情报学报 ISSN 1000-0135
第35卷第5期530-538，2016年5月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000-0135
Vol. 35 No. 5, 530-538 May 2016

doi:10.3772/j.issn.1000-0135.2016.005.009

学术文本的结构功能识别——基于段落的识别[1)]

黄 永 陆 伟 程齐凯 桂思思

（武汉大学信息管理学院，信息检索与知识挖掘研究所，武汉 430072）

摘要 学术文本的结构功能识别是学术文本章节层次的文本分类问题，其本质就是识别章节的结构功能。本文将基于段落的学术文本结构功能识别分为两个子问题：段落位置识别及基于段落投票的章节结构功能识别。在自动构建的大规模数据集上的实验结果表明，虽然基于段落的结构功能识别效果不如基于章节整体内容的识别，但仍然取得了不错的效果。本文结合实验结果着重分析了影响基于段落的识别效果的两个重要因素：段落长度及章节中段落数量，并在最后对学术文本结构功能识别的三个层次做了总结，指出了拟进一步探讨的问题和方向。

关键词 结构功能 文本分类 文本挖掘

The Structure Function Recognition of Academic Text ——Paragraph-based Recognition

Huang Yong, Lu Wei, Cheng Qikai and Gui Sisi

(*School of Information Management, Wuhan University, Wuhan* 430072)

Abstract The structure function recognition of academic text is a text categorization problem on section level, of which essence is to recognize the structure function of sections. In this paper, we have divide the paragraph-based recognition into two subtasks: the recognition of paragraph position and the structure function recognition based on majority voting by paragraphs in sections. Experiments were conducted on datasets constructed automatically. Though the results were not as good as the recognition based on section content, it proved that it is feasible to recognize structure function based on paragraph. Also we analyzed the reasons from the aspects of the length of paragraph and the number of paragraphs in sections. Finally, we summarized the research works of structure function recognition briefly and some potential application are recommended.

Keywords structure function, text categorization, text mining

1 引 言

学术文本中存在不同层次的结构单元，如全文、章节、段落等，不同层次结构单元反应不同内容，如学术文献全文可以反映文献主题思想，章节给出文献的方法或者结论，段落给出论据或论证等[1]。学术文本的结构功能是对学术文献的结构和章节功能

收稿日期：2015年11月10日

作者简介：**黄永**，男，1991年生，武汉大学信息管理学院，博士研究生，主要研究方向：信息检索、数据挖掘，E-mail：huangyng@gmail.com。**陆伟**，男，1974年生，武汉大学信息管理学院，博士，副院长，教授，主要研究方向：信息检索、知识管理、数据挖掘等，E-mail：weilu@whu.edu.cn。**程齐凯**，男，1989年生，武汉大学信息管理学院，博士研究生，主要研究方向：信息检索、数据挖掘，E-mail：chengqikai0806@gmail.com。**桂思思**，女，1992年生，武汉大学信息管理学院，博士研究生，主要研究方向：用户兴趣、查询专指度、信息检索，E-mail：sgui0229@whu.edu.cn。

1） 本文系国家自然科学基金面上项目“面向词汇功能的学术文本语义识别与知识图谱构建”（项目编号：71473183）；教育部人文社会科学基地重大项目“面向细粒度的网络信息检索模型及框架构建研究”（项目编号：10JJD630014）的研究成果之一。

的概括，包括“引言”、“相关研究”、“方法”、“实验”、“结论”五类标签，其自动识别是章节层次的分类问题，根据不同的结构单元，其自动识别主要分为三个层次：基于章节标题、基于章节内容、基于段落[2]。笔者曾在章节标题的层次使用序列标注的思想完成结构功能的自动识别，但是该方法存在局限性，如数据集构建困难、对于含有未登录词的章节识别效果较差等；基于章节内容层次的结构功能识别克服了上述缺陷，它使用文本分类方法进行学术文本功能识别，并取得了令人满意的实验效果[3]。本文则是探讨学术文本结构功能识别的第三个层次：基于段落的识别，并将其分为以下两个子问题：

（1）段落位置识别：针对给定段落，判断属于何种结构功能的章节。例如给定如下一段文字：

“本文后续结构如下：第二部分对相关研究进行了调研，第三部分对所提出的方法进行阐述，第四部分对数据集的构建、实验的设计以及最终的实验结果进行了论述，最后总结工作，并对下一步的研究工作做出展望。”

