



西安交通大学图书馆
XI'AN JIAOTONG UNIVERSITY LIBRARY

2023 CALIS第二十一届引进数据库培训周

“数字学术服务创新与发展”研究课题 效用视角下核心引文识别与数字资源评估初探

Preliminary Study on Core Citation Recognition and Digital Resource Evaluation from the Perspective of Utility

西安交通大学图书馆

李凌

2023年5月



汇报提要

1

研究背景

相关研究介绍，包括：引文分类体系、引文分类标注方法、引文分类应用等研究。

2

研究思路

引文效用分类特征选择与自动提取，核心引文识别模型构建及其在数字资源评估的应用。

3

研究主要内容

引文特征自动提取与核心引文识别主要步骤及方法介绍。

4

研究应用拓展

学科资源保障分析
数据库资源保障分析

5

启发及后续研究

引文特征识别与自动分类及其实践应用新思路。





研究背景

传统引文分析主要基于**引用频次**来量化分析文献间的引用关系，并将其作为评判资源质量、保障率、以及学术影响力的重要标准。

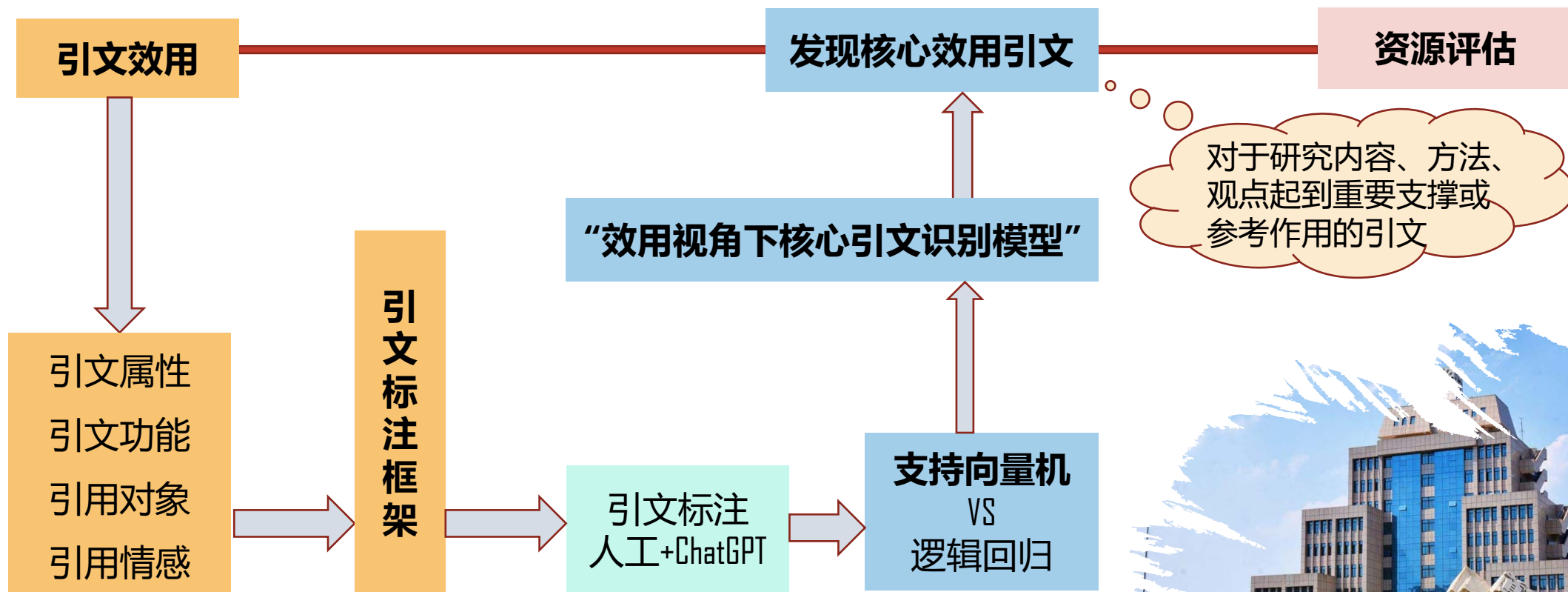
基于引用频次
简单、直观
忽略了引文效用差异

为避免单一频次分析带来的弊端，将引文分析逐步拓展到内容层面，通过对引文内容特征、分类方法等方面的探索，发现引用行为特点、引文功能差异等，并以此为基础开展多种引文评价及影响力评估的研究与实践。

