基于表格检索和机器学习二阶段的文献表格相关文本自动识别*

黄佳妮 于丰畅
（武汉大学信息管理学院，武汉 430072）

摘要：学术文献中的表格以结构化的形式高度凝练地展示了文献中的核心知识。主流文献检索引擎中已逐步开始使用表格内容作为文字摘要的补充，以帮助科研人员快速掌握研究工作核心，提升科研工作效率。但是在仅展示表格而不提供表格的相关信息（对表格进行描述或解释的文本）的情况下，读者往往难以充分理解表格内容，阻碍文献阅读效率的进一步提升。本文提出一种基于表格检索和机器学习二阶段的表格相关文本识别方法，阶段一运用表格内容进行全文检索，获取潜在相关文本；阶段二构建机器学习模型，判断表格与潜在相关文本间的相关性，从而实现文献中表格相关文本的自动识别。以Text Retrieval Conference会议论文数据集为例，验证本文所提出的方法的有效性，证明该方法能够快速抽取文献中与图表相关的文本，为现有的论文图表抽取式摘要相关研究提供借鉴，对提高科研人员文献调研效率具有重要的现实意义。

关键词：文献表格；表格理解；机器学习
中图分类号：TP391　DOI：10.3772/j.issn.1673-2286.2022.11.009
引文格式：黄佳妮，于丰畅. 基于表格检索和机器学习二阶段的文献表格相关文本自动识别[J]. 数字图书馆论坛，2022 (11)：34-42.

近年来，通信、网络技术的不断进步促进了学术交流，加速了学术成果的产出，学术文献数量也呈指数级增长。2015年，仅在医学领域，平均每天就有约2 200篇新论文发表[1]。海量文献对科研人员的文献调研、阅读学习等科研工作提出了挑战，如何在尽可能少的时间内，从文献中获取尽可能多的有效信息，成为亟待解决的问题。深度学习、自然语言处理等技术的兴起，为海量学术文献的自动化处理、论文核心知识的自动抽取提供了可能[2-4]。

当前的大多数研究主要关注学术文献的正文，往往忽略对图像、表格和其他半结构化信息的分析和处理。图像、表格是学术文献的重要组成部分，它们集中体现了学术研究的主要研究内容，图表内容也常用于支撑文献核心观点。多项研究表明，表格通常用于呈现实验的设置和结果，以及已有实验、背景或术语定义的相关信息[5-6]。Futrelle[7]以生物科学领域的文献为例展开研究，发现学术文献中的图表及其相关文本描述占整篇论文的50%。相较于文字内容，图表内容因其简洁明了的视觉特性，在阅读速度上有较大的优势。因此，以图表内容作为科技文献摘要的补充信息，是一种辅助科研人员快速定位、理解文献的可行手段。包括Springer、CNKI、Semantic Scholar在内的多家科技文献服务商也逐步将文献中的图表纳入检索范围，提供文献图表检索功能。

然而，当前此类服务尚不完善，其主要原因在于：学术文献中的图表，特别是表格，以结构化的形式高度概括了文献的实验流程、研究成果等关键知识，其表现形式具有一定的抽象性，要求读者具备相关的领域知识。在缺乏表格相关上下文信息的情况下，读者往往很难充分理解表格内容[8]。Yu等[9]的实证研究表明，仅提供表标题而不提供相关补充信息将显著降低受试者对表格的理解程度。读者无法通过图表理解文献的主要内容时，只能重新走上通篇阅读文献的老路，阅读效率

* 本研究得到2021年度湖北省博士后创新研究岗位项目“基于迁移学习的开放领域非格式化文档理解”（编号：211000090）资助。

无法得到有效提高。因此，从学术文献全文中识别出与图表相关的信息，对帮助读者充分理解表格的含义，节省文献阅读时间有重要的现实意义。

为实现自动识别学术文献中与表格相关的信息、辅助科研人员快速理解表格内容、提升学术调研工作效率，本文提出一种基于表格检索和机器学习的二阶段方法，阶段一运用表格内容进行全文检索，将检索结果作为与表格内容潜在相关的文本；阶段二在学习检索特征和文本特征的基础上，使用机器学习分类方法对潜在相关文本与具体表格的相关性进行判断。然后通过收集的Text Retrieval Conference会议论文数据，对本文提出的方法进行验证，得到了较好的结果。

