基于PageRank的新闻关键词提取算法

顾亦然，许梦馨

(南京邮电大学自动化学院 南京 210023)

【摘要】现有的基于复杂网络的关键词提取算法在构建加权文本网络时没有考虑文本的自然语言特性，且在提取关键词时较少涉及复杂网络领域经典算法。本文引入词频分享权重，利用词频特性为节点之间的连边加权。在此基础上，基于PageRank算法，并结合人类语言习惯特性定义位置权重系数，提出了一个新的新闻关键词提取算法——LTWPR算法，综合考虑了文本网络的局部特征和全局特征。采用新浪新闻语料进行了大量实验，结果表明该算法能够快速有效的覆盖新闻作者标注的关键词，且提取效果更佳。

关 键 词 复杂网络；关键词提取；自然语言；PageRank；词频分享权重

中图分类号 TP311；TP391.1 **文献标志码** A doi:10.3969/j.issn.1001-0548.2017.05.021

Keyword Extraction from News Articles Based on PageRank Algorithm

GU Yi-ran and XU Meng-xin

(College of Automation, Nanjing University of Posts and Telecommunications Nanjing 210023)

Abstract Most of the existing methods of extracting keyword based on complex networks ignore the natural language characters when building the weighted text network. In the meantime, they involve less the classical algorithms in complex network field. Based on PageRank algorithm, we propose a keyword extraction method, named LTWPR (located and TF-weighted PageRank), which takes into consideration term-frequency character and human language characters. The algorithm creates a term-frequency-shared weight in order to share the node's term-frequency value to its links, and defines a position weight coefficient to express different importance of words in different positions of news articles. LTWPR brings text networks' local and global features into consideration, making the results more accurate. Comprehensive experiments are conducted based on news articles grabbed from Sina News. Experimental results show that LTWPR algorithm is more effective and can better cover the keywords tagged by authors.

Key words complex networks; keyword extraction; natural language; PageRank; term-frequency-shared weight

随着信息时代的到来以及互联网的蓬勃发展，关键词成为用户搜索信息必不可少的工具。关键词以凝练简洁的形式对文本主题进行有效概括，通过提取关键词，可以结构化地表示目标文本，提高人们的文献管理与检索效率。

关键词在新闻领域有十分重要的作用。目前，网页新闻如新浪新闻会在网页源代码中标注keywords或tags属性的词语，并在网页新闻下端贴出标签或文章关键词，使得用户在搜索相关新闻时能快速定位。由于个体语言的差异性，手动标注关键词可能存在不规范或不准确的问题，且核对工作较繁琐。因此，找到一种规范化、合理高效的文本关键词自动提取方法具有十分重要的意义。

传统的关键词提取算法是基于TF-IDF[1-2]计算词语的特征权重，利用词频TF发现高频词，再通过引入逆文本频率指数IDF[3]来降低高频却不具代表性的词语对文本的重要度，提高提取关键词的准确率，算法思想十分简单。但此方法计算复杂度较高，需将所有文本均考虑在内才能计算词语的逆文本频率指数，因此，其提取关键词准确度受文档集合大小的影响较大。另一经典算法是以TextRank[4]为典型代表的基于词图模型的关键词抽取算法。受著名的Google网页排名算法PageRank[5]的启发，文献[4]把词看做网页，将词与词之间的语义关系看作链接，

收稿日期：2016－09－16；修回日期： 2016－12－26
基金项目：教育部人文社会科学研究规划基金(15YJZH016)
作者简介：顾亦然(1972－)，女，博士，教授，主要从事复杂网络理论与应用、嵌入式系统及通信网络等方面的研究.

开发了TextRank算法，因其不需要事先对多篇文本进行训练，仅利用单篇文档本身就能实现关键词提取，实现方法简单高效并得到广泛应用。然而，该方法采用的是词语节点影响力均分的无权图模型，在进行关键词抽取时仅考虑了词语的词性信息，未考虑词语节点之间的相互影响力，导致非重要词语吸收的贡献值相对增加。

