基于文献计量和众包技术的前沿科技关键词挖掘

吕佳高[1] 梁奎阳[2] 蔡 伟[3]

(北京航空航天大学软件开发环境国家重点实验室 北京 100191)[1]

(北京国科知源科技有限公司 北京 100191)[2] (北京市科技信息中心 北京 100085)[3]

摘 要 随着科学技术高速发展,科技文献的数量与日俱增,从海量的文献数据中挖掘出前沿科技关键词是一个新的挑战,由专家进行人工分析是一种常见而传统的方式,但这种方式的效率低且成本高。文中提出了一种将文献计量与众包技术相结合的算法:首先利用自然语言处理的词性标注技术处理并获取文献中的名词,然后通过基于文献计量的科技监测方法筛选出潜在的科技关键词,最后利用众包平台的数据进一步筛选潜在的科技关键词。采用计算机领域和生物医药领域的英文文献数据进行实验,结果表明所提算法有一定的效果,其效率比人工分析的方式高,能为专家人工分析起到辅助作用。所提算法能够更好地指导前沿技术关键词的挖掘,为未来更加自动和智能的前沿技术关键词挖掘提供参考。

关键词 文献计量,关键词挖掘,众包技术

中图法分类号 TP391.1 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.03.041

Frontier Scientific Keyword Extraction Based on Bibliometric and Crowdsourcing

LV Jia-gao[1] LIANG Kui-yang[2] CAI Wei[3]

(State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, China)[1]

(Beijing Guoke Zhiyuan Technology Co., Ltd., Beijing 100191, China)[2]

(Beijing Sci-Tech Information Center, Beijing 100085, China)[3]

Abstract With the rapid development of science, the annual amount of scientific papers is growing, and new challenge is to extract the frontier scientific keywords from lots of papers. In traditional way, the extraction work is done by experts, which is inefficient and costs much. A new algorithm based on bibliometric analysis and crowdsourcing technique was proposed in this paper. Part-of-speech tagging is used to obtain the nouns from scientific papers, and potential scientific keywords are selected from these nouns by bibliometric analysis. The last procedure is using data from crowdsourcing platform to check potential scientific keywords and get results. English scientific papers in computer science and biomedicine are used to conduct experiments. The experiment results suggest that the proposed algorithm has effect on extraction, and it's more efficient than expert extraction procedure, so it can assist the expert to analysis frontier scientific keywords. In conclusion, this algorithm can do automatic extraction and show possibility of more automatic and intelligent extraction procedure in the future.

Keywords Bibliometric, Keyword extraction, Crowdsourcing

1 引言

当今社会科学技术发展十分迅速,据统计,至2009年世界上总计有超过5000万篇科技论文[1],而每年新增的期刊论文数量超250万[2]。随着科技论文数量的增大,出现了越来越多的在线数据库和学术检索引擎,以方便研究人员快速获取到自己需要的论文资源。面对这样大量而繁杂的数据,如何有效地对它们进行分析,并提取出当前最有价值的前沿科技关键词,是一个值得研究的问题。

前沿科技关键词挖掘及趋势分析,就是通过对顶级期刊或会议的文献数据进行深层次的分析,挖掘出其中最新出现的或者是关注热度急速增长的技术名词。通过分析获得前沿科技关键词,可以帮助研究人员了解当前某领域的最新前沿,以发现新的研究方向,从而进行针对性的研究,帮助初涉研究者了解领域的最新动态,帮助专业研究人员寻找突破创新点。

现阶段,分析和获取前沿科技关键词的方式主要是依靠

到稿日期:2018-02-03 返修日期:2018-05-28 本文受国家重点研发计划项目课题:众智化专业知识协同开发技术及应用(2017YFB1402403)资助。

吕佳高(1995—),男,硕士生,主要研究方向为数据挖掘;**梁奎阳**(1969—),男,硕士,中级工程师,主要研究方向为科技情报与信息处理,E-mail:liangky@escience.gov.cn(通信作者);**蔡 伟**(1976—),女,助理研究员,主要研究方向为科技数据分析。

领域内的专家分析大量文献。由专家阅读大量文献内容是一个耗时耗力的过程,这种方式效率低且成本高,而且每位专家都有自己侧重的研究方向,仅根据少数专家的判断很难对整个领域很好的把握。因此,希望能找到一种较为自动化的方式,利用机器从文献中筛选出合适的前沿科技关键词。

当前在基于文献计量的科技监测方法中,对挖掘科技热点关键词的研究较为丰富,而对挖掘前沿科技关键词的研究较少。科技热点关键词往往具有较高的关注热度,这会反映在关键词的出现频率、文献引用次数等特征中,而前沿科技关键词不一定具有类似的热度特征,新出现的前沿科技没有获得广泛关注也是很常见的情况。因此,科技热点关键词的挖掘方式不适合用来挖掘前沿科技关键词。

本文的主要贡献如下:

(1)提出了一种前沿科技关键词的挖掘方法,该方法通过自然语言处理、文献计量等多种方式筛选和获取科技关键词,具有一定的效果。

(2)引入众包平台的数据,与传统的文献计量方式进行综合分析,进一步提升了前沿科技关键词的挖掘效果。

(3)挖掘算法为自动挖掘前沿技术主题词提供了可能性,也为未来更加高效和智能的前沿技术关键词挖掘提供了参考。

2 相关工作

2.1 前沿科技的定义

国内外众多学者从多个不同的角度对前沿科技进行了不同的定义。19 世纪 60 年代,科学计量学之父普赖斯提出了研究前沿的概念,参考文献标志着科学研究前沿的本质[3],某个领域的科学研究前沿是由最近发表的且得到大量引用的文章所体现的。有的学者对其的定义较为笼统,认为领域内的大部分科学研究人员关注的相关问题和概念都是科学研究前沿[4]。之后出现了共被引文献的概念,有学者认为共被引文献聚类可表示研究前沿[5],也有学者认为共被引文献簇只能代表知识基础,共被引文献簇的引证文献才能有效地代表研究前沿[6]。另外,有我国学者认为研究前沿是领域内突然出现的新主题和新理论[7],认为前沿科技是最为先进和最具潜力的领域主题或领域方向[8]。

2.2 基于文献计量的科技监测方法

(1)基于文献的外部特征统计。一个文献的外部特征值指的是题名、关键词、作者等可以直接从文献中提取的内容。通过分析外部特征值在一段时间内的变化,可以发现一段时间内科技词汇的变化;也可以从空间的角度发现不同国家地域的热点科技的异同。一种最直观的挖掘方法是词频统计。突发检测算法[9]正是利用了词频统计的方法,该算法寻找短时间内有明显增长的词语作为热点科技词汇。周文杰[10]使用了 Sci^2 提供的突发词检测算法,探测了动物学领域的研究前沿。这种算法的优点是简单易行,易于实现,但其缺点也很明显,即词频增长的幅度在各个领域是不同的,无法通过一种比较科学的方法来确定一个临界值,往往需要进行人工判断;同时,其会过滤掉刚刚出现的低频词汇,而这些低频词汇中可能会有新的研究方向和技术,从而对结果造成影响。

(2)通过引文分析发现热点科技文献。每篇论文都会引用其他不同的文献,引文分析通过对这种引用关系进行统计来分析引用数量等特征,对科学发展趋势进行评价与预测[11]。引文分析中较为常用和重要的一种方法是共引分析法。两篇文献有共引关系是指它们同时被其他文献引用,共引关系的强弱取决于它们被多少篇文献同时引用。那些具有较多共引关系的文献往往代表了领域内得到普遍认可的概念,或者是普遍使用的方法[12]。2013 年,盛立[13]使用了 BibExcel 软件构建了文献共被引文献网络,并用 VOSviewer 展示了生物医药领域的知识图谱。共引分析以一种非常科学的计量方式来进行统计和分析,但其只根据文章的外部特征进行分析而忽略了对文章具体内容的分析,同时也会漏掉一些刚出现且没有得到较高引用的前沿科技文献。

(3)通过共词分析来获得热点技术词汇。共词分析采用的方法是,以两个词为一组,统计在某篇文献中的词频以及位置等信息,然后在许多组词的基础上进行聚类,通过分析这些词的关系得到领域内科技词汇的结构和变化[14]。郑彦宁等[15]在共词分析的基础上,加入了滑动窗口切分数据,有效地追踪了研究主题的变化过程。共词分析能够将结果聚焦到词语上,但共词分析也存在一些问题,由于是人工进行词语的选择,因此往往存在一些主观的因素,而且同一个词在不同的语境中的含意有所不同,若脱离了文章上下文内容则无法进行准确的判断[16]。

在引用分析、共词分析后,往往会得到复杂的网络,因此可以在文献计量学的基础上应用网络分析方法来揭示科学结构。常见的方法有优先链接、社团结构、传播动力学等[17]。

近几年,也有学者尝试用主题模型进行科学前沿的挖掘。主题模型是对文档中隐含主题的一种建模方法,能够基于文本语料库识别出潜在的主题。2017 年,冯佳等[18]使用了 LDA 模型,获取了文献中的多个主题,并通过计算主题强度和新颖度,筛选得到科学前沿主题。白如江等[19]对多数据源数据进行了聚类分析,并综合考虑主题相似度、新型度和热点度等指标来确定研究前沿主题。使用主题模型能够合理地将主题词划分到各个主题之中,能够起到一定的挖掘效果。但通过主题模型算法得到的主题没有确定的名称,往往需要通过人工进行命名。同时,一个前沿主题中所包含的词语,也并非全部都是前沿科技主题词,这也是主题模型算法的不足之处。

除此之外,周群等[20]认为应当基于更有时效性的科技媒体新闻报道、评论和观点文章筛选重要施引文献,识别和预测具有潜在影响力的最新研究成果或进展。孙震[21]利用文献的外部特征统计,辅以学术出版商提供的论文下载统计数据,来判断和识别前沿的研究主题。

2.3 众包技术

2005 年,Howe 和 Robinson 提出了“众包”这个概念,用来描述企业利用互联网将工作外包给一群人的过程。首次发表众包相关内容的论文把众包定义成“在线地分布式地解决问题或进行某项生产的过程”[22]。在某些情况下,贡献者会获得金钱或者某种认可作为奖励,但更多的情况下,众包行为

