《审稿人讲座》

浅谈科研

分享人：张文涛

北京大学信息科学技术学院网络所

分享内容部分参考自北邮邵蓥侠教授的《研究生做科研的几点建议》

个人简介

张文涛，北京大学，信息科学技术学院计算机系博士生

科研

- **论文**：在数据库、数据挖掘和机器学习等领域发表论文10余篇，其中，CCF A类论文12篇，包括SIGMOD, VLDB, KDD和NeurIPS等国际一流会议
- **系统**：参与多个开源系统，包括分布式机器学习系统Angel (6.4k stars)，自动化机器学习系统MindWare和黑盒优化系统OpenBox
- **审稿**：担任多个国际一流会议和高质量期刊的审稿人

荣誉奖励

- 苹果学者（亚太地区1人，全球15人）
- 百度奖学金全球前20强
- 北京市三好学生
- 北京大学年度人物候选人、国家奖学金、廖凯原奖学金和学术创新奖等

主页: https://zwt233.github.io/

提纲

1. **什么是科研?**
2. **怎么做科研?**
    a) 如何选方向?
    b) 如何看论文?
    c) 如何发现新问题?
    d) 如何提出新方法?
    e) 如何验证新方法?
3. **怎么写论文?**
    a) 如何讲故事?
    b) 如何提升语言表达?

1. 什么是科研？

科研是一个创造新知识的过程

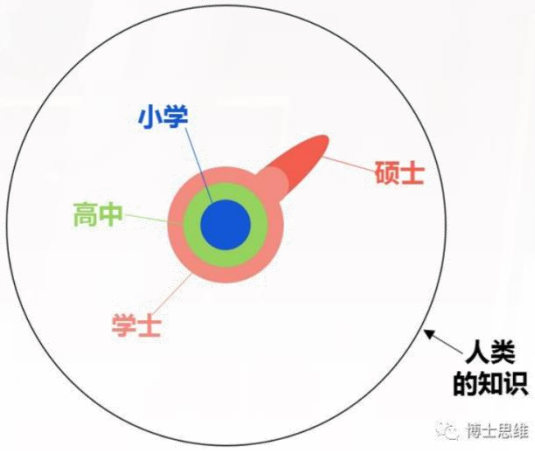
- **Research** is creative and systematic work undertaken to increase the stock of knowledge.

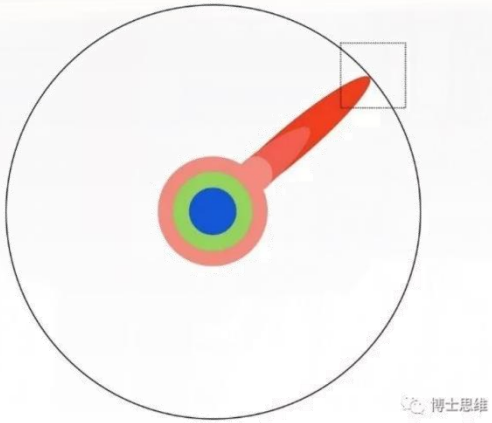
-- from Wikipedia.

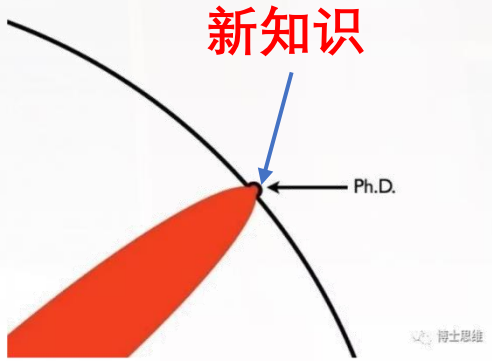
- **科学研究**指在发现问题后，经过分析找到可能解决问题的方案，并利用科研实验和分析，对相关问题的内在本质和规律而进行的调查研究、实验、分析等一系列的活动，为创造发明新产品和新技术提供理论依据，或获得新发明、新技术、新产品。

