doi:10.3772/j.issn.2095-915x.2021.06.007

面向中医文献的短语挖掘方法

谢永红[1,2] 蒋彦钊[1,2] 贾麒[1,2] 范欣欣[1,2]

1. 北京科技大学计算机与通信工程学院 北京 100083；
2. 材料领域知识工程北京市重点实验室 北京 100083

开放科学（资源服务）标识码（OSID）

摘要：[目的 / 意义] 在中医文献中存在大量的短语，目前的短语挖掘方法在中医文献上效果差强人意，针对这个问题，提出了面向中医文献的短语挖掘方法。[方法 / 过程] 该方法在中医文献分词器基础上，利用中医领域新语言知识库，训练得到短语质量评分模型，并在此基础上利用词性标签信息构建短语分割模型对文献进行挖掘，提高中医文献中短语挖掘的准确率。并在《中医古代名医医案》上进行实验。[结果 / 结论] 选取挖掘短语的 Top300 对其进行精确率的评估，其准确率为 84.96%。实验证明中医文献分词器 + 短语分割模型的挖掘方法在中医领域文献上的短语挖掘效果优于其他挖掘方法。

关键词：中医文献短语挖掘；短语挖掘；高质量短语；中医文献分词器；短语质量评分模型；词性标签；短语分割模型

中图分类号：TP391.1；G35

Phrase Mining Technology for TCM Literature

XIE Yonghong[1,2] JIANG Yanzhao[1,2] JIA Qi[1,2] FAN Xinxin[1,2]

基金项目 中国科学技术信息研究所情报工程实验室 2019 年开放基金项目“中医药古代文献的知识术语挖掘关键技术研究”。

作者简介 谢永红（1970-），副教授，研究方向为知识发现与知识工程、数据库技术，Email：xieyh@ustb.edu.cn；蒋彦钊（1995-），硕士研究生，研究方向为知识工程、语言技术；贾麒（1989-），博士研究生，研究方向为知识图谱、自然语言处理；范欣欣（1996-），硕士研究生，研究方向为知识图谱、自然语言处理。

引用格式 谢永红 , 蒋彦钊 , 贾麒 , 等 . 面向中医文献的短语挖掘方法 [J]. 情报工程 , 2021, 7(6): 76-87.

1. School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China;
2. Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing 100083, China

Abstract: [Objective/ Significance] There are a large number of high-quality phrases in traditional Chinese medicine literature. The current phrase mining method is unsatisfactory in traditional Chinese medicine literature. In response to this problem, a phrase mining method for traditional Chinese medicine literature is proposed in this research. [Methods/Process] The method uses the new language knowledge based on the field of traditional Chinese medicine to train the phrase quality scoring model on the basis of text segmentation. On this basis, it uses the part-of-speech tag information to construct a phrase segmentation model to mine the literature and improve the precision of phrase mining. Then, this research conduct experiments in "Ancient Chinese Medical Cases of Traditional Chinese Medicine". [Results /Conclusions] The Top300 mining phrase was selected to evaluate the precision, and the precision was 84.96%. The experiment proves that the text segmentation for the traditional Chinese medicine literature + phrase segmentation model is superior to other mining methods in the literature of traditional Chinese medicine.
Keywords: TCM literature; phrase mining; high-quality phrases; text Segmentation for TCM; phrase quality scoring model; part-of-speech tag; phrase segmentation model

引言

自然语言处理是目前人工智能研究的一个重要研究方向，而高质量的短语知识库对自然语言处理中的信息抽取尤为重要。高质量的语言知识库不仅可以让语料的标注更加准确，而且使得自然语言处理模型获得更加准确的语义信息。为了获得高质量的语言知识库，短语挖掘[1]成为研究者们研究的热点。

短语挖掘指的是从给定的语料中挖掘出其中包含的高质量的短语。从文本中挖掘出高质量的短语不仅可以对文献进行客观的评价，而且也有利于人们对文献的理解以及下一步的研究。

中医作为中国传统医学，经过长期的发展，积累了大量的书籍文献著作，很多重要文献多成书于古代，使用文言文和古人的口语。因为成书年代不同，表达方式多种多样，与现代汉语差异较大，其中大量的方剂名称、症状信息、疾病名称等多数是由几个词组成的短语形式，这些都增加了自动化处理中医文献、正确抽取结构化中医知识的难度。所以研究提高中医短语挖掘准确率的自动化方法显得尤为重要。

