情报学报 2017年6月 第36卷 第6期
Journal of the China Society for Scientific and Technical Information, June 2017, 36(6): 599-605

学术文本的结构功能识别——在关键词自动抽取中的应用

方 龙，李 信，黄 永，陆 伟

（武汉大学信息管理学院，信息检索与知识挖掘实验所，武汉 430072）

摘 要 当前的关键词自动提取研究大多基于候选词的词频、文档频率等统计信息，往往忽略了候选词所在的学术文本的内在结构，导致关键词提取的效果不佳。本文将学术文本看作是5个结构功能域的集合，提出了融合学术文本结构功能特征的多特征组合提取方法，并利用学术文本的章节标题对其结构功能进行识别，然后通过SVM二分类和LambdaMART学习排序算法分别在计算机语言学领域的文献集上进行了实现。实验结果表明，本文提出的组合特征方法相比基准特征在关键词提取的效果上取得了较大的提升，尤其在分类实验中准确率的相对提升上达到10.75%，证明了学术文本结构功能特征在关键词自动提取上的重要性。

关键词 结构功能；关键词提取；学术文本；支持向量机；学习排序

Structure-Function Recognition of Academic Text —Application in Automatic Keywords Extraction

Fang Long, Li Xin, Huang Yong and Lu Wei

(*School of Information Management, Information Retrieval and Knowledge Mining Laboratory, Wuhan University, Wuhan* 430072)

Abstract: The current approach for automatic extraction of keywords is mostly based on the frequency and documentation frequency of the alternative words, which are ignored in the inner structure of academic text and leads to bad work of keywords extraction. In this article, we considered an academic text as a collection of five structure-function models, proposed a mixed feature method with academic text structure features, and implemented it in the computer science documents by a classification technique called support vector machine (SVM) and a ranking model named LamdbaMART. The results show that the methods we put forward are more effective than the baseline model with base features and a 10.75% relative improvement has been observed on the precision in classification of experiment, which proves that the academic text structure is important for automatic keywords extraction.

Key words: structure-function; keywords extraction; academic text; support vector machine; learning to rank

1 引 言

近年来，随着网络上学术资源的爆炸式增长，科研工作者迫切地需要对海量学术资源进行检索、分类、聚类、信息过滤、监督及自动摘要等操作，以提高对资源的访问效率和识别、区分对自己有价

收稿日期：2016-09-28；**修回日期**：2016-11-23

基金项目：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”（71473183），国家自然科学基金面上项目“基于多语义信息融合的学术文献引文推荐研究”（71673211）。

作者简介：方龙，男，1993年生，硕士研究生，主要研究方向为信息检索；李信，男，1991年生，博士研究生，主要研究方向为信息检索、文本挖掘；黄永，男，1991年生，博士研究生，主要研究方向为信息检索、数据挖掘；陆伟，男，1974年生，博士，副院长，教授，主要研究方向为信息检索、知识管理、数据挖掘等，E-mail: weilu@whu.edu.cn。

值的数据和信息[1-2]。关键词能够高度地概括科研文献的内容、精炼文献的主题信息，是实现上述操作的基础[3]。目前，除了大部分的科研论文包含有作者关键词外，绝大多数如网页、短文本等格式的学术资源存在关键词缺失的情况。同时，由于不同的作者对于词义的理解不同，导致人工标注关键词的结果带有强烈的主观意愿；而若采用统一标准进行人工标注，则耗时耗力、无法即时地完成标注任务。为了解决上述问题，关键词自动提取研究应运而生，即通过计算机采用统一的标准自动地在学术资源内容的基础上生成关键词。通过文献调研发现，计算机科学、图书情报学、信息科学及数据挖掘领域的前辈们已对该课题进行了广泛的探讨，如 Witten 等[4]提出了经典的 KEA 模型，Mihalcea 等[5]提出了 TextRank 算法，Beliga 等[6]提出了一种基于网络的关键词提取算法 SBKE 等，均取得了一定的效果。但是，在先前的大多数研究中没有考虑不同位置结构的文本内容在语义上蕴含的差异性，仅有少数研究涉及词汇在文本中出现的位置信息，却没有系统、深入地探讨其所在位置的结构功能对关键词抽取的影响；然而，实际上文本的结构功能体现了文章各部分的语义作用，处于不同结构功能的词汇蕴含的语义信息量是不同的。