-- 百度百科

- 科学研究的基本任务就是探索、认识未知和创新。







误区1：做科研 = 写论文

误区

- 我不想做科研，因为不想/不会写论文，只想写代码。
- 我想做科研，但是不会/不擅长写论文。
- ……

正解

➢ **论文是科研成果的一种载体**，用于跟公众/同行分享新知识，做科研通常需要写论文。

- 核心：学会与他人分享你的科研成果。
- 发明专利、软件著作、(开源)软件产品等

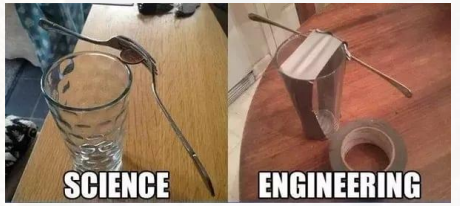
➢ **写论文不是做科研的全部。**

- 读：读文献
- 写：写论文
- 做：做实验
- 听：听报告
- 说：交流，作报告

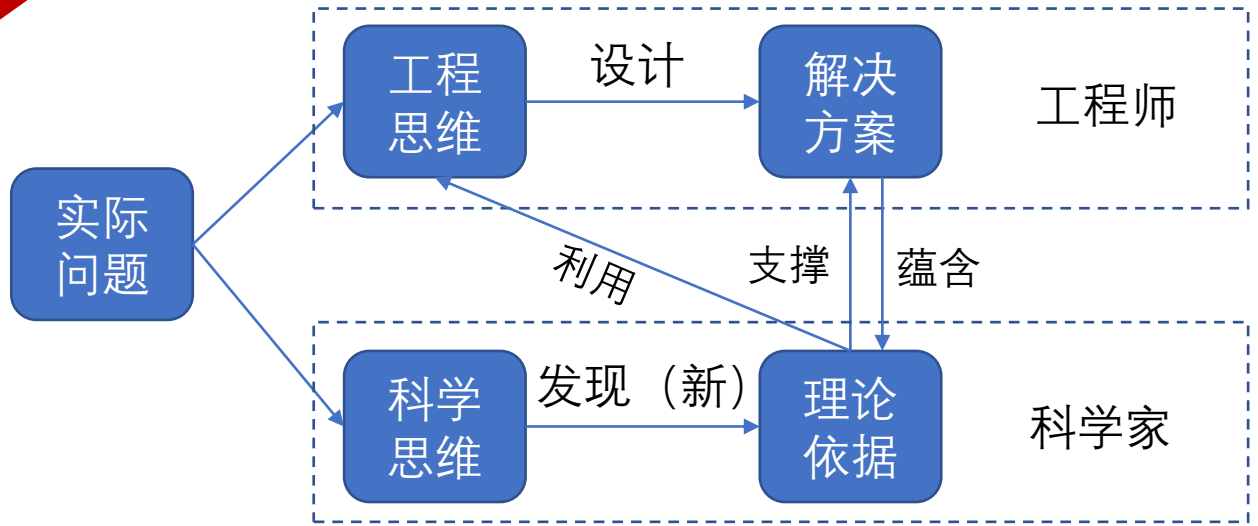
误区2：科研与工程对立

误区

- 我不想做科研，我想做工程，写代码。
- 做工程就是写代码。
- ……



正解



科研和工程都需要具有“解决问题的能力”，科研要在解决问题基础上，进行理论分析。

2. 怎么做科研？

做科研的方法论

1. **如何选定一个感兴趣的研究方向？**
2. **如何增加知识储备，积累阅读量？**
3. **如何确定一个感兴趣的研究问题？**
   - 目前没有被很好解决的（前沿）问题
4. **思考是否有比state-of-the-art更优的方法，如何提出新方法？**
5. **如何验证新方法的有效性？**
   - 根据真实应用场景，进行实例分析
   - 通过理论分析，证明新方法的优越性
   - 在一些数据集上通过实验进行实证

1.如何选方向？

- 如何选定自己的研究领域
  - 导师给你一个确定的研究方向
  - 自力更生
    - 收集近几年相关的优秀论文，整理研究领域的现状，从而找到一个感兴趣的方向
    - 由近及远收集感兴趣方向的全部文献，精读/粗读每篇文献，写出读书笔记
    - 总结出目前存在的问题，已有解的问题和新的研究问题，确定最终科研方向
- 选择最适合你的科研方向
  - 你自己的兴趣
    - 这个是最重要的！
  - 自己的知识积累
    - 数学基础、算法基础、编程基础
  - 能否获得必要的资源
    - 数据
    - 硬件、软件支撑
    - 导师或者实验室是否熟悉该方向

