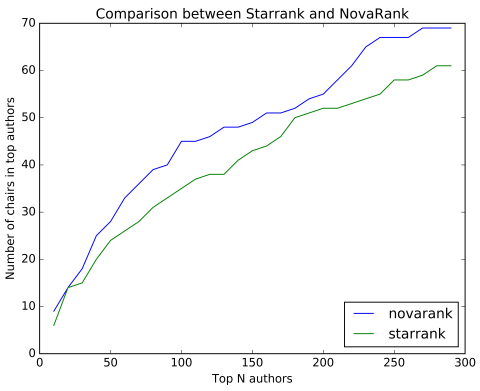
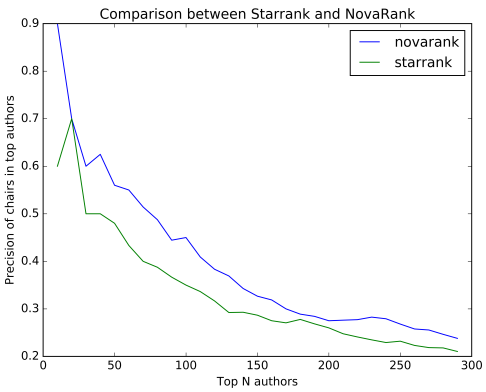
之前我尝试用从学术图谱中获取的整个 cs 网络来进行排序，但是效果不好。主要是引用网络太复杂，存在大量噪音。
所以我只采用了 nlp 领域五个顶级会议(ACL,NAACL,EMNLP,EACL,COLING)的发表数据构建数据集。
此外，我建立了从 2006-2016 年之间的这五个会议的 chair 的数据集作为我们的 groudtruth 来验证算法。

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
	

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)