ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 35,No. 3,May 2020,pp. 458－473
DOI:10. 16337/j. 1004-9037. 2020. 03. 009

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于统计特征的Quality Phrase挖掘方法

杨欢欢[1,2,3]，赵书良[1,2,3]，李文斌[4]，武永亮[5]，田国强[1,2,3]

(1. 河北师范大学计算机与网络空间安全学院，石家庄，050024；2. 河北师范大学河北省供应链大数据分析与数据安全工程研究中心，石家庄，050024；3. 河北师范大学河北省网络与信息安全重点实验室，石家庄，050024；4. 河北地质大学信息工程学院，石家庄，050031；5. 河北师范大学数学科学学院，石家庄，050024)

摘　要：Quality Phrase挖掘是从文本语料库中提取有意义短语的过程，是文档摘要、信息检索等任务的基础。然而现有的无监督短语挖掘方法存在候选短语质量不高、Quality Phrase的特征权重平均分配的问题。本文提出基于统计特征的Quality Phrase挖掘方法，将频繁*N*-Gram挖掘、多词短语组合性约束及单词短语拼写检查相结合，保证了候选短语的质量；引入公共知识库对候选短语添加类别标签，实现了Quality Phrase特征权重的分配，并考虑特征之间相互影响设置惩罚因子调整权重比例；按照候选短语的特征加权函数得分排序，提取Quality Phrase。实验结果表明，基于统计特征的Quality Phrase挖掘方法明显提高了短语挖掘的精度，与最优的无监督短语挖掘方法相比，精确率、召回率及F1-Score分别提升了5.97%，1.77%和4.02%。

关键词：文本挖掘；Quality Phrase；统计特征；候选短语；特征加权

中图分类号：TP391　　**文献标志码**：A

Quality Phrase Mining Method Based on Statistic Features

YANG Huanhuan[1,2,3], ZHAO Shuliang[1,2,3], LI Wenbin[4], WU Yongliang[5], TIAN Guoqiang[1,2,3]

(1. College of Computer and Cyber Security, Hebei Normal University, Shijiazhuang, 050024, China; 2. Hebei Provincial Engineering Research Center for Supply Chain Big Data Analytics & Data Security, Hebei Normal University, Shijiazhuang, 050024, China; 3. Key Laboratory of Network & Information Security, Hebei Normal University, Shijiazhuang, 050024, China; 4. College of Information Engineering, Hebei GEO University, Shijiazhuang, 050031, China; 5. School of Mathematical Sciences, Hebei Normal University, Shijiazhuang, 050024, China)

Abstract: Quality Phrase mining is a process of extracting meaningful phrases from text corpus, which is the basis of tasks such as document summary and information retrieval. However, the existing unsupervised phrase mining methods have problems of low quality of candidate phrases and average distribution of feature weight of Quality Phrase. Therefore, a Quality Phrase mining method based on statistic features is proposed. This method combines frequent *N*-Gram mining, combinatorial constraints of multi-word phrases, and spell checking to ensure the quality of candidate phrases. The public knowledge base is introduced to add labels to the candidate phrases, and the weight distribution of Quality Phrase is realized. The penalty factor is set to adjust the weight ratio considering the mutual influence between the features. The Quality Phrase is extracted according to the score of the feature weighting function of the

基金项目：国家社会科学基金重大(13&ZD091，18ZDA200)资助项目。
收稿日期：2019-09-19；**修订日期**：2019-12-11

candidate phrases. Experimental results show that the Quality Phrase mining method based on statistic features significantly improves the precision of phrase mining. Compared with the optimal unsupervised phrase mining methods, the precision, recall and F1-Score values are improved by 5.97%, 1.77%, and 4.02%, respectively.

Key words: text mining; Quality Phrase; statistic features; candidate phrases; feature weighting

引　言

互联网技术的发展带来了大量的非结构化文本数据,如新闻、日志及社交评论等。如何将文本数据表示成计算机可以处理的形式是文本挖掘任务的首要问题。词袋模型是经典的文本表示方法,但是它忽略了词出现的顺序,不能表达完整的语义并且存在维数灾难的问题。因此,近几年的研究从词语粒度转变为短语粒度,实现了对文本更好的表示。从文档中挖掘短语不仅可以帮助人们迅速准确地理解文档主旨,还可以应用到主题建模、文档摘要和信息检索等任务中,具有重要的研究意义。

短语挖掘起源于自然语言处理中的自动术语识别,后来出现了基于语法规则的短语挖掘,Li等[1]针对Twitter数据提出基于分段的词性标注分割框架HybridSeg。Shang等[2]从公共知识库中获得Quality Phrase,构建了一种自动化短语挖掘框架AutoPhrase,并开发POS(Part of speech)引导的短语分割模型进一步提高性能。Lin等[3]基于概念知识树建立双宾短语的语义表达模型,使计算机能够理解并处理双宾短语。但是,基于语法规则的短语挖掘不容易迁移到其他语言,不适合分析与语法无关的文本数据,因此实现了基于数据驱动的短语挖掘。Noraset等[4]利用边缘估计来改进递归神经网络语言模型,通过训练边缘值匹配语料库中的N-Gram概率。Chi等[5]旨在挖掘具有区别力的、非冗余的文本最长闭频繁序列模式。然而,基于数据的方法需要设定阈值,超过阈值认为符合短语,可想这种方式会丢失小于阈值但有意义的短语,为了改进提出基于统计的短语挖掘方法。该方法的思想是计算候选短语得分,从大到小排序来提高短语质量。Parameswaran等[6]将从数据集中提取概念的问题视为超市购物篮问题,采用支持度和置信度作为统计指标来抽取概念。El-Kishky等[7]提出根据索引找到大于最小支持度阈值的短语作为候选,自底向上地按照Sig得分进行合并,最后将一篇文档表示成短语袋的形式。Nedelina等[8]改进Topical PageRank算法,将主题特异性和语料库特异性作为短语显著性得分指标,按照排名顺序提取短语。

最新研究关注关键短语的生成与提取。文献[9-11]基于Seq2Seq架构,将语言的相关约束融入到关键短语生成中。文献[12-14]研究得出神经关键短语的生成方法。Debanjan等[15]使用短语嵌入提取关键短语。Florian[16]实现了基于多部图的无监督关键短语提取。Zhang等[17]将人的注意力集成到关键短语提取模型中,有效改进了Twitter数据集。

虽然有学者致力于关键短语的研究,但关键短语个数较少,不能很好地表示文档。因此,本文以Quality Phrase为挖掘目标,针对基于统计的短语挖掘存在候选短语质量不高、Quality Phrase特征权重分配不恰当的问题,提出基于统计特征的Quality Phrase挖掘方法(Quality Phrase mining method based on statistic features,QPMSF)。该方法将频繁N-Gram挖掘、多词短语的组合性约束和单词短语的拼写检查相结合,保证了候选短语的质量;引入公共知识库对候选短语添加类别标签,实现了根据特征对Quality Phrase贡献程度的不同来分配相应的权重;最后按照候选短语的特征加权函数得分排序,提取Quality Phrase。

