相关论文见 https://gipplab.org/pub**论文预印本：**Meuschke, N. & Jagdale, A. & Spinde, T. & Mitrovi´c, J. & Gipp, B., "PDF信息提取工具基准测试：面向学术文档的多任务多领域评估框架"，载于《面向更美好世界的信息：常态、虚拟性、物理性、包容性》，LNCS，第13972卷，Cham：施普林格·自然瑞士，2023年，第383–405页，DOI：10.1007/978-3-031-28032-0 31。**点击下载：**BibTeX

arXiv:2303.09957v1 [cs.IR] 17 Mar 2023

面向学术文档的多任务多领域评估框架：PDF信息提取工具基准测试

NormanMeuschke[1], [ORCID], Apurva Jagdale[2],Timo Spinde[1], [ORCID], Jelena Mitrovi´c[2, 3], [ORCID]和Bela Gipp[1], [ORCID]

[1] 德国哥廷根大学，37073 哥廷根¨{meuschke、spinde、gipp}@uni-goettingen.de[2] 帕绍大学，94032 帕绍，德国{apurva.jagdale、jelena.mitrovic}@uni-passau.de[3] 塞尔维亚人工智能研发研究所，21000 诺维萨德，塞尔维亚

摘要。从学术 PDF 文档中提取信息，对于众多索引、检索和分析应用至关重要。由于可用工具种类繁多、技术差异显著，而近期的性能基准测试又十分罕见，因此很难选出提取特定内容元素的最佳工具。此外，这类基准测试通常只涵盖少数内容元素，例如页眉元数据或书目参考文献，并且使用来自特定学科的小型数据集。我们提供了一个规模大且多样化的评估框架，支持的提取任务比大多数相关数据集更多。该框架基于 DocBank 构建，这是一个多领域数据集，包含从 arXiv 上 50 万页研究论文中提取的 150 万个已标注内容元素。借助这一新框架，我们对十种免费工具进行了基准测试，评估它们从学术 PDF 文档中提取文档元数据、书目参考文献、表格及其他内容元素的能力。GROBID 的元数据和参考文献提取效果最佳，其次是 CERMINE 和 Science Parse。在表格抽取方面，Adobe Extract 的表现优于其他工具，但其效果远低于其他内容元素的提取效果。所有工具在提取列表、页脚和公式方面都表现不佳。我们认为，要使大多数内容元素的提取质量达到令人满意的水平，还需要开展更多研究，以改进并组合使用这些工具。我们提出的这类评估数据集和框架有助于推动这一研究方向。为促进这一目标，我们公开了数据和代码。

关键词：PDF · 信息抽取 · 基准测试 · 评估。

1 引言

便携式文档格式（PDF）是学术文档最常见的编码格式。从学术 PDF 文档中提取信息，对于许多索引、检索和分析任务至关重要。文档搜索、推荐、摘要生成、分类、知识库构建、问答和文献计量分析只是其中几个例子 [31]。

然而，这种格式的技术设计使信息抽取颇具挑战。Adobe 通过扩展 PostScript 页面描述语言，设计了 PDF，使其成为一种跨平台的固定版式格式。[24] PDF 着重编码文档的视觉布局，以确保文档在不同软件和硬件平台上外观一致，但几乎不包含有关文档元素的结构和语义信息。

自 1993 年 PDF 格式问世以来，已有众多用于从 PDF 文档中进行信息抽取（IE）的工具问世。这些工具的开发经历了快速的技术演进，抽取方法从基于规则的算法发展到统计机器学习（ML），再到深度学习（DL）模型（参见第 2 章）。目前，要找到最适合从 PDF 文档中提取特定内容元素的工具仍然很困难，原因如下：

1. 通常，工具仅支持提取学术文档中部分内容元素，例如标题、作者、段落、文内引文、图注、表格、图、公式或参考文献。2. 许多信息抽取工具（例如，我们研究中考察的 35 种工具中有 12 种）已不再维护或已经过时。3. 以往对信息抽取工具的评估通常只考察特定内容元素，或使用领域特定语料库，因此评估结果难以比较。此外，最新的综合信息抽取工具基准测试分别于 2015 年针对元数据、[4] [55],2017 年针对正文 [6],以及 2018 年针对参考文献[5] [54]，发布。这些评估未能反映该领域最新的技术进展。

为弥补这一知识缺口，并帮助人们找到从学术 PDF 文档中提取特定元素的最佳工具，我们基于一个包含来自 arXiv 文档的 50 万页、涵盖多个领域的数据集，对十种先进的非商业工具进行了全面评估。这些工具支持处理十一种内容元素。

我们的代码、数据和资源均已公开发布于
http://pdf-benchmark.gipplab.org

2 相关工作

本章节介绍 PDF 信息抽取方法（第2.1节）、适用于训练和评估 PDF 信息抽取方法的标注数据集，以及对信息抽取工具的先前评估（第2.2节）。

2.1 从 PDF 文档中提取信息

表格1总结了自 1999 年以来有关 PDF 信息抽取的出版物。对于每篇出版物，表格均展示其主要技术方法以及

[4] 例如作者、标题、所属机构、地址、电子邮箱等[5] 指提取书目参考文献的各个组成部分，例如作者、标题、出版场所、编辑、卷、期、页码范围、出版年份等。

表1：从 PDF 文档中进行信息抽取的出版物。

出版物[1]	年份	任务[2]	方法	训练数据集[3]
Palermo [44]	1999	M，目录	规则	100份文档
Klink [27]	2000	M	规则	979页
Giuffrida [18]	2000	M	规则	1,000份文档
Aiello [2]	2002	罗马尼亚语，标题	规则	1,000页
Mao [37]	2004	M	OCR，规则	309份文档
Peng [45]	2004	M, R	CRF	CORA（500条参考文献）
Day [14]	2007	M, R	模板	160,000条引文
Hetzner [23]	2008	R	HMM	CORA（500条参考文献）
Councill [12]	2008	R	CRF	CORA（200 篇参考文献），CiteSeer（200 篇参考文献）
Lopez [36]	2009	B、M、R	条件随机场、深度学习	None
Cui [13]	2010	M	HMM	400份文档
Ojokoh [42]	2010	M	HMM	CORA（500条参考文献），ManCreat
Kern [25]	2012	M	HMM	FLUX-CiM（300条参考文献），电子预印本、Mendeley、PubMed（1.9万条记录）
Bast [5]	2013	B、M、R	规则	DBLP（690份文档）、PubMed（500份文档）
Souza [53]	2014	M	CRF	100份文档
Anzaroot [3]	2014	R	CRF	马萨诸塞大学（1,800条参考文献）
Vilnis [57]	2015	R	CRF	马萨诸塞大学（1,800条参考文献）
Tkaczyk [55]	2015	B、M、R	CRF，规则，SVM	CiteSeer（4,000条参考文献）、CORA（500条参考文献）、GROTOAP、PMC（53K篇文档）
Bhardwaj [7]	2017	R	FCN	5,090条参考文献
Rodrigues [49]	2018	R	双向长短期记忆网络	40,000 条参考文献
Prasad [46]	2018	M, R	条件随机场、深度学习	FLUX-CiM（300 条参考文献）、CiteSeer（4,000 条参考文献）
Jahongir [4]	2018	M	规则	10,000 份文档
Torre [15]	2018	B, M	规则	300 份文档
Rizvi [47]	2020	R	区域卷积神经网络	40,000 条参考文献
Hashmi [22]	2020	M	规则	45 份文档
Ahmed [1]	2020	M	规则	150 份文档
Nikolaos [33]	2021	B、M、R	注意力，双向长短期记忆网络	3,000 份文档

[1]按时间顺序排列的出版物；标签仅表示第一作者。[2] (B) 正文，(M) 元数据，(R) 参考文献，(RO) 阅读顺序，(ToC) 目录[3] 领域特定数据集：计算机科学：CiteSeer [43], CORA [39], DBLP [52], FLUX- CiM [10,11], ManCreat [42]；健康科学：PubMed [40], PMC [41]

训练数据集。27 种方法中有 18 种（67%）采用机器学习或深度学习（DL）技术，其余则采用基于规则的抽取（规则）。早期工具依赖人工编写的规则 [44]。第二代工具采用统计机器学习，例如基于隐马尔可夫模型（HMM）的方法 [8],条件随机场（CRF） [29], 和最大熵 [26]。最新的信息抽取工具采用 Transformer 模型 [56]。