是贡献者的一种善意，或者贡献者能从中获得一种满足感。

一个最为常见的众包案例就是 Captcha 验证码[23]。网络中每天都会有大量的验证码被使用，用来判断请求是否来自一个真实的人。验证码往往也会被处理成各种形状，并添加背景干扰来增大机器识别的难度。而 Captcha 验证码则利用了人们的大量输入，实现了免费的古籍文字的数字化工作。Captcha 一次会展示两个古籍文字的验证码，一个验证码是已知答案的，另一个是不确定答案。用户输入两个验证码后，Captcha 会检测已知答案的部分是否正确，如果正确，则记录下另一部分的答案。Captcha 验证码就是一个众包平台，将古籍文字的识别工作分发给输入验证码的用户。

3　前沿科技关键词挖掘算法的框架

前沿科技关键词挖掘算法主要包含科技文献的预处理、科技关键词的粗提取、前沿科技关键词的筛选与提取 3 个部分。算法的主要框架如图 1 所示。

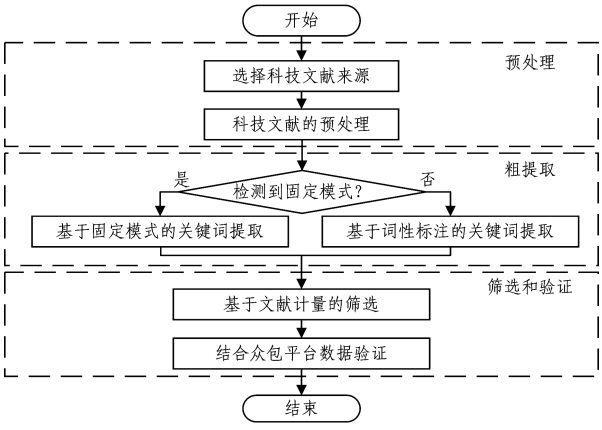

图 1　前沿科技关键词挖掘算法的框架

Fig. 1　Framework of frontier scientific keyword extraction algorithm

在科技文献的预处理阶段，算法根据指定的文献所属领域与文献来源，提取出相应的文献。之后算法会对这些文献中的句子进行词性标注，并生成句法树，从而为后续的步骤做准备。

在科技关键词的粗提取阶段，算法通过检测文献中的句子是否出现固定的句子模式，来选择使用基于固定模式的关键词提取，或基于词性标注的关键词提取，以完成粗提取操作。完成粗提取步骤后，得到的结果即为潜在的科技关键词。

在前沿科技关键词的筛选与提取阶段，算法对上一步获得的潜在的科技关键词进行了文献计量的筛选，并进一步结合众包平台的数据进行了验证，从而得到最终的前沿科技关键词结果。

下文将详细介绍算法各个部分的详细内容。

4　科技文献的选取和预处理

4.1　前沿科技关键词的定义

由于众多学者对前沿科技的定义有各种不同的见解，本文首先需要确定前沿科技关键词的定义，以免给读者造成不必要的误解和困惑。通过对许多学者的不同观点进行整合，并结合数据库中现有的学术论文的实际情况，本文中的前沿关键词指的是具有如下特性的科技主题词：

(1)高端性。前沿科技关键词应当选取自领域内最为权威或最具有影响力的会议、期刊文献，而不应从大量的文献中进行挑选。

(2)时效性。前沿科技关键词应当出现在最近一段时间(一般不超过一年)的文献中，以保证选取的前沿科技关键词能代表最新的领域内的研究内容。

(3)代表性。前沿科技关键词能够代表一篇论文(例如，关键词是某种新提出的研究方法)或一类论文(例如，关键词是某种新提出的研究模型)。

(4)非凡性。前沿科技关键词应当是一个不常见的词汇，在日常生活中很少使用这类词汇；或者是一个日常词汇，但含义与日常词汇的含义不相同。

4.2　文献来源的选取

根据本文对前沿科技关键词的定义，在文献来源的选取上须有一定的要求。这里以计算机科学领域和生物医药领域为例，给出选取文献来源的一些参考建议。

在计算机科学领域，主要的前沿技术与顶级的研究成果往往是通过会议发表的。因此在选取文献来源时，应主要挑选在顶级会议上发表的文献。顶级会议的列表可以参见《中国计算机学会推荐国际学术会议和期刊目录》。

生物医药领域的主要研究成果都发表在各大期刊上，因此可以选取领域内影响因子较高的期刊作为前沿科技关键词的文献来源。类似地，对于其他的领域，也可以参考领域内权威的排名或影响因子的排名来选取文献来源。

4.3　科技文献的预处理

由于本文需要从科技文献中提取出前沿科技关键词，因此需要对文献进行预处理，以便后续算法能提取出合适的前沿科技关键词。

每篇科技文献的篇幅往往在 10 页以上，提取前沿科技关键词的过程并不需要对文献的全部内容做完整的处理，只需要提取其中的重要信息即可。而文章的标题和摘要往往指明了文献所使用的关键技术和方法，文献的关键词通常也会指明文献所使用的技术，因此预处理的首要步骤是提取出文献标题、摘要和关键词。

前沿科技关键词，往往指一项新的研究技术或模型，其词性应为名词。在预处理过程中，可以对文献的标题和摘要中的每一句话分别做自然语言处理中的词性标注，从而得到每一个词语的词性，并生成句法树，以便后续算法进行基于词性的处理。

5　科技关键词的粗提取

科技关键词的粗提取是将预处理得到的文献标题和摘要内容进行分析和处理，从而得到科技关键词的过程。可以将预处理获得的文献关键词直接作为科技关键词，并加入到粗提取结果中。

