节点分析社区挖掘等机器学习任务——基于学术文献网络

黄之豪 [1]
20110980005@fudan.edu.cn
周笑宇 [1]
20210980066@fudan.edu.cn
王嘉炜 [2]
20210840012@fudan.edu.cn

[1] 大数据学院
[2] 数学科学学院

摘要

本次社交网络分析作业中，我们小组主要完成了如下工作：1. 通过选取关键词及爬虫，获取了金融、CS 领域 10 个细分关键词的总共 229 万条文献数据；2. 计算网络统计属性；3. 进行网络节点分析；4. 基于异构文献网络学术合作预测；5. 通用预训练文献网络 Embedding；6. 文献 Abstract 相似性分析；7. 网页、作图等项目可视化。

1 Data

首先，小组成为根据专业特点和课程特点（如：social network analysis）确定了金融和 CS 两大主题和两大主题下 10 个细分主题：financial pricing, financial risk management, financial risk measurement, behavior finance, corporate finance, fintech, financial supervision, deep learning, machine learning, social network analysis。

将这些关键词作为 web of science 搜索的 keywords 进行文章爬取，去除没有 doi 的文章，总共获得种子文章 263,698 篇。进一步抓取种子文章的参考文献，最终总共获得了 2,298,732 篇不重复的文献数据。考虑到 mongo 在修改数据、数据库结构方面便捷性优势较大，使用 mongo 数据库。

字段及相关说明如下表所示 (网盘中有"deep learning" 为关键词的详细数据):

部分数据及详细代码见百度网盘：https://pan.baidu.com/s/1D92YDNL8fhqTaM61GOSMqw 提取码：su5w ，如有问题可联系三位同学，项目网址：http://10.192.7.62:8080/index.html

字段名	说明	字段名	说明
_id	doi	publication_type	文献类型 (Journal,book 等)
authors	作者	authors_full_name	作者全名
title	题目	num_id	自己给予的 unique 编号
pub_year	发表年份	research_areas	作者研究领域
abstract	摘要	wos_id	Web Of Science 文章编码
issn	issn 编号	big_topic	Finance / CS
journal	发表期刊	topic_field	10 大细分领域中的一个或多个
keywords	关键词	author_address	作者工作地址
issue	期刊 issue	cited_number	其他文章引用该文章数
crawl_time	爬取时间	cite_references_number	引用其他文章数
vol	期刊 vol	category	WOS 领域分类
lang	语言	keywords_plus	更多关键词
publisher	出版商	reference	文章引用文章的 doi list
pub_date	发表日期		

表 1: 数据库字段及说明

2 网络的统计属性

在网络的统计属性上，我们抽取主题为 Deep Learning 领域的文章构建作者合作网络，计算了平均度，平均聚类系数等属性，如表所示。同时，作出作者合作网络 (图 1 左) 及文献引用网络 (图 1 右) 中的度分布图，可明显看出度是服从幂律分布的。

进一步，我们采用 Louvain 算法对作者合作网络进行了社区划分，该算法目标是最大化模块度，能够发现层次化的社区结构。可视化展现前几个社区，可见整个作者的合作网络是比较离散的，几位作者之间往往会构成学术团体，团体内部合作紧密，外部合作较少。

平均度	5.003
平均聚类系数	0.994
平均路径长度	1.16
最大联通子图直径	3

表 2: 作者合作网络属性

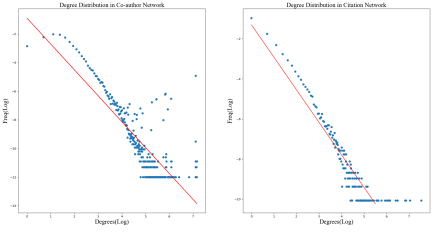

图 1: 度分布

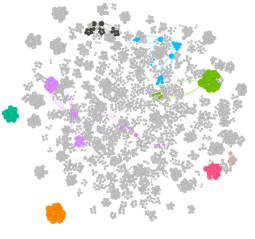

图 2: 社区可视化

3 网络节点分析

3.1 针对论文节点

在论文引用网络中，包含节点（论文）和有向边（引用，由引用指向被引用）。这些图的性质能够帮助我们更好的度量节点的中心性，从而找出影响力最大的论文。在项目中我们使用的计算中心性的算法包括：