分词是短语挖掘的基础，相较于传统文献，中医文献的语法和表达方式具有一定的特殊性，比如通假字等。所以本文提出了一种能够较好处理具有古文特征的中医文献分词器，使得短语挖掘在较好的分词基础上进行；针对多数短语挖掘方法往往依赖于大量的专家指导和人工标注训练集的问题，本文在现有的语言知识库基础上添加中医文献高质量短语，构建中医领域的新语言知识库；并基于此构建训练集，训

doi:10.3772/j.issn.2095-915x.2021.06.007

练短语质量评分模型；然后在分词的基础上结合词性标签，构建词性标签序列质量评分模型；最终形成中医文献分词器 + 短语分割模型的中医文献短语挖掘模型。

1 相关研究

国内外针对短语挖掘已经有很长时间的研究 [2-4]，按照发展时间大致分为基于规则的短语挖掘方法，统计学习的短语挖掘方法和基于深度学习的短语挖掘方法。

基于规则的短语挖掘方法指的是根据短语的词法、语法等规则及文本特征构建相应的格式化短语识别模板，并利用模式匹配的方法进行短语挖掘 [5]。基于统计学习的短语挖掘方法是指从大量的文本中统计短语的特征，根据短语的各个特征信息进行短语挖掘的方法。如 Sarasvady 等 [6] 根据短语出现的频率等信息挖掘高质量的短语。基于神经网络的短语挖掘方法指的是利用神经网络学习文本中所存在的语法、句法以及语义特征，根据这些特征对文本进行短语挖掘。如 Xu 等 [7] 基于协同训练进行的电商领域短语挖掘。但是这些方法在进行短语挖掘时都是针对单一文档进行挖掘，所以其挖掘有一定的局限性。为解决此问题，有人进行了相关研究，如 K. Frantzi 等 [8]、Y. Park 等 [9] 在进行的术语提取任务时对多文档进行了挖掘。

当前中文短语挖掘大多所挖掘的语料是现代文 [10,11]。由于中医文献大量源于古籍，其语法有其特性，所以需要利用特有的分词方法及词性标签等语法信息，寻找在中医文献中短语存在的规律，并利用这种规律来提高短语挖掘获得的短语的质量。

2 方法

本文的目标是从大量的中医语料中挖掘出一些高质量短语，高质量短语应该具有以下几个特性。

（1）高频率：指高质量短语应该在待挖掘文献中出现足够多的次数。（2）一致性：指词组作为短语出现的频率要高于偶然出现的频率。例如“金银花 颗粒”和“忍冬 颗粒”，假设“金银花”和“忍冬”这两个词的出现频率相似，但是在日常生活中，“金银花 颗粒”作为短语出现的频率要高于“忍冬 颗粒”偶然出现的频率，所以“金银花 颗粒”更符合高质量短语的一致性。（3）信息性：高质量短语应该具有实际的意义。（4）完整性：完整地表达了一个含义。在一些文本中，因为主题的不同，一个短语和它的子短语可能都具有完整性。如“桂枝汤”和“桂枝”。所以在出现桂枝汤的时候希望其可以以“桂枝 汤”的形式挖掘出来。低质量短语即为不完全包含上述四个特征的短语，如：克 乳香、一钱 青等。

为了更好地进行高质量的短语挖掘，需要构建一个包含中医高质量短语的新语言知识库。因此，本文在现有的通用语言知识库基础上，结合在中医领域研究的前期积累，添加了大量中医文献高质量短语，建立了中医领域新语言知识库。

首先中医文献分词器对待挖掘语料进行分词。其次，在分词的结果上采用 N-gram[12] 的方法提取出大量的候选短语，这些短语包

含一些低质量短语。再次，从候选短语中过滤掉已经存在于中医领域新语言知识库中的短语，剩下的作为训练短语质量评分模型的负样本，并且将中医领域新语言知识库筛选出的高质量短语作为正样本形成训练集。从次，利用这个训练集训练得到短语质量评分模型；同时利用分词后每个词的词性标签，构建词性标签序列质量评分模型；利用短语质量评分模型以及词性标签序列质量评分模型构建短语分割模型，对分词后语料进行短语分割。最后，对分割后的短语进行质量评分，获取高评分的短语作为挖掘出的高质量短语。具体细节方法在接下的章节中介绍。短语挖掘流程示意图如图 1 所示。