5.1　基于固定模式的关键词提取

这种方式主要应用于标题文本。对于论文的标题，经常会出现以下固定模式：

(1)SparkR:Scaling R Programs with Spark

(2)ReproZip:Computational Reproducibility With Ease

(3)Shasta:Interactive Reporting At Scale

(4)EmptyHeaded:A Relational Engine for Graph Processing

(5)iOLAP:Managing Uncertainty for Efficient Incremental OLAP

(6)Simba:Efficient In-Memory Spatial Analytics

从上述例子中可以发现,这些标题中都含有冒号。冒号之前的内容是作者给自己发现的前沿技术的命名,或者是新的研究模型的命名;冒号之后的内容是对前沿技术的简单描述,或者是对研究模型的简单介绍。因此,冒号之前的内容可以认为是前沿科技关键词,冒号之后的内容往往是解释性的文字说明。应用这种固定模式的标题,能方便快捷地提取出一些科技关键词。

但是,并不是所有满足该固定模式的标题都能正确提取出前沿科技关键词,这里举出一个反例:

To Join or Not to Join?:Thinking Twice about Joins before Feature Selection

可以看出,在这类反例标题中,冒号之前的内容较长,多数时候冒号之前的内容不是一个名词或名词词组,因此可以通过判断冒号之前的内容是否是一个完整的句子,并且检查冒号之前的内容的长度,来筛选掉这样的反例。

该部分的伪代码如图 2 所示。

```
Input:sentence
Output:keyword
if sentence 中包含冒号 then
    s ← sentence 中冒号之前的部分
    if s 不是一个句子 and s 的长度<=5 then
        keyword ← s
        return keyword
    end if
end if
```

图 2 基于固定模式的关键词提取伪代码

Fig. 2 Pseudo code of keyword extraction based on fixed pattern

5.2 基于词性标注的关键词提取

对于没有固定模式的标题,以及摘要中的文字内容,就需要借助基于词性标注的关键词提取方法来提取前沿科技关键词。

在预处理阶段,已经完成了在句子中识别词语的词性的工作,下面在此基础上提取关键词或词组。

通过词性的标注可以轻松地找到句子中的名词。图 3 给出了一个词性标注的例子。在 Penn Treebank 词性标签中,有关名词的标记有 4 个:NN(名词的单数形式或不可数名词)、NNS(名词的复数形式)、NNP(专有名词的单数形式)、NNPS(专有名词的复数形式)。在句子的词性标注中,连续的名词标记意味着名词词组的出现。在图 3 中,可以提取到的名词和名词词组有:Query Construction,Manipulation 和 Nested Relational Results。

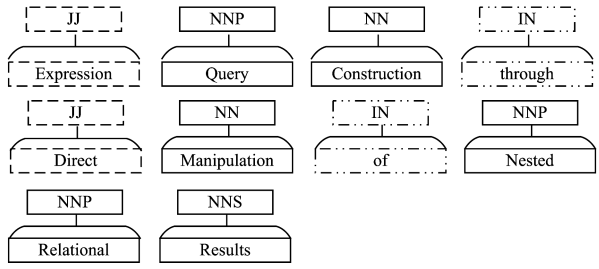

图 3 词性标注结果(1)

Fig. 3 POS tagging result(1)

在科技关键词中,名词或名词词组是不可缺少的一部分。同时,关键词中需要有定语对名词进行限定,这样才能使关键词的含义更加准确和清晰。在 Penn Treebank 词性标签中,有关定语的词性标记有 JJ(形容词的普通形式)、JJR(形容词的比较级)、JJS(形容词的最高级)、VBG(动名词或动词的现在进行时)、VBN(动词的过去分词)等。给上述的名词词组添加定语,可以得到更加准确的关键词。在图 3 中,可以提取到的关键词有:Expressive Query Construction,Direct Manipulation 和 Nested Relational Results。

总而言之,在句子中查找关键词的方法是寻找句子的词性标注序列中的固定形式,其正则表达式如下:

(JJ | JJR | JJS | VBN | VBG)?(NN | NNS | NNP | NNPS)+

基于自然语言处理的关键词提取的伪代码如图 4 所示。

```
Input:sentence
Output:keywords
nounPhrases←sentence 中的名词组
for nounPhrase in nounPhrases do
  word ←nounPrase 的前一个单词
  if word 是定语 then
      keyword ←word + nounPhrase
  else
      keyword ←nounPhrase
  end if
  keywords←keywords+keyword
end for
return keywords
```

图 4 基于自然语言处理的关键词提取伪代码

Fig. 4 Pseudo code of keywords extraction based on NLP

6 前沿科技关键词的筛选与提取

前沿科技关键词是科技关键词的一个部分,通过基于文献计量的科技检测方法,并结合众包平台数据验证的方式,对粗提取得到的科技关键词进行判断和筛选,从而获得最终的前沿科技关键词。

6.1 基于文献计量的筛选方法

对粗提取过程中产生的科技关键词进行文献计量分析,根据相应的计量学特征,筛选掉不符合前沿科技关键词特征的词语。

将粗提取过程中产生的科技关键词作为检索词,并在文

献数据库中进行检索，将命中的文献数量按年份进行聚类，可以得到每一年的文献命中数量 H_i。将每一年的文献命中数量 H_i 除以每一年数据库收录的文献总数 S_i，并进行归一化处理。为了方便计算和处理，再将每一年的比值乘上一个常数 C，即可得到数据 K_i：

$$K_i = C \cdot \frac{H_i}{S_i} \tag{1}$$

可以认为，数据 K_i 是该科技关键词的年度热度值。因为 K_i 表示了某一年中检索词在单位文献中的命中次数，将每一年的热度数据绘制成折线图后，可以得到科技关键词的趋势热度图。

