面向应急需求的科技文献推荐与知识抽取

作者：安璐 魏辰瑜　　单位：武汉大学信息管理学院　　汇报人：魏辰瑜

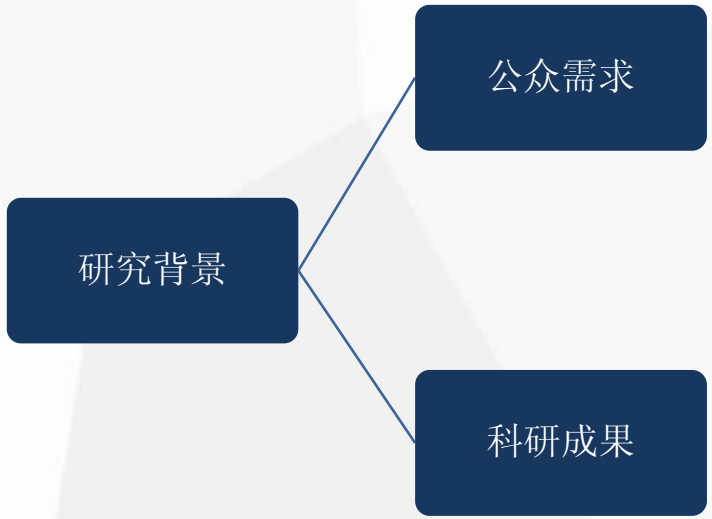
目 录
Contens

绪论

1 研究背景与意义

2 国内外研究现状

研究背景与意义



在突发事件下，公众的信息需求往往并**不能被很好地满足**，这易使公众对事件的认知出现不确定性，进而引发社会恐慌，甚至造成破坏社会和谐稳定的严重后果。

自进入知识经济时代，每日有海量的科技文献资源被公开发表，然而大量的科技知识却不能被公众所知晓和吸收，这造成了大量**科研成果的浪费**。

丰富公众应急需求研究。本文针对公众应急需求的研究有助于丰富应急情境下不同主体信息需求的内容，为应急管理提供可靠依据。

一方面能及时**满足公众的应急需求**，减少用户查找信息的时间和精力；另一方面**实现科研成果的利用**，提高知识利用率，为突发事件下的应急管理提供一定的建议和启发。

国内外研究现状

现存问题

突发事件下的公众应急需求挖掘

①目前有关突发事件的舆情研究大多数基于社交媒体，多从**信息供给方（媒体）**视角来进行研究，而针对**信息接收方（公众）**的需求挖掘研究较少。
②已有研究多将用户的信息需求概括为**宏观**层面的某几个方面。