1. 特征向量中心性

通过结合无向图中的邻居节点，或者有向图中的输入邻居节点的重要性来概括中心度。

2. PageRank

   将中心性除以节点的外连接（出度）的数目，使得每个邻居节点获得源结点中心性的一部分。

3. 接近中心性

   若一个节点越趋于中心，则其能够更快速到达其它节点，即它与其他节点之间有最小平均最短距离。

3.2 针对作者节点

通过上述方式我们计算出论文的影响力（中心性），记为 $IF(u_i)$。我们希望通过论文的影响力来衡量作者影响力，我们设计了如下算法：

1. 将每篇论文的影响力 $IF(u_i)$ 均分给该论文的所有作者，每位作者获得的影响力为 $\frac{IF(u_i)}{n(u_i)}$，$n(u_i)$ 为论文 u_i 的作者数。

2. 将作者 v_i 从其发表的论文中所获得的影响力求和，$\sum_{v_i} \frac{IF(u_i)}{n(u_i)}$，作为 v_i 的初始影响力。

3. 根据作者合作网络，利用改进的 PageRank 算法计算，迭代过程如下：

$$C_p(v_i) = \alpha \sum_{j=1}^{n} A_{i,j} \frac{C_p(v_j)}{d_j^{out}} + \beta(v_i)$$

$$A_{i,j} = \ln(k_{i,j}) + 1$$

$$\beta(v_i) = \sum_{v_i} \frac{IF(u_i)}{n(u_i)}$$

   其中 $k_{i,j}$ 表示两个作者之间的合作次数，$u_i \in \{$ 作者 v_i 写过的论文 $\}$。

4. 直到 $C_p(v_i)$ 收敛，迭代终止，得到所有作者最终的中心性值 (影响力)。

为了衡量我们算法的效果，根据我们计算的中心性值对作者进行排序，将 Google Scholar 中作者的总被引用次数的排名结果作为参照集，取前十五位学者，结果如下，两种排序的 Spearman 系数为 0.86。

Pr 值排序	1	2	3	4	5	6	7	8	9	10	...
总被引用数	1	2	4	3	8	9	7	6	5	14	...

表 3: 两种排名比较

4 学术合作预测

针对学术合作预测问题，传统的同构网络只包含单一类型的节点，这种网络结构包含的拓扑信息有限，往往难以揭露深层次的规律。因此，为了充分利用已知信息来对未来合作进行预测，我们考虑异构学术网络。在异构网络中不仅包含作者节点，还包含论文节点、主题节点、期刊节点以及连接它们的边，节点之间的关系也更加复杂。受到 2011 年发表于 ASONAM 上论文的启发，异构网络中不同节点间复杂的关系衍生出了元路径的概念，针对每一种元路径，都可以计算该路径的异构拓扑特征，如 Path Count、Symmetric Random Walk 等，最后将这些特征输入到二分类模型中，来对作者合作与否做出预测。

在实验过程中，我们选取了“作者-论文-共同合作者-论文-作者”，“作者-论文-共同引用论文-论文-作者”，“作者-论文-期刊-论文-作者”这三类元路径，在下文中我们简记为“A-P-A-P-A”，“A-P$\rightarrow P\leftarrow$P-A”，“A-P-V-P-A”。我们计算这三类关系的 Symmetric Random Walk(SRW) 值作为特征，计算方式如下：

$$\begin{aligned} SRW_R(a_i,a_j) &= RW_R(a_i,a_j) + RW_{R^{-1}}(a_j,a_i) \\ &= \frac{PC_R(a_i,a_j)}{PC_R(a_i,\cdot)} + \frac{PC_{R^{-1}}(a_j,a_i)}{PC_{R^{-1}}(\cdot,a_j)} \end{aligned}$$