研究方向：图数据挖掘

新应用模型 **人工智能时代涌现了大量基于图数据挖掘和分析的新应用及新模型。**

社交网络分析 | 意见领袖挖掘 | 商品推荐 | 智慧教育 | 智能交通 | 新药发现 | 基因测序

图表示学习	图神经网络	图处理	图分析
(DeepWalk, Node2vec)	(GraphSAGE, GCN)	(图匹配、图采样、图查询)	(图遍历、相似度估计)

图模型 **高效可扩展的图挖掘算法**

研究成果：

- 大规模图学习(NeurIPS '21)
- 图模型蒸馏(KDD '21)
- 深层图神经网络 (KDD '21, TKDE '21)
- 基于图的推荐系统(CSUR '21)

图系统 **自动化的大规模图系统**

研究成果：

- 自动化机器学习 (VLDB '21)
- 可扩展的自动化图学习 (WWW '22)
- 分布式图嵌入(JOS'21)
- 调参与黑盒优化(KDD '21)

图数据 **图数据的标注、清洗、生成、增强和隐私保护等**

社交网络

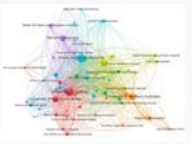
知识图谱

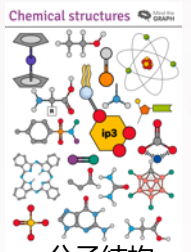

基因网络


交通网络


万维网

物联网


合作网络



分子结构

研究成果：

- 针对超大图的高效标注(SIGMOD '21, VLDB '21)
- 针对稀疏场景的图数据蒸馏(SIGMOD '20)
- 众包场景下带噪声的图标注（ NeurIPS '21 ）

现实世界的图有稀疏、不均衡、噪声以及类型多样化等特点

研究历程

1. 以工业界需求为导向，问题驱动

2. 从数据、模型和系统三个层面来研究图数据挖掘

3. 以开源系统形式来实现学术研究的转化和落地

Angel

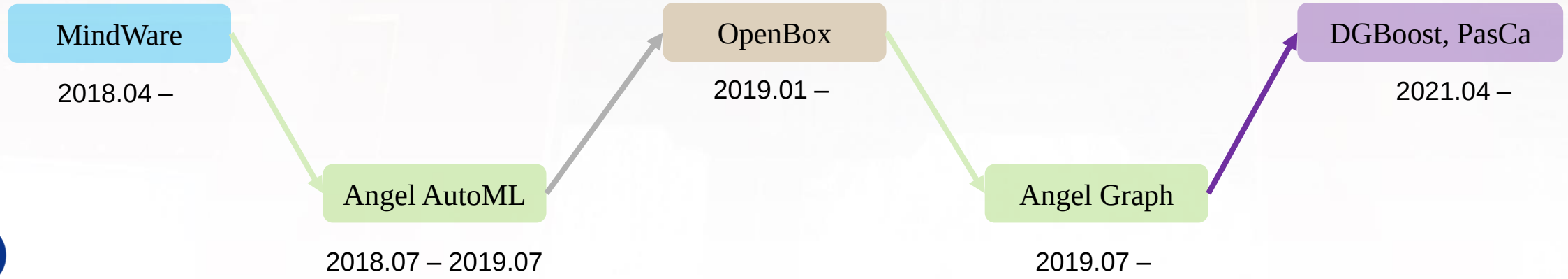
https://github.com/Angel-ML/angel

MindWare
TOOLS FOR DATA INTELLIGENCE

https://github.com/PKU-DAIR/mindware

OPENBOX

https://github.com/PKU-DAIR/open-box



2.如何看论文?

1.被动过滤：略读一些过滤后的潜在高质量工作

现在顶会论文很多，有一些工作并不值得花大量时间精读。很多媒体会把近期某个方向的高质量论文做总结，可以闲暇时间看看，以便快读提升阅读广度。**工具**：公众号，知乎，博客和Google学术推荐等

2.主动检索：主动查询第一步过滤出的感兴趣的工作或本领域课题组工作，略读Title, Abstract, Introduction和Conclusion等。**工具**：Google

3.精读：带着问题看原文中不理解的部分，如：技术细节，实验设置和方案，文章结构和写法等 (跑开源代码前建议先看下issue)