基于此，本文试图将学术文本的结构功能作为特征，融合进关键词自动提取时的特征组合中，同时将关键词自动提取视为一个有监督的机器学习任务，在实验过程中对词汇所处位置的结构功能进行加权处理，以探讨学术文本的结构功能这一特征对关键词自动提取技术的影响。

2 相关研究

关键词自动提取的研究最早始于 1957 年 Luhn[7] 提出的利用词频的关键词抽取技术，随后国内外学者相继提出了多种关键词自动提取方法。从流程上来看，关键词自动提取的过程主要分为文本预处理、获取关键词候选词集和抽取关键词三个步骤；从方法上看，关键词自动提取则主要可分为基于统计的方法、基于机器学习的方法和基于语言学的方法三类。

2.1 基于统计的方法

基于统计的关键词自动提取，指通过对词汇的非语义特征信息进行统计计算，将特征值的加权得分排序。Luhn 等[7]最早的提出的基于词频统计的方法，Salton 等[8]引入 TF-IDF 特征的方法及 Matsuo 等[3]采用词共现的方法，都属于基于统计的提取方法。这类方法的特点是简单易用、计算方便，但是准确率较低。

2.2 基于机器学习的方法

基于机器学习的关键词自动提取，主要指将关键词提取转化为分类问题或机器学习排序问题。随着近年机器学习研究的高速发展，这类方法被广泛地使用，在抽取的准确率和召回率上有了较大的提升。例如，Witten 等[4]基于朴素贝叶斯提出的经典算法 KEA，取得了 33.27%的准确率；Zhang 等[9]则构造特征向量，利用支持向量机（SVM）模型进行关键词提取；Mihalcea 等[5]提出的 TextRank 算法对关键词进行排序；Yih 等[10]则训练最大熵模型进行关键词提取，取得了 46.97%的准确率；Jo 等[11]训练神经网络模型，结果准确率高达 64.7%。

2.3 基于语义的方法

基于语义的关键词抽取，主要指从更深的语义层次来对文本内容进行分析，利用词语的语义特征来判定候选词是否为关键词。例如，Ercan 等[12]使用了词法分析进行关键词提取，Hulth[13]使用了句法分析进行关键词提取等。虽然该方法较好地契合了人类的逻辑思维，但是由于目前还无法很好地解决词义消歧、同义词冗余表达等问题，因此该类方法在实践中单独使用较少，往往作为前两种方法的辅助手段来提高关键词提取的准确率。

总之，文献调研结果显示，机器学习是目前对学术文献关键词自动提取研究中应用最广泛、也最有效的方法。因此，本文拟选取机器学习的方法，提出一种融合学术文本结构功能的多特征关键词自动提取方法。本文研究的创新点如下：①将学术文本结构功能应用于关键词抽取，为关键词自动抽取提供了一种新的方法和思路；②通过实验验证了学术文本结构功能在关键词自动抽取上的效果提升上具有积极作用；③研究了学术文献中不同类型的关键词在各个学术文本结构功能域中分布情况进行了描绘。本文的下一部分将对学术文本的结构功能识别和关键词自动提取具体过程进行详细阐述。

3 研究方法描述

3.1 学术文本结构功能及其自动识别

学术文本的结构功能，指学术文本的不同章节

内容所体现的语义功能，反映各个章节的目的性和功能性[14]。陆伟等通过对大量计算机领域的科研文献结构进行总结分析，提出了一个较为完善学术文本结构功能的框架，即学术文本在结构上由“引言”、“相关研究”、“方法”、“实验”和“结论”等五个功能域构成[14]。如表 1 所示，每一种结构功能反映了相应章节体现的逻辑功能和目的，也构成了学术文本的逻辑结构，使学术文本的结构更加语义化。

表 1　学术文本结构功能框架

结构域	含义
引言	研究的引入、研究动机、研究目的等，是对研究问题的引入。
相关研究	研究综述、相关工作、背景介绍等，介绍相关研究和研究背景。
方法	提出的方法、框架、概念、设计及采用的工具等，主要是对论文研究思路的表达。
实验	实验数据、过程、评测、结果、发现、讨论及系统的设计、实现等，是对实验过程及结果的描述。
总结	全文的总结、下一步研究工作的展望及应用等。