文献推荐研究

现有文献推荐算法的推荐对象多为**论文作者**，而**鲜有**根据**公众的应急需求**进行文献推荐的研究。

解决方案

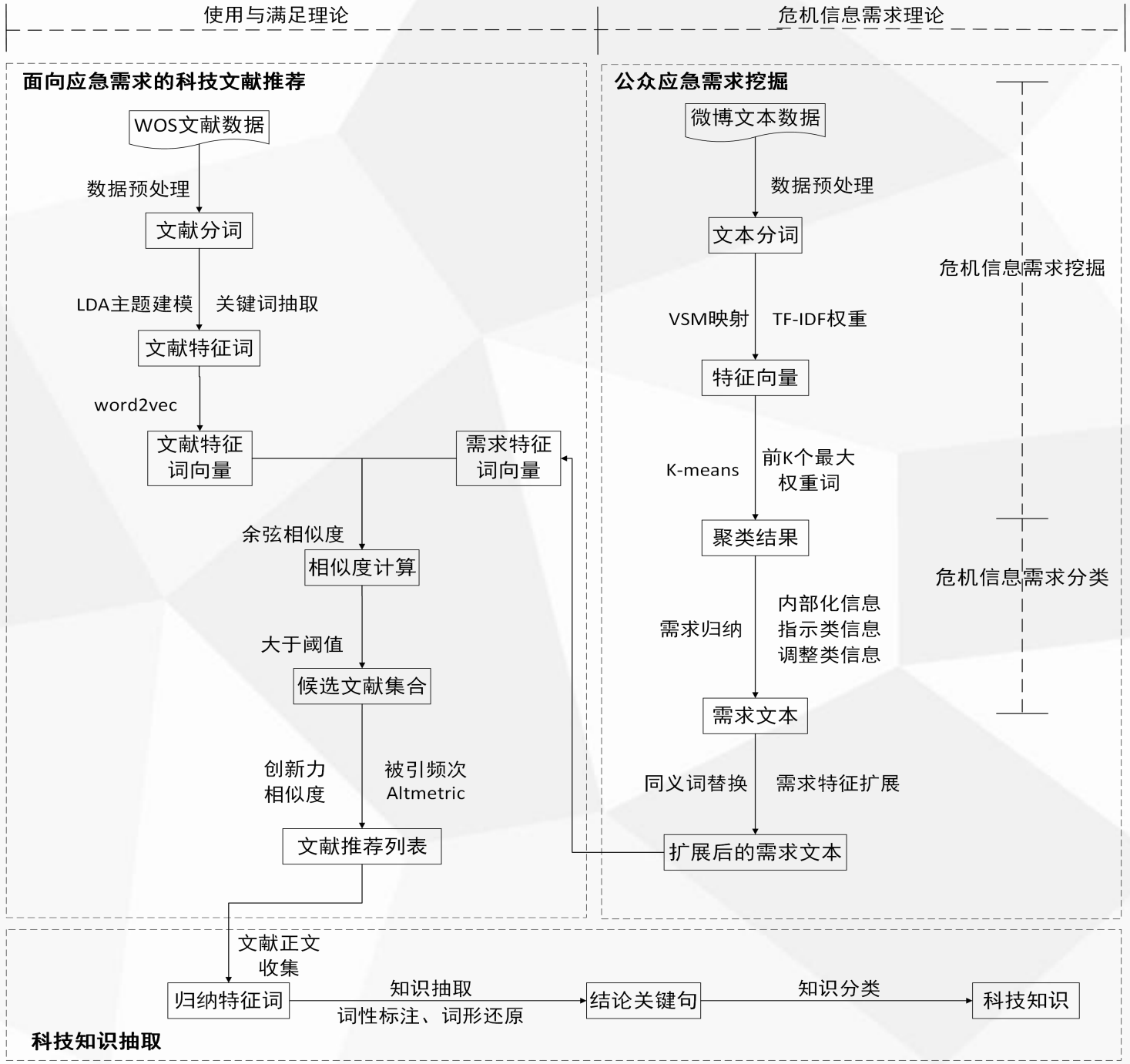
本研究对用户需求进行**细粒度**需求挖掘，反应用户在某一时间段的具体需求。

根据**公众的应急需求**为公众进行文献推荐，实现公众需求和文献两种异构信息的匹配。

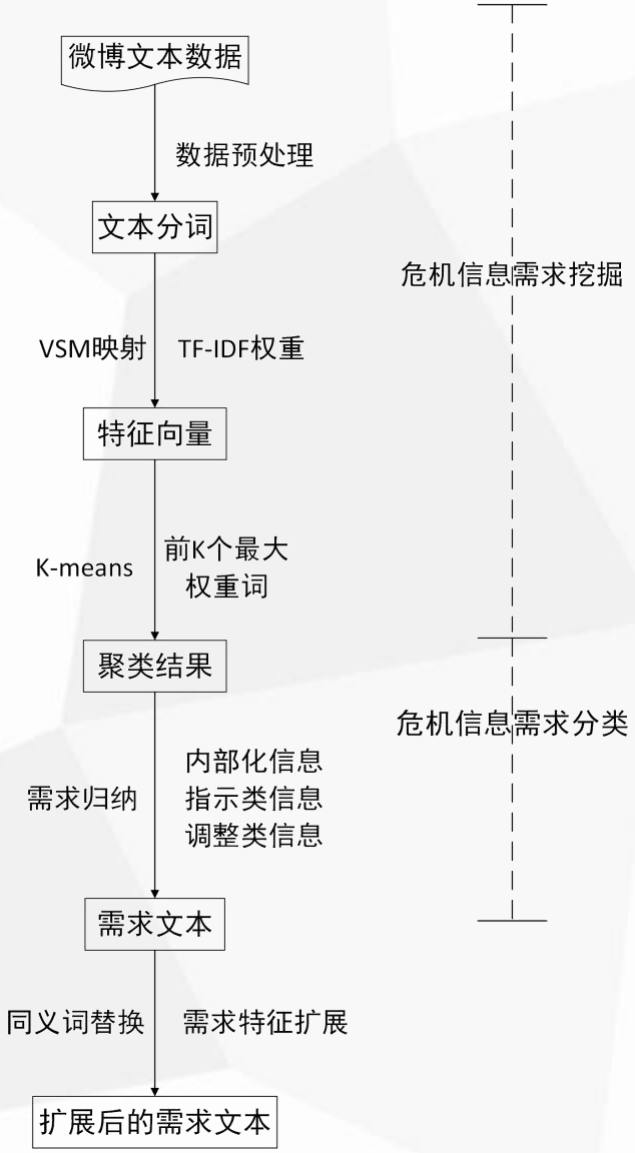
研究设计

1. 公众应急需求挖掘
2. 面向应急需求的科技文献匹配
3. 科技知识抽取

研究框架


使用与满足理论
危机信息需求理论
面向应急需求的科技文献推荐
WOS文献数据
数据预处理
文献分词
LDA主题建模
关键词抽取
文献特征词
word2vec
文献特征词向量
需求特征词向量
余弦相似度
相似度计算
大于阈值
候选文献集合
创新力相似度
被引频次Altmetric
文献推荐列表
公众应急需求挖掘
微博文本数据
数据预处理
文本分词
VSM映射
TF-IDF权重
特征向量
K-means
前K个最大权重词
聚类结果
需求归纳
内部化信息
指示类信息
调整类信息
需求文本
同义词替换
需求特征扩展
扩展后的需求文本
危机信息需求挖掘
危机信息需求分类
文献正文收集
归纳特征词
知识抽取
词性标注、词形还原
结论关键句
知识分类
科技知识
科技知识抽取