段落位置识别的目的就是识别出其应该属于具有“引言”功能的章节。

（2）基于段落投票的章节结构功能识别：使用投票法，综合某章节内所有段落的位置识别结果来识别该章节的结构功能。

本文在自动构建的数据集上进行了实验研究，着重从段落长度、章节中段落数量两个方面对识别效果的影响进行了分析。最终的实验结果表明，基于段落的学术文本结构功能识别效果虽然不如基于章节内容的学术文本结构识别，但仍然取得了不错的效果，整体上是可行的。

本文之后的内容安排如下：在第二部分将语言学理论及结构功能应用的相关研究做简单梳理；第三部分给出了本文提出的基于段落识别的总体框架及实验流程；在第四部分通过实验对比分析了基于段落与基于章节的差异；最后一部分对本文进行总结并且指出了结构功能识别的后续应用。

2 相关研究

本部分主要从学术文本结构功能的语言学相关理论以及其相关应用两方面进行调研。学术文本结构功能是对学术文本的内部结构在章节层次的描述与概括，相较于语言学篇章体裁分析中经典的IMRD[4]模型，结构功能加入了“相关研究”，从而能够更加完整的描述研究性论文结构和章节功能。语言学中语步分析（move analysis）[4]则是相对应于段落层次的研究，其主要内容是在论文结构 IMRD 下，将章节内容分成多个语步（move），从而形成论文整体结构的概括。最初语步分析主要应用于论文写作的指导中，随后用于不同领域论文的结构分析。在不同的领域具有不同的分析模式[5]，并且在多个领域如计算机[6]、医学[7]等都有实证研究。例如文献[7]对于不同的章节分为多个语步（图 1），并且对语步在论文中的分布做了相关的统计分析。在他们的研究中，无论是结构还是语步的分析研究一般是在少量文本（10～50 篇以内）上进行，使用人工识别方法并没有通过自动学习的手段进行大规模的实验，更多的应用于语言学的调研与论文写作的指导当中。

通过上述分析可以看出，段落相互独立而又相互衔接的构成一个章节，同一结构功能章节中的段落具有不同的作用，不同结构功能章节中段落功能构成不同。因此通过文本分类方法判断段落属于何种功能的章节，根据段落识别断章节功能是可行的，并且语言学中语步分析对于对学术文本的理解研究也极具启发意义，如何更加细粒度的识别学术文本语义单元将是未来研究的重要方向。

Move	Discourse function	
1:	Presenting Background Information	The Introduction Section
2:	Reviewing Related Research	
3:	Presenting New Research	
4:	Describing Data Collection Procedure	The Methods Section
5:	Describing Experimental Procedure	
6:	Describing Data-Analysis Procedure	
7:	Indicating Consistent Observations	The Results Section
8:	Indicating Non-Consistent Observations	
9:	Highlighting Overall Research Outcome	The Discussion Section
10:	Explaining Specific Research Outcomes	
11:	Stating Research Conclusions	

