

(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
2019 年 8 月 15 日 (15.08.2019)



(10) 国际公布号
WO 2019/153552 A1

- (51) 国际专利分类号:
G06F 17/30 (2006.01)
- (21) 国际申请号: PCT/CN2018/085348
- (22) 国际申请日: 2018 年 5 月 2 日 (02.05.2018)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201810145692.7 2018年2月12日 (12.02.2018) 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八卦岭八卦三路平安大厦六楼, Guangdong 518000 (CN)。
- (72) 发明人: 陈海涛(CHEN, Haitao); 中国广东省深圳市福田区八卦岭八卦三路平安大厦六楼, Guangdong 518000 (CN)。
- (74) 代理人: 深圳市精英专利事务所(SHENZHEN TALENT PATENT SERVICE); 中国广东省深圳市福田区深南中路6009号绿景广场B栋20层B, Guangdong 518000 (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, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, 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,

(54) Title: AUTOMATIC TAGGING METHOD AND APPARATUS, AND COMPUTER DEVICE AND STORAGE MEDIUM

(54) 发明名称: 自动打标签的方法、装置、计算机设备及存储介质

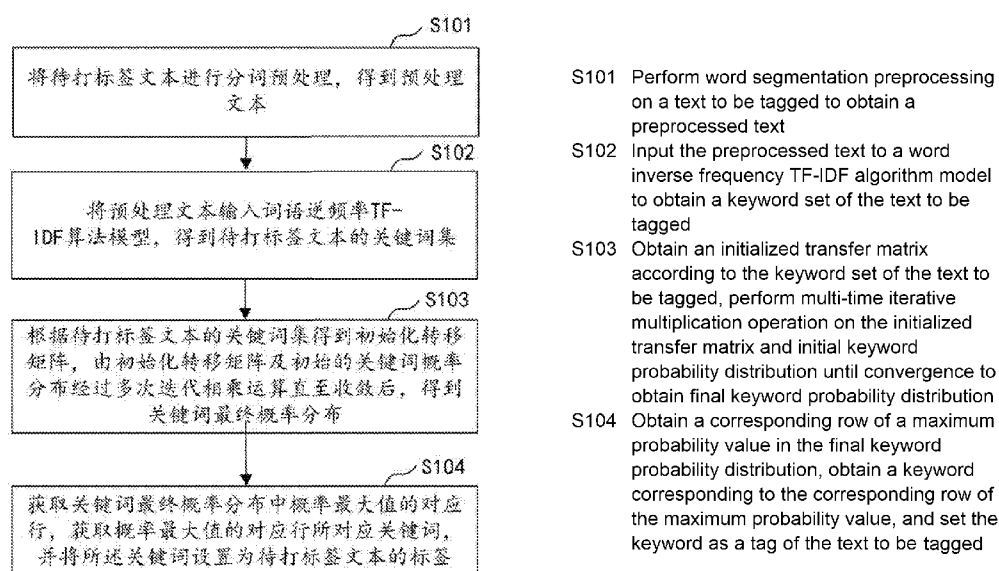


图 1

(57) Abstract: An automatic tagging method and apparatus, and a computer device and a storage medium. The method comprises: performing word segmentation preprocessing on a text to be tagged to obtain a preprocessed text (S101); inputting the preprocessed text to a word inverse frequency TF-IDF algorithm model to obtain a keyword set of the text to be tagged (S102); obtaining an initialized transfer matrix according to the keyword set of the text to be tagged, performing multi-time iterative multiplication operation on the initialized transfer matrix and initial keyword probability distribution until convergence to obtain final keyword probability distribution (S103); and obtaining a corresponding row of a maximum probability value in the final keyword probability distribution, obtaining a keyword corresponding to the corresponding row of the maximum probability value, and setting the keyword as a tag of the text to be tagged

SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG,
US, UZ, VC, VN, 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))。

tagged (S104). According to the method, an article is tagged by means of automatic learning, so that manual tagging is avoided, the tagging efficiency is improved, and the manpower cost is reduced.

(57) 摘要: 一种自动打标签的方法、装置、计算机设备及存储介质。该方法包括: 将待打标签文本进行分词预处理, 得到预处理文本 (S101); 将预处理文本输入词语逆频率TF-IDF算法模型, 得到待打标签文本的关键词集 (S102); 根据待打标签文本的关键词集得到初始化转移矩阵, 由初始化转移矩阵及初始的关键词概率分布经过多次迭代相乘运算直至收敛后, 得到关键词最终概率分布 (S103); 获取关键词最终概率分布中概率最大值的对应行, 获取概率最大值的对应行所对应关键词, 并将所述关键词设置为待打标签文本的标签 (S104)。该方法通过自动学习的方式对文章打标签, 避免了手动打标, 提高打标效率且节省人力成本。