- ? 思考一：分类体系缺乏统一标准、指标粒度差异明显、通用性不高
- ? 思考二：研究内容丰富，但基于引文效用分析的应用实践较少
- ? 思考三：多以人工标引为主，利用深度学习算法来实现引用功能自动分类，但其分类效果往往与标注语料的规模和质量密切相关。由于大规模语料标注较为困难，因此深度学习方法分类效果的优劣还有待于进一步验证。

西安交大创新港图书阅览中心

研究思路





研究主要步骤



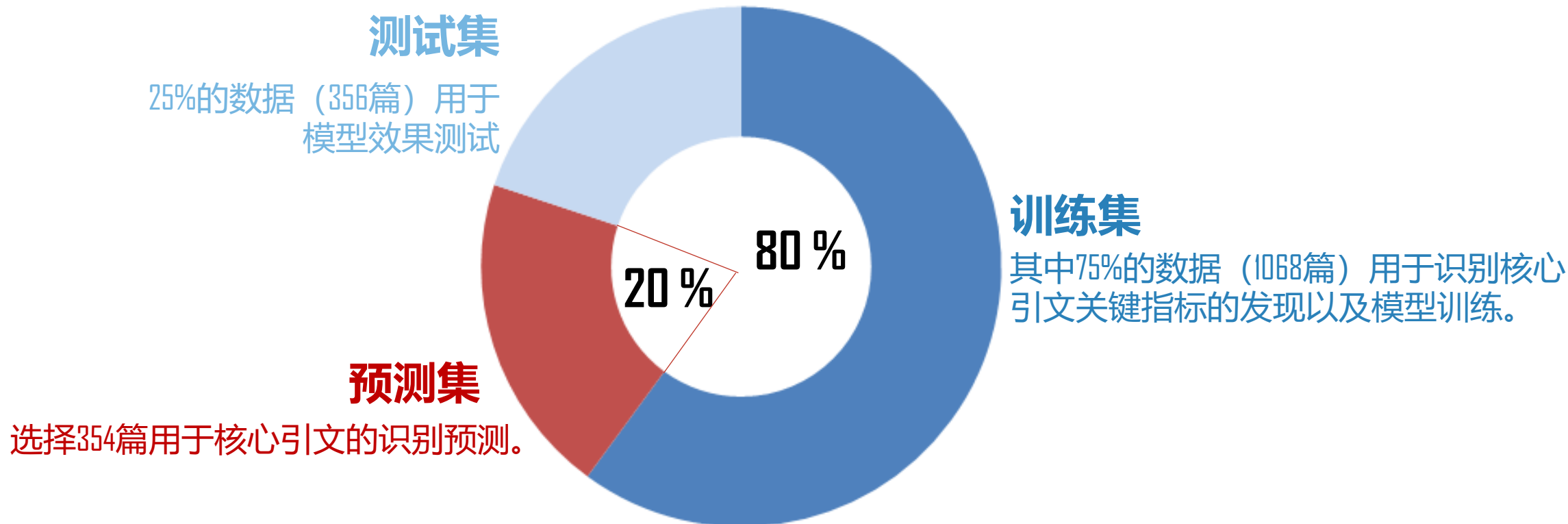


1. 数据获取

- 西安交通大学机构知识库
- 医学部2022-2023年高质量论文
- 选择 “article”的引文

1778篇引文

随机选取1424篇 (80%) 用作模型构建
354篇 (20%) 用于预测



2.引文内容标注框架设计



引文特征属性

年份、所在期刊、引文出现章节名称、引文出现频次、引文所在语句。



引文功能属性

研究基础、相关研究、背景资料
(基于、使用、支撑、比较、相关研究、未来启发、背景介绍)