图 1 语步分析实例

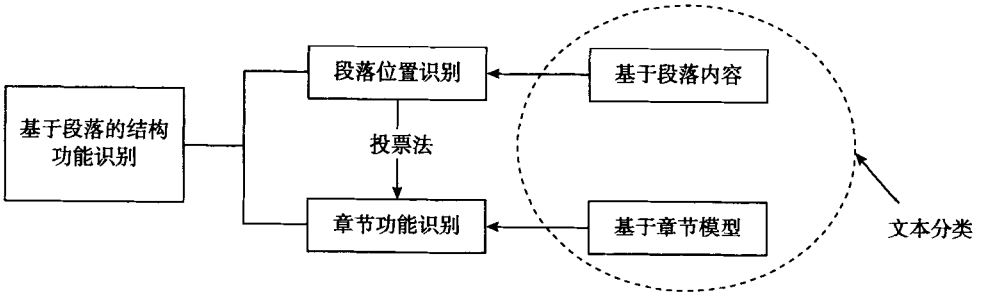

图2 基于段落的学术文本结构功能识别框架图

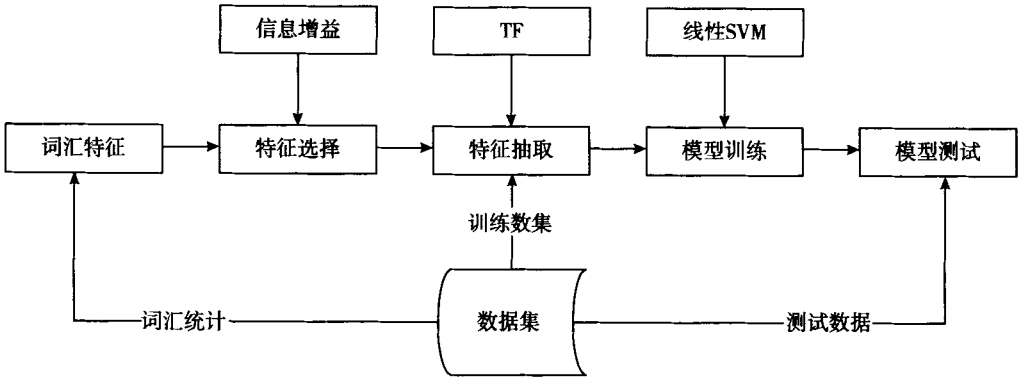

图3 基于段落内容的文本分类的流程图

除语言学领域之外,在计算机科学领域,也有使用学术文本的内部结构进行作弊论文的自动生成研究例如 scigen①。其主要原理是从不同结构功能的章节中抽取通用词汇或者通用句子构成论文主体框架,而论文的主题词汇或者方法则是随机抽取生成。在图书科学领域,文章内部结构最近逐渐获得重视被用于文献计量[8,9]、引文评价[10]、引文推荐[11,12]、引文功能识别[13]等任务当中。

3 框架及方法

本文将基于段落的学术文本结构功能识别分为两个子问题:段落位置识别、基于段落投票的章节结构功能识别(图2)。其中段落位置识别主要使用基于段落内容的识别方法,章节功能识别可以使用两种方法即投票法和基于章节模型[3]两种方法。基于段落内容和基于章节的模型都是一种文本分类问题。

3.1 段落位置识别

段落位置识别是根据段落内容识别出其应属于具有何种功能的章节。本文将之转换为文本分类问题,可有两种处理方式:①基于章节的分类模型[3],即将段落认为是一个独立的章节,使用基于章节内容的数据训练得到的模型识别其结构功能,将之作为段落位置识别结果;②基于段落内容的分类模型,即基于段落内容以其所属章节的结构功能为标签,将之转化为文本分类问题,使用文本分类方法(图3)进行模型训练和预测,其具体步骤如下:

(1)数据准备:根据文章的章节标题可自动构建学术文本结构功能的训练数据集[3]。章节标题可以认为是作者对于章节结构功能的标注,因此本文使用以下标题(表1)对文献中的章节进行筛选,并将筛选得到的章节标注为对应的结构功能。

表1 使用章节标题筛选结构功能

结构功能	对应的章节标题
引言	introduction
相关研究	related work, literature review, background
方法	method, methodology, model
实验	experiment, result, data
结论	conclusion, conclusion and discussion, discussion

本文对 ScienceDirect 中 2000 ~ 2013 年的计算

① https://pdos.csail.mit.edu/archive/scigen/

机领域128本期刊26万篇论文全文按照上述规则对论文中的段落进行抽取，构建了三个数据集：①段落位置数据集：分别从每种结构功能的段落中抽取40 000个段落构建一个大小为20万的平衡数据集，该数据集作为段落位置识别数据集；②章节功能数据集：每种结构功能的章节随机抽取5000个，构建一个大小为25 000的章节功能数据集；③基于段落的章节功能数据集：以段落为单位并与数据集②中章节一一对应，用于研究基于段落的章节结构功能识别的效果。

(2)特征选择：统计数据集中所有的词汇，使用信息增益[14]方法进行词汇特征选择，然后选择信息增益值最大的N个词汇作为词汇特征。N从区间[1000,42000]中以1000为间隔进行取值，根据分类效果确定N值。

(3)特征抽取：对于信息增益值最大的N个词汇，在训练数据集中抽取词汇频率作为特征值。

(4)模型训练：由于文本分类的特征维度高，训练数据大，线性SVM可以克服该缺点，且分类效果与其他核函数的SVM相差不大，所以本文使用LIBLINEAR作为分类器。LIBLINEAR是从由Lin等开发[15]的LIBSVM独立出来的用于文本分类的线性SVM工具。

(5)模型测试：在训练数据训练得到模型之后，在测试数据上进行测试。本文使用5折交叉检验进行模型的测试，使用准确率P(预测正确数/预测数)、召回率R(正确数/所有应该正确数)及$F1[2*P*R/(P+R)]$值作为评价指标。

