

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2022 年 5 月 27 日 (27.05.2022)



(10) 国际公布号
WO 2022/105178 A1

- (51) 国际专利分类号:
G06F 16/33 (2019.01)
- (21) 国际申请号: PCT/CN2021/097021
- (22) 国际申请日: 2021 年 5 月 29 日 (29.05.2021)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
202011321892.7 2020 年 11 月 23 日 (23.11.2020) CN
- (71) 申请人: 平安科技(深圳)有限公司(PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN]; 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。
- (72) 发明人: 李弦(LI, Xian); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。 阮晓雯

(RUAN, Xiaowen); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。 徐亮(XU, Liang); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。 洪博然(HONG, Boran); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(74) 代理人: 广州三环专利商标代理有限公司(SCIHEAD IP LAW FIRM); 中国广东省广州市越秀区先烈中路 80 号汇华商贸大厦 1508 室, Guangdong 510070 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK,

(54) Title: KEYWORD EXTRACTION METHOD AND RELATED DEVICE

(54) 发明名称: 一种关键词提取的方法及相关装置

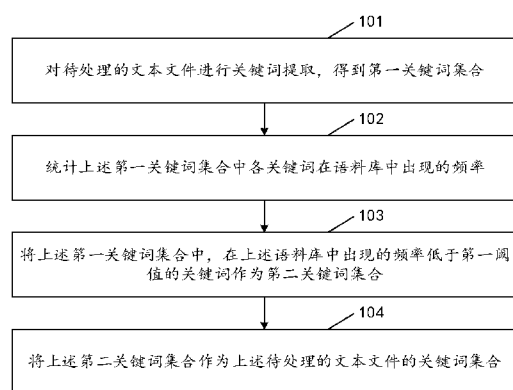


图 1

- 101 Perform keyword extraction on a text file to be processed, to obtain a first keyword set
- 102 Calculate the frequency with which each keyword in the first keyword set occurs in a corpus
- 103 Take keywords in the first keyword set that have an occurrence frequency in the corpus that is lower than a first threshold as a second keyword set
- 104 Take the second keyword set as a keyword set of the text file to be processed

(57) Abstract: A keyword extraction method and a related device. The method comprises: performing keyword extraction on a text file to be processed, to obtain a first keyword set (101); calculating the frequency with which each keyword in the first keyword set occurs in a corpus (102); taking keywords in the first keyword set that have an occurrence frequency in the corpus that is lower than a first threshold as a second keyword set (103), a text file contained in the corpus and the text file to be processed belonging to the same file type; and taking the second keyword set as a keyword set of the text file to be processed (104). According to the method and the device, the accuracy of keyword extraction is improved by improving the keyword selection method during keyword extraction.

(57) 摘要: 一种关键词提取的方法及相关装置, 该方法包括: 对待处理的文本文件进行关键词提取, 得到第一关键词集合(101); 统计上述第一关键词集合中各关键词在语料库中出现的频率(102); 将上述第一关键词集合中, 在上述语料库中出现的频率低于第一阈值的关键词作为第二关键词集合(103), 上述语料库内包含的文本文件与上述待处理的文本文件属于相同文件类型; 将上述第二关键词集合作为上述待处理的文本文件的关键词集合(104)。该方法和装置通过改进关键词提取时关键词的选择方法, 提高关键词提取的精度。

LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX,
MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL,
PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,
ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US,
UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国 (除另有指明, 要求每一种可提供的地区
保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM,
AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG,
CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU,
IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT,
RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI,
CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告 (条约第21条(3))。

一种关键词提取的方法及相关装置

本申请要求于 2020 年 11 月 23 日提交中国专利局、申请号为 202011321892.7, 发明名称为“一种关键词提取的方法及相关装置”的中国专利申请的优先权, 其全部内容通过引用结合在本申请中。

5 技术领域

本申请实施例涉及自然语言处理领域, 具体涉及一种关键词提取的方法及相关装置。

背景技术

10 关键词是指单个媒体在制作使用索引时用到的词汇, 对文本文件进行关键词提取一直都是业界的研究热点。词频-逆文档频率 (Term Frequency-Inverse Document Frequency, TF-IDF) 方法基于词频对文本文件的关键词进行提取: 首先对需要提取关键词的文本文件进行分词, 再统计上述每个分词的词频和逆文档频率, 最后将词频乘以逆文档频率的结果作为分词的权重值, 将分词按照权重值由大到小进行排序, 排名靠前的分词就可以作为文本文件的关键词。

15 发明人意识到, 上述 TF-IDF 方法中, 词的重要性与该词在文本文件中出现的次数成正比, 与该词在语料库的文章中出现的次数成反比, 在需要提取关键词的文章所属领域的语料缺乏的情况下, 上述 TF-IDF 方法提取出来的关键词可能不具有代表性。

发明内容

本申请实施例公开了一种关键词提取的方法及相关装置, 通过改进关键词提取时关键词的选择方法, 提高关键词提取的精度。

20 第一方面, 本申请实例公开了一种关键词提取的方法, 包括:

对待处理的文本文件进行关键词提取, 得到第一关键词集合;

统计上述第一关键词集合中各关键词在语料库中出现的频率, 将上述第一关键词集合中, 在上述语料库中出现的频率低于第一阈值的关键词作为第二关键词集合, 上述语料库内包含的文本文件与上述待处理的文本文件属于相同文件类型;

25 将上述第二关键词集合作为上述待处理的文本文件的关键词集合。

第二方面, 本申请实施例公开了一种关键词提取的装置, 包括:

提取单元, 用于对待处理的文本文件进行关键词提取, 得到第一关键词集合;

统计单元, 用于统计上述第一关键词集合中各关键词在语料库中出现的频率, 上述语料库内包含的文本文件与上述待处理的文本文件属于相同文件类型;

30 确定单元, 用于将上述第一关键词集合中, 在上述语料库中出现的频率低于第一阈值的关键词作为第二关键词集合; 将上述第二关键词集合作为上述待处理的文本文件的关键词集合。

35 第三方面, 本申请实施例公开了一种服务器, 包括: 处理器和存储器, 其中, 上述存储器中存储有计算机程序, 上述处理器调用上述存储器中存储的计算机程序, 用于执行以下方法:

对待处理的文本文件进行关键词提取, 得到第一关键词集合;

统计上述第一关键词集合中各关键词在语料库中出现的频率, 将上述第一关键词集合中, 在上述语料库中出现的频率低于第一阈值的关键词作为第二关键词集合, 上述语料库内包含的文本文件与上述待处理的文本文件属于相同文件类型;

40 将上述第二关键词集合作为上述待处理的文本文件的关键词集合。

第四方面, 本申请实施例公开了一种计算机可读存储介质, 上述计算机可读存储介质中存储有计算机程序, 当上述计算机程序在一个或多个处理器上运行时, 执行以下方法:

对待处理的文本文件进行关键词提取, 得到第一关键词集合;

统计上述第一关键词集合中各关键词在语料库中出现的频率, 将上述第一关键词集合

中,在上述语料库中出现的频率低于第一阈值的关键词作为第二关键词集合,,上述语料库内包含的文本文件与上述待处理的文本文件属于相同文件类型;

将上述第二关键词集合作为上述待处理的文本文件的关键词集合。

本申请实施例首先改进关键词的选择方法,再通过考虑关键词在待处理的文本文件中的分布,优化关键词集合中各关键词的排序,最后通过关键词与相邻的分词按照在待处理的文本文件中先后顺序进行组合,提取出因为分词而被拆分的组合词,从而提高关键词提取的精度。

附图说明

为了更清楚地说明本申请实施例或背景技术中的技术方案,下面将对本申请实施例或背景技术中所需要使用的附图作简单的介绍。

图1是本申请实施例公开的一种关键词提取方法的流程示意图;

图2是本申请实施例公开的另一种关键词提取方法的流程示意图;

图3是本申请实施例公开的又一种关键词提取方法的流程示意图;

图4是本申请实施例公开的一种关键词提取的装置的结构示意图;

图5是本申请实施例公开的一种服务器的结构示意图。

具体实施方式

为了使本申请的目的、技术方案和优点更加清楚,下面将结合附图对本申请作进一步地描述。

本申请的说明书、权利要求书及附图中的术语“第一”和“第二”等仅用于区别不同对象,而不是用于描述特定顺序。此外,术语“包括”和“具有”以及它们的任何变形,意图在于覆盖不排他的包含。例如包含了一系列步骤或单元的过程、方法、系统、产品或设备等,没有限定于已列出的步骤或单元,而是可选地还包括没有列出的步骤或单元等,或可选地还包括对于这些过程、方法、产品或设备等固有的其它步骤或单元。