引用对象属性

指施引文献引用引文的对象，本研究将引用对象归纳为“概念理论、方法、数据、结论、其他”。

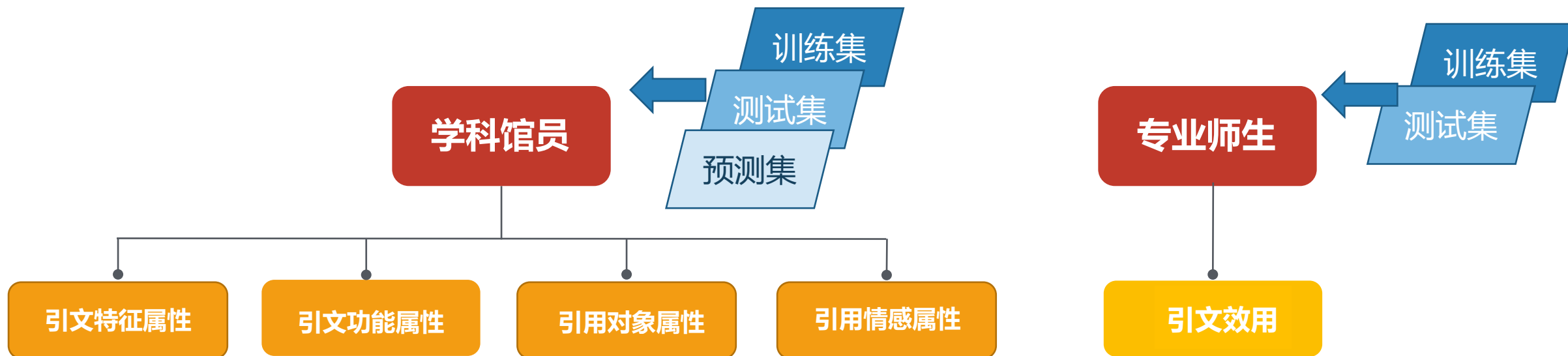


引用情感属性

引用情感在传统的三级分类法基础上进行了部分定义扩充，认为对施引文献起正面支撑作用的引文属于“积极”，与研究结果相悖、指出现有成果不足、研究局限等引文表示“消极”，其余则均标注为“中立”。

| | 引文特征 | 英文描述 | 特征定义与描述 |
|--------|----------|--------------------|--|
| 引文特征属性 | 引文年份 | Year | 引文发表的年份 |
| | 引文所在期刊 | Source | 引文所在的期刊 |
| | 引文所在章节名称 | Chapter | 引文在施引文献中出现的章节名称。如引言、背景介绍、方法、结果结论、讨论等。 |
| | 引文出现频次 | Frequency | 引文在施引文献中出现的频次。 |
| | 引文所在语句 | Text | 引文在施引文献中的语句。 |
| 引用功能属性 | 研究基础 | research basis | 施引文献研究使用了引文的研究方法、成果、数据、结果、结论等，或引用内容对施引文献研究或结论的形成起到支撑作用。 |
| | 相关研究 | Related research | 引文研究是施引文献研究方法、结论、数据、效果、案例等的内容介绍或对比，并不对本研究起到支撑作用。 |
| | 背景资料 | Background | 引文对施引文献进一步研究有所启发，或在未来研究中可能有所涉及与施引文献工作相关的时代背景、研究背景、学术背景、文化背景等 |
| 引用对象属性 | 概念定义 | Concept definition | 施引文献引用了引文中提出的概念、定义、介绍等。 |
| | 方法 | method | 施引文献研究中使用了引文中的研究方法、试验方法、模型、算法、工具、方案等。 |
| | 数据 | Date | 施引文献使用了引文中的研究数据、公式、实验结果、统计数据等。 |
| | 结论 | conclusion | 施引文献使用引文中得出的研究结论。 |
| | 其他 | other | 施引文献中未引用引文中任何内容。 |
| 引用情感属性 | 积极 | Support | 引文内容、主要观点或结论对施引文献的论证方法、观点、结论起支撑作用。 |
| | 中立 | neutral | 除表现出“积极”“消极”的其他论述。 |
| | 消极 | disapprove | 引文内容、主要观点或结论与施引文献研究结果相悖、引文指出现有成果的不足、研究局限。 |

3.引文标注



- 原始数据由不同专业背景的学科馆员依据引文标注框架，分别对所有引文内容特征属性、功能属性及引用对象等属性进行标注。
- 邀请医学专业教师及研究生若干，对训练集与测试集引文仅进行引文效用判断及标注。

高质量语料库

模型训练准确性

与自动标引结果对比



4.特征筛选和计算 特征指标筛选

引文标注体系包括：引文题目，引文期刊，引文年份，引文所在章节名称，引文出现频次，引用情感，引文功能，引用对象类型，引文重要性9种引文特征指标，其中引文重要性为目标指标，通过相关性矩阵和散点图矩阵，观察各指标和目标间的相关性。