其中 $PC_R(a_i,\cdot)$ 表示以一定规则的从 a_i 出发的所有路径之和，$PC_R(\cdot,a_j)$ 表示以一定规则的以 a_j 为结尾的所有路径之和。

采取逻辑回归模型进行分类，首先将数据集按照时间段分成三部分:T_0=[2003 年,2008 年],T_1=[2009 年,2014 年],T_2=[2015 年,2018 年], 将 T_0 和 T_1 的合作数据记录下来作为训练集，将 T_1 和 T_2 的合作数据记录下来作为测试集。训练模型中，给出特征的显著性如下：

Meta Path + Measure	Coefficient	P-value
A-P-A-P-A + SRW	0.9366	0.016
A-P$\rightarrow P\leftarrow$P-A + SRW	0.5285	0.060
A-P-V-P-A + SRW	0.4482	0.142

由此可见，作者之间发生学术合作的概率与他们有共同合作者高度相关，相关性次之的是共同引用同一篇论文，最后是在同一种期刊发表过论文。这也与实

际情况相符合：一般而言，若两人有共同合作者，则他们无论是在学术交流还是在现实生活中接触的概率都要远大于另两种情况，因此这种特征在分类模型中表现极为显著，而引用同一篇论文则暗含两者有相同或相近的研究方向，这比只是在同一领域的以期刊为桥梁的关系无疑要更为密切。

最后，我们根据训练的逻辑回归模型对未来学术合作进行预测，输出的混淆矩阵如下图所示：

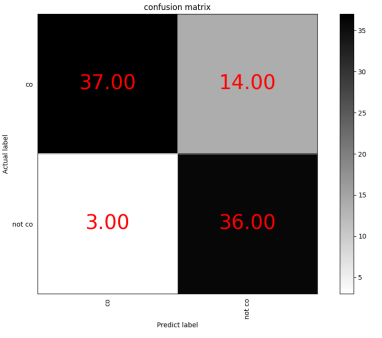

图 3: 预测结果

可以看到，我们的模型对于未来不会发生合作的情况预测的非常准确，对于未来会发生合作的学者预测也有较高的准确率。

5 Metapath2vec

Metapath2vec 是 Yuxiao Dong 等 [1] 于 2017 年提出的一种用于异构信息网络 (Heterogeneous Information Network, HIN) 的顶点嵌入方法。metapath2vec 使用基于 meta-path 的 random walks 来构建每个顶点的异构邻域，然后用 Skip-Gram 模型来完成顶点的嵌入。在 metapath2vec 的基础上，作者还提出了 metapath2vec++ 来同时实现异构网络中的结构和语义关联的建模。

本文的主要贡献在于：先前针对同构网络提出的 embedding 方法无法保留异构网络中的” 节点上下文 (word-context)” 信息，使用 metapath 方法最大化保留给定异构网络的结构和语义。开发了一种基于异构负采样的，称为 metapath2vec++ 的方法，该方法可以准确而有效地预测节点的异构邻域。

5.1 The Metapath2vec Framework

Metapath2vec 模型思想、代码部分很大程度都借鉴了 Mikolov 等 (2013) word2vec[2] 的思想。以下主要根据 Metapath2vec 模型的几处创新点进行模型