1　相关定义

定义1　短语。短语P是由文本文档d中从第i个单词位置开始、长度为l的不间断单词序列组成,

即$P=d[i,i+l-1]=w_i,w_{i+1},\cdots,w_{i+l-1}$。短语是介于单词和句子之间的语言单位。

定义2 Quality Phrase。文本文档d中能够表达特定的主题或含义,可以作为完整语义单元出现的短语称为Quality Phrase。

定义3 Quality Phrase挖掘。指从任意长度、任意形式的文本语料库C(如商业评论语料库、论文摘要语料库、新闻语料库等)中,提取每篇文本文档内出现的短语,对短语赋予一定的质量得分,按照得分由高到低进行排序,选择得分较大的前m个短语组成Quality Phrase的集合$QP=\{QP_1,\cdots,QP_m\}$,即为Quality Phrase挖掘的结果。Quality Phrase挖掘可以作为主题建模、文本分类以及信息检索等文本任务的基础。

定义4 点互信息PMI(Pointwise mutual information)。假设离散随机变量P_i,P_j相互独立,PMI量化了联合概率$p(P_i,P_j)$与个体分布$p(P_i),p(P_j)$之间的差异,衡量了两个随机变量之间的相关性,具体表示为

$$\mathrm{PMI}(P_i;P_j)=\log\frac{p(P_i,P_j)}{p(P_i)p(P_j)}=\log\frac{p(P_i|P_j)}{p(P_i)}=\log\frac{p(P_j|P_i)}{p(P_j)} \tag{1}$$

2 Quality Phrase评价准则

在短语挖掘 领域,“Quality Phrase”还没有一个通用的评价标准。本文根据文献[18-21]的研究,总结得出较为客观的短语质量评价准则:频繁性、组合性、信息性和完整性,用于后续的Quality Phrase挖掘方法。下面依次阐述各准则的具体含义,并给出数学公式来量化准则。

准则1 频繁性。如果短语P在文本语料库C中总共出现的次数$f(P)$大于某个特定的阈值f_t,则短语P满足频繁性。阈值f_t根据语料库的规模、内容不断变化。

从统计的角度来说,高频短语具有相对重要的作用,低频短语可能是不重要的。例如,一篇科技论文中多次写到“deep learning”,那么该短语是Quality Phrase的可能性极大。

准则2 组合性。短语P满足组合性准则,当且仅当存在一组P_i,P_j,使得$P=P_i\oplus P_j$并且$\mathrm{function}(P_i,P_j)=\mathrm{true}$。其中,$\oplus$指短语$P$的词序列可以由$P_i,P_j$两部分组成,function是某个特定统计意义度量(互信息、卡方检验、T检验等)下的布尔函数,决定P_i,P_j能否合并。

通俗地讲,组合性指如果两个相邻的单词或短语共同出现比单独出现更有意义,则这两个相邻的单词或短语可以合并成一个新的短语。例如,New和York共同出现比York单独出现的频率更高,表达的意义更完整,因此在进行Quality Phrase挖掘时,将New York识别为具有组合性准则的短语。

本文利用信息论中的互信息相关理论来表示组合性的布尔函数function。互信息可以度量一个随机变量中包含另一个随机变量的信息量。由互信息衍生出来的“点互信息”这一概念更适合应用到组合性准则的度量中。由定义4可知,等式的最后一项可以理解为在P_i出现的情况下P_j出现的概率除以P_j单独出现的概率。比值越大,说明P_i,P_j共同出现更有意义,可以合并,短语P满足组合性。其中$p(x)$的计算方式为x在语料库中出现的次数与所有频繁短语出现总次数的比值。

$$p(x)=\frac{f(x)}{\sum\limits_{x'\in Pop}f(x')} \tag{2}$$

式中Pop表示文本语料库中所有符合准则1的频繁短语的集合。因此,组合性的布尔函数表示为

$$\mathrm{function}(P_i,P_j)=\begin{cases}\mathrm{true} & \mathrm{PMI}(P_i;P_j)\geqslant\alpha\\ \mathrm{false} & \text{其他}\end{cases} \tag{3}$$

准则3 信息性。若短语P在文本文档d中能够表达特定的主题或准确的概念,则称该短语满足信息性准则。例如,一篇科技论文中,“text classification”这一短语要比“the paper”具有更好的信息性。

本文根据逆文档频率这一统计值来量化信息性的概念。短语P的逆文档频率可以由语料库中文本文档总数与包含该短语的文件数目比值的对数来表示

$$\mathrm{IDF}(P)=\log\frac{|D|}{|\{d\in D:P\in d\}|} \tag{4}$$

准则4 完整性。在文本文档d中,短语$P=d[i,i+l-1]=w_i,w_{i+1},\cdots,w_{i+l-1}$满足完整性,当且仅当不存在短语$Q=d[j,j+k-1]=w_j,w_{j+1},\cdots,w_{j+k-1}$,使得$[j,j+k-1]$是$[i,i+l-1]$的子集。

完整性准则要求短语能够表达一个完整的语义单元。例如,文本文档中同时存在“Naive Bayesian Model”“Naive Bayesian”“Bayesian Model”3个短语。那么,该准则认为“Naive Bayesian Model”是一个完整的短语或者说具有较高的完整性得分,“Naive Bayesian”和“Bayesian Model”不属于完整性短语或具有较低的得分,即短语的子集不是完整的表达。当然,如果文档中出现“Naive Bayesian”,同时没有比它更完整的短语,那么“Naive Bayesian”被认为满足完整性。

本文将完整性准则看作条件概率的问题。假设短语P_{sub}存在比它更完整的短语P_{com},那么条件概率$p(P_{\mathrm{com}}|P_{\mathrm{sub}})$表示在短语$P_{\mathrm{sub}}$出现的情况下$P_{\mathrm{com}}$出现的概率。条件概率越大,短语$P_{\mathrm{sub}}$越不完整。由条件概率公式可得,$p(P_{\mathrm{com}}|P_{\mathrm{sub}})=\dfrac{p(P_{\mathrm{com}}\cap P_{\mathrm{sub}})}{p(P_{\mathrm{sub}})}$,分子部分为$P_{\mathrm{com}}$和$P_{\mathrm{sub}}$同时出现的概率,本质是$P_{\mathrm{com}}$出现的概率,分母部分则表示$P_{\mathrm{sub}}$出现的概率,即公式可以简化为$p(P_{\mathrm{com}}|P_{\mathrm{sub}})=\dfrac{f(P_{\mathrm{com}})}{f(P_{\mathrm{sub}})}$。再者,将式子取倒数$\dfrac{1}{p(P_{\mathrm{com}}|P_{\mathrm{sub}})}=\dfrac{f(P_{\mathrm{sub}})}{f(P_{\mathrm{com}})}$的含义是比值越大,短语$P_{\mathrm{sub}}$越完整。

