(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 109726298 B
(45)授权公告日 2020.12.29

(21)申请号 201910015944.9

(22)申请日 2019.01.08

(65)同一申请的已公布的文献号
申请公布号 CN 109726298 A

(43)申请公布日 2019.05.07

(73)专利权人 上海市研发公共服务平台管理中心
地址 200235 上海市徐汇区钦州路100号2号楼4楼

(72)发明人 刘晋元 胡寅骏 朱悦 赵燕 徐旻昕 王茜

(74)专利代理机构 上海光华专利事务所(普通合伙) 31219
代理人 高彦

(51)Int.Cl.
G06F 16/36(2019.01)
G06F 16/34(2019.01)

(56)对比文件
CN 106844658 A,2017.06.13
CN 106776711 A,2017.05.31
焦晓静等.知识图谱在科技情报研究中的应用模型构建.《图书情报知识》.2017,(第3期),

审查员 杨鹏

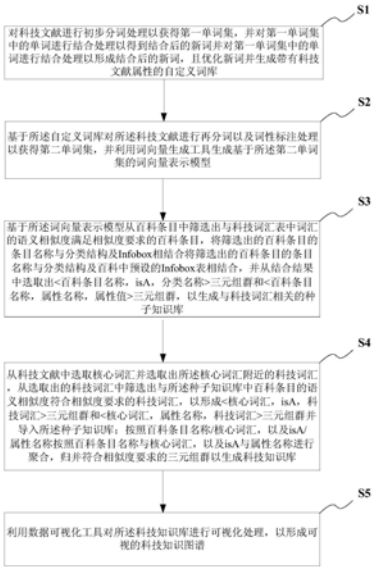
权利要求书3页 说明书10页 附图6页

(54)发明名称

适用于科技文献的知识图谱构建方法、系统、终端及介质

(57)摘要

本发明提供适用于科技文献的知识图谱构建方法、系统、终端及介质，用于根据特定来源数据构建相应的适用于科技文献的知识图谱。本发明提供的技术方案与科技文献的非结构化文本特性有极高的契合度，分词单元由结合单词以及人工校验来完善，有利于科技文献中的专业符合词的抽取，并利用百科自动抽取种子知识库，从而节省了大量的初期手工维护种子知识库的成本。



CN 109726298 B

1.一种适用于科技文献的知识图谱构建方法,其特征在于,包括:

S1:对科技文献进行初步分词处理以获得第一单词集,并对第一单词集中的单词进行结合处理以形成结合后的新词,且修正新词并生成带有科技文献属性的自定义词库;

S2:基于所述自定义词库对所述科技文献进行再分词以及词性标注处理以获得第二单词集,并利用词向量生成工具生成基于所述第二单词集的词向量表示模型;

S3:基于所述词向量表示模型从百科条目中筛选出与科技词汇表中词汇的语义相似度满足相似度要求的百科条目,将筛选出的百科条目的条目名称与分类结构及百科中预设的Infobox表相结合,并从结合结果中选取出<百科条目名称,isA,分类名称>三元组群和<百科条目名称,属性名称,属性值>三元组群,以生成与科技词汇相关的种子知识库;

S4:从科技文献中选取核心词汇并选取出所述核心词汇附近的科技词汇,从选取出的科技词汇中筛选出与所述种子知识库中百科条目的语义相似度符合相似度要求的科技词汇,以形成<核心词汇,isA,科技词汇>三元组群和<核心词汇,属性名称,科技词汇>三元组群并导入所述种子知识库;按照百科条目名称与核心词汇,以及isA与属性名称进行聚合,归并符合相似度要求的三元组群以生成科技知识库;

S5:利用数据可视化工具对所述科技知识库进行可视化处理,以形成可视的科技知识图谱;

其中,步骤S1包括:

S101:利用分词工具对科技文献进行初步分词处理,以获得所述第一单词集;

S102:计算所述第一单词集中单词之间的互信息量,并将满足结合条件的互信息量所对应的单词进行结合处理;

S103:重复上述步骤S101或者S102;

S104:利用正则表达式修正结合后的新词并将修正后的新词导入所述分词工具的自定义词库中;

S105:基于所述自定义词库对科技文献再次进行分词处理,并根据词性对此次分词后的单词进行结合处理且将结合后的词导入所述自定义词库中,以生成所述带有科技文献属性的自定义词库。

2.根据权利要求1所述的适用于科技文献的知识图谱构建方法,其特征在于,步骤S102中单词之间的互信息量的计算公式为:

$$Score(w_i, w_j) = \frac{freq(w_i, w_j) - \delta}{freq(w_i) freq(w_j)}$$;当Score(w_i,w_j)大于预设阈值时将单词w_j和单词w_j做结合处理;

其中,freq(w_i)、freq(w_j)以及freq(w_i,w_j)分别表示单词w_i的出现频度、单词w_j的出现频度以及单词w_i和单词w_j前后同时出现的频度;δ是防止特别低频的单词被结合到一起的折现系数。

3.根据权利要求1所述的适用于科技文献的知识图谱构建方法,其特征在于,步骤S104中利用正则表达式修正结合后的新词的方式包括:

利用正则表达式选取出带有特殊字符的单词以进行修正;其中,所述特殊字符包括:以“该”、“及”、“使”、“其”、“为”、“或”、“的”、“和”、“在”、“将”、“与”或“用”词开头或结尾的单词。

4.根据权利要求1所述的适用于科技文献的知识图谱构建方法，其特征在于，步骤S2包括：

S201：利用分句工具对科技文献按照句子进行切分；

S202：利用分词工具对切分后的句子进行分词处理并按照自定义词库进行词性标注处理，将分词及词性标注处理的处理结果作为科技文献语料库输入至词向量生成工具中，以生成科技文献的词向量表示模型。

5.根据权利要求1所述的适用于科技文献的知识图谱构建方法，其特征在于，步骤S3包括：

S301：从经分词处理后的科技文献中选取名词、名词短语及动名词短语以建立科技词汇表；

S302：获取百科中预设的条目dump文件以及分类链接dump文件并导入数据库中；

S303：从百科条目中筛选出与科技词汇表中词汇的语义相似度大于0.6且百科中的预设字段page_namespace的值为0的条目，并通过与百科中预设的revision表、text表连接的方式从筛选出的条目中选取百科的内部编号、条目名称及文本字段，以生成百科的page_refined表；

S304：利用百科中预设的categorylinks表和所述page_refined表选出<百科条目名称，分类名称>二元组群并转化为所述<百科条目名称，isA，分类名称>三元组群；

S305：利用所述page_refined表中用于表示正文内容的字段找到与Infobox表相关联的数据资源，选取其中的<属性名称，属性值>二元组群并转化为所述<百科条目名称，属性名称，属性值>三元组群。

6.根据权利要求1所述的适用于科技文献的知识图谱构建方法，其特征在于，步骤S4包括：

S401：选取科技文献中的科技词汇并计算科技词汇表中所有单词的tf-idf值；其中，所述tf-idf值的计算公式为：$idf(w) = log\frac{1+|D|}{1+df(w)}$；|D|为文献总数，df(w)为包含有单词w的文献数量；

S402：计算科技文献每个段落中的科技词汇的tf-idf值并按倒序排序，并选取每个段落的核心词汇；

S403：计算所述核心词汇与所述种子知识库中百科条目的语义相似度，选取语义相似度大于0.5的核心词汇所在的句子；

S404：计算句子中其它科技词汇与所述核心词汇所对应的三元组群中分类名词或属性值的语义相似度；其中，若满足分类名称语义相似度要求，则组建<核心词汇，isA，科技词汇>三元组群；若满足属性值相似度要求，则组建<核心词汇，属性名称，科技词汇>三元组群；

S405：将<核心词汇，isA，科技词汇>三元组群和<核心词汇，属性名称，科技词汇>三元组群加入所述种子知识库；将所述种子知识库中的三元群组统一视为<科技词汇1，词关系，科技词汇2>，并按照科技词汇1与词关系进行聚合；将聚合后的同一族群中语义相似度满足相似度要求的科技词汇2或者满足字符串相似度要求的字符串进行归并，且选择族群中字符串长度最长者作为代表词汇并记录科技词汇2的多种表达，以最终形成科技知识库。