4.存档：保存精读后的论文笔记，比如标记一些表达地道的语句模仿学习。**工具**：Mendeley

5.总结：文章的Motivation和Insight是啥？有哪些不足？能否启发自己正在做的工作？

完善已经有的idea并记录新的idea，定期总结和优化。 **工具**：OneNote和PPT

3.如何发现新问题？

https://www.zhihu.com/question/21646993

- 读论文
  - 找到已发论文的缺点/弱点（方法/实验设置），并解决
- 交叉组合
  - 不同的平台（分布式vs单机，外存vs内存）
  - 不同的数据类型和任务（同构vs异构，无监督vs有监督）
  - 工业界的实际需求 （稀疏，大规模）
  - 复杂的现实环境（考虑一些复杂因素：如带噪声场景下的数据标注）
  - 多读论文——读一些不是本领域的论文
  - 实际生活实际应用中产生的想法
- 实现过程中的分析
  - 只有亲手实现过，才能知道算法/系统的细节
  - 仔细检查每个细节，可能会有所发现
- 讨论
  - 组会报告、他人提问、导师把控

发现问题比提出方法重要得多！
一个好的insight能motivate一系列工作

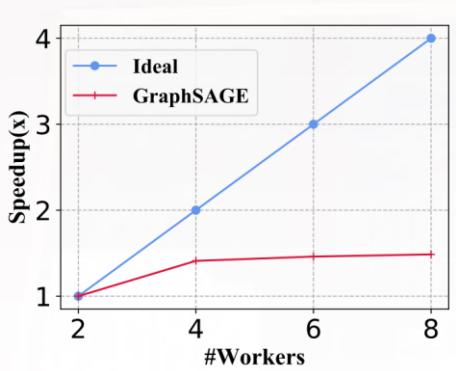
举例：图神经网络的可扩展性

1. 单机场景的存储和效率瓶颈

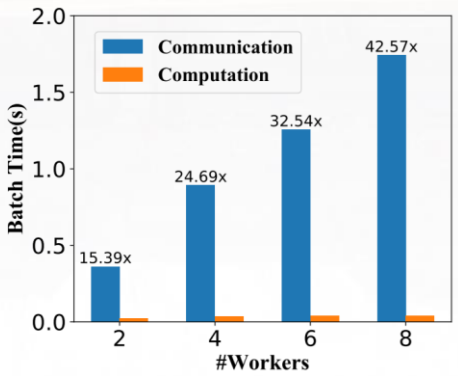
- 受GPU显存限制很难存储超大图
- 大矩阵乘法很耗时

2. 分布式场景的通信瓶颈

- 每批次训练，K层GNN需要拉取K阶以内邻居特征
- 加速比受限制于多机之间的高昂通讯代价



Speedup



Bottleneck

4.如何思考新方法？

- **问题定义**
  - 形式化定义
- **阅读相关文献**
  - 与定义“问题”相关的工作
- **探索方法**
  1. A方法迁移到B问题： Attention机制迁移到GCN → GAT
  2. AB组合法：AutoML组合GNN → GraphNAS
  3. 理论/实验分析+反推方法

举例：理论分析

过平滑问题

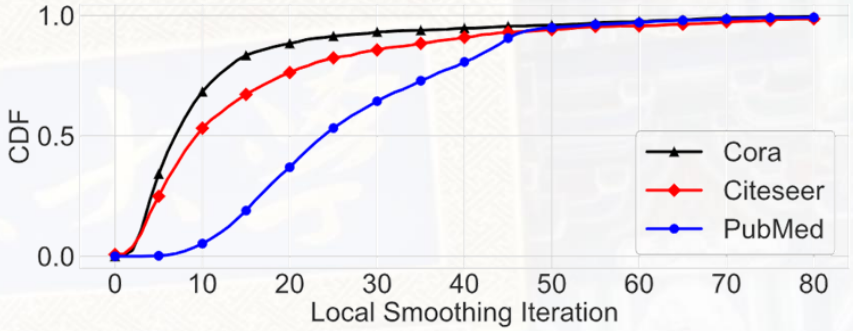
- GNN smooths the representation of each node via aggregating its own representations and the ones of its neighbors.
- Suppose $\widetilde{\mathrm{D}}$ is the diagonal node degree matrix with self loops, and $\widehat{\mathrm{A}} = \widetilde{\mathrm{D}}^{r-1}\widetilde{\mathrm{A}}\widetilde{\mathrm{D}}^{-r}$ is the normalized adjacency matrix. By continually smoothing the node feature with infinite number of propagation in SGC, the final smoothed feature is

$$\mathrm{X}^{(\infty)} = \widehat{\mathrm{A}}^{\infty}\mathrm{X}, \qquad \widehat{\mathrm{A}}^{\infty}_{i,j} = \frac{(d_i+1)^r\left(d_j+1\right)^{1-r}}{2m+n}$$

The final feature is over-smoothed and unable to capture the full graph structure information since it only relates with the node degrees of target nodes and source nodes.

