南京大学
NANJING UNIVERSITY

生成式情报学术语自动抽取与多维关联知识挖掘研究

汇报：闫晓慧　　导师：邓三鸿

作者：胡昊天 邓三鸿 闫晓慧 杨文霞 王东波

南京大学信息管理学院

江苏省数据工程与知识服务重点实验室

南京农业大学信息管理学院

2023.07.13

研究背景与内容

➢ **情报学术语**

- 术语是概念的语言表征，是特定专业领域中一般概念的词语指称
- 情报学术语是情报学学科知识体系的核心构成要素，承载情报学研究历程
- 梳理研究重点与技术演化路径，完善情报学理论方法术语体系

➢ **低资源场景术语抽取**

- 以BERT为代表的自然语言理解（NLU）模型通常需要大规模数据集
- 情报学领域文本缺少高质量的有标签术语语料库

➢ **低资源场景术语抽取**

- 提出低资源场景下生成式自动术语抽取方法
- 开展了情报学领域术语发现及多维知识挖掘

数据源与预处理

➢ **数据源**

- Journal of Informetrics (JOI)
- 2007-2021年间1088篇论文全文数据
- 51,273个段落，252,162个句子，6,074,778个词汇

➢ **数据标注**

- 人工标注成本较高，难以保证标注一致性
- 结合术语词典匹配和人工校对，完成小规模数据的自动化标注
- 领域术语词典：中国人文社会科学术语索引（CSSTI）
- 随机采样12000条句子，最终保留2373条句子作为有标签数据集

面向低资源场景的生成式术语抽取方法

- GPT-3：生成式模型（NLG），可通过输入提示文本（Prompt），以小样本甚至零样本的方式进行交互并生成回答（Answer）。
- 直接将通用GPT-3用于情报学术语抽取容易在生成术语的同时引入噪声。
- 构建少量针对特定任务的情报学“提示-回答”文本对，进行有监督领域微调，提升模型在低资源场景下的学习与泛化能力。

提示（Prompt）：	回答（Answer）：
“by a storytelling-based semantic path analysis”	“path analysis”
“a distance or similarity measure remains open”	“similarity measure”
“was retrieved from this database in September”	“database”
“.....”	“...”

“提示-回答”文本对示例

实验结果分析

- 在低资源场景下，可以有效引导GPT-3输出期望的文本内容与格式，在一定程度上减少人工劳力投入，降低数据成本。
- 通过构造针对特定任务的“提示-回答”文本对微调模型，可以使得面向生成式任务的模型具备信息抽取的能力。
- 远程操作，无法本地部署，在应用成本和数据安全性方面，存在一定缺陷。