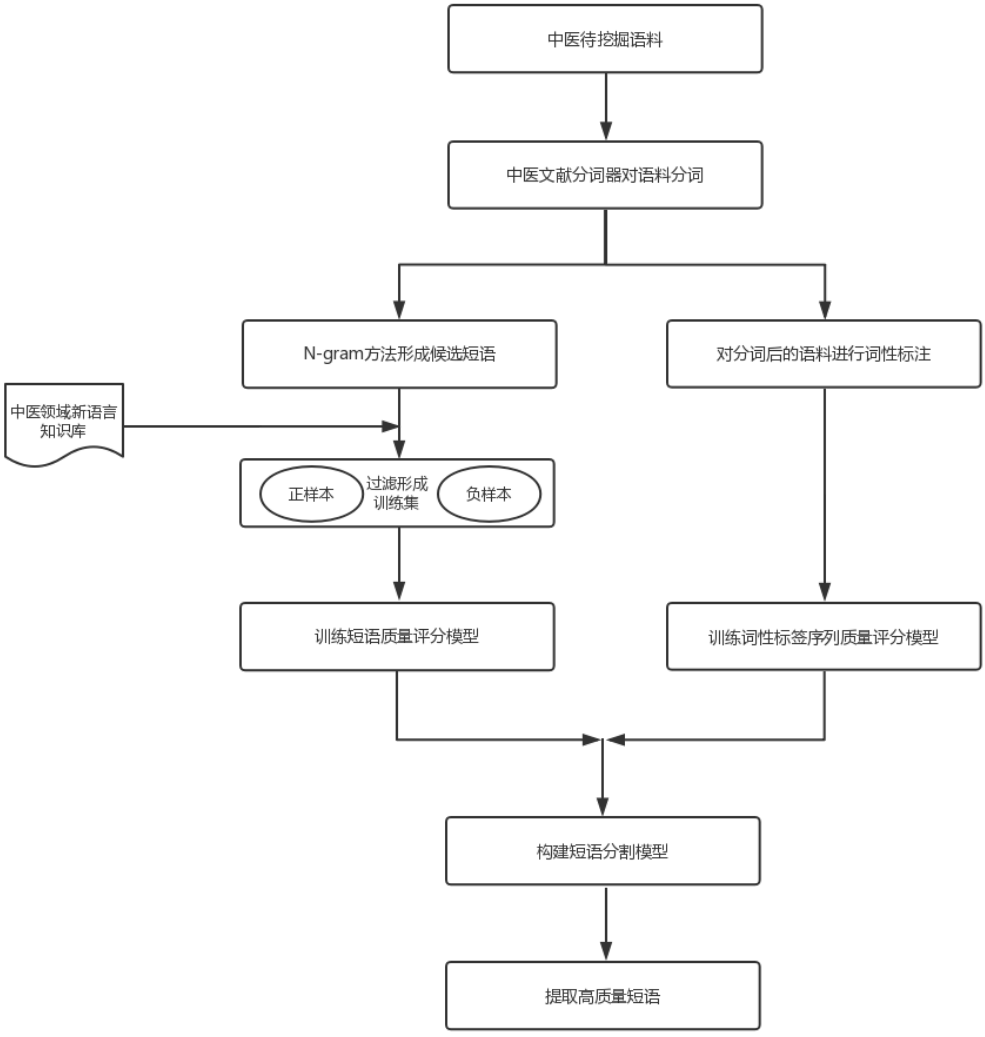


图 1　短语挖掘流程示意图

2.1　中医文献分词器

中医文献大多数成书于古代，表达方式多种多样，并有其独特的语法特性，所以建立专门的中医文献分词器 [13] 对于提高中医短语挖掘及中医语料的处理是非常重要的。中医文献分词器基于统计学习的分词方法，使用 N-gram 语言模型结合隐马尔可夫模型（HMM[14]）进行中医文献分词。分词流程如图 2 所示。

doi:10.3772/j.issn.2095-915x.2021.06.007

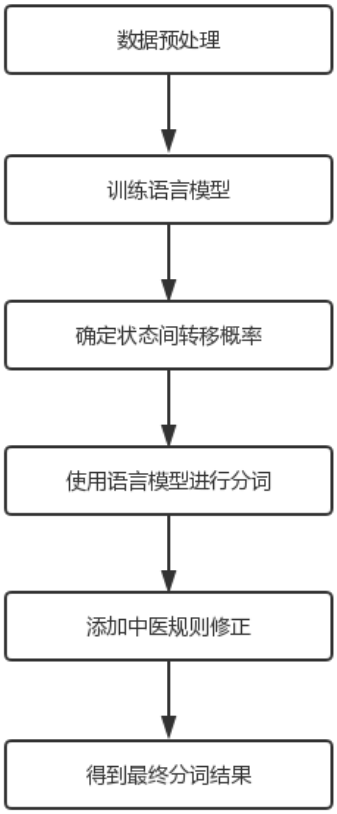


图 2　分词流程示意图

（1）数据预处理首先将大量的电子化中医文献进行预处理，包括去目录、调整文件编码格式等，并将处理后文献的每一个字后面添加一个空格，作为语言模型的训练集。

（2）训练语言模型。中医文献大多是用文言文撰写的，单字词的情况比较普遍，但考虑到中医领域有很多专业术语，如方剂名称、症状信息、疾病名称，为了在最大化保证准确率的前提下又尽可能节约算力和时间，选择了 4-gram 语言模型。也就是第 i 个字出现的概率仅与前 3 个字出现的概率有关。