举例：实验分析

(a) Two nodes with different local structures



(b) The CDF of LSI in different graphs

Figure 1: (Left) The local graph structures for two nodes in different regions; the node in dense region has larger smoothed area within two iterations of propagation. (Right) The CDF of LSI in three citation networks.

Idea: Local-Smoothing Iteration(LSI, parameterized by ϵ) is defined as

$$K(i,\epsilon) = \min\left\{k: \left\|\tilde{I}_i - I(k)_i\right\|_2 < \epsilon\right\},$$

Where ϵ is an arbitrary small constant with $\epsilon > 0$, $I(k)_{i,j} = \frac{\partial \hat{X}_{ih}^{(k)}}{\partial \hat{X}_{jh}^{(0)}}, \forall h \in \{1,2,\cdots,f\}$, and $\tilde{I} = I(\infty)$.

Insight: how to control the smoothness in a node dependent way ?

举例：设计方法

Node Dependent Local Smoothing

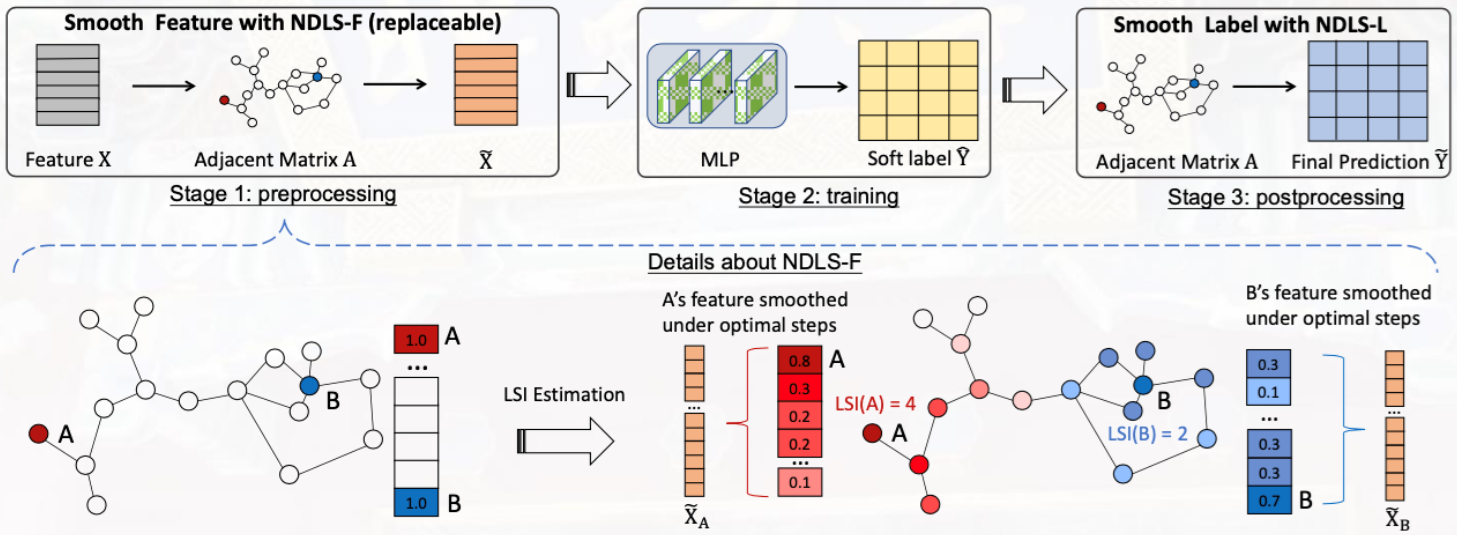


Figure 2: Overview of the proposed NDLS method, containing 1) feature smoothing with NDLS (NDLS-F); 2) model training with smoothed features; 3) label smoothing with NDLS (NDLS-L). NDLS-F and NDLS-L are pre-processing and post-processing steps, respectively.