7.一种适用于科技文献的知识图谱构建系统，其特征在于，包括：

词库生成模块，用于对科技文献进行初步分词处理以获得第一单词集，并对第一单词集中的单词进行结合处理以形成结合后的新词，且优化新词并生成带有科技文献属性的自定义词库；

词向量生成模块，用于基于所述自定义词库对所述科技文献进行再分词以及词性标注处理以获得第二单词集，并利用词向量生成工具生成基于所述第二单词集的词向量表示模型；

种子知识库生成模块，用于基于所述词向量表示模型从百科条目中筛选出与科技词汇表中词汇的语义相似度满足相似度要求的百科条目，将筛选出的百科条目的条目名称与分类结构及百科中预设的Infobox表相结合，并从结合结果中选取出<百科条目名称，isA，分类名称>三元组群和<百科条目名称，属性名称，属性值>三元组群，以形成与科技词汇相关的种子知识库；

科技知识库生成模块，用于从科技文献中选取核心词汇并选取出所述核心词汇附近的科技词汇，从选取出的科技词汇中筛选出与所述种子知识库中百科条目的语义相似度符合相似度要求的科技词汇，以形成<核心词汇，isA，科技词汇>三元组群和<核心词汇，属性名称，科技词汇>三元组群并导入所述种子知识库；按照百科条目名称与核心词汇，以及isA与属性名称进行聚合，归并符合相似度要求的三元组群，以形成科技知识库；

知识图谱生成模块，用于利用数据可视化工具对所述科技知识库进行可视化处理，以生成可视的科技知识图谱。

8.一种计算机可读存储介质，其上存储有计算机程序，其特征在于，该程序被处理器执行时实现权利要求1至6中任一项所述的适用于科技文献的知识图谱构建方法。

9.一种电子终端，其特征在于，包括：处理器及存储器；

所述存储器用于存储计算机程序，所述处理器用于执行所述存储器存储的计算机程序，以使所述电子终端执行如权利要求1至6中任一项所述的适用于科技文献的知识图谱构建方法。

适用于科技文献的知识图谱构建方法、系统、终端及介质

技术领域

[0001] 本发明涉及知识图谱构建领域，特别是涉及适用于科技文献的知识图谱构建方法、系统、终端及介质。

背景技术

[0002] 知识图谱是显示知识发展进程与结构关系的一系列各种不同的图形，用可视化技术描述知识资源及其载体，挖掘、分析、构建、绘制和显示知识及它们之间的相互关系，是一种利用可视化技术描述知识资源及其载体的语义网络，所形成的语义网络可以用于解决精准搜索、计算文本语义相似度、制作对话机器人或智能问答系统等人工智能领域的课题。近几年来，随着深度学习等新一代机器学习技术的发展，将知识图谱作为深度学习的输入甚至约束的研究也较为盛行。

[0003] 但是，在根据特定来源数据性质制定相应的构建知识图谱方法等知识图谱方面则没有较好的技术解决方案。

发明内容

[0004] 鉴于以上所述现有技术的缺点，本发明的目的在于提供适用于科技文献的知识图谱构建方法、系统、终端及介质，用于解决现有技术无法根据特定来源数据性质制定相应的构建知识图谱等技术问题。

[0005] 为实现上述目的及其他相关目的，本发明提供一种适用于科技文献的知识图谱构建方法，其包括：S1：对科技文献进行初步分词处理以获得第一单词集，并对第一单词集中的单词进行结合处理以形成结合后的新词，且优化新词并生成带有科技文献属性的自定义词库；S2：基于所述自定义词库对所述科技文献进行再分词以及词性标注处理以获得第二单词集，并利用词向量生成工具生成基于所述第二单词集的词向量表示模型；S3：基于所述词向量表示模型从百科条目中筛选出与科技词汇表中词汇的语义相似度满足相似度要求的百科条目，将筛选出的百科条目的条目名称与分类结构及百科中预设的Infobox表相结合，并从结合结果中选取出<百科条目名称，isA，分类名称>三元组群和<百科条目名称，属性名称，属性值>三元组群，以生成与科技词汇相关的种子知识库；S4：从科技文献中选取核心词汇并选取出所述核心词汇附近的科技词汇，从选取出的科技词汇中筛选出与所述种子知识库中百科条目的语义相似度符合相似度要求的科技词汇，以形成<核心词汇，isA，科技词汇>三元组群和<核心词汇，属性名称，科技词汇>三元组群并导入所述种子知识库；按照百科条目名称与核心词汇，以及isA与属性名称进行聚合，归并符合相似度要求的三元组群，以生成科技知识库；S5：利用数据可视化工具对所述科技知识库进行可视化处理，以形成可视的科技知识图谱。

[0006] 于本发明的一实施例中，步骤S1包括：S101：利用分词工具对科技文献进行初步分词处理，以获得所述第一单词集；S102：计算所述第一单词集中单词之间的互信息量，并将满足结合条件的互信息量所对应的单词进行结合处理；S103：重复上述步骤S101或者S102；

S104:利用正则表达式修正结合后的新词并将修正后的新词导入所述分词工具的自定义词库中;S105:基于所述自定义词库对科技文献再次进行分词处理,并根据词性对此次分词后的单词进行结合处理且将结合后的词导入所述自定义词库中,以生成所述带有科技文献属性的自定义词库。

[0007] 于本发明的一实施例中,步骤S102中单词之间的互信息量的计算公式为:$Score(w_i, w_j) = \frac{freq(w_i, w_j)-\delta}{freq(w_i)freq(w_j)}$;当Score($w_i$,$w_j$)大于预设阈值时将单词$w_j$和单词$w_j$做结合处理;其中,freq($w_i$)、freq($w_j$)以及freq($w_i$,$w_j$)分别表示单词$w_i$的出现频度、单词$w_j$的出现频度以及单词$w_i$和单词$w_j$前后同时出现的频度;δ是防止特别低频的单词被结合到一起的折现系数。

[0008] 于本发明的一实施例中,步骤S104中利用正则表达式修正结合后的新词的方式包括:利用正则表达式选取出带有特殊字符的单词以进行修正;其中,所述特殊字符包括:以“该”、“及”、“使”、“其”、“为”、“或”、“的”、“和”、“在”、“将”、“与”或“用”词开头或结尾的单词。

[0009] 于本发明的一实施例中,步骤S2包括:S201:利用分句工具对科技文献按照句子进行切分;S202:利用分词工具对切分后的句子进行分词处理并按照自定义词库进行词性标注处理,将分词及词性标注处理的处理结果作为科技文献语料库输入至词向量生成工具中,以生成科技文献的词向量表示模型。

[0010] 于本发明的一实施例中,步骤S3包括:S301:从进行分词处理后的科技文献中选取名词、名词短语及动名词短语以建立科技词汇表;S302:获取百科中预设的的条目dump文件以及分类链接dump文件并导入数据库中;S303:从百科条目中筛选出与科技词汇表中词汇的语义相似度大于0.6且百科中的预设字段page_namespace的值为0的条目,并通过与百科中预设的revision表、text表连接的方式从筛选出的条目中选取百科的内部编号、条目名称及文本字段,以生成百科的page_refined表;S304:利用百科中预设的categorylinks表和所述page_refined表选出<百科条目名称,分类名称>二元组群并转化为所述<百科条目名称,isA,分类名称>三元组群;S305:利用所述page_refined表中用于表示正文内容的字段找到与Infobox表相关联的数据资源,选取其中的<属性名称,属性值>二元组群并转化为所述<百科条目名称,属性名称,属性值>三元组群。

[0011] 于本发明的一实施例中,步骤S4包括:S401:选取科技文献中的科技词汇并计算科技词汇表中所有单词的tf-idf值;其中,所述tf-idf值的计算公式为:$idf(w) = log\frac{1+|D|}{1+df(w)}$;|D|为文献总数,df(w)为包含有单词w的文献数量;S402:计算科技文献每个段落中的科技词汇的tf-idf值并按倒序排序,并选取每个段落的核心词汇;S403:计算所述核心词汇与所述种子知识库中百科条目的语义相似度,选取语义相似度大于0.5的核心词汇所在的句子;S404:计算句子中其它科技词汇与所述核心词汇所对应的三元组群中分类名词或属性值的语义相似度;其中,若满足分类名称语义相似度要求,则组建<核心词汇,isA,科技词汇>三元组群;若满足属性值相似度要求,则组建<核心词汇,属性名称,科技词汇>三元组群;S405:将<核心词汇,isA,科技词汇>三元组群和<核心词汇,属性名称,科技词汇>三元组群加入所述种子知识库;将所述种子知识库中的三元群组统一视为<科技