基于复杂网络的关键词提取方法是近年来随着复杂网络研究的兴起而出现的一种新的关键词提取算法。文本网络已被证实具有小世界特性[6]，可以使用复杂网络理论进行关键词的提取。在一个文本网络中，词语被视为节点，词语之间的联系抽象为连边。所有词语和连边即构成一个文本复杂网络。已被研究出的基于复杂网络的关键词提取方法大多基于词语在同一句子中共现次数为连边加权[7]，再应用复杂网络统计参数度、聚类系数、介数、接近中心性、最短路径等[8]两参数加权或三参数加权计算得到词语节点的特征权重，进而得到一篇文章的关键词。文献[9]利用特征词共现次数为连边加权，通过节点的加权聚类系数和介数两参数加权计算节点的综合特征值。文献[10]利用词语共现次数为连边加权，通过加权度及聚类系数两参数加权计算节点的特征权重。文献[11]利用两个词语在同一句话中共现次数的倒数为连边加权，通过节点的加权中心度和介数两参数加权计算节点的综合特征值，从而提取文本关键词。文献[12]提出应用语义加权网络提取中文关键词的方法，利用词共现频率和语义相似度构造语义加权网络，通过节点的介数、聚类系数变化值和平均最短路径变化值三参数加权计算得到节点的综合特征值。文献[13]提出的基于复杂网络的关键词提取方法也通过词语共现次数的倒数为连边加权，利用偏向中心性和度中心性两参数归一再加权计算节点的综合特征值。文献[14]依据词汇在文本中的共现关系构造词汇概念复杂网络，提出了一种利用词汇概念本身频率以及其相邻节点的数量及重要性指标为节点加权，计算出文本词汇的重要性指标获取候选关键词集。

已有的基于复杂网络的关键词提取研究大部分只应用了复杂网络统计参数，较少应用复杂网络经典方法计算词语节点的权重，且忽略了可以利用自然语言词频特性对节点和节点之间的连边赋予权重的方法。本文针对上述问题进行研究分析，在较好构建文本复杂网络的基础上，基于PageRank算法提出一种新的新闻关键词提取算法，实现对关键词的有效提取。

1 基于PageRank的新闻关键词排序算法LTWPR

本文算法思想可概括为以下3点：

1) 考虑到词频对新闻主题的重要性，引入词频分享权重概念，将目标节点的词频值根据邻居节点对其的重要度来分配给相应的连边，实现对连边加权；

2) 考虑邻居节点对目标节点的重要度贡献，即：与关键词联系越紧密的词语其成为关键词的可能性也越大，基于PageRank算法提出本文LTWPR (located and TF-weighted PageRank)算法；

3) 考虑词语位置对词语重要度的影响，增加位置权重系数。显然，对新闻文本而言，出现在标题中的词语能够更好地反映主题，相比于出现在正文中，其重要性更高，因此对位于新闻标题的词语赋予更高的权重系数。

本文算法摒弃了大部分研究使用的复杂网络两参数或三参数加权的方法，直接将最原始的词频指标与复杂网络经典方法PageRank相融合，并考虑了人类语言习惯而设计位置权重系数。在时间复杂度上，由于传统的TF-IDF算法在计算每个词语的IDF值时均需要遍历整个新闻文本集合，时间复杂度可谓非常高。而本文算法在读入新闻文本集合后，一次只需处理一篇，时间复杂度较TF-IDF算法更低。同时，与已有的基于复杂网络的关键词提取算法相比，本文算法省去了对两参数或三参数的调节过程与时间，获取结果更为直接。

1.1 基于连边规则的文本网络构建

文本网络构建过程最重要的是分词和连边关系的创建。首先，分词获得所有词语节点。目前，文献[15]的NLPIR分词软件已经相对成熟，故分词步骤直接调用NLPIR软件的接口来实现。

其次，根据一定规则创建词语节点之间的连边。经过大量学者的研究发现，距离为1或2的词语之间构建一条连边在文本复杂网络中最为常见，且效果最佳[9]。因此，本文采用该方法构建连边规则，即同一个句子中，距离小于等于2的有效词语之间产生一条连边且连边方向为当前词语单向指向位于其后的词语。其中，有效词语指可能作为关键词的词语。同时，为了对该连边规则进行精确定义，本文提出了标点符号识别策略，对“同一个句子”和“距离”定义如下：

1) 同一个句子定义为：以句号、叹号、逗号、分号、省略号分隔。

2) 距离定义为：以括号、引号、顿号、冒号、破折号分隔，间距记为1；以百分号、千分号、单位符号分隔，间距记为0。因为分词软件会直接将此类标点与其修饰的数字分为同一个词，如5.7%作为数词词性存在，故此3种标点无需考虑。