从本质上来讲，结构功能其实是对学术文本中章节的功能和目的的标注，因此可以将其识别问题转化为一种分类问题。黄永等[15-16]在之前的研究中将结构功能的自动识别分为三个层次：第一，引入序列标注的思想使用条件随机场，根据章节标题来识别各个章节对应的结构功能；第二，在使用自然标注的方法构建大规模的数据集的基础上，利用深度学习的方法基于章节全部内容对结构功能进行识别；第三，进一步细化探讨章节中各个段落对于结构功能识别的作用，即基于章节中段落内容的结构功能识别方法，均取得了一定的效果。

考虑到本研究使用的数据集均来自于计算机语言学领域，该领域内论文结构较为固定，因此本文采用了操作简便、对内容结构固定文本识别准确率较高的基于章节标题的学术文本结构功能识别方法。

3.2　结合学术文本结构功能的多特征关键词自动提取

本文采用机器学习的思想，将关键词自动提取转化为基于多特征的二分类问题或机器学习排序问题。关键词提取的主要流程分为以下几个步骤：①抽取领域全部文献的作者关键词，经过去重处理，构建先验知识库；②获取候选词；③特征选取；④确定训练集，训练模型；⑤最终关键词提取。整体流程如图 1 所示。

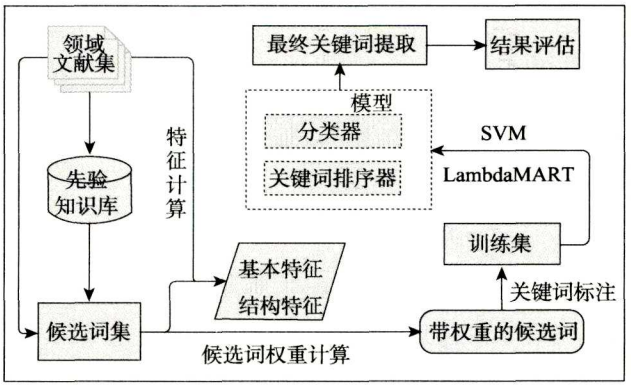

图 1　结合学术文本结构功能的多特征关键词自动提取

3.2.1　候选词的获取

对某一个特定领域的学术文献进行关键词提取时，利用作者关键词构建先验知识库来选取候选词集可以在一定上提高关键词提取的准确率[17]。因此，本文首先从含 N 篇文献的领域文献集 $D=[d_1,d_2,d_3,\cdots,d_N]$中抽取所有作者关键词，经过去重等处理，选取词频大于 1 的词汇构成先验知识库。然后，对每一篇文档 d 进行小写转换、词干提取等文本预处理，先验知识库进行对比匹配，将存在先验知识库中的词加入候选词集 W。最后，对于每一篇文献 d，得到一个候选词集 $W=[w_1,w_2,w_3,\cdots,w_m]$。为了后续的分类实验，对每个候选词需要根据先验知识库进行标注，如果候选词是作者关键词则标记为 1，否则，标记为 0。

3.2.2　特征选取

特征选取是关键词自动提取的关键步骤，很多算法都是通过对原有特征的改进或添加新的特征来优化提取效果[3,5,6,9]。本研究在对经典的 KEA 算法[4]分析的基础上，选取其使用的 TF*IDF 和词汇在文本中首次的位置作为候选词的基准特征，并将这两个特征与前文识别得到的“引言”、“相关研究”、“方法”等 5 种结构功能相结合，得到候选词的结构功能特征。下面将对基准特征和结构功能特征进行具体阐释。

（1）基准特征

基准特征包含“TF*IDF”和“词汇在文本中首次出现的位置”（FirstLocation）两个特征。TF*IDF 是在信息检索和数据挖掘领域被广泛使用的一种加权指标，它可以评估候选词语对于文献集中的某一篇文献的重要程度，即词汇的重要性随着它在文献中出现的次数而增加，但同时随着它在文献集中出现的频率而下降。