词汇1,词关系,科技词汇2>,并按照科技词汇1与词关系进行聚合;将聚合后的同一族群中语义相似度满足相似度要求的科技词汇2或者满足字符串相似度要求的字符串进行归并,且选择族群中字符串长度最长者作为代表词汇并记录科技词汇2的多种表达,以最终形成科技知识库。

[0012] 为实现上述目的及其他相关目的,本发明提供一种适用于科技文献的知识图谱构建系统,其包括:词库生成模块,用于对科技文献进行初步分词处理以获得第一单词集,并对第一单词集中的单词进行结合处理以形成结合后的新词,优化新词并生成带有科技文献属性的自定义词库;词向量生成模块,用于基于所述自定义词库对所述科技文献进行再分词以及词性标注处理以获得第二单词集,并利用词向量生成工具生成基于所述第二单词集的词向量表示模型;种子知识库生成模块,用于基于所述词向量表示模型从百科条目中筛选出与科技词汇表中词汇的语义相似度满足相似度要求的百科条目,将筛选出的百科条目的条目名称与分类结构及百科中预设的Infobox表相结合,并从结合结果中选取出<百科条目名称,isA,分类名称>三元组群和<百科条目名称,属性名称,属性值>三元组群,以形成与科技词汇相关的种子知识库;科技知识库生成模块,用于从科技文献中选取核心词汇并选取出所述核心词汇附近的科技词汇,从选取出的科技词汇中筛选出与所述种子知识库中百科条目的语义相似度符合相似度要求的科技词汇,以形成<核心词汇,isA,科技词汇>三元组群和<核心词汇,属性名称,科技词汇>三元组群并导入所述种子知识库;按照百科条目名称与核心词汇,以及isA与属性名称进行聚合,归并符合相似度要求的三元组群,以形成科技知识库;知识图谱生成模块,用于利用数据可视化工具对所述科技知识库进行可视化处理,以生成可视的科技知识图谱。

[0013] 为实现上述目的及其他相关目的,本发明提供一种计算机可读存储介质,其上存储有计算机程序,所述计算机程序被处理器执行时实现所述适用于科技文献的知识图谱构建方法。

[0014] 为实现上述目的及其他相关目的,本发明提供一种电子终端,包括:处理器及存储器;所述存储器用于存储计算机程序,所述处理器用于执行所述存储器存储的计算机程序,以使所述电子终端执行所述适用于科技文献的知识图谱构建方法。

[0015] 如上所述,本发明的适用于科技文献的知识图谱构建方法、系统、终端及介质,具有以下有益效果:本发明提供的技术方案与科技文献的非结构化文本特性有极高的契合度,分词单元由结合单词以及人工校验来完善,有利于科技文献中的专业符合词的抽取,并利用百科自动抽取种子知识库,从而节省了大量的初期手工维护种子知识库的成本。

附图说明

[0016] 图1a显示为本发明一实施例中适用于科技文献的知识图谱构建方法的流程示意图。

[0017] 图1b显示为本发明一实施例中生成带有科技文献属性的自定义词库的流程示意图。

[0018] 图2显示为本发明一实施例中生成词向量表示模型的流程示意图。

[0019] 图3显示为本发明一实施例中生成种子知识库的流程示意图。

[0020] 图4显示为本发明一实施例中生成科技知识库的流程示意图。

[0021] 图5显示为本发明一实施例中适用于科技文献的知识图谱构建系统的结构示意图。

[0022] 图6显示为本发明一实施例中电子终端的结构示意图。

具体实施方式

[0023] 以下通过特定的具体实例说明本发明的实施方式，本领域技术人员可由本说明书所揭露的内容轻易地了解本发明的其他优点与功效。本发明还可以通过另外不同的具体实施方式加以实施或应用，本说明书中的各项细节也可以基于不同观点与应用，在没有背离本发明的精神下进行各种修饰或改变。需说明的是，在不冲突的情况下，以下实施例及实施例中的特征可以相互组合。

[0024] 需要说明的是，在下述描述中，参考附图，附图描述了本申请的若干实施例。应当理解，还可使用其他实施例，并且可以在不背离本申请的精神和范围的情况下进行机械组成、结构、电气以及操作上的改变。下面的详细描述不应该被认为是限制性的，并且本申请的实施例的范围仅由公布的专利的权利要求书所限定。这里使用的术语仅是为了描述特定实施例，而并非旨在限制本申请。空间相关的术语，例如“上”、“下”、“左”、“右”、“下面”、“下方”、““下部”、“上方”、“上部”等，可在文中使用以便于说明图中所示的一个元件或特征与另一元件或特征的关系。

[0025] 再者，如同在本文中所使用的，单数形式“一”、““一个”和“该”旨在也包括复数形式，除非上下文中有相反的指示。应当进一步理解，术语“包含”、“包括”表明存在所述的特征、操作、元件、组件、项目、种类、和/或组，但不排除一个或多个其他特征、操作、元件、组件、项目、种类、和/或组的存在、出现或添加。此处使用的术语“或”和“和/或”被解释为包括性的，或意味着任一个或任何组合。因此，“A、B或C”或者“A、B和/或C”意味着“以下任一个：A；B；C；A和B；A和C；B和C；A、B和C”。仅当元件、功能或操作的组合在某些方式下内在地互相排斥时，才会出现该定义的例外。

[0026] 本发明提供适用于科技文献的知识图谱构建方法、系统、终端及介质，用于根据特定来源数据构建相应的适用于科技文献的知识图谱。本发明提供的技术方案与科技文献的非结构化文本特性有极高的契合度，分词单元由结合单词以及人工校验来完善，有利于科技文献中的专业符合词的抽取，并利用百科自动抽取种子知识库，从而节省了大量的初期手工维护种子知识库的成本。于下文中，将结合具体的实施例说明本发明技术方案的实施方式以及工作原理。

[0027] 如图1a所示，展示本发明一实施例中适用于科技文献的知识图谱构建方法的流程示意图。所述方法可应用于智能终端或者控制器；本发明所指的智能终端例如可采用台式电脑、本地服务器或者云端服务器等固定智能终端，也可采用手机、pad电脑、笔记本电脑、智能手环等移动智能终端；本发明所指的控制器例如可采用MCU控制器、FPGA控制器、DSP控制器、SoC控制器或者ARM控制器等等。所述适用于科技文献的知识图谱构建方法具体包括如下所述的步骤S1～S5：

[0028] S1：对科技文献进行初步分词处理以获得第一单词集，并对第一单词集中的单词进行结合处理以形成结合后的新词，且优化新词并生成带有科技文献属性的自定义词库。

[0029] 如图1b所示，展示本发明一实施例中生成带有科技文献属性的自定义词库的流程

示意图。于本实施例中，步骤S1分别由步骤S101～S105这五个子步骤实现，其包括：

[0030] S101：利用分词工具对科技文献进行初步分词处理，以获得所述第一单词集。因语言结构不同，针对中文和非英文通常采用不同的分词工具。例如：利用jieba分词工具对中文语言的科技文献进行分词处理，利用波特词干算法对英文语言的科技文献进行分词处理。

[0031] 用于对中文语言的科技文献进行分词处理的分词工具包括但不限于jieba分词工具，还可采用例如：NLPIR分词工具、Ansj分词工具、LTP分词工具、FNLP分词工或者THULAC分词工具等等，本发明对此不作限定。

[0032] 步骤S102：计算所述第一单词集中单词之间的互信息量，并将满足结合条件的互信息量所对应的单词进行结合处理。具体的，计算两个单词之间的互信息量，判断该互信息量是否大于预设阈值，若大于则表示该互信息量满足结合条件，故可将对应的两个单词相互结合。

[0033] 在一实施例中，两个单词之间的互信息量被表示为：

$Score(w_i, w_j) = \frac{freq(w_i, w_j)-\delta}{freq(w_i)freq(w_j)}$；其中，$freq(w_i)$、$freq(w_j)$以及$freq(w_i,w_j)$分别表示单词$w_i$的出现频度、单词$w_j$的出现频度以及单词$w_i$和单词$w_j$前后同时出现的频度；δ是防止特别低频的单词被结合到一起的折现系数，本实施例中设为5。