当然,短语P_{sub}的超短语可能不止一个,假设超短语共有n个,则完整性概率公式为

$$P_{\mathrm{sub}}=\frac{f(P_{\mathrm{com}_1})}{\sum_{i=1}^{n}f(P_{\mathrm{com}_i})}\times\frac{f(P_{\mathrm{sub}})}{f(P_{\mathrm{com}_1})}+\cdots+\frac{f(P_{\mathrm{com}_n})}{\sum_{i=1}^{n}f(P_{\mathrm{com}_i})}\times\frac{f(P_{\mathrm{sub}})}{f(P_{\mathrm{com}_n})} \tag{5}$$

每个条件概率的权值是该超短语出现的次数除以所有超短语出现的总次数。

Quality Phrase挖掘的目的就是从文本语料库中提取符合上述4个准则的短语,将无结构化的文本数据转化为有意义的模式。

3 基于统计特征的候选短语挖掘方法

面对无结构的文本语料库,如何提取短语是需要解决的首要问题。自然语言处理中一种经典方法是N-Gram语言模型。它的基本思想是将文本文档的内容按照单词顺序进行大小为N的窗口滑动操作,形成很多个长度为N的短语。例如,用N-Gram对“Transactions on Knowledge and Data Engineering”进行短语提取的结果如图1所示(以$N=1,2,3$为例)。该模型的优点是不会丢失任何一个短语,但产生的代价巨大。随着语料库中每篇文本文档字数的增加,将会产生指数级的短语数量,给短语挖掘带来困难。因此,本文将频繁性准则融合到N-Gram的提取过程中。

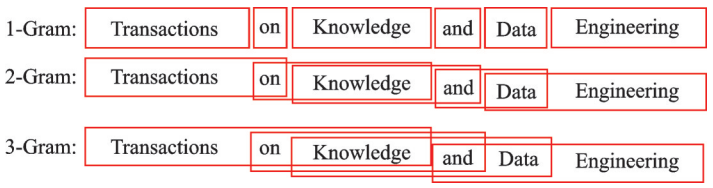


图1 N-Gram语言模型示例

Fig.1 Example of the N-Gram language model

3.1 频繁*N*-Gram短语挖掘

将频繁性准则应用到*N*-Gram短语提取过程中,可以过滤掉文本语料库中频数小于阈值f_t的词语序列,有效减少短语的数量。本文根据短语的索引位置提出基于索引信息的频繁*N*-Gram挖掘算法。其中用到2个键值对字典表:第1个为索引字典表,将短语字符串作为键,短语出现的所有位置的列表作为值;第2个为频数字典表,短语字符串为键,该短语在文本语料库中出现的频数作为值。代码如算法1所示。

算法1 基于索引信息的频繁*N*-Gram挖掘算法

输入:文本语料库C,频繁性阈值f_t,最大短语长度ω

输出:短语频数字典表frequence

```
1. index = null /*初始化索引字典表*/
2. frequence = null /*初始化频数字典表*/
3. /*将语料库中单词的索引保存到字典表中*/
4. for i = 1 to |C| do
5.     for j = 1 to |C[i]| do
6.         index[key] = index[key] ∪ [i,j]
7.     end for
8.   end for
9. /*按照长度由小到大的迭代顺序处理短语*/
10. for len = 1 to ω do
11.   for key in index do
12.     if |index[key]| ≥ f_t then
13.       frequence[key] = |index[key]| /*频繁短语*/
14.       /*将超短语的索引保存到临时字典表*/
15.       for [m,n] ∈ index[key] do
16.         newkey = key ⊕ C[m,n+1]
17.         index′[newkey] = index′[newkey] ∪ [m,n+1]
18.       end for
19.     end if
20.   end for
21.   index = null
22.   index = index′
23.   index′ = null
24. end for
25. return frequence
```

算法1以数据预处理后的文本语料库为基础,第1~2步初始化字典表;第3~8步使用双层循环将文本语料库中所有单词的索引信息保存在索引字典表中;第9~24步的最外层循环根据短语的长度从小到大依次迭代,目的是若长度为len的短语不满足频繁性,则以它开始的长度为$len+1$的短语必定不频繁,可以按照短语长度排除该短语的所有超短语,减少算法的时间消耗。算法1中ω是一个很小的输入常数,不会对时间复杂度产生影响。对于满足频繁性准则的短语,第12~19步表示将频繁短语放入频数字典表,并将该短语的超短语(长度增加1)索引信息保存到临时索引字典表,待当前长度的所有短

语判断完毕后,进入下一轮迭代。

3.2 多词短语组合性约束

3.1节挖掘得到的短语满足频繁性准则,在此过程中,频繁性阈值的大小起到了决定性作用。当频繁性阈值设置较大时,从文本语料库提取的短语数量相对较少、质量很高,但是这种方式会过滤掉一些出现次数不多却对文章主旨有意义的短语;若频繁性阈值设置较小,可以保证不遗漏出现频数少却有意义的短语,但随之带来的缺点是短语数量增多,为后续的Quality Phrase挖掘带来时间上的消耗。为解决频繁性阈值带来的问题,本文在候选短语挖掘阶段加入组合性约束,在低频短语的基础上,需满足统计意义度量函数,既减少了低频短语的数量,也确保了候选短语的质量。

由组合性准则可知,判断短语是否符合组合性,关键是能否找到一个有效的分割位置,使得左右两个子短语的统计意义得分为真。寻找最优分割位置的方式根据自底向上、逐层迭代的思想,按照统计意义得分最大的原则,对每一层产生的两个相邻单元(可能是单字词语或多词短语)进行合并,最外层统计意义得分的结果决定该短语是否满足组合性。迭代过程中小于组合性阈值继续执行。

下面以"Automated Phrase Mining from Text Corpora"为例,详细说明组合性准则的评判过程。算法初始时,假设组合性阈值$\alpha=3$,短语以空格为分隔,每个单词自然形成一个单元。迭代的第一步计算任意两个相邻单元(Automated和Phrase,Phrase和Mining,Mining和from,from和Text,Text和Corpora)的PMI值,选取最大值Text和Corpora,将2个单词合并到1个单元。在此基础上进行第2层的迭代,计算相邻单元(Automated和Phrase,Phrase和Mining,Mining和from,from和Text Corpora)的PMI值,此时计算结果显示Phrase和Mining的紧密程度最大,将Phrase和Mining合并到同一个单元。第3层的迭代共4个单元、3组候选组合,按照最大值原则合并的单元为from和Text Corpora。第4层迭代得出Phrase Mining和from Text Corpora之间的统计意义得分高于Automated和Phrase Mining。最后,计算单元Automated和单元Phrase Mining from Text Corpora之间的PMI值为3.3,高于设定的阈值α,因此该短语满足组合性准则。判断过程中,保存相邻单元的PMI值,避免重复计算。