1.2 文本网络构建算法

文本网络构建算法流程图如图1所示。具体步骤描述如下。

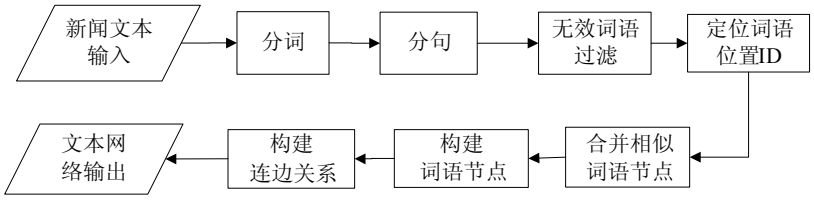


图1　文本网络构建算法流程图

输入数据：新闻文本。

1) 分词。读入用户词典，利用NLPIR分词软件输出“词语+词性”组合列表，存入List<String> WordList，记录词语内容strWord，词性intType；

2) 分句。根据标点符号将目标文本分成句子，存入List<String> Sentence；

3) 无效词语过滤。若命中停用词表和无用词性表的词语标记为Ignore；

4) 定位每个未被标记Ignore的词语在句子中的位置。根据Sentence确定词语位置intPositon，和所属句子索引intLineNum，作为词语独一无二的位置ID；

5) 合并词语。扫描词语，对同义词进行合并，将被兼并词语标记为Merged，指向兼并它的词语；

6) 构建节点。将未被标记为Ignore的剩余词语构造为节点，存入List<kNode> NodeList中；

7) 构建网络。根据1.1节定义的连边规则，依据位置intPosition和所属句子索引intLineNum对所有节点进行加边，对于标记为Merged的兼并节点，将边和邻居节点加载到其宗主节点上，两节点之间的相应边权不变。

输出数据：新闻文本网络。

1.3 新闻关键词提取算法LTWPR

新闻关键词提取算法LTWPR具体步骤描述如下。

1) 构建文本网络，如1.2节描述。

2) 词频计算。根据List<String> WordList中未被标记为Ignore的词语，计算每个词语节点 i 的词频 TF_i 值。

3) 迭代计算词语节点 i 的初步特征权重 TWPR_i 值，如式(1)。其中，TWPR(0)的初始取值规则为：若目标节点的度值非0，则初值设为网络中度值非零节点总数的倒数，否则设为0。迭代终止条件为全部节点的最后两次TWPR值误差精度 $E \leqslant 0.001$。

$$\mathrm{TWPR}_i(k) = s\sum_{j\in\Gamma(i)}\mathrm{TWRP}_j(k-1)\frac{\mathrm{TWeight}_{ji}}{\mathrm{TW}_j}+(1-s)\frac{1}{N} \tag{1}$$

$$\mathrm{TWeight}_{ij} = \mathrm{TF}_i\frac{\mathrm{TF}_j}{\sum_{k\in\Gamma(i)}\mathrm{TF}_k}+\mathrm{TF}_j\frac{\mathrm{TF}_i}{\sum_{k\in\Gamma(j)}\mathrm{TF}_k} \tag{2}$$

$$\mathrm{TW}_j = \sum_{p\in\Gamma(j)}\mathrm{TWeight}_{pi} \tag{3}$$

式中，k 为当前迭代步数；N 为总节点数；s 为系数因子，一般取 $s=0.85$；$\mathrm{TWeight}_{ij}$ 为节点 i 与其邻居节点 j 对两者连边的TF分享权重，计算方法如式(2)，其中，TF_i 为节点 i 的TF值，$\Gamma(i)$ 为节点 i 的邻居节点集合；TW_j 为与节点 j 相连的所有连边权重之和，计算方法如式(3)，其中，$\Gamma(j)$ 为节点 j 的邻居节点集合。

4) 计算节点综合特征值 KWW_i：

$$\mathrm{KWW}_i = \begin{cases}\alpha * \mathrm{TWPR}_i & \mathrm{Key}=0\\ (1-\alpha) * \mathrm{TWPR}_i & \mathrm{Key}=1\end{cases} \tag{4}$$

式中，α 为词语节点的语义权重系数；$\mathrm{Key}=0$ 代表节点位于标题句子位置，$\mathrm{Key}=1$ 代表节点位于新闻正文；KWW_i 值越大，说明词语节点 i 成为关键词的可能性越高。本文实验中统一取 $\alpha=0.65$。