[0034] 于本实施例中，用于判断计算得到的两个单词能否相互结合的结合条件被表示为：总token数/500。也即，当$Score(w_i,w_j)$大于总token数/500时，便可将单词w_i和单词w_j结合在一起，反之则不将单词w_i和单词w_j结合在一起。其中，token数是指本用以分析能否相互结合的单词的总数量。需要说明的是，本实施例中设置的阈值，即总token数/500，是由经验所得，即从过往的测试数据中挑选出的最优值作为用以判断单词能否结合的阈值。

[0035] 步骤S103：重复步骤S101、步骤S102。由于中文文本的词与词之间不像英文文本那样有空格分隔，因此，针对中文语言的科技文献重复一次步骤S102，而针对英文语言的科技文献则重复两次步骤S101即可。

[0036] 步骤S104：利用正则表达式修正结合后的新词并将修正后的新词导入所述分词工具的自定义词库中。

[0037] 具体的，选取并清洗由步骤S101～步骤S103所获的结合后的新词，将清洗后的新词导入分词工具的自定义词库。以jieba分词工具为例，jieba分词工具按照词性可分为多个类别，例如：词性为c的名词、词性为n的名词、词性为nt的机构团体及词性为q的量词等等，本实施例将jieba分词工具中词性为nz的其它专有名词作为自定义词库。

[0038] 在一实施例中，清洗由步骤S101～步骤S103获得的结合词的方式为利用正则表达式选取出以预设词开头或结尾的单词并对这些词进行手动加工修正。具体的，可利用正则表达式选取出带有特殊字符的单词以进行修正，所述特殊字符包括但不限于：以“该”、“及”、“使”、“其”、“为”、“或”、“的”、“和”、“在”、“将”、“与”、“用”等词开头或结尾的单词。

[0039] 步骤105：基于所述自定义词库对科技文献再次进行分词处理，并根据词性对此次分词后的单词进行结合处理且将结合后的词导入所述自定义词库中，以生成所述带有科技文献属性的自定义词库。

[0040] 具体的，基于已增添用户自定义词库的自然语言处理工具，重新对科技文献进行

分词，并对前后均为名词、名词短语、动名词短语的单词进行结合。将前后单词均属于词性为n开头的名词或名词短语，或词性为vn的动名词短语结合形成新的名词短语，其中，所述n开头的名词或名词短语例如为n-名词、nt-机构团体或nz-其他专有名词。可选的，名词、名词短语及动名词短语前为m(数词)、q(量词)或m(数词)与q(量词)组合的则一并将其进行结合，并导至jieba分词工具的用户自定义词库中，从而生成带有科技文献属性的自定义词库。

[0041] S2：基于所述自定义词库对所述科技文献进行再分词以及词性标注处理以获得第二单词集，并利用词向量生成工具生成基于所述第二单词集的词向量表示模型。

[0042] 如图2所示，展示本发明一实施例中生成词向量表示模型的流程示意图。于本实施例中，步骤S2分别由步骤S201～S202这两个子步骤实现，其包括：

[0043] S201：利用分句工具对科技文献按照句子进行切分。

[0044] S202：利用分词工具对切分后的句子进行分词处理并按照自定义词库进行词性标注处理，将分词及词性标注处理的处理结果作为科技文献语料库输入至词向量生成工具中，以生成科技文献的词向量表示模型。

[0045] 在一实施例中，可采用Punkt分句工具对科技文献进行句子切分；对于中文科技文献而言，可利用jieba分词工具基于所述用户自定义词库对分割后的各句子进行分词和词性标注处理，并将处理结果作为科技文献语料库输入到word2vec工具中，从而生成科技文献的词向量表示模型。对于英文科技文献而言，则可采用波特词干算法提取词干后再利用word2phrase工具结合单词，基于所述自定义词库并使用Standford Parser工具进行词性标注处理，将处理结果作为科技文献语料库输入到word2vec工具中，从而生成科技文献的词向量表示模型。

[0046] S3：基于所述词向量表示模型从百科条目中筛选出与科技词汇表中词汇的语义相似度满足相似度要求的百科条目，将筛选出的百科条目的条目名称与分类结构及百科中预设的Infobox表相结合，并从结合结果中选取出<百科条目名称，isA，分类名称>三元组群和<百科条目名称，属性名称，属性值>三元组群，以生成与科技词汇相关的种子知识库。

[0047] 如图3所示，展示本发明一实施例中生成种子知识库的流程示意图。于本实施例中，步骤S3分别由步骤S301～S305这五个子步骤实现，其包括：

[0048] S301：从经分词处理后的科技文献中选取名词、名词短语及动名词短语以建立科技词汇表。具体的，选取步骤S2中分词后的科技文献中词性为n的名词(例如n-名词、nt-机构团体、nz-其他专有名词)或词性为vn的动名词短语(英文则是N开头的NN-名词、NP-名词短语、NR-固有名词)，并选出DF值大于5且小于文档总数20％的科技词汇以生成科技词汇表，其中，所述DF值为出现特定科技词汇的文档数量。

[0049] 步骤S302：获取百科中预设的的条目dump文件以及分类链接dump文件并导入数据库中。以维基百科为例，下载维基百科的条目dump文件(如zhwiki-20180801-pages-articles.xml.bz2)和分类链接dump文件(如zhwiki-20180801-categorylinks.sql.gz)，并导入MySQL中。

[0050] 步骤S303：从百科条目中筛选出与科技词汇表中词汇的语义相似度大于0.6且百科中的预设字段page_namespace为0的条目，并通过与百科中预设的revision表、text表连接的方式从筛选出的条目中选取百科的内部编号、条目名称及文本字段，以生成百科的

page_refined表。

[0051] 于本实施例中，对维基百科的page表进行优化，选取与科技词汇表中的单词有语义相似度大于0.6且page_namespace为0(即维基百科的实体)的条目，与revision表、text表连接抽出维基百科的内部编号(page.page_id)、条目名称(page.page_title)、文本(text.old_text)两个字段，并另存为page_refined表。

[0052] 步骤S304：利用百科中预设的categorylinks表和所述page_refined表选出<百科条目名称，分类名称>二元组群并转化为所述<百科条目名称，isA，分类名称>三元组群。

[0053] 于本实施例中，用categorylinks表和page_refined表抽出<维基百科条目名称，分类名称>二元组转化为<维基百科条目名称，isA，分类名称>三元组，筛选出分类名称与科技词汇表中的单词语义相似度大于0.5的三元组，作为种子知识库的数据来源之一。选取SQL问具体如下：

[0054] SELECT B.page_title,A.cl_to FROM categorylinks AS A LEFT JOIN page AS B ON

[0055] A.cl_from＝B.page_id WHERE B.page_namespace＝0AND B.page_title IS NOT NULL AND

[0056] A.cl_to<>B.page_title AND B.page_title。

[0057] 其中，B.page_title即为维基百科条目名称，A.cl_to则是分类名称。

[0058] 步骤S305：利用所述page_refined表中用于表示正文内容的字段找到与Infobox表相关联的数据资源，选取其中的<属性名称，属性值>二元组群并转化为所述<百科条目名称，属性名称，属性值>三元组群。

[0059] 于本实施例中，从维基百科正文内容中即从pagee_refined表中的text_old字段中找到与Infobox相关的数据资源，并选取其中的<属性名称，属性值>对转化为<维基百科条目名称，属性名称，属性值>三元组加入到种子知识库。需要说明的是，具体如何找到Infobox以及如何选取出与<属性名称，属性值>均为现有，故不再赘述。

[0060] 另外需要说明的是，本实施例中所涉的文件、表、字段，例如条目dump文件、分类链接dump文件、预设字段page_namespace、revision表、text表、page refined表等等，均是维基百科中的文件、表、字段。

[0061] S4：从科技文献中选取核心词汇并选取出所述核心词汇附近的科技词汇，从选取出的科技词汇中筛选出与所述种子知识库中百科条目的语义相似度符合相似度要求的科技词汇，以形成<核心词汇，isA，科技词汇>三元组群和<核心词汇，属性名称，科技词汇>三元组群并导入所述种子知识库；按照百科条目名称与核心词汇，以及isA与属性名称进行聚合，归并符合相似度要求的三元组群，以生成科技知识库。