从理论上说，相较于针对具体案例的基于规则的算法，人们更倾向于采用灵活性和适应性更强的机器学习与深度学习技术，这一趋势可从表 1中看出。然而，许多训练数据集具有特定领域属性，例如，其中只包含来自计算机科学或健康科学领域的文档，且文档数量少于 500 篇。这两个因素使得相应信息抽取方法的泛化能力受到质疑。值得注意的例外包括 Ojokoh 等人 [42],Kern 等人 [25], 以及 Tkaczyk 等人 [55] ，他们使用涵盖不同领域的多个数据集进行训练和评估。然而，这些方法针对的是特定任务，即页眉元数据提取、参考文献抽取，或两者兼有。

此外，Mao 等人的文献综述指出，大多数 PDF 文本抽取方法都没有说明所使用的真实数据和性能指标，这阻碍了性能比较 [38]。一个积极的例外是 Bast 等人发表的一项研究 [5], ，该研究提出了一个全面的 PDF 文本抽取评估框架，其中对所采用的性能度量进行了细致的说明。

2.2 已标注数据集与先前的基准测试

表格2 总结了可用于训练和评估 PDF 信息抽取方法的数据集，并按其提供的真实标签类型进行分类。大多数数据集仅提供文档元数据、参考文献或两者的标签。

表 2：标注 用于信息抽取的数据集 PDF 文档。

出版物[1]	**规模**	**真实标签**
Fan [16]	147份文档	元数据
F¨arber [17]	9万份文档	参考文献
Grennan [21]	10亿条参考文献	参考文献
Saier [51,50]	100万份文档。	参考文献
Ley [30,52]	600万份文档	元数据、参考文献
Mccallum [39]	935 文档	元数据、参考文献
Kyle [34]	810万份文档	元数据、参考文献
Ororbia [43]	600万份文档。	元数据、参考文献
Bast [6]	12,098 份文档	正文、章节、标题
Li [31]	50 万页	图注、公式、图、页脚
		列表、元数据、段落,
		参考文献，章节 ，表格

[1]标签仅指第一作者。

只有 Li 等人的 DocBank 数据集 [31] 提供了学术文档中 12 种不同内容元素的注释，包括图、公式、表格和图注。迄今为止，这些内容元素大多尚未用于基准评估。DocBank 的规模相对较大（涵盖四年间发表于 arXiv 的研究论文中的 50 万页）。DocBank 数据集的一个缺点是其参考文献标签粒度较粗，未对书目条目的作者、出版商、卷或日期等字段进行标注；而专门针对参考文献的数据集，如 unarXive [21]或 S2ORC [34]则进行了此类标注。

表格 3 展示了自 1999 年以来开展的 PDF 信息提取基准测试。此类研究寥寥无几，而且很少重复开展或更新；考虑到许多工具会频繁更新，这种情况并不理想。其他工具则在技术上逐渐过时，或无人维护。例如，pdf-extract[6]，lapdftext[7], PDF-SSA4MET[8], 和 PDFMeat[9] 目前已不再积极维护，而 ParsCit[10]已由 NeuralParsCit 取代[11]，SciWING[12]也已如此。

表 3：PDF 信息提取方法的基准测试评估。

出版物[1]	数据集	指标[2]	工具	标签[3]
Granitzer [19]	电子预印本（2,452 篇文档），Mendeley（20,672 篇文档）	P, R	2	M
Lipinski [32]	arXiv（1,253 篇文档）	Acc	7	M
Bast [6]	arXiv（12,098 篇文档）	自定义	14	荷兰语、旁遮普语 RO, W
K¨orner [28]	100（德语文档）	P, R, F_1	4	Ref
Tkaczyk [54]	9,491 份文档	P, R, F_1	10	Ref
Rizvi [48]	8,766 条参考文献	F_1	4	Ref

[1]标签仅表示第一作者。[2] (P) 精确率，(R) 召回率，(F_1) F_1-分数，(Acc) 准确率[3] (M) 元数据，(NL) 新行，(Pa)段落，(Ref) 参考文献，(RO) 阅读顺序，(W) 词

如表3 所示，迄今用于评估 PDF 信息抽取工具的最大数据集约包含 24,000 份文档。与此任务可用的数据集规模相比，这一数量较小，相关数据集见表2。大多数研究仅专注于评估元数据和参考文献抽取（另见表3）。一个例外是 Bast 和 Korzen 提出的基准测试

[6] https://github.com/CrossRef/pdfextract[7]
https://github.com/BMKEG/lapdftext[8]
https://github.com/eliask/pdfssa4met[9]
https://github.com/dimatura/pdfmeat[10]
https://github.com/knmnyn/ParsCit[11]
https://github.com/WING-NUS/Neural-ParsCit[12]
https://github.com/abhinavkashyap/sciwing

[6], 该基准测试评估了 14 种工具产生的多余词、缺失词、多余段落、缺失段落、多余新行和缺失新行，但使用的数据集规模相对较小，约包含 1 万份文档。

通过回顾相关工作，我们得出以下结论：（1）近期针对 PDF 信息抽取工具的基准测试很少；（2）大多只分析元数据抽取；（3）使用的领域特定数据集规模较小；（4）纳入的工具已过时或不再维护。（5）尚未使用多种适当标注的数据集来评估 PDF 文档的信息抽取工具。因此，我们认为有必要在一个规模较大的学术文档标注数据集上，对最先进的 PDF 信息提取工具进行基准测试。该数据集涵盖多个领域，并包含多种内容元素。

3 方法

本章节介绍我们的实验设置，包括对所评估工具的说明（章节3.1）、所用数据集（章节 3.2）以及所遵循的流程（章节3.3）。

3.1 评估的工具

我们选择了十种积极维护的非商业开源工具，并根据其提取任务对其进行分类。

1. **元数据提取** 包含用于提取标题、作者、摘要及其他类似文档元数据的工具。
2. **参考文献抽取** 包含用于访问书目参考文献字符串并将其解析为作者姓名、出版物标题和出版场所等字段的工具。3. **表格抽取** 指可用于访问表格结构和数据的工具。4. **通用抽取**涵盖用于提取段落、章节、图、图注、公式、列表或页脚等内容的工具。

对于我们评估的每种工具，表 4 展示了其版本、支持的抽取任务、主要技术方法和输出格式。下文将简要介绍各工具，重点说明其技术方法。

Adobe Extract[13] 是一种基于云的 API，可用于提取表格以及归属于通用抽取 类别的众多其他内容元素。该 API 使用 Adobe Sensei[14] 人工智能和机器学习平台来理解 PDF 文档的结构。为评估 Adobe Extract API，我们使用 Adobe PDFServices Python SDK[15] 来访问该 API 提供的服务。

Apache Tika[16] 支持以 XML 格式提取元数据和内容。我们使用 tika-python[17] 客户端访问 Tika REST API。遗憾的是，我们发现 tika-python 仅支持提取内容（段落）。

[13] https://www.adobe.io/apis/documentcloud/dcsdk/pdf-extract.html[14] https://www.adobe.com/de/sensei.html[15] https://github.com/adobe/pdfservices-python-sdk-samples[16] https://tika.apache.org/[17] https://github.com/chrismattmann/tika-python

表4：已评估的信息抽取工具概览。

Tool	版本	任务[1]	技术	输出
Adobe Extract	1.0	G, T	Adobe Sensei AI 框架	JSON, XLSX
Apache Tika	2.0.0	G	Apache PDFBox	TXT
Camelot	0.10.1	T	OpenCV，PDFMiner	CSV，数据框
CERMINE	1.13	G, M, R	CRF，iText ，规则 ，SVM	JATS
GROBID	0.7.0	G, M,R, T	条件随机场、深度学习、Pdfalto	TEI XML
PdfAct	n/a	G, M, R, T	pdftotext、规则	JSON、TXT、XML
PyMuPDF	1.19.1	G	光学字符识别、Tesseract	TXT
RefExtract	0.2.5	R	pdftotext，规则	TXT
ScienceParse	1.0	G、M、R、	条件随机场、pdffigures2、规则	JSON
Tabula	1.2.1	T	PDFBox ，规则	CSV，数据框