自动打标签的方法、装置、计算机设备及存储介质

本申请要求于2018年2月12日提交中国专利局、申请号为201810145692.7、申请名称为“自动打标签的方法、装置、计算机设备及存储介质”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。

技术领域

本申请涉及文章分类技术领域，尤其涉及一种自动打标签的方法、装置、计算机设备及存储介质。

背景技术

文章的标签有助于文章的搜索以及分类，目前常用的方式是手动打标签，即作者为自己的文章编辑标签，但是并非所有作者都为自己的文章打标签。若海量的未打标签的文章都通过手动打标的方式来实现标签的添加，则效率极其低下，而且极大的增加了人力成本。

发明内容

本申请提供了一种自动打标签的方法、装置、计算机设备及存储介质，旨在解决现有技术海量的未打标签的文章都通过手动打标的方式来实现标签的添加，导致效率极其低下，而且极大增加了人力成本的问题。

第一方面，本申请提供了一种自动打标签的方法，其包括：将待打标签文本进行分词预处理，得到预处理文本；将预处理文本输入词语逆频率 TF-IDF 算法模型，得到待打标签文本的关键词集；根据待打标签文本的关键词集得到初始化转移矩阵，由初始化转移矩阵及初始的关键词概率分布经过多次迭代相乘运算直至收敛后，得到关键词最终概率分布；获取关键词最终概率分布中概率最大值的对应行，获取概率最大值的对应行所对应关键词，并将所述关键词设置为待打标签文本的标签。

第二方面，本申请提供了一种自动打标签的装置，其包括：

文本预处理单元，用于将待打标签文本进行分词预处理，得到预处理文本；

关键词集获取单元，用于将预处理文本输入词语逆频率 TF-IDF 算法模型，得到待打标签文本的关键词集；

最终概率分布获取单元，用于根据待打标签文本的关键词集得到初始化转移矩阵，由初始化转移矩阵及初始的关键词概率分布经过多次迭代相乘运算直至收敛后，得到关键词最终概率分布；

打标单元，用于获取关键词最终概率分布中概率最大值的对应行，获取概率最大值的对应行所对应关键词，并将所述关键词设置为待打标签文本的标签。

第三方面，本申请又提供了一种计算机设备，包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，所述处理器执行所述计算机程序时实现本申请提供的任一项所述的自动打标签的方法。

第四方面，本申请还提供了一种存储介质，其中所述存储介质存储有计算机程序，所述计算机程序包括程序指令，所述程序指令当被处理器执行时使所述处理器执行本申请提供的任一项所述的自动打标签的方法。

本申请提供一种自动打标签的方法、装置、计算机设备及存储介质。该方法通过自动学习的方式对文章打标签，避免了手动打标，提高打标效率且节省人力成本。

附图说明

为了更清楚地说明本申请实施例技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

图1为本申请实施例提供的一种自动打标签的方法的示意图；

图2为本申请实施例提供的一种自动打标签的方法的子流程示意图；

图3是本申请实施例提供的一种自动打标签的方法的另一子流程示意图；

图4为本申请实施例提供的一种自动打标签的装置的示意性框图；

图5为本申请实施例提供的一种自动打标签的装置的子单元示意性框图；

图6为本申请实施例提供的一种自动打标签的装置的另一子单元示意性框图；

图7为本申请实施例提供的一种计算机设备的示意性框图。

具体实施方式

下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

请参阅图 1，图 1 是本申请实施例提供的一种自动打标签的方法的示意图。该方法应用于台式电脑、手提电脑、平板电脑等终端中。如图 1 所示，该方法包括步骤 S101~S104。

S101、将待打标签文本进行分词预处理，得到预处理文本。

如图 2 所示，所述步骤 S101 包括以下步骤：

S1011、对待打标签文本进行分词，得到分词后文本。

在本实施例中，是基于概率统计模型的分词方法对待打标签文本进行分词。基于概率统计模型的分词方法的步骤如下：

步骤十一、对一个待分词的子串 S ，按照从左到右的顺序取出全部候选词 $w_1, w_2, \dots, w_i, \dots, w_n$ ；

步骤十二、到词典中查出每个候选词的概率值 $P(w_i)$ ，并记录每个候选词的全部左邻词；

步骤十三、计算每个候选词的累计概率，同时比较得到每个候选词的最佳左邻词；

步骤十四、如果当前词 w_n 是字串 S 的尾词，且累计概率 $P(w_n)$ 最大，则 w_n 就是 S 的终点词；

步骤十五、从 w_n 开始，按照从右到左顺序，依次将每个词的最佳左邻词输出，即 S 的分词结果。

S1012、对分词后文本包括的分词一一设置加权值。

在本实施例中，以已进行分词的待打标签文本中分词来进行加权处理，也就是已进行分词的待打标签文本中是可以视作由多个分词组成，此时对整篇已进行分词的待打标签文本从头至尾按位置、词性、长度等因素对文本中的各分词进行加权处理，按如下规则：