[0062] 如图4所示，展示本发明一实施例中生成科技知识库的流程示意图。于本实施例中，步骤S4分别由步骤S401～S405这五个子步骤实现，其包括：

[0063] 步骤S401：选取科技文献中的科技词汇并计算科技词汇表中所有单词的tf-idf值；其中，所述tf-idf值的计算公式为：idf(w)＝log(1+|D|)/(1+df(w))；|D|为文献总数，df(w)为包含有单词w的文献数量。

[0064] 于本实施例中，将每篇科技文献视作一个文档，选取科技文献全文数据中的科技词汇，计算科技词汇表中所有单词的tf-idf值，计算公式如下：

$tf-idf(w)=log\frac{1+|D|}{1+df(w)}$；其中，|D|为文献总数，df(w)为包含有单词w的文档数量。

[0065] 步骤S402：计算科技文献每个段落中的科技词汇的tf-idf值并按倒序排序，并选取每个段落的核心词汇。

[0066] 于本实施例中，选取科技文献每个段落中的科技词汇，计算其tf-idf值并按照倒序进行排序，并选取每个段落的核心词汇。具体地，本发明的实施例中获取段落中的句子数量L，将倒序排序中前L位作为该段落的核心词汇。

[0067] 步骤S403：计算所述核心词汇与所述种子知识库中百科条目的语义相似度，选取语义相似度大于0.5的核心词汇所在的句子。

[0068] 于本实施例中，计算核心词汇与种子知识库中维基百科条目的条目名称的语义相似度，抽出含有语义相似度大于预设阈值的核心词汇所在的句子。于本实施例中，抽出含有语义相似度大于0.5的核心词汇所在的句子。

[0069] 步骤S404：计算句子中其它科技词汇与所述核心词汇所对应的三元组群中分类名词或属性值的语义相似度；其中，若满足分类名称语义相似度要求，则组建<核心词汇，isA，科技词汇>三元组群；若满足属性值相似度要求，则组建<核心词汇，属性名称，科技词汇>三元组群。

[0070] 于本实施例中，计算句子中其他科技词汇与核心词汇所对应三元组群中分类名称或属性值的语义相似度，如与分类名称的语义相似度大于0.5则组建<核心词汇，isA，科技词汇>三元组，如与属性值的语义相似度匹配则组建<核心词汇，属性名称，科技词汇>三元组。

[0071] 步骤S405：将<核心词汇，isA，科技词汇>三元组群和<核心词汇，属性名称，科技词汇>三元组群加入所述种子知识库；将所述种子知识库中的三元群组统一视为<科技词汇1，词关系，科技词汇2>，并按照科技词汇1与词关系进行聚合；将聚合后的同一族群中语义相似度满足相似度要求的科技词汇2或者满足字符串相似度要求的字符串进行归并，且选择族群中字符串长度最长者作为代表词汇并记录科技词汇2的多种表达，以最终形成科技知识库。

[0072] 于本实施例中，将抽出的核心词汇三元组加入种子知识库，将种子知识库中的三元组视为<科技词汇1，词关系，科技词汇2>，并按照科技词汇1、词关系进行聚合，若聚合后同一族群中科技词汇2的语义相似度大于一语义相似度阈值或者字符串相似度大于一字符串相似度阈值，则将同一族群中的科技词汇2进行归并处理，并选择族群中字符串长度最长者作为代表词汇并记录科技词汇2的多种表达，最终形成可用于绘制、生成知识图谱的科技知识库。于本实施例中，同一族群中科技词汇2的语义相似度大于0.7或者字符串相似度大于0.85，则将同一族群中的科技词汇2进行归并处理。

[0073] S5：利用数据可视化工具对所述科技知识库进行可视化处理，以形成可视的科技知识图谱。具体的，利用数据可视化工具对所述科技知识库进行可视化处理，以形成可视的科技知识图谱。所述可视化工具例如为Netdraw软件或者基于d3.js的数据可视化软件等等。

[0074] 本领域普通技术人员可以理解：实现上述各方法实施例的全部或部分步骤可以通过计算机程序相关的硬件来完成。前述的计算机程序可以存储于一计算机可读存储介质

中。该程序在执行时，执行包括上述各方法实施例的步骤；而前述的存储介质包括：ROM、RAM、磁碟或者光盘等各种可以存储程序代码的介质。

[0075] 如图5所示，展示本发明一实施例中适用于科技文献的知识图谱构建系统的结构示意图。于本实施例中，所述系统包括词库生成模块51、词向量生成模块52、种子知识库生成模块53、科技知识库生成模块54以及知识图谱生成模块55。

[0076] 所述词库生成模块51用于对科技文献进行初步分词处理以获得第一单词集，并对第一单词集中的单词进行结合处理以形成结合后的新词，且优化新词并生成带有科技文献属性的自定义词库。所述词向量生成模块52用于基于所述自定义词库对所述科技文献进行再分词以及词性标注处理以获得第二单词集，并利用词向量生成工具生成基于所述第二单词集的词向量表示模型。所述种子知识库生成模块53用于基于所述词向量表示模型从百科条目中筛选出与科技词汇表中词汇的语义相似度满足相似度要求的百科条目，将筛选出的百科条目的条目名称与分类结构及百科中预设的Infobox表相结合，并从结合结果中选取出<百科条目名称，isA，分类名称>三元组群和<百科条目名称，属性名称，属性值>三元组群，以形成与科技词汇相关的种子知识库。所述科技知识库生成模块54用于从科技文献中选取核心词汇并选取出所述核心词汇附近的科技词汇，从选取出的科技词汇中筛选出与所述种子知识库中百科条目的语义相似度符合相似度要求的科技词汇，以形成<核心词汇，isA，科技词汇>三元组群和<核心词汇，属性名称，科技词汇>三元组群并导入所述种子知识库；按照百科条目名称与核心词汇，以及isA与属性名称进行聚合，归并符合相似度要求的三元组群，以形成科技知识库。所述知识图谱生成模块55用于利用数据可视化工具对所述科技知识库进行可视化处理，以生成可视的科技知识图谱。

[0077] 需要说明的是，应理解以上装置的各个模块的划分仅仅是一种逻辑功能的划分，实际实现时可以全部或部分集成到一个物理实体上，也可以物理上分开。且这些模块可以全部以软件通过处理元件调用的形式实现；也可以全部以硬件的形式实现；还可以部分模块通过处理元件调用软件的形式实现，部分模块通过硬件的形式实现。例如，知识图谱生成模块可以为单独设立的处理元件，也可以集成在上述装置的某一个芯片中实现，此外，也可以以程序代码的形式存储于上述装置的存储器中，由上述装置的某一个处理元件调用并执行以上知识图谱生成模块的功能。其它模块的实现与之类似。此外这些模块全部或部分可以集成在一起，也可以独立实现。这里所述的处理元件可以是一种集成电路，具有信号的处理能力。在实现过程中，上述方法的各步骤或以上各个模块可以通过处理器元件中的硬件的集成逻辑电路或者软件形式的指令完成。

[0078] 例如，以上这些模块可以是被配置成实施以上方法的一个或多个集成电路，例如：一个或多个特定集成电路(Application Specific Integrated Circuit，简称ASIC)，或，一个或多个微处理器(digital signal processor，简称DSP)，或，一个或者多个现场可编程门阵列(Field Programmable Gate Array，简称FPGA)等。再如，当以上某个模块通过处理元件调度程序代码的形式实现时，该处理元件可以是通用处理器，例如中央处理器(Central Processing Unit，简称CPU)或其它可以调用程序代码的处理器。再如，这些模块可以集成在一起，以片上系统(system-on-a-chip，简称SOC)的形式实现。

[0079] 如图6所示，展示本发明一实施例中电子终端的结构示意图。本实施例提供的电子终端包括：处理器61、存储器62、收发器63、通信接口64和系统总线65；存储器62和通信接口

64通过系统总线65与处理器61和收发器63连接并完成相互间的通信，存储器62用于存储计算机程序，通信接口64和收发器63用于和其他设备进行通信，处理器61用于运行计算机程序，使电子终端执行如上知识图谱构建方法的各个步骤。