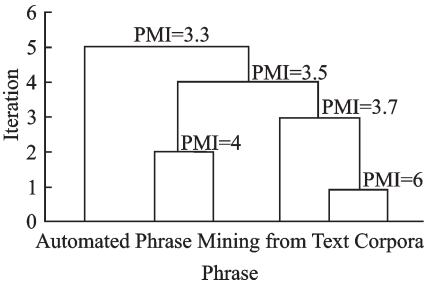


图2 短语的组合性判断示例

Fig.2 Example of combinatorial judgment of phrase

3.3 单词短语拼写检查

组合性准则适用于2个及以上单词组成的短语,然而单词的质量也会影响候选短语挖掘的精度。为改进暴力算法对单词进行拼写检查的缺陷,本文引入Trie单词查找树结构,利用字符串的公共前缀来减少查询时间,最大限度地减少字符串的比较次数,同时提升挖掘准确率。

Trie单词查找树具有3个基本性质:(1)根节点不包含字符,除根节点外的每一个节点都只包含1个字符;(2)从根节点到某一节点,路径上经过的字符连接起来为该节点对应的字符串;(3)每个节点的所有子节点包含的字符都不相同。图3是以adj,adv,auto,but,bus,from和feel为例构建的Trie单词查找树。候选短语挖掘的单词短语拼写检查过程中,先构建单词表的Trie结构,然后检查频繁1-Gram是否在Trie中出现,若出现,则加入候选短语列表。

综合频繁N-Gram挖掘、多词短语的组合性约束和单词短语的拼写检查,提出基于统计特征的候选短语挖掘方法。具体代码实现如下。

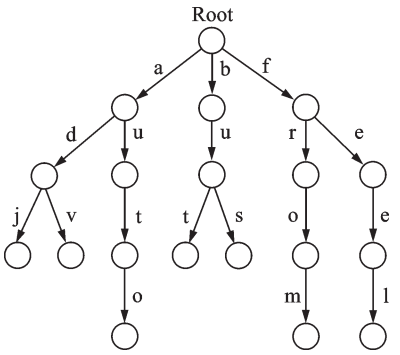


图3 Trie单词查找树结构

Fig.3 Structure of Trie tree

算法2 基于统计特征的候选短语挖掘方法

输入: 短语频数字典表 frequence,单词表 vocabulary

输出: 候选短语字典表 candidate

```
1.for word in vocabulary do
2. insertTrie( word )/*构建 Trie 单词查找树*/
3.end for
4.for key in frequence do
5. if |key|= 1 then
6.   if trieIsExist( key )= true then /*拼写检查*/
7.    candidate [ key ]= α
8.   end if
9. else if |key|= 2 then/*计算长度为 2 短语的 PMI*/
10.   if PMI( key [ 1 ]; key [ 2 ])≥ α then
11.    candidate [ key ]= PMI( key [ 1 ]; key [ 2 ])
12.   end if
13. else/*长度大于 2 的短语执行图 2 的操作*/
14.   length = |key|
15.   while length > 2 do
16.    for i = 1 to length − 1 do
17.     /*寻找两个相邻单元的 PMI 最大值*/
18.     if PMI( key [ i ]; key [ i + 1 ])> PMI( key [ i + 1 ]; key [ i + 2 ]) then
19.      loc = i
20.     else
21.      loc = i + 1
22.     end if
23.     key [ loc ]= key [ loc ]+ key [ loc + 1 ]/*合并单元*/
24.     length = length − 1
25.    end for
26.   end while
27.   /*计算最外层 2 个单元的 PMI 值,若大于阈值则为候选短语*/
28.   if PMI( key [ 1 ]; key [ 2 ])≥ α then
29.    candidate [ key ]= PMI( key [ 1 ]; key [ 2 ])
30.   end if
31. end if
32.end for
```

将频繁 *N*-Gram 作为算法 2 的输入,第 1~3 步结合外部字典表构建 Trie 单词查找树;从第 4 步开始,对频繁短语进行逐个判断。若短语长度为 1,执行第 5~8 步,通过拼写检查的单词放入候选短语列表;若长度为 2,执行第 9~12 步,计算左右两个单词的 PMI 是否大于 α,若大于阈值则放入候选短语列表。当短语长度大于 2 时,执行第 13 步的条件分支。如图 2 的判断过程,短语长度决定第 15 步的迭代次数,每次迭代先找到相邻两个单元 PMI 的最大值进行合并(第 16~25 步),当只剩下 2 个单元时,执行第

28～30步计算PMI值,若大于阈值α加入候选短语列表。

4 基于统计特征的Quality Phrase选择方法

在上述候选短语挖掘的基础上,计算频繁性、组合性、信息性和完整性4个准则的值作为候选短语的4个特征,根据4个特征对Quality Phrase的贡献程度得到对应的权重,考虑特征之间的相互影响引入惩罚因子调整权重分配。最后按照候选短语的特征加权函数得分排序,选择Quality Phrase。

4.1 特征对Quality Phrase的贡献程度

在Quality Phrase挖掘中,特征加权的目的是得到频繁性、组合性、信息性和完整性4个特征对Quality Phrase贡献程度的大小,使得对Quality Phrase影响较大的特征占较大的权重,对Quality Phrase影响较小的特征占较小的权重,按照特征加权函数得分进行排序,提高短语挖掘的性能。

本文根据类别信息计算权重,具体思路如下:

(1) 计算每个候选短语的频数,PMI,IDF,P_{sub},标记候选短语的类别(以公共语料库维基百科为基准,若候选短语在维基百科中出现,则标记1;否则标记为0),得到特征矩阵,矩阵的每行代表一个候选短语,5列分别代表频次,PMI,IDF,P_{sub}和类别。

$$P=\begin{bmatrix} \text{frequence}_1 & \text{PMI}_1 & \text{IDF}_1 & P_{\text{sub}_1} & c_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \text{frequence}_n & \text{PMI}_n & \text{IDF}_n & P_{\text{sub}_n} & c_n \end{bmatrix} \tag{6}$$

(2) 定义如下参数:

ta:类别为1的候选短语在当前特征维度的特征值之和;

tb:类别为1的候选短语在除当前特征之外的其他特征维度上的特征值之和;

tc:类别为0的候选短语在当前特征维度的特征值之和;

td:类别为0的候选短语在除当前特征之外的其他特征维度上的特征值之和。

根据参数,本文提出"特征贡献程度"的概念表示4个特征对Quality Phrase的影响程度,表达式为

$$\text{contribution}=\frac{ta}{ta+tc} \tag{7}$$

特征对类别的区分能力直接取决于类别为1的候选短语在当前特征维度的特征值之和与所有候选短语在当前特征维度的特征值之和,比值越大,当前特征属于类别1的概率越大。

(3) 对4个特征的贡献程度进行归一化处理,得到候选短语的特征加权函数

$$sig=\omega_1*\text{frequence}+\omega_2*\text{PMI}+\omega_3*\text{IDF}+\omega_4*P_{\text{sub}} \tag{8}$$