文本第一个词是标题，赋予权值 8^* ；段首第一个词等于“摘要”，则赋予

权值 5*；段首第一个词等于“关键词”或“结论”，则赋予权值 5*；词语长度等于 2，赋予权值 3*；词性为名词，赋予权值 2*；其他，每段首赋予权值 1*。

S1013、删除分词后文本中的停用词，并统计各分词的词频，得到第一三元组。

三元组 $\langle w_i, \text{fre}_i, v_i \rangle$ 表示待打标签文本经处理后的结果集，其中 w_i 是词语， fre_i 是词语 w_i 加权后出现的次数， v_i 是词语在文本中的位置权重；其中，当对分词后文本包括的分词一一设置加权值后，需删除其中的停用词（停用词包括虚拟词、语气组词、副词、符号、一个字的词，这些停用词不会作为关键词的候选词），能准确的筛选出候选的关键词进行后续处理。

S1014、获取第一三元组中词频大于预设词频阈值所对分词之间的词语相似度；

其中，通过词语相似度计算，计算第一三元组 $\langle w_i, \text{fre}_i, v_i \rangle$ 中词频 $\text{fre}_i > 2$ 的所有词语相似度 sim_{ij} ；当 $\text{sim}_{ij} > 0.9$ 则认为两个词语的相似度极高，在文本中可以替换，将返回四元组 $\langle w_i, w_j, \text{sim}_{ij}, \text{fre}_i + \text{fre}_j \rangle$ ，并删除第一三元组里的词语 w_j 。四元组 $\langle w_i, w_j, \text{sim}_{ij}, \text{fre}_i + \text{fre}_j \rangle$ 表示对三元组中部分词语计算相似度后的集合，其中 sim_{ij} 表示词语 w_i 、 w_j 的相似度， $\text{fre}_i + \text{fre}_j$ 表示两个词语的词频之和。

S1015、若分词之间的词语相似度大于预设词语相似度阈值，保留其中任意一个分词，得到第二三元组，并将第二三元组作为预处理文本。

其中，在第一三元组 $\langle w_i, \text{fre}_i, v_i \rangle$ 中，查找四元组 $\langle w_i, w_j, \text{sim}_{ij}, \text{fre}_i + \text{fre}_j \rangle$ 中的词语；当三元组的 fre_i 替换为四元组中的 $\text{fre}_i + \text{fre}_j$ ，重新组成第二三元组 $\langle w_i, \text{fre}_i + \text{fre}_j, v_i \rangle$ ，该第二三元组 $\langle w_i, \text{fre}_i + \text{fre}_j, v_i \rangle$ 即为预处理文本。通过文本预处理，所得到的预处理文本满足关键词筛选模型的输入标准（即将文本进行了向量化处理），能更为准确的得到文章的关键词。当对待打标签文本进行分词后，就能对各词语的词频进行统计，以作为文章标签的候选词。

S102、将预处理文本输入词语逆频率 TF-IDF 算法模型，得到待打标签文本的关键词集。

在一实施例中，所述词语逆频率 TF-IDF 算法模型为：

$$\text{TF-IDF}_{i,j} \rightarrow \text{TF}_{i,j} \times \text{IDF}_i = \frac{n_{i,j}}{\sum_k n_{i,j}} \times \log \frac{\sum_{i=1}^m n_{t_i}}{n_{t_i}}$$

其中，TF 部分分子 $n_{i,j}$ 表示词语 t_i 在文本 j 中出现的次数，分母表示文本 j

中所有的词语频词和, IWF 部分分子表示语料库中所有词语频数之和, nt_i 表示词语 t_i 在语料库中出现的总频数。

如图 3 所示, 所述步骤 S102 包括以下步骤:

S1021、生成语料库词语统计结果集;

S1022、获取预处理文本;

S1023、将预处理文本及语料库词语统计结果集输入词语逆频率 TF-IDF 算法模型, 得到待打标签文本的关键词集。

在本实施例中, 语料库词语统计结果集是基于语料库得到的。语料库是用户预先挑选一定数量文章(如 2000 篇), 采用步骤 S10111-S10115 中的文本预处理算法, 忽略步骤 S10114 中相似度计算的步骤, 对语料库中的文章分别进行处理, 得到二元组 $\langle w_i, fre_i \rangle$, 其中 w_i 是词语, fre_i 是词语 w_i 加权后出现的频次。合并所有二元组 $\langle w_i, fre_i \rangle$, 得到 $\langle w_i, fre_{isum} \rangle$, 其中 fre_{isum} 是词语 w_i 在语料库中出现的总频次, 也即 $\langle w_i, fre_{isum} \rangle$ 即为所生成的语料库词语统计结果集。将预处理文本及语料库词语统计结果集输入词语逆频率 TF-IDF 算法模型, 得到待打标签文本的关键词集, 所得到的关键词集准确度较高。