[0080] 上述提到的系统总线可以是外设部件互连标准(Peripheral Pomponent Interconnect，简称PCI)总线或扩展工业标准结构(Extended Industry Standard Architecture，简称EISA)总线等。该系统总线可以分为地址总线、数据总线、控制总线等。为便于表示，图中仅用一条粗线表示，但并不表示仅有一根总线或一种类型的总线。通信接口用于实现数据库访问装置与其他设备(例如客户端、读写库和只读库)之间的通信。存储器可能包含随机存取存储器(Random Access Memory，简称RAM)，也可能还包括非易失性存储器(non-volatile memory)，例如至少一个磁盘存储器。

[0081] 上述的处理器可以是通用处理器，包括中央处理器(Central Processing Unit，简称CPU)、网络处理器(Network Processor，简称NP)等；还可以是数字信号处理器(Digital Signal Processing，简称DSP)、专用集成电路(Application Specific Integrated Circuit，简称ASIC)、现场可编程门阵列(Field－Programmable Gate Array，简称FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件。

[0082] 综上所述，本发明提供的适用于科技文献的知识图谱构建方法、系统、终端及介质，本发明提供的技术方案与科技文献的非结构化文本特性有极高的契合度，分词单元由结合单词以及人工校验来完善，有利于科技文献中的专业符合词的抽取，并利用百科自动抽取种子知识库，从而节省了大量的初期手工维护种子知识库的成本。所以，本发明有效克服了现有技术中的种种缺点而具高度产业利用价值。

[0083] 上述实施例仅例示性说明本发明的原理及其功效，而非用于限制本发明。任何熟悉此技术的人士皆可在不违背本发明的精神及范畴下，对上述实施例进行修饰或改变。因此，举凡所属技术领域中具有通常知识者在未脱离本发明所揭示的精神与技术思想下所完成的一切等效修饰或改变，仍应由本发明的权利要求所涵盖。

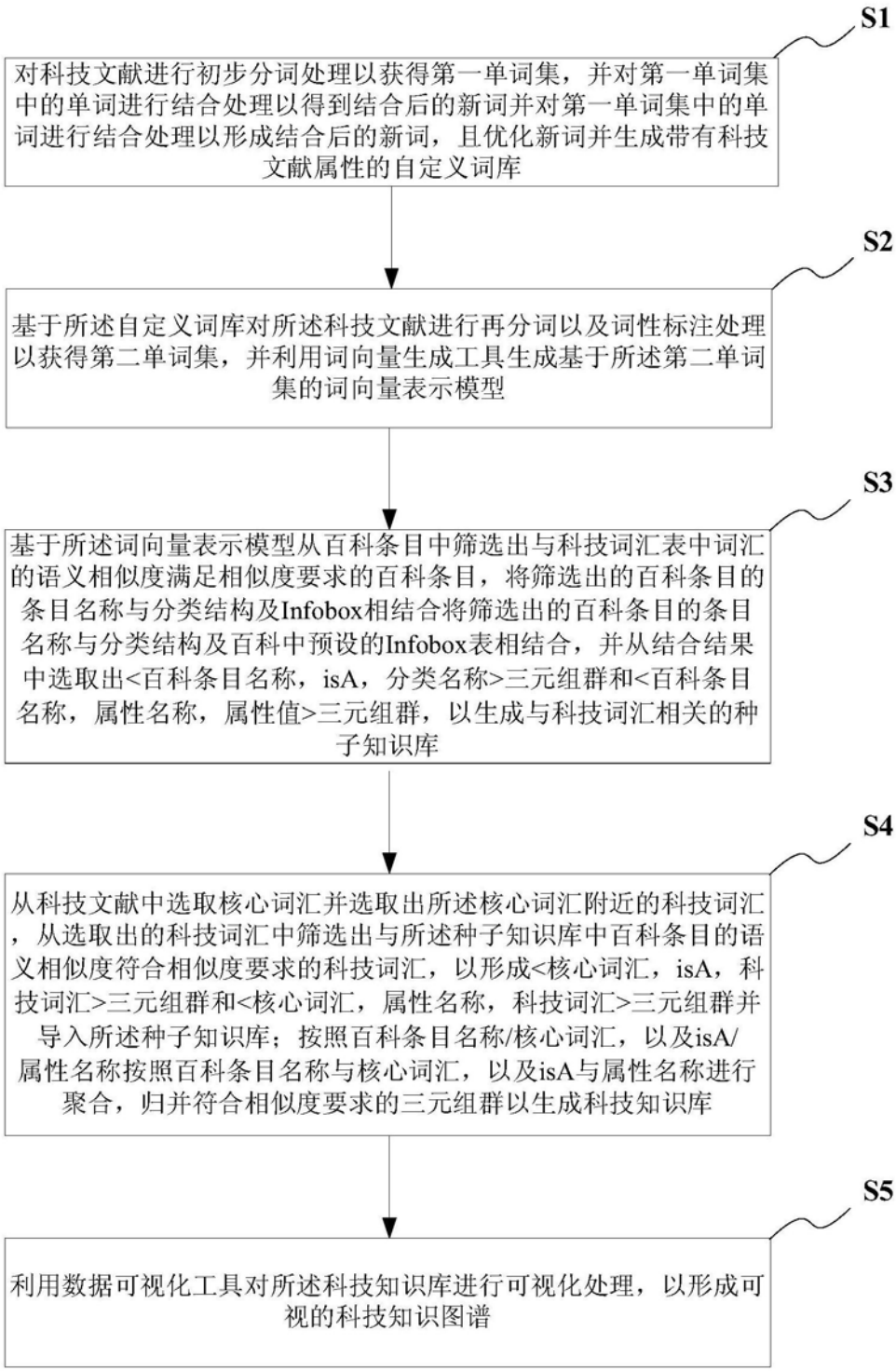

S1
对科技文献进行初步分词处理以获得第一单词集，并对第一单词集中的单词进行结合处理以得到结合后的新词并对第一单词集中的单词进行结合处理以形成结合后的新词，且优化新词并生成带有科技文献属性的自定义词库
S2
基于所述自定义词库对所述科技文献进行再分词以及词性标注处理以获得第二单词集，并利用词向量生成工具生成基于所述第二单词集的词向量表示模型
S3
基于所述词向量表示模型从百科条目中筛选出与科技词汇表中词汇的语义相似度满足相似度要求的百科条目，将筛选出的百科条目的条目名称与分类结构及Infobox相结合将筛选出的百科条目的条目名称与分类结构及百科中预设的Infobox表相结合，并从结合结果中选取出<百科条目名称，isA，分类名称>三元组群和<百科条目名称，属性名称，属性值>三元组群，以生成与科技词汇相关的种子知识库
S4
从科技文献中选取核心词汇并选取出所述核心词汇附近的科技词汇，从选取出的科技词汇中筛选出与所述种子知识库中百科条目的语义相似度符合相似度要求的科技词汇，以形成<核心词汇，isA，科技词汇>三元组群和<核心词汇，属性名称，科技词汇>三元组群并导入所述种子知识库；按照百科条目名称/核心词汇，以及isA/属性名称按照百科条目名称与核心词汇，以及isA与属性名称进行聚合，归并符合相似度要求的三元组群以生成科技知识库
S5
利用数据可视化工具对所述科技知识库进行可视化处理，以形成可视的科技知识图谱