公众应急需求挖掘



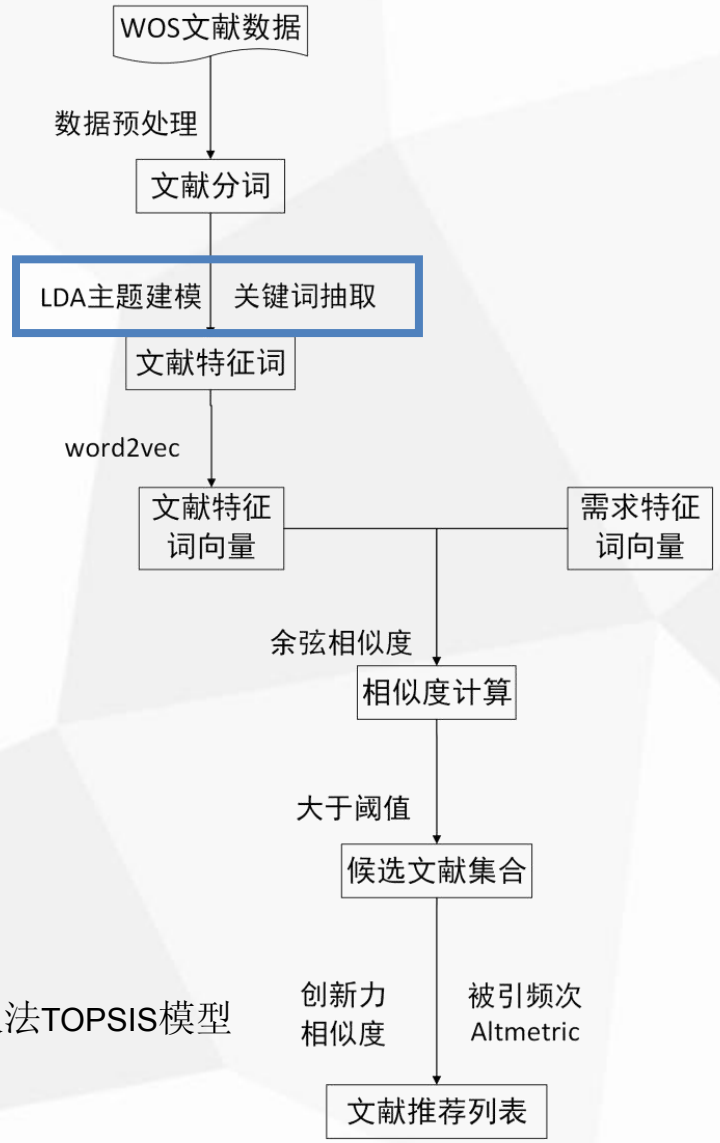
K-Means聚类算法：依据**手肘法**（The Elbow Method）预先确定聚类簇数。手肘法原理为：随着聚类数的增多，当整个数据集的误差平方和（SSE）不再显著下降，则认为聚类趋于稳定。

每个聚类簇返回TF-IDF权重最大的前k个词作为需求关键词。

同义词替换：依据**Sinomed中国生物医学文献数据库**的**主题检索**功能进行同义词替换。

需求特征扩展：首先基于WOS语料库使用Word2Vec训练词向量，将与待扩展词在词向量模型中语义**距离最近**的若干词加入公众需求。

面向应急需求的科技文献推荐



熵权法TOPSIS模型

基于LDA模型的关键词抽取步骤：

①LDA模型 主题Z_k的词汇ω_n分布和文档D_m的主题Z_k分布

$$P(\omega_n|Z_k) = \varphi_{k,n}$$

$$P(Z_k|D_m) = \theta_{m,k}$$

②将原始主题–词矩阵转化为**词汇在主题上的分布向量**，$W_c=[W_{c,1}, W_{c,2}, \ldots, W_{c,n}]$，$W_{c,i}$ 代表词语c在主题i 上的概率分布,K表示主题总数

$$W_{c,i} = \frac{\varphi_{i,c}}{\sum_{t=1}^{K} \varphi_{t,c}}$$

③**余弦相似度**计算**词$W_{c,i}$和文档D_m**的相似度

$$sim(i,j) = cos\,(u,v) = \frac{u \cdot v}{|u||v|}$$

④取相似度最高的前5个词作为文档D_m的关键词

面向应急需求的科技文献推荐

熵权法TOPSIS模型

被引频次通过WOS直接导出
Altmetric值通过Digital Science公司开发的浏览器扩展工具查询得到

文献创新性计算步骤：

①通过CiteSpace获取某学科领域的科学研究前沿主题；

表 1 “long covid-19”研究领域研究前沿主题词

主题词	突显强度
coronavirus disease	10.62
viral shedding	8.36
supply chain	7.9
quarantine	7.23
computed tomography	7
long-term care facilities	6.77
contact tracing	5.87
food security	5.87

②将论文的**研究主题**（通过LDA抽取得到）与科学研究**前沿主题**进行**Jaccard**相似度计算，得出的值为论文的创新力值。A表示需求的关键词集合，B表示文献的关键词集合。

$$Jaccard\ Cofficient = \frac{C(A \cap B)}{C(A \cup B)}$$

熵权法TOPSIS模型计算得分步骤

1.标准归一化处理:

$$a_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{n} {x_{ij}}^2}}$$

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \cdots & a_{ij} & \cdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix}$$

其中，0≤j≤m，m 表示评价指标的个数，n 表示总文献数，x_{ij}表示位于第 i 行第 j 列的原始数据。

2.熵权法确定各指标权重:

①计算概率矩阵P_{ij}

$$P_{ij} = \frac{Z_{ij}}{\sum_{i=1}^{n} Z_{ij}}$$

②计算各指标信息熵e_j

$$e_j = -\frac{1}{\ln(n)} * \sum_{i=1}^{n} \left(p_{ij} \ln(p_{ij})\right)$$

若p_{ij}，则$\lim_{p_{ij} \to 0} p_{ij} * \ln(p_{ij}) = 0$。

③计算指标权重w_j

$$w_j = \frac{1 - e_j}{\sum_{j=1}^{m} (1 - e_j)}$$

3.TOPSIS法建模计算得分：

（1）为评价结果的客观性，根据指标权重w_j创建规范化分析矩阵 B。

$$B = |b_{ij}|_{n \times m} = |w_j \times a_{ij}|_{n \times m}$$

（2）确定正负理想值。正理想解Z^+和负理想解Z^-分别表示第 j 个指标在第 i 篇文献中的最大值和最小值。

$$Z^+ = max\{b_{ij}\}$$

$$Z^- = min\{b_{ij}\}$$

（5）计算距离。选用欧式距离计算文献评价指标到正负理想的距离，D_i^+为第 i 个评价对象到Z^+的距离，D_i^-为 i 个评价对象到Z^-的距离，具体公式为：

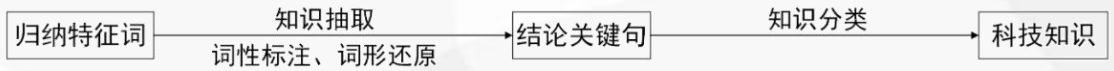
$$D_i^+ = \sqrt{\sum_{j=1}^{m} (Z_i^+ - Z_{ij})^2}$$

