基于“属性-结构”特征融合的交叉领域新兴社群预测

分享人：向荣荣

华中师范大学信息管理学院

Central China Normal University

01 研究目标

研究意义

学科交叉点是当今科技创新的重要突破点。及时有效地识别学科交叉研究领域中的新兴趋势是科技创新关键任务，有助于科研人员和管理人员及早把握相关领域研究方向。

研究现状

1、现在研究主要从主题、社群、网络来理解领域知识内容的变化，注重其语义内容和网络结构的当前揭示和演化趋势的未来预测。

2、社群演化预测多侧重于网络结构特征观测和**演化事件类型识别和预判**，缺乏对**特定类型社群**形成的内在机理揭示且考虑的特征维度较为单一。

研究目标：构建交叉领域**新兴社群识别和预测模型**，助力及时高效地科学知识发现。

02 分析框架

1、社群发现

以e-health作为研究案例，从论文标题、摘要和关键词中提取词汇并逐年筛选累积词汇构建共词网络，并借助Leiden算法进行社群发现；

2、热点/新兴社群判定

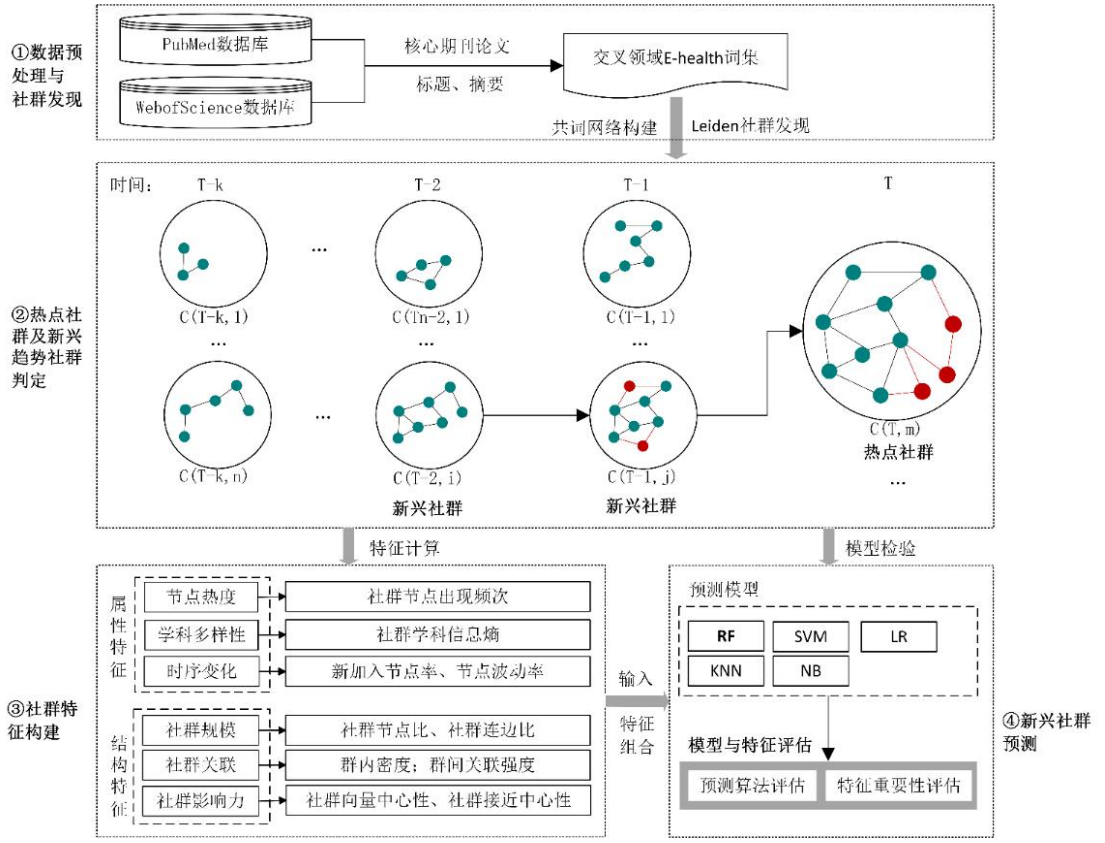
基于逐年时间窗口下的社群，对结果年份进行热点社群判定，并根据每个热点社群回溯历史高相似度且具有高发展潜力社群为新兴社群；