1 相关研究

本研究基于表格检索和机器学习，实现表格相关文本的识别，主要涉及表格识别与定位以及表格相关文本抽取这两个领域，其相关研究状况如下。

1.1 表格识别与定位

国内外学者对PDF文件中表格的识别、定位进行了多方面、深入的研究。窦方坤等[10]以药学文献为主要研究对象，抽取文献PDF中的所有文本元素，确定表标题文本所在区域，将表标题以下的区域看作表格所在区域。于丰畅等[11]运用机器视觉技术和PDF解析技术，从底层编码分析和图片理解两种视角获取图表范围的先验信息，通过对PDF中的几何对象进行聚类来确定图表坐标。田翠华等[12]基于pdfplumber，设计了一款基于Python平台对PDF文档中的表格进行识别和提取的软件。Siegel等[13]提出一种在大量科学文献中为图表形成高质量标签的无监督方法，并使用这个数据集训练了一个用于表格检测的深度神经网络。近年来深度学习技术发展迅速，越来越多的研究人员使用深度学习方法实现PDF中的表格检测。这些研究根据网络类型，可以分为目标检测算法（Faster R-CNN[14]、Mask R-CNN[15]、YOLO[16]等）、卷积神经网络（CNN[17]）和图神经网络（GNN[18]）等。

1.2 表格相关文本抽取

表格相关文本提取的相关研究主要围绕为表格生成提取式摘要展开。1999年，Futrelle等[19]手工构建了4个科技文献中图表摘要的例子，讨论了自动化图表摘要生成的流程和相关算法。Jain等[20]提出了一种基于注意力机制的混合分层Encoder-Decoder模型，该模型能够利用表内容之外的结构，但其局限性在于只能对固定模式的表格进行处理。Yu等[21]采用分层聚类技术，基于词汇相似性对学术文献中的句子和图表进行聚类，根据聚类结果确定与图表相关的文本。Agarwal等[22]构建了自动为生物医学文献中的图表生成结构化文本摘要系统FigSum，生成的结构化摘要由4类句子生成，包括图表的背景信息、实现图表所示内容的方法等。不过，FigSum的实验数据仅基于44个生物医学领域论文中的图表，因此模型的泛化能力有待进一步研究。Bhatia等[23]找到文档文本中引用图表的句子，计算学术文献中每个句子与引用句的相似度和接近度，从而确定与图表相关的文本；此外，还研究了如何选择最佳的图表摘要大小，以在信息的完备度和生成的摘要长度间取得平衡。Takeshima等[24]提出一种权重传播机制，在“单词重要性估计”和“句子权重更新”等过程中确定与图表相关度最高的句子。Park等[25]提出了一种基于本体的、从论文正文中提取图表描述性文本的方法，为句子构建知识表示，采用本体语义来辅助图表相关信息的概念识别。Saini等[26]提出了一种新的无监督方法（FigSum++），使用多目标进化算法对生物科学领域的文章自动生成图表摘要。也有学者对现有的自动生成图表摘要系统进行了对比评估。如Polepalli等[27]通过从19种不同的期刊中选取94个带注释的图表，对一系列FigSum+系统进行评价，并通过准确性、召回率、F1和ROUGE分数来评估测试结果，结果表明：最好的FigSum+系统是基于无监督方法的系统，F1得分为0.66，ROUGE-1得分为0.97。

通过文献调研发现，现有相关研究仍存在可改进之处。已有研究往往针对单一表结构展开，且以学术文献中的表格作为研究对象的较少，没有充分利用表格标题、注释等表格相关文本。所使用的方法也较为局限，如基于文本相似度、基于本体、基于规则等，往往依赖大量人工处理，难以扩展到大规模学术文献数据集上。本研究所提出的方法不受表格结构、格式的影响，可扩展应用于不同布局的表格。此外，本研究将表格标题纳入表格全文检索的检索字段中，辅助表格相关文本的识别。

尽管本文所提出的方法旨在识别学术文献原文中

与表格相关的描述信息，不属于文本生成任务，但可为图表摘要的自动生成相关研究提供借鉴。此外，表格相关文本的自动识别还能在辅助科研人员快速理解表格内容、提升学术调研工作效率、增强图表检索效果等多方面发挥作用。

2 基于表格检索和机器学习二阶段的文献表格相关文本识别

2.1 问题界定

本文所提出的方法旨在自动识别并抽取学术文献PDF中的表格以及对表格进行描述、解释的表格相关文本。本研究将表格相关文本定义为一组对表格进行描述或解释的句子，如图1所示，示例中阴影部分即为与表格Table 2相关的文本。该表格主要对模型在各个实验指标上的效果进行展示，表格相关文本的主要内容是对不同模型以及指标数值的阐述和对比分析。

2.2 研究思路

在规范的学术文献正文中，作者为了对实验指标或实验配置进行具体展示、描述或讨论，会使用诸如“如表1所示，某指标……”之类明确的关于表格中具体内容的引用。基于这种写作规范，本文将表格内容作为检索词，通过检索的方式在正文中查找表格内容可能相关的信息。需要指出的是，由于表格中文字内容数量较少且专有名词占比较大，检索结果不可避免包含并非与具体表格存在直接关联的内容，如相关研究章节中对于其他研究中指标性能的介绍。因此，仍需要从语义角度对检索到的文本是否与具体表格相关进行判断。

