情报学报 ISSN 1000-0135
第35卷第4期425-431，2016年4月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000-0135
Vol. 35 No. 4, 425-431 April 2016

doi:10.3772/j.issn.1000-0135.2016.004.009

学术文本的结构功能识别——在学术搜索中的应用[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 在学术大数据环境下，学术文本挖掘研究向细粒度和语义化方向发展。学术文本的结构功能是对学术文本正文的结构及章节功能的概括。为探讨结构功能在学术搜索中的作用，本文将学术文本看作是个结构功能域的集合，使用域加权语言模型对学术文本结构功能进行加权，并以一般语言模型为基准，在INEX04数据上进行了文档级检索实验。实验结果表明本文所提出的模型取得了较大的提升，尤其在P@5上的相对提升达到13.93%。根据模型中各个结构功能域的权重参数分析可以得知，引言功能作用最大，相关研究、方法的作用次之，实验及结论的作用最小。本文的实验也证明了学术文本的结构功能在学术搜索中的应用价值。

关键词 结构功能 学术搜索 域加权 语言模型

The Structure Function Recognition of Academic Text ——Application in Academic Search

Huang Yong, Lu Wei, Cheng Qikai and Gui Sisi

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

Abstract In scholar big data environment, academic text mining is becoming more fine-grained and semantic. The structure function is the summary of the academic's structure and function of sections. From the main content respect, we use a field-weighted language model based on structure function, and conduct document level retrieval experiments on the datasets of INEX04. The results of experiments demonstrate that our model is more effective than the baseline model and a 13.93% relative improvement has gotten on the P@5. The parameters of fields in model show that the introduction is the most important function, and the related work, method follows, the experiment and conclusion are at last. The experiment of this paper also demonstrate the application value of structure function in academic search.

Keywords structure function, academic search, field weighted, language model

1 前 言

近年来，学术文本的结构化呈现越来越受到重视，学术数据库（如ScienceDirect、Wiley、Springer等）都将PDF全文数据转化为html格式，这使得格式化的学术文本全文数据更容易获得，也促进了学术文本挖掘的细粒度及语义化研究。学术文本的结构功能使用引言、相关研究、方法、实验、结论五个标签对学术文本的内部结构及章节功能进行了概括[1]。结构功能使得学术文本的内部结构更具语义化，但是如何有效地利用学术文本的结构功能进

收稿日期：2015年12月16日

作者简介：**黄永**，男，1991年生，博士研究生，主要研究方向：信息检索、数据挖掘。**陆伟**，男，1974年生，教授，博士生导师，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，博士研究生，主要研究方向：信息检索、数据挖掘。**桂思思**，女，1992年生，博士研究生，主要研究方向：用户兴趣、查询专制度、信息检索。

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）和武汉大学与中国科技信息研究所合作项目“科学文献的语义功能识别与深度利用”研究成果之一。

行检索也成为了一种新的挑战。以检索结果排序的角度来看,学术文本是由结构功能构成的结构化文档,对学术文本的评分也就是各结构功能得分的加权和。

关于结构化文本检索模型的研究,Robertson等[2]在针对一般文档的概率检索模型BM25[3]的基础上进行改进,提出了BM25的域加权版本BM25F模型。Lu等[4]对BM25F进行改造,提出了BM25E模型,该模型可用于XML检索,并在INEX04数据集上对学术文本的题目、摘要以及章节标题进行加权,取得了很好的实验效果。Ogilvie等[5]则是使用混合域语言模型[6]对学术文本中的题目、题注、段落等多种结构进行加权,并提升了检索效果。

本文则是从学术文本正文角度出发,使用域加权语言模型对学术文本中各结构功能进行加权,探讨不同结构功能在学术文本文档级检索中的作用。本文在INEX04数据[7]上的实验表明,相较于基准结果,融入结构功能的学术搜索在MAP、nDCG、P@5、P@10、P@20五个评测指标上均获得了较大提升。

本文第二部分主要阐述结构化文档检索的相关研究;第三部分具体描述本文所使用的方法,包括学术功能结构识别、基于结构功能域加权的语言模型以及参数优化;第四部分描述本文使用的数据以及实验结果;最后一部分总结全文所做工作,并给出下一步研究方向。