图1a

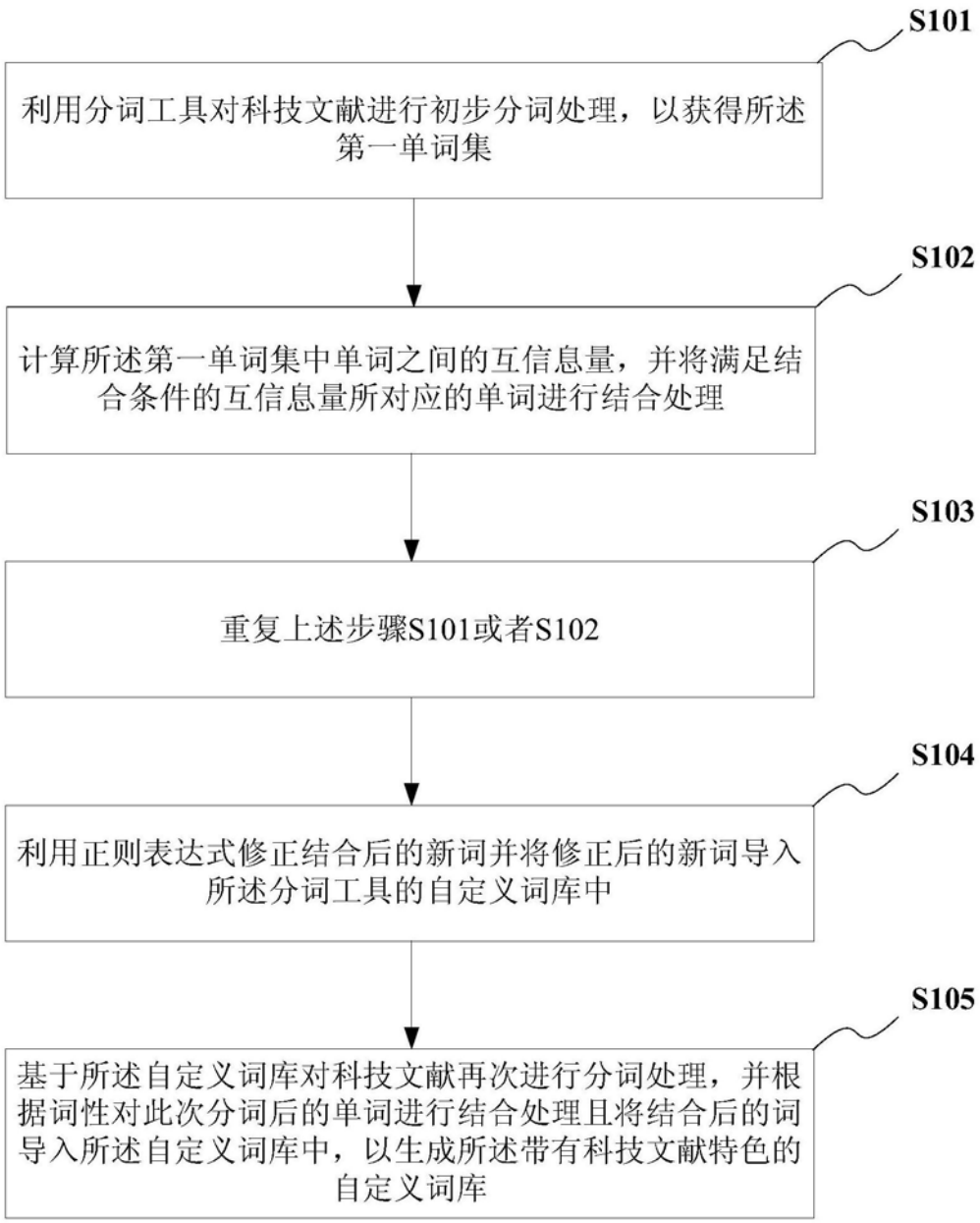

S101
利用分词工具对科技文献进行初步分词处理，以获得所述第一单词集
S102
计算所述第一单词集中单词之间的互信息量，并将满足结合条件的互信息量所对应的单词进行结合处理
S103
重复上述步骤S101或者S102
S104
利用正则表达式修正结合后的新词并将修正后的新词导入所述分词工具的自定义词库中
S105
基于所述自定义词库对科技文献再次进行分词处理，并根据词性对此次分词后的单词进行结合处理且将结合后的词导入所述自定义词库中，以生成所述带有科技文献特色的自定义词库

图1b

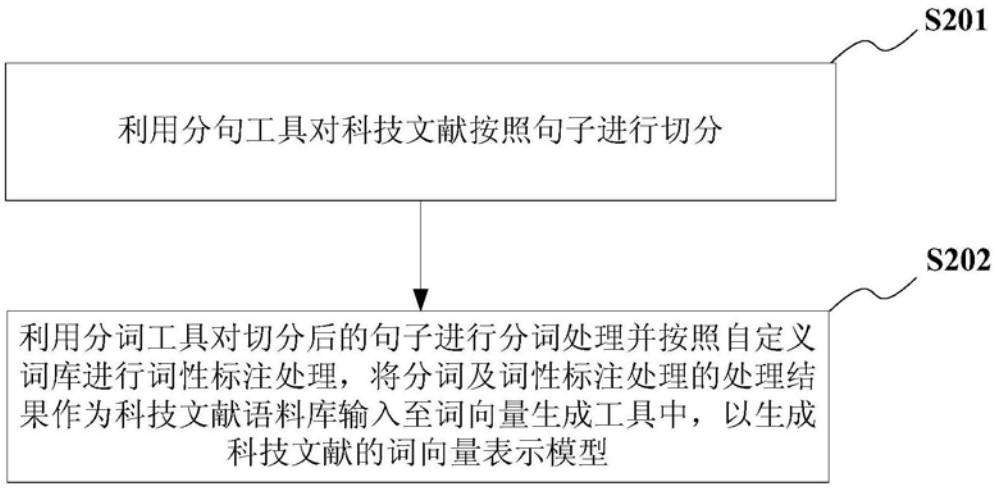

S201
利用分句工具对科技文献按照句子进行切分
S202
利用分词工具对切分后的句子进行分词处理并按照自定义词库进行词性标注处理，将分词及词性标注处理的处理结果作为科技文献语料库输入至词向量生成工具中，以生成科技文献的词向量表示模型