在本文中提及的“实施例”意味着,结合实施例描述的特定特征、结构或特性可以包含在本申请的至少一个实施例中。在说明书中的各个位置出现上述短语并不一定均是指相同的实施例,也不是与其它实施例互斥的独立的或备选的实施例。本领域技术人员可以显式地和隐式地理解的是,本文所描述的实施例可以与其它实施例相结合。

在本申请中,“至少一个(项)”是指一个或者多个,“多个”是指两个或两个以上,“至少两个(项)”是指两个或三个及三个以上,“和/或”,用于描述关联对象的关联关系,表示可以存在三种关系,例如,“A和/或B”可以表示:只存在A,只存在B以及同时存在A和B三种情况,其中A,B可以是单数或者复数。字符“/”一般表示前后关联对象是一种“或”的关系。“以下至少一项(个)”或其类似表达,是指这些项中的任意组合。例如,a,b或c中的至少一项(个),可以表示:a,b,c,“a和b”,“a和c”,“b和c”,或“a和b和c”。

本申请的技术方案涉及人工智能和/或大数据技术领域,如可具体涉及自然语言处理技术。本申请可应用于文本处理等场景,以实现关键词提取,使得提高关键词提取的精度,从而推动智慧城市的建设。可选的,本申请涉及的数据如文本文件和/或各关键词等可存储于数据库中,或者可以存储于区块链中,比如通过区块链分布式存储,本申请不做限定。

本申请实施例适用于与待处理的文本文件所属的领域类型相同的文本文件不多,即语料缺乏的情况下对待处理的文本文件进行关键词提取;本申请是对传统关键词提取方法TF-IDF的优化,通过改进关键词提取时关键词的选择方法,提高关键词提取的精度。为了更清楚地描述本申请的方案,下面先介绍一些与TF-IDF方法相关的知识。

语料库:指经过取样、加工的大规模电子文本库,即存放文本文件的数据库。

词频TF:指给定的词语在当前文本文件中出现的频率。由于同一个词语在长文件中的

词频可能比在短文件中有更高的词频，因此需要根据文本文件的长度对给定的词语的词频进行归一化，那么上述词频就为给定的词语在当前文本文件中出现的次数除以当前文本文件的总词数，那么词频的公式可以表示为：

$$\text{词频 TF} = \frac{\text{给定的词语在当前文本文件中出现的次数}}{\text{当前文本文件的总词数}}$$

逆文档频率 IDF：是对给定的词语的普遍重要性的度量。即如果给定的词语只在语料库中越少的文本文件中出现，那么上述给定的词语更能够代表文本文件的主旨，其权重值应当越大；如果给定的词语在语料库的大量的文本文件中都出现，那么上述给定的词语就无法代表文本文件的主旨，也就是上述给定的词语无法清楚地表示出其代表的内容，其权重值应当小，那么逆文档频率的公式可以表示为：

$$\text{逆文档频率 IDF} = \log \frac{\text{语料库包含的文本文件总数}}{\text{包含给定的词语的文本文件总数} + 1}$$

在关键词提取方法 TF-IDF 中，一个词语的重要性与其在当前文本文件中出现的次数成正比，与其在语料库的文本文件中出现的频率成反比。那么，如果某个给定的词语在待处理的文本文件中出现频率越高，并且在语料库的文本文件中出现频率越低，该词语就能够较好地代表当前文本文件的含义，成为上述文本文件的关键词。

使用 TF-IDF 方法对待处理的文本文件进行关键词提取时，通常按照以下步骤：

1、对待处理的文本文件进行预处理。

对待处理的文本文件的预处理操作包括分词、词性标注、去除停用词。在分词部分，可以采用的分词工具有很多，包括结巴分词、盘古分词等，本部分可以采用最常用的结巴分词对待处理的文本文件进行分词，结巴分词基于前缀词典实现高效的词图扫描，生成句子中汉字所有可能成词情况构成的有向无环图，再动态规划查找最大概率路径，找出基于词频的最大切分组合，由于上述结巴分词是一种非常典型的分词工具，具体原理这里不再赘述。

词性标注指为每个分词加上合适的词性标签，以便于对句子的分析和从分词集合中去除停用词，比如保留分词集合中词性为名词、专有名词、动词的分词，由于上述词性标注和去除停用词是非常典型的处理步骤，具体原理这里不再赘述。

这样，可以得到包含 n 个分词的候选关键词集合，记为：

2、计算候选关键词集合中每个分词在待处理的文本文件中的词频 TF。

3、计算候选关键词集合中每个分词在整个语料库中的逆文档频率 IDF。

4、计算 TF 乘以 IDF 得到候选关键词集合中每个关键词的 TF-IDF 值。

5、将候选关键词集合中每个分词按照 TF-IDF 值由大到小进行排序，排名靠前的关键词就可以作为上述待处理的文本文件的关键词。

由于 TF-IDF 方法依赖于语料库，在与待处理的文本文件属于同一领域的文本文件数量很少的情况下，整个语料库中包含大量与待处理的文本文件领域不相关的文本文件，通过 TF-IDF 方法提取出来的关键词在待处理的文本文件所属领域很有可能不具有代表性。本申请提供了一种新的关键词提取的方法，首先改进关键词的选择方法，过滤掉在待处理的文本文件所属领域中不具有代表性的关键词；再通过综合考虑关键词在待处理的文本文件中的分布和关键词的逆文档频率，优化关键词集合中各关键词的排名的名次，让更加具有代表性的关键词排名靠前；最后通过关键词与相邻的分词按照在待处理的文本文件中先后顺序进行组合，提取出因为分词而被拆分的组合词，从而提高关键词提取的精度。

接下来结合本申请实施例中的附图对本申请实施例进行描述。

请参阅图 1，图 1 是本申请实施例公开的一种关键词提取方法的流程示意图，如图所示，上述方法包括：

S101：对待处理的文本文件进行关键词提取，得到第一关键词集合。

在对关键词的选择方法进行优化之前，首先需要对待处理的文本文件进行关键词的初步提取，得到一组关键词集合，可以将上述关键词集合记为第一关键词集合。

在上述步骤中，具体的关键词提取工具可以为基于大规模语料和 TF-IDF 算法的关键词提取开源程序，例如，利用 jieba.analyse.extract_tags 算法包对待处理的文本文件进行关键词提取。

如果语料库内的文本文件数量、类型较多，并且语料库中同时包括与待处理的文本文件类型相同和不同的文本文件，那么上述语料库可以称为大规模语料库，大规模语料库内的文本文件可以称为大规模语料。其中，提取上述第一关键词集合对应的语料库为大规模语料库。

为了不遗漏关键词、方便后续对关键词集合中各个关键词的选择，可以将关键词的数量设置为大于或等于 20。

S102：统计上述第一关键词集合中各关键词在语料库中出现的频率。

上述步骤中，统计上述第一关键词集合中各关键词在上述语料库中出现的频率，其中，语料库内包含的文本文件与上述待处理的文本文件属于相同文件类型。

在步骤 S101 中，得到上述第一关键词集合所采用的 TF-IDF 方法基于大规模语料库，上述第一关键词集合中的关键词可能在大规模语料中具有代表性，但是在待处理的文本文件所属的领域内却不具有代表性，通过将上述待处理的文本文件所属领域文件类型相同的文本文件作为语料库的语料，可以刻画关键词在上述待处理的文本文件所属领域中的重要程度。

关键词在语料库中出现的频率可以通过不同的维度来刻画，比如基于文本文件，关键词在语料库中出现的频率可以表示为语料库包含关键词的文本文件的总数除以语料库中包含的文本文件的总数。比如基于词语，关键词在语料库中出现的频率可以表示为语料库中关键词出现的总次数除以语料库中词语的总数。关键词在语料库中出现的频率越大，关键词就越无法代表文本文件的主旨。

S103：将上述第一关键词集合中，在上述语料库中出现的频率低于第一阈值的关键词作为第二关键词集合。

在语料库内包括的文本文件与待处理的文本文件属于相同文件类型时，统计的上述第一关键词集合中各关键词在语料库中的频率就可以很好的刻画关键词在上述待处理的文本文件所属领域的重要程度，在上述语料库中出现的频率越小，上述关键词在上述待处理的文本文件所属领域越具有代表性。