2 相关研究

本部分主要从学术文本内部结构在检索中的应用和域加权模型两个方面进行相关研究的调研和论述。

学术文本具有复杂的内部结构,如学术文本元数据(题目、摘要、关键词、参考文献等)以及学术文本内部元素(引用上下文、章节标题、题注等)。Lu等对学术文本题目、摘要、章节标题的域加权进行片段级[3]和文档级[8]检索,证明了上述三种结构的重要性;Ogilvie等探讨了文本中的题目、题注、段落、注释等多种元素的作用[5];相较于学术文本中固有结构,引文上下文可以看作是作者对被引文献的语义化总结。Robertson等将引用看作是学术文档间的继承[9],Ritchie等则将引文上下文加入到被引文献中进行索引和检索[10],并取得了不错的效果提升。还有一些研究[11]将外部工具生成的结构化摘要应用于结构化检索中。

现有域加权信息检索模型主要有BM25F[2]、混合域语言模型(Mixture of Field Language Model, MLFM)[6]、面向半结构化数据的概率检索模型(Probabilistic Retrieval Model for Semi-structured data, PRMS)[12]。BM25F是Robertson等在BM25的基础上进行修改并实现的域加权版本[2]。MLFM[6]是Ogilvie等通过文档内各个域语言模型的线性混合在结构化文档集合上进行检索实验。PRMS[12]由Kim和Croft提出,使用各文本域相对于查询词的条件概率代替了MLFM中的域权重参数,从而使得每一个查询词与每一个域之间都有一个权重参数。Croft等在文献[13]中使用相关反馈法对上述概率映射进行估计。BM25F、MLFM两种模型中查询共享域参数,而PRMS可以根据查询词的不同分别对文本域进行加权。

现有研究证明了对学术文本内部结构的有效利用能够提升检索效果,但是还没有研究对学术文本结构功能在学术搜索中的作用进行探索,因此本文将学术文本中的不同结构功能看作是学术文本的不同域(结构功能域),使用域加权语言模型对学术功能域进行加权,用以探讨学术文本结构功能在学术搜索中的作用。本文的下一部分将对学术文本的结构功能识别及使用的检索模型进行具体的阐述。

3 方法描述

3.1 学术文本结构功能及其自动识别

学术文本往往具有严谨的逻辑结构,从研究问题引入、研究背景介绍、解决方法的提出、验证到最终得出结论,各个章节都具有很强的目的性和功能性。文献[1]通过对计算机领域的研究性论文结构进行总结分析,将学术文本结构功能分为5种,即引言、相关研究、方法、实验、结论。这五种功能既反映了学术文本章节的逻辑功能和目的,又构成了学术文本的逻辑结构,这种结构功能使学术文本结构更加语义化。

如图1所示,一篇学术文本有多个章节构成,左边是文章中的章节,右边是章节对应的结构功能。结构功能就是将学术文本中对学术文本中章节的功能和目的的标注,其自动识别本质上是一种分类问题,根据研究对象的不同,其自动识别分为三个层次[1]。第一,基于章节标题的结构功能识别方法:

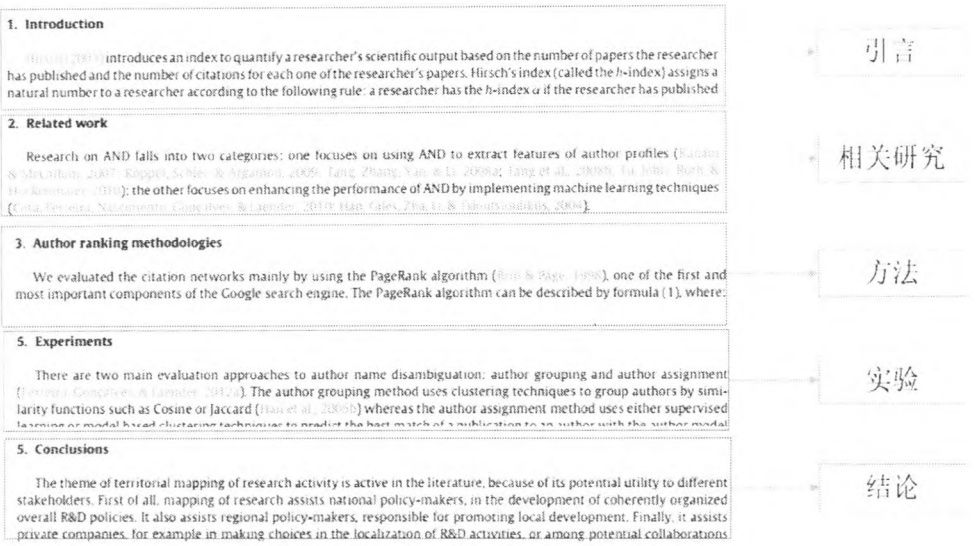

图 1　学术文本中的章节及对应结构功能

在只有学术文本章节标题的情况下，根据章节标题给出其对应的结构功能，以序列标注的思想使用条件随机场模型进行识别[1]；第二，基于章节全部内容的结构功能识别方法，使用自然标注的方法构建大规模数据集，将词汇特征与深度学习特征结合，使用文本分类方法进行结构功能识别[14]；第三，基于章节中段落内容的结构功能识别方法，探讨段落对于结构功能识别的作用[15]。

在自动识别的三个层次中，基于章节标题的识别方法简单易行，且对内容结构比较固定的学术文本识别准确率较高。本文的实验数据来自同一领域，该领域内论文结构比较固定，因此本文使用基于章节标题的结构功能识别方法对数据集中的学术文本进行识别。

3.2　结构功能域加权语言模型

首先对本文中使用的符号进行说明：假设一个查询由 m 个词构成，表示为 $Q=(q_1,q_2,\cdots,q_m)$；假设文档集合 C 中一共存在 n 个结构功能域，表示为 $S=(S_1,S_2,\cdots,S_n)$，那么文档集合中的每一个文档 D 是由 $(S_1,S_2,\cdots,S_n)$ 中的结构功能域构成，同时将每一个域的权重表示 $(w_1,w_2,\cdots,w_n)$。

语言模型[16]是目前最常用的信息检索模型之一。它通过文档 D 生成查询 $Q=(q_1,q_2,\cdots,q_m)$ 的概率 $P(Q|D)$ 对文档进行评分，生成概率是查询中每个词在文档中出现的概率乘积，其计算公式如下：

$$P(Q\mid D)=\prod_{i=1}^{m}P(q_i\mid D) \tag{1}$$

在对学术文本的结构功能进行识别后，学术文本 D 则可以看作是由一个或者多个结构功能域 $S=(S_1,S_2,\cdots,S_n)$ 组成的结构化文档。当学术文本表示成为由结构功能域组成的结构化文本之后，需要在传统的语言模型中加入结构功能域与查询之间的概率映射，也即是

$$P(S_j|q_i,C)=\frac{P(q_i|S_j,C)P(S_j|C)}{P(q_i|C)} \tag{2}$$

$$P(q_i\mid C)=\sum_{S_k\in S}P(q_i\mid S_k,C)P(S_k\mid C) \tag{3}$$

$$P(S_j\mid q_i,C)=\frac{P(q_i\mid S_j,C)P(S_j\mid C)}{\sum_{S_k\in S}P(q_i\mid S_k,C)P(S_k\mid C)} \tag{4}$$

在上述公式中，C 表示整个数据集，q_i 表示查询 Q 中的第 i 个词。如公式(6)所示，查询词 q_i 与结构功能域 S_j 之间的概率映射 $P(S_j|q_i,C)$ 等于查询 q_i 在结构功能域 S_j 中出现的概率与查询词 q_i 在整个数据集中出现的概率 $P(q_i|C)$ 的比值。$P(q_i|C)$ 可以通过查询词 q_i 在各结构功能域中出现的概率之和进行计算[公式(7)]。$P(S_k|C)$ 表示结构功能域 S_k 在数据集中的先验概率。综合公式(6)和公式(7)，可得到公式(8)，用以计算查询词及结构功能域间的概率。

加入结构功能域与查询之间的概率映射之后的语言模型可以表示成：

$$P(Q\mid D)=\prod_{i=1}^{m}\sum_{j=1}^{n}P(S_j\mid q_i,C)P(q_i\mid S_j,D) \tag{5}$$

其中，$P(q_i|S_j,D)$ 是在结构功能域上使用特定平滑方法后构建的语言模型，在本文中平滑方法使用的是狄利克雷平滑[17,18]。$P(S_j|q_i,C)$ 是上面所述的