根据科技关键词的热度趋势，可以对该关键词进行基本的判断。如果近几年内该关键词的热度较高，则该关键词应当被认为是热点关键词，而不应当是前沿技术关键词。本文所期望的前沿技术关键词的热度趋势应当满足如下条件：

(1)近几年内，年度热度的平均值应当处于较低水平；

(2)当年的热度值应大于 0；

(3)热度曲线的走势呈向上或保持水平的趋势。

即前沿技术趋势图应类似于如图 5 所示的形状。对于条件(1)，若要确定年度热度值是否较低，则需要确定阈值，如果低于该阈值则认为年度热度值较低。对于不同的领域，相应的阈值也应当不同，领域内阈值的确定需要在具体的使用场景中进行尝试和确定。

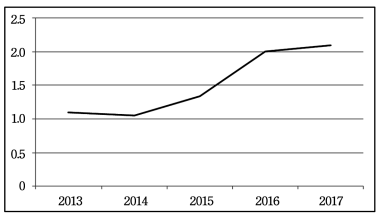

图 5　热度趋势示例

Fig. 5　Example of heat trend

基于文献计量的关键词筛选伪代码如图 6 所示。

```
Input:word,threshold1,threshold2
Output:isNewTech
trend←word 的近 5 年趋势热度值
if trend 的均值 >threshold1 then
    isNewTech=false
    return isNewTech
end if
if trend 的当年热度值=0 then
    isNewTech=false
    return isNewTech
end if
if trend 的前 4 年热度值均<0 then
    isNewTech=true
    return isNewTech
end if
isNewTech=false
return keywords
```

图 6　基于文献计量的关键词筛选伪代码

Fig. 6　Pseudo code of keyword screeing based on bibliometric

6.2　结合众包平台的数据验证

在完成基于文献计量的关键词筛选之后，需要结合众包平台的数据进行验证，以得到更优的结果。

通过上述的关键词提取和筛选流程，可以从句子中提取出关键词，即含有定语的名词词组。在很多情况下，这个关键词是十分普通且常见的或者涵义十分宽泛的，这样的词语往往不能作为前沿技术关键词。图 7 给出了一个例子，经过科技关键词提取流程处理后，得到的词语是 Fast Algorithm，这个词在日常生活中是十分常见的，不能作为前沿技术关键词。

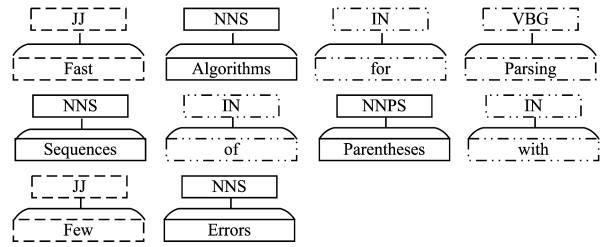

图 7　词性标注结果(2)

Fig. 7　POS tagging result (2)

为了判断一个词语是否是普通且常见的，可以借助 Google Trends 上的数据。Google 趋势是 Google 公司提供的一项公众服务，是针对谷歌搜索数据的统计和分析平台，提供统计谷歌搜索中的检索词的检索频率的功能。其在网页中用一个图表进行展示，横轴表示时间，纵轴表示检索词的检索频率。除此之外，还可以通过它查看不同地区、不同语言的检索频率统计，以及下载检索频率的数据。如果一个词语是日常生活中的常见词，则人们往往会在搜索引擎中检索该词。如果一个词在很长一段时间内在谷歌搜索中被大量搜索，则可以认为这个词是一个常见词。图 8 给出了 Fast Algorithm 在 Google Trends 中的查询结果，由此可以判断出 Fast Algorithm 是一个常见词。

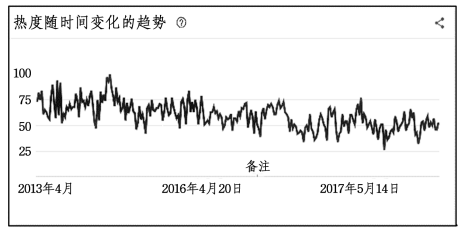

图 8　Google Trends 中的查询结果

Fig. 8　Query result of Google Trends

在验证得到关键词不适合作为前沿技术主题词时，就需要调整词语，以得到一个更好的关键词。在进行词性标注的预处理时，得到了句法树。根据句法树，不仅可以知道标注的各个词的词性，还能知道句子中的每个词组和相应的词性成分。图 9 给出了一个句法树的例子。

有些前沿技术不是一个全新的概念，而是对已有技术的改善和创新。这样的前沿技术对应的主题词往往包含许多定语。通过向已有的名词短语中添加定语，可以避免遗漏这类前沿技术主题词。具体的做法是：向前向后扩展关键词长度，增加定语，并将增长的词语进行 Google Trends 验证，重复此