3、社群特征构建

结合领域社群的属性和结构特征，构建新兴社群相关特征；

4、新兴社群预测

基于多元特征和预测算法进行新兴社群预测，并对输入特征和预测算法进行评估。



02 分析框架——新兴社群判定

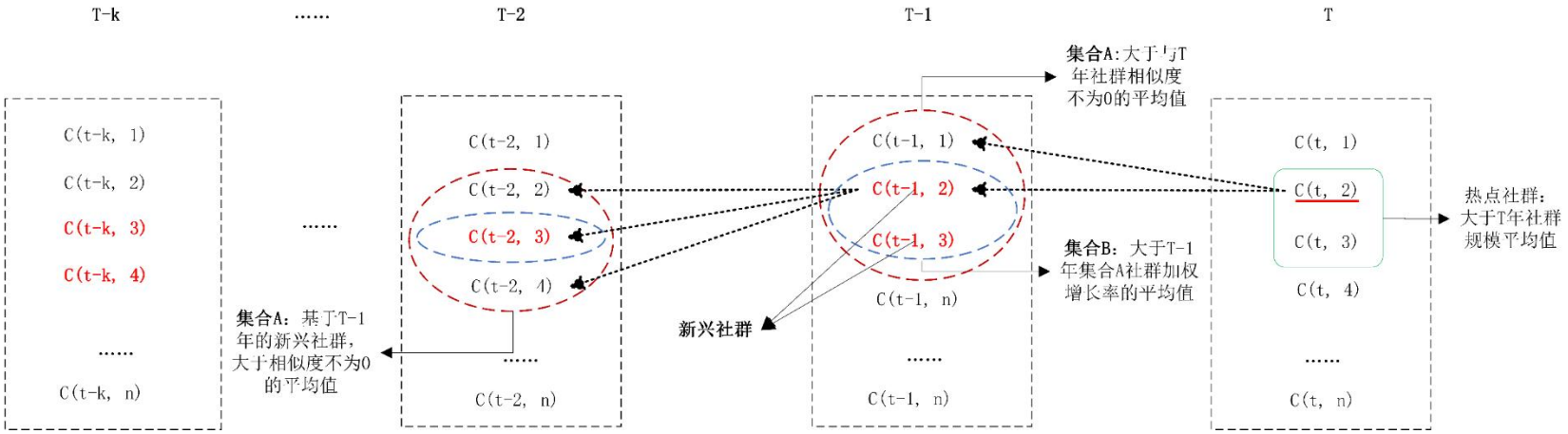
1、新兴社群定义

本文定义的新兴社群指在**未来很有可能成为热点社群**,具有**持续性、快速增长性和发展潜力**的社群

2、新兴社群判定流程

1）一次筛选：**Jaccard系数**找到与热点社群具有较高相似度的社群（*可能成为热点社群、持续性*）

2）二次筛选：基于社群流行度、社群波动率指标并加入词频因素，提出**加权增长率**指标，筛选出处于快速成长阶段，且发展潜力较大的社群。（*快速增长性和发展潜力*）



02 分析框架——社群特征构建

1、社群特征维度划分

结构特征被用来阐释基于网络关系的结构趋势,例如网络中心性、网络密度等；而属性特征是除了结构特征以外的其他属性，例如角色属性、学科属性等。

2、结构特征

1）社群规模：直观描述
2）社群关联度：内部和外部
3）社群影响力：局部和全局

3、属性特征

1）学科特征：学科交叉
2）热度特征：出现频次
3）序列特征：长、短时间变化

特征维度	维度细分	特征说明
属性特征	热度特征	社群节点的出现率
	学科交叉特征	社群节点的学科交叉特征
	序列特征	邻接时间社群节点变化特征
		时间段内社群节点变化特征
结构特征	规模特征	社群节点数量占比
		社群连边数量占比
	关联度特征	社群内节点的紧密程度
		社群间的紧密程度
	影响力特征	社群连接其他社群的能力
		社群被其他社群连接的程度
		社群与其他社群的远近程度
		社群控制其他社群的能力

03 实证分析——描述性统计

本文的数据跨度为21年，首先对2019年的社群进行热点社群判定，得到2019年大于节点数量平均值的热点社群22个。基于社群相似度和增长率，利用2019年的热点社群追溯2000年-2018年的新兴社群演化路径，筛选得到2000年-2018年的新兴社群284个，新兴社群分布情况具体见图3。需要指出的是，社群增长率需要计算社群新加入节点，因此1999年的社群不纳入新兴社群判定范围。