Rank	QUANTITATIVE FINANCE	MACHINE LEARNING
1	Complexity,Metastability and Nonextensivity	2018IEEE/ACM40TH INTERNATIONAL CONFERENCE ON SOFTWARE ENGINEERING
2	MARKET MICROSTRUCTURE AND LIQUIDITY	ECAI2010-19TH EUROPEAN CONFERENCE ON ARTIFICIAL INTELLIGENCE
3	MATHEMATICAL CONTROL THEORY AND FINANCE	PROCEEDINGS OF THE 12TH INTERNATIONAL CONFERENCE ON WEB INFORMATION SYSTEMS AND TECHNOLOGIES
4	NUMERICAL COMPUTATIONS:THEORY AND ALGORITHMS	MATHEMATICAL PROGRAMMING COMPUTATION
5	FROM STOCHASTIC CALCULUS TO MATHEMATICAL FINANCE	ARTIFICIAL INTELLIGENCE IN MEDICINE

表 4: 训练的 Embedding 效果展示

框架概述：

该文章最大贡献在于根据异构网络特性，提出基于预先指定的元路径 (meta-path) 来进行随机游走，预先构造路径能够保持” 节点上下文” 的概念。meta-path scheme 定义 $\mathcal{P}: V_1 \xrightarrow{R_1} V_2 \xrightarrow{R_2} \cdots V_t \xrightarrow{R_t} V_{t+1} \cdots \xrightarrow{R_{l-1}} V_l$ ，其中，$R = R_1 \circ R_2 \circ \cdots \circ R_{l-1}$ 代表 V_1 到 V_l 这些节点间的关系。如图所示，例如”APA”(作者-文章-作者) 就是一条元路径 (工程上来说论文源代码和本次 PJ 都使用” 期刊-作者-期刊” 的元路径)

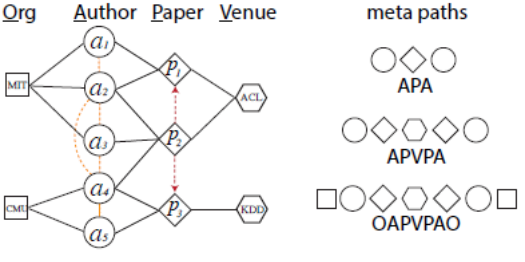

图 4: 文献 meta-path 示例

文章第二大贡献在于提出基于异构网络的负采样，即 metapath2vec++，具体地，通过具体到某个类型节点的邻居取代所有类型进行负采样。

生成随机游走: 1000 次随机游走, 每次 10 步。Embedding 参数: 128 维, skip-gram window 7, negative sampling 5。实验效果：通过使用 metapath2vec 训练 Embedding 后，通过相似度匹配，找到与某个期刊最相近的期刊，QUANTITATIVE FINANCE 和 MACHINE LEARNING 为例，得到与之最为相似的期刊如下：

根据先验知识，量化金融确实与数值计算、数学控制、随机分析等有着密不可分的关系，结合网盘中的 Embedding 降维成三维之后展示效果，可以看出，Embedding 质量较好。

6 摘要相似度

这里选取了 Zhelezniak 等 (2019)[3] 的 baseline 文本相似度分析方法，通过 word2vec, Glove, ELMo 三种词向量模型和 avg_cosine, pearson, spearman, kendall, apsyn, apsynp 六个向量相似度度量指标来计算摘要的相似度，部分结果如可视化网页所示。

7 小组分工

周笑宇：负责整体模型思路的设计，网络的统计属性和社区划分的编程实现，PPT 以及讲稿制作。
黄之豪：负责爬虫，Metapath2Vec 以及摘要相似度的编程实现以及节点三维可视化，项目的网页可视化。
王嘉炜：负责节点分析、作者学术合作预测的编程实现。

References

[1] Yuxiao Dong, Nitesh V. Chawla, and Ananthram Swami. Metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 135 – 144, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450348874. doi: 10.1145/3097983.3098036. URL https://doi.org/10.1145/3097983.3098036.

[2] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111 – 3119, Red Hook, NY, USA, 2013. Curran Associates Inc.

[3] Vitalii Zhelezniak, Aleksandar Savkov, April Shen, and Nils Y. Hammerla. Correlation Coefficients and Semantic Textual Similarity. *arXiv e-prints*, art. arXiv:1905.07790, May 2019.