为上述关键词在语料库中出现的频率设置阈值，即第一阈值，频率低于上述第一阈值的关键词就可以作为在上述待处理的文本文件所属领域具有代表性的关键词；最终得到的频率为 0 到 1 之间的数，可以根据实验需要将上述第一频率设置为大于 0 小于 1 之间的任意数，只要保证第一关键词集合中存在关键词的频率低于上述第一阈值即可。

S104：将上述第二关键词集合作为上述待处理的文本文件的关键词集合。

第二关键词集合中的关键词是从第一关键词集合中筛选和过滤而得到的，通过每个关键词在包含的文本文件与上述待处理的文本文件属于相同文件类型的语料库中出现的频率，得到在待处理的文本文件所属领域具有代表性的关键词，将上述第二关键词集合作为上述待处理的文本文件的关键词集合，关键词的代表性更强。

本申请的方案还包括，统计第一关键词集合中每个关键词在语料库的逆文档频率，语料库内包含的文本文件与上述待处理的文本文件属于相同文件类型。为了与传统 TF-IDF 方

法中获取的逆文档频率进行区分，可以将上述逆文档频率称为类内逆文档频率，根据上述逆文档频率的公式，类内逆文档频率越小，关键词在待处理的文本文件所属领域越具有代表性，为上述类内频率设置阈值，即第二阈值，将类内逆文档频率高于上述阈值的关键词作为第三关键词集合，将上述第三关键词集合作为上述待处理文本文件的关键词集合，从而过滤掉在大规模语料中具有代表性，但是在待处理的文本文件所属领域不具有代表性的关键词，从而提高关键词提取的精度。

需要说明的是，这里的第二阈值与第一阈值之间没有任何关系，上述第二阈值为大于0的数，具体取值可以通过实验结果进行调整，只需要保证第一关键词集合中存在关键词的类内逆文档频率高于上述第二阈值即可。

本申请的方案还包括，统计第一关键词集合中各关键词的类内逆文档频率之后，按照类内逆文档频率高到低对第一关键词集合中各关键词进行排序，得到上述第一关键词集合中各个关键词的排名的名次，按照排名的名次从小到大选择关键词作为第四关键词集合，将第四关键词集合作为上述待处理的文本文件的关键词集合，具体的选择方式可以是排名的第一个开始选择个数符合要求的关键词。比如，第一关键词词中包含25个关键词，按照类内逆文档频率由高到低进行排序后，选择前5个，或前10个，或前15个关键词作为第四关键词集合，只要选择的关键词个数小于第一关键词集合中关键词的个数即可，具体选择可以根据实验结果进行调整。

例如，在中美贸易战时期、疫情时期，某个市的工业和信息化局发布的公文中会带有该时期的特殊词汇，本身工业和信息化局发布的公文所属的领域的相关语料资源较少，通过传统的TF-IDF方法，基于大规模语料对上述公文提取的关键词使，如果提取的关键词为20个，这20个关键词中往往会包括关键词“企业”或者“税收”等，但是上述关键词在工业和信息化局发布的公文的领域内是一个非常常见的词语，不具有代表性，我们希望得到的是与贸易战、疫情相关的关键词；将语料库内的文本文件替换为工业和信息化局发布的公文同领域的文本文件，上述关键词“企业”或者“税收”在语料库内的文本文件中大量存在，那么其类内逆文档频率就会较低，可以只保留频率大于0.5的关键词；或者对上述20个关键词按照类内逆文档频率进行排序，只保留排名前10的关键词，作为上述公文的关键词，这样就可以过滤掉在上述公文所属领域中不具有代表性的关键词。

具体的选择方式还可以是从排名的名次大于1的关键词开始，依次选择个数符合实验要求的关键词。特别地，为了提高提取的关键词在上述待处理的文本文件所属领域的特异性，可以统计第一关键词集合中各关键词的类内逆文档频率，将上述关键词按照上述类内逆文档频率由高到低进行排序，将排名前第一百分比的关键词作为第一候选关键词集合，将排名前第二百分比，但是不属于上述第一候选关键词集合的关键词作为第二候选关键词集合，上述第一百分比和上述第二百分比的取值为0到100%之间，但是上述第一百分比小于上述第二百分比，再从上述第一候选关键词集合和第二候选关键词集合中选择关键词作为第二关键词集合。

比如，第一关键词集合包含20个关键词，将低级关键词集合中每个关键词按照类内类内逆文档频率由高到低进行排序之后，将排名前40%的5个关键词作为第一候选关键词集合，将不包含上述第一候选关键词集合中的关键词且排名前80%的关键词，即排名在40%至80%的11个关键词作为第二候选关键词集合，再根据实验需要选择关键词，比如，总共需要提取出10个关键词，可以从第一候选关键词中选择5个关键词，再从第二候选关键词中选择5个关键词即可。

通过将待处理的文本文件同领域的文本文件作为语料库，以关键词在语料库中出现的频率来筛选出在待处理的文本文件所属领域具有代表性的关键词，可以提高关键词的提取精度，除了上述基于词频的方法来筛选关键词，关键词在待处理的文本文件中的分布也

是衡量关键词的重要程度的因素，有的关键词的权重值很高，但是却只在待处理的文本文件的一句话或比较集中的几句话出现，这样的关键词的代表性不强；有的关键词权重值不是很高，但是在待处理的文本文件中多个地方都出现，这样的关键词代表性更强。请参阅图 2，图 2 是本申请实施例公开的又一种关键词提取方法的流程示意图，如图所示，上述方法包括：

S201：确定第一关键词在上述待处理的文本文件中出现的不同段落的个数，得到上述第一关键词对应的段落数。

其中，上述从第四关键词集合中获取的第一关键词不是第四关键词集合中排名第一的关键词，这里的第一关键词指的是第二关键词集合中任意一个关键词，没有特定的顺序。对于从第四关键词集合中获取的任意一个关键词，需要确定上述关键词在待处理的文本文件中所在不同的段落的个数，得到上述关键词的段落数。例如，待处理的文本文件一共有 5 个段落，对于关键词“基础设施”来说，在第一个段落出现 3 次，第二个段落出现 5 次，第 4 个和第 5 个段落分别出现 7 次，那么上述关键词“基础设施”在待处理的文本文件中所在的不同的段落数为 4。

S202：确定上述段落数在上述第四关键词集合中各关键词对应的段落数的排序，得到第一排序值。

上述第四关键词集合中每个关键词都对应各自的段落数，将上述第四关键词集合中各关键词按照段落数由多到少排序可以得到一组段落数排序，可以记为 rank1；确定上述第一关键词在上述排序中的排序值，上述排序值可以称为第一排序值，需要说明的是，上述第一排序值指的是上述第一关键词对应的排序值，上述“第一”并没有特殊的顺序含义。

S203：计算上述第一关键词在上述待处理的文本文件中的词频和在语料库中的第二逆文档频率的乘积，得到上述第一关键词的权重值，上述语料库包括与上述待处理的文本文件类型相同和不同的文本文件。

在初步提取关键词时，需要计算候选关键词的词频 TF 乘以逆文档频率 IDF 得到每个候选关键词的 TF-IDF 值，上述步骤中，计算上述第一关键词在待处理的文本文件中的词频 TF 乘以第二逆文档频率，得到上述第一关键词的权重值，上述第二逆文档频率指的是语料库对应的文本文件类型同时包括与上述待处理的文本文件相同和不同的文件类型，即上述大规模语料。

S204：确定上述权重值在上述第四关键词集合中各关键词对应的权重值的排序，得到第二排序值。

与上述第一关键词的段落数类似，第四关键词集合中的每个关键词都对应一个权重值，将上述第四关键词集合中各关键词按照权重值由大到小排序可以得到一组权重值排序，可以记为 rank2；上述第一关键词在上述 rank2 中排名的名次可以记为第二排序值，需要说明的是，上述第二排序值指的是上述第一关键词对应的排序值，上述“第二”并没有特殊的顺序含义。

S205：将上述第一排序值与上述第二排序值的加权和作为上述第一关键词的排序参考值。

第一排序值是上述第一关键词在上述根据段落数多少进行排序得到的排名的名次，第二排序值是上述第一关键词在上述根据权重值大小进行排序得到的排名的名次，对上述第一排序值和第二排序值分别赋予加权值，上述加权值可以通过实验进行调整，计算上述排名的名次与加权值的加权和，上述加权和就为上述第一关键词进行排序的参考值。