不同的词汇在文本中首次出现的位置往往不同，Witten 等[4]认为一般出现在科研论文前面位置如摘要、引言等部分的词语，成为关键词的可能性较其他位置的词语大，因为科研论文前面的部分常常可以概括论文的主题。因此，本文将词汇在文本中首次出现的位置作为特征之一，计算公式如下：

$$FirstLocation(w,d) = Location(w,d)/size(d) \quad (1)$$

其中，Location(w,d)为候选词w在测试文献 d 中首次出现的位置，即出现在第多少个单词的位置；size(d)为测试文献 d 中的总词汇数。

（2）结构功能特征

基准特征仅从文档级的层面对候选词进行了特征表示，引入结构功能则可以从更为细粒度的层面对候选词进行特征表示，可能提高关键词自动提取的效率。根据陆伟等[14]学术文本结构功能的划分，科研论文存在 5 个结构功能域，表示 $S=[s_1,s_2,s_3,s_4,s_5]$，那么文献集中的任一篇文献 d 是由$[s_1,s_2,s_3,s_4,s_5]$中的结构功能域构成。因此，对于每一个结构功能域，分别计算其对应的 5 个 TF*IDF 值和 1 个针对结构功能域的 FirstLocation 值，作为候选词的结构功能特征。计算公式如下：

$$TF * IDF_s = tf_s(w) * \log[n_s/df_s(w)] \quad (2)$$

$$FirstLocatio(w,s) = Location(w,s)/size(s) \quad (3)$$

其中，$tf_s(w)$是 w 在测试文献结构功能域s中的词频，n_s 为训练样本中结构功能域s的数目，$df_s(w)$是训练样本中包含有词汇 w 的结构功能域 s 的数目 Location(w,s)为候选词 w 首次出现的结构功能域 s 中的位置，size(s)为结构功能域 s 的总词汇数。

3.2.3 模型训练

关键词自动提取可以看作一个典型的二分类问题[18-19]，即对候选词进行二值判断，属于关键词或不属于关键词。同时，关键词自动提取也可以看做是一个排序问题[5,20]，即根据特征权重得分对候选词汇进行排序，获取 TopN 个词作为文档的关键词。基于此，本文使用林智仁教授[21]等开发的 LIBSVM 工具和 Dang[22]开发的 RankLib 工具，实现了 SVM 算法和 LambdaMART 算法来分别训练关键词自动提取的二分类模型和机器学习排序模型。

4 实验与结果分析

4.1 数据集

本文使用了 2000 年至 2014 年 ScienceDirect 数据库中 13 多万篇的计算机语言学领域期刊论文作为文献集，抽取该领域文献集的全部作者关键词，进行词干提取、去重、词频统计等处理，保留词频大于 1 的关键词，得到含有 74723 个词汇的领域先验知识库。然后，随机选取 4000 篇文献作为本次实验的实验文献集，并随机选择其中 3000 篇文献作为训练文献集，剩余的 1000 篇文献作为测试文献集。然后随机抽取 20000 个标记为 0 的候选词和训练文献集中所有标记为 1 的候选词及其特征表示作为模型训练所需的训练集，将剩余 1000 篇文献中标记为 1 的候选词及其特征表示作为模型预测评估所需的测试集。

实验文献集中共有 19087 个作者关键词，其文档频率分布如图 2 所示，平均每篇文献 4.77 个，其中有 82.27%的关键词（15703 个）在全文中出现（不去重），平均每篇文献有 3.93 个关键词。对于未出现在全文的 3117 个（去重）作者关键词，统计其在文献集中的文档频率如表 2 所示，可以发现大部分在全文中未出现的作者关键词在领域知识库中是存在的。以上统计结果说明，从领域文献集抽取作者关键词构建先验知识库用于关键词自动提取是合理的。

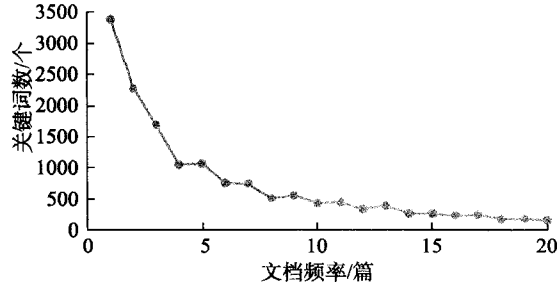

图 2 实验文献集中作者关键词文档频率分布图

表 2 未在全文出现的作者关键词在文献集中的文档频率分布表

文档频率	关键词数	关键词占比/%
df>1	1660	53.3
df>2	1259	40.4
df>3	1053	33.8
df>4	911	29.2
df>5	817	26.2
df>6	756	24.3