4.2 特征之间相互影响

考虑到特征之间相互影响存在冗余的情况,本节采用皮尔逊相关系数度量2个特征之间的相关程度,加入惩罚因子表示特征的冗余量对Quality Phrase造成的负面影响,改进特征贡献程度,完善候选短语的特征加权函数,提取Quality Phrase。

两个特征之间的皮尔逊相关系数为

$$\rho_{i,j}=\frac{\text{cov}(f_i,f_j)}{\delta_{f_i}\delta_{f_j}} \tag{9}$$

式中:f_i,f_j表示两个特征;$\text{cov}(f_i,f_j)$为两个特征的协方差;δ_{f_i}和δ_{f_j}分别为标准差。皮尔逊相关系数越大表明特征之间存在的冗余越多。由于本文只考虑特征之间的影响程度,与正负无关,因此取皮尔逊相关系数的绝对值$|\rho_{i,j}|$作为度量指标。那么,特征的冗余度计算方式为该特征与特征空间F中其他3个特征的皮尔逊相关系数的绝对值之和的平均值,表达式为

$$\rho_i' = \frac{1}{3}\sum_{j\in F, j\neq i}|\rho_{i,j}| \tag{10}$$

式中ρ_i'的取值分为以下几种情况：当$\rho_i'=0$时，特征f_i没有冗余，即该特征的贡献程度不需要减弱；当$\rho_i'\neq 0$时，值越大表示该特征越冗余，需要调整特征f_i的贡献程度；当$\rho_i'=1$时，表明该特征可以被其他特征所替代，为减少冗余可取消该特征。因此，引入惩罚因子β_i来改进特征之间的相互影响

$$\beta_i = \begin{cases} 0 & \rho_i'=1 \\ \dfrac{1}{1+3*\rho_i'} & \text{其他} \end{cases} \tag{11}$$

改进后，特征的权重为

$$\omega_i' = \beta_i * \omega_i \tag{12}$$

最终，候选短语的特征加权函数为

$$sig = \omega_1'*\text{frequence} + \omega_2'*\text{PMI} + \omega_3'*\text{IDF} + \omega_4'*P_{\text{sub}} \tag{13}$$

结合4.1节特征对Quality Phrase的贡献程度和4.2节特征之间相互影响，本文提出基于统计特征的Quality Phrase选择方法，算法如下。

算法3 基于统计特征的Quality Phrase选择方法

输入：短语频数字典表frequence，候选短语字典表candidate，维基百科实体表Wiki-entity，Quality Phrase的个数rankid

输出：Quality Phrase列表Quality Phrase List

1.for $i=1$ to 4 do/*计算特征贡献程度*/

2. for count = 1 to |candidate| do

3. if candidate [count] in wiki-entity then

4. $ta += P$ [count][f_i]

5. else

6. $tc += P$ [count][f_i]

7. end if

8. end for

9. contribution$_i = ta/(ta+tc)$

10.end for

11.for $i=1$ to 4 do/*根据惩罚因子调整权值*/

12. $\omega_i = \text{contribution}_i / \sum_{j=1}^{4}\text{contribution}_j$

13. for $j=1, j\neq i$ to 4 do

14. $\rho_{i,j} = \text{cov}(f_i, f_j)/\delta_{f_i}\delta_{f_j}$

15. $\rho_i' += \rho_{i,j}$

16. end for

17. $\rho_i' = \rho_i'/3$

18. if $\rho_i'=1$ then

19. $\beta_i = 0$

20. else

21. $\beta_i = 1/(1+3*\rho_i')$

22. end if

23. $\omega_i' = \beta_i * \omega_i$

24.end for

25./*计算候选短语的特征加权函数得分*/

26.for count = 1 to |candidate| do

27. $sig = \omega_1' * \text{frequence} + \omega_2' * \text{PMI} + \omega_3' * \text{IDF} + \omega_4' * P_{\text{sub}}$

28.end for

29./*按得分排序,选择前rankid的候选短语*/

30.if sort(sig)$\geqslant$ rankid then

31. return qualityPhraseList

32.end if

在算法3中,第1~10步根据式(7)的定义计算频繁性、组合性、信息性和完整性4个特征对Quality Phrase的贡献程度。第11~24步根据式(9)—(12)加入惩罚因子去除特征之间的冗余,调整特征的权重比例。第25~28步根据式(13)计算每个候选短语的特征加权函数得分,第30步对函数得分排序,选择排名前rankid的候选短语作为Quality Phrase输出。将第3节的候选短语挖掘和第4节的特征加权方法合并,即为本文提出的基于统计特征的Quality Phrase挖掘方法,方法框架如图4所示。

基于统计特征的Quality Phrase挖掘方法框架分为4部分:第1部分实现对文本语料库的预处理工作,包括去特殊字符、去停用词以及提取词元信息等;第2部分融合频繁N-Gram短语挖掘、多词短语组合性约束和单词短语拼写检查来实现候选短语挖掘;第3部分的主要目的是得到频繁性、组合性、信息性、完整性的特征权重,使得对Quality Phrase贡献较大的特征占较大的比重,对Quality Phrase贡献小的特征占较小的权重;第4部分根据候选短语的特征加权函数得分排序,提取排名靠前的短语作为Quality Phrase。

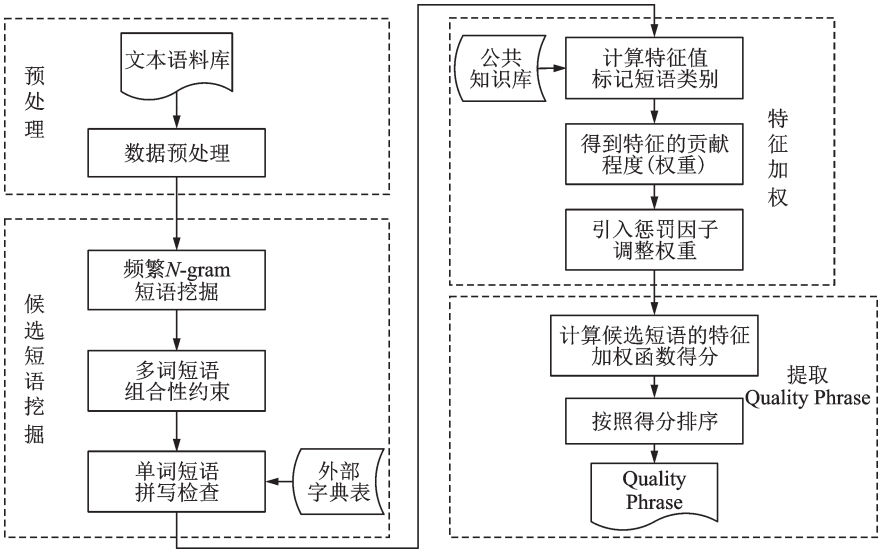


图4 基于统计特征的Quality Phrase挖掘方法框架

Fig.4 Framework of Quality Phrase mining method based on statistic features

5 实验验证

本节对基于统计特征的Quality Phrase挖掘方法进行实验分析，并与其他短语挖掘方法做比较。本文所有实验环境：操作系统为Windows 10专业版（64位操作系统）、处理器为Intel(R) Core(TM) i7-2600 @ 3.40 GHz、内存8 GB、硬盘1 TB，实验算法使用Java语言编写，开发工具为Eclipse Luna Service Release2，jdk1.8。