过程直到找到一个较为不常见的词。例如，验证得到 Fast Algorithm 是一个常见词，则需要扩展关键词的长度，从而得到新的关键词 Fast Algorithm for Parsing Sequences，继续验证 Fast Algorithm for Parsing Sequences，发现 Google Trends 查询到的搜索数据很少，则可以认为这个关键词是一个不常见的词。

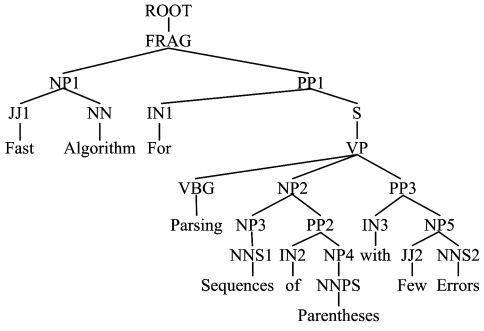

图 9 句法树标注结果

Fig. 9 Annotation result of syntax tree

算法的伪代码如图 10 所示。

```
Input:word,sentence
Output:keyword
while word 在 Google Trends 中出现频繁 do
    word ←在 sentence 中扩展 word 长度
end while
return keyword
```

图 10 结合众包平台数据验证过程的伪代码

Fig. 10 Pseudo code of crowdsourcing checking procedure

7 实验与评价

本文使用提出的前沿技术关键词挖掘方法来测试英文文献中的前沿技术关键词提取效果，并通过专家评定的方式衡量提取的准确度。

实验所采用的文献数据是计算机领域和生物医药领域的部分会议和期刊文献数据。计算机领域的数据来源为《中国计算机学会推荐国际学术会议和期刊目录》中的 A 类会议文献；生物医药领域的数据来源为 SCI 影响因子排名前 20 的期刊文献。

实验所采用的词性标注工具是 Stanford CoreNLP[24] 工具包，它提供了一套自然语言分析的工具，包括词元化、标注词性、标注词语间的依赖关系、情感分析、实体提取、标准化日期时间等。通过自定义一个处理流水线，可以定制相应的处理流程。本文利用了 CoreNLP 提供的标注词性和句法树的功能。

基于文献计量的筛选方法采用的数据库是收录有三千多万条期刊、会议、学位论文、项目文献数据的综合性数据库。

7.1 度量标准

本文采用查准率、查全率和 F-score 作为评估标准，其定义如下：

$$p=\frac{TP}{TP+FP}\times 100\% \tag{2}$$

$$r=\frac{TP}{TP+FN}\times 100\% \tag{3}$$

$$F=\frac{2pr}{p+r} \tag{4}$$

其中，TP 表示挖掘算法和专家共同认为的前沿科技关键词数量，FP 表示挖掘算法认为是前沿科技关键词而专家不认同的词语数量，FN 表示专家认为是前沿科技关键词而挖掘算法没有挖掘出的词语数量，TN 表示挖掘算法和专家共同认为的不属于前沿科技关键词的数量，$TP+FP$ 表示挖掘算法挖掘出的所有前沿科技关键词的数量，$TP+FN$ 表示专家选出的前沿科技关键词的数量。

7.2 挖掘算法的挖掘效果实验

通过随机抽取的方式，在计算机领域和生物医药领域中各选取 5 个子领域，每个子领域随机抽取 50 篇文献作为一组，通过专家阅读外文文献，来判断前沿技术关键词。

计算机领域选取的子领域为：数据库与数据挖掘、人工智能、计算机网络、网络与信息安全、计算机图形学与多媒体。这些子领域分别为计算机领域的 1－5 组。

生物医药领域选取的子领域为：癌症、免疫、肿瘤、神经内科、心理学。这些子领域分别为生物医药领域的 1－5 组。

将人工判别前沿科技关键词的结果与前沿科技关键词挖掘算法得到的结果进行比较，得到的正确率如表 1 所列。

表 1 挖掘正确性统计

Table 1 Mining correctness statistics

领域	组号	p	r	F
计算机	1	0.615	0.930	0.741
	2	0.727	0.842	0.780
	3	0.778	0.897	0.833
	4	0.745	0.844	0.792
	5	0.784	0.763	0.773
生物医药	1	0.619	0.650	0.634
	2	0.784	0.558	0.652
	3	0.854	0.761	0.805
	4	0.703	0.605	0.650
	5	0.750	0.489	0.592

从表 1 中可以看出，在不同的领域，前沿科技关键词挖掘算法的效果也不尽相同，在计算机领域得到的挖掘效果要优于生物医药领域的挖掘效果。在同一领域中的不同子领域，挖掘效果的差异也较大。这是因为在不同的领域中，文献的内容差异很大；在不同的文献来源中，对文献的格式与内容要求各有不同；对于不同的作者，行文的风格也各不相同。这些因素都会对基于自然语言处理的过程造成一定的影响，从而使挖掘效果在不同的领域和子领域间各有不同。

通过比较人工标注的结果和前沿科技挖掘算法产生的结果，能够发现一些造成前沿科技挖掘算法正确率降低的原因：1)受数据库收录数据的影响，在进行文献计量的筛选时，数据库收录数据不全会导致热度趋势不准确，从而影响挖掘算法的正确率；2)CoreNLP 工具包对词性的标注有时不准确，导致前沿技术挖掘算法在提取关键词时会出现错误，从而影响正确率，因此后续可以采用外文科技文献专门训练的词性标注工具，以提高正确率。