此外，关键词对表述文献内容有不同的作用，在文献中有不同的功能，例如，有些关键词反映了文献中的研究问题，有些关键词说明了文献中使用的研究方法，Dutta 等[23]提出了关键词分类的框架，

在此基础上，本文将关键词分为问题关键词、方法关键词和其他关键词等类型词，并随机抽取 50 篇文献，将这些文献的作者关键词（232 个）进行标注，分别统计每个类型关键词在各个结构功能域中的分布情况。表 3~表 5 描述了 50 篇作者关键词的在各结构功能域中出现的词数及其占该类词数的比例，其对应的分布图如图 3~图 5 所示。对比分析可发现无论是哪种类型的关键词均在引言部分的分布最多，都超过 55%，问题类型关键词最高达 66.3%；同时，无论是哪种类型的关键词在相关研究部分的分布最少，其他类型关键词最低仅 6.5%；此外，不同类型的关键词在各个结构功能域中的分布也存在较为明显的差异，如问题类型的关键词在相关研究部分占 22.4%，而其他类型的关键词在相关研究中的占比仅为 6.5%。以上现象说明将学术结构功能加入候选词的特征表示中对于关键词自动提取是有意义的、可行的。

表 3　问题类型关键词(98 个)在结构功能域中的分布情况

	引言	相关研究	方法	实验	结论
关键词占比/%	66.3	22.4	39.8	42.9	42.9
关键词数	65	22	39	42	42

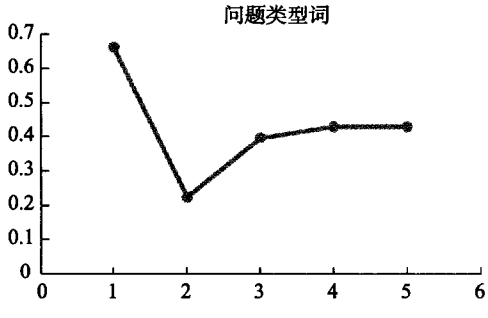

图 3　问题类型词在结构功能域中的分布图

表 4　方法类型关键词（88 个）在结构功能域中的分布情况

	引言	相关研究	方法	实验	结论
关键词占比/%	63.6	14.8	46.6	37.5	36.4
关键词数	56	13	41	33	32

4.2　实验结果及评测

基于相同的训练文献集和测试文献集，本文分别利用候选词的“基准特征”和“基准特征+结构功能特征”先后实现了基于 SVM 二分类的关键词自动提取和基于 LambdaMART 机器学习排序的关键词自动提取，并以“基准特征”的结果为 Baseline，对分类及排序结果进行了评估比较。

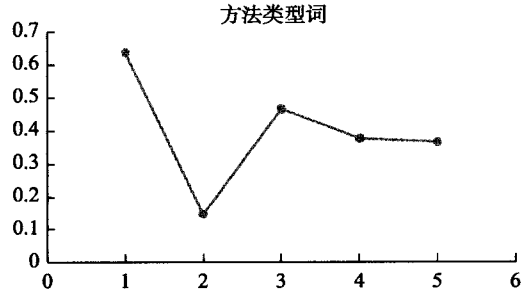

图 4　方法类型词在结构功能域中的分布图

表 5　其他类型关键词（46 个）在结构功能域中的分布情况

	引言	相关研究	方法	实验	结论
关键词占比/%	58.7	6.5	45.7	39.1	45.7
关键词数	27	3	21	18	21

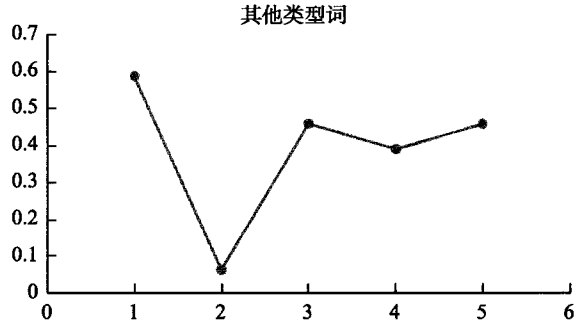

图 5　其他类型词在结构功能域中的分布图

（1）SVM 二分类结果评估

本文主要采用准确率来对分类结果进行评估，评估结果见表 6。

表 6　SVM 分类实验结果评价　　%

评价指标	基准特征（2 个）	基准特征+结构功能特征（8 个）	相对提升
准确率	48.67	53.90	10.75

从表 4 和表 5 可以发现，融合了结构功能特征的关键词提取的准确率比仅用基准特征的效果提升了 10.75%，达到 53.90%。这说明学术结构功能特征能够提升基于二分类的关键词自动提取的效果。