在整个集合中查询词与结构功能域的概率映射，即文档中各个结构功能相对于查询词的权重。

3.3 参数估计

由公式(8)和公式(9)可以看出，根据查询所包含词汇的不同，每一个查询对结构功能域的加权也就不同，因此增加了模型对词汇与结构功能域之间的表现力的同时无疑也增加了参数训练的难度。在本文的参数训练中，对上述模型的 $P(S_j|q_i,C)$ 进行简化，假设所有的查询词共享结构功能域参数，模型也就可以表示成为

$$P(Q|D) = \prod_{i=1}^{m}\sum_{j=1}^{n} w_j P(q_i|S_j,D) \quad (6)$$

文档得分是查询 Q 中每一个词 q_i 在各学术结构功能域中概率加权和的乘积。其中结构功能域的加权参数满足以下条件：

$$\sum_{j=1}^{n} w_j = 1(w_j \in [0,1]) \quad (7)$$

也即学术文本正文中不同的结构功能域的权重和为1。

不同领域的学术文本，其结构功能构成有所不同，但都可以通过上述模型进行结构功能域加权。

语言模型中的参数选择可以通过一些启发式方法实现[3]。本文使用 GridSearch 方法对参数 w_j 进行估计：对每一个参数在区间[0,1]中每隔0.05进行取值，并且需要保证所有的参数和为1，将得到的每一组的参数代入本文提出的语言模型中，通过五折交叉检验的检索效果评估出最优的参数。

4 实验及结果

4.1 数据

本文使用的是 XML 检索会议 INEX 在2004年 ad-hoc 检索所使用的数据集 inex-1.4，主要包括了 IEEE Computer Society publications 从1995~2002年的12 107篇文档，涵盖74个主题，其中 CO (Content Only) 主题40个，CAS (Content and Structure) 主题34个，一般包括主题编号、类型、主题(title)、描述及关键词等信息。CO 和 CAS 主题的区别在于后者对检索主题进行了结构限制，由于本文只探讨文档级别的 XML 检索，因此只采用 CO 主题。与正式的 INEX 要求一样，只利用主题(title)项作为查询词进行检索，且本文不考虑查询词内部逻辑操作符。对于每一个查询主题，都有一个相关结果集，该结果集是由参与者人工筛选得来，每一条相关结果都会给出其相关的内部元素，每篇文档的相关元素的相关性是由两个属性评定分别是 exhaustiveness 和 specificity，具体详见[7]。两个属性值都大于0表示该元素相关，由于本文需要进行文档级检索实验，因此认为包含相关元素的文档为相关文档。本文的实验以学术文本的正文为检索对象，因此将文档中的元数据、参考文献等内容去除，只保留正文内容。

4.2 结构功能识别

根据 INEX 文档的内部结构，本文抽取每一篇学术文本的正文，并且将正文根据 sec 标签分为不同的章节，并从不同的章节中根据标签 st 抽取一级标题。根据正文中的章节标题，本文使用文献[1]中所提出的基于章节标题的结构功能识别方法对章节的结构功能进行识别。该方法是一种基于序列标注思想的识别方法，简单方便，并且具有较高的准确率。

4.3 数据索引

对 INEX 数据中的正文进行结构功能识别之后，文档中的章节标签 sec 被章节结构功能标签所替代。本文使用 Indri① 对数据进行索引，使用 smart_stopwords② 作为停用词表，词干提取方法设置为 porter stemmer[19]，并且添加五个结构功能域。

4.4 参数的训练

本模型涉及两种参数：五个结构功能域的权重参数以及语言模型的参数。

对于五个结构功能域的权重参数，本文按照上文所述的权重参数计算方法得到1万多组参数，对每一组参数在 INEX04 的主题上进行五折交叉检验，选择最优的结果作为最优参数。语言模型中使用狄利克雷平滑，其只有一个参数 mu，考虑到实验时间消耗，本文不对 mu 进行优化，在所有实验中将 mu 设置为默认参数2500。