5.1 数据集

本文选择5个真实数据集作为实验的文本语料库：(1)5Conf包含AI，DB，DM，IR，ML五个领域科技论文的标题文本信息；(2)DBLP Abstracts收集了计算机类文章的摘要信息；(3)AP News是TREC 1998年的新闻文本数据集；(4)AMiner-Titles将AMiner-Paper[22]（研究学术信息网络的数据集）中的标题信息抽取出来；(5)AMiner-Abstracts为从AMiner-Paper中抽取出来的摘要语料库。对5个文本语料库进行数据预处理，去除特殊符号、去除停用词、提取词元信息后的数据集基本情况如表1所示。

表1 数据集总体情况

Table 1 Data set in general

数据集	文档数目	文档长度范围
5Conf	20 000	4～24
DBLP Abstracts	5 209	57～900
AP News	1 795	26～571
AMiner-Titles	19 229	3～27
AMiner-Abstracts	4 787	105～1 962

5.2 对比算法

为证明本文方法的有效性以及采用本文方法挖掘的Quality Phrase精度最高，将本文方法与现有的无监督短语挖掘算法进行比较。

无监督短语挖掘算法分为3大类[23]：基于统计的、基于网络图的和基于主题的。由于QPMSF框架总体趋向于无监督的方式，因此对比算法选取3类无监督短语挖掘的代表性算法。TF-IDF属于基于统计的短语挖掘算法，TextRank是基于网络图的代表，TopMine是短语挖掘领域性能很好的基于主题的方法。另外，选择最新的关键短语提取算法ParaNet＋CoAtt和最新的主题短语挖掘方法CQMine进行对比。

(1)TF-IDF是经典的短语挖掘算法，根据短语的频率以及逆文档频率提取Quality Phrase。

(2)TextRank[24]由PageRank网页重要性排序算法衍生而来，它的基本思想是如果一个单词出现在很多单词后面，说明这个单词比较重要。

(3)TopMine采用自底向上的方式将1篇文档分割成单词短语或多词短语，并应用于文档主题生成模型。

(4)ParaNet＋CoAtt将关键短语的语言约束集成到Seq2Seq网络，通过覆盖机制减少重叠短语的产生。

(5)CQMine是一种高效的主题短语挖掘方法，运用动态规划和分块的思想提取短语。

5.3 评价指标

目前短语挖掘领域判定一个短语是否为Quality Phrase有两种方式：(1)以维基百科实体为基准；(2)通过领域专家进行评价。本文选择第1种方式，若算法挖掘的短语能够在维基百科中找到一条对应的实体，则认为该短语是真正的Quality Phrase。由这种方式产生的混淆矩阵如表2所示。根据混淆矩

阵计算准确率(Accuracy)、精确率(Precision)、召回率(Recall)和F1-Score,作为评价算法优劣的指标。

表2 混淆矩阵

Table 2 Confusion matrix

预测结果	真实结果	
	维基百科实体	不是维基百科实体
算法挖掘的Quality Phrase	True positive(TP)	False positive(FP)
算法判定不是Quality Phrase	False negative(FN)	True negative(TN)

5.4 实验结果

5.4.1 组合性统计意义度量选择

组合性准则应用于候选短语挖掘的多词短语组合性检验过程,是候选短语挖掘乃至Quality Phrase挖掘的重要环节。其中,统计意义度量函数的选择直接决定了多词短语的质量。实验通过对比统计学中的点互信息PMI、卡方检验CHI、T检验3种度量方式,选择效果最好的函数作为判断多词短语能否组合的依据。由于3种度量函数在5个语料库上的模型准确率均能达到99%以上,没有明显差异,不再给出图表对比。图5用精确率和召回率的综合指标F1-Score来展示CHI,T-test和PMI的对比结果。

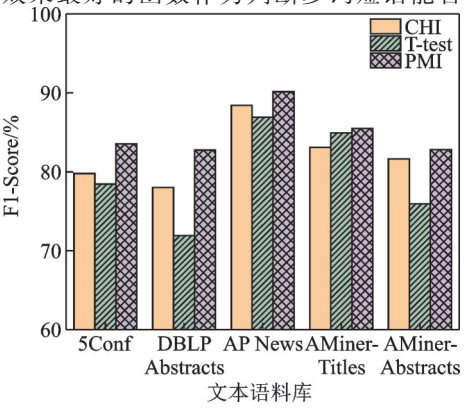


图5 组合性统计意义度量方式的对比

Fig.5 Comparison of combinatorial statistical significance measures

图5中的数据显示,应用组合性准则挖掘5个不同文本语料库的候选短语,F1-Score的范围为70%~90%,能够很好过滤低质量的短语。点互信息PMI在5个文本语料库上的表现最好,相比于卡方检验CHI平均提升2.96%,比T-test平均提升5.53%。因此,本文在候选短语挖掘阶段选择点互信息PMI作为组合性统计度量函数,保证多词短语的精度最高。

5.4.2 候选短语挖掘阶段实验结果对比

在候选短语挖掘阶段,本文综合了频繁*N*-Gram挖掘、多词短语的组合性约束和单词短语的拼写检查3种思想,提出基于统计特征的候选短语挖掘方法。为确保每一个步骤对候选短语挖掘的有效性,本文在各个环节都进行了实验验证。通过不断变换频繁性阈值,5个文本语料库的Precision-Recall曲线如图6所示。由于实验所用的5个文本语料库的规模不同,所以频繁性阈值的变化区间也不相同。

从5个文本语料库呈现的PR曲线可以看出,单使用频繁性一个准则进行*N*-Gram短语挖掘时,精确率和召回率的结果受频繁性阈值的影响极大,鲁棒性差。若设置的频繁性阈值较大,精确率可以达到较高的水平,但是召回率很低;相反,频繁性阈值设置较小时,召回率很高,可以达到90%以上,但是精确率明显下降。所以,单使用频繁性一个准则,很难达到精确率和召回率的平衡。为解决频繁性准则带来的问题,加入多词短语的组合性约束,两者的结合可以使5个数据集上的精确率稳定在一个较好的区间,基本维持在90%~100%,此时保证在精确率不明显下降的情况下,不断调节频繁性阈值来提升召回率,寻找最优的参数组合。在频繁*N*-Gram挖掘和多词短语组合性约束的基础上,增加单词短语