5) 根据KWW值对所有词语节点进行从大到小排序，输出排名前 n 的词语即为提取的关键词。

2 实验及结果分析

本文爬取了新浪新闻(http://news.sina.com.cn/)下6个领域类别(财经、国际、国内、政务、军事、社会)共2 536篇新闻作为实验数据集，并与经典的TF-IDF算法和TextRank算法进行对比。采集的新闻文本格式如图2所示。

由于个体语言的差异性，新闻中作者标注的关

键词集难免存在不合理的情况，且对同一事件的报导不同新闻媒体的叙述方式不同标注的关键词也不尽相同。然而，考虑到新浪新闻的权威性、正式性和凝练性，本文默认大部分新闻文本数据的关键词标注是合理的。因此，在实验数据的准备工作中，只对部分关键词标注明显不合理的新闻文本进行关键词手动微调，去掉确实不合理或者有重复性质的关键词。

标题：日媒：中国军舰若无其事经过日本领海恐成常态
关键词：中国海军 中国军舰 日本领海
正文：日本右翼媒体《产经新闻》6月15日称，针对中国海军军舰15日进入日本“领海”一事，日本政府并未提出抗议，而仅仅是向中方表达担忧。《产经新闻》声称，“无害通航”不能成为中国海军在日本领海内通行的凭证。如果日方再次允许这种事情发生，今后中国军舰会再次“若无其事地经过日本领海”，这种航行恐怕也会成为“常态”。对此，中国海军军事学术研究所研究员李杰在接受环球网采访时明确指出，日方试图在国际法所认可的“无害通航”上搞双重标准。李杰强调，中国海军正常的海洋活动，他国无可指责。

产经声称，此次中国海军派出的是一艘情报收集舰，中国海军已经“事实上”构成在日本领海内“从事军事行动”。产经说，日本政府“不能放任事态置之不理，任由中国海军从事搜集情报”。

对此，中国海军军事学术研究所研究员李杰指出，国际法中的“无害通航”原则不仅适用于商船，也同样适用于军舰，尽管部分国家对军舰是否拥有“无害通航”的权利持不同立场，但日本政府奉行国际法中关于军舰“无害通航”的条款，比如美国的军舰就可以在日本领海“无害通航”航行。李杰表示，日本现在提出中国军舰不适于“无害通航”，明显是在搞双重标准。

产经还称，从九州南端到冲绳与那国岛之间的西南诸岛水域是中国海军进出太平洋的重要关口，也是日本自卫队与美军共同监控中国海军的“要塞”，如果此次对中国海军的行动采取默认立场，阻挡中国海军的要塞将成为“摆设”。

在李杰看来，日本媒体的说法是强词夺理。“中国海军在太平洋的活动中必然要探索各种道路，寻找最为便利的航线。中国海军走哪条道，过哪个海峡，那都是中国自己的问题。中方没有危害到他国的安全，中国海军正常的海洋活动，他国无可指责。”李杰表示。

针对日本炒作中国海军军舰驶入日本“领海”一事，15日下午，中国国防部新闻局在回应《环球时报》问询时表示，吐噶喇海峡是用于国际航行的领海海峡，中国军舰通过该海峡符合《联合国海洋法公约》规定的航行自由原则。（环球网）

图2 新闻文本格式

实验时采用召回率 R、准确率 P 和综合 F 值作为算法的评价指标。3个指标的定义分别如式(5)～(7)所示：

$$R=\frac{S\cap E}{S} \tag{5}$$

$$P=\frac{S\cap E}{E} \tag{6}$$

$$F=\frac{2RP}{R+P} \tag{7}$$

式中，S 表示作者标注的关键词集；E 表示应用相关算法自动提取的关键词集。在考虑整体的性能指标时，采用平均召回率、平均准确率和平均综合 F 值进行评价。

本文首先随机选取单篇新闻进行对比实验说明本文算法优于TF-IDF算法和TextRank算法之处。进一步的，为了验证基于复杂网络的新闻关键词排序算法的有效性，针对爬取的2 536篇新浪新闻语料，首先将所有新闻语料汇总，进行上述3种算法对比实验，并分析设置提取关键词个数的合理性，在此基础上将新闻按领域分类，再进行综合对比实验并分析结果。

2.1 单篇新闻实验

首先采用图2所示的新闻“日媒：中国军舰若无其事经过日本领海恐成常态”进行对比实验。3种算法提取出的关键词及相关指标如表1所示。

表1 3种算法提取结果对比

算法	关键词	召回率R/%	准确率P/%	综合F值/%
TF-IDF	中国海军、无害通航 李杰、日本领海	66.667	50	57.143
TextRank	中国海军、军舰 无害通航、日本	33.333	25	28.571
LTWPR	中国海军、日本领海 军舰、中国军舰	100	75	85.714