4.5 查询构建

由于本文使用的是基于域加权的语言模型，因

① http://www.lemurproject.org/indri/

② http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop

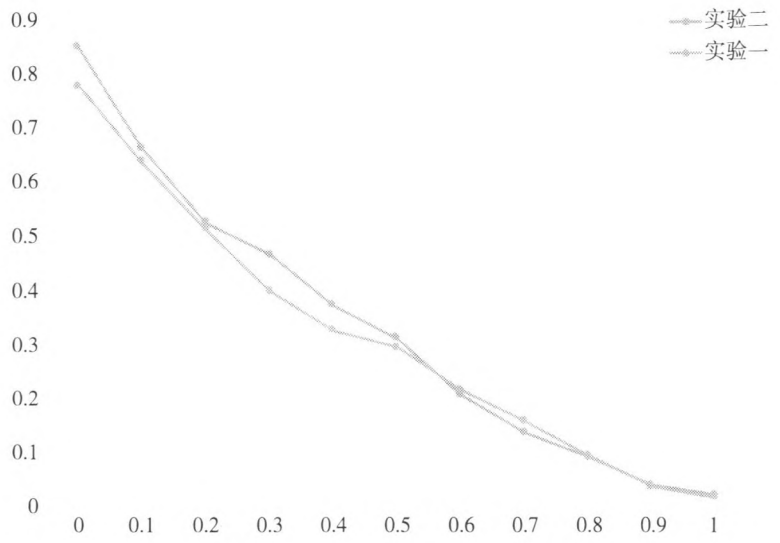

图 2　两种模型随着召回率的查准率曲线图

此不能简单使用 Indri 的查询语言[20]#combine，所以本文对于每一个查询使用#wsum 语言根据域加权参数构建查询句。例如检索词为 information retrieval，对于一般语言模型而言，Indri 查询为 #combine (information retrieval)，而对于域加权模型而言，假设五个不同的域的权重参数分别为(0.4,0.2,0.2,0.1,0.1)，那么其 Indri 的查询为：

```
#combine(
#wsum(0.4 information.(intro)
        0.2 information(.rw)
        0.2 information.(.md)
        0.1 inforamtion.(rlt)
        0.1 information.(cncl))
#wsum(0.4 retrieval.(intro)
        0.2 retrieval(.rw)
        0.2 retrieval.(.md)
        0.1 retrieval.(rlt)
        0.1 retrieval.(cncl))
)
```

其中位于查询词前面的小数表示该域的权重，在查询词和点号“.”后括号内的则是五种结构功能域名，上述只是作为查询构建的示例，本文使用的是 INEX04 数据集的 40 个主题作为查询。

4.6　实验设置

本文以学术文本正文的结构功能为研究对象，探讨学术文本结构功能对于检索的作用，因此本文设置两组实验：

(1)语言模型(狄利克雷平滑)(基准实验)。

(2)结构功能域加权语言模型。

在上述两组实验在 INEX04 的 40 个查询主题上进行五折交叉检验，将每折最优结果取平均作为各组的实验结果。

4.7　实验结果

本文使用 MAP、nDCG、P@5、P@10、P@20 五个评测指标对两组数据进行比较，得到以下结果如表 1 所示。

表 1　两组实验对比试验结果

	MAP	nDCG	P@5	P@10	P@20
语言模型	0.2838	0.5726	0.4389	0.4111	0.3324
结构功域加权	0.2980	0.5966	0.5000	0.4345	0.3633
相对提升	5.02%	4.21%	13.93%	5.68%	9.31%

从表 1 中可以看出，相对与传统语言模型，本文提出结构功能域加权模型在各项指标中均有提升，其中 P@5 提升最大，达到 13.93%。因此，学术文本的结构功能能够提升学术搜索的准确率。上述是两组实验中针对查准率的结果统计，本文又对两组实验的其召回率与准确率关系图进行统计，如下图 2 所示。

从图 2 中，可以看出本文所提出的模型在召回率在 50% 之前都有较为明显的提升，说明本模型在总体上也具有较好的效果。

为进一步探讨各种结构功能的作用，本文将五折交叉检验中每一折最优的结构功能域参数进行统计得到表 2：

表2 五折交叉检验中格结构功能域的最优权重参数

folder	引言	相关研究	方法	实验	结论
1	0.25	0.2	0.25	0.2	0.1
2	0.3	0.15	0.25	0.15	0.15
3	0.3	0.25	0.2	0.1	0.15
4	0.3	0.2	0.25	0.15	0.1
5	0.3	0.25	0.2	0.1	0.15
avg	0.29	0.21	0.23	0.14	0.13