各模型性能对比

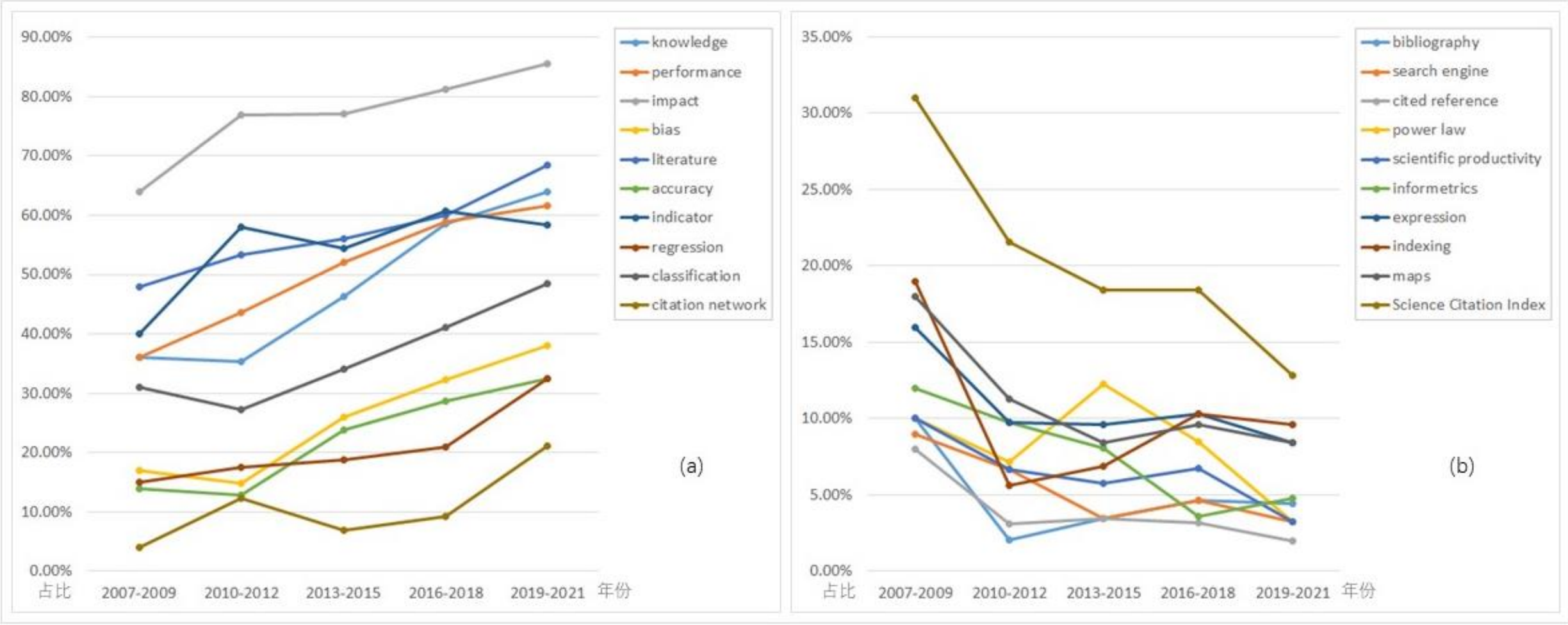
模型	Precision	Recall	F1
BERT	85.71%	86.60%	86.15%
SciBERT	86.15%	87.63%	86.88%
RoBERTa	77.45%	81.44%	79.40%
GPT-3 (curie)	93.75%	82.47%	87.75%
GPT-3 (davinci)	93.77%	82.82%	**87.96%**

新术语发现

出现频次最高的前15个新发现候选术语

序号	新发现术语	新发现术语(译文)
1	citation network analysis	引文网络分析
2	(natural) language processing	自然语言处理
3	hierarchical classification system	层次分类系统
4	coupling network analysis	耦合网络分析
5	HITS authority score	超链诱导主题搜索权威值
6	documentation system	文件系统，文献系统
7	topic mining	主题挖掘
8	neural network learning	神经网络学习
9	big data analysis	大数据分析
10	random search	随机搜索
11	social media mining	社交媒体挖掘
12	corpus tagging	语料库标注
13	smoothing factor	平滑因子
14	term weights	术语权重
15	interview analysis	访谈分析