（2）LambdaMART 机器学习排序结果评估

对于 LambdaMART 机器学习排序的结果，本文主要采用了 MAP、P@5 和 NDCG@5 三个指标对训练得到的模型进行评估，结果如表 7 所示。

从表 7 中可以看出，融合了结构功能特征的排序模型在关键词提取结果的 MAP、P@5 和 NDCG@5 值上相对于仅用基准特征的排序模型均有提高，其中 P@5 的提升最大，达到 5.08%。因此，学术文本

结构功能特征能够提升基于机器学习排序的关键词自动提取的效果。

表7 lambda MART机器学习排序结果评价

评价指标	基准特征（2个）	基准特征+结构功能特征（8个）	相对提升（%）
MAP	0.3394	0.352	3.71
P@5	0.1832	0.1925	5.08
NDCG@5	0.3539	0.3713	4.92

此外，假设返回的结果为5个关键词，那么分类结果中有2.7（5×53.90%）个结果是正确的，而机器学习排序的最好结果只有0.96（5×0.1925）个结果是正确的，因此，对比分类实验和机器学习排序的结果，可以发现学术文本结构功能特征对于分类实验的效果提升优于基于机器学习排序的关键词自动提取的实验效果。

4.3 分类预测错误分析

本文选取了两个分类实验的预测错误进行分析，其中仅用基准特征的分类实验结果中有655篇文献关键词判别错误，融合了结构功能特征的分类实验中出现关键词判别错误的文献有614篇，即利用结构功能特征的关键词提取方法将错误率降低了6.26%。

同时，统计得到两个实验的预测错误中有596篇文献是相同的，经过关键词去重，得到出错关键词不同的文献122篇。对于这122篇文献，72.13%（88篇）的文献在仅使用基准特征时的出错关键词个数比融合了结构功能特征的出错关键词个数多；且融合结构功能特征正确提取的关键词比仅用基准特征正确提取的关键词多109个，其中有85个关键词（占比77.98%）是出现在文献不同的结构功能域中。上述现象说明，融合学术文本结构功能域的关键词自动提取方法较好地解决了基准特征的局限性，在准确率和召回率上都有较好的提升。

此外，本文还对上述596篇文献中的关键词在领域文献集中文档频率分布进行了描绘，如图6所示。与图2对比可以发现，这些关键词的文档频率普遍偏低，且不是领域的热点关键词，有可能是随着领域的发展交叉而产生的新词，这给关键词自动提取带来了更大的挑战。

5 结 语

本文为探讨学术文本结构功能在关键词自提取

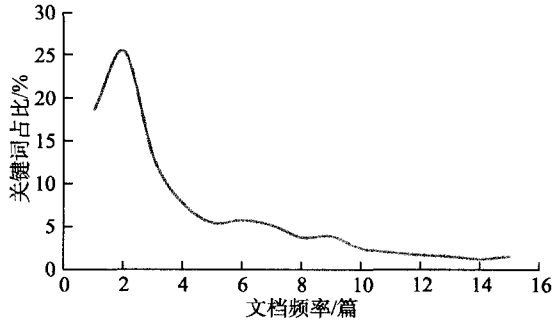

图6 预测错误文献中关键词在领域文献中的文档频率占比图

中的作用，将学术文本看作是结构功能域的集合，把文本的结构功能特征加入候选词的特征组合中，并在计算机语言学领域的文献机上分别利用SVM和Lambda MART进行分类实验和机器学习排序实验来提取文本关键词。实验结果表明，融合结构功能特征的多特征关键词提取方法在分类结果的准确率和排序结果的MAP、P@5、NDCG@5等各个评测指标上均有较大提升；并且根据预测错误结果分析发现，融合学术文本结构功能域的关键词自动提取方法较好地解决了基准特征在结构语义上局限性，在准确率和召回率上都有一定提升。