S103、根据待打标签文本的关键词集得到初始化转移矩阵, 由初始化转移矩阵及初始的关键词概率分布经过多次迭代相乘运算直至收敛后, 得到关键词最终概率分布。

在一实施例中, 所述根据待打标签文本的关键词集得到初始化转移矩阵中, 所述初始化转移矩阵为 n 维方阵, n 维方阵的维数与关键词集中关键词总个数相等; 所述初始的关键词概率分布为每一行值均为 $1/n$ 的 n 维列向量; 其中, n 为与关键词集中关键词总个数相等的正整数;

所述由初始化转移矩阵及初始的关键词概率分布经过多次迭代相乘运算记为 $V_m = MV_{m-1}$, 其中, m 为正整数, V_0 为初始的关键词概率分布, M 为初始化转移矩阵。

在本实施例中, 例如, 关键词集中关键词的总个数为 4 个, 分别记为 A, B, C, D。根据关键词的位置、词性、长度等因素得到初始化转移矩阵 M 为:

$$M = \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$$

假设每一个关键词为待打标签文本的最终标签的概率都是相等的，即 $1/n$ ；故初始的关键词概率分布就是一个所有值都为 $1/n$ 的 n 维列向量 V_0 ；用 $V_n = MV_{n-1}$ 计算得到关键词最终概率分布 V_n （此初始化转移矩阵 M 乘以 V_{n-1} ，经过不断迭代（一般是 30 次左右）最终会收敛，不会出现一直迭代的情况）。

S104、获取关键词最终概率分布中概率最大值的对应行，获取概率最大值的对应行所对应关键词，并将所述关键词设置为待打标签文本的标签。

作为步骤 S103-S104 的另一实施例，也可以是根据待打标签文本的关键词集得到初始化转移矩阵，由初始化转移矩阵及初始的关键词概率分布经过多次迭代相乘运算，当 V_n 这一 n 维列向量种有一行出现大于预设概率值时，则停止迭代并将该行对应的关键词作为待打标签文本的标签。若同时出现了多行大于预设概率值的情况，则将这些行对应的关键词同时作为待打标签文本的标签。

在一实施例中，所述步骤 S101 之前，还包括：

步骤一、爬取待打标签文本，并存储至 MongoDB 数据库中。即原始数据从网上爬取，得到待打标签文本，存放到 MongoDB 数据库。通过爬取数据，可设置一筛选条件，即爬取未设置标签的文本从而进行打标签。

该方法通过自动学习的方式对文章打标签，避免了手动打标，提高打标效率且节省人力成本。

本申请实施例还提供一种自动打标签的装置，该自动打标签的装置用于执行前述任一项自动打标签的方法。具体地，请参阅图 4，图 4 是本申请实施例提供的一种自动打标签的装置的示意性框图。自动打标签的装置 100 可以安装于台式电脑、平板电脑、手提电脑、等终端中。

如图 4 所示，自动打标签的装置 100 包括文本预处理单元 101、关键词集获取单元 102、最终概率分布获取单元 103、打标单元 104。

文本预处理单元 101，用于将待打标签文本进行分词预处理，得到预处理文本。

如图 5 所示，所述文本预处理单元 101 包括以下子单元：

分词单元 1011, 用于对待打标签文本进行分词, 得到分词后文本。

在本实施例中, 是基于概率统计模型的分词方法对待打标签文本进行分词。基于概率统计模型的分词方法如下:

1) 对一个待分词的子串 S , 按照从左到右的顺序取出全部候选词 $w_1, w_2, \dots, w_i, \dots, w_n$; 2) 到词典中查出每个候选词的概率值 $P(w_i)$, 并记录每个候选词的全部左邻词; 3) 计算每个候选词的累计概率, 同时比较得到每个候选词的最佳左邻词; 4) 如果当前词 w_n 是子串 S 的尾词, 且累计概率 $P(w_n)$ 最大, 则 w_n 就是 S 的终点词; 5) 从 w_n 开始, 按照从右到左顺序, 依次将每个词的最佳左邻词输出, 即 S 的分词结果。