由于采用的是 4-gram 语言模型，所以每个字就存在 4 种状态：第一种为单字词或者多字词的首字；第二种为多字词的第二字；第三种为多字词的第三字；第四种为多字词的其余部分。将这 4 种状态分别标记为 a,b,c,d。在由 n+1 个字（c_i, i=0…n）组成的句子 c_0, c_1, c_2, ..., c_n 中，对于 c_k 来说，其对应的四种状态的概率分别为：

$$P(c_k|a)=P(c_k) \quad (1)$$

$$P(c_k|b)=P(c_k|c_{k-1}) \quad (2)$$

$$P(c_k|c)=P(c_k|c_{k-1}, c_{k-2}) \quad (3)$$

$$P(c_k|d)=P(c_k|c_{k-1}, c_{k-2}, c_{k-3}) \quad (4)$$

（3）确定状态间转移概率。由于单字词的后面只能是单字词或多字词的词首，多字词的首字后面只能是多字词的第二字，多字词的第二字后面只能是多字词的第三字或单字词或多字词的首字，多字词的第三字后面只能是多字词的其余部分或单字词或多字词的首字，多字词的其余部分后面能是单字词或多字词的首字或多字词的其余部分，那么除上述转移状态，其余转移概率为零。因此，非零状态下的条件转移概率有 8 种，即：

$$P(a|a), P(b|a), P(c|b), P(a|b), P(a|c), P(d|c), P(a|d), P(d|d) \quad (5)$$

通过对大量的中医文献进行统计，得到上述转移概率如下：

$$P(a|a)=0.96, P(b|a)=0.2, P(c|b)=0.009, P(a|b)=0.9, P(a|c)=1, P(d|c)=0.005, P(a|d)=1, P(d|d)=0.0001$$

（4）使用语言模型进行分词。根据所得的转移概率以及 4-gram 语言模型计算各个邻接字的各种情况的条件概率，可以使用 HMM 的方法找到最优状态路径，作为切分结果，得到最初的分词结果。在由 n+1 个字（c_i, i=0…n）组成的句子 c_0, c_1, c_2, ..., c_n 中，每个字符对应的 a,b,c,d 四种可能存在状态的切分方式概率如图 3 所示：

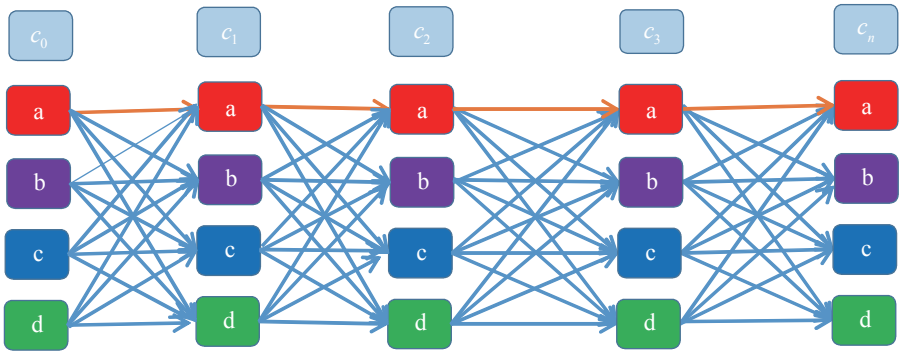


图 3 切分方式概率示意图

其中图 3 一共存在 4n 条路径，对每条路径可以算出这条路径不同状态序列所对应的概率，如可以得出 c_0, c_1, c_2, c_3, ..., c_n 全为状态 a 的这条路径其概率 P 为：

$$P=P(c_0|a)P(a|a)P(c_1|a)P(a|a)P(c_2|a)P(a|a)P(c_3|a)$$
$$\cdots\cdots P(a|a)P(c_n|a) \quad (6)$$

类似地可以求出左右路径的概率值，把所得概率最大的那条路径作为初步的切分路径，就可得到初步分词结果。

（5）添加中医规则修正。由于中医文献中有其存在的特殊语法，所以需要对初步切分的结果进行优化处理。具体做法为根据词性和中医方面语言学知识编写规则文件，再根据规则文件对分词结果进行进一步的切分。具体规则含义如表 1 所示：

表 1 规则含义对照表（部分数据）

存储样式	含义	部分样例数据
某字符|*	从某字符后面进行切割	《|* 》|*
*|某字符	从某字符前面进行切割	*|与 *|曰
|某字符|	从某字符前面和后面都进行切割	*|之|* *|如|*
某字符|某字符	在这两个字符之间进行切割	枣|一枚