例如，关键词“基础设施”对应的段落数为 4，在第二关键词集合中各关键词对应的段落数由多到少的排名，即上述 rank1 中排名第 3；关键词“基础设施”对应的权重值在上述 rank2 中排名第 5；分别对上述 rank1 和 rank2 赋予加权值 0.5 和 0.6，那么上述关键

词“基础设施”最终的排序参考值为 $4 \times 0.5 + 3 \times 0.6$ ，为 3.8。

特别地，对上述第一排序值和第二排序值分别赋予加权值时，将上述第一排序值的加权值与上述第二排序值的加权值的和设置为 1，并且，上述第一排序值的加权值大于 0 且小于或等于 0.5，这样，可以让最终的排序以权重值为主，段落数分布为辅。

5 S206：按照上述第一关键词的排序参考值的大小确定上述第一关键词在上述第四关键词集合中的顺序。

对于上述第二关键词集合中任意一个关键词，即上述第一关键词来说，上述排序参考值的大小就是其在上述第四关键词集合中的排序的依据；对于整个上述第四关键词集合来说，每个关键词都对应一个排序参考值，将上述第四关键词集合中各关键词按照其排序参考值由大到小进行排序，可以调整上述第二关键词集合中各关键词的顺序，让真正具有代
10 表性的关键词靠前。

提取关键词时，往往都会对待处理的文本文件进行分词处理，经过分词之后，某些本身可以作为上述待处理的文本文件的关键词的组词或者新词可能会被分成 2 个或以上的词，但是关键词提取方法往往只能提取其中一个词，不能提取完整的关键词。请参阅图 3，
15 图 3 是本申请实施例公开的又一种关键词提取方法的流程示意图，如图所示，上述方法包括：

S301：将与上述第二关键词之间相差分词个数小于第三阈值的分词，按照在上述待处理的文本文件中的先后顺序与上述第二关键词进行组合得到组词集合；

其中，上述第二关键词是上述第四关键词集合中任意一个关键词，没有特殊的顺序含义。组词经过分词有可能被分成多个分词，上述多个分词在上述待处理的文本文件中必定是相邻的。上述步骤中，可以首先定位上述第二关键词在原文中的位置，再选择与上述位置左右相邻的分词与上述第二关键词进行组合，得到组词集合。其中选择与上述位置左右相邻的分词时，分词的个数小于阈值，即第三阈值，上述第三阈值为大于 0 的正整数；与上述第二关键词进行组合时，分词与上述第二关键词的顺序必须在按照上述待处理的文
20 本文件中的先后顺序进行组合得到组词。

例如，上述第二关键词记为 wordn，选择在上述待处理的文本文件中与上述 wordn 之间相差分词个数小于 4 的分词进行组合，即选择与上述第二关键词左右三个分词进行组合，上述选择的分词与上述第二关键词可以记为 [wordn-3, wordn-2, wordn-1, wordn, wordn+1, wordn+2, wordn+3]，对关键词和分词进行组合时，按照在待处理的文本文件中的先后顺序
30 进行组合，即可以按照 [wordn-3, wordn-2, wordn-1, wordn]、[wordn-2, wordn-1, wordn]、[wordn-1, wordn]、[wordn, wordn+1]、[wordn, wordn+1, wordn+2]、[wordn, wordn+1, wordn+2, wordn+3]、[wordn-1, wordn, wordn+1] 等进行组合得到组词集合。

需要说明的是，上述分词的个数可以通过实验进行调整，上述例子中选择前后 3 个分词主要考虑到实际情况中的组词一般最多由 4 个单独的分词进行组合得到；上述第二关键词在上述待处理的文本文件中出现的位置可能有多个，对每个位置都采取上述步骤，进行分词选择和组合得到组词集合。
35

S302：在组词在上述待处理的文本文件中的词频除以上述第二关键词在上述待处理的文本文件中的词频大于第四阈值的情况下，将上述组词作为上述第二关键词。

其中，上述组词为上述组词集合中任意一个组词将上述组词在上述待处理的文本文件中进行全文遍历，得到上述组词在上述待处理的文本文件中的词频，如果上述组词的词频除以上述第二关键词的词频大于阈值，那么上述组词对于上述待处理的文本文件来说更加具有代表性，那么就可以将上述组词作为上述第二关键词，即用上述组词替换上述第二关键词。
40

其中，上述第四阈值为大于 0.5 小于 1 之间的数，为了保证组词在待处理文本文件

的代表性,上述阈值可以设为小于1但是大于或等于0.75的数,也可以通过实验结果调整,本申请补做任何限制,通过将上述组合词在上述待处理的文本文件中进行全文遍历,再利用词频来衡量组合词的方法,可以保证上述组合词在上述待处理的文本文件中的重要性,即只有在上述组合词相较于上述关键词在文中出现的频率超过一定值,才可以把上述组合词作为关键词,这样提取出来的组合词是具有代表性的关键词。

特别地,在完成上述第四关键词集合中所有关键词与分词进行组合和筛选之后,若上述第四关键词集合的关键词之间存在包含关系,对上述第四关键词集合中的关键词进行分词,将不重复的分词作为第四关键词集合,将上述第四关键词集合作为上述待处理的文本文件的关键词集合。

综上所述,本申请提供的关键词提取的方法,首先改进关键词的选择方法,过滤掉在待处理的文本文件所属领域中不具有代表性的关键词;再通过综合考虑关键词在待处理的文本文件中的分布和关键词的逆文档频率,优化关键词集合中各关键词的排名的名次,让更加具有代表性的关键词排名靠前;最后通过关键词与相邻的分词按照在待处理的文本文件中先后顺序进行组合,提取出因为分词而被拆分的组合词,从而提高关键词提取的精度。

上述详细阐述了本申请实施例的方法,下面提供本申请实施例的装置。

图4是本申请实施例公开的一种关键词提取的装置的结构示意图;上述关键词提取的装置40可以包括提取单元401、统计单元402、确定单元403,其中,各个单元的描述如下:

提取单元401,用于对待处理的文本文件进行关键词提取,得到第一关键词集合;

统计单元402,用于统计上述第一关键词集合中各关键词在语料库中出现的频率,上述语料库内包含的文本文件与上述待处理的文本文件属于相同文件类型;

确定单元403,用于将上述第一关键词集合中,在上述语料库中出现的频率低于第一阈值的关键词作为第二关键词集合;将上述第二关键词集合作为上述待处理的文本文件的关键词集合。

在一种可能的实施方式中,上述统计单元402,还用于统计上述第一关键词集合中各关键词在语料库中的第一逆文档频率,上述第一逆文档频率对应的语料库内包含的文本文件与上述待处理的文本文件属于相同文件类型;

上述确定单元403,上述确定单元,还用于将上述第一关键词集合中,第一逆文档频率高于第二阈值的关键词作为第三关键词集合;将上述第三关键词集合作为上述待处理的文本文件的关键词集合。

在一种可能的实施方式中,上述装置还包括:

排序单元404,用于将上述第一关键词集合按照上述第一逆文档频率由高到低进行排序;

上述确定单元403,还用于依据各关键词的排名的名次从小到大选择关键词作为第四关键词集合;将上述第四关键词集合作为上述待处理的文本文件的关键词集合。

在一种可能的实施方式中,上述确定单元403,还用于将排名位于前第一百分比的关键词作为第一候选关键词集合,将不包括所述第一候选关键词集合中的关键词且排名位于前第二百分比的关键词作为第二候选关键词集合;所述第二百分比大于所述第一百分比;从所述第一候选关键词集合和第二候选关键词集合中选择关键词作为第四关键词集合。

在一种可能的实施方式中,上述装置还包括:

上述确定单元403,还用于确定第一关键词在所述待处理的文本文件中出现的不同段落的个数,得到所述第一关键词对应的段落数;确定所述段落数在所述第四关键词集合中各关键词对应的段落数的排序,得到第一排序值;所述第一关键词为所述第四关键词集合

中任意一个关键词；

计算单元 405，用于计算上述第一关键词在上述待处理的文本文件中的词频和在语料库中的第二逆文档频率的乘积，得到上述第一关键词的权重值，上述语料库包括与上述待处理的文本文件类型相同和不同的文本文件；

5 上述确定单元 403，还用于确定上述权重值在上述第四关键词集合中各关键词对应的权重值的排序，得到第二排序值；将上述第一排序值与上述第二排序值的加权和作为上述第一关键词的排序参考值；按照上述第一关键词的排序参考值的大小确定上述第一关键词在上述第四关键词集合中的顺序。