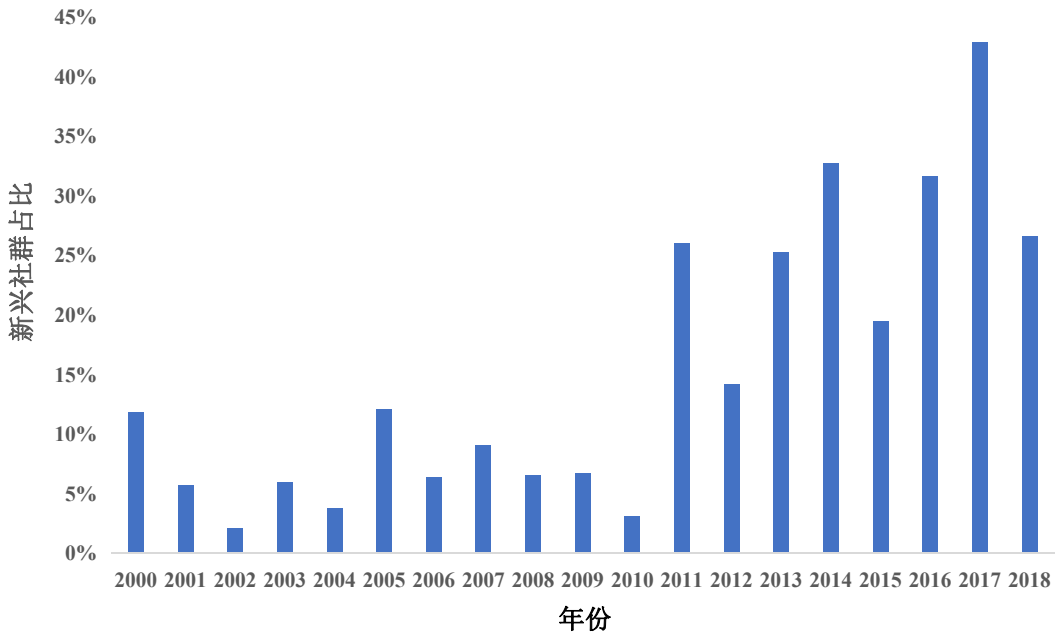


图 2000-2018年新兴社群分布情况

03 实证分析——算法评估

1、基于**属性+结构特征**维度的**RF算法**在新兴社群预测效果上表现最优，体现出特征维度多样性和多预测算法比较的重要性。

表 不同特征维度下新兴社群预测结果

特征维度	预测算法	precision	recall	f1-score
属性特征	LR	80.49	57.89	67.35
	SVM	73.47	62.07	67.29
	NB	75.86	32.35	45.36
	KNN	83.67	63.08	71.93
	RF	91.11	63.08	74.55
结构特征	LR	87.50	16.28	27.45
	SVM	46.67	12.07	19.18
	NB	33.64	75.51	46.54
	KNN	68.75	34.38	45.83
	RF	58.54	42.86	49.48
属性+结构特征	LR	74.51	65.52	69.72
	SVM	88.64	67.24	76.47
	NB	36.09	75.00	48.73
	KNN	88.37	65.52	75.25
	RF	94.44	68.00	79.07

注：加粗部分为同特征维度下F1-score最高的预测算法；评估指标的单位为百分比(%)

03 实证分析——算法评估

2、**3种特征维度下预测算法表现效果比较接近，且评估结果排序高度一致。**其中，所有预测算法均在基于属性+结构特征的特征维度下表现效果最好，NB算法对特征表现最不敏感。

