2021学期汇报

导师：林欣

51194506013 郭晨亮

01

上学期的总结

- (1) 关于作者消歧方法的研究

①投稿论文《基于异构网络的无监督作者名称消歧》 华东师范大学学报（自然科学版） 已录用

②用更多数据集、复现其它论文方法如HRFAENE等改进，但效果不好。

- (2) 关于学术词筛选和分类的研究

①学习了一些关于节点分类的方法，主要是对图网络表示进行研究，包括GNN、随机游走、数据增强等方法。

②对海洋领域的新发现的26535个词进行分类标注，包含9924个学术词和16611个非学术词，按1:1划分训练集，进行筛选学术词训练，达到平均85%的整体准确率。

01 上学期的总结

- (1) 作者消歧

文献数据：(作者author、机构organization、出版机构venue、年份year、标题title、摘要abstract、内容paper、关键词keyword)

附加数据：作者的个人主页、邮箱、地址、文献的引用关系

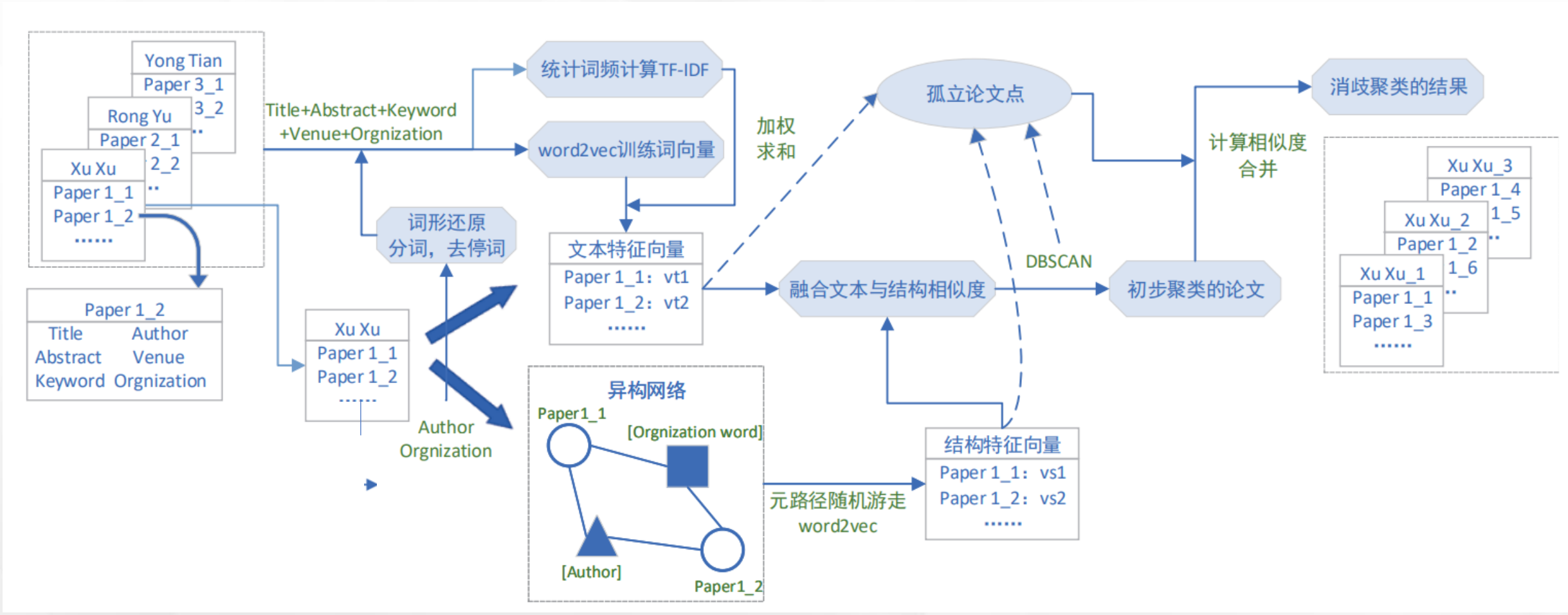
实体：作者、文献、机构、关键词（学术词）

歧义：同一名字的作者有不同文本表示 (词序、缩写、错字)

多个人具有同一名字 (名字本身相同、缩写/出错后相同)