3.2 基于段落投票的章节结构功能识别

基于段落的章节功能识别是根据章节中所有段落的位置识别结果，使用投票法进行章节功能的识别。本文使用以下几种投票法[16]进行基于段落的章节结构功能识别：

(1)计数法：对于整个章节中每个自然段对章节的结构功能进行投票，票数多的结构功能为该章节的结构功能。

(2)加权法：在计数法的基础上，根据段落的长度对每个自然段的长度进行加权，一个段落的长度越长其权重越大。本文使用log对段落的长度进行平滑处理。

(3)二次分类：对于章节内的段落投票不直接进行数值上的判断，而是以每一种结构功能作为一维特征，其所得票数的归一化值作为特征值，使用分类器进行二次分类。

上述三种方法前两种根据章节位置识别结果使用朴素的投票规则即票多者胜进行投票，第三种则是根据识别结果训练出新的投票模型，在之后的实验将会验证三种方法的效果。段落位置识别是基于段落投票的章节结构功能识别的基础，前者识别效果的好坏也直接影响到后者。

4 实 验

4.1 段落位置识别

从前文已知，段落位置识别可使用两种文本分类方法：基于章节的分类模型和基于段落内容的分类模型，因此本文分别使用上述两种方法做了两组对比实验：

(1)使用数据集①进行基于段落内容的文本分类。

(2)使用数据集②进行模型训练，将每一个段落认为是一个独立的章节，使用基于章节的分类模型对段落所属位置进行识别。

上述两组实验按照上一节文本分类的流程进行文本分类实验。在两组实验特征选择中，不同的信息增益最大N个词汇取得的分类效果如图4所示。

从图4可以看出，两组实验随着N值的增大，识别准确率在不断提升，但是在$N=3000$之后变化幅度较小，因此本文中，两组实验都是用信息增益最大的3000个词作为特征进行模型的训练，并且在数据集①上进行五折交叉检验，得到以下结果，如表2所示。

从表2可以看出，使用基于段落内容的分类模型识别效果比基于章节的分类模型效果好；从效果上来看，识别准确率、召回率均在65%左右。从各种结构功能来看，实验功能区的的效果最好，准确率达到70%，结论及相关研究的效果次之，引言效果最差。在引言功能中的段落的作用是提出问题，描述方法，给出结论等，与后文章节内容有一定的重复，导致在基于段落层次的分类，引言功能识别效果较差；方法、实验、结论与其他功能章节重复度较低，因此效果较好。

本文基于段落内容的分类模型的分类结果进行统计(表3)，表格的每一行表示一种结构功能被识别为五种功能的比例。可以从表中看出，引言功能的段落主要被错分为相关研究、结论，比例分别为

21.61%和17.66%；相关研究主要被错分为引言、方法、结论，比例分别为10.35%、7.88%和7.67%；方法主要被错分为实验，比例为15.22%；实验主要被错分为方法，比例为15.73%；结论主要被错分为引言、相关研究，比例分别为9.32%和8.14%。由上所基于段落内容的段落位置识别结果分析可知：引言、相关研究、结论三种结构功能更容易相互错分，实验及方法两者之间更容易错分。这与基于章节内容的识别结论一致[3]。

4.2 基于段落投票的章节结构功能识别

在上一节最优实验结果组的基础上，在数据集③上分别使用计数法、加权法以及二次分类法进行基于段落投票的章节结构功能识别实验（实验三至实验五），并以基于章节内容的结构功能识别[3]作为对比实验（实验六）。