加权单元 1012, 用于对分词后文本包括的分词一一设置加权值。

在本实施例中, 以已进行分词的待打标签文本中分词来进行加权处理, 也就是已进行分词的待打标签文本中是可以视作由多个分词组成, 此时对整篇已进行分词的待打标签文本从头至尾按位置、词性、长度等因素对文本中的各分词进行加权处理, 按如下规则: 文本第一个词是标题, 赋予权值 8^* ; 段首第一个词等于“摘要”, 则赋予权值 5^* ; 段首第一个词等于“关键词”或“结论”, 则赋予权值 5^* ; 词语长度等于 2, 赋予权值 3^* ; 词性为名词, 赋予权值 2^* ; 其他, 每段首赋予权值 1^* 。

统计单元 1013, 用于删除分词后文本中的停用词, 并统计各分词的词频, 得到第一三元组。

三元组 $\langle w_i, \text{fre}_i, v_i \rangle$ 表示待打标签文本经处理后的结果集, 其中 w_i 是词语, fre_i 是词语 w_i 加权后出现的次数, v_i 是词语在文本中的位置权重; 其中, 当对分词后文本包括的分词一一设置加权值后, 需删除其中的停用词 (停用词包括虚拟词、语气组词、副词、符号、一个字的词, 这些停用词不会作为关键词的候选词), 能准确的筛选出候选的关键词进行后续处理。

相似度获取单元 1014, 用于获取第一三元组中词频大于预设词频阈值所对分词之间的词语相似度。

其中, 通过词语相似度计算, 计算第一三元组 $\langle w_i, \text{fre}_i, v_i \rangle$ 中词频 $\text{fre}_i > 2$ 的所有词语相似度 sim_{ij} ; 当 $\text{sim}_{ij} > 0.9$ 则认为两个词语的相似度极高, 在文本中可以替换, 将返回四元组 $\langle w_i, w_j, \text{sim}_{ij}, \text{fre}_i + \text{fre}_j \rangle$, 并删除第一三元组里的词语 w_j 。四元组 $\langle w_i, w_j, \text{sim}_{ij}, \text{fre}_i + \text{fre}_j \rangle$ 表示对三元组中部分词语计算相似度后的集

合，其中 sim_{ij} 表示词语 w_i 、 w_j 的相似度， $\text{fre}_i + \text{fre}_j$ 表示两个词语的词频之和。

删词单元 1015，用于若分词之间的词语相似度大于预设词语相似度阈值，保留其中任意一个分词，得到第二三元组，并将第二三元组作为预处理文本。

其中，在第一三元组 $\langle w_i, \text{fre}_i, v_i \rangle$ 中，查找四元组 $\langle w_i, w_j, \text{sim}_{ij}, \text{fre}_i + \text{fre}_j \rangle$ 中的词语；当三元组的 fre_i 替换为四元组中的 $\text{fre}_i + \text{fre}_j$ ，重新组成第二三元组 $\langle w_i, \text{fre}_i + \text{fre}_j, v_i \rangle$ ，该第二三元组 $\langle w_i, \text{fre}_i + \text{fre}_j, v_i \rangle$ 即为预处理文本。通过文本预处理，所得到的预处理文本满足关键词筛选模型的输入标准（即将文本进行了向量化处理），能更为准确的得到文章的关键词。当对待打标签文本进行分词后，就能对各词语的词频进行统计，以作为文章标签的候选词。

关键词集获取单元 102，用于将预处理文本输入词语逆频率 TF-IDF 算法模型，得到待打标签文本的关键词集。

在一实施例中，所述词语逆频率 TF-IDF 算法模型为：

$$\text{TF-IDF}_{i,j} \rightarrow \text{TF}_{i,j} \times \text{IDF}_i = \frac{n_{i,j}}{\sum_k n_{k,j}} \times \log \frac{\sum_{i=1}^m n_{t_i}}{n_{t_i}}$$

其中，TF 部分分子 $n_{i,j}$ 表示词语 t_i 在文本 j 中出现的次数，分母表示文本 j 中所有的词语频词和，IDF 部分分子表示语料库中所有词语频数之和， n_{t_i} 表示词语 t_i 在语料库中出现的总频数。

如图 6 所示，所述关键词集获取单元 102 包括以下子单元：

第一处理单元 1021，用于生成语料库词语统计结果集；

第二处理单元 1022，用于获取预处理文本；

关键词集计算单元 1023，用于将预处理文本及语料库词语统计结果集输入词语逆频率 TF-IDF 算法模型，得到待打标签文本的关键词集。

在本实施例中，语料库词语统计结果集是基于语料库得到的。语料库是用户预先挑选一定数量文章（如 2000 篇），采用文本预处理单元 101 中的文本预处理算法，忽略相似度计算的步骤，对语料库中的文章分别进行处理，得到二元组 $\langle w_i, \text{fre}_i \rangle$ ，其中 w_i 是词语， fre_i 是词语 w_i 加权后出现的频次。合并所有二元组 $\langle w_i, \text{fre}_i \rangle$ ，得到 $\langle w_i, \text{fre}_{\text{isum}} \rangle$ ，其中 fre_{isum} 是词语 w_i 在语料库中出现的总频次，也即 $\langle w_i, \text{fre}_{\text{isum}} \rangle$ 即为所生成的语料库词语统计结果集。将预处理文本及语料库词语统计结果集输入词语逆频率 TF-IDF 算法模型，得到待打标签文本的关键词集，所得到的关键词集准确度较高。