冷启动问题：对于包含同一或接近的作者文本名称的一组文献进行聚类，得到的每个聚类代表一个作者实体。

增量更新问题：在已有聚类的基础上增加新的文献，将作者分类到已有的实体聚类结果中或产生新的实体聚类。



(1)分别学习论文的结构特征、文本特征并融合，根据不同特征的相似度完成聚类

(2)对作者、机构、标题、关键词进行词形还原、标准化的预处理，在词向量的基础上使用了TF-IDF、随机打乱的方法学习文本特征，融合时搜索最优权重

(3)在Aminer数据集、sci数据上进行测试

01

上学期的总结

- (1) 作者消歧

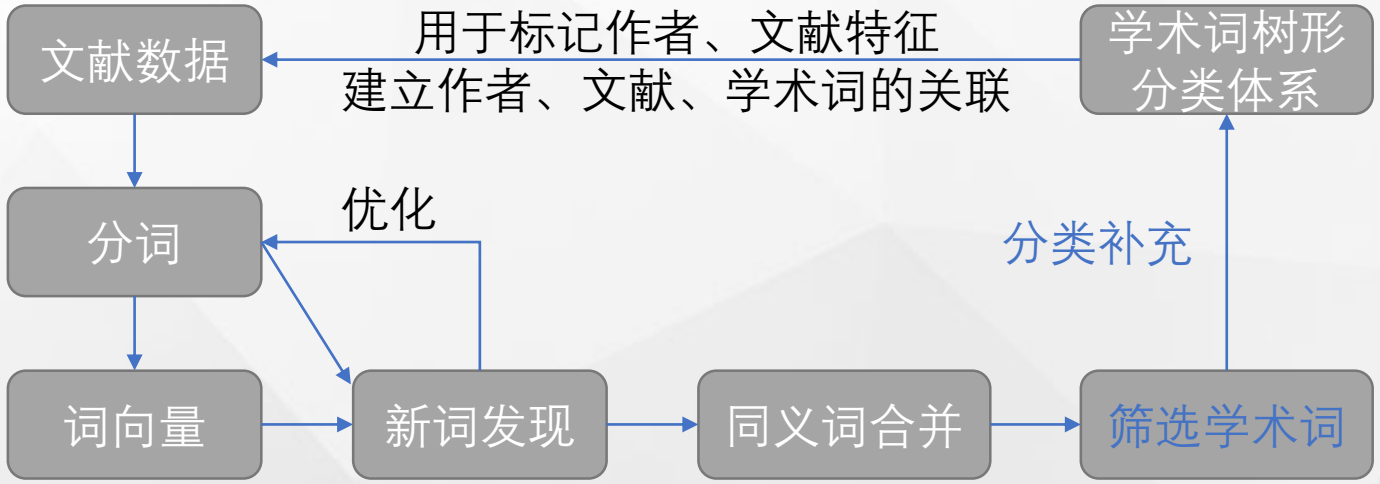
作者名称	本文方法			Aminer 方法			OAG 比赛第一名		
	Rec	Prec	F1	Rec	Prec	F1	Rec	Prec	F1
15 个的均值	45.23	74.79	**56.37**	46.44	70.85	56.10	50.05	52.98	51.47
100 个的均值	65.77	75.21	**70.17**	63.03	77.96	67.79	73.36	60.14	66.10

作者名称	Rec	Prec	F1
abbas h	100.0	100.0	100.0
aalkjaer christian	90.00	100.0	94.74
abel robert	93.69	43.24	59.18
aarabi mahmoud	100.0	100.0	100.0
aamir muhammad	87.50	100.0	93.33
abe yuki	83.33	100.0	90.91
abbasi s a	100.0	100.0	100.0
abe kazuo	87.76	100.0	93.48
abdullah	65.22	48.39	55.56
abab j	80.00	66.67	72.73
均值	88.75	85.83	87.27

	Rec	Prec	F1
原始模型	65.77	75.21	**70.17**
只用结构特征	60.10	65.65	62.75
只用文本特征	87.65	41.04	55.90
去除词形还原	61.25	77.97	68.61
去除 TF-IDF 加权	62.99	75.72	68.77
去除词向量的随机打乱	55.92	78.76	65.40
去除关键词	63.00	75.79	68.81
去除来源	62.24	77.32	68.96
去除摘要	61.18	77.24	68.28