此外，本文提出的融合结构功能特征的多特征关键词自动提取方法虽然取得了较好的实验结果，但仍存在一些问题需要进一步探索：本文在候选词的获取时使用了先验知识库，对于特定领域的关键词提取有积极作用，但不适合领域无关的关键词自动提取，因此需进一步针对领域无关的关键词提取优化候选词获取的方法；本文证明了学术文本结构功能在关键词自动提取中具有一定的应用价值，故应进一步发掘其应用场景。

参 考 文 献

[1] Thakur A, Lal A, Lim J, et al. PostHat and all that: automating abstract interpretation[J]. Electronic Notes in Theoretical Computer Science, 2015, 311: 15-32.

[2] Tkaczyk D, Szostek P, Fedoryszak M, et al. CERMINE: automatic extraction of structured metadata from scientific literature[J]. International Journal on Document Analysis and Recognition (IJDAR), 2015, 18(4): 317-335.

[3] Matsuo Y, Ishizuka M. Keyword extraction from a single document using word co-occurrence statistical information[J]. Transactions of the Japanese Society for Artificial Intelligence, 2002, 17(3): 217-223.

[4] Witten I H, Paynter G W, Frank E, et al. Automatic keyphrase

extraction using probabilistic prediction[C]//Proceedings of the ACM Conference on Digital Libraries. New York: ACM Press, 1999: 254-255.

[5] Mihalcea R, Tarau P. TextRank: Bringing order into texts[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP, 2004: 404-411.

[6] Beliga S, Meštrović A, Martinčić-Ipšić S. Selectivity-based keyword extraction method[J]. International Journal on Semantic Web & Information Systems, 2016, 12(3): 1-26.

[7] Luhn H P. A statistical approach to mechanized encoding and searching of literary information[J]. IBM Journal of Research and Development, 1957, 1(4): 309-317.

[8] Salton G, Yang C S. On the specification of term values in automatic indexing[J]. Journal of Documentation, 1973, 29(4): 351-372.

[9] Zhang K, Xu H, Tang J, et al. Keyword extraction using support vector machine[C]// Proceedings of the International Conference on Advances in Web-Age Information Management. Heidelberg: Springer-Verlag, 2006: 85-96.

[10] Yih W T, Goodman J, Carvalho V R. Finding advertising keywords on web pages[C]// Proceedings of the 15th International Conference on World Wide Web. New York: ACM Press, 2006: 213-222.

[11] Jo T, Lee M, Gatton T M. Keyword extraction from documents using a neural network model[C]// Proceedings of International Conference on Hybrid Information Technology. Los Alamitos: IEEE Computer Society Press, 2006: 194-197.

[12] Ercan G, Cicekli I. Using lexical chains for keyword extraction[J]. Information Processing & Management, 2007, 43(6): 1705-1714.

[13] Hulth A. Automatic keyword extraction: Combining machine learning and natural language processing[M]. Saarbrücken: VDM Verlag, 2008.

[14] 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.

[15] 黄永, 陆伟, 程齐凯. 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.

[16] 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, 2016, 35(5): 530-537

[17] 宋宇, 真溱. 关键词自动抽取技术综述[J]. 情报理论与实践, 2016, 39(4): 141-144.

[18] Frank E, Paynter G W, Witten I H, et al. Domain-specific keyphrase extraction[C]// Proceedings of the 16th International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers, 1999: 668-673.

[19] Turney P D. Learning algorithms for keyphrase extraction[J]. Information Retrieval Journal, 2000, 2(4): 303-336.

[20] Li L, Su C, Sun Y Q, et al. Hashtag biased ranking for keyword extraction from microblog posts[C]// Proceedings of the International Conference on Knowledge Science, Engineering and Management. Springer International Publishing, 2015: 348-359.

[21] Chang C C, Lin C J. LIBSVM: A library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): Article No. 27.

[22] Dang V. Ranklib[EB/OL]. [2016-08-08]. http://sourceforge.net/p/lemur/wiki/ranklib/.

[23] Dutta B, Majumder K, Sen B K. Classification of keywords extracted from research articles published in science journals[J]. Annals of Library and Information Studies, 2008, 55(4): 317-333.

（责任编辑　车　尧）