最终概率分布获取单元 103, 用于根据待打标签文本的关键词集得到初始化转移矩阵, 由初始化转移矩阵及初始的关键词概率分布经过多次迭代相乘运算直至收敛后, 得到关键词最终概率分布。

在一实施例中, 所述根据待打标签文本的关键词集得到初始化转移矩阵中, 所述初始化转移矩阵为 n 维方阵, n 维方阵的维数与关键词集中关键词总个数相等; 所述初始的关键词概率分布为每一行值均为 $1/n$ 的 n 维列向量; 其中, n 为与关键词集中关键词总个数相等的正整数;

所述由初始化转移矩阵及初始的关键词概率分布经过多次迭代相乘运算记为 $V_m = MV_{m-1}$, 其中, m 为正整数, V_0 为初始的关键词概率分布, M 为初始化转移矩阵。

在本实施例中, 例如, 关键词集中关键词的总个数为 4 个, 分别记为 A, B, C, D。根据关键词的位置、词性、长度等因素得到初始化转移矩阵 M 为:

$$M = \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$$

假设每一个关键词为待打标签文本的最终标签的概率都是相等的, 即 $1/n$; 故初始的关键词概率分布就是一个所有值都为 $1/n$ 的 n 维列向量 V_0 ; 用 $V_n = MV_{n-1}$ 计算得到关键词最终概率分布 V_n (此初始化转移矩阵 M 乘以 V_{n-1} , 经过不断迭代 (一般是 30 次左右) 最终会收敛, 不会出现一直迭代的情况)。

打标单元 104, 用于获取关键词最终概率分布中概率最大值的对应行, 获取概率最大值的对应行所对应关键词, 并将所述关键词设置为待打标签文本的标签。

作为最终概率分布获取单元 103 和打标单元 104 的另一实施例, 也可以是根据待打标签文本的关键词集得到初始化转移矩阵, 由初始化转移矩阵及初始的关键词概率分布经过多次迭代相乘运算, 当 V_n 这一 n 维列向量种有一行出现大于预设概率值时, 则停止迭代并将该行对应的关键词作为待打标签文本的标签。若同时出现了多行大于预设概率值的情况, 则将这些行对应的关键词同时作为待打标签文本的标签。

在一实施例中, 自动打标签的装置 100, 还包括:

爬取单元，用于爬取待打标签文本，并存储至 MongoDB 数据库中。即原始数据从网上爬取，得到待打标签文本，存放到 MongoDB 数据库。通过爬取数据，可设置一筛选条件，即爬取未设置标签的文本从而进行打标签。

可见，该装置通过自动学习的方式对文章打标签，避免了手动打标，提高打标效率且节省人力成本。

上述自动打标签的装置可以实现为一种计算机程序的形式，该计算机程序可以在如图 7 所示的计算机设备上运行。

请参阅图 7，图 7 是本申请实施例提供的一种计算机设备的示意性框图。该计算机设备 500 设备可以是终端。该终端可以是平板电脑、笔记本电脑、台式电脑、个人数字助理等电子设备。

参阅图 7，该计算机设备 500 包括通过系统总线 501 连接的处理器 502、存储器和网络接口 505，其中，存储器可以包括非易失性存储介质 503 和内存储器 504。该非易失性存储介质 503 可存储操作系统 5031 和计算机程序 5032。该计算机程序 5032 包括程序指令，该程序指令被执行时，可使得处理器 502 执行一种自动打标签的方法。该处理器 502 用于提供计算和控制能力，支撑整个计算机设备 500 的运行。该内存储器 504 为非易失性存储介质 503 中的计算机程序 5032 的运行提供环境，该计算机程序 5032 被处理器 502 执行时，可使得处理器 502 执行一种自动打标签的方法。该网络接口 505 用于进行网络通信，如发送分配的任务等。本领域技术人员可以理解，图 7 中示出的结构，仅仅是与本申请方案相关的部分结构的框图，并不构成对本申请方案所应用于其上的计算机设备 500 的限定，具体的计算机设备 500 可以包括比图中所示更多或更少的部件，或者组合某些部件，或者具有不同的部件布置。