从表2中可以看出，五种结构功能中权重最高的是引言功能，其平均权重为0.29，相关研究、方法权重相近，平均权重分别为0.21和0.23，实验及结论的权重相近，平均权重分别为0.14和0.13。因此可以看出学术文本的检索中，不同的结构功能的贡献有所不同，其中引言部分作用最大，相关研究及方法部分作用次之，实验及结论部分作用最小。

5 结 论

本文为探讨不同结构功能在学术搜索中的作用，将学术文本看作是个结构功能域的集合，使用域加权语言模型对学术文本中结构功能域进行加权，并在INEX04数据的正文上进行文档级检索实验。实验表明，融入结构功能的语言模型相较于一般语言模型在MAP、nDCG、P@5、P@10、P@20等五个评测指标上均有较大的提升，并且根据模型中各结构功能域的加权参数值可以发现：在学术文本的正文检索中，引言功能最为重要，相关研究、方法作用次之，实验及结论的作用相对较小。

本文提出的基于结构功能域加权的语言模型取得了较好的实验效果，并且对不同的结构功能的作用做了初步探索，但仍有以下工作需要进一步研究：本文为了简化参数优化过程，假设所有查询共享结构功能域参数，这虽然在一定程度上解决了参数优化问题，但仍需要一种参数优化方法对公式(9)中的参数进行估计；本文证明了学术文本的结构功能在学术搜索中具有一定的应用价值，但其具体应用场景仍需进一步挖掘。

参 考 文 献

[1] 陆伟，黄永，程齐凯. 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，2014，33(9)：979-985.

[2] Robertson S，Zaragoza H，Taylor M. Simple BM25 extension to multiple weighted fields [C]//Proceedings of the thirteenth ACM international conference on Information and knowledge management. ACM，2004：42-49.

[3] Robertson S，Zaragoza H. The probabilistic relevance framework：BM25 and beyond[M]. Now Publishers Inc，2009.

[4] Lu W，Robertson S，MacFarlane A. Field-weighted XML retrieval based on BM25[M]. Springer Berlin Heidelberg，2006.

[5] Ogilvie P. Retrieval using Document Structure and Annotations [D]. Carnegie Mellon University，2010.

[6] Ogilvie P，Callan J. Combining document representations for known-item search [C]//Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. ACM，2003：143-150.

[7] Malik S，Lalmas M，Fuhr N. Overview of INEX 2004 [M]. Springer Berlin Heidelberg，2005.

[8] 陆伟. 基于域加权词频法的XML文档级检索实现与评价[J]. 中国图书馆学报，2007，32(6)：57-60.

[9] Robertson S，Lu W，MacFarlane A. XML-structured documents：retrievable units and inheritance [M]// Flexible Query Answering Systems. Springer Berlin Heidelberg，2006：121-132.

[10] Ritchie A，Robertson S，Teufel S. Comparing citation contexts for information retrieval [C]//Proceedings of the 17th ACM conference on Information and knowledge management. ACM，2008：213-222.

[11] Ali M S，Consens M P，Gu X，et al. Efficient，effective and flexible XML retrieval using summaries [C]// INEX. 2006：89-103.

[12] Kim J，Xue X，Croft W B. A probabilistic retrieval model for semi-structured data [M]//Advances in Information Retrieval. Springer Berlin Heidelberg，2009：228-239.

[13] Kim J Y，Croft W B. A field relevance model for structured document retrieval [M]//Advances in Information Retrieval. Springer Berlin Heidelberg，2012：97-108.

[14] 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，2016 35(3)：293-300.

[15] 黄永，陆伟，程齐凯，等. 学术文本的结构功能识别——基于段落的识别[J]. 情报学报(待见刊).

[16] Ponte J M，Croft W B. A language modeling approach to information retrieval [C]//Proceedings of the 21st annual international ACM SIGIR conference on

Research and development in information retrieval. ACM, 1998: 275-281.

[17] Zhai C X, Lafferty J. Two-stage language models for information retrieval [C]//Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2002: 49-56.

[18] Zhai C, Lafferty J. A study of smoothing methods for language models applied to ad hoc information retrieval [C]//Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2001: 334-342.

[19] Porter M F. Snowball: A language for stemming algorithms. October 2001[J]. Retrieved March, 2001, 1: 2014.

[20] Strohman T, Metzler D, Turtle H, et al. Indri: A language model-based search engine for complex queries [C]//Proceedings of the International Conference on Intelligent Analysis. 2005, 2(6): 2-6.

(责任编辑　刘志辉)