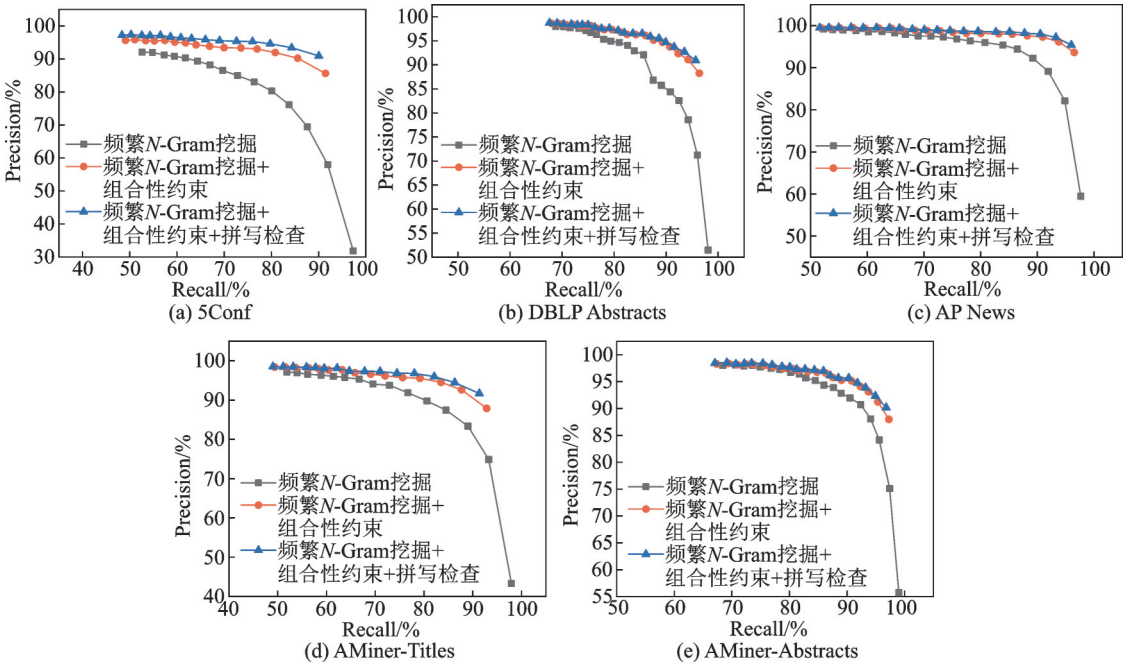


图6 候选短语挖掘阶段精确率-召回率曲线

Fig.6 Precision-Recall curves of candidate phrase mining phase

的拼写检查可以进一步提高精确率,虽然拼写检查错误地排除了一部分真正的单词,导致召回率有所下降,但综合性指标F1-Score仍然是上升趋势。下面将5个文本语料库在3种递进组合(方法①表示频繁*N*-Gram挖掘,方法②表示频繁*N*-Gram挖掘+多词短语组合性约束,方法③表示频繁*N*-Gram挖掘+多词短语组合性约束+单词短语拼写检查)上的最优结果列举如表3所示。

表3数据显示,方法③的准确率与方法①②基本持平,最多提高0.07%。由于精确率和召回率相互矛盾,所以方法③只能保证在一方面表现最好的情况下,另一方面不明显降低。因此以精确率和召回率的综合指标F1-Score为考量,方法③比方法①提高2.06%~10.31%,平均提高5.45%。方法③比方法②提高0.17%~2.05%,平均提高0.93%。可见,候选短语挖掘将频繁*N*-Gram挖掘、组合性约束和拼写检查相结合的思路正确,很大程度上提高了候选短语的质量。

5.4.3 本文算法与其他算法的对比

短语挖掘领域已有多种算法,本文将TF-IDF,TextRank,TopMine,ParaNet+CoAtt,CQMine与本文QPMSF方法对比。各方法在文本语料库上的F1-Score分布情况如图7所示。

对比发现,ParaNet+CoAtt算法的F1-Score最低,原因是虽然提取的关键短语精确度很高,但同时排除了大量的Quality Phrase,导致召回率下降,整体的F1-Score只能维持在29.6%~36.96%。其次,TextRank表现最差,在5个语料库上的平均F1-Score为60.35%,需要进一步提高。 TF-IDF和TopMine的总体表现相似。但是观察发现,在短文本语料库(5conf,AMiner-Titles)上,TopMine要好于TF-IDF方法,在长文本语料库(DBLP Abstracts,AP News,AMiner-Abstracts)上,TF-IDF好于TopMine。因为TopMine采用自底向上的方式分割文本,标题语料库中的Quality Phrase比较集中,而新闻、摘要等语料库中短语的凝练程度不高,相比于TF-IDF来说,不适合使用TopMine。尽管CQMine

表3 候选短语挖掘3个阶段对比

Table 3 Comparison of three stages of candidate phrase mining %

数据集	挖掘方法	准确率	精确率	召回率	F1-Score
5conf	方法①	99.65	80.29	80.08	80.18
	方法②	99.62	85.62	**91.45**	88.44
	方法③	**99.72**	**90.94**	90.04	**90.49**
DBLP Abstracts	方法①	99.97	**92.08**	85.68	88.76
	方法②	99.96	91.11	94.23	92.64
	方法③	**99.97**	90.91	**95.68**	**93.23**
AP News	方法①	99.95	92.25	89.21	90.71
	方法②	99.94	93.62	**96.51**	95.05
	方法③	**99.96**	**95.39**	95.99	**95.69**
AMiner-Titles	方法①	99.63	83.36	89.02	86.10
	方法②	99.58	87.87	**92.86**	90.30
	方法③	**99.68**	**91.67**	91.35	**91.51**
AMiner-Abstracts	方法①	99.97	90.72	92.38	91.54
	方法②	99.97	**93.13**	93.73	93.43
	方法③	**99.97**	92.32	**94.93**	**93.60**

方法在每个数据集上都有所提升,但是提升明显的是短文本数据集,因为重叠短语分割方法对短文本更加有效。本文QPMSF方法与5种方法的最优值相比,精确率提升了5.97%,召回率提升了1.77%,F1-Score提高了4.02%。可见,在候选短语的基础上对短语准则进行加权处理能够有效提高Quality Phrase挖掘的精度。