（6）得到最终分词结果。根据规则文件对初步分词结果进行切分，得到第二次分词结果，然后利用中医领域常见词表，对结果进行修正得到最终的分词结果。部分分词结果如图 4 所示：

医学 之 要 ， 莫先于 切脉 ， 脉 候 不 真 ， 则 虚实 莫辨 ， 攻补 妄施 ， 鲜 不 夭人 寿命 者 。
其 次 则 当 明 药性 ， 如 病 在 某经 当用 某药 ， 或 有 因 此 经 而 旁达 他经 者， 是 以 补母泻子 ， 扶弱抑强 ， 义 有 多 端， 指 不 一 定。
自 非 兼 贯 博通 ， 析微洞奥 ， 不 但 呼应不灵 ， 或 反 致邪失正 。
先 正 云 用药 如 用 兵， 诚不 可以 不 慎 也。
古今 著 《本草》 者， 无虑 数百 家 。
其 中 精 且 详 者， 莫如 李氏 《纲目》 ， 考究 渊博 ， 指示 周 明 ， 所以 嘉惠 斯人 之 心 ， 良 云 切 至 。
第 卷帙浩繁 ， 卒难究殚 ， 舟车 之 上 ， 携 取 为 难 ， 备 则 备 矣， 而 未能 要 也。
他 如 主治 、 指掌 、 药性 歌赋 ， 聊以 便 初学 之 诵习 ， 要 则 要 矣， 而 未能 备 也。
近 如 《蒙筌》 、 《经疏》 ， 世 称善 本 ， 《蒙筌》 附 类， 颇 著 精 义， 然 文 拘对偶 ， 辞太繁缛 ， 而 阙略 尚多 ；
《经疏》 发明 主治 之 理， 制方 参互 之 义， 又 著简误 以 究其 失 ， 可 谓 尽 善 ， 然 未暇 详 地道 ， 明 制 治 ， 辨真赝 ， 解 处 偶有 傅会 ， 常 品 时 多 芟黜，均 为 千虑之一 失 。

图 4 部分分词结果

doi:10.3772/j.issn.2095-915x.2021.06.007

2.2 构建训练样本集

2.2.1 构建中医领域新语言知识库

为了得到中医文献高质量短语，同时也为了减少人力和时间的花销，在通用语言知识库基础上，利用实验室长期进行中医领域研究积累的大量短语集合及从开放知识库中获取的高质量中医短语集合，建立中医领域新语言知识库。中医领域新语言知识库部分数据如图 5 所示：

开封
竹林七贤
乔木
纸张
曾健成
咸阳市
芒根汤
安胎四物饮
艾叶
神情虚弱
神验黄柏散
圣灵丸
苍耳草膏
大乌金丸
丁香茱萸汤
豆蔻子汤
耳内外肿痛
伏铁饮
谷食不化
海桐散

图 5 中医领域新语言知识库部分数据

2.2.2 形成候选短语

为了构建正负样本集，要使用 N-gram 的方法来形成候选短语，由于大多数短语由两个词组成，我们选用 2-gram 形成候选短语，也就是以 2 为长度分割分词后的句子。如对于分词后的文本“用 藿香 正气散 一服 愈”，那么使用 2-gram 的到的候选短语有｛“用 藿香”，“藿香 正气散”，“正气散 一服”，“一服 愈”｝。

2.2.3 构建正负样本集

构建的候选短语中包含一些低质量短语和高质量短语，利用中医领域新语言知识库将候选短语分成两个集合，一个是存在于中医领域新语言知识库中的高质量短语集合，作为训练短语质量评分模型的正样本集；另一个是不在中医领域新语言知识库中的可能质量较差的短语集合，作为训练的负样本集。

2.3 训练短语质量评分模型

构建好训练集之后采用随机森林 [15] 的方法训练短语质量评分模型，随机森林是一种由多棵决策树组成的集成分类器。集成分类器用到的主要思想是集成学习，因为单一的分类器的精度很容易遇到瓶颈难以提升且容易出现过拟合现象，因此通过聚集多个模型来提高预测精度，获得更好的分类结果。

采用随机森林训练质量评分模型对短语进行特征选择是至关重要的，选择合适的特征会使得模型最后达到较好的效果。针对上述提到的高质量短语的四个属性，利用统计学方法选择以下的信息作为短语特征。