7.3　文献计量法和众包验证法的筛选效果实验

采用与 7.2 节中相同的样本分组，分别统计粗筛选之后的科技关键词数目、文献计量法筛选出的科技关键词数目，以及众包验证法筛选得到的最终结果数目，并进行分析。计算机领域和生物医药领域的筛选效果图如图 11、图 12 所示。这里以组号为横坐标，以文献计量法筛选出的科技关键词数目占粗筛选的数目百分比、众包验证法筛选得到的最终结果数目占粗筛选之后的数目百分比为纵坐标。

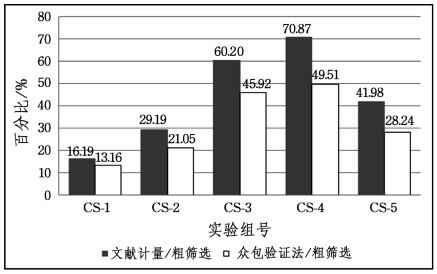

图 11　计算机领域筛选效果柱状图

Fig. 11　Histogram of screening results in CS Field

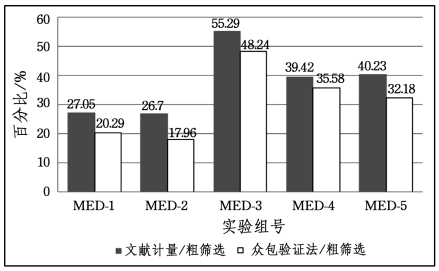

图 12　生物医药领域筛选效果柱状图

Fig. 12　Histogram of screening results in BioMed Field

从图中可以看出，基于文献计量法的筛选具有很好的筛选效果，众包验证法在文献计量法的筛选结果之后，起到了进一步的筛选作用。在文献计量法和众包验证法的共同作用下，能够从粗筛选的结果中剔除大量不符合前沿技术条件的词语。

7.4　文献计量法的参数选择实验

选取 7.2 节中的 250 篇计算机领域文献和 250 篇生物医药领域文献作为样本，选择不同的阈值，比较最终得到的前沿科技关键词结果数目的查准率、查全率和 F-score。计算机领域和生物医药领域的参数选择效果如图 13、图 14 所示。图中以阈值为横坐标，以查准率、查全率和 F-score 为纵坐标。

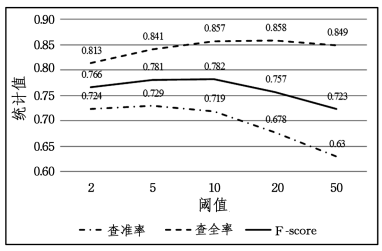

图 13　计算机领域参数选择效果折线图

Fig. 13　Line chart of parameter selection in CS Field

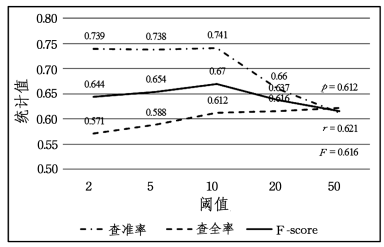

图 14　生物医药领域参数选择效果折线图

Fig. 14　Line chart of parameter selection in BioMed Field

从图中可以看出，阈值只有选取在一个合适的范围内，才能得到更优的结果。如果阈值选取得过小，将导致查准率降低，影响最终的结果；如果阈值选取得过大，将使查全率有所下降，从而使最终的 F-score 降低。

7.5　挖掘算法评价

通过基于固定模式的关键词提取、基于自然语言处理的关键词提取和结合众包平台数据验证 3 种方法，可以实现从数据库的文献信息中提取前沿技术关键词。其优点是方法简单易行，适用于小型数据规模或没有引用信息的文献数据；不足之处在于每次验证关键词是否是前沿技术关键词时，都需要联网抓取数据，效率不高。

结束语　本文对国内外的前沿科技挖掘方法和挖掘平台的现状进行了分析。通过前沿技术关键词挖掘算法可以从外文文献中提取出前沿技术关键词，结合了众包平台的数据验证，能够取得不错的效果。通过分析得到前沿科技关键词，可以帮助研究人员了解当前某领域的最新前沿，有助于他们发现新的研究方向，并进行针对性的研究，帮助初涉研究者了解领域最新动态，帮助专业研究人员寻找突破创新点。

但是，现阶段仍然有一些地方需要改进。前沿科技关键词挖掘算法能够比较准确地提取前沿科技关键词，在后续提高准确率的工作中，可以从增加固定模式和提高词性标注的准确度两个方面入手。需要进一步对外文文献进行分析，找出更多的固定模式，增加提取前沿科技关键词的准确性。同时，后续可以利用外文科技文献书籍进行专门的训练，以得到一个质量更好的词性标注工具。

参 考 文 献

[1]　JINHA A E. Article 50 million: an estimate of the number of scholarly articles in existence[J]. Learned Publishing, 2010, 23(3): 258-263.

[2]　WARE M, MABE M. The STM report: An overview of scientific and scholarly journal publishing[R]. Nebraska: Digital Commons at University of Nebraska-Lincoln, 2015.

[3]　SCOTT J. Social Networks : Critical Concepts in Sociology (Vol. 4)[M]. London: Routledge, 2002: 328-331.

[4]　BRAAM R R, MOED H F, VAN RAAN A F J. Mapping of science by combined co-citation and word analysis I. Structural aspects[J]. Journal of the American Society for information Science, 1991, 42(4): 233.