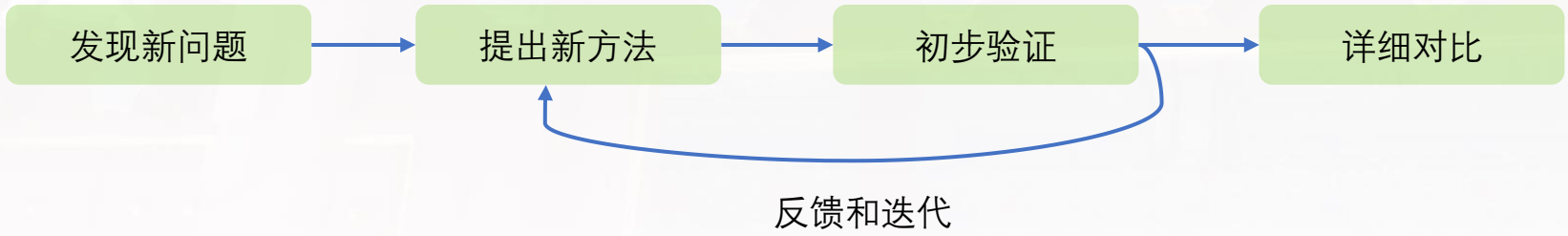
Node Dependent Propagation → P(NDLS-F)-T-P(NDLS-L) ← Node Dependent Propagation

↑

All ML models (MLP, LR, XGB, ect.) can be applied

5.如何验证新方法？

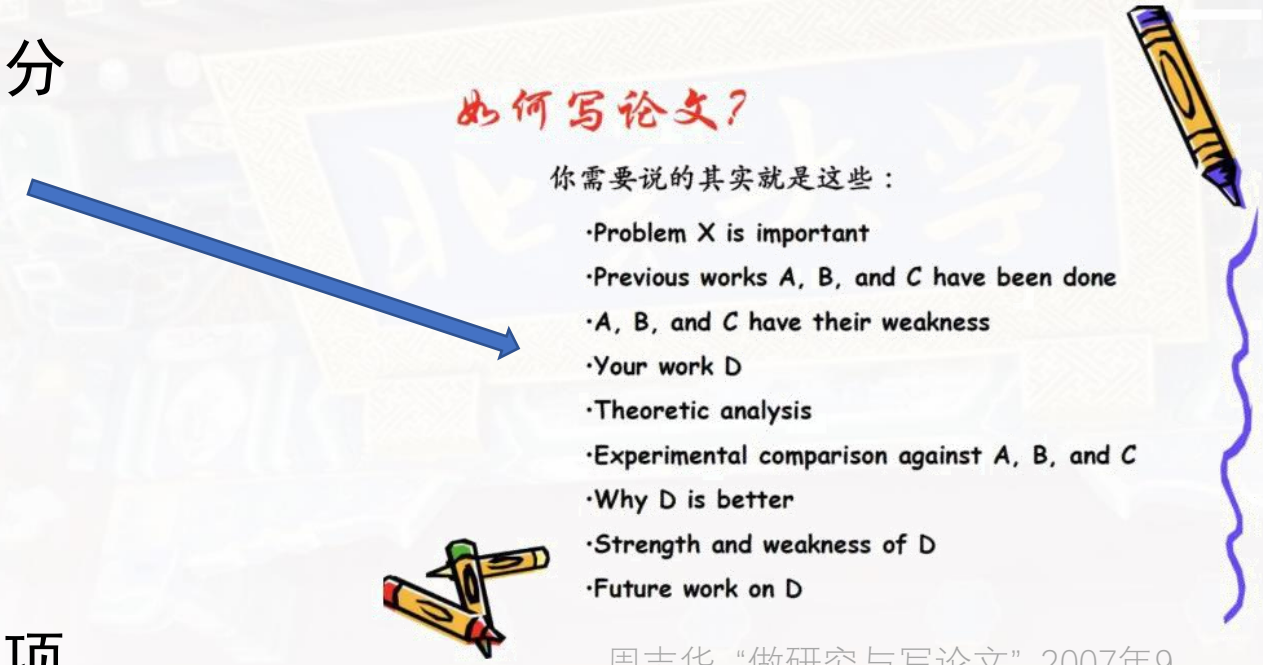
1. 在部分数据集上进行初步测试
2. 根据测试结果调整改进方法
3. 反复迭代1和2
4. 在更多数据集/任务上测试，并对比SOTA方法



3. 怎么写论文？

写论文的方法论

- 论文组成部分
  - 摘要
  - 引言
  - 相关工作
  - 技术部分
  - 实验
  - 结论



周志华, “做研究与写论文”, 2007年9

- 写作注意事项
  - 好的写作不一定能让你的论文变成一篇高水平论文，但高水平论文一定需要好的写作
  - 站在读者的角度去写，从多角度去批判
    - 使用例子和图片说明复杂问题
    - 使用简单句，不要使用复杂句
    - 导师修改以后，多读几遍，研究写作

多读多练!