本文提出一种基于表格检索和机器学习二阶段的文献表格相关文本识别方法，研究思路如下。

阶段一：基于表格检索的潜在相关文本获取。识别并抽取文献中的表格数据内容，在表格数据、表格标题的基础上构建检索词，进行全文检索，获取与表格内容有潜在相关关系的文本。

2.3 Documents as visits

The guidelines for the Medical Records Track stated that the unit of retrieval should be a single patient visit. A visit is a single admission for a single patient — if the same patient is admitted on two different occasions these will be viewed as two separate visits. Our approach was to treat individual reports as sub-documents and compile them together with all the other reports pertaining to a single patient admission into a single larger document. The unit of retrieval is then a 'patient visit' rather than individual medical reports. As all reports for a single visit are concatenated together we make no distinction as to the different reports type — radiology, discharge summary, etc.

3 Results and Analysis

Table 2 reports the results we obtained in comparison to median values obtained across all systems. Overall, our concept-based approaches demonstrate improvements over the median of the TREC systems.

	infAP (%Δ)	infNDCG (%Δ)	R-prec (%Δ)	Prec@10 (%Δ)	Recall
Median	0.1689	0.4243	0.2960	0.4702	
AEHRC0	0.2130 (+26%)	0.4630 (+9%)	0.3251 (+10%)	0.4894 (+4%)	0.6406
AEHRCsub	0.2163 (+28%)	0.4682 (+10%)	0.3314 (+12%)	0.5043 (+7%)	0.6702
ARHRC1	0.2066 (+22%)	0.4614 (+8%)	0.3140 (+7%)	0.4596 (-2%)	0.6675
ARHRC2	0.2128 (+25%)	0.4608 (+8%)	0.3205 (+9%)	0.4553 (-3%)	0.6803

Table 2: Comparison of our concept-based approaches to the median result obtained across all TREC systems.

3.1 Concept-based IR contribution

Results are heavily dependent on the quality of concept extraction provided by the MetaMap system. MetaMap only identifies UMLS concepts, which are then mapped to SNOMED CT concepts. Mapping between terminologies may result in a loss in meaning from the original query or document. Certain UMLS concepts have no equivalent in SNOMED CT, such cases were found in the worst performing queries, e.g. query 178 "Patients with metastatic breast cancer". Nevertheless, our concept-based baseline (AEHRC0) provides consistent improvements over the median of TREC systems.

3.2 Effect of Weighting Concept Relationships

We now consider the contribution of combining weights of query concepts and related concepts for scoring a document. Empirical results show that considering related concepts along with the original query concepts can improve retrieval effectiveness; which concept relationships to consider and how to weight these is however a challenging issue. Our AEHRC1 and AEHRC2 submission have proved that not all concept relationships lead to improvements of retrieval

图1 文献表格相关文本识别任务示例

阶段二：基于机器学习的相关性判断。构建检索特征和文本语义特征融合的机器学习模型，学习文本检索结果是否与表格内容相关，若相关则进一步判断具体与哪一个表格相关。将检索特征和文本特征拼接作为机器学习模型的输入，机器学习模型输出输入特征所表征的文本是否与文献中的任一表格相关。若相关，则根据阶段一中表格的全文检索结果进一步确定其具体与哪一表格相关。整体研究流程如图2所示。

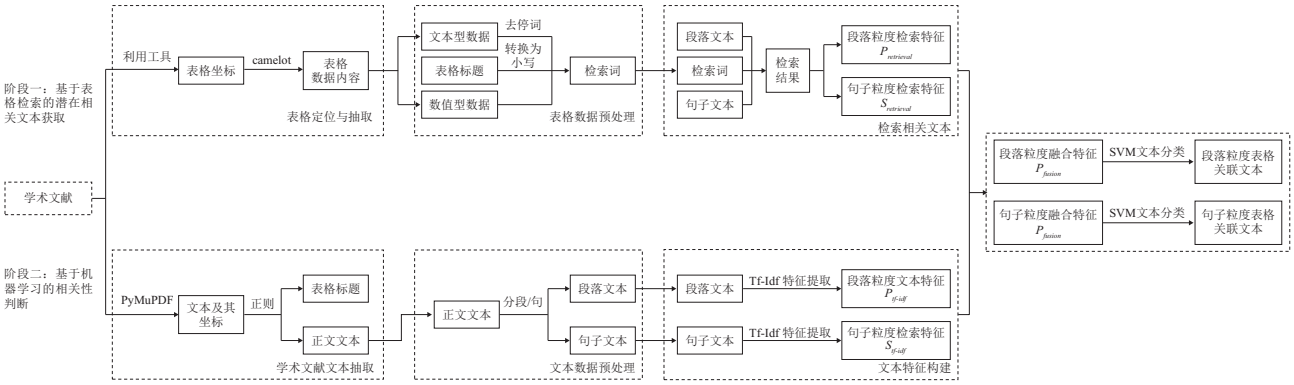

图2　整体研究流程

（1）表格定位与抽取。表格定位是指识别PDF文件中的表格，并获取表格在PDF中的坐标。本研究以于丰畅等[11]提出的基于机器视觉的PDF学术文献结构识别方法作为本任务的表格定位算法，获取表格坐标并存入结构化数据库。