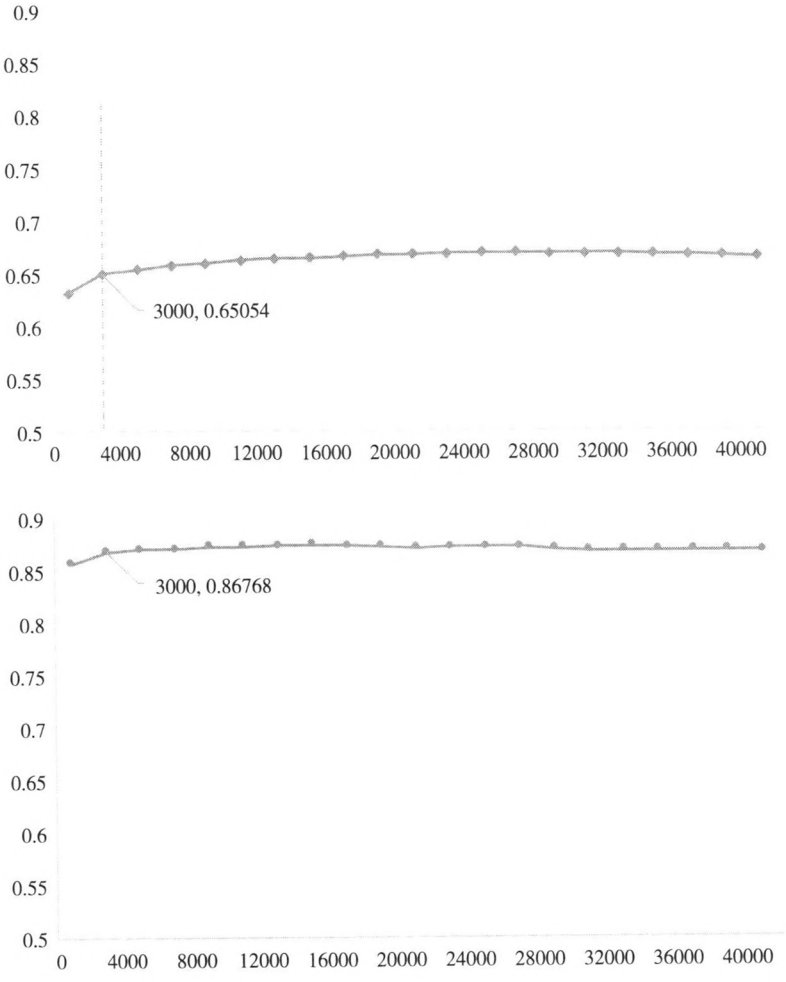

图4 特征选择中不同的N值的分类效果

表2 基于段落识别和基于章节模型识别的对比实验结果

	实验一			实验二		
	P	*R*	*F*1	*P*	*R*	*F*1
引言	**0.6145**	0.4765	**0.5367**	0.3606	**0.5758**	0.4435
相关研究	**0.6338**	**0.7074**	**0.6686**	0.5813	0.6075	0.5941
方法	0.6353	**0.6529**	**0.6440**	**0.6650**	0.5338	0.5922
实验	**0.7076**	**0.7029**	**0.7052**	0.6044	0.6636	0.6326
结论	0.6557	**0.7129**	**0.6831**	**0.7209**	0.5720	0.6379
整体	**0.6494**	**0.6505**	**0.6500**	0.5864	0.5905	0.5885

注：表中结果较好的指标已加黑

表3 段落位置识别错分表

	引言	相关研究	方法	实验	结论	总计
引言	47.65%	21.61%	9.36%	3.73%	17.66%	100%
相关研究	10.35%	70.74%	7.88%	3.36%	7.67%	100%
方法	6.78%	6.99%	65.29%	15.22%	5.71%	100%
实验	3.45%	4.13%	15.73%	70.29%	6.40%	100%
结论	9.32%	8.14%	4.52%	6.74%	71.30%	100%
总计	77.54%	111.61%	102.77%	99.35%	108.73%	500%

表4 三种方法实验结果

实验设置		引言	相关研究	方法	实验	结论	整体
实验三	*P*	0.7407	**0.8610**	0.6849	**0.9259**	**0.7940**	**0.8013**
	R	**0.6958**	0.6902	**0.9440**	0.8228	0.7876	**0.7881**
	*F*1	**0.7175**	0.7662	0.7938	**0.8713**	**0.7908**	**0.7947**
实验四	*P*	**0.7709**	0.8357	**0.7009**	0.8931	0.7454	0.7892
	R	0.5908	**0.7160**	0.9338	**0.8494**	**0.8152**	0.7810
	*F*1	0.6689	**0.7712**	**0.8008**	0.8707	0.7788	0.7851
实验五	*P*	0.6500	0.4977	0.7399	0.5657	0.8859	0.6678
	R	0.4528	0.9400	0.1138	0.8468	0.6770	0.6061
	*F*1	0.5338	0.6508	0.1973	0.6783	0.7675	0.6370
实验六	*P*	0.8138	0.8619	0.8500	0.9163	0.8990	0.8682
	R	0.8332	0.8622	0.8976	0.9046	0.8384	0.8672
	*F*1	0.8234	0.8620	0.8732	0.9104	0.8676	0.8677