观察表1，新闻作者标注的标准关键词为“中国海军”、“中国军舰”、“日本领海”3个，实验设置提取的关键词个数为4。其中，TF-IDF方法提取出的关键词为“中国海军”、“无害通航”、“李杰”、“日本领海”。观察图2中新闻文本，发现词语“无害通航”在文中出现7次，“李杰”在文中出现6次，使得这两个词语成为关键词的可能性大大提高，干扰了最终的提取结果。TextRank算法提取的出关键词为“中国海军”、“军舰”、“无害通航”、“日本”，对标准关键词集的覆盖率较低。分析发现“中国军舰”和“日本领海”组合词词频较低，邻居节点数量小，而“军舰”和“日本”词频相对高一些，在文本网络中拥有更多的邻居节点为其贡献重要度，因而排名靠前，导致最终各项提取指标值较低。而本文LTWPR算法提取出的关键词为“中国海军”、“日本领海”、“军舰”和“中国军舰”，基本覆盖标准关键词集。说明本文算法在一定程度上优于TF-IDF算法和TextRank算法。

2.2 按关键词提取个数对比实验

通过观察爬取的2 536篇新浪新闻语料发现，作者标注的关键词(以下称标准关键词)个数在1～5范围之间。关键词分布如图3所示。

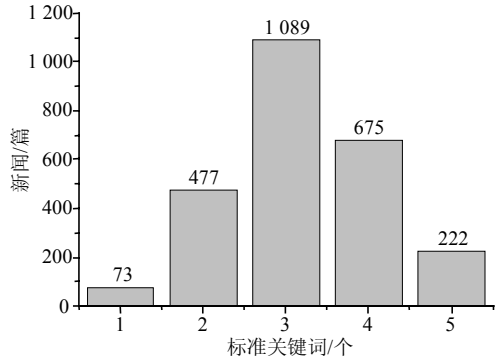
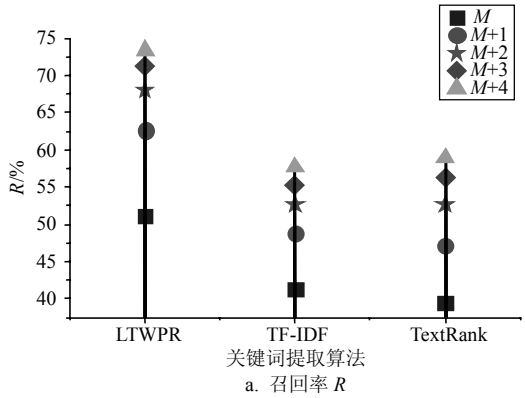


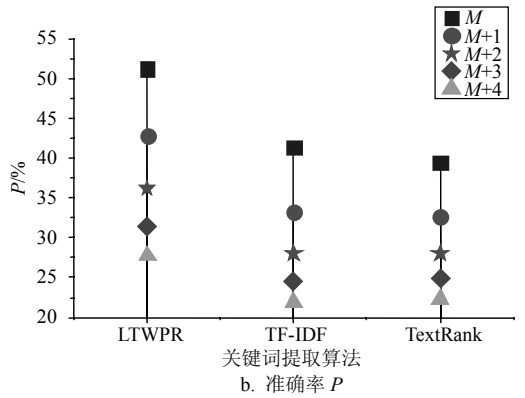
图3　新闻标准关键词分布

从图3中可以看出，对于新闻这种简洁明了的文本(500～800字最佳)，标准关键词设置3个的最为普遍，设置4个和2个的次之。标准关键词设置为1个时，新闻主题概括较不明确，在用户输入关键词搜索新闻时，较难匹配到目标新闻，较大程度降低搜索的准确性。而标准关键词设置为5个或更多时，新闻主题特征分散或不明显，容易干扰搜索结果，导致搜索有效性的降低。