1. “引文出现频次”和引文效用几乎没有相关性。
2. “引文功能”与引文效用相关性最高。
3. 最终选定的特征为：
引文所在章节名称，引用情感，引文功能，引用对象类型。



| Row ID | D 引文... | D 引文... | D 引用情感 | D 引文功能 | D 引用对象类型 | D ▼ 是否重要 |
|----------|--------------|---------|--------------------|-------------------|-------------------|--------------------|
| 是否重要 | 0.4082830... | ? | 0.2492376699656... | 0.722071434700... | 0.426052628760... | 1.0 |
| 引文功能 | 0.2665492... | ? | 0.2562750273182761 | 1.0 | 0.244650071567... | 0.7220714347001006 |
| 引用对象类型 | 0.3061776... | ? | 0.2686970836013... | 0.244650071567... | 1.0 | 0.4260526287607... |
| 引文所在章... | 1.0 | ? | 0.1897132239300... | 0.266549213391... | 0.306177628361... | 0.4082830573885... |
| 引用情感 | 0.1897132... | ? | 1.0 | 0.256275027318... | 0.268697083601... | 0.2492376699656... |
| 引文出现频次 | ? | 1.0 | ? | ? | ? | ? |

4.特征筛选和计算 模型训练及验证评价

逻辑回归模型训练 & 支持向量机SVM模型训练

建模方法：选用**梯度下降算法 (SAG)**，设置学习率为**0.001**，评价指标选用**混淆矩阵 (Scorer节点)**。其中,SVM需通过**独热编码(OneHotEncoder)**对非数值类型数据进行转化处理。两种方法使用同样的训练集与测试集。

评价参数：**准确率 (Accuracy)**、**Cohen`s kappa**、及**ROC曲线**

| 是否重要 \ Prediction (是否重要) | 否 | 是 |
|--------------------------|-----|-----|
| 否 | 167 | 18 |
| 是 | 14 | 157 |

Correct classified: 324 Wrong classified: 32
Accuracy: 91.011% Error: 8.989%
Cohen`s kappa (K): 0.82%

逻辑回归

效果对比：核心误判 **32**篇
Accuracy: **91.011%**
Cohen`s kappa: **0.82%**

| 是否重要 \ Prediction (是否重要) | 否 | 是 |
|--------------------------|-----|-----|
| 否 | 179 | 6 |
| 是 | 21 | 150 |

Correct classified: 329 Wrong classified: 27
Accuracy: 92.416% Error: 7.584%
Cohen`s kappa (K): 0.848%