术语随时间变化趋势分析



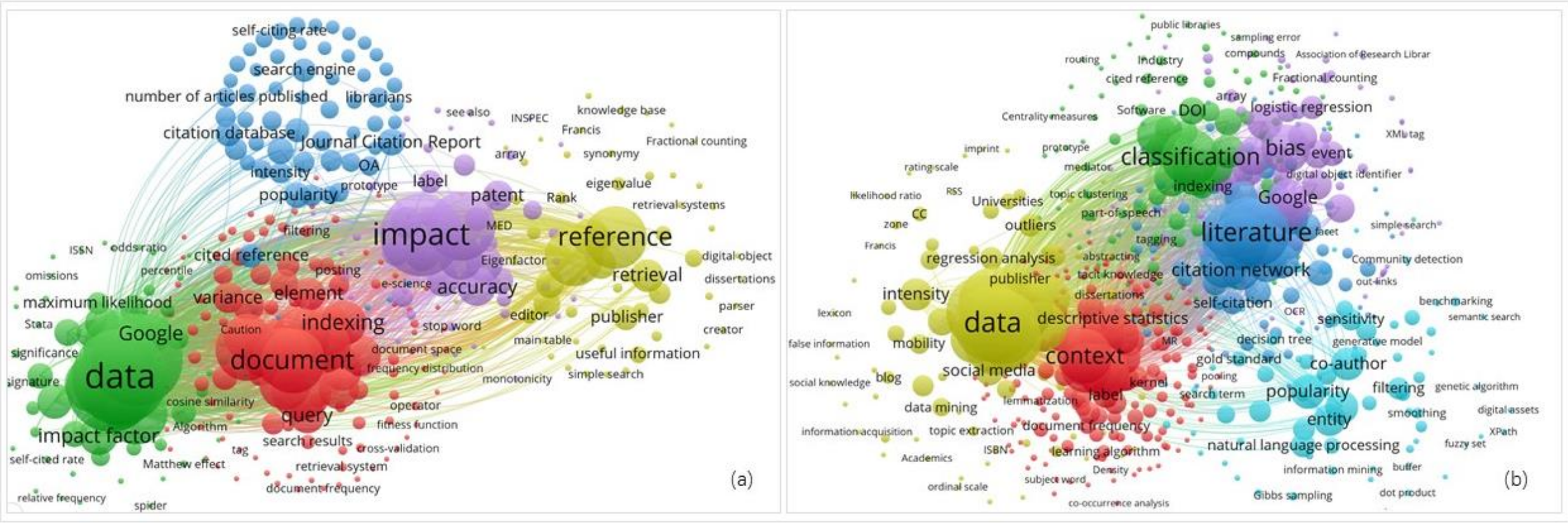
期刊Journal of Informetrics中部分术语变化率折线图

消亡术语与新兴术语分析

期刊Journal of Informetrics部分消亡与新兴术语

序号	消亡术语	对应中文术语	新兴术语	对应中文术语
1	webometrics	网络计量学	feature engineering	特征工程
2	chaining	链化	topic classification	主题分类
3	omissions	漏检	information redundancy	信息冗余
4	e-science	科学研究信息化	information mining	信息挖掘
5	bibliographic record	书目记录	topic clustering	主题聚类
6	intellectual property rights	知识产权	self-efficacy	自我效能
7	document length	文档长度	trigram	三元组
8	synonymy	一义多词	constructivism	建构主义
9	catalog	目录，书目	co-clustering	协同聚类
10	document space	文档空间	multiclass classification	多类别分类
11	ordinal regression	序回归	kappa statistic	kappa 统计量
12	Dewey Decimal Classification	杜威十进分类法	Jaccard coefficient	雅可比系数
13	automatic text classification	文本自动分类	out-links	出链接
14	International Standard Serial Number	国际标准连续出版物号	basic metadata	基本元数据
15	immediacy index	即年指标	response time	响应时间
16	OPAC	联机公共检索目录	Chinese Library Classification	中国图书馆分类法
17	Lotka's Law	洛特卡定律	think tank	智库
18	Rand index	兰德指数	digital assets	数字资产
19	hierarchic clustering	层次聚类	discriminant validity	区别效度
20	MARC	机读目录	reciprocal links	互链

时间维度下研究热点变迁



不同时间区间内情报学术语聚类网络

时间维度下研究热点变迁

- 对信息检索，尤其是搜索引擎相关研究显著减少。
- 更加关注分类与聚类等算法，尤其是算法效率与语言特征，出现了以神经网络和LDA等为代表的自然语言处理技术与文本挖掘方法。
- 依旧注重对引文分析和影响力计算，从引文和元数据分析转向了内容分析与文本分析。
- 对于网络分析，衍生出细化的社交网络分析、引文网络分析，对共作者、共被引、耦合等多种维度的分析。
- 科学计量和文献计量的研究占比更多，而信息计量占比下降。

南京大学
NANJING UNIVERSITY

谢谢

闫晓慧　790751615@qq.com

胡昊天　hhtdlam@126.com

南京大学信息管理学院

2023.07.13