获得表格坐标后，需要根据坐标抽取表格的数据内容。本研究调用camelot第三方库抽取特定坐标的表格所对应的数据内容。

（2）文本抽取与表格、表标题匹配。利用PyMuPDF库读取PDF文件，获取PDF全文文本及坐标，根据学术论文中表格标题特征编写正则表达式，对获取的PDF全文文本按句匹配，得到表格标题文本。计算表格标题文本与表格的欧氏距离，实现表格标题与表格的一一对应。

（3）候选检索词生成。针对表格数据内容，若数据值为数字，则直接输出为检索词；若数据值为文本，则去除停用词、转换为小写后输出为检索词。此外，对表格标题进行分词、删除“table”、去除停用词等预处理，形成表格候选检索词。

（4）潜在相关文本获取与文本特征构建。使用上一步获得的检索词，运用全文检索技术，对段落粒度、句子粒度的文本进行检索，并构建$n\times1$维的检索特征向量，其中n是文献中段落/句子的数量。对于一篇学术文献中的所有句子或段落，若其出现在该文献中任一表格的检索结果中，则检索特征值为1；若其未出现在检索结果中，则检索特征值为0。经过此步骤，可以确定与表格有潜在相关关系的文本。

本文构造了基于TF-IDF的语义特征向量，对学术文献正文文本进行特征提取，获得学术文献段落粒度文本、句子粒度文本的TF-IDF特征值。将检索特征与文本特征拼接，得到融合特征，作为后续文本与表格的关联性预测的输入特征。

（5）文本关联性预测与表格关联文本确定。文本关联性预测是二分类文本任务，即将与表格有潜在相关关系的文本分为“与文献中的任一表格相关”或“与文献中所有表格无关”两类。本研究采用支持向量机（Support Vector Machines，SVM）算法完成分类任务。运用Scikit-learn机器学习库中提供的SVM模型，以“潜在相关文本获取与文本特征构建”步骤中构建的融合特征作为输入，实现文本关联性预测。

对预测结果为“与文献中的任一表格相关”的文本t，结合表格全文检索结果，若其被某一表格的检索词检索到，则认为文本t是该表格的关联文本。

3　实验与讨论

3.1　数据来源

本研究数据来源为Text Retrieval Conference会议论文集，论文采集情况如表1所示。

表1 Text Retrieval Conference会议论文集

时间跨度/年	论文数量/篇	页面数量/页	表格数量/个	论文语言	文件格式
2012—2019	635	4 224	1 229	英语	PDF

3.2 实验设置

3.2.1 数据预处理

数据预处理步骤如下：①调用学术文献解析工具Grobid[28]对学术文献的正文进行识别，对作者姓名与机构、参考文献等与实验无关的文本进行过滤；②对从学术论文PDF中抽取的段落粒度的文本，运用Nltk库[29]进行句子分割，形成句子粒度的文本数据集；③将文本中的换行符、制表符、单词跨行连字符删除，并统一转换为小写。

3.2.2 数据标注

由于时间有限，且标注全部数据的人力成本较高，实验按照表格布局、表格数据类型等特征对Text Retrieval Conference会议论文集进行采样，共选取263篇论文中的303个表格进行标注，标注情况如表2所示。后续研究将进一步扩充数据集，增加标注数据量。从人工标注结果来看，在数据集中与表格无关的文本占比较大，存在明显的数据不平衡问题，会对机器学习模型的分类性能造成影响。因此，设与表格相关的文本为正例，与表格无关的文本为负例，按照正例∶负例=3∶1的比例对负例进行下采样。

表2 文献表格相关文本数据集

粒 度	学术文献数/篇	表格数/个	段落数/个	正例数/个	负例数/个	负例抽样数/个
段落	263	303	4 950	423	4 527	1 269
句子	263	303	11 469	2 149	9 320	6 447