1）短语出现的频率：每一个候选短语在待挖掘语料中出现的频率。

2）逐点互信息：可以用来衡量两个事物的相关性。逐点互信息的结果越大，表示相关性越高。

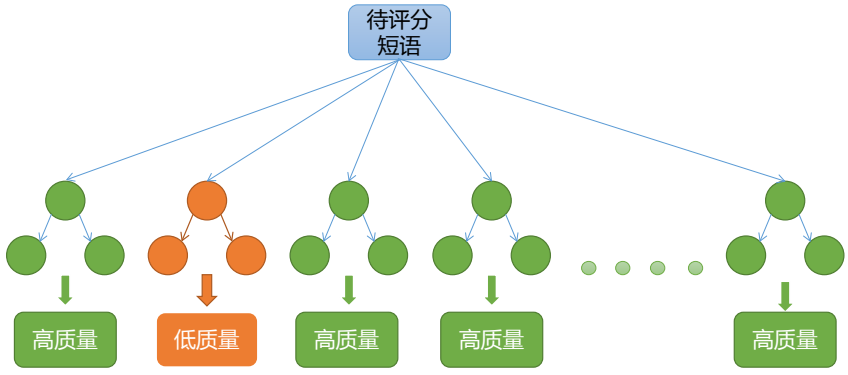
3）KL 散度 [16,17]：KL 散度（Kullback-Leibler divergence）又被称作相对熵，是两个概率分布间差异的非对称性度量。KL 散度越小，说明候选短语的质量越高。

4）TF-IDF[18]：TF-IDF 是一种统计方法，用于评估一个字或者一个词对一个文件集或一个语料库中一份文件的重要程度。

5）短语中标点符号的使用：短语标点的使

用分为不同的两类。第一类是指候选短语中间带有标点符号，那么不论是何种标点符号，该候选短语的质量都会比较低，可以将其特征值设置为 0。第二类是指候选短语在引号、括号之间，在这类标点符号之间的候选短语具有较高的概率是提供了完整信息的高质量短语，可以将其特征值设置为 1。

选取好特征之后，就可以使用随机森林训练短语质量评分模型，这里的随机森林包含 1000 棵树，每棵树分别从正样本集合中和负样本集合中分别随机选取 100 个候选短语作为该树的训练集（这里采取有放回抽样）。当随机森林中所有决策树训练完成之后，一个短语的质量评分即为随机森林中判断该短语为高质量短语的决策树的比例。例如：对于短语“藿香正气散”，如果 1000 棵树中有 800 棵树认为其为高质量短语，那么该短语的质量评分就为 0.8。短语质量评分模型示意图如图 6 所示。



$$\text{短语质量评分} = \frac{\text{判断为高质量短语的决策树数量}}{\text{总的决策树数量}}$$

图 6　短语质量评分模型示意图

2.4　利用词性标签分割短语

短语作为一个完整的语义单元是有一些语法规律的 [19]，从句子的语法角度进行分析，词性之间的组合方式往往有一定规律可循。因此可以利用词性等信息来对句子的分词结果进行短语分割，从而提高被挖掘短语的质量。

假设有一条由 n+1 个词（w_i，i=0…n）的待挖掘的语料的分词结果 $w_0, w_1, w_2,..., w_n$。每个 w_i 词可以打上其词性标签 t_i（i=0…n）。假设其中包含一个词性标签子序列 $t_{[1,r)} \in t_1....t_{r-1}$，对其进行词性标签序列质量评分，其质量评分越高表示其相关单词序列作为一个完整语义的可能性越高。假如存在一条词性标签序列 ‘n,n,n,v,n’，其中 n 表示词性为名词，v 表示词性为动词。那么对于 ‘n,n’ 这个词性标签子序列的质量评分可能会较高，而对于 ‘n,v’ 这个词性标签子序列评分就会较低，因为大部分短语都不会以动词作为结尾。

假设将带有词性标签的句子分割为 m 个短语，其边界设置为 $b_0, b_1, b_2,..., b_m$ $(m \leqslant n)$，

doi:10.3772/j.issn.2095-915x.2021.06.007

其中 b_i 表示第 i+1 个短语的起始位置，那么词性标签序列的质量评分模型 T 则可用以下公式表示：

$$T(t_{[b_{r-1},b_r)}) = (1 - P(t_{b_r} | t_{b_r-1})) \prod_{j=b_{r-1}+1}^{b_r-1} P(t_j | t_{j-1}) \quad (7)$$

其中$P(t_{b_r} | t_{b_r-1})$表示在短语边界 b_r 左边词性标签为 t_{b_r-1} 出现的条件下短语边界右边词性标签为 t_{b_r} 出现的概率，该概率可以在分词后进行词性标注的文档中统计得到的。这样就可以得到任意的词性标签序列的质量评分。