[1](G) 通用信息，(M) 元数据，(R) 参考文献，(T) 表格

Camelot[18] 可以使用 *Stream* 或 *Lattice* 模式提取表格。前者利用单元格之间的空白，后者则利用表格边框来识别表格单元格。在我们的实验中，我们仅使用 Stream 模式，因为我们的测试文档是学术论文，其中的表格通常利用空白而非单元格边框来划分单元格。Stream 模式内部使用 PDFMiner 库[19] 提取字符，随后根据空白间距将字符组合成词和句子。

CERMINE [55] 提供元数据、参考文献和通用信息抽取功能。该工具使用 iText PDF 工具包[20] 提取字符，并使用 Docstrum[21] 图像分割算法对文档图像进行页面分割。CERMINE 使用通过 LibSVM[22]库实现的支持向量机分类器，以及用于元数据提取的基于规则的算法。对于参考文献提取，该工具采用 k-均值聚类，以及通过 MALLET[23] 工具包实现的条件随机场进行序列标注。CERMINE 会返回一个 XML 文件，其中包含整个 PDF 的注释。我们使用 Beautiful Soup[24] 库筛选 CERMINE 的输出文件，以提取与评估相关的注释。

GROBID[25] [35] 支持全部四项提取任务。该工具允许使用特征工程条件随机场（默认设置），或使用基于 TensorFlow 和 Keras 的 DeLFT[26] 深度学习库实现的条件随机场与深度学习模型组合。GROBID 针对不同组件采用序列标注模型级联。模型级联中的模型使用各自独立的标签序列

[18] https://github.com/camelot-dev/camelot[19] https://github.com/pdfminer/pdfminer.six[20] https://github.com/itext[21] https://github.com/chulwoopack/docstrum[22] https://github.com/cjlin1/libsvm[23] http://mallet.cs.umass.edu/sequences.php[24] https://www.crummy.com/software/BeautifulSoup/bs4/doc/[25] https://github.com/kermitt2/grobid[26] https://github.com/kermitt2/delft

算法和特征；部分模型使用分词器。这种方法通过允许调整模型来提供灵活性，并提高模型的可维护性。我们使用生产环境设置评估默认条件随机场模型（根据该工具的文档，这是提升 GROBID 服务器性能和可用性的推荐设置[27]）。

PdfAct 原名 Icecite [5]，是一款基于规则的工具，支持全部四种提取任务，包括附录、致谢和目录的提取。该工具使用 PDFBox[28] 和 pdftotext[29] PDF 操作与内容提取库。我们使用该工具的 JAR 发行版[30]。

PyMuPDF[31] 扩展了 MuPDF[32] 查看器库，增加了字体和图像提取、PDF 合并以及文件嵌入功能。PyMuPDF 使用 Tesseract[33] 进行光学字符识别。PyMuPDF 无法处理文件名中包含特殊字符的文件。

RefExtract[34] 是一款使用 pdftotext[35] 和正则表达式的参考文献抽取工具。RefExtract 会返回文档整个参考文献目录的标注。然而，我们数据集中的真实标注（参见章节3.2）针对的是文档的单独页面，并不总是涵盖整篇文档。如果某篇文档只有部分参考文献具有真实标注，我们会使用正则表达式筛选 RefExtract 的输出，仅保留具有真实标签的参考文献。

Science Parse[36] 使用基于 GROBID 数据训练的条件随机场（CRF）模型来提取标题、作者和参考文献。它还采用 Clark 和 Divvala [9] 提出的基于规则的算法，以 JSON 格式提取章节和段落。

Tabula[37] 是一款表格抽取工具。与 Camelot 类似，Tabula 提供使用 PDFBox 实现的*Stream* 模式，以及使用 OpenCV 实现、用于识别表格单元格的*Lattice* 模式。

3.2 数据集

我们使用 DocBank[38] 数据集，该数据集由 Li 等人创建， [31], 用于我们的实验。图1 展示了数据集的编制过程。首先，创建者收集了 arXiv 文档，这些文档同时提供 PDF 和 LaTeX 源代码。随后，Li 等人编辑 LaTeX 代码，以便对文档 PDF 版本中的内容元素进行准确的自动标注。为此，他们插入了命令，将内容元素格式化为特定

[27] https://GROBID.readthedocs.io/en/latest/Troubleshooting/[28] http://pdfbox.apache.org/[29] https://github.com/jalan/pdftotext[30] https://github.com/ad-freiburg/pdfact[31] https://github.com/pymupdf/PyMuPDF[32] https://mupdf.com/[33] https://github.com/tesseract-ocr/tesseract[34] https://github.com/inspirehep/refextract[35] https://linux.die.net/man1/pdftotext[36] https://github.com/allenai/science-parse[37] https://github.com/chezou/tabula-py[38] https://github.com/doc-analysis/DocBank

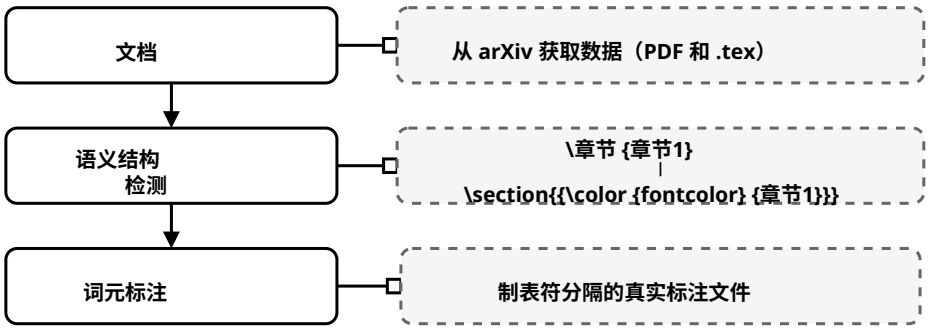


图 1：生成 DocBank 数据集的流程。

颜色。图 3 的中心部分展示了内容元素与颜色的对应关系。在最后一步，数据集的创建者使用 PDFPlumber[39] 和 PDFMiner，根据颜色提取并标注相关内容元素。DocBank 将标注作为单独的文件提供，数据集中的每个文档页面对应一个文件。

表格 5 展示了制表符分隔的真实标注文件的结构。文件中的每一行对应页面上的一个组件，结构如下。索引 0 表示词元本身，例如一个单词。索引 1-4 表示词元的边界框信息，其中 (x0, y0) 表示 PDF 坐标空间中词元的左上角，(x1, y1) 表示右下角。索引 5-7 反映词元以 RGB 表示法表示的颜色，索引 8 表示词元的字体，索引 9 表示内容元素类型的标签。每个真实标注文件都遵循图 2所示的命名规则。

表格 5：DocBank 纯文本真实标注文件的结构。

索引	0	1	2	3	4	5	6	7	8	9
内容标记		x0	y0	x1	y1	R	G	B	字体名称	标签

来源：https://doc-analysis.github.io/docbank-page/index.html。

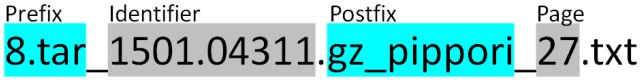


图 2：DocBank 真实标注文件的命名方案。

[39] https://github.com/jsvine/pdfplumber

DocBank 数据集为 50 万页上的 150 万个内容元素提供真实标注。Li 等人从 2014年至 2018年间发表的 arXiv 论文中提取了这些页面，论文涵盖物理学、数学、计算机科学及众多其他领域。DocBank 数据集规模庞大、时效性强、所含文档种类多、标注的内容元素数量众多，并且得益于弱监督标注方法，标注质量很高，因此是我们用途的理想选择。