标注示例如图3所示，标注文件中保存段落/句子粒度的学术文本、所属章节、文本是否与文献中的某一表格相关。此标注文件用于评价文献表格有效信息关联任务的最终效果。

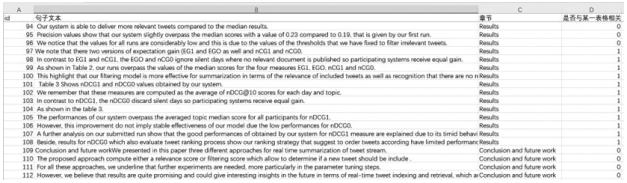

图3 “相关文本-表格”标注文件示例

在标注完成后，对每一文献中的所有表格标注文件进行遍历，对于文献中的每一条文本记录，只要与任一表格相关，就标记为表格相关文本，用于文本关联性预测任务的评价。

3.2.3 实验细节和参数设置

实验各流程的相关细节和参数设置如表3所示。

3.3 实验结果和分析

3.3.1 消融实验

实验包含两个子阶段，分别是表格相关文本预测

表3 实验参数设置

实验环节	相关工具	参数设置	补充说明
表格抽取	camelot	Flavor=stream	在以空格分隔单元格的表格上也有较好的效果
文本抽取	PyMuPDF	-	获取段落粒度、句子粒度文本
	Grobid	-	识别文本中的作者名称、参考文献等并删除
全文检索	Elastic Search	匹配策略：match	minimum_should_macth表示检索结果必须包含检索语句中70%的词汇；检索结果分数范围为多次实验经验值
		minimum_should_match：70%	
		检索结果分数：3~30	
文本特征构建	Scikit-learn（Tfidf）	默认参数设置	-
文本相关性预测	Scikit-learn（SVM）	Kernel=’linear’，	GridSearch选择最佳参数
		class_weight=1∶2，C=2	

和“相关文本-表格”关联关系确定。为探究本文提出的基于表格检索与机器学习二阶段模型的有效性以及不同粒度文本对实验结果的影响，开展以下对比实验：①以仅使用TF-IDF特征的SVM模型作为表格相关文本预测实验的baseline模型，与提出的二阶段模型的实验结果对比，观察结合检索特征是否能提升相关文本预测的准确率、召回率等指标。②以仅通过表格检索获得潜在相关文本的模型作为“相关文本-表格”关联关系确定实验的baseline，与提出的二阶段模型的实验结果对比，观察在潜在相关文本的基础上进一步进行文本分类是否能提升文本、表格间一一对应的效果。实验结果如表4、表5所示。其中，0代表文本与表格无关，1代表文本与表格相关。

表4　表格相关文本预测实验结果

方　法	粒　度	评价指标				
		精确率	召回率	准确率	$F1_{macro}$	$F1_{weighted}$
TF-IDF+SVM（baseline）	段落	0.92	0.77	0.88	0.81	0.87
检索+SVM二阶段（模型一）	段落	0.89	0.83	0.90	0.86	0.90
TF-IDF+SVM（baseline）	句子	0.82	0.76	0.85	0.78	0.84
检索+SVM二阶段（模型二）	句子	0.84	0.80	0.86	0.81	0.86

表5　“相关文本-表格”关联实验结果

方　法	粒　度	结果类别	评价指标				
			精确率	召回率	准确率	$F1_{macro}$	$F1_{weighted}$
检索（baseline）	段落	0	0.98	0.85	0.84	0.56	0.89
		1	0.13	0.59			
模型一+检索	段落	0	0.83	0.66	0.68	0.67	0.69
		1	0.51	0.72			
检索（baseline）	句子	0	0.95	0.90	0.86	0.62	0.88
		1	0.25	0.40			
模型二+检索	句子	0	0.82	0.83	0.74	0.67	0.74
		1	0.54	0.52			

从以上两表可知，在文本相关性预测实验中，将表格检索结果与文本特征结合的方法使得召回率、精确率、*F1*都有所提升。在“相关文本-表格”关联实验中，将检索得到的潜在相关文本直接作为最终表格相关文本的baseline模型，在确定文本、表格间的一一对应关系上的效果较差，准确率、召回率低于本文提出的二阶段方法。

3.3.2 结果分析

（1）不同粒度效果。在表格相关文本预测实验中，段落粒度的实验结果优于句子粒度，推测原因是与表格相关的段落，包含多个与表格关联的句子，分类特征更为明显。

“相关文本-表格”关联实验的结果则相反，在句子粒度上实验结果更佳。例如，“模型二+检索”的实验精确率为0.74，高于“模型一+检索”的0.68。可能的原因是段落容易受到多个句子信息融合的影响，而句子包含的关于特定表格的信息明确，因此更容易判断和哪个表格有关联。

（2）表格相关文本预测效果。表4的实验结果表明，相较于基线模型，结合表格检索结果的方法有显著提升，$F1_{macro}$提升了5%，由此推断：将表格检索的结果与文本特征拼接能够改进机器学习模型在表格相关文本预测实验中的效果。

（3）“相关文本-表格”关联效果。由表5可知，不通过表格相关文本预测实验筛选相关文本，直接根据检索结果确定文本与表格之间的关联关系的方法召回率、准确率均较低。该方法是无监督过程，可以使用全部的表格、文本数据，而数据中负例占比较大，且准确率、召回率高，因此整体的精确率数值较高。