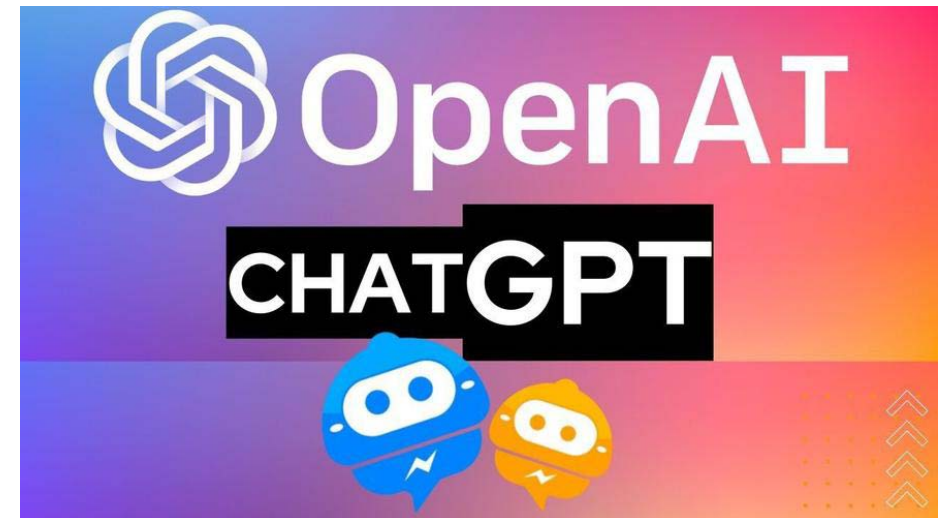
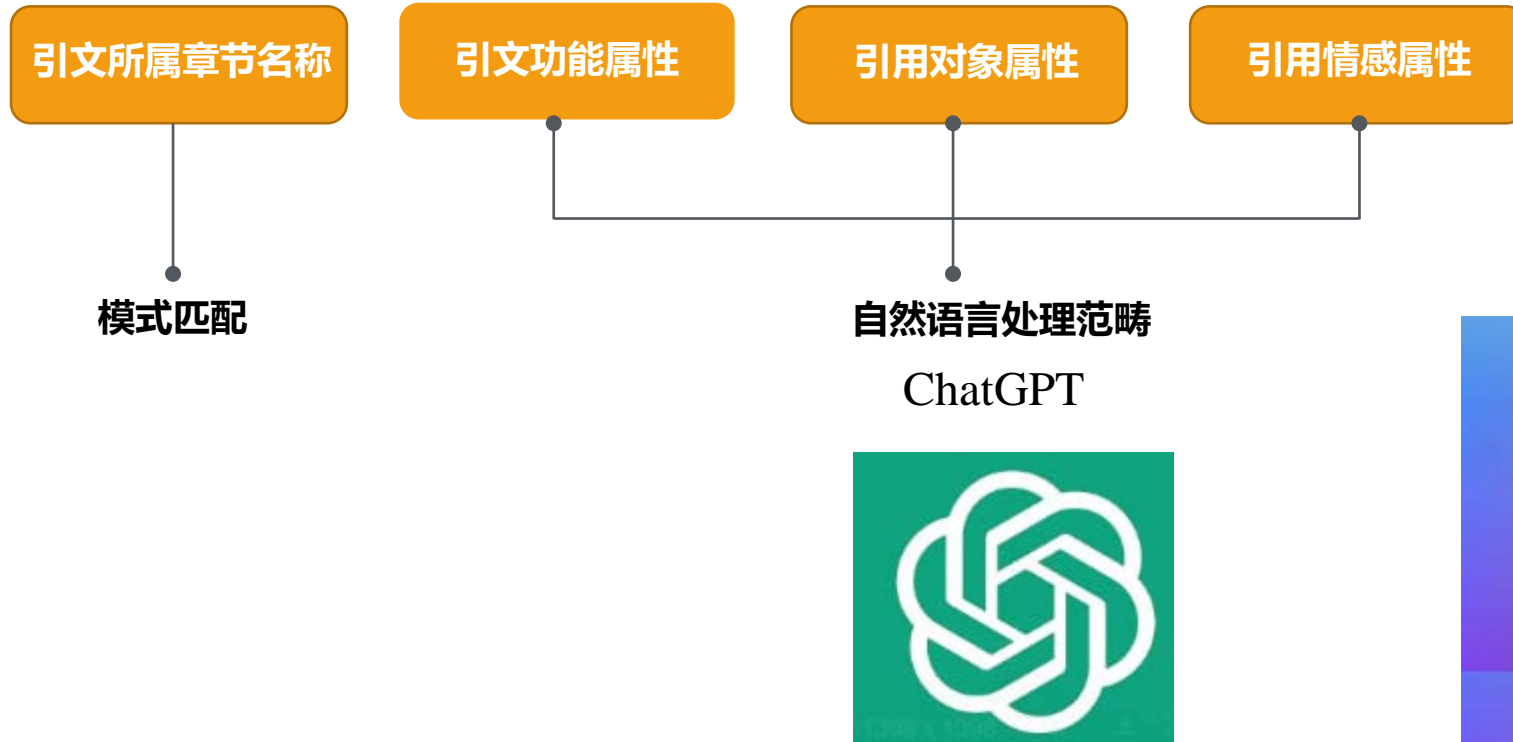
支持向量机 (SVM)

核心误判 **27**篇
Accuracy: **92.416%**
Cohen`s kappa : **0.848%**

5.自动获取引文特征及建模效果对比

ChatGPT文本分类、语义角色标注、机器阅读理解等方面的优势与引文内容特征分析、提取、标注与分类的需求较为契合。

ChatGPT能否直接根据引文内容对其效用进行分类？





比



引用功能可以分为以下几种:

1. 研究基础: 表示施引文献研究以引文为起点、受到启发或基于引文开展的深入研究, 施引文献研究使用了引文研究的方法、成果、实验数据、结果、研究结论等, 引文的结论、数据、研究内容对施引文献研究或结论的形成起到支撑作用。
2. 相关研究: 表示施引文献研究与引文研究在结论、方法、结果等方面形成对比。与施引文献研究相关的研究方法、结论、数据、案例等介绍。
3. 背景资料: 表示引文对施引文献进一步研究有所启发, 或在未来研究中可能有所涉及, 与施引文献工作相关的时代背景、研究背景、学术背景、文化背景等。



好的, 请您提供需要分类的10个引用。



11. Hu D, Hiley CR, Fan TP. Comparative studies of the angiogenic activity of vasoactive intestinal peptide, endothelin-1 and -3 and angiotensin II in a rat sponge model. *Br J Pharmacol* 1996;117:545-551.
12. Yang J, Zong CH, Zhao ZH, et al. Vasoactive intestinal peptide in rats with focal cerebral ischemia enhances angiogenesis. *Neuroscience* 2009;161:413-421.
13. Yang J, Shi QD, Song TB, et al. Vasoactive intestinal peptide increases VEGF expression to promote proliferation of brain vascular endothelial cells via the cAMP/PKA pathway after ischemic insult in vitro. *Peptides* 2013;42:105-111.
14. Antonawich FJ, Said SI. Vasoactive intestinal peptide attenuates cytochrome c translocation in cultured neurons. *J Neurosci* 2003;23:1011-1018.



Send a message.

Free Research Preview. ChatGPT may produce inaccurate information about people, places, or facts. [ChatGPT May 3 Version](#)



1. 11, 12, 13, 17, 18: 研究基础, 都是关于血管活性肽的研究, 探讨其在促进血管生成以及保护缺血性损伤方面的作用。
2. 14, 15, 16, 19, 20: 相关研究, 这些文献探讨了与血管活性肽相关的不同主题, 如其对细胞凋亡、移植体生存、神经炎症和脑损伤的影响等。



1. Lakhan SE, Kirchgessner A, Tepper D, Leonard A. Matrix metalloproteinases and blood-brain barrier disruption in acute ischemic stroke. *Front Neurol* 2013;4:32.
2. Cardoso FL, Brites D, Brito MA. Looking at the blood-brain barrier: molecular anatomy and possible investigation approaches. *Brain Res Rev* 2010;64:328-363.
3. Obermeier B, Daneman R, Ransohoff RM. Development, maintenance and disruption of the blood-brain barrier. *Nat Med* 2013;19:1584-1596.
4. Khatri R, McKinney AM, Swenson B, Janardhan V. Blood-brain barrier, reperfusion injury, and hemorrhagic transformation in acute ischemic stroke. *Neurology* 2012;79(13 Suppl 1):52.

1. 研究基础: Lakhan等人的研究使用了引文中提到的矩阵金属蛋白酶和血脑屏障破坏的方法和数据, 并且引文出现在施引文献的方法和讨论部分。
2. 相关研究: Cardoso等人的研究与施引文献有关, 但并没有直接支持施引文献的工作, 因此引文多出现在介绍部分。
3. 背景资料: Obermeier等人的研究对施引文献进一步的背景知识有所启发, 因此被用作背景资料进行引用。
4. 相关研究: Khatri等人的研究与施引文献有关, 但并没有直接支持施引文献的工作, 因此引文多出现在介绍部分。

5.自动获取引文特征及建模效果对比

模型分类效果对比

待完成自动化特征分类以及格式化输入后，检测自动特征提取的分类效果：使用SVM及相同的训练集和测试集进行模型训练，并与人工标引数据训练结果进行了对比。

对比结果

ChatGPT分类数据预测效果无法达到人工标注的高准确率，但可有效**提升特征提取效率**。

| 是否重要 \ Prediction (是否重要) | 否 | 是 |
|--------------------------|-----|-----|
| 否 | 179 | 6 |
| 是 | 21 | 150 |

Correct classified: 329
Accuracy: 92.416%
Cohen's kappa (K): 0.848%

核心误判 **27篇**
Accuracy: **92.416%**
Cohen's kappa : **0.848%**

人工标记数据SVM模型评价结果

| 是否重要 \ Prediction (是否重要) | 否 | 是 |
|--------------------------|-----|-----|
| 否 | 179 | 6 |
| 是 | 27 | 144 |