[5]　SMALL H, GRIFFITH B C. The structure of scientific litera-

tures I:Identifying and graphing specialties[J]. Science studies, 1974,4(1):17-40.

[6] PERSSON O. The intellectual base and research fronts of "jasis" 1986-1990[J]. Journal of the American Society for Information Science,1994,45(1):31.

[7] CHEN C. CiteSpace II:Detecting and visualizing emerging trends and transient patterns in scientific literature[J]. Journal of the American Society for information Science and Technology,2006,57(3):359-377.

[8] ZHU L,ZHAO R X,KOU Y T,et al. Study on Integrated Mode of Science and Technology Monitoring Base on Literature[J]. Digital Library Forum,2015(10):53-57. (in Chinese)
朱亮,赵瑞雪,寇远涛,等.一种基于文献的综合科技监测模式研究[J].数字图书馆论坛,2015(10):53-57.

[9] KLEINBERG J. Bursty and hierarchical structure in streams [C]// Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2002: 91-101.

[10] ZHOU W J. The Criterion Related Validity of Research Frontier Exploration:a Co-words Analysis based on the Natural Language Processing[J]. Library and Information,2018,38(1):1-7. (in Chinese)
周文杰.研究前沿探测的效标关联效度研究:基于自然语言处理[J].图书与情报,2018,38(1):1-7.

[11] GENG H Y,XIAO X T. The Research Progress and Trends of Cocitation Analysis in Foreign Countries[J]. Journal of Information,2006,25(12):68-70. (in Chinese)
耿海英,肖仙桃.国外共引分析研究进展及发展趋势[J].情报杂志,2006,25(12):68-70.

[12] SMALL H. A SCI-MAP case study:Building a map of AIDS research[J]. Scientometrics,1994,30(1):229-241.

[13] SHENG L. Recognize the Fronts and Trends of Biology and Medical Research Domain [D]. Beijing: Academy of Military Medical Sciences,2013. (in Chinese)
盛立.生物医学领域研究前沿识别与趋势预测[D].北京:中国人民解放军军事医学科学院,2013.

[14] JIANG Y. A Co-Word Analysis of Bibliometric in 1995 ~ 2004 [J]. Journal of the China Society for Scientific and Technical Information. 2006,25(4):504-512. (in Chinese)
蒋颖.1995~ 2004 年文献计量学研究的共词分析[J].情报学报,2006,25(4):504-512.

[15] ZHENG Y N,XU X Y,LIU Z H. Study on the Method of Identifying Research Fronts Based on Keywords Co-occurrence[J]. Library and Information Service, 2016, 60(4): 85-92. (in Chinese)
郑彦宁,许晓阳,刘志辉.基于关键词共现的研究前沿识别方法研究[J].图书情报工作,2016,60(4):85-92.

[16] AN X Y,ZHONG H. The Theoretical Summary of Scienceand Technology Monitoring and the Comparative Analysis of Application System [J]. Information Studies: Theory & Application, 2010,33(5):124-128. (in Chinese)
安新颖,钟华.科技监测的理论综述与应用系统对比分析[J].情报理论与实践,2010,33(5):124-128.

[17] ZHONG H X. Review on Emerging Trend Detection[J]. Journal of Modern Information,2017,37(12):28. (in Chinese)
钟辉新.新兴趋势探测研究综述[J].现代情报,2017,37(12):28.

[18] FENG J,ZHANG Y Q. Research on the Method of Detecting and Analyzing Scientific Fronts Based on LDA and Ontology [J]. Information Studies: Theory & Application, 2017, 40(8): 49-54. (in Chinese)
冯佳,张云秋.基于 LDA 和本体的科学前沿识别与分析方法研究[J].情报理论与实践,2017,40(8):49-54.

[19] BAI R J,LENG F H,LIAO J H. A Method of Detecting Research Front Based on Subjects Comparison of Multiple Data Sources[J]. Information Studies: Theory & Application, 2017, 40(8): 43-48. (in Chinese)
白如江,冷伏海,廖君华.一种基于多数据源主题对比的科学研究前沿识别方法[J].情报理论与实践,2017,40(8):43-48.

[20] ZHOU Q,ZHOU Q J,LENG F H. Research and Demonstration of the Method of Identifying Research Fronts Based on Media of Science and Technology [J]. Journal of Modern Information, 2018,38(2):62-68. (in Chinese)
周群,周秋菊,冷伏海.基于科技媒体视角的研究前沿识别方法研究与实证[J].现代情报,2018,38(2):62-68.

[21] SUN Z. Study on the Integrated Model of Research Front Based on the Multi-Source Data of Scientific Papers[J]. Journal of Intelligence,2016,35(8):95-100. (in Chinese)
孙震.基于科学论文多源数据的研究前沿集成识别模型研究[J].情报杂志,2016,35(8):95-100.

[22] BRABHAM D C. Crowdsourcing as a model for problem solving:An introduction and cases[J]. Convergence, 2008, 14(1): 75-90.

[23] KAZAI G. In search of quality in crowdsourcing for search engine evaluation[C]// European Conference on Information Retrieval. Springer Berlin Heidelberg,2011:165-176.

[24] MANNING C,SURDEANU M,BAUER J,et al. The Stanford CoreNLP natural language processing toolkit[C]// Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics:System Demonstrations. 2014:55-60.