1.如何讲故事(写Intro)?

1. Why：交代问题背景，为什么要做这个问题?
2. How：这个问题有什么challenge？现有方法有啥缺点?
3. Observation(可选)：介绍方法背后的Insight
4. What：介绍方法和实验效果
5. So what：总结贡献点

举例: Story Line

1 Introduction

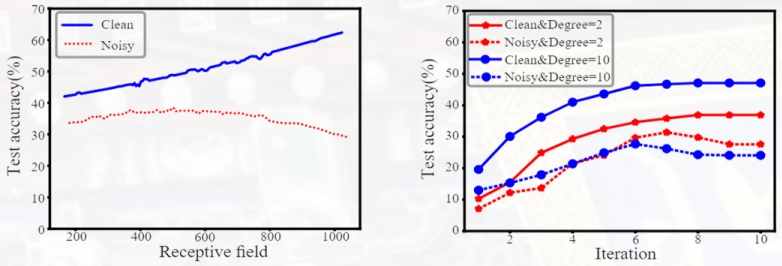
Graphs are ubiquitous in the real world, such as social, academic, recommendation, and biological networks [38, 30, 31, 10, 28]. Unlike the independent and identically distributed (i.i.d) data, nodes are connected by edges in the graph. Due to the ability to capture the graph information, message passing is the core of many existing graph models assuming that labels and features vary smoothly over the edges of the graph. Particularly in Graph Convolutional Neural Network (GCN) [16], the feature of each node is propagated along edges and transformed through neural networks. In Label Propagation (LP) [25], node labels are propagated and aggregated along edges in the graph.

The message passing typically requires a large amount of labeled data to achieve satisfactory performance. However, labeling data, be it by specialists or crowd-sourcing, often consumes too much time and money. The process is also tedious and error-prone. As a result, it is desirable to achieve good classification results with labeled data that is both few and unreliable. Active Learning (AL) [1] is a promising strategy to tackle this challenge, which minimizes the labeling cost by prioritizing the selection of data in order to improve the model performance as much as possible. Unfortunately, conventional AL methods [3, 8, 20, 41, 21, 27] treat message passing and AL independently without *explicitly* considering their impact on each other. In this paper, we advocate that a better AL method should unify node selection and message passing towards minimizing labeling cost, and we make two contributions towards that end.

1.交代问题背景：图模型里的Message Passing

2.交代问题的challenge和现有方法缺点

举例: Story Line



(a) Influence of GCN (b) Influence of LP

Figure 1: The influence between feature/label propagation and test accuracy with clean/noisy label.

The first contribution is that we quantify node influence by how much the initial feature/label of label node v influences the output feature/label of node u in GCN/LP, and then connect AL with influence maximization, e.g., the problem of finding a small set of seed nodes in a network that maximizes the spread of influence. To demonstrate this idea, we randomly select different sets of $|\mathcal{V}_l| = 20$ labeled nodes under the clean label and train a 2-layer GCN with a different labeled set on the Cora dataset [16]. For LP, we select two sets with the average node degree of 2 and 10. As shown in Figure 1, the model performance in both GCN/LP tends to increase along with the receptive field/node degree under the clean label, implying the potential gain of increasing the node influence.

Note that in real life, both humans and automated systems are prone to mistakes. To examine the impact of label noise, we set the label accuracy to 50%, and Figure 1(a) and Figure 1(b) show that the test accuracy could even drop with the increase of node influence under the noisy label. This is because the noise of labels will also be widely propagated with node influence increasing, thus diminishing the benefit of influence maximization. Therefore, our second contribution is that we further propose to maximize the reliable influence spread when label noise is taken into consideration. Specifically, each node is associated with a new parameter called the *quality* factor, indicating the probability that the label given by the oracle is correct. We recursively infer the quality of newly selected nodes based on the smoothing features/labels of previously selected nodes across the graph's edges, i.e., nodes that share similar features or graph structure are likely to have the same label.