10 在一种可能的实施方式中，上述第一排序值的加权值与上述第二排序值的加权值的和为 1，上述第一排序值的加权值大于 0 且小于或等于 0.5。

本申请实施例中，对上述第一排序值和第二排序值分别赋予加权值时，将上述第一排序值的加权值与上述第二排序值的加权值的和设置为 1，并且，上述第一排序值的加权值大于 0 且小于或等于 0.5，这样，可以让最终的排序以权重值为主，段落数分布为辅。

在一种可能的实施方式中，上述装置还包括：

15 组合单元 406，用于将与第二关键词之间相差分词个数小于第三阈值的分词，按照在所述待处理的文本文件中的先后顺序与所述第二关键词进行组合得到组合词集合，所述第二关键词为所述第四关键词集合中任意一个关键词；

20 上述确定单元 403，还用于在组合词在所述待处理的文本文件中的词频除以所述第二关键词在所述待处理的文本文件中的词频大于第四阈值的情况下，将所述组合词作为所述第二关键词，所述组合词为所述组合词集合中任意一个组合词。

25 本申请实施例中，组合词经过分词有可能被分成多个分词，上述多个分词在上述待处理的文本文件中必定是相邻的。上述步骤中，可以首先定位上述第四关键词在原文中的位置，再选择与上述位置左右相邻的分词与上述第二关键词进行组合，得到组合词集合。其中选择与上述位置左右相邻的分词时，分词的个数小于阈值；与上述第二关键词进行组合时，分词与上述第二关键词的顺序必须在按照上述待处理的文本文件中的先后顺序进行组合得到组合词。

30 将上述组合词在上述待处理的文本文件中进行全文遍历，得到上述组合词在上述待处理的文本文件中的词频，如果上述组合词的词频除以上述第二关键词的词频大于阈值，那么上述组合词对于上述待处理的文本文件来说更加具有代表性，那么就可以将上述组合词作为上述第二关键词，即用上述组合词替换上述第二关键词。

35 其中，上述阈值可以设为 0.75，也可以通过实验结果调整，通过将上述组合词在上述待处理的文本文件中进行全文遍历，再利用词频来衡量组合词的方法，可以保证上述组合词在上述待处理的文本文件中的重要性，即只有在上述组合词相较于上述关键词在文中出现的频率超过一定值，才可以把上述组合词作为关键词，这样提取出来的组合词是具有代表性的关键词。

需要说明的是，上述分词的个数可以通过实验进行调整，上述例子中选择前后 3 个分词主要考虑到实际情况中的组合词一般最多由 4 个单独的分词进行组合得到。

上述第二关键词在上述待处理的文本文件中出现的位置可能有多个，对每个位置都采取上述步骤，进行分词选择和组合得到组合词集合。

40 特别地，在完成上述第四关键词集合中所有关键词与分词进行组合和筛选之后，若上述第四关键词集合的关键词之间存在包含关系，对上述第二关键词集合中的关键词进行分词，将不重复的分词作为第二关键词集合，将上述第二关键词集合作为上述待处理的文本文件的关键词集合。

综上所述，本申请提供的关键词提取的方法，首先改进关键词的选择方法，再通过考虑关键词在待处理的文本文件中的分布，优化关键词集合中各关键词的排序，最后通过关键词与相邻的分词按照在待处理的文本文件中先后顺序进行组合，提取出因为分词而被拆分的组合词，从而提高关键词提取的精度。

5 请参阅图 5，图 5 是本申请实施例公开的一种服务器的结构示意图。上述服务器 50 可以包括存储器 501、处理器 502。进一步可选的，还可以包含通信接口 503 以及总线 504，其中，存储器 501、处理器 502 以及通信接口 503 通过总线 504 实现彼此之间的通信连接。通信接口 503 用于与上述关键词提取的装置 40 进行数据交互。

10 其中，存储器 501 用于提供存储空间，存储空间中可以存储操作系统和计算机程序等数据。存储器 501 包括但不限于是随机存储记忆体 (random access memory, RAM)、只读存储器 (read-only memory, ROM)、可擦除可编程只读存储器 (erasable programmable read only memory, EPROM)、或便携式只读存储器 (compact disc read-only memory, CD-ROM)。

15 处理器 502 是进行算术运算和逻辑运算的模块，可以是中央处理器 (central processing unit, CPU)、显卡处理器 (graphics processing unit, GPU) 或微处理器 (microprocessor unit, MPU) 等处理模块中的一种或者多种的组合。

存储器 501 中存储有计算机程序，处理器 502 调用存储器 501 中存储的计算机程序，以执行以下操作：

对待处理的文本文件进行关键词提取，得到第一关键词集合；

20 统计上述第一关键词集合中各关键词在语料库中出现的频率，上述语料库内包含的文本文件与上述待处理的文本文件属于相同文件类型；

将上述第一关键词集合中，在上述语料库中出现的频率低于第一阈值的关键词作为第二关键词集合；

将上述第二关键词集合作为上述待处理的文本文件的关键词集合。

25 需要说明的是，服务器 50 的具体实现还可以对应参照图 2、图 3 所示的方法实施例的相应描述。

本申请实施例还提供一种计算机可读存储介质，上述计算机可读存储介质中存储有计算机程序，当上述计算机程序在一个或多个处理器上运行时，可以实现图 1、图 2 以及图 3 所示的关键词提取的方法。

30 可选的，本申请涉及的存储介质如计算机可读存储介质可以是非易失性的，也可以是易失性的。

35 综上所述，首先改进关键词的选择方法，过滤掉在待处理的文本文件所属领域中不具有代表性的关键词；再通过综合考虑关键词在待处理的文本文件中的分布和关键词的逆文档频率，优化关键词集合中各关键词的排名的名次，让更加具有代表性的关键词排名靠前；最后通过关键词与相邻的分词按照在待处理的文本文件中先后顺序进行组合，提取出因为分词而被拆分的组合词，从而提高关键词提取的精度。

本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，上述流程可以由计算机程序相关的硬件完成，上述计算机程序可存储于计算机可读存储介质中，上述计算机程序在执行时，可包括如上述各方法实施例的流程。而前述的存储介质包括：只读存储器 ROM 或随机存储记忆体 RAM、磁碟或者光盘等各种可存储计算机程序代码的介质。

权利要求

1、一种关键词提取的方法，包括：

对待处理的文本文件进行关键词提取，得到第一关键词集合；

5 统计所述第一关键词集合中各关键词在语料库中出现的频率，将所述第一关键词集合中，在所述语料库中出现的频率低于第一阈值的关键词作为第二关键词集合，所述语料库中包含的文本文件与所述待处理的文本文件属于相同文件类型；

将所述第二关键词集合作为所述待处理的文本文件的关键词集合。

2、根据权利要求1所述的方法，其中，所述方法还包括：

10 统计所述第一关键词集合中各关键词在语料库中的第一逆文档频率，所述第一逆文档频率对应的语料库内包含的文本文件与所述待处理的文本文件属于相同文件类型；

将所述第一关键词集合中，第一逆文档频率高于第二阈值的关键词作为第三关键词集合；

将所述第三关键词集合作为所述待处理的文本文件的关键词集合。

3、根据权利要求2所述的方法，其中，所述方法还包括：

15 将所述第一关键词集合按照所述第一逆文档频率由高到低进行排序，得到所述第一关键词集合中各关键词排名的名次；

依据各关键词的排名的名次从小到大选择关键词作为第四关键词集合；

将所述第四关键词集合作为所述待处理的文本文件的关键词集合。

20 4、根据权利要求3所述的方法，其中，所述依据各关键词的排名的名次从小到大选择关键词作为第四关键词集合，包括：

将排名位于前第一百分比的关键词作为第一候选关键词集合，将不包括所述第一候选关键词集合中的关键词且排名位于前第二百分比的关键词作为第二候选关键词集合；所述第二百分比大于所述第一百分比；

从所述第一候选关键词集合和第二候选关键词集合中选择关键词作为第四关键词集合。

25 5、根据权利要求3所述的方法，其中，所述方法还包括：

确定第一关键词在所述待处理的文本文件中出现的不同段落的个数，得到所述第一关键词对应的段落数；确定所述段落数在所述第四关键词集合中各关键词对应的段落数的排序，得到第一排序值；所述第一关键词为所述第四关键词集合中任意一个关键词；