Correct classified: 323
Accuracy: 90.73%
Cohen's kappa (K): 0.813%

核心误判 **33篇**
Accuracy: **90.73%**
Cohen's kappa : **0.813%**

自动获取数据SVM模型评价结果



研究应用拓展

学科资源保障分析

利用“效用视角下核心引文识别模型”对全部数据进行分析，识别出核心引文共计**803**篇，分布于**315**个期刊。

资源保障

JCR核心期刊保障率达到**79.88%**

引文期刊保障率达到**83.77%**

核心引文期刊保障率达到**95.24%**

分析结果

我校重点保障了该学科的高质量资源。
我校对师生发文引文期刊保障率也达到较高水平。
对发文核心引文资源的保障率达到95%以上。

应用启发

针对我校某学科、科研团队、研究专题，从科研发文引文效用的角度，分析我校对**核心支撑资源**的保障力度。

基于ERS的学科期刊保障分析

| 指标 | 数量 | 占比 |
|---------------|-------|--------|
| 学科期刊总量 | 22933 | 100% |
| 已保障资源 | 9823 | 42.83% |
| 学科 JCR 核心期刊保障 | 3518 | 79.88% |

*依据教育部学科与期刊学科映射关系，“医学”学科共映射期刊22933种

学科发文保障分析

| 指标 | 总数量 | 已保障数量 | 占比 |
|----------------|------|----------------|--------|
| 发文期刊 | 1047 | 904 | 86.34% |
| 发文的引用期刊 | 6625 | 5550 | 83.77% |
| 本研究中发文“核心引文”期刊 | 315 | 285(已购)+15(OA) | 95.24% |





研究应用拓展

数据库资源保障分析

本研究中识别的核心引文所在的315个期刊分布于30个数据库中，分别对其进行三种指标的计算。

指标设定

学科期刊保障率A = $\frac{\text{该数据库包含该学科期刊数量}}{\text{该学科期刊总量}} * 100\%$

A反映出该库对于学科资源的整体保障情况，不区分核心资源。

核心引文期刊的保障率B = $\frac{\text{该数据库包含核心引文期刊数量}}{\text{核心引文分布的所有期刊量 (315)}} * 100\%$

B是基于本研究识别的核心引文期刊保障情况，值越高，说明该库对学科成果产出的支撑力度越大。

核心引文期刊占学科期刊比率C = $\frac{\text{该数据库包含核心引文期刊数量}}{\text{该数据库中该学科期刊数量}} * 100\%$

C反映出数据库所包含的学科期刊中核心期刊占比。

应用启发

对数据库效用评估方面，不能仅依据单一指标，应依据研究或应用目的，选择相应指标综合判断。

| | 保障核心引文期刊数量 | 学科期刊数量 | 学科期刊保障率 A | 核心引文期刊保障率 B | 核心引文期刊/学科期刊量 C |
|---|------------|--------|-----------|-------------|----------------|
| 1 | 98 | 2850 | 29.01% | 31.11% | 3.44% |
| 2 | 86 | 1047 | 10.66% | 27.30% | 8.21% |



启示与思考

不同粒度的引文标注框架可依据研究目的各抒所长

虽然引文分类指标体系缺乏统一标准的问题仍客观存在，但可根据具体研究与应用目的，发挥不同体系所长。

“ChatGPT+SVM”为实现引文特征识别与自动分类提供新思路

基于ChatGPT+SVM模型预测结果虽无法达到人工标注数据构建模型的高准确率，却可在保证较好预测效果的同时，大大提升特征识别与分类效率。

利用核心效用资源，客观评价学科资源保障及数据库支撑表现

引文效用可更加科学地反应出文献资源对于研究本身的价值，依此做出核心资源的判断、开展的资源保障评估等应用将更加科学。

未来可基于Fine tuning构建引文自动分类模型

本研究采用的是小样本提示（prompts），需要反复给出Promet进行预训练。随着ChatGPT API正式开放，可利用ChatGPT提供的fine tuning功能，实现功能需求定制化模型。





致谢

感谢各位评审专家对此次课题研究的鼓励和认可！
感谢CALIS及深圳大学为大家创造了学习交流的机会！
感谢各位老师不吝赐教，为我们的研究提出更多的宝贵意见！



西安交通大学图书馆

李凌