然后可以利用词性标签序列质量评分模型 T 和之前训练好的短语质量评分模型 Q 构建短语分割模型对语料进行短语边界的重新划分。对于一句话给定第一个短语边界 b_0（通常为句首），那么下一个短语边界 b_1，有：

$$P(b_1|b_0) = T(t_{[b_0,b_1)}) \times Q(w_{b_0} \cdots w_{b_1-1}) \quad (8)$$

选择一个 b_1 使得这个函数取得最大值，那么 b_0 与 b_1 之间的单词组合就作为第一个划分出的短语。对于短语边界 b_2 采取同样的方法，这样就完成了短语边界的重新划分。

在短语分割之后，利用短语质量评分模型 Q 对出现频次高于某个阈值的短语进行评分。然后将短语以及质量评分写入文件中。

3 结果实验结果对比与分析

3.1 实验设置

针对《中医古代名医医案》文献提取其中的高质量短语；将短语出现频率阈值设置为 10，即出现 10 次以上的短语才能作为候选短语集中的短语；依据中医文献的分词统计结果，分词后的词长多为 2~3，且短语通常由两个词组成，所以将短语最大字长设置为 6。

3.2 对比实验结果

将中医文献分词器 + 短语分割模型与 TF-IDF、TextRank[20] 和 ANSJ① 分词方法 + 短语分割模型结果进行对比。其中，TF-IDF 方法指的是根据词频等信息挖掘文章中的短语；TextRank 方法是通过词之间的相邻关系构建网络，然后用 PageRank[21] 迭代计算每个节点的 rank 值，排序 rank 值挖掘文章的短语。

本实验选取挖掘出的前 300 个短语作为高质量短语。选用常用的精确率（Precision）作为评估指标对四种方法进行评判，精确率的计算方式为挖掘出的短语中真正的高质量短语数除以挖掘出的高质量短语数目。因无法对语料中所有的高质量短语精确统计，所以无法对其召回率及 F1 值进行计算。实验评估结果如表 2 所示。

表 2 四种方法评估结果

方法	TF-IDF	Text-Rank	Ansj分词+短语分割模型	中医文献分词器+短语分割模型
精确率	74.36%	76.27%	80.55%	84.96%

本实验每个方法都是返回的一个抽取的短语的列表。部分结果如表 3 所示。

① https://github.com/NLPchina/ansj_seg

表 3　四种方法挖掘部分结果

TF-IDF	TextRank	Ansj分词+短语分割模型	中医文献分词器+短语分割模型
三钱	不能	藿香 正气散	藿香 正气散
一钱	半夏	面色 萎黄	面色 萎黄
二钱	咳嗽	四肢 逆冷	大汗 淋漓
五分	茯苓	寿 山	肌肉 消瘦
半夏	白芍	神识 昏迷	骨节 疼痛
四钱	舌苔	小溲 短少	胸膈 满闷
白芍	腹痛	肌肉 消瘦	六味 地黄汤
杏仁	甘草	大汗 淋漓	沈 祖
茯苓	杏仁	精神 疲倦	木火 刑金
甘草	饮食	口 渴欲饮	精神 疲倦
钱半	发热	清阳 不升	州 都
12	脉象	刘某 某	竹叶 石膏汤
大便	不可	胸膈 满闷	竹 苑
五钱	呕吐	匀 两次	四肢 厥冷
枳壳	处方	脱 肛	杭 菊花
…	…	…	…

3.3　实验结果分析

根据表 3 可以发现，ANSJ 分词 + 短语分割模型和中医文献分词器 + 短语分割模型挖掘出的质量短语语义更加完整，因为这两种方法在分词的基础上利用词性标签浅层语法信息对分词后的文本进行了短语的重新划分，从而使得挖掘到的短语语义更加完整。而 TD-IDF 和 TextRank 是在分词后直接利用词频等信息进行短语挖掘，所以中医文献分词器 + 短语分割模型相对于 TD-IDF 和 TextRank 这两种方法挖掘出的短语语义信息是更加完整的。

根据表 2 可以发现中医文献分词器 + 短语分割模型方法的准确率更高，原因是我们所使用的分词器是针对中医文献专门的分词器，其在分词的过程中加入了中医方面语言学知识编写规则文件对初次分词结果进行修正，所以在针对于中医文献的分词上会更加准确，从而导致分词的误差传递会更小从而在使得最后挖掘出的短语精确率更高。