其中，所述处理器 502 用于运行存储在存储器中的计算机程序 5032，以实现如下功能：将待打标签文本进行分词预处理，得到预处理文本；将预处理文本输入词语逆频率 TF-IDF 算法模型，得到待打标签文本的关键词集；根据待打标签文本的关键词集得到初始化转移矩阵，由初始化转移矩阵及初始的关键词概率分布经过多次迭代相乘运算直至收敛后，得到关键词最终概率分布；获取关键词最终概率分布中概率最大值的对应行，获取概率最大值的对应行所对应关键词，并将所述关键词设置为待打标签文本的标签。

在一实施例中，处理器 502 还执行如下操作：对待打标签文本进行分词，

得到分词后文本；对分词后文本包括的分词一一设置加权值；删除分词后文本中的停用词，并统计各分词的词频，得到第一三元组；获取第一三元组中词频大于预设词频阈值所对分词之间的词语相似度；若分词之间的词语相似度大于预设词语相似度阈值，保留其中任意一个分词，得到第二三元组，并将第二三元组作为预处理文本。

在一实施例中，处理器 502 还执行如下操作：生成语料库词语统计结果集；获取预处理文本；将预处理文本及语料库词语统计结果集输入词语逆频率 TF-IDF 算法模型，得到待打标签文本的关键词集；其中，所述词语逆频率 TF-IDF 算法模型为：

$$\text{TF-IWF}_{i,j} \rightarrow \text{TF}_{i,j} \times \text{IWF}_i = \frac{n_{i,j}}{\sum_k n_{i,j}} \times \log \frac{\sum_{i=1}^m n_{t_i}}{n_{t_i}}$$

其中，TF 部分分子 $n_{i,j}$ 表示词语 t_i 在文本 j 中出现的次数，分母表示文本 j 中所有的词语频词和，IWF 部分分子表示语料库中所有词语频数之和， n_{t_i} 表示词语 t_i 在语料库中出现的总频数。

在一实施例中，所述根据待打标签文本的关键词集得到初始化转移矩阵中，所述初始化转移矩阵为 n 维方阵， n 维方阵的维数与关键词集中关键词总个数相等；所述初始的关键词概率分布为每一行值均为 $1/n$ 的 n 维列向量；其中， n 为与关键词集中关键词总个数相等的正整数；

所述由初始化转移矩阵及初始的关键词概率分布经过多次迭代相乘运算记为 $V_m = MV_{m-1}$ ，其中， m 为正整数， V_0 为初始的关键词概率分布， M 为初始化转移矩阵。

在一实施例中，处理器 502 还执行如下操作：爬取待打标签文本，并存储至 MongoDB 数据库中。

本领域技术人员可以理解，图 7 中示出的计算机设备的实施例并不构成对计算机设备具体构成的限定，在其他实施例中，计算机设备可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件布置。例如，在一些实施例中，计算机设备可以仅包括存储器及处理器，在这样的实施例中，存储器及处理器的结构及功能与图 7 所示实施例一致，在此不再赘述。

应当理解，在本申请实施例中，处理器 502 可以是中央处理单元 (Central Processing Unit, CPU)，该处理器 502 还可以是其他通用处理器、数字信号处理

器 (Digital Signal Processor, DSP)、专用集成电路 (Application Specific Integrated Circuit, ASIC)、现成可编程门阵列 (Field-Programmable Gate Array, FPGA) 或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。其中, 通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

在本申请的另一实施例中提供一种存储介质。该存储介质可以为非易失性的计算机可读存储介质。该存储介质存储有计算机程序, 其中计算机程序包括程序指令。该程序指令被处理器执行时实现本申请实施例的自动打标签的方法。

所述存储介质可以是前述设备的内部存储单元, 例如设备的硬盘或内存。所述存储介质也可以是所述设备的外部存储设备, 例如所述设备上配备的插接式硬盘, 智能存储卡 (Smart Media Card, SMC), 安全数字 (Secure Digital, SD) 卡, 闪存卡 (Flash Card) 等。进一步地, 所述存储介质还可以既包括所述设备的内部存储单元也包括外部存储设备。

所属领域的技术人员可以清楚地了解到, 为了描述的方便和简洁, 上述描述的设备、装置和单元的具体工作过程, 可以参考前述方法实施例中的对应过程, 在此不再赘述。

以上所述, 仅为本申请的具体实施方式, 但本申请的保护范围并不局限于此, 任何熟悉本技术领域的技术人员在本申请揭露的技术范围内, 可轻易想到各种等效的修改或替换, 这些修改或替换都应涵盖在本申请的保护范围之内。因此, 本申请的保护范围应以权利要求的保护范围为准。