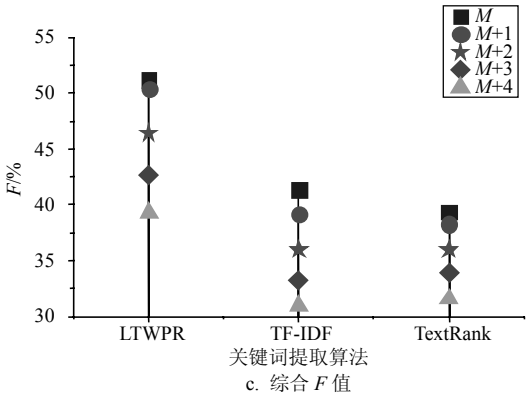
本文按照标准关键词个数将新闻语料分为汇总1～汇总5五个类别，假设新闻语料标准关键词个数为 M 个，设置提取的关键词个数分别为 $M \sim M+4$ 个，进行对比实验，实验结果如图4所示。



a. 召回率 R



b. 准确率 P



c. 综合 F 值

图4　按关键词提取个数实验结果

观察图4可得，在提取的关键词个数设置为 $M \sim M+4$ 各种情况下，本文LTWPR算法与经典的TF-IDF算法和TextRank算法相比，召回率 R、准确率 P 和综合 F 值均更高，且上述5种情况下，平均召回率 R、平均准确率 P 和平均综合 F 值较TF-IDF算法分别提高14.182%，8.056%和9.868%，较TextRank算法分别提高14.513%，8.384%和10.198%，提取效果均更佳。本文LTWPR算法在提取关键词个数为 $M \sim M+4$ 五种情况下，召回率 R 最高达到73.451%，而TF-IDF算法和TextRank算法分别只有57.675%和58.908%。

另一方面，从图4中不难发现，准确率 P 值随着提取关键词个数的增加而减小。提取关键词个数设置为 M 时，准确率 P 值最大，而由于提取的关键词个数最少，导致此时召回率 R 值最小，提取的关键词个数设置为M+4时则恰好相反。去除上述这两种极端的情况，同时考虑到 F 值是准确率 P 和召回率 R 两个参数的综合指标，通过对比分析发现，当提取关键词个数为 $M+1$ 和 $M+2$ 时，F 值相对最大，且此时准确率 P 值相对较高。因此本文得出以下结论：提取的关键词个数设置为 $M+1$ 和 $M+2$ 时最合理，同时也更加符合实际情况下人类的行为特性，能够在实际应用于批量提取新闻关键词的情况下，做到保留选择余地的同时不容易干扰作者最终的关键词标注结果。

2.3　按新闻领域对比实验

爬取的2 536篇新浪新闻语料各个领域类别篇数分布大致相同。根据2.2节的实验结果，得到提取的关键词个数设置为 $M+1$ 和 $M+2$ 时最合理。因此，本节只对提取关键词个数为 $M+1$ 和 $M+2$ 的新闻语料进行按新闻类别分类的对比实验，结果分别如图5和图6所示。