对比表5中段落粒度的两个模型的实验效果，运用SVM机器学习方法的模型在正例（结果类别为1）的精确率提升38%，召回率提升13%。在句子粒度上，SVM机器学习模型也优于基线模型。SVM机器学习模型在段落粒度、句子粒度上均优于基线模型，因此得出结论，在进行“相关文本-表格”的关联之前，先通过表格相关文本预测实验筛选相关文本可以提升模型效果。图4为表格示例，图5为表格相关文本示例，其中阴影部分为本文提出的方法所识别出的与表格相关的文本。

TABLE 1. RESULTS OF OUR TEAM

Run id	MAP	R-Prec	bpref	P@30	Methods
BJUTFreq	0.1088	0.1610	0.1891	0.2328	Frequency
BJUTCnor	0.0729	0.1137	0.1431	0.1822	C measure
BJUTEntr	0.0731	0.1174	0.1493	0.1639	Entropy Differences

图4 表格示例

As we cannot get the whole collection, we only deal with each topic one by one through our system. When we get the final results, the tweets of different topics are independent of each other. And our system mainly contains three parts as the Fig.1 shows.

- The first part is corpus and preprocessing. We already detailed introduce this part in previous chapter.
- The second part is query expansion. It is the most important part. First of all, we make use of the top 100 tweets in the new corpus file of one topic to get the expansion words through our methods that will be introduced detailed in next chapter. Then we combine the original query words with the expansion words to refine search with the following formula:
  The new query words = the original query words + α (the expansion words).
- The final part is getting the results. The number of result tweets is 1000 by one topic, and one result contains the topic number, an unused column, a tweet id, the rank, the score and the run tag such as 'MB111 Q0 311248486941200386 1 -4.13049 BJUTFreq'. And we set the score with tf-idf model.

III. QUERY EXPANSION

In our system, we use three keywords extraction measures for query expansion, which are respectively based on the frequency, words' spatial distributions along the text and the theory of Shannon's entropy difference between the intrinsic and extrinsic modes which refer to the fact that relevant words significantly reflect the author's writing intentions.

For every topic we import the top 100 twitters which are pre-processed through our system into one file which we consider as an initial whole text. All of our three algorithms are based on above hypothesis.

A. Frequency Measure

This frequency based method is set as baseline. In our system, we count every word which appears in the corpus, and build a map structure to save statistical data. We filtered the stop words in our corpus by using an English stop words list. After above all works have done, we arrange all the words in reverse chronological order and select the top-5 words as the expanded key query words.

B. C Measure

P. Carpenas et al. [4] suggest using the statistical analysis of spatial distributions, i.e. spectra, to detect relevant words with the same occurrence frequency. They claim that long-range correlation or clustering (self-attraction to each other) in the spatial distribution of relevant keywords is an important feature of human-written texts, in spite of random occurrences of irrelevant words. Normalized standard deviation (C) of the nearest neighbor spacing is used to characterize the spatial distribution of a particular word.

$$C(\sigma_{nor},n)=\frac{\sigma_{nor}-<\sigma_{nor}>(n)}{sd(\sigma_{nor})(n)} \quad (1)$$

Where $\sigma_{nor}=\frac{\sigma}{\sqrt{1-p}}$ and the parameter σ is defined as $\sigma=s/<d>$ with $<d>$ being the average distance and $s=\sqrt{<d^2>-<d>^2}$.

The C score is our purpose value. $C=0$ indicates that the word appears at randomly, $C>0$ that the word is clustered, and $C<0$ that the word repels itself. In addition, two words with the same C value can have different clustering, but the same statistical significance.

C. Entropy Differences Measure

Yang et al. [5] propose a new metric 'Entropy difference' to evaluate and rank the relevance of words in a text. The method uses the Shannon's entropy difference between the intrinsic and extrinsic mode, which refers to the fact that relevant words significantly reflect the author's writing intentions, i.e., their occurrences are modulated by the author's purposes, while the irrelevant words are distributed randomly in the text.

The idea of intrinsic-extrinsic mode is based on the general idea that relevant words are clustered, and therefore the set of distances between consecutive appearances of a word should consist of small intra-cluster distances and large inter-cluster distances.

A simple way is suggested to distinguish the intrinsic and extrinsic mode, the positions of the word occurrences in a text with frequency m are denoted by $t_1, t_2, t_3,\ldots t_m$. The distance between two successive occurrences of a word, can be written as $d_i=t_{i+1}-t_i$. The arrival time differences belongs to the intrinsic mode d^A if $d_i \leqq \mu$. Thus the intrinsic modes entropy of a word is defined as:

$$H(d^A)=-\sum_{d\in d^A} P_d \log_2 P_d \quad (2)$$

Let $d^B=\{d_i \mid d_i>\mu\}$ be the union set for all $d_i>\mu$. We can define the extrinsic mode entropy of a word as

$$H(d^B)=-\sum_{d\in d^B} P_d \log_2 P_d \quad (3)$$

Thus the entropy differences between the intrinsic and extrinsic mode can be define as follows:

$$ED^q(d)=(H(d^A))^q-(H(d^B))^q \quad (4)$$

Obviously, a word with different p randomly placed in a text would have a different entropy difference between the intrinsic and extrinsic mode, the ED_{geo}, the normalized entropy difference measure ED_{nor} is defined as

$$ED^q_{nor}(d)=ED^q(d)/|ED^q_{geo}(d)| \quad (4)$$

When we have calculated the ED_{nor} for every word, we can sort words by using it. A word with larger ED_{nor} explains that it plays more important role in the text.