注:已加粗较好的实验结果,不包括实验六

由表3中实验三至实验五可以看出,三种投票方法中计数法和加权法的识别效果好于二次分类法。计数法和加权法识别效果相近,但从总体来看,计数法效果更好。数据集中每种结构功能的章节所包含的段落个数较多(图5右),从而使得单纯使用计数法进行投票取得了较好识别效果,其准确率为80%,召回率为78.81%,*F*1值为79.47%。从识别准确率上来看,基于段落投票的章节结构功能识别法是整体上是可行的。但与实验六对比而言,识别效果仍然不如基于章节内容的学术文本结构功能识别[3],其主要原因还是受限于段落位置识别的效果。其中相关研究、方法、实验的召回率与准确率差异较大:方法的识别准确率只有68%而召回率为94%,召回率远高于准确率说明其他功能容易被误分为方法;相关研究的识别准确率86%而召回率为69%,实验识别准确率为92%,召回率82%,召回率低于准确率说明相关研究、实验容易被错分为其他功能。为进一步研究结构功能之间的错分关系,对实验三结果进行统计得到错分表如表5所示。

从整体来看,识别结果中方法比例为137%,分别由其他四种功能错分比例为9.38%、11.5%、15.20%、7.36%,而方法分为其他功能的比例较少,说明其他功能更容易错分为方法。从各结构功能来看,引言主要错分为相关研究、方法、结论;相关研究错分为引言和方法;实验主要错分为方法;结论错分为引言和方法。引言、结论更容易相互错分,其他功能错分为方法,这与基于章节内容的识别结论[3]稍有差异。

表5 实验三结果错分表

	引言	相关研究	方法	实验	结论	总计
引言	69.58%	8.04%	9.38%	0.70%	12.30%	100.00%
相关研究	12.80%	69.02%	11.50%	0.98%	5.70%	100.00%
方法	1.68%	0.52%	94.40%	2.50%	0.90%	100.00%
实验	0.80%	0.18%	15.20%	82.28%	1.54%	100.00%
结论	9.08%	2.40%	7.36%	2.40%	78.76%	100.00%
总计	93.94%	80.16%	137.84%	88.86%	99.20%	500.00%