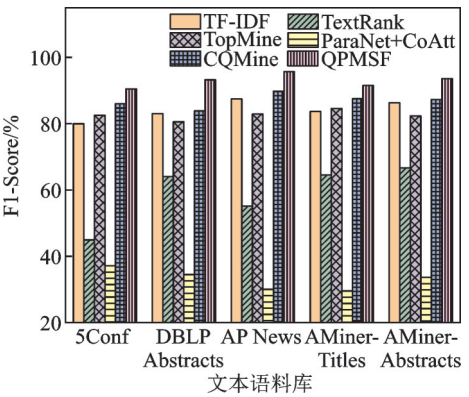


图7 本文算法与其他算法的对比

Fig.7 Comparison of QPMSF with other algorithms

6 结束语

本文针对无监督短语挖掘过程中候选短语质量不高和Quality Phrase的特征权重不恰当分配的问题,提出基于统计特征的Quality Phrase挖掘方法。结合频繁的*N*-Gram挖掘、组合性约束和拼写检查提高了候选短语的质量;引入公共知识库维基百科中的实体对候选短语添加类别标签,实现了Quality Phrase的特征加权。与最优的短语挖掘方法相比,基于统计特征的Quality Phrase挖掘方法明显提高了Quality Phrase挖掘的精度,精确率、召回率和F1-Score分别提升了5.97%,1.77%和4.02%。本文将特征加权应用到Quality Phrase的挖掘过程,为无监督短语挖掘打开了新的思路。然而,短语挖掘领域对短语质量的评价方式还有待深入研究,以维基百科的实体为基准判断一个短语属于或不属于Quality Phrase还不够准确。另外,实现多语言通用的Quality Phrase挖掘方法,并应用到文本分类、主题建模等任务是本文今后研究的方向。

参考文献：

[1] LI Chenliang, SUN Aixin, WENG Jianshu, et al. Tweet segmentation and its application to named entity recognition[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 27(2): 558-570.

[2] SHANG Jingbo, LIU Jialu, JIANG Meng, et al. Automated phrase mining from massive text corpora[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(10): 1825-1837.

[3] 林子琦,倪晚成,赵美静,等.基于概念知识树的双宾短语分析[J].中文信息学报,2017, 31(5): 21-31, 49.
LIN Ziqi, NI Wancheng, ZHAO Meijing, et al. Parsing of double-object phrases based on concept knowledge tree[J]. Journal of Chinese Information Processing, 2017, 31(5): 21-31, 49.

[4] NORASET T, DOWNEY D, BING L. Estimating marginal probabilities of n-grams for recurrent neural language models [C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2018: 2930-2935.

[5] 池云仙. 基于最长闭频繁序列模式的文本分类方法[D]. 石家庄: 河北师范大学, 2017.
CHI Yunxian. Text classification method based on the longest closed frequent sequential patterns[D].Shijiazhuang: Hebei Normal University, 2017.

[6] PARAMESWARAN A, GARCIA-MOLINA H, RAJARAMAN A. Towards the web of concepts: Extracting concepts from large datasets [J]. Proceedings of the VLDB Endowment, 2010, 3(1): 566-577.

[7] EL-KISHKY A, SONG Yanglei, WANG Chi, et al. Scalable Topical phrase mining from text corpora[J]. Proceedings of the VLDB Endowment, 2014, 8(3): 305-316.

[8] TENEVA N, CHENG Weiwei. Salience rank: Efficient keyphrase extraction with topic modeling[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics(Short Papers). Stroudsburg, PA: ACL, 2017: 530-535.

[9] ZHAO Jing, ZHANG Yuxiang. Incorporating linguistic constraints into keyphrase generation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: ACL, 2019: 5224-5233.

[10] CHEN Jun, ZHANG Xiaoming, WU Yu, et al. Keyphrase generation with correlation constraints[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: ACL, 2018: 4057-4066.

[11] WANG Yue, LI Jing, HOU Pong-Chan, et al. Topic-aware neural keyphrase generation for social media language[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: ACL, 2019: 2516-2526.

[12] CHEN Wang, HOU Pong-Chan, LI Piji, et al. An integrated approach for keyphrase generation via exploring the power of retrieval and extraction[C]//Proceedings of NAACL-HLT. Minneapolis, Minnesota: ACL, 2019: 2846-2856.

[13] HOU Pong-Chan, CHEN Wang, WANG Lu, et al. Neural keyphrase generation via reinforcement learning with adaptive rewards[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: ACL, 2019: 2163-2174.

[14] YE Hai, WANG Lu. Semi-supervised learning for neural keyphrase generation[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: ACL, 2019: 4142-4153.

[15] DEBANJAN M, JOHN K, RAJIV R S, et al. Key2Vec: Automatic ranked keyphrase extraction from scientific articles using phrase embeddings[C]// Proceedings of NAACL-HLT. New Orleans, Louisiana: ACL, 2018: 634-639.

[16] FLORIAN B. Unsupervised keyphrase extraction with multipartite graphs[C]//Proceedings of NAACL-HLT. New Orleans, Louisiana: ACL, 2018: 667-672.

[17] ZHANG Yingyi, ZHANG Chengzhi. Using human attention to extract keyphrase from microblog post[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: ACL, 2019: 5867-5872.

[18] LIU Jialu, SHANG Jingbo, HAN Jiawei. Phrase mining from massive text and its applications[J]. Synthesis Lectures on Data

Mining and Knowledge Discovery, 2017, 9(1): 1-89.

[19] LI Bing, YANG Xiaochun, ZHOU Rui, et al. An efficient method for high quality and cohesive topical phrase mining[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 31(1): 120-137.

[20] 杨玥, 张德生. 中文文本的主题关键短语提取技术[J]. 计算机科学, 2017, 44(11A): 432-436.
YANG Yue, ZHANG Desheng. Technology of extracting topical keyphrases from Chinese corpora[J]. Computer Science, 2017, 44(11A): 432-436.

[21] LIU Jialu, SHANG Jingbo, WANG Chi, et al. Mining quality phrases from massive text corpora [C]// Proc ACM SIGMOD Int Conf Manag Data. New York, USA: ACM, 2015: 1729-1744.

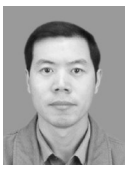
[22] TANG Jie, ZHANG Jing, YAO Limin, et al. ArnetMiner: Extraction and mining of academic social networks[C]// Proc of the Fourteenth ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York, USA: ACM, 2008: 990-998.

[23] 赵京胜, 朱巧明, 周国栋, 等. 自动关键词抽取研究综述[J]. 软件学报, 2017, 28(9): 2431-2449.
ZHAO Jingsheng, ZHU Qiaoming, ZHOU Guodong, et al. Review of research in automatic keyword extraction[J]. Journal of Software, 2017, 28(9): 2431-2449.

[24] LI Wengen, ZHAO Jiabao. TextRank algorithm by exploiting Wikipedia for short text keywords extraction[C]//Proceedings of 2016 3rd International Conference on Information Science & Control Engineering. Piscataway, NJ: IEEE, 2016: 683-686.

作者简介:

杨欢欢(1993-),女,硕士研究生,研究方向:数据挖掘与智能信息处理,E-mail:yanghuan_june@163.com。



赵书良(1967-),**通信作者**,男,教授,博士生导师,研究方向:数据挖掘与智能信息处理。

李文斌(1974-),男,教授,研究方向:机器学习、演化计算、数据挖掘。

武永亮(1986-),男,博士研究生,研究方向:数据挖掘与自然语言处理。

田国强(1995-),男,硕士研究生,研究方向:数据挖掘与智能信息处理。

(编辑:张黄群)