IV. RESULTS

In this year's TREC Microblog Track, we submit 3 versions of runs which are shown in the Tab. 1.

图5 表格相关文本示例

（4）整体实验效果。综合表格相关文本预测和“相关文本-表格”关联关系确定两个实验的实验结果可以看出，基于SVM和表格检索模型的段落粒度实验效果最好，与基线实验相比，在各个指标上的提升最明显。在“相关文本-表格”关联关系确定实验中，本研究仅使用表格检索的结果，精确率仍有待提高。

根据表4结果，段落粒度实验在各项指标上优于句子粒度实验，$F1_{weighted}$提高了4%。对比表5中不同粒度模型的实验效果可以发现，两阶段模型在正例分类上相较于baseline模型都有提升，段落粒度最为明显，准确率和召回率分别提升38%、13%，句子粒度为32%、12%。综上，基于SVM和表格检索模型的段落粒度实验效果最好，与基线实验相比，在各个指标上的提升最明显。此外，“相关文本-表格”关联关系确定实验精确率尚有待提高，当前实验的主要不足在于，对被预测为与任一表格关联的文本和具体表格之间的匹配问题，仅使用表格检索的结果，特征不足。后续研究将考虑增加后处理步骤或挖掘其他特征，实现更精确的“相关文本-表格”关联关系确定。

4 结论与局限性

本文提出了一种基于表格检索和机器学习，在学术文献全文中识别表格相关文本的方法，在Text Retrieval Conference数据集上从段落粒度、句子粒度对表格相关文本识别进行了验证。由实验结果可知，本文提出的方法能够对现有的图表摘要进行有效的补充，对提高文献阅读效率具有重要的现实意义。但本研究仍存在不足之处，例如本文使用的机器学习模型对于自然语言理解能力尚有欠缺，且实验效果受表格抽取工具精确度的影响。未来考虑在更加广泛的多学科数据上，使用深度学习自然语言模型作进一步的改进研究。

参考文献

[1] MEDLINE/PubMed Resources [EB/OL]. [2022-11-21]. http://www.nlm.nih.gov/bsd/stats/cit_added.html.

[2] CARVAILLO J C, BAROUKI R, COUMOUL X, et al. Linking bisphenol S to adverse outcome pathways using a combined text mining and systems biology approach [J]. Environmental health perspectives, 2019, 127 (4): 047005.

[3] KVELER K, STAROSVETSKY E, ZIV-KENET A, et al. Immune-centric network of cytokines and cells in disease context identified by computational mining of PubMed [J]. Nature Biotechnology, 2018, 36 (7): 651-659.

[4] TCHOUA R B, CHARD K, AUDUS D, et al. A hybrid human-computer approach to the extraction of scientific facts from the literature [J]. Procedia Computer Science, 2016, 80: 386-397.

[5] YEPES A J, VERSPOOR K. Towards automatic large-scale curation of genomic variation: improving coverage based on supplementary material [J]. BioLINK SIG, 2013, 2013: 39-43.

[6] WONG W, MARTINEZ D, CAVEDON L. Extraction of named entities from tables in gene mutation literature [C] // Proceedings of the BioNLP 2009 Workshop. Stroudsburg: Association for Computational Linguistics, 2009: 46-54.

[7] FUTRELLE R P. Handling figures in document summarization [C] //Text Summarization Branches Out. Association for Computational Linguistics, 2004: 61-65.

[8] SANDUSKY R J, TENOPIR C. Finding and using journal-article components: Impacts of disaggregation on teaching and research practice [J]. Journal of the American Society for Information Science and Technology, 2008, 59 (6): 970-982.

[9] YU H, AGARWAL S, JOHNSTON M, et al. Are figure legends sufficient? Evaluating the contribution of associated text to biomedical figure comprehension [J]. Journal of Biomedical Discovery and Collaboration, 2009, 4 (1): 1-10.

[10] 窦方坤，曹皓伟，徐建良. 基于文本元素的PDF表格区域识别方法研究 [J]. 软件导刊，2020，19 (1)：113-116.

[11] 于丰畅，程齐凯，陆伟. 基于几何对象聚类的学术文献图表定位研究 [J]. 数据分析与知识发现，2020，5 (1)：140-149.

[12] 田翠华，张一平，胡志钢，等. PDF文档表格信息的识别与提取 [J]. 厦门理工学院学报，2020，28 (3)：70-76.