30 计算所述第一关键词在所述待处理的文本文件中的词频和在语料库中的第二逆文档频率的乘积，得到所述第一关键词的权重值，所述语料库包括与所述待处理的文本文件类型相同和不同的文本文件；确定所述权重值在所述第四关键词集合中各关键词对应的权重值的排序，得到第二排序值；

将所述第一排序值与所述第二排序值的加权和作为所述第一关键词的排序参考值；

35 按照所述第一关键词的排序参考值的大小确定所述第一关键词在所述第四关键词集合中的顺序。

6、根据权利要求5所述的方法，其中，所述方法还包括：

所述第一排序值的加权值与所述第二排序值的加权值的和为1，所述第一排序值的加权值大于0且小于或等于0.5。

7、根据权利要求6所述的方法，其中，所述方法还包括：

40 将与第二关键词之间相差分词个数小于第三阈值的分词，按照在所述待处理的文本文件中的先后顺序与所述第二关键词进行组合得到组合词集合，所述第二关键词为所述第四关键词集合中任意一个关键词；

在组合词在所述待处理的文本文件中的词频除以所述第二关键词在所述待处理的文本文件中的词频大于第四阈值的情况下，将所述组合词作为所述第二关键词，所述组合词为

所述组合词集合中任意一个组合词。

8、一种关键词提取的装置，其中，所述装置包括：

提取单元，用于对待处理的文本文件进行关键词提取，得到第一关键词集合；

5 统计单元，用于统计所述第一关键词集合中各关键词在语料库中出现的频率，所述语料库内包含的文本文件与所述待处理的文本文件属于相同文件类型；

确定单元，用于将所述第一关键词集合中，在所述语料库中出现的频率低于第一阈值的关键词作为第二关键词集合；将所述第二关键词集合作为所述待处理的文本文件的关键词集合。

10 9、一种服务器，其中，所述服务器包括处理器和存储器，其中，所述存储器中存储有计算机程序，所述处理器调用所述存储器中存储的计算机程序，用于执行以下方法：

对待处理的文本文件进行关键词提取，得到第一关键词集合；

统计所述第一关键词集合中各关键词在语料库中出现的频率，将所述第一关键词集合中，在所述语料库中出现的频率低于第一阈值的关键词作为第二关键词集合，所述语料库内包含的文本文件与所述待处理的文本文件属于相同文件类型；

15 将所述第二关键词集合作为所述待处理的文本文件的关键词集合。

10、根据权利要求 9 所述的服务器，其中，所述处理器还用于执行：

统计所述第一关键词集合中各关键词在语料库中的第一逆文档频率，所述第一逆文档频率对应的语料库内包含的文本文件与所述待处理的文本文件属于相同文件类型；

20 将所述第一关键词集合中，第一逆文档频率高于第二阈值的关键词作为第三关键词集合；

将所述第三关键词集合作为所述待处理的文本文件的关键词集合。

11、根据权利要求 10 所述的服务器，其中，所述处理器还用于执行：

将所述第一关键词集合按照所述第一逆文档频率由高到低进行排序，得到所述第一关键词集合中各关键词排名的名次；

25 依据各关键词的排名的名次从小到大选择关键词作为第四关键词集合；

将所述第四关键词集合作为所述待处理的文本文件的关键词集合。

12、根据权利要求 11 所述的服务器，其中，执行所述依据各关键词的排名的名次从小到大选择关键词作为第四关键词集合，包括：

30 将排名位于前第一百分比的关键词作为第一候选关键词集合，将不包括所述第一候选关键词集合中的关键词且排名位于前第二百分比的关键词作为第二候选关键词集合；所述第二百分比大于所述第一百分比；

从所述第一候选关键词集合和第二候选关键词集合中选择关键词作为第四关键词集合。

13、根据权利要求 11 所述的服务器，其中，所述处理器还用于执行：

35 确定第一关键词在所述待处理的文本文件中出现的不同段落的个数，得到所述第一关键词对应的段落数；确定所述段落数在所述第四关键词集合中各关键词对应的段落数的排序，得到第一排序值；所述第一关键词为所述第四关键词集合中任意一个关键词；

计算所述第一关键词在所述待处理的文本文件中的词频和在语料库中的第二逆文档频率的乘积，得到所述第一关键词的权重值，所述语料库包括与所述待处理的文本文件类型相同和不同的文本文件；确定所述权重值在所述第四关键词集合中各关键词对应的权重值的排序，得到第二排序值；

40 将所述第一排序值与所述第二排序值的加权和作为所述第一关键词的排序参考值；

按照所述第一关键词的排序参考值的大小确定所述第一关键词在所述第四关键词集合中的顺序。

14、根据权利要求 13 所述的服务器，其中，所述第一排序值的加权值与所述第二排序

值的加权值的和为 1, 所述第一排序值的加权值大于 0 且小于或等于 0.5;

所述处理器还用于执行:

将与第二关键词之间相差分词个数小于第三阈值的分词, 按照在所述待处理的文本文件中的先后顺序与所述第二关键词进行组合得到组合词集合, 所述第二关键词为所述第四关键词集合中任意一个关键词;

在组合词在所述待处理的文本文件中的词频除以所述第二关键词在所述待处理的文本文件中的词频大于第四阈值的情况下, 将所述组合词作为所述第二关键词, 所述组合词为所述组合词集合中任意一个组合词。

15、一种计算机可读存储介质, 其中, 所述计算机可读存储介质中存储有计算机程序, 当所述计算机程序在一个或多个处理器上运行时, 执行以下方法:

对待处理的文本文件进行关键词提取, 得到第一关键词集合;

统计所述第一关键词集合中各关键词在语料库中出现的频率, 将所述第一关键词集合中, 在所述语料库中出现的频率低于第一阈值的关键词作为第二关键词集合, 所述语料库内包含的文本文件与所述待处理的文本文件属于相同文件类型;

将所述第二关键词集合作为所述待处理的文本文件的关键词集合。

16、根据权利要求 15 所述的计算机可读存储介质, 其中, 当所述计算机程序在一个或多个处理器上运行时, 还用于执行:

统计所述第一关键词集合中各关键词在语料库中的第一逆文档频率, 所述第一逆文档频率对应的语料库内包含的文本文件与所述待处理的文本文件属于相同文件类型;

将所述第一关键词集合中, 第一逆文档频率高于第二阈值的关键词作为第三关键词集合;

将所述第三关键词集合作为所述待处理的文本文件的关键词集合。

17、根据权利要求 16 所述的计算机可读存储介质, 其中, 当所述计算机程序在一个或多个处理器上运行时, 还用于执行:

将所述第一关键词集合按照所述第一逆文档频率由高到低进行排序, 得到所述第一关键词集合中各关键词排名的名次;

依据各关键词的排名的名次从小到大选择关键词作为第四关键词集合;

将所述第四关键词集合作为所述待处理的文本文件的关键词集合。

18、根据权利要求 17 所述的计算机可读存储介质, 其中, 执行所述依据各关键词的排名的名次从小到大选择关键词作为第四关键词集合, 包括:

将排名位于前第一百分比的关键词作为第一候选关键词集合, 将不包括所述第一候选关键词集合中的关键词且排名位于前第二百分比的关键词作为第二候选关键词集合; 所述第二百分比大于所述第一百分比;

从所述第一候选关键词集合和第二候选关键词集合中选择关键词作为第四关键词集合。

19、根据权利要求 17 所述的计算机可读存储介质, 其中, 当所述计算机程序在一个或多个处理器上运行时, 还用于执行:

确定第一关键词在所述待处理的文本文件中出现的不同段落的个数, 得到所述第一关键词对应的段落数; 确定所述段落数在所述第四关键词集合中各关键词对应的段落数的排序, 得到第一排序值; 所述第一关键词为所述第四关键词集合中任意一个关键词;

计算所述第一关键词在所述待处理的文本文件中的词频和在语料库中的第二逆文档频率的乘积, 得到所述第一关键词的权重值, 所述语料库包括与所述待处理的文本文件类型相同和不同的文本文件; 确定所述权重值在所述第四关键词集合中各关键词对应的权重值的排序, 得到第二排序值;

将所述第一排序值与所述第二排序值的加权和作为所述第一关键词的排序参考值;

按照所述第一关键词的排序参考值的大小确定所述第一关键词在所述第四关键词集合中的顺序。

20、根据权利要求 19 所述的计算机可读存储介质，其中，所述第一排序值的加权值与所述第二排序值的加权值的和为 1，所述第一排序值的加权值大于 0 且小于或等于 0.5；