3.3 评估流程

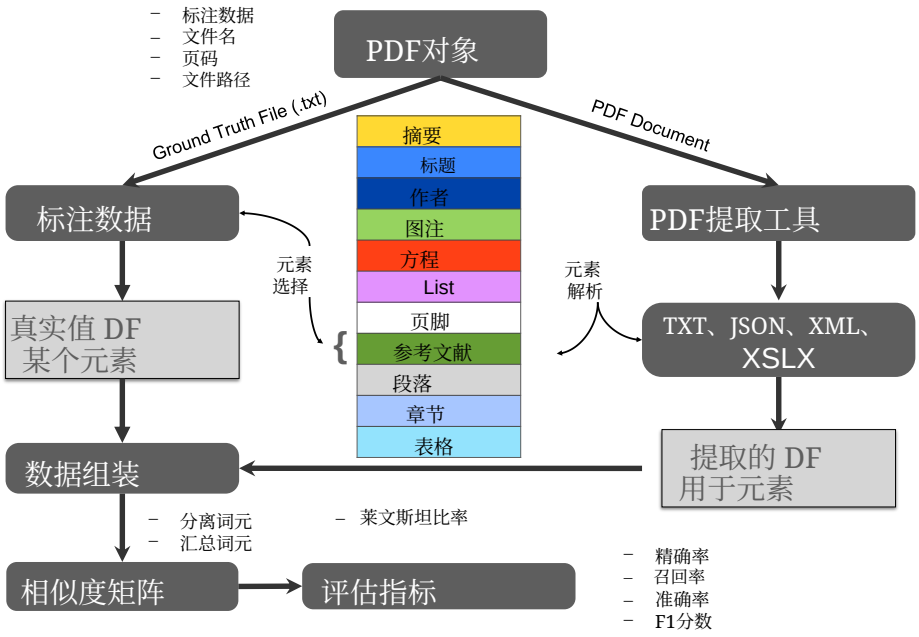


图 3：比较信息抽取工具提取的内容元素与真实标注并计算评估指标的流程概览。

图 3 展示了我们的评估流程。首先，我们选择关联的真实标注文件中包含相关标签的 PDF 文件。例如，为了评估参考文献抽取工具，我们会查找包含 参考文献 标记的真实标注文件。我们将 PDF 文件、真实标注文件、可从文件名中获取的文档 ID 和页码（参见图 2），以及文件路径纳入一个自定义的 Python 对象中（见图 *PDF* 对象 图 3）。

随后，评估过程分为两个分支，目标是创建两个 pandas 数据框——一个用于存放相关的真实数据，另一个用于存放信息抽取工具的输出结果。为此，系统会对真实标注文件和信息抽取工具的输出文件进行解析，并筛选出

相关的内容元素。例如，为了评估通过 CERMINE 进行的参考文献抽取，我们仅从 CERMINE 的 XML 输出文件中解析参考文献标签，并将其导入数据框（见 抽取结果数据框 图 3）。

最后，我们将真实数据数据框和提取数据数据框转换为两种格式，以便进行比较并计算性能指标。第一种是分离词元格式，其中每个词元在数据框中各占一行。第二种是汇总词元格式，其中所有词元合并为数据框中以空格分隔的一行。分离词元用于计算严格分数，以衡量词元级抽取质量；而汇总词元得出的分数则更宽松，旨在反映某类内容元素的平均抽取质量。我们将在下文解释这两种分数的概念及其计算方法。

我们采用莱文斯坦比率来量化分离词元和汇总词元两种格式下，提取出的词元与真实数据之间的相似度。方程（1）定义了提取词元 t_e 与真实词元 t_g之间的Levenshtein 距离。

$$lev_{t_e,t_g}(i,j) = \begin{cases} max(i,j), & \text{若 } min(i,j) = 0, \\ min \begin{cases} lev_{t_e,t_g}(i-1,j)+1 \\ lev_{t_e,t_g}(i,j-1)+1 \\ lev_{t_e,t_g}(i-1,j-1)+1_{(t_{ei}\neq t_{ej})} \end{cases} & \text{否则。} \end{cases} \tag{1}$$

方程（2）定义了派生的莱文斯坦比率分数（γ）。

$$\gamma\left(t_e,t_g\right) = 1 - \frac{lev_{t_e,t_g}(i,j)}{|t_e|+|t_g|} \tag{2}$$

方程（3）展示了文档（Δ^d）的相似度矩阵 （d），其中包含提取的数据框中每个分离词元与大小为γ的分离词元 E^s 以及大小为 m 的真实数据框中分离词元 G^s 的莱文斯坦比率 n。

$$\Delta^d_{m\times n} = \gamma\left[E^s_i, G^s_j\right]^{m,n}_{i,j} \tag{3}$$

使用 $m \times n$ 相似度矩阵，我们根据方程（*4*）和方程（*5*），分别计算精确率和召回率 R^d 分数。分子采用莱文斯坦比率大于或等于 0.7 的提取词元数量。为与 Granitzer 等人的实验保持一致，我们选择了这一阈值。 [19]。然后，我们根据方程（ F_1^d 6）计算分数，以词元级分数衡量工具的提取质量。

$$P^d = \frac{\#\Delta^d_{i,j} \geq 0.7}{m} \tag{4}$$

$$R^d = \frac{\#\Delta^d_{i,j} \geq 0.7}{n} \tag{5}$$

$$F_1{}^d = \frac{2 \times P^d \times R^d}{P^d + R^d} \tag{6}$$

此外，我们还计算准确率分数 A^d，以反映工具对某类词元的平均提取质量。为获得 A^d，我们根据方程（7），计算合并令牌格式中提取词元与真实标记的莱文斯坦比率。

$$A^d = \gamma\left[E^c, G^c\right] \tag{7}$$

图4 和图5 分别展示了使用独立词元和汇总词元时，作者姓名“Yuta”“Hamada”“Gary”和“Shiu”的相似度矩阵。图4还展示了字符串 $Gary$ 和 $Yuta$ 的莱文斯坦比率计算示例。这两个字符串的 Levenshtein 距离为 6，字符串总长度为 8，因此莱文斯坦比率为 0.25，并将其填入相似度矩阵。图5 则以类似方式展示了使用汇总词元计算两个字符串准确率分数的示例。

	Yuta	Hamada	Gary	Shiu1,
Yuta	1.0	0.2	0.25	0.2
Hamada	0.2	1.0	0.2	0.0
Gary	0.25	0.2	1.0	0.0
Shiu	0.25	0.0	0.0	0.8

		Y	u	t	a
	0	1	2	3	4
G	1	2	3	4	5
a	2	3	4	5	4
r	3	4	5	6	5
y	4	3	4	5	⑥

图 4：左：使用分离词元计算的作者姓名相似度矩阵。右：使用动态规划计算两个作者姓名之间的 Levenshtein 距离 (6) 及最优编辑脚本（黄色高亮部分）。

	Yuta Hamada Gary Shiu1,
Yuta Hamada Gary Shiu	0.957

图 5：使用汇总词元得到的两组作者姓名的相似度矩阵。

4 结果

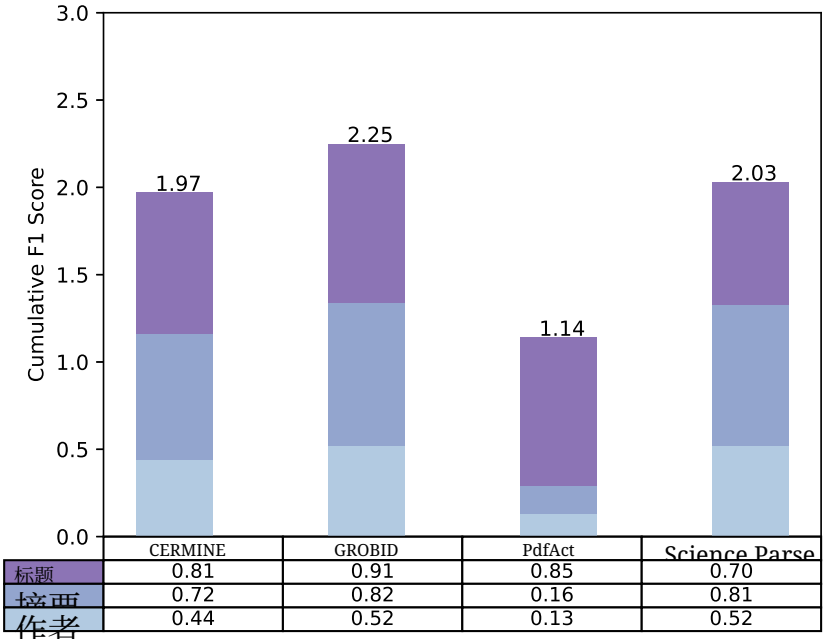
我们按提取任务（见图 6–9）和工具（见表格6）分组展示评估结果。这种双重划分方式有助于识别针对特定提取任务或内容元素表现最佳的工具，也让人们更容易衡量各工具的优缺点。请注意，针对特定任务的结果可视化（图 6–9）仅包含支持相应提取任务的工具。有关受评估工具及其支持的提取任务的概述，请参见表格 4 。