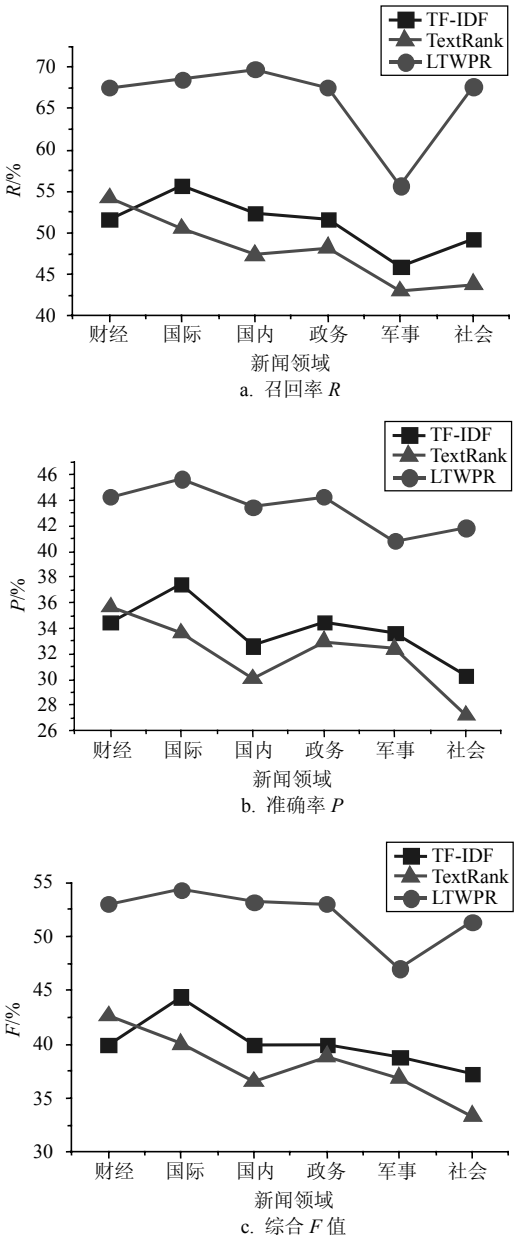


图5 提取关键词M+1时各项指标结果

观察图5和图6，本文提出的新闻关键词提取算法LTWPR的提取召回率 R 、准确率 P 和综合 F 值均高于经典的TF-IDF算法和TextRank算法。其中，对于提取 $M+1$ 个关键词的情况，LTWPR算法与TF-IFD算法相比，在各个新闻领域的平均召回率 R 、平均准确率 P 和平均综合 F 值分别提高了15.005%、9.56%和11.933%；与TextRank算法相比，分别提高了18.287%，11.405%和13.949%。对于提取 $M+2$ 个关键词的情况，LTWPR算法在各个新闻领域的平均召回率 R 、平均准确率 P 和平均综合 F 值较TF-IDF算法分别提高了18.222%、9.094%和11.96%，较TextRank算法分别提高了20.074%，9.309%和12.564%，其中，召回率 R 提升幅度最大。因此，本文LTWPR算法在新闻关键词提取方面优于TF-IDF算法和TextRank算法，具有良好的有效性和实用性。

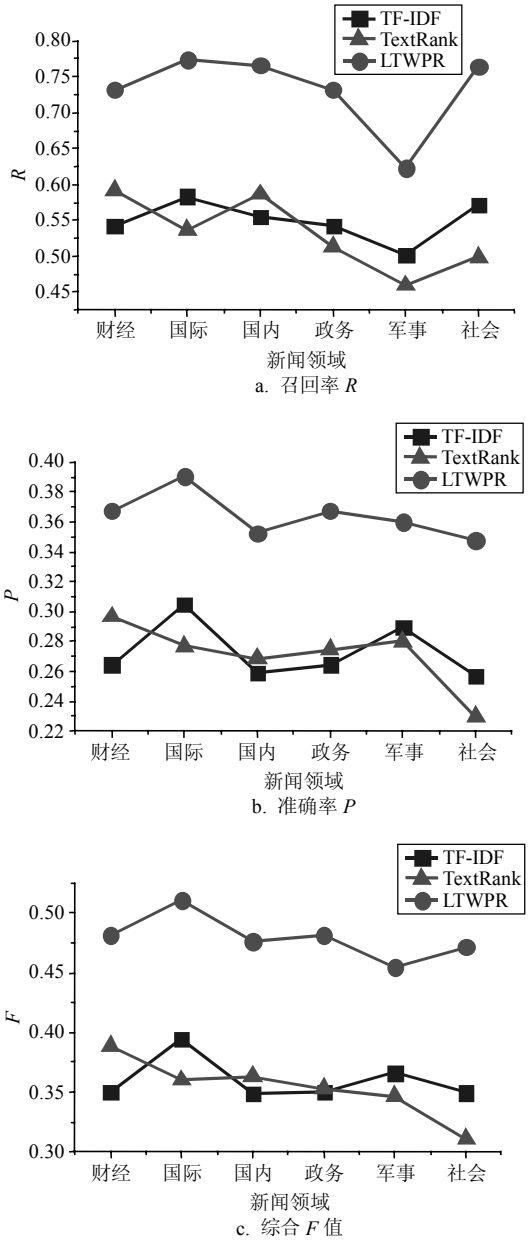


图6 提取关键词M+2时各项指标结果

3 结 束 语

本文的LTWPR算法基于PageRank算法，在考虑词频重要性的同时，融合了邻居节点对目标节点的重要度贡献，同时将人类的语言习惯列入考虑，对位于标题的词语赋予更高的语义权重系数，构造出一种新的新闻关键词提取算法。通过将爬取的新浪新闻语料抽象为文本复杂网络进行实验验证，利用召回率 R 、准确率 P 、和综合 F 值3个指标评价本文所构造算法的有效性。本文主要得出结论如下：

1) 将新闻语料按关键词提取个数分类进行的

实验结果表明，不管提取的关键词个数多或者少，本文LTWPR算法提取出的关键词能够更好地覆盖新闻作者标注的关键词，提取结果优于TF-IDF算法和TextRank算法，具有较高的有效性；