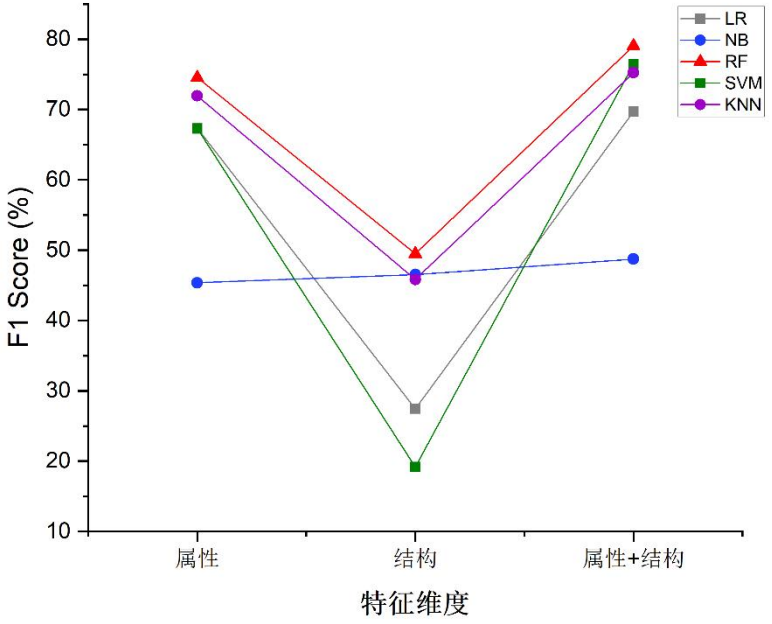


图 不同特征维度下算法预测结果比较

03 实证分析——特征评估

1、针对社群特征**重要性**排序，指标中社群新加入节点比率、社群波动率、社群关系强度贡献度最大，特征中序列特征、热度特征贡献度较大，特征分类中属性特征的重要性大于结构特征。

2、针对社群特征**影响方向**，社群序列特征、社群内关联度、社群全局影响力产生正向影响，社群规模、社群局部影响力、社群热度产生负向影响。

3、针对社群特征**分布差异**，新加入节点比率、社群波动率对新兴社群预测的区分度属于第一梯队。

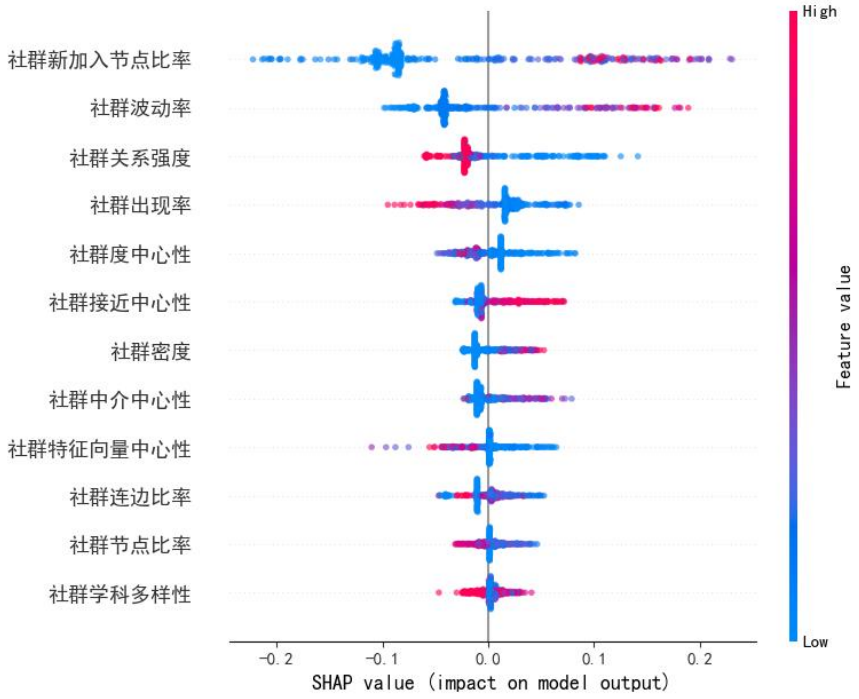


图 基于属性+结构特征维度的RF特征重要性分布

04 研究结论

研究问题:如何判定并预测新兴社群，新兴社群具备什么特征？

01 **新兴社群最优预测方法**

同时考虑属性和结构特征的RF算法模型在新兴社群预测效果上表现最优；

02 **社群特征贡献度**

对交叉领域新兴社群预测而言，社群属性特征重要性大于结构特征，其中社群新加入节点比率和社群波动率对新兴社群预测的贡献度最大且区分度最为明显。

03 **新兴社群特征表现**

社群活跃性更强、内部联系更紧密、全局影响力更大的社群更容易成为新兴社群。

04 研究不足和进一步展望

仅对e-health领域实证

仅采用e-health领域进行实证分析，不同领域的新兴社群可能会呈现出不同的特征，未来可以继续探究其他研究领域；

社群特征构建不全

社群特征构建缺乏社群潜在特征，未来可以融合多学科理论和方法深入挖掘社群潜在特征，如社群语义、角色、行为等特征。

Q&A

谢 谢！

分享人：向荣荣

华中师范大学信息管理学院

Central China Normal University