图6 展示了 F_1 CERMINE、GROBID、PdfAct 和 Science Parse 在元数据提取任务中的累计得分，即提取标题、摘要和作者。因此，累计得分的理论最高值为三分。 F_1 总体而言，GROBID 的表现最佳，累计 F_1 得分为 2.25，各项得分分别为 F_1 标题提取 得分为 0.91，摘要提取得分为 0.82，作者提取得分为 0.52。Science Parse（2.03）和 CERMINE（1.97）的累计 F_1得分相近，而 PdfAct 的累计 F_1 得分最低，为 1.14。不过，PdfAct 的标题提取表现排名第二， F_1 得分为 0.85。所有工具在作者提取方面的表现都不如标题和摘要提取。看来，CERMINE、GROBID 和 Science Parse 采用的基于机器学习的方法，在元数据提取方面优于 PdfAct 所采用的基于规则的算法[40]。

图7 展示了参考文献提取任务的结果。GROBID 的 F_1 得分为 0.79，在该任务中的表现同样最佳。CERMINE 排名第二， F_1 得分为 0.74；Science Parse 和 RefExtract 并列第三， F_1 得分均为 0.49。与元数据提取任务一样，PdfAct 在参考文献提取任务中的 F_1 得分也最低，仅为 0.15。尽管 RefExtract 和 PdfAct 都使用 pdftotext 和正则表达式，GROBID 却能为不同组件高效地划分级联序列标注模型[41] ，这可能是其性能更优的原因。 [36]。

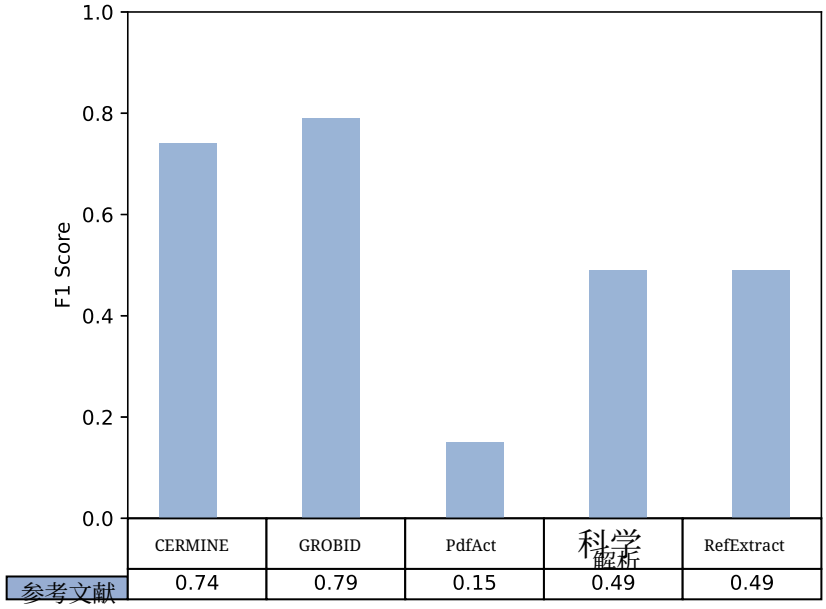
图8 展示了表格提取任务的结果。Adobe Extract 的表现优于其他工具，F_1 得分为 0.47。Camelot（$F_1 = 0.30$）、Tabula（$F_1 = 0.28$）和 GROBID（$F_1 = 0.23$）的表现明显不如 Adobe Extract。Camelot 和 Tabula 都会错误地将双栏文章视为表格，并把表格标题当作表格区域的一部分，从而拉低性能得分。Camelot 和 Tabula 使用了相似的$Stream$ 和$Lattice$ 模式（参见第3.1节），这可能导致两种工具的结果高度相似。尽管 PdfAct 声称支持表格提取，但它并未对任何包含表格的测试文档生成输出。所有工具在表格提取方面的性能都明显低于提取其他内容元素时的表现，这可能是因为还需要提取额外的结构信息。表格提取的难度也体现在 GROBID GitHub 仓库中用户就此提交的大量问题上[42]。

[40] 更多有关这些工具提取方法的信息，请参见表格 4 。[41] https://grobid.readthedocs.io/en/latest/Principles/[42] https://github.com/kermitt2/grobid/issues/340



	CERMINE	GROBID	PdfAct	Science Parse
标题	0.81	0.91	0.85	0.70
摘要	0.72	0.82	0.16	0.81
作者	0.44	0.52	0.13	0.52

图 6：元数据提取结果。



	CERMINE	GROBID	PdfAct	科学解析	RefExtract
参考文献	0.74	0.79	0.15	0.49	0.49

图 7：参考文献抽取结果。

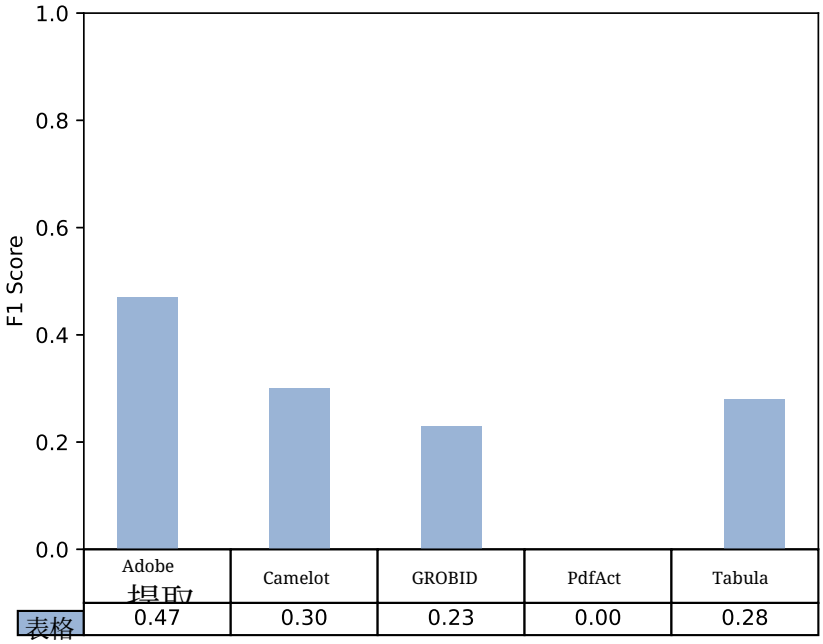


图 8：表格抽取结果。

图 9 展示了通用提取任务的结果。GROBID 的累计得分最高， F_1 为 2.38，其次是 PdfAct（累计$F_1 = 1.66$）。Science Parse 的累计 F_1 得分为 1.25（仅支持段落提取和章节提取），CERMINE 的得分为 1.20，均远低于 GROBID，与 PdfAct 的得分相近。Apache Tika、PyMuPDF 和 Adobe Extract 只能提取段落。

在段落提取方面，GROBID（0.9）、CERMINE（0.85）和 PdfAct（0.85）得分较高，Science Parse（0.76）和 Adobe Extract（0.74）紧随其后。Apache Tika（0.52）和 PyMuPDF（0.51）的得分明显较低，因为这些工具会将章节、图注、列表、页脚和公式等其他元素也纳入段落。

值得注意的是，只有 GROBID 在章节提取方面取得了颇有希望的 F_1 分数 0.74。GROBID 和 PdfAct 是仅有的能够部分提取图注的工具。没有任何工具能够提取列表。只有 PdfAct 支持提取页脚，但其 F_1 分数较低，仅为 0.20。只有 GROBID 支持公式提取，但提取质量相对较低（$F_1 = 0.25$）。为减少评估工作量，我们首先使用覆盖 2014年1月和二月的数据两个月样本，测试列表、页脚和公式的提取效果。如果某个工具的性能分数始终为 0，我们便不再继续评估该工具。按照这一流程，我们仅在完整数据集上评估了 GROBID 和 PdfAct。