01 上学期的总结

- (2)学术词筛选和分类



1	无海洋学科特征
2	因分词问题不构成词（过短，分词错误）
3	因分词问题不构成词（和、的、与等过长词组）
4	含义宽泛不是专业术语（生活用语，宽泛无科学特征）
5	颗粒度太小（有海洋特征但不适合做关键词）
6	颗粒度过大（有海洋特征但为上层词）
7	参数词、变量、计量单位等
8	有错别字/错误符号
9	翻译不当/表述不准
10	部分/全部英文

在三角形网格的三维表面模型基础上，提出了一种基于aif的三角形网格切割算法。 三角网格
试验表明，本文可有效地提高物方三角网的精度，从而有效地改善重建三 维模型的效果。 三角网格
坐标转换简化为投影点与投影平面均匀三角网格之间的空间关系。 三角网格

对虾养殖塘
浊流滑塌沉积体系
亚热带辐合带
海冰密集度反演
海洋盐差能
水下锚
海底原油管道

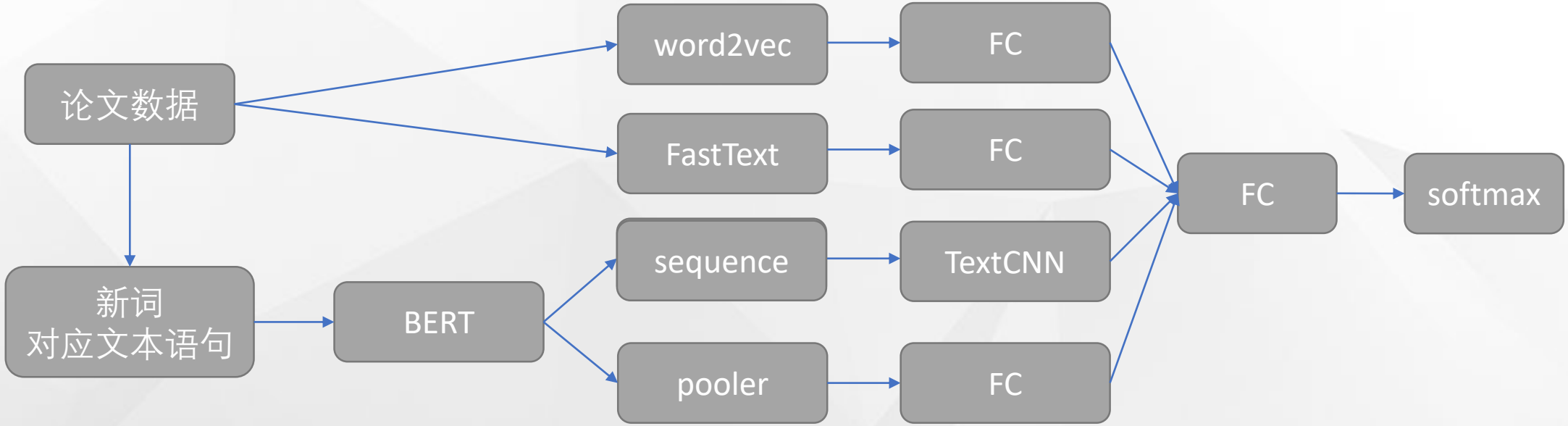
客观分析方法
锆石年代学
模糊算法
科里奥利加速度
粮食生产者
产卵盛期
氦氖激光

D06	海洋
D0601	物理海洋学
D060101	描述物理海洋学
T06010100400	盐楔
……	
D060102	海洋波动
T060102003500	罗斯贝波
……	
D060103	潮汐与潮流
T060103000000	气象潮
……	

01

上学期的总结

- (2)学术词筛选和分类




```
sel: 30 ac: 0.8472854109796785 kc: 0.907940395991835
sel: 33 ac: 0.854185623293906 kc: 0.898754847928148
sel: 36 ac: 0.858431907794965 kc: 0.889977546438048
sel: 39 ac: 0.861919927206551 kc: 0.880791998367013
sel: 42 ac: 0.865862905671822 kc: 0.873647683200653
sel: 45 ac: 0.868516833484986 kc: 0.863645641967748
sel: 48 ac: 0.871094934789202 kc: 0.852622984282506
```

02

接下来要做的研究工作

(1) 阅读短文本分类领域相关文献，将现有方法与其它方法在数据集上进行对比。

(2) 优化现有模型，尝试加入注意力机制、RNN等方法改善效果。

(3) 探索训练数据划分方式对模型效果的影响，用于探索如何选取有代表性的数据进行标注。

(4) 将现有筛选学术词方法应用到学术词分类。