[13] SIEGEL N, LOURIE N, PORWER R, et al. Extracting scientific figures with distantly supervised neural networks [C] // Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries. New York: Association for Computing Machinery, 2018: 223-232.

[14] SCHREIBER S, AGNE S, WOLF I, et al. Deepdesrt: Deep learning for detection and structure recognition of tables in document images [C] //2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2017: 1162-1167.

[15] SAHA R, MONDAL A, JAWAHAR C V. Graphical object detection in document images [C] //2019 International Conference on Document Analysis and Recognition (ICDAR). Piscataway: IEEE, 2019: 51-58.

[16] HUANG Y, YAN Q, LI Y, et al. A YOLO-based table detection method [C] //2019 International Conference on Document Analysis and Recognition (ICDAR). Piscataway: IEEE, 2019: 813-818.

[17] KAVASIDIS I, PINO C, PALAZZO S, et al. A saliency-based convolutional neural network for table and chart detection in digitized documents [C] //International Conference on Image Analysis and Processing. Cham: Springer, 2019: 292-302.

[18] RIBA P, DUTTA A, GOLDMANN L, et al. Table detection in invoice documents by graph neural networks [C] //2019 International Conference on Document Analysis and Recognition (ICDAR). Piscataway: IEEE, 2019: 122-127.

[19] FUTRELLE R P. Summarization of diagrams in documents [J]. Advances in Automated Text Summarization, 1999: 403-421.

[20] JAIN P, LAHA A, SANKARANARAYANAN K, et al. A mixed hierarchical attention based encoder-decoder approach for standard table summarization [J/OL]. arXiv preprint arXiv: 1804.07790 [2022-11-21]. DOI:10.18653/v1/N18-2098.

[21] YU H. Towards answering biological questions with experimental evidence: automatically identifying text that summarize image content in full-text articles [C] //AMIA Annual Symposium Proceedings. Bethesda: American Medical Informatics Association, 2006: 834.

[22] AGARWAL S, YU H. FigSum: automatically generating structured text summaries for figures in biomedical literature [C] //AMIA Annual Symposium Proceedings. Bethesda: American Medical Informatics Association, 2009: 6.

[23] BHATIA S, MITRA P. Summarizing figures, tables, and algorithms in scientific publications to augment search results [J]. ACM Transactions on Information Systems, 2012, 30 (1): 1-24.

[24] TAKESHIMA R, WATANBE T. The Extraction of Figure-Related Sentences to Effectively Understand Figures [M] // KACPRZYK J.Innovations in Intelligent Machines-2. Berlin: Springer Berlin Heidelberg, 2012: 19-31.

[25] PARK G, RAYZ J T, POUCHARD L. Figure descriptive text extraction using ontological representation [C] //The Thirty-Third International Flairs Conference. Palo Alto: AAAI Press, 2020.

[26] SAINI N, SAHA S, BHATTACHARYYA P, et al. Textual entailment-based figure summarization for biomedical articles [J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2020, 16 (1s): 1-24.

[27] POLEPALLI R B, SETHI R J, YU H. Figure-associated text summarization and evaluation [J]. PloS One, 2015, 10 (2): e0115671.

[28] LOPEZ P. GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications [C] //International Conference on Theory and Practice of Digital Libraries. Berli, Springer, 2009: 473-474.

[29] BIRD S, LOPER E. Nltk: The natural language toolkit [C] // Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions. 2004: 31.

作者简介

黄佳妮，女，1999年生，硕士研究生，研究方向：文本挖掘。
于丰畅，男，1990年生，博士后，通信作者，研究方向：信息抽取、机器学习，E-mail：yufc2002@whu.edu.cn。

Automatic Recognition of Table-related Text in Literature Based on Table Retrieval and Machine Learning Two-stage Method

HUANG JiaNi YU FengChang
(School of Information Management, Wuhan University, Wuhan 430072, P. R. China)

Abstract: The tables in academic literature concisely represent the core knowledge in the literature in a structured form. Numerous academic search engines have integrated tables into retrieval results, which may help researchers quickly grasp the core knowledge and improve the research efficiency. However, while solely displaying the table without offering related information about it, readers frequently fail to fully understand the table's content, hindering further improvement of literature reading efficiency. We propose a two-stage table-related text recognition method based on machine learning and table retrieval. Stage 1 uses the table content to perform a full-text retrieval, and the retrieval results are regarded as the text potentially related to the table. Stage 2 builds a machine learning model to determine the correlation between the table and potentially relevant text, thereby realizing the automatic recognition of relevant text in the literature. This study utilizes the dataset from the Text Retrieval Conference as an example to verify the effectiveness of the method proposed in this paper. This method can easily extract text related to tables in the literature, which can provide a reference for the existing research on extractive summary of scientific tables and it is of great practical significance for improving the efficiency of literature research.

Keywords: Scientific Table; Table Understanding; Machine Learning

（收稿日期：2022-11-06）