对于通用提取任务，GROBID 凭借其分割模型优于其他工具[43]，；该模型根据布局特征检测文档的主要区域。因此，通过将不平衡类别分配给不同模型，段落等常见内容元素就不会影响从正文以外区域提取稀有元素。GROBID 使用的级联模型也便于调整各个模型。以布局和结构为流程基础，有助于关联更简单的训练数据。

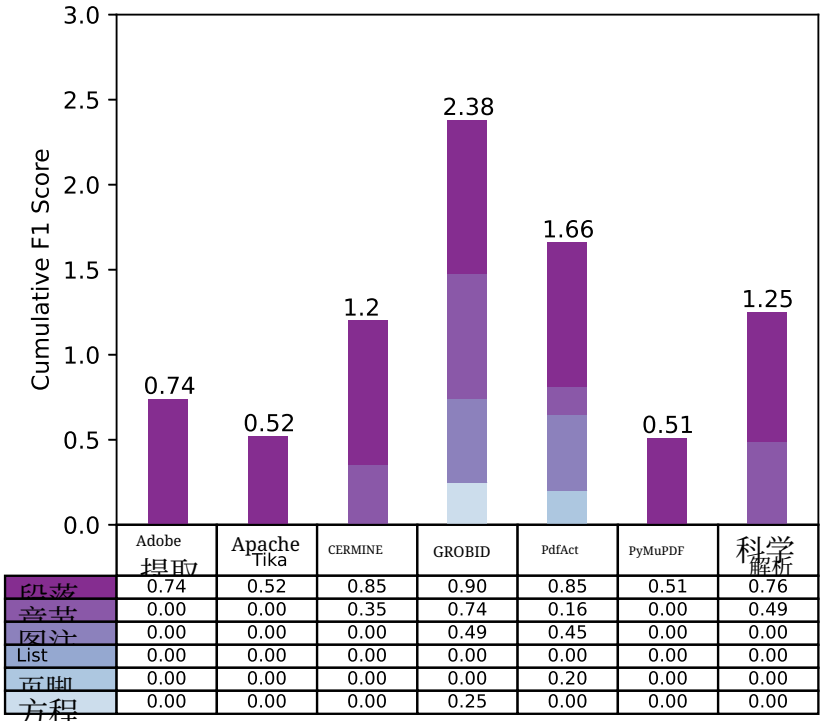


图9：通用数据提取结果。

表格中按工具展示的结果细分情况6 凸显了各项提取任务结果展示的核心结论。不同工具针对不同内容元素的结果差异很大。当然，没有任何工具能在所有元素上表现最佳；相反，即使整体表现良好的工具，也可能在某些提取任务上完全失效。大量内容元素无法提取，或只能以较差的质量提取，这表明未来工作仍有很大的改进空间。

[43] https://grobid.readthedocs.io/en/latest/Principles/

表6：按抽取工具分组的结果。

Tool[1]	标签	# 检-测	# 精-处理[2]	Acc	F_1	P	R
Adobe 提取 表格		1,635	736	**0.52**	**0.47**	**0.45**	**0.49**
	段落 3,985		3,088	0.85	0.74	0.72	0.76
Apache Tika	段落 339,603		258,582	0.55	0.52	0.43	0.65
Camelot	表格	16,289	**11,628**	0.27	0.30	0.23	0.44
CERMINE	标题	16,196	14,501	0.84	0.81	0.81	0.81
	作者	**19,788**	14,797	0.43	0.44	0.44	0.46
	摘要	19,342	16,716	0.71	0.72	0.68	0.76
	参考文献	**40,333**	35,193	0.80	0.74	0.71	0.77
	段落 361,273		348,160	0.89	0.85	0.83	0.87
	章节	**163,077**	139 ,921	0.40	0.35	0.32	0.38
GROBID	标题	16,196	16,018	**0.92**	**0.91**	**0.91**	**0.92**
	作者	**19,788**	**19,563**	**0.54**	**0.52**	**0.52**	**0.53**
	摘要	19,342	**18,714**	0.82	**0.82**	**0.81**	0.83
	参考文献	**40,333**	**36,020**	**0.82**	**0.79**	**0.79**	**0.80**
	段落 361,273		**358,730**	**0.90**	**0.90**	**0.89**	**0.91**
	章节	**163,077**	**163 ,037**	**0.77**	**0.74**	**0.73**	**0.76**
	图注	**90,606**	**62,445**	**0.57**	**0.49**	**0.47**	0.51
	表格	**16,740**	8,633	0.24	0.23	0.23	0.23
	方程	**142,736**	**96,560**	**0.26**	**0.25**	**0.20**	**0.32**
PdfAct	标题	**17,670**	**16,834**	0.85	0.85	0.85	0.86
	作者	13,110	2,187	0.14	0.13	0.12	0.18
	摘要	**21,470**	4,683	0.17	0.16	0.15	0.20
	参考文献	30,263	12,705	0.19	0.15	0.17	0.20
	段落 **361,318**		357,905	0.85	0.85	0.80	0.89
	章节	129,361	87,605	0.21	0.16	0.12	0.25
	图注	83,435	53,314	0.45	0.45	0.40	**0.52**
	页脚	**32,457**	**26,252**	**0.23**	**0.20**	**0.25**	**0.16**
PyMuPDF	段落 339,650		258,383	0.55	0.51	0.41	0.65
RefExtract	参考文献	**40,333**	38,405	0.55	0.49	0.44	0.55
Science Parse	标题	11,696	11,687	0.79	0.70	0.70	0.70
	作者	471	471	**0.54**	**0.52**	**0.52**	**0.53**
	摘要	14,150	14,149	**0.83**	0.81	0.73	**0.90**
	参考文献	**40,333**	35,200	0.55	0.49	0.49	0.50
	段落 **361,318**		355,529	0.79	0.76	0.76	0.76
	章节	**163,077**	158 ,556	0.54	0.49	0.49	0.50
Tabula	表格	10,361	9,456	0.29	0.28	0.20	0.46

[1]粗体表示每种内容元素类型的最佳值。[2] 检测到的项目数与处理的项目数之差，是由于 PDF 读取异常或警告所致。如果某个项目的 F_1 score 不为零，我们就将其标记为已处理。

5 结论与未来工作

我们提出了一个用于从学术 PDF 文档中进行信息抽取的开放评估框架。我们的框架采用 DocBank 数据集 [31] 该数据集涵盖 12 种内容元素类型，并包含从多个学科的 arXiv 论文 50 万页中标注的 150 万个实例。与大多数相关数据集相比，该数据集规模更大、主题更多样，并支持更多提取任务。

我们使用新开发的框架，对十种免费工具在学术 PDF 文档中提取文档元数据、书目参考文献、表格及其他内容元素的性能进行了基准测试。在元数据和参考文献提取任务中，GROBID 表现最佳，其次是 CERMINE 和 Science Parse。在表格抽取方面，Adobe Extract 的表现优于其他工具，尽管其性能远低于其他内容元素的提取效果。所有工具在提取列表、页脚和公式时都表现不佳。

虽然 DocBank 涵盖的学科比其他数据集更多，但我们认为，未来研究中进一步丰富其学科、文档类型和内容元素的多样性是一项有价值的工作。表格 2 显示，目前还有更多适用于从 PDF 文档中进行信息抽取的数据集尚未得到利用。用于创建 DocBank 数据集的弱监督标注方法可以迁移到其他 LaTeX 文档集合中。

除了数据集之外，我们的框架还可以集成其他工具，并能在工具更新时轻松替换。我们计划在未来更新并扩展性能基准测试。

表格、公式、页脚、列表及类似内容元素的提取，是我们基准测试中的工具面临的最大挑战。近期，Grennan 等人[20]表明，使用合成数据集训练模型可以改进引文解析。类似的方法也有望帮助改进对当前难以提取的内容元素的获取。

结合不同的提取方法，或可开发出一种适用于各种场景的抽取工具，我们认为这是理想的方向。Sciencebeam-pipelines[44] 项目目前正朝着这一目标迈出初步步伐。我们希望，我们的评估框架能够通过将信息抽取工具的性能基准测试纳入持续开发与集成流程，助力这一研究方向。