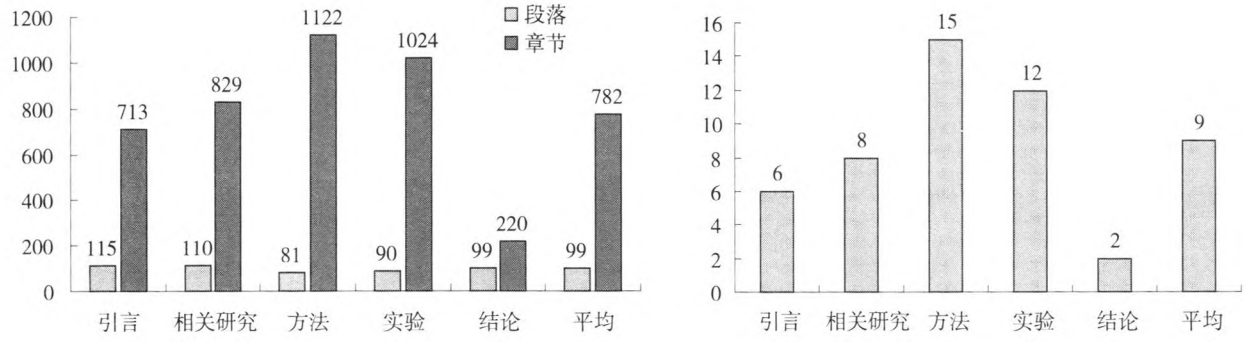

图5 数据集②和③中段落和章节的长度差异及各结构功能章节段落平均个数

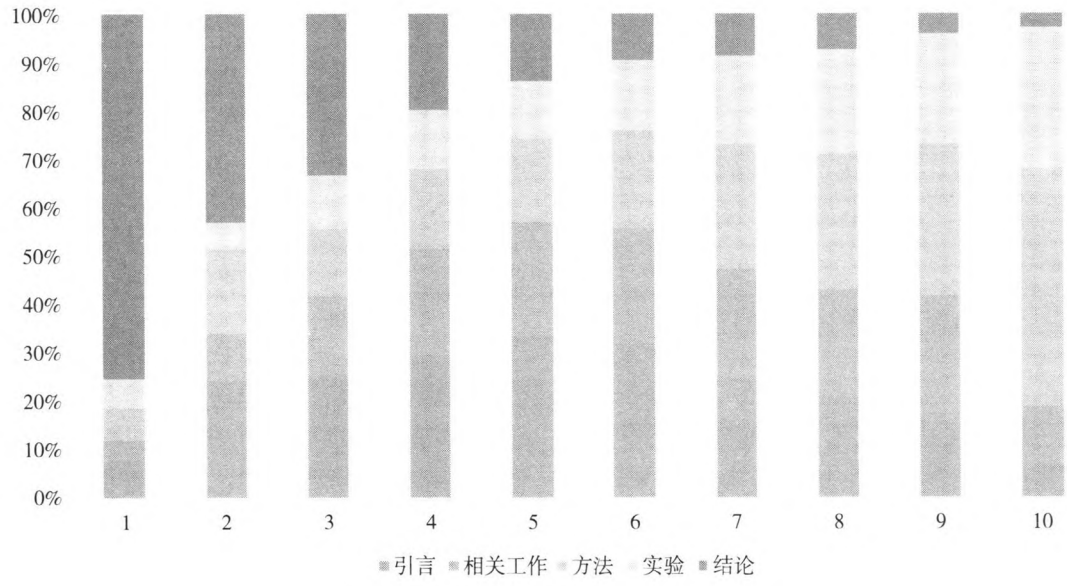

图6 不同结构功能章节中段落个数分布

4.3 讨论

从上述两小节可以看出,段落位置识别整体准确率为65%,其效果高于基于章节训练的模型;基于段落位置投票的章节结构功能识别法的准确率达到了80%,但是整体效果仍低于基于章节内容的结构功能识别。

基于段落位置投票的章节结构功能识别的效果依赖于段落位置识别效果,而段落的长度(包含词的个数)较短,能够提供的分类线索十分有限,从而导致段落位置识别效率较低。如图6所示,段落平均长度只有99,而章节平均长度为782,段落的平均长度大概是章节平均长度的1/8,一个章节平均包含9个段落。段落长度与章节长度具有较大差异,因此直接使用基于章节的分类模型进行段落位置识别效果较差。从第二部分可知,段落在同一结构功能的章节中的作用不同,且不同功能章节中的段落构成也不同。不同章节中的段落对于章节功能的作

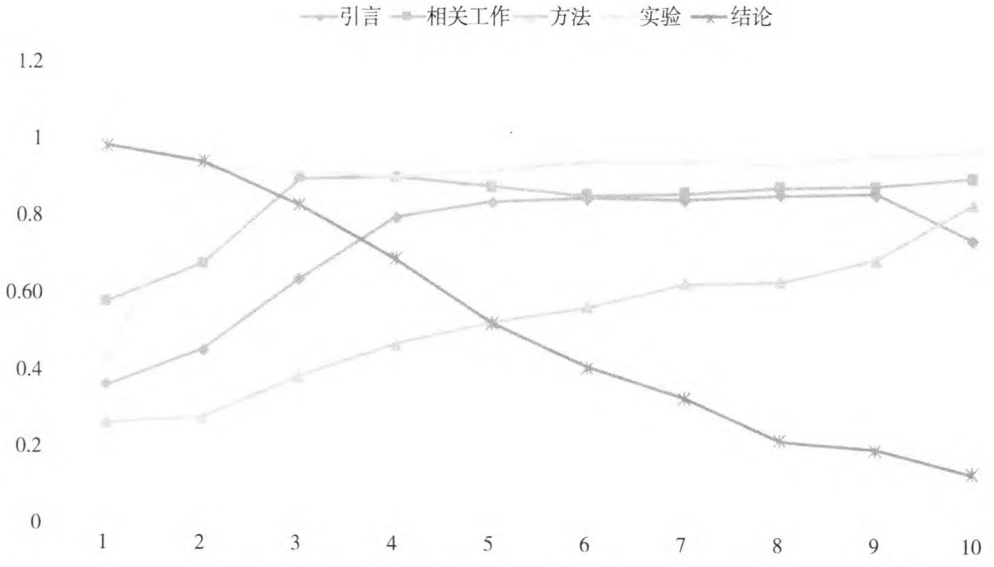

图 7　基于段落投票的结构功能识别准确率与段落个数折线图

用是相互独立的，仅仅使用词汇特征难以描述段落的作用，因此基于词汇特征的段落位置识别的效果并不是特别的理想。

同样基于段落投票的章节结构功能识别法因受限于段落位置识别效果，其学术文本功能识别效果不如基于章节内容的结构功能识别，其识别准确率受到章节中段落个数的影响。如图 7 给出数据集中各结构功能章节中段落个数的比例（如果章节段落个数超过 10 个按计算）。其中可以看出具有“结论”功能章节的数量随着其包含段落个数增多而不断降低，具有其他结构功能的章节中段落个数分布较为均匀；而从图 6 可以看出随着段落个数的增多，“总结”功能的章节识别准确率随之降低，而其他结构功能识别准确率则是不断提高。在“总结”功能的章节中段落长度较长，因此训练得到模型对于长段落识别准确率较高，也因此“总结”功能章节在段落数目比较少时准确率较高。其他功能的章节在各个段落长度分布比较均匀，从而使得训练得到的模型对于各种长度的段落识别准确率相近，在识别准确率相近的情况下，段落数量越多，基于段落投票的章节结构功能识别的准确率越高。