图2

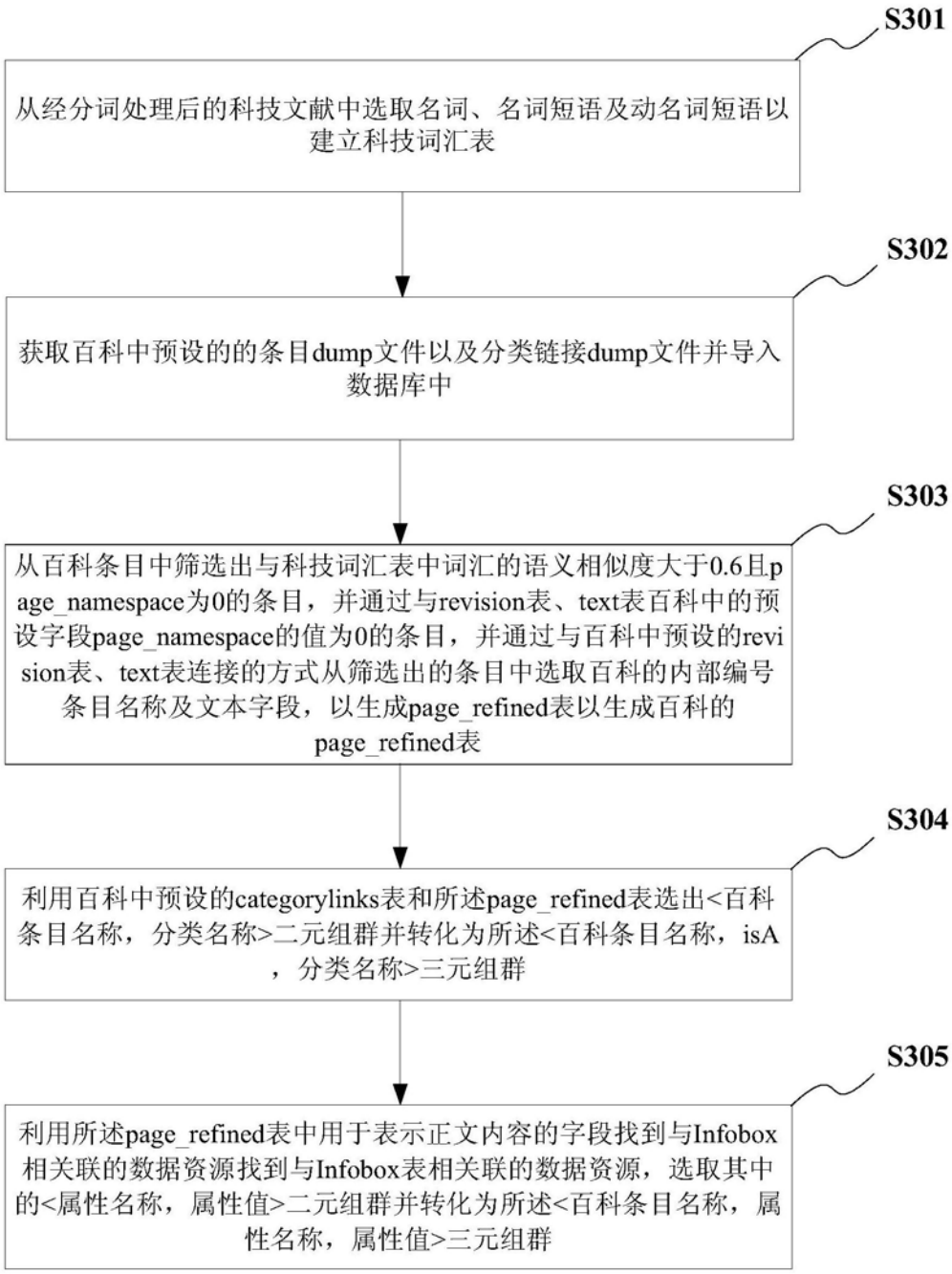

S301
从经分词处理后的科技文献中选取名词、名词短语及动名词短语以建立科技词汇表
S302
获取百科中预设的的条目dump文件以及分类链接dump文件并导入数据库中
S303
从百科条目中筛选出与科技词汇表中词汇的语义相似度大于0.6且page_namespace为0的条目，并通过与revision表、text表百科中的预设字段page_namespace的值为0的条目，并通过与百科中预设的revision表、text表连接的方式从筛选出的条目中选取百科的内部编号条目名称及文本字段，以生成page_refined表以生成百科的page_refined表
S304
利用百科中预设的categorylinks表和所述page_refined表选出<百科条目名称，分类名称>二元组群并转化为所述<百科条目名称，isA，分类名称>三元组群
S305
利用所述page_refined表中用于表示正文内容的字段找到与Infobox相关联的数据资源找到与Infobox表相关联的数据资源，选取其中的<属性名称，属性值>二元组群并转化为所述<百科条目名称，属性名称，属性值>三元组群

图3

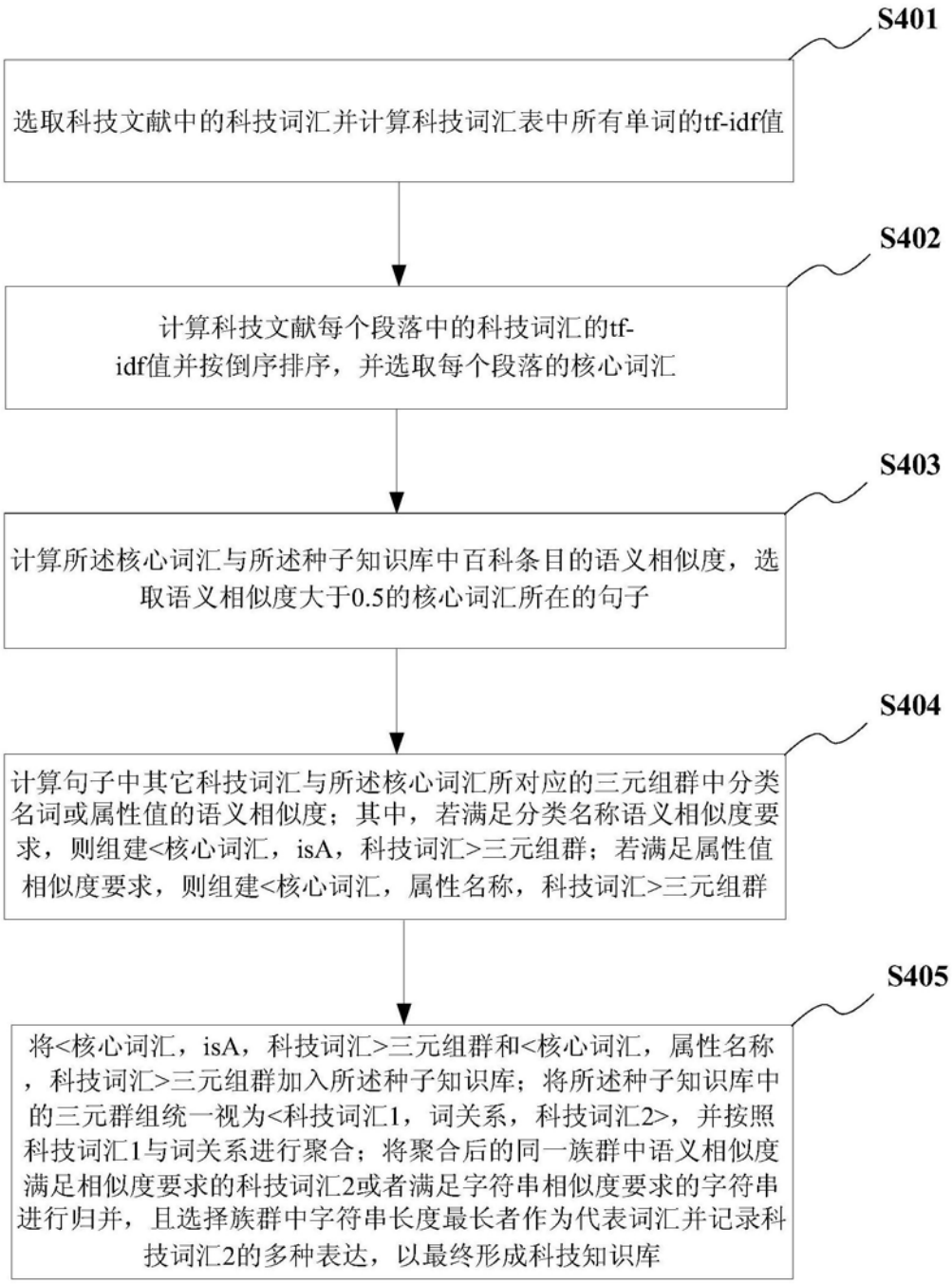

S401
选取科技文献中的科技词汇并计算科技词汇表中所有单词的tf-idf值
S402
计算科技文献每个段落中的科技词汇的tf-idf值并按倒序排序，并选取每个段落的核心词汇
S403
计算所述核心词汇与所述种子知识库中百科条目的语义相似度，选取语义相似度大于0.5的核心词汇所在的句子
S404
计算句子中其它科技词汇与所述核心词汇所对应的三元组群中分类名词或属性值的语义相似度；其中，若满足分类名称语义相似度要求，则组建<核心词汇，isA，科技词汇>三元组群；若满足属性值相似度要求，则组建<核心词汇，属性名称，科技词汇>三元组群
S405
将<核心词汇，isA，科技词汇>三元组群和<核心词汇，属性名称，科技词汇>三元组群加入所述种子知识库；将所述种子知识库中的三元群组统一视为<科技词汇1，词关系，科技词汇2>，并按照科技词汇1与词关系进行聚合；将聚合后的同一族群中语义相似度满足相似度要求的科技词汇2或者满足字符串相似度要求的字符串进行归并，且选择族群中字符串长度最长者作为代表词汇并记录科技词汇2的多种表达，以最终形成科技知识库

图4


知识图谱构建系统
词库生成模块
51
词向量生成模块
52
种子知识库生成模块
53
科技知识库生成模块
54
知识图谱生成模块
55

图5

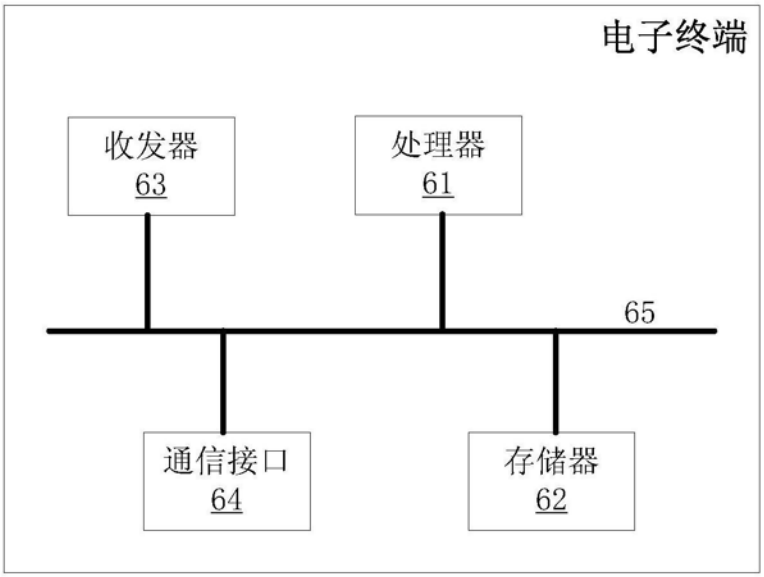

电子终端
收发器
63
处理器
61
65
通信接口
64
存储器
62

图6