$$D_i^- = \sqrt{\sum_{j=1}^{m} (Z_i^- - Z_{ij})^2}$$

（6）计算得分S_i，即计算各评价对象 i 与最优距离的接近程度，S_i越大表示文献价值越高。

$$S_i = \frac{D_i^-}{D_i^+ + D_i^-}$$

基于规则的科技知识抽取



结论关键句判定标准：

（1）从语义上看，句子的内容是对以上或以下内容的简明概括。

（2）句子是对实验结果数据的直接描述，也可以是基于对实验结果的推理和定性的解释。

（3）句子的主语不是引文文献中的对象，如引文作者。

新发现：阐述论文发现的新规律/联系；
新贡献：作者根据研究结论为公众提出的建议和对策。

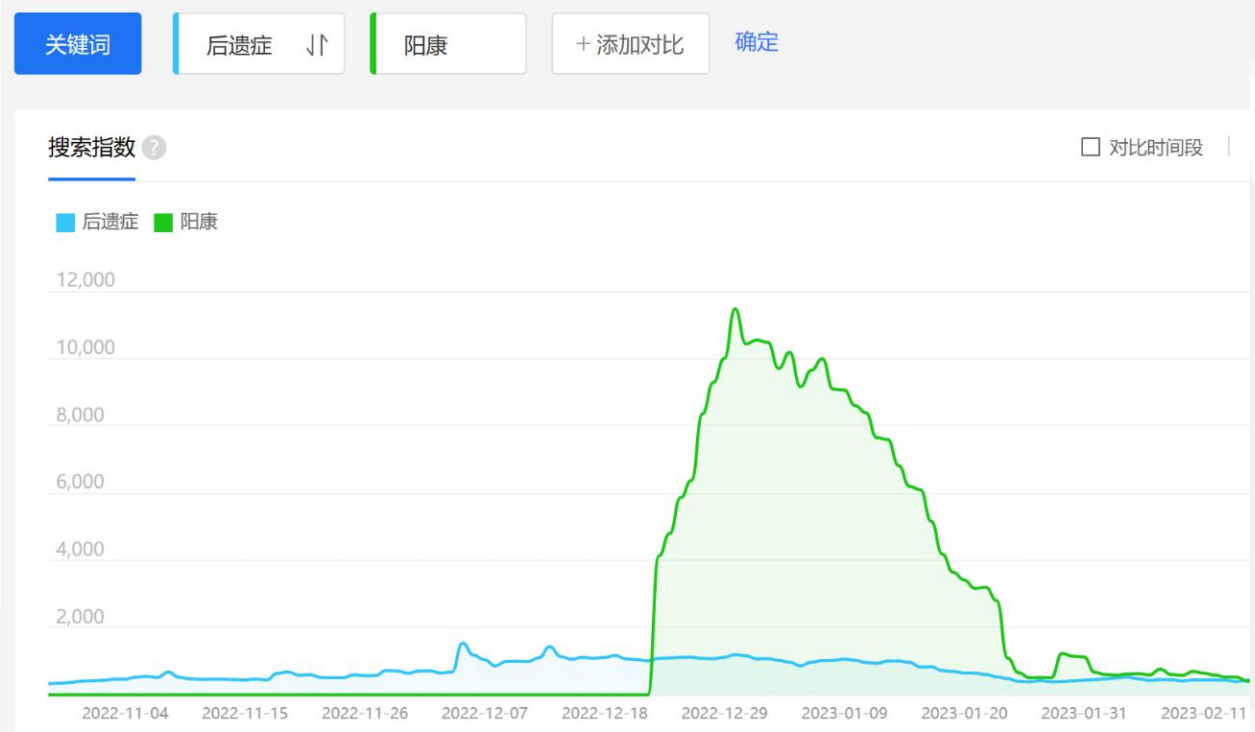
表2 结论特征引导词及例句

类型	序号	结论特征引导词	例句
新发现	1	conclusion	In conclusion, our results seem to indicate that…
	2	identify	this study is the first to identify both…
	3	result	based on our results it seems that infected patients…
	4	suggest	these results suggest that…
	5	found	We found that around 72.3% of our patients….
	6	reveal	the analysis of our sample revealed …
	7	report	as these groups more often reported…
	8	highlight	the data from our study highlight a significant rate…
	9	show	Our study showed a significant association between…
	10	detect	statistically significant differences in …were detected…
	11	confirm	his finding confirmed that …
	12	indicate	the results indicate…
	13	support	our results support that …
	14	contribute	it has been highlighted to contribute to the good health…
	15	recommend	the meddiet-related foods have been recently recommended to be…
新贡献	16	advise	in such conditions…should be… or advised strict self-isolation
	17	advice	our data support the proposal that public health advice…
	18	suggestion	thus supporting our previous suggestion that…

实证结果

1. 公众应急需求挖掘
2. 面向应急需求的科技文献匹配
3. 科技知识抽取

公众应急需求挖掘



需求文本数据来源

平台：微博

工具：Python爬虫

时间范围：2022年12月15日至2023年1月31日

文本检索：按关键词检索的方法爬取了文本中包含 “阳康”“后遗症”与新冠感染相关的微博

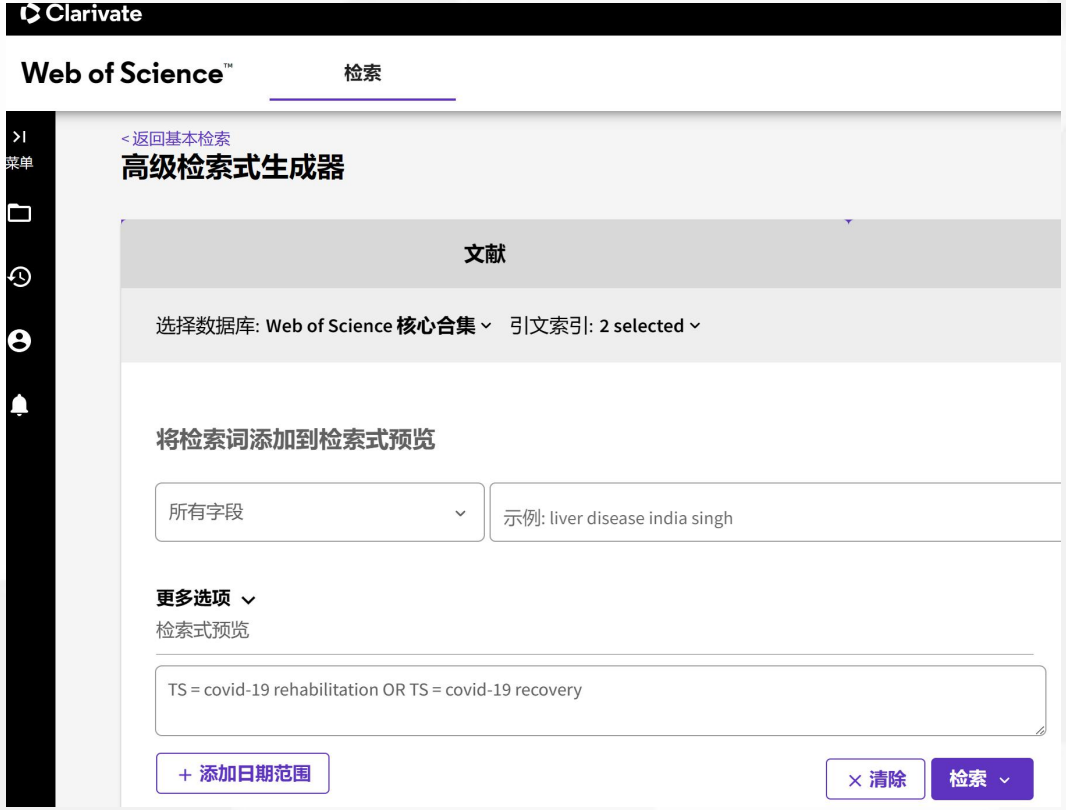
数量：共计**111,934**条数据，其中包含“阳康”一词的共有38,268条数据，包含“后遗症”一词的共有73,666条数据。剔除无关数据后，共剩余**74,787**条数据。