参考文献

1. Ahmed, M.W., Afzal, M.T.：FLAG-PDFe：面向科学出版物的特征导向元数据提取框架。IEEE Access **8**，99458–99469（2020年5月）。https://doi.org/10.1109/ACCESS.2020.2997907 2. Aiello, M., Monz, C., Todoran, L., Worring, M.：面向广泛文档类别的文档理解。《国际文档分析与识别期刊》**5(1)**（2002年8月）。https://doi.org/10.1007/s10032-002-0080-x 3. Anzaroot, S., Passos, A., Belanger, D., McCallum, A.：学习软线性约束及其在引文字段提取中的应用。载于：第52届

[44] https://github.com/elifesciences/sciencebeam-pipelines

计算语言学协会年会（第1卷：长篇论文）。第593–602页。计算语言学协会，巴尔的摩，马里兰州（2014 1056 4 10 3115 1 14）。https://doi.org/ . /v /P - . Azimjonov, J., Alikhanov, J.：基于规则的学术文章元数据提取框架。arXiv CoRR 1807.09009v1 [cs.IR]，1–10（2018）。https://doi.org/10 09009 5 48550 1807 . /arXiv. . . Bast, H., Korzen, C.：Icecite研究论文管理系统。载于：Web信息系统工程——WISE 2013，第8181卷，第396–409页。Springer Berlin，海德堡，南京，中国（2013）。https://doi.org/10.1007/978-3-642- 41154-0 30 6. Bast, H., Korzen, C.：PDF文本抽取的基准测试与评估。载于：2017年ACM/IEEE数字图书馆联合会议（JCDL）。第1–10页。IEEE，多伦多，安大略省，加拿大（2017 7991564 7 10 1109 2017）。https://doi.org/ . /JCDL. . . Bhardwaj, A., Mercier, D., Dengel, A., Ahmed, S.：DeepBIBX：用于基于图像的书目数据抽取的深度学习。载于：第24届神经信息处理国际会议论文集。LNCS，第10635卷，第286–293页。Springer，中国广州（2017）。https://doi.org/10.1007/978-3-319- 70096-0 30 8. Borkar, V., Deshmukh, K., Sarawagi, S.：将文本自动分割为结构化记录。SIGMOD Record 30(2), 175–186（2001年6月）。https://doi.org/10 375682 9 1145 376284 2 0 . / . . Clark, C., Divvala, S.：PDFFigures：从研究论文中挖掘图。载于：第16届ACM/IEEE-CS数字图书馆联合会议论文集。第143–152页。JCDL '16，计算机协会，美国纽约州纽约（2016 2910904 10 10 1145 2910896）。https://doi.org/ . / . . Cortez, E., da Silva, A.S., Gonçalves, M.A., Mesquita, F., de Moura, E.S.：FLUX-CIM：引文元数据的灵活无监督抽取。载于：第7届ACM/IEEE-CS数字图书馆联合会议论文集。第215–224页。JCDL '07，计算机协会，美国纽约州纽约（2007）。https://doi.org/10 1255219 11 1145 1255175 . / . . Cortez, E., da Silva, A.S., Gonçalves, M.A., Mesquita, F., de Moura, E.S.：一种从书目引文中提取元数据的灵活方法。JASIST 60 21049 12 6 1144 1158 2009 10 1002（），–（2009年6月）。https://doi.org/ . /asi. . Councill, I., Giles, C.L., Kan, M.Y.：ParsCit：开源条件随机场参考文献字符串解析包。载于：第六届语言资源与评测国际会议论文集。欧洲语言资源协会，摩洛哥马拉喀什（2008），https://aclanthology.org/L08-1291/ 13. Cui, B.G., Chen, X.：一种改进的文献元数据提取隐马尔可夫模型。载于：高级智能计算理论与应用，第6215卷，第205–212页。Springer Berlin Heidelberg，中国长沙（2010）。https://doi.org/10 1 26 14 1007 978 3 642 14922 . / - - - - . Day, M.Y., Tsai, R.T.H., Sung, C.L., Hsieh, C.C., Lee, C.W., Wu, S.H., Wu, K.P., Ong, C.S., Hsu, W.L.：采用分层知识表示框架的参考文献元数据提取。决策支持系统 43(1), 152–167（2007年2月 006 15 10 1016 2006 08）。https://doi.org/ . /j.dss. . . . De La Torre, M., Aguirre, C., Anshutz, B., Hsu, W.：MATESC：面向科学出版物的元数据分析文本抽取器与章节分类器。载于：第10届知识发现、知识工程与知识管理国际联合会议论文集。第1卷，第261–267页。SciTePress（2018）。https://doi.org/10.5220/0006937702610267

16. Fan, T., Liu, J., Qiu, Y., Jiang, C., Zhang, J., Zhang, W., Wan, J.: PARDA：用于评估学术 PDF 文档元数据提取的数据集。载于：协作计算：网络、应用与协同工作。第 417–431 页。Springer，Cham（ 2019 10 1007 978 3 030 12981 1 ¯29 ）。
17 ¨
https://doi.org/ . / - - - - . Farber, M., Thiemann, A., Jatowt, A.: 面向引文相关任务的高质量金标准。载于：第十一届国际语言资源与评测会议论文集。欧洲语言资源协会，日本宫崎（2018 1296 18 18 ），https://aclanthology.org/L - . Giuffrida, G., Shek, E.C., Yang, J.: 从 PostScript 文件中基于知识提取元数据。载于：第五届 ACM 数字图书馆会议论文集。第 77–84 页。DL '00，计算机协会，纽约州纽约，美国（2000 336639 19 10 1145 336597 ）。https://doi.org/ . / . . Granitzer, M., Hristakeva, M., Jack, K., Knight, R.: 面向众包书目元数据管理的元数据提取技术比较。载于：第 27 届 ACM 应用计算年度研讨会论文集。第 962–964 页。SAC '12，计算机协会，纽约州纽约，美国（2012）。https://doi.org/10 2245462 20 1145 2245276 . / . . Grennan, M., Beel, J.: 用于引文解析的合成与真实参考文献字符串，以及重新训练和样本外数据对于有意义评估的重要性：使用 GROBID、GIANT 和 CORA 开展的实验。载于：第八届科学出版物挖掘国际研讨会论文集。第 27–35 页。计算语言学协会，中国武汉（2020），https: // aclanthology.org/2020 4 21 1.wosp- . . Grennan, M., Schibel, M., Collins, A., Beel, J.: GIANT：用于深度引文解析的十亿条带标注合成书目参考文献字符串数据集 [数据]（2019）。https://doi.org/10.7910/DVN/LXQXAO 22. Hashmi, A.M., Afzal, M.T., ur Rehman, S.: 从科学类PDF文档中提取元数据的基于规则的方法。载于：2020年第五届智能系统与工业应用创新技术国际会议（CITISIA）。第 1–4 页。IEEE，澳大利亚悉尼（2020）。https://doi.org/10 9371784 23 1109 50690 2020 . /CITISIA . . . Hetzner, E.: 使用隐马尔可夫模型提取引文元数据的简易方法。载于：第八届 ACM/IEEE-CS 数字图书馆联合会议论文集。第 280–284 页。JCDL '08，计算机协会，纽约州纽约，美国（2008 1378937 24 10 1145 1378889 ）。https://doi.org/ . / . . Kasdorf, W.E.:《哥伦比亚数字出版指南》。哥伦比亚大学出版社，美国（2003）25. Kern, R., Jack, K., Hristakeva, M.: TeamBeam——科学文献元数据提取。D-Lib 杂志 18(7/8)（2012 年 7 月）。https://doi.org/10.1045/july2012-kern 26. Klein, D., Manning, C.D.: 自然语言处理模型中的条件结构与条件估计。载于：2002 年自然语言处理实证方法会议论文集（EMNLP）。第 9–16 页。计算语言学协会，宾夕法尼亚州费城，美国（2002）。https://doi.org/10 1118695 27 3115 1118693 . / . . Klink, S., Dengel, A., Kieninger, T.: 基于版面与文本特征的文档结构分析。载于：IAPR 文档分析系统国际研讨会。国际模式识别协会，巴西里约热内卢（2000）28. K¨orner, M., Ghavimi, B., Mayr, P., Hartmann, H., Staab, S.: 使用基于行的条件随机场评估参考文献字符串提取：德语出版物案例研究