5 当所述计算机程序在一个或多个处理器上运行时，还用于执行：

将与第二关键词之间相差分词个数小于第三阈值的分词，按照在所述待处理的文本文件中的先后顺序与所述第二关键词进行组合得到组合词集合，所述第二关键词为所述第四关键词集合中任意一个关键词；

10 在组合词在所述待处理的文本文件中的词频除以所述第二关键词在所述待处理的文本文件中的词频大于第四阈值的情况下，将所述组合词作为所述第二关键词，所述组合词为所述组合词集合中任意一个组合词。

— 1/2 —

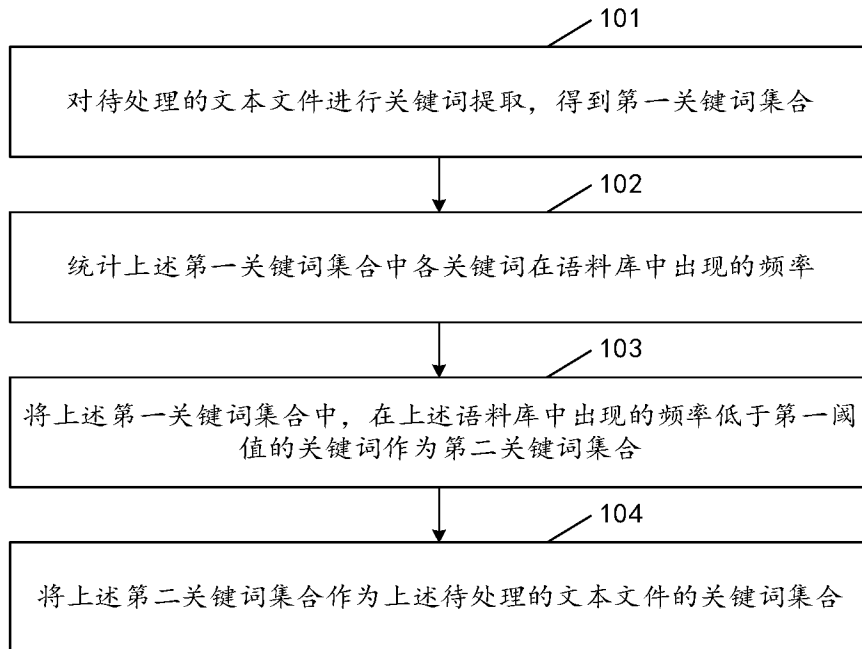


图 1

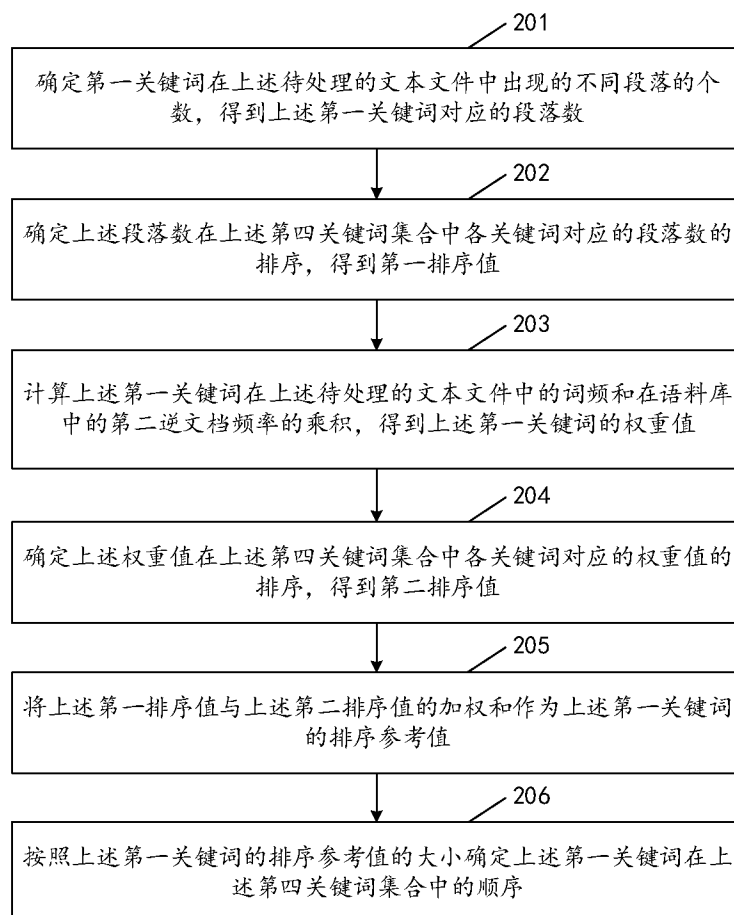


图 2

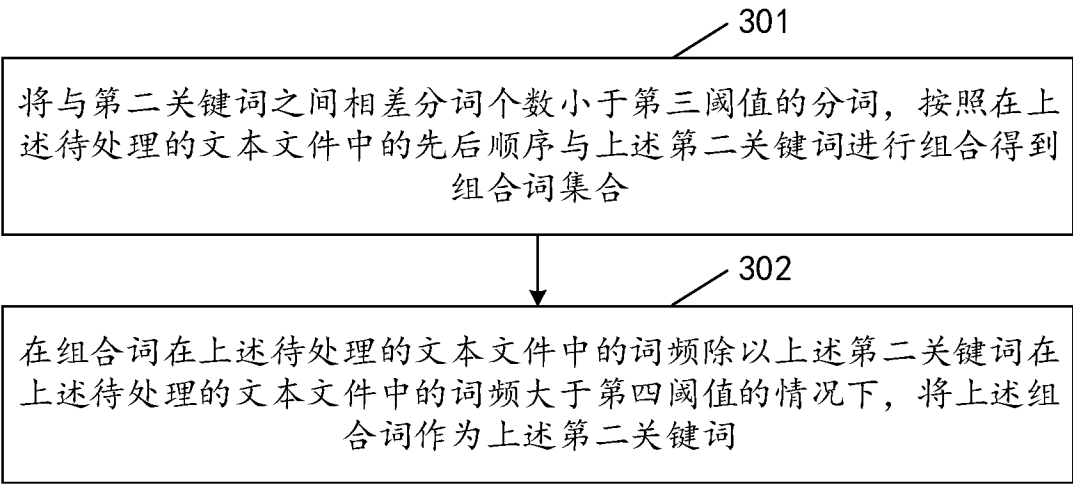


图 3

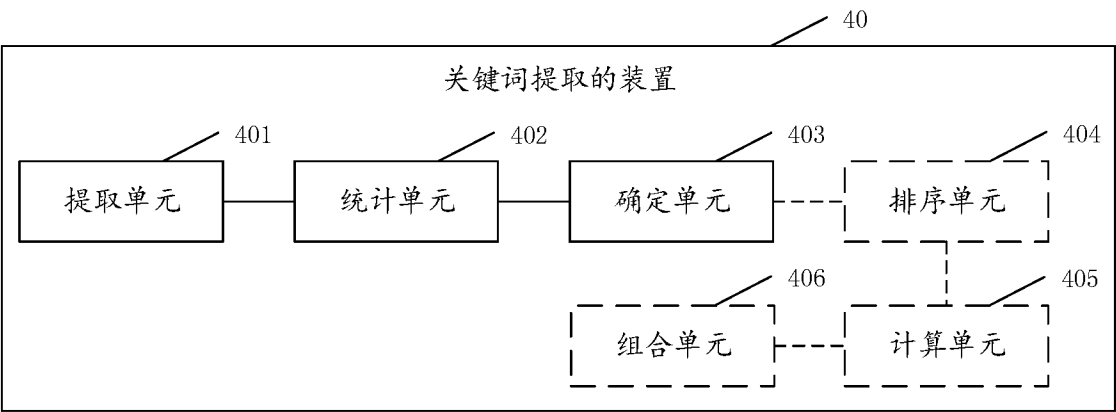


图 4

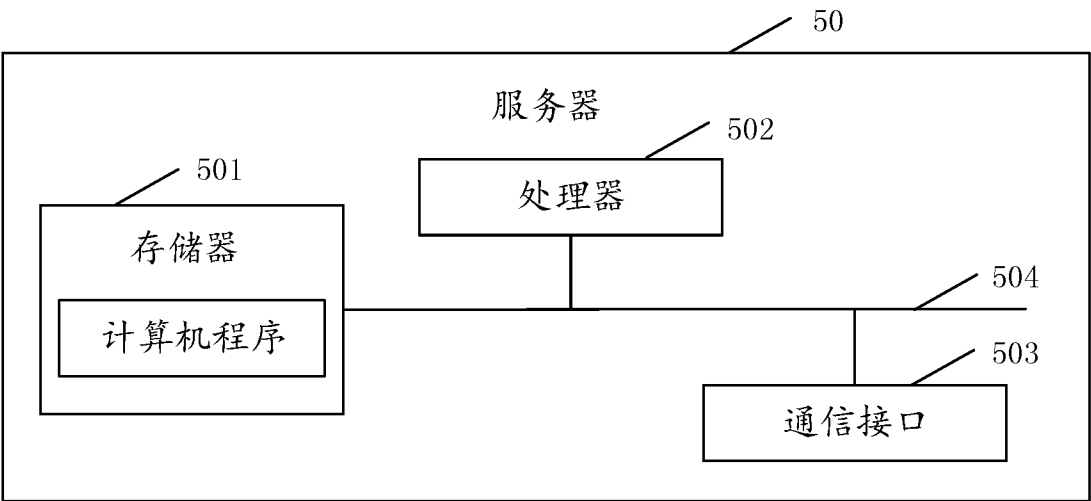


图 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2021/097021

A. CLASSIFICATION OF SUBJECT MATTER G06F 16/33(2019.01)i According to International Patent Classification (IPC) or to both national classification and IPC		
B. FIELDS SEARCHED Minimum documentation searched (classification system followed by classification symbols) G06F Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) WPI, EPODOC, BAIDU, CNPAT, CNKI, IEEE: 关键词, 抽取, 提取, 语料库, 出现, 频率, 次数, 概率, 阈值, 词频, 逆文档频率, 段落, 组合词, keyword, extract, corpus, occurrence, frequency, threshold, TF, IDF, paragraph, compound word		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 112417101 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 26 February 2021 (2021-02-26) claims 1-10, description paragraphs [0005]-[0162]	1-20
Y	CN 106328147 A (UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA et al.) 11 January 2017 (2017-01-11) description paragraphs [0056]-[0075]	1-20
Y	CN 104346418 A (INTERNATIONAL BUSINESS MACHINES CORPORATION) 11 February 2015 (2015-02-11) description, paragraphs [0054]-[0067]	1-20
A	CN 108399165 A (GUANGDONG POLYTECHNIC NORMAL UNIVERSITY) 14 August 2018 (2018-08-14) entire document	1-20
A	CN 110377724 A (XIAMEN MEIYUZHONGYANG INFORMATION TECHNOLOGY CO., LTD.) 25 October 2019 (2019-10-25) entire document	1-20
☒ Further documents are listed in the continuation of Box C. ☒ See patent family annex.		
* Special categories of cited documents: "A" document defining the general state of the art which is not considered to be of particular relevance "E" earlier application or patent but published on or after the international filing date "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) "O" document referring to an oral disclosure, use, exhibition or other means "P" document published prior to the international filing date but later than the priority date claimed "T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention "X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone "Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art "&" document member of the same patent family		
Date of the actual completion of the international search 18 August 2021		Date of mailing of the international search report 26 August 2021
Name and mailing address of the ISA/CN China National Intellectual Property Administration (ISA/CN) No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088, China Facsimile No. (86-10)62019451		Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2021/097021

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2018342241 A1 (BAIDU ONLINE NETWORK TECHNOLOGY BEIJING CO., LTD.) 29 November 2018 (2018-11-29) entire document	1-20

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2021/097021

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	112417101	A	26 February 2021	None			
CN	106328147	A	11 January 2017	None			
CN	104346418	A	11 February 2015	US	2014283097	A1	18 September 2014
CN	108399165	A	14 August 2018	None			
CN	110377724	A	25 October 2019	None			
US	2018342241	A1	29 November 2018	CN	107193973	A	22 September 2017

国际检索报告

国际申请号

PCT/CN2021/097021