公众应急需求挖掘

表 3 公众应急需求聚类结果

信息类型	需求类型	序号	主题描述	关键词
指示类信息	轻型症状	1	失眠	睡不着 失眠 晚上 睡眠 半夜
		2	嗜睡	嗜睡 睡不醒 失眠 好困 睡觉
		3	嗅觉失灵	味觉 嗅觉 恢复 味道 失灵
		4	咳嗽	咳嗽 嗓子 厉害 ct 不停 白肺
		5	疼痛	疼痛 咽痛 酸痛 肌痛 关节痛
		6	食欲不振	吃饭 恶心 胃口 不香 食欲
		7	头晕乏力	症状 乏力 头晕 疲惫 没力气
		8	听力问题	耳鸣 耳闷 听力下降 听力障碍 中耳炎
		9	眼部问题	眼部表现 结膜炎 畏光 近视 视力下降
		10	心慌	心慌 胸闷 心肌炎 心率 心跳
调整类信息	康后运动	11	剧烈运动引发的症状	剧烈运动 猝死 心肌炎 不宜 低钾血症
		12	科学恢复锻炼	恢复 跑步 锻炼 训练 慢慢
	防护措施	13	人口流动防护	防护 返乡 人口流动 春节 人口流出
		14	居家防护	个人防护 防护措施 预防措施 防护 感染风险
	二次感染	15	复阳	复阳 二次感染 复检阳性 核酸复阳 复阳病例
	心理健康	16	焦虑抑郁	焦虑 抑郁 担心 害怕 难过
	营养管理	17	营养饮食	减肥 运动 肥胖 营养 营养不良

面向应急需求的科技文献推荐



文献数据来源

检索数据库：Web of Science中的SCI和SSCI数据库

检索式：“TS = covid-19 rehabilitation OR TS = covid-19 recovery”

“TS = long covid OR TS = post-covid syndrome”

文献类型：论文

出版时间：2020-2023年

文献总数：共**29378**篇，其中第一条检索式得到的文献为8919篇，第二条检索式得到的文献为20459篇。剔除重复文献后共剩余**28905**篇文献。

面向应急需求的科技文献推荐

基于LDA的主题提取

图1 主题——一致性变化情况

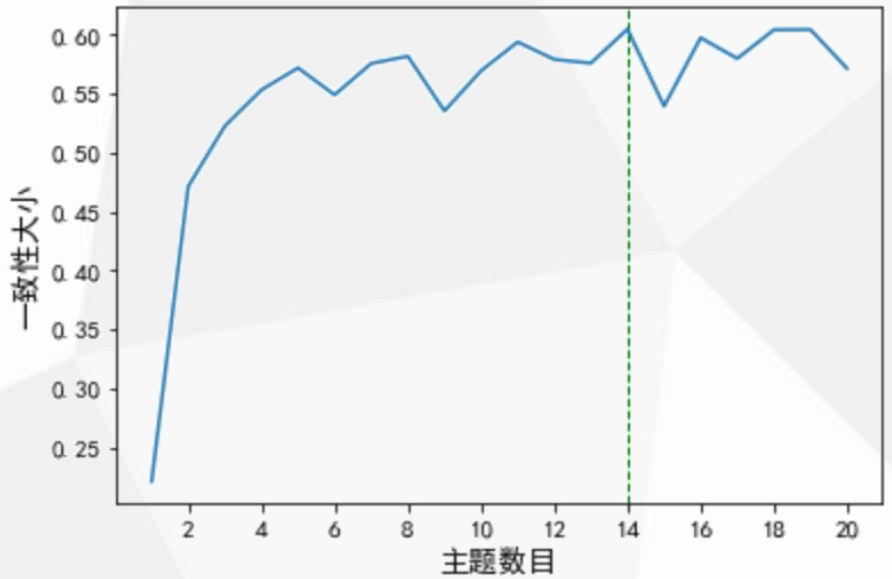


表4 基于LDA模型抽取的文献特征词

序号	文献标题	文献特征词
1	when covid does not heal: post-covid courses with fatigue and exercise intolerance	exercise infectious diseases fatigue several
2	a comparison of pain, fatigue, and function between post-covid-19 condition, fibromyalgia, and chronic fatigue syndrome: a survey study	impacted physical anxiety psychological cognitive
3	the impact of healthcare setting on post-covid mood disorders: a single-centerperspective from southern Italy respiratory intensive care unit	physical phase evaluation mood disorders
4	the kids are not alright: a preliminary report of post-covid syndrome in university students	exercise olfactory pain online students
5	multidimensional phenotyping of the post-covid-19 syndrome: a swiss survey study	pain media headache fatigue online
...	...	...
28905	social isolation in a crisis in hartmann von aue's iwein	physical recovery loneliness moral passive

面向应急需求的科技文献推荐

相似度计算

余弦相似度 Word2vec

表5 需求（嗅觉失灵）与候选文献的相似度计算值

序号	文献标题	相似度
1	exploring trajectory curves from loss of smell and taste in previously hospitalized covid-19 survivors: the long-covid-exp-cm multicenter study	0.497655779
2	prevalence of persistent olfactory disorders in patients with covid-19: a psychophysical case-control study with 1-year follow-up	0.493175699
3	third of patients have gustatory dysfunction 9 months after sars-cov-2 infection: the anosvid study	0.482354901
4	olfaction, anosmia and olfactory rehabilitation	0.463412854
...	...	...
71	post-acute sequelae of sars-cov-2 infection: caring for the 'long-haulers'	0.300152179

表 6 各需求候选相关文献分布

需求编号	需求名称	候选相关文献数量
1	营养管理	6
2	康后运动	777
3	心理健康	859
4	防护措施	34
5	二次感染	744
6	嗅觉失灵	71

表 7 “long covid-19”研究领域研究前沿主题词

主题词	突显强度
coronavirus disease	10.62
viral shedding	8.36
supply chain	7.9
quarantine	7.23
computed tomography	7
long-term care facilities	6.77
contact tracing	5.87
food security	5.87