6　总　结

本文从段落层次对学术文本结构功能识别进行研究，研究可分为段落位置识别及基于段落投票的章节结构功能识别。在本文构建的数据集上的实验表明，段落位置识别准确率为 65%，基于段落投票的结构功能识别准确率为 80%，虽然效果不如基于章节内容的结构功能识别效果，但是本文所提方法仍然是可用的。

从学术文本的结构功能的提出到自动识别的三个层次，笔者希望通过对学术文本的不同层次的结构单元用于结构功能识别的探索，以期找到最有效的结构功能识别方法，并且使之能够应用于其他的学术文本挖掘研究中。从目前的研究来看，基于章节标题的识别方法更简单方便，但是容易受到数据集的影响；基于章节内容识别则在具有较高的准确率同时具有更强的适用性。

学术文本的结构功能是从更深层次或者偏重于语义的角度去理解学术文本的结构，可将不同的学术文本的结构框架进行统一，从而可应用于更多的学术文本挖掘研究中。学术文本结构功能在未来的应用包括：基于学术文本结构功能层次的学术文献的计量，例如基于结构功能的引文分布、加入结构功能的引文网络等，结构功能为文献计量提供了新的角度去理解文本；融入学术文本功能结构的学术文本检索；学术文本中词汇权重计算，一个词汇在不同结构功能章节中出现权重应该有所不同；除应用于上述研究问题，在移动互联网时代，结构功能框架在移动阅读中更具有应用价值，如学术文献的导航、推荐等。完成学术文本结构功能识别之后，如何进一步应用将是下一步工作的重点。

参　考　文　献

[1] Leydesdorff L. The Challenge of Scientometrics: The Development, Measurement, and Self-organization of Scientific Communications [M]. Universal-Publishers,

2001.

[2] 陆伟,黄永,程齐凯,等.学术文本的结构功能识别——功能框架及基于章节标题的识别[J].情报学报,2014(9):979-985.

[3] 黄永,陆伟,程齐凯,等.学术文本的结构功能识别——基于章节内容的识别[J].情报学报(待见刊).

[4] Pufahl I, Swales J M. Genre analysis: english in academic and research settings[J]. Language, 1993, 69.

[5] Zhang Lei. Grasping the structure of journal articles: utilizing the functions of information units[J]. Journal of the American Society for Information Science & Technology, 2012, 63(3):469-480.

[6] Posteguillo S. The schematic structure of computer science research articles[J]. English for Specific Purposes An International Journal of Esp, 1999, 18:139-160.

[7] Nwogu K N. The medical research paper: Structure and functions[J]. English for Specific Purposes, 1997, 16(2):119-138.

[8] Hu Z, Chen C, Liu Z. Where are citations located in the body of scientific articles? A study of the distributions of citation locations[J]. Journal of Informetrics, 2013, 7(4): 887-896.

[9] Ding Y, Liu X, Guo C, et al. The distribution of references across texts: Some implications for citation analysis[J]. Journal of Informetrics, 2013, 7(3): 583-592.

[10] Zhu X, Turney P, Lemire D, et al. Measuring academic influence: Not all citations are equal[J]. Journal of the Association for Information Science and Technology, 2014, doi: 10.1002/asi.23179.

[11] He Q, Pei J, Kifer D, et al. Context-aware citation recommendation [C]//Proceedings of the 19th international conference on World wide web. ACM, 2010:421-430.

[12] He J, Nie J Y, Lu Y, et al. Position-aligned translation model for citation recommendation [C]//String Processing and Information Retrieval. Springer Berlin Heidelberg, 2012: 251-263.

[13] Teufel S, Siddharthan A, Tidhar D. Automatic classification of citation function [C]//Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2006: 103-110.

[14] Yang Y, Pedersen J O. A comparative study on feature selection in text categorization[C]//ICML, 1997, 97: 412-420.

[15] Fan R E, Chang K W, Hsieh C J, et al. LIBLINEAR: A library for large linear classification[J]. The Journal of Machine Learning Research, 2008, 9: 1871-1874.

[16] Vens C. Majority Voting[M]. Springer New York, 2013.

(责任编辑 车 尧)