2) 将新闻语料按领域分类进行的实验结果表明，在财经、国际、国内、政务、军事、社会6个领域，本文LTWPR算法的关键词提取有效性更高，应用于新闻关键词的提取时实用性更强，且对于国际领域新闻的提取效果最优。

本文研究工作还存在一些不足。由于基于复杂网络的关键词提取方法仍依赖于分词软件进行分词才能构造文本网络，而构建文本网络的优劣将直接影响提取关键词的各项指标。本文算法在提取关键词时，提取结果也在一定程度上受分词软件分词准确率的影响。因此，在分词软件分词准确率受限的情况下，如何提高文本网络构建的完备性仍需进一步研究。

参 考 文 献

[1] SALTON G. Developments in automatic text retrieval[J]. Science, 1991, 253(5023): 974-979.

[2] 杨凯艳. 基于改进的TFIDF关键词自动提取算法研究[D]. 湖南, 湘潭: 湘潭大学, 2015.
YANG Kai-yan. Research on automatic keyword extraction algorithm based on improved TFIDF[D]. Xiangtan, Hunan: Xiangtan University, 2015.

[3] GUO A, YANG T. Research and improvement of feature words weight based on TFIDF algorithm[C]//Proceedings of the Information Technology, Networking, Electronic and Automation Control Conference(ITNEC 2016). Chongqing, China: IEEE, 2016: 415-419.

[4] MIHALCEA R, TARAU P. TextRank: Bringing order into texts[C]//Conference on Empirical Methods in Natural Language Processing, EMNLP 2004. Barcelona, Spain: [s.n.], 2004: 404-411.

[5] BRIN S, PAGE L. The anatomy of a large-scale hyper textual web search engine[C]//Proceedings of the 7th World Wide Web Conference (WWW7). Brisbane, Australia: [s.n.], 1998: 107-117.

[6] CANCHO R F I, SOLÉ R V. The small world of human language[J]. Proceedings Biological Sciences, 2001, 268(1482): 2261-2266.

[7] 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, 2011, 13(17): 217-223.

[8] 任晓龙, 吕琳媛. 网络重要节点排序方法综述[J]. 科学通报, 2014, 59(13): 1175-1197.
REN Xiao-long, LÜ Lin-yuan. Review of ranking nodes in complex networks[J]. Chin Sci Bull, 2014, 59(13): 1175-1197.

[9] 谢凤宏, 张大为, 黄丹, 等. 基于加权复杂网络的文本关键词提取[J]. 系统科学与数学, 2010, 30(11): 1592-1596.
XIE Feng-hong, ZHANG Da-wei, HUANG Dan, et al. Keywords extraction based on weighted complex network[J]. Journal of Systems Science and Mathematical Sciences, 2010, 30(11): 1592-1596.

[10] 唐俊. 复杂网络在新闻网页关键词提取中的应用[J]. 云南民族大学学报(自然科学版), 2012, 21(4): 305-308.
TANG Jun. Application of complex networks to keyword extraction of news web pages[J]. Journal of Yunnan Nationalities University: Natural Sciences Edition, 2012, 21(4): 305-308.

[11] 左晓飞. 基于复杂网络的关键词提取研究[D]. 西安: 西安电子科技大学, 2013.
ZUO Xiao-fei. Research on keyword extraction based on complex network[D]. Xian: XiDian University, 2013.

[12] CHEN Q, JIANG Z, BIAN J. Chinese keyword extraction using semantically weighted network[C]//International Conference on Intelligent Human-Machine Systems & Cybernetics. [S.l.]: IEEE, 2014: 83-86.

[13] NAN J, XIAO B, LIN Z, et al. Keywords extraction from Chinese document based on complex network theory[C]// 2014 Seventh International Symposium on Computational Intelligence and Design (ISCID). [S.l.]: IEEE, 2015: 383-386.

[14] 刘通. 基于复杂网络的文本关键词提取算法研究[J]. 计算机应用研究, 2016, 33(2): 365-369.
LIU Tong. Algorithm research of text key word extraction based on complex networks[J]. Application Research of Computers, 2016, 33(2): 365-369.

[15] 张华平. ICTCLAS汉语分词系统[EB/OL]. [2014-06-25]. http://ictclas.nlpir.org/.
ZHANG Hua-ping. ICTCLAS Chinese word segmentation system[EB/OL]. [2014-06-25]. http://ictclas.nlpir.org/.

编 辑 蒋 晓