3 .方法背后的Observation

4 .基于Observation得到的第一个Insight

5 .基于Observation得到的第二个Insight

举例: Story Line

Based on the above insights, we propose a fundamentally new AL selection criterion for GCN and LP–*reliable influence maximization* (RIM)–by considering both quantity and quality of influence simultaneously. Under a high-quality factor, we enforce the influence of selected label nodes for large overall reaches, while under a low-quality factor, we make mistake penalization to limit the selected node influence. RIM also maintains some nice properties such as submodularity, which allows a greedy approximation algorithm for maximizing reliable influence to reach an approximation ratio of $1 - \frac{1}{e}$ compared with the optimum. Empirical studies on public datasets demonstrate that RIM significantly outperforms the state-of-the-art methods GPA [13] by 2.2%-5.1% in terms of predictive accuracy when the labeling accuracy is 70%, even if it is enhanced with the anti-noise mechanism PTA [6]. Furthermore, in terms of efficiency, RIM achieves up to 4670× and 18652× end-to-end runtime speedups compared to GPA in GPU and CPU, respectively.

6．概括方法和实验结果

In summary, the core contributions of this paper are 1) We open up a novel perspective for efficient and effective AL for GCN and LP by enforcing the feature/label influence with a connection to *social influence maximization* [18, 9, 4]; 2) To the best of our knowledge, we are the first to consider the influence quality in graph-based AL, and we propose a new method to estimate the influence quantity based on the feature/label smoothing; 3) We combine the influence quality and quantity in a unified RIM framework. The empirical study on both GCNs and LP demonstrates that RIM significantly outperforms the compared baselines in performance and efficiency.

7．总结贡献

2.如何提升语言表达？

1. **逻辑组织**：提前设计文章结构模板

误区：想到哪儿写哪儿

2. **细节填充**：模仿学习和总结

1. 打开和待投稿会议相关的3篇文章
2. 总结共性，模仿写作风格（句式和表达）
3. 粗略填充完全文
4. Grammly过语法，细节优化（实验数值和配图）
5. 导师或师兄师姐给反馈
6. 迭代45，直到满意

3.审稿人视角

希望看到的:

1. New Problem – 很重要，但是没被挖掘出来的新问题
2. New Method – 解决某类问题的新方法
3. Nice Story – Intro里吸引人的故事
4. Nice Explanation – 新概念的定义和漂亮的配图

不希望看到的:

1. 问题定义不清晰/不现实
2. 问题老旧且方法不够impressive
3. 方法没有Insight

   不能从理论/实验给出设计方法的Motivation
4. 实验设置不合理

   没有和SOTA方法对比，Toy数据集

PKU-DAIR实验室

北京大学 PEKING UNIVERSITY

北京大学数据与智能实验室(Data and Intelligence Research Lab at Peking University)：
实验室由北京大学计算机系长江学者特聘教授崔斌老师带领，多年来主要在人工智能、大数据等领域进行前沿研究，在理论和技术创新以及系统研发上取得多项成果，已在国际顶级学术会议和期刊发表学术论文100余篇。

开源项目：实验室围绕机器学习系统已经展开了多方面的研究工作，包括机器学习/深度学习系统优化、 AutoML、图机器学习、AI4DB等，发布了多个开源项目：

黑盒优化系统OpenBox
https://github.com/PKU-DAIR/open-box

自动化机器学习系统MindWare
https://github.com/PKU-DAIR/mindware

分布式深度学习系统河图（Hetu）
https://github.com/PKU-DAIR/hetu

OGB榜一解决方案GAMLP
https://github.com/PKU-DAIR/GAMLP

数据库自动调参技术测评与分析
https://github.com/PKU-DAIR/KnobsTuningEA

分布式机器学习平台Angel
https://github.com/Angel-ML

OPENBOX

MindWare
TOOLS FOR DATA INTELLIGENCE

Hetu 河图

Angel

欢迎感兴趣的同学联系实习或交流问题！

扫一扫上面的二维码图案，加我微信

微信：z1299799152

邮箱：wentao.zhang@pku.edu.cn

企业合作:
2017年，课题组与腾讯公司成立北京大学-腾讯协同创新实验室，深度合作并开源了分布式机器学习平台Angel。另外，实验室还与阿里巴巴、苹果、百度、快手、中兴通讯等多家知名企业开展项目合作和前沿探索，解决实际问题，进行科研成果的转化落地。

PKU-DAIR

Q&A

希望能对各位有点帮助!

谢谢!