A. 主题的分类 G06F 16/33 (2019.01) i 按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类																							
B. 检索领域 检索的最低限度文献 (标明分类系统和分类号) G06F 包含在检索领域中的除最低限度文献以外的检索文献 在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用)) WPI, EPDOC, BAIDU, CNPAT, CNKI, IEEE: 关键词, 抽取, 提取, 语料库, 出现, 频率, 次数, 概率, 阈值, 词频, 逆文档频率, 段落, 组词, keyword, extract, corpus, occurrence, frequency, threshold, TF, IDF, paragraph, compound word																							
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>PX</td> <td>CN 112417101 A (平安科技深圳有限公司) 2021年 2月 26日 (2021 - 02 - 26) 权利要求1-10、说明书第[0005]-[0162]段</td> <td>1-20</td> </tr> <tr> <td>Y</td> <td>CN 106328147 A (中国科学技术大学 等) 2017年 1月 11日 (2017 - 01 - 11) 说明书第[0056]-[0075]段</td> <td>1-20</td> </tr> <tr> <td>Y</td> <td>CN 104346418 A (国际商业机器公司) 2015年 2月 11日 (2015 - 02 - 11) 说明书第[0054]-[0067]段</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>CN 108399165 A (广东技术师范学院) 2018年 8月 14日 (2018 - 08 - 14) 全文</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>CN 110377724 A (厦门美域中央信息科技有限公司) 2019年 10月 25日 (2019 - 10 - 25) 全文</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>US 2018342241 A1 (BAIDU ONLINE NETWORK TECHNOLOGY BEIJING CO., LTD.) 2018年 11月 29日 (2018 - 11 - 29) 全文</td> <td>1-20</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	PX	CN 112417101 A (平安科技深圳有限公司) 2021年 2月 26日 (2021 - 02 - 26) 权利要求1-10、说明书第[0005]-[0162]段	1-20	Y	CN 106328147 A (中国科学技术大学 等) 2017年 1月 11日 (2017 - 01 - 11) 说明书第[0056]-[0075]段	1-20	Y	CN 104346418 A (国际商业机器公司) 2015年 2月 11日 (2015 - 02 - 11) 说明书第[0054]-[0067]段	1-20	A	CN 108399165 A (广东技术师范学院) 2018年 8月 14日 (2018 - 08 - 14) 全文	1-20	A	CN 110377724 A (厦门美域中央信息科技有限公司) 2019年 10月 25日 (2019 - 10 - 25) 全文	1-20	A	US 2018342241 A1 (BAIDU ONLINE NETWORK TECHNOLOGY BEIJING CO., LTD.) 2018年 11月 29日 (2018 - 11 - 29) 全文	1-20
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																					
PX	CN 112417101 A (平安科技深圳有限公司) 2021年 2月 26日 (2021 - 02 - 26) 权利要求1-10、说明书第[0005]-[0162]段	1-20																					
Y	CN 106328147 A (中国科学技术大学 等) 2017年 1月 11日 (2017 - 01 - 11) 说明书第[0056]-[0075]段	1-20																					
Y	CN 104346418 A (国际商业机器公司) 2015年 2月 11日 (2015 - 02 - 11) 说明书第[0054]-[0067]段	1-20																					
A	CN 108399165 A (广东技术师范学院) 2018年 8月 14日 (2018 - 08 - 14) 全文	1-20																					
A	CN 110377724 A (厦门美域中央信息科技有限公司) 2019年 10月 25日 (2019 - 10 - 25) 全文	1-20																					
A	US 2018342241 A1 (BAIDU ONLINE NETWORK TECHNOLOGY BEIJING CO., LTD.) 2018年 11月 29日 (2018 - 11 - 29) 全文	1-20																					
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																							
<table border="0"> <tr> <td> * 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件 </td> <td> “T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件 </td> </tr> </table>			* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件	“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件																			
* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件	“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件																						
国际检索实际完成的日期 2021年 8月 18日		国际检索报告邮寄日期 2021年 8月 26日																					
ISA/CN的名称和邮寄地址 中国国家知识产权局 (ISA/CN) 中国 北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451		受权官员 吴少鸿 电话号码 86-(10)-53961533																					

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2021/097021

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	112417101	A	2021年 2月 26日	无			
CN	106328147	A	2017年 1月 11日	无			
CN	104346418	A	2015年 2月 11日	US	2014283097	A1	2014年 9月 18日
CN	108399165	A	2018年 8月 14日	无			
CN	110377724	A	2019年 10月 25日	无			
US	2018342241	A1	2018年 11月 29日	CN	107193973	A	2017年 9月 22日