每种方法的结果列表都是按照质量从高到低顺序进行排列的，在此只使用了每种方法的前 300 个结果进行评估，因此，方法整体的精确率会略微低于表 2 中的数据。但是根据表 2 中的结果，在前 300 个短语中中医文献分词器 + 短语分割模型方法准确率更高，也就表明在这 300 个短语中中医文献分词器 + 短语分割模型挖掘出的高质量短语比重更大，所以也就证明了对于同一篇中医文献使用中医文献分词器 + 短语分割模型可以挖掘出更多的高质量短语。

doi:10.3772/j.issn.2095-915x.2021.06.007

4 总结

本文阐述了中医短语挖掘的意义、实用性和重要性，详细介绍了中医文献分词器＋短语分割模型的短语挖掘方法原理和流程。通过对比实验，从多个角度分析表明了中医文献分词器＋短语分割模型方法能在中医文献的高质量短语挖掘任务中取得较好的效果；定性分析结果也表明，增加了词性标签的方法在中医领域短语挖掘的任务上表现更好，能获得更完整语义的高质量短语，为其他任务（如词表的扩充、命名实体识别等任务）打下了坚实的基础。方法还有进一步改进的空间，在分词器、短语质量评分模型及词性标签范围等方面还可进一步优化，这也是我们未来研究的目标。

参 考 文 献

[1] Shang J, Liu J, Jiang M, et al. Automated phrase mining from massive text corpora[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(10):1825-1837.

[2] Witten I H, Paynter G W, Frank E, et al. KEA: Practical Automatic Keyphrase Extraction[C]. Fourth ACM Conference on Digital Libraries. ACM. 1999:254–255.

[3] Liu Z, Chen X, Zheng Y, et al. Automatic Keyphrase Extraction by Bridging Vocabulary Gap[C]. Fifteenth Conference on Computational Natural Language Learning. 2011.

[4] Han P, Du J, Chen L. Web opinion mining based on sentiment phrase classification vector[C]. IEEE International Conference on Network Infrastructure & Digital Content. 2010.

[5] Gaizauskas R, Demetriou G, Humphreys K. Term Recognition and Classification in Biological Science Journal Articles[C]. Computional Terminology for Medical & Biological Applications Workshop of the 2 Nd International Conference on NLP. 2000.

[6] Sarasvady S A, Vijayakumar P B, Jocabs D C, et al. Mining and computing phrase weight in texts[C]. Third International Conference on Innovative Computing Technology. IEEE, 2013.

[7] Xu Y, Liu J, Xiao Y, et al. Phrase Mining in Ecommerce Based on Cooperative Training[J]. Computer Engineering, 2020, 46(4):70-76,84.

[8] Frantzi K, Ananiadou S, Mima H . Automatic recognition of multi-word terms: the C-value/NC-value method[J]. International Journal on Digital Libraries, 2000, 3(2):115-130.

[9] Park Y, Byrd R J, Boguraev B K . Automatic Glossary Extraction: Beyond Terminology Identification[C]. Proceedings of the 19th international conference on Computational linguistics. Association for Computational Linguistics. 2002.

[10] 刘晨晖 , 张德生 , 胡钢 . 基于 TAKE 的中文关键短语提取算法研究 [J]. 计算机工程与应用 , 2020, 56(10):115-121.

[11] 黄九鸣 , 吴泉源 , 张圣栋 , 等 . 基于 AC-Trie 的在线社交网络文本流热点短语挖掘 [J]. 电子学报 , 2016, 44(10):2466-2470.

[12] Suzuki M, Yamagishi N, Tsai Y C, et al. Multilingual text categorization using Character N-gram[C]. 2008 IEEE Conference on Soft Computing in Industrial Applications. IEEE, 2008:49-54.

[13] Jia Q, Xie Y, Xu C, et al. Unsupervised Traditional Chinese Medicine Text Segmentation Combined with Domain Dictionary[C]. International Conference on Artificial Intelligence and Security. Springer, Cham, 2019:304-314.

[14] Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition[J]. Process of IEEE, 77(2):257-286.

[15] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees[J]. Machine Learning, 2006, 63(1):3-42.

[16] Kullback S, Leibler R A. On information and sufficiency[J]. The Annals of Mathematical Statistics, 1951(22):79-86.

[17] Huang A. Similarity measures for text document clustering[C]. The New Zealand Computer Science Research Student Conference (NZCS RSC), New Zealand. 2008:49-56.

[18] Salton G. The SMART Retrieval System-Experimentsin Automatic Document Processing[M]. New Jersey : Prentice Hall Inc, 1971.

[19] Finch G . Linguistic Terms and Concepts[M]. Oxford: Macmillan Education UK, 2000.

[20] Mihalcea R, Tarau P. TextRank: Bringing Order into Texts[C]. Proceeding Conference on Empirical Methods in Natural Language Processing. 2004.

[21] Page L, Brin S, Motwani R, et al. The PageRank Citation Ranking: Bringing Order to the Web[J]. Stanford Digital Libraries Working Paper, 1998.