——德语出版物案例研究。载于：数据库与信息系统新趋势。第 137–145 页。Springer，Cham（2017）。https://doi.org/10.1007/978-3-319- 67162-8 15 29. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: 条件随机场：用于序列数据分段与标注的概率模型。载于：第十八届国际机器学习会议论文集。第 282–289 页。ICML '01，Morgan Kaufmann Publishers Inc.，美国加利福尼亚州旧金山（2001），https://dl.acm.org/doi/10 655813 30 5555 645530. / . . Ley, M.: DBLP：一些经验教训。VLDB基金会会议录 2(2)，1493–1500（2009 年 8 月 1687577 31 10 14778 1687553 ）。https://doi.org/ . / . . Li, M., Xu, Y., Cui, L., Huang, S., Wei, F., Li, Z., Zhou, M.: DocBank：文档版面分析基准数据集。载于：第28届国际计算语言学会议论文集。第 949–960 页。国际计算语言学委员会，西班牙巴塞罗那（线上）（2020）。https://doi.org/10 82 32 18653 1 2020 . /v / .coling-main. . Lipinski, M., Yao, K., Breitinger, C., Beel, J., Gipp, B.: 面向科学类PDF文档的页眉元数据提取方法与工具评估。载于：第13届ACM/IEEE-CS数字图书馆联合会议论文集。第 385–386 页。JCDL '13，计算机协会，纽约州纽约，美国（2013 2467753 33 10 1145 2467696 ）。https://doi.org/ . / . . Livathinos, N., Berrospi, C., Lysak, M., Kuropiatnyk, V., Nassar, A., Carvalho, A., Dolfi, M., Auer, C., Dinkla, K., Staar, P.: 使用循环神经网络实现稳健的 PDF 文档转换。《AAAI人工智能会议论文集》35(17)，15137–15145（2021 年 5 月）。https://doi.org/10 17777 34 1609 35 17 . /aaai.v i . . Lo, K., Wang, L.L., Neumann, M., Kinney, R., Weld, D.: S2ORC：语义学者开放研究语料库。载于：计算语言学协会第58届年会论文集。第 4969–4983 页。计算语言学协会，线上（2020）。https://doi.org/10.18653/v1/2020.acl- main.447 35. Lopez, P.: GROBID（2008），https://github.com/kermitt2/grobid 36. Lopez, P.: GROBID：结合自动书目数据识别与学术出版物术语提取。载于：数字图书馆研究与先进技术，LNCS，第 5714 卷，第 473–474 页。Springer Berlin Heidelberg（2009 8 62 37 10 1007 978 3 642 04346 ）。https://doi.org/ . / - - - - . Mao, S., Kim, J., Thoma, G.R.: 用于数字资料保存的自动元数据提取动态特征生成系统。载于：第一届文档图像分析与图书馆国际研讨会。第 225–232 页。IEEE计算机学会，美国加利福尼亚州帕洛阿尔托（2004）。https://doi.org/10 1263251 38 1109 2004 . /DIAL. . . Mao, S., Rosenfeld, A., Kanungo, T.: 文档结构分析算法：文献综述。载于：文档识别与检索X会议论文集。SPIE会议论文集，第 5010 卷，第 197–207 页。SPIE，美国加利福尼亚州圣克拉拉（2003 年 1 月 476326 39 10 1117 12 ）。https://doi.org/ . / . . McCallum, A.K., Nigam, K., Rennie, J., Seymore, K.: 使用机器学习实现互联网门户自动构建。《信息检索》3(2)，127–163 1009953814988 40 2000 10 1023（7 月）。https://doi.org/ . /A: . 美国国家医学图书馆：PubMed，https://pubmed.ncbi.nlm.nih.gov/ 41. 美国国家医学图书馆：PubMed Central，https://www.ncbi.nlm.nih.gov/ pmc/

42. Ojokoh, B., Zhang, M., Tang, J.：用于从异构参考文献中提取元数据的三元组隐马尔可夫模型。《信息科学》181(9)，1538–1551（2011年5月）。https://doi.org/10.1016/j.ins.2011.01.014 43. Ororbia, A.G., Wu, J., Khabsa, M., WIlliams, K., Giles, C.L.：CiteSeerX中的大型学术数据：从万维网中进行信息抽取。载于：第24届万维网国际会议论文集。第597–602页。WWW ’15 Companion，计算机协会，纽约州纽约（2015）。https://doi.org/10.1145/2740908.2741736 44. Palmero, G., Dimitriadis, Y.：使用一种新型循环模糊神经系统进行结构化文档标注和规则提取。载于：第五届文档分析与识别国际会议论文集。第181–184页。Springer，印度班加罗尔（1999）。https://doi.org/10.1109/ICDAR.1999.791754 45. Peng, F., McCallum, A.：使用条件随机场从研究论文中准确提取信息。载于：计算语言学协会北美分会人类语言技术会议论文集：HLT-NAACL。第329–336页。计算语言学协会，美国马萨诸塞州波士顿（2004），https://aclanthology.org/N04-1042 46. Prasad, A., Kaur, M., Kan, M.Y.：Neural ParsCit：一种基于深度学习的参考文献字符串解析器。《国际数字图书馆杂志》19(4)，323–337（2018年11月）。https://doi.org/10.1007/s00799-018-0242-1 47. Rizvi, S.T.R., Dengel, A., Ahmed, S.：一种用于书目参考文献提取的混合方法与统一框架。IEEE Access 8，217231–217245（2020年12月）。https://doi.org/10.1109/ACCESS.2020.3042455 48. Rizvi, S.T.R., Lucieri, A., Dengel, A., Ahmed, S.：对用于文档图像中基于图像的参考文献检测的目标检测网络进行基准测试。载于：2019年数字图像计算：技术与应用（DICTA）。第1–8页。IEEE，澳大利亚珀斯（2019）。https://doi.org/10.1109/DICTA.2019.8945991 49. Rodrigues Alves, D., Colavizza, G., Kaplan, F.：从艺术与人文科学文献中深度挖掘参考文献。《研究计量与分析前沿》 **3**，21 2018 10 3389 2018 00021（7月）。50 https://doi.org/10.3389/frma.2020.00001 50. Saier, T., Farber, M.：Bibliometric-Enhanced arXiv：用于基于论文和基于引文的任务的数据集。载于：第8届计量学增强信息检索国际研讨会（BIR）论文集。CEUR Workshop Proceedings，第2345卷，第14–26页。CEUR-WS.org，德国科隆（2019），http://ceur-ws.org/Vol-2345/paper2.pdf 51. Saier, T., F¨arber, M.：unarXive：一个大型学术数据集，包含出版物全文、带注释的文内引文以及指向元数据的链接。《科学计量学》125(3)，3085–3108（2020年12月）。https://doi.org/10.1007/s11192-020-03382-z 52. 达格施图尔城堡—莱布尼茨信息学中心、特里尔大学：dblp：计算机科学书目，https://dblp.org/ 53. Souza, A., Moreira, V., Heuser, C.：ARCTIC：使用双层条件随机场从PDF科学论文中提取元数据。载于：2014年ACM文档工程研讨会论文集。第121–130页。DocEng ’14，计算机协会，纽约州纽约（2014）。https://doi.org/10.1145/2644866.2644872 54. Tkaczyk, D., Collins, A., Sheridan, P., Beel, J.：机器学习与规则、开箱即用与重新训练：开源书目参考文献和引文解析器评估。载于：第18届ACM/IEEE数字图书馆联合会议论文集。第99–108页。JCDL ’18，计算机协会，纽约州纽约（2018）。https://doi.org/10.1145/3197026.3197048

55. Tkaczyk, D.,Szostek,P., Fedoryszak, M., Dendek,P.J., Bolikowski, L.：CER MINE：从科学文献中自动提取结构化元数据。《国际文档分析与识别杂志》（IJDAR）18(4)，317–335（2015年12月8 56101007100320150249）。https://doi.org/./s---.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.：Attention Is All You Need。载于：第31届神经信息处理系统国际会议论文集。第6000–6010页。NIPS'17，Curran Associates Inc.，美国纽约州雷德胡克（2017），https://dl.acm.org/doi/103295349 5755553295222./..
Vilnis, L., Belanger, D., Sheldon, D., McCallum, A.：用于非局部推理的Bethe投影。载于：第三十一届人工智能不确定性会议论文集。第892–901页。UAI'15，AUAI出版社，弗吉尼亚州阿灵顿（2015）。https://doi.org/10.48550/arXiv.1503.01397