表 8 需求（心理健康）创新力计算结果

序号	文献标题	创新力值
1	influence of social contacts on corona concerns in the nursing home population quantitative cross-sectional survey	0.64
2	distant from others, but close to home: the relationship between home attachment and mental health during covid-19	0.615384615
3	the enemy who sealed the world: effects quarantine due to the covid-19 on sleep quality, anxiety, and psychological distress in the Italian population	0.592592593
4	stress and associated factors among frontline healthcare workers in the covid-19 epicenter of da nang city, Vietnam	0.592592593
5	anxiety severity, perceived risk of covid-19 and individual functioning in emerging adults facing the pandemic	0.592592593
...	...	...
859	the impact of loosening covid-19 restrictions and live-in partner status on sexual and mental health in a Canadian sample	0

面向应急需求的科技文献匹配

文献价值量化与排序-文献推荐结果

表 9 需求（心理健康）推荐论文的指标数值信息

排名	论文标题	相似度	创新性	被引频次	Altmetric	综合得分
1	the enemy who sealed the world: effects quarantine due to the covid-19 on sleep quality, anxiety, and psychological distress in the Italian population	0.5090	0.5926	401	114	0.6492
2	influence of social contacts on corona concerns in the nursing home population quantitative cross-sectional survey	0.4525	0.6400	1	1	0.5485
3	stress and associated factors among frontline healthcare workers in the covid-19 epicenter of da nang city, Vietnam	0.4604	0.5926	8	0	0.5182
4	distant from others, but close to home: the relationship between home attachment and mental health during covid-19	0.3383	0.6154	21	23	0.4926
5	anxiety severity, perceived risk of covid-19 and individual functioning in emerging adults facing the pandemic	0.3305	0.5926	21	1	0.4678
6	the impact of social quarantine on the living status and mental health of the elderly in the Wuhan community: one year after Wuhan covid-19 blockade	0.3305	0.5926	0	1	0.4645
7	impact of the covid-19 lockdown on a long-term care facility: the role of social contact	0.3141	0.5926	7	3	0.4602
8	the role of psychosocial factors in explaining sex differences in major depression and generalized anxiety during the covid-19 pandemic	0.6094	0.2963	1	3	0.3416
9	two-years follow-up of symptoms and return to work in complex post-covid-19 patients	0.3550	0.2857	0	847	0.3317
10	psychological impact and coping strategies of frontline medical staff in Hunan between January and march 2020 during the outbreak of coronavirus disease 2019 (covid-19) in Hubei, China	0.3359	0.2857	442	3	0.2870

科技知识抽取

表10 关键句分布统计

需求编号	新发现	新贡献	结论关键句总数
1	53	4	57
2	66	5	71
3	59	2	61
4	55	3	58
5	43	0	43
6	56	4	60
平均数	55.3	3	58.3

表 11 嗅觉失灵主题的知识抽取结果（部分）

文献id	新发现	新贡献
20939	1.This study is conducted to identify the prevalence of olfactory and/or gustatory dysfunction in patients with coronavirus disease in northern part of India. 2.Which suggest that self-reported od or gd underestimates the frequency of chemosensitive disorders. 3.We found that most patients with od and/or gd in age group of 18-35 years with male gender predominance. 4.cultural socio-economic status level of education/awareness and willingness to participate in the study are possible reasons for these results. 5.All of these limitations should be considered in future studies to investigate and characterize the olfactory and gustatory dysfunctions in covid-19 patients with a possibility of developing interventions which can result in early recovery of symptoms. 6.However our study has shown that new onset od and gd are early and frequent presenting symptom in covid-19 patients.	1.As on date, the European rhinologic society (ers) and the ent UK association advises against giving systemic corticosteroids to patients with sudden olfactory dysfunction since recovery can occur in the first weeks after onset. 2.Administration of intranasal corticosteroids (incs) in patients with anosmia is also controversial and ers recommends against its use. 3.It is recommended that otorhinolaryngologists and even general practitioners should focus on recent onset od and/or gd as significant symptom.

新发现：阐述论文发现的新规律/联系；
新贡献：作者根据研究结论为公众提出的建议和对策。

研究总结

1 工作总结

2 不足与展望

研究总结

挖掘公众应急需求

根据危机信息需求理论将公众应急需求归纳为六类，分别为**轻型症状、康后运动、防护措施、二次感染、心理健康、营养管理**。

面向应急需求的科技文献推荐

基于LDA主题模型筛选出了与需求相关的文献，应用熵权法TOPSIS模型对文献的**相似度、创新性、学术影响力、Altmetric**四项指标进行综合评价

科技知识抽取

本研究基于**规则抽取**的方法，抽取科技文献的结论句作为科技知识。

不足与展望

需求挖掘存在局限

仅对新冠流行期间某一时间段的公众需求进行挖掘。未来可扩大时间范围，对不同时间段的公众需求进行挖掘，从而归纳不同时期公众的应急需求。此外，也可挖掘不同突发事件下的公众需求，丰富研究主题。

未进行细粒度的知识抽取

本研究基于规则抽取的方法，抽取科技文献的结论句作为科技知识。未来针对科技知识的抽取可细化至具体的知识单元，建立知识图谱，更精准、全面地满足公众应急需求。

参考文献

[1] 李月琳,张建伟,包虹虹.突发公共卫生事件情境下大学生的信息需求及满足程度[J].图书情报工作,2020,64(22):85-95.

[2] Fang Y B, Jia J W, Li J Z, et al. Analysis of public information demand during the COVID-19 pandemic based on four-stage crisis model [J]. Frontiers in Physics, 2022. 10.

[3] 杨康，杨超，朱庆华. 基于社交媒体的突发公共卫生事件公众信息需求与危机治理研究[J]. 情报理论与实践. 2021, 44(03): 59-68.

[4] Tang S, Wu X, Chen J, et al. Release and demand of public health information in social media during the outbreak of covid-19 in China [J]. Frontiers in Public Health. 2022, 9: 2296-2565.

[5] 陈长华，李小涛，邹小筑，等. 融合Word2vec与时间因素的馆藏学术论文推荐算法[J]. 图书馆论坛. 2019, 39(05): 110-117.

[6] 熊回香，孟璇，叶佳鑫. 基于关键词语义类型和文献老化的学术论文推荐[J]. 现代情报. 2021, 41(01): 13-23.

[7] 李晓敏，王昊，李跃艳. 基于细粒度语义实体的学术论文推荐研究[J]. 情报科学. 2022, 40(04): 156-165.

[8] Haruna K, Ismail M A, Bichi A B, et al. A citation-based recommender system for scholarly paper recommendation[C]// Computational Science and Its Applications – ICCSA 2018. Berlin: Springer, Cham, 2018: 514-525.

[9] 朱祥,张云秋,惠秋悦.基于学科异构知识网络的学术文献推荐方法研究[J].图书馆杂志,2020,39(08):103-110.

[10] 潘峰, 怀丽波, 崔荣一. 基于分布式图计算的学术论文推荐算法[J]. 计算机应用研究, 2019,36(06):1629-1632.

[11] 索传军,盖双双.单篇学术论文的评价本质、问题及新视角分析[J].情报杂志,2018,37(06):102-107.

[12] 楼雯,蔡蓁.科学论文评价的涵义与方式研究综述[J].情报杂志,2021,40(05):171-177.

[13] 罗卓然,王玉琦,钱佳佳等.学术论文创新性评价研究综述[J].情报学报,2021,40(07):780-790.

[14] 杨建林,钱玲飞.基于关键词对逆文档频率的主题新颖度度量方法[J].情报理论与实践,2013,36(03):99-102.

[15] 潘菲,王效岳,白如江等.研究主题视域下零被引与高被引论文分析——以环境科学领域为例[J].图书情报工作,2018,62(20):77-87.

[16] Zahedi Z, Costas R, Wouters P. How well developed are Altmetrics? a cross-disciplinary analysis of the presence of ‘alternative metrics’ in scientific publications[J]. Scientometrics, 2014(101):1491-1513.

参考文献

[17] Shuai X, Pepe A, Bollen J. How the scientific community reacts to newly submitted preprints: article downloads, twitter mentions, and citations[J]. Plos One, 2012,7(11).

[18] 郑美莺,梁飞豹,梁嘉熹.单篇论文评价方法——PaperRank算法[J].科技与出版,2016(07):94-98.

[19] 李冬梅，张扬，李东远，等. 实体关系抽取方法研究综述[J]. 计算机研究与发展. 2020, 57(07): 1424-1448.

[20] 张帆，乐小虬. 面向领域科技文献的句子级创新点抽取研究[J]. 现代图书情报技术. 2014(09): 15-21.

[21] 孙明珠,马静,钱玲飞.基于文档主题结构和词图迭代的关键词抽取方法研究[J].数据分析与知识发现,2019,3(08):68-76.

[22] 杨京,王芳,白如江.一种基于研究主题对比的单篇学术论文创新力评价方法[J].图书情报工作,2018,62(17):75-83.

[23]奉国和,周榕鑫,武佳佳.基于熵权TOPSIS及因子分析的学术期刊综合评价研究[J].图书情报工作,2018,62(17):84-95.

[24] 沈雪莹,欧石燕.科学文献知识单元抽取及应用研究：梳理与展望[J].情报理论与实践,2022,45(12):1.95-207.

[25] 李瑛, 周立. 科技期刊论文创新点合理呈现的价值及理想模式[J]. 中国科技期刊研究, 2018,29(10):993-999.

[26] 温有奎,吴广印.碎片化科研创新点动态挖掘研究[J].数字图书馆论坛,2014,122(07):25-32.

[27] 曹树金,闫欣阳,张倩等.中外情报学论文创新性特征研究[J].图书情报工作,2020,64(01):80-92.

[28] Lin L, Wang D, Shen S. Extraction of thesis research conclusion sentences in academic literature[C]// The ACM/IEEE Joint Conference on Digital Libraries 2021. Proceedings of the 2nd Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents. IEEE ,2021:74-76.

[29] 唐晓波,高和璇.基于关键词词向量特征扩展的健康问句分类研究[J]. 数据分析与知识发现, 2020,4(07):66-75.

论文汇报完毕，感谢各位老师聆听!

安璐：anlu97@163.com　　魏辰瑜：2019301040119@whu.edu.cn