权 利 要 求 书

1. 一种自动打标签的方法，其特征在于，包括：

将待打标签文本进行分词预处理，得到预处理文本；

将预处理文本输入词语逆频率 TF-IDF 算法模型，得到待打标签文本的关键词集；

根据待打标签文本的关键词集得到初始化转移矩阵，由初始化转移矩阵及初始的关键词概率分布经过多次迭代相乘运算直至收敛后，得到关键词最终概率分布；

获取关键词最终概率分布中概率最大值的对应行，获取概率最大值的对应行所对应关键词，并将所述关键词设置为待打标签文本的标签。

2. 根据权利要求 1 所述的自动打标签的方法，其特征在于，所述将待打标签文本进行分词预处理，得到预处理文本，包括：

对待打标签文本进行分词，得到分词后文本；

对分词后文本包括的分词一一设置加权重值；

删除分词后文本中的停用词，并统计各分词的词频，得到第一三元组；

获取第一三元组中词频大于预设词频阈值所对分词之间的词语相似度；

若分词之间的词语相似度大于预设词语相似度阈值，保留其中任意一个分词，得到第二三元组，并将第二三元组作为预处理文本。

3. 根据权利要求 2 所述的自动打标签的方法，其特征在于，所述词语逆频率 TF-IDF 算法模型为：

$$TF-IWF_{i,j} \rightarrow TF_{i,j} \times IWF_i = \frac{n_{i,j}}{\sum_k n_{i,j}} \times \log \frac{\sum_{i=1}^m nt_i}{nt_i}$$

其中，TF 部分分子 $n_{i,j}$ 表示词语 t_i 在文本 j 中出现的次数，分母表示文本 j 中所有的词语频词和，IWF 部分分子表示语料库中所有词语频数之和， nt_i 表示词语 t_i 在语料库中出现的总频数；

所述将预处理文本输入词语逆频率 TF-IDF 算法模型，得到待打标签文本的关键词集，包括：

生成语料库词语统计结果集；

获取预处理文本；

将预处理文本及语料库词语统计结果集输入词语逆频率 TF-IDF 算法模型，得到待打标签文本的关键词集。

4. 根据权利要求 1 所述的自动打标签的方法，其特征在于，所述根据待打标签文本的关键词集得到初始化转移矩阵中，所述初始化转移矩阵为 n 维方阵， n 维方阵的维数与关键词集中关键词总个数相等；所述初始的关键词概率分布为每一行值均为 $1/n$ 的 n 维列向量；其中， n 为与关键词集中关键词总个数相等的正整数；

所述由初始化转移矩阵及初始的关键词概率分布经过多次迭代相乘运算记为 $V_m = MV_{m-1}$ ，其中， m 为正整数， V_0 为初始的关键词概率分布， M 为初始化转移矩阵。

5. 根据权利要求 4 所述的自动打标签的方法，其特征在于，所述将待打标签文本进行分词预处理，得到预处理文本，之前包括：

爬取待打标签文本，并存储至 MongoDB 数据库中。

6. 一种自动打标签的装置，其特征在于，包括：

文本预处理单元，用于将待打标签文本进行分词预处理，得到预处理文本；

关键词集获取单元，用于将预处理文本输入词语逆频率 TF-IDF 算法模型，得到待打标签文本的关键词集；

最终概率分布获取单元，用于根据待打标签文本的关键词集得到初始化转移矩阵，由初始化转移矩阵及初始的关键词概率分布经过多次迭代相乘运算直至收敛后，得到关键词最终概率分布；

打标单元，用于获取关键词最终概率分布中概率最大值的对应行，获取概率最大值的对应行所对应关键词，并将所述关键词设置为待打标签文本的标签。

7. 根据权利要求 6 所述的自动打标签的装置，其特征在于，所述文本预处理单元，包括：

分词单元，用于对待打标签文本进行分词，得到分词后文本；

加权单元，用于对分词后文本包括的分词一一设置加权值；

统计单元，用于删除分词后文本中的停用词，并统计各分词的词频，得到第一三元组；

相似度获取单元，用于获取第一三元组中词频大于预设词频阈值所对分词之间的词语相似度；

删词单元，用于若分词之间的词语相似度大于预设词语相似度阈值，保留其中任意一个分词，得到第二三元组，并将第二三元组作为预处理文本。

8. 根据权利要求 7 所述的自动打标签的装置，其特征在于，所述词语逆频率 TF-IDF 算法模型为：

$$TF\text{-}IWF_{i,j} \rightarrow TF_{i,j} \times IWF_i = \frac{n_{i,j}}{\sum_k n_{k,j}} \times \log \frac{\sum_{i=1}^m nt_i}{nt_i}$$

其中，TF 部分分子 $n_{i,j}$ 表示词语 t_i 在文本 j 中出现的次数，分母表示文本 j 中所有的词语频词和，IWF 部分分子表示语料库中所有词语频数之和， nt_i 表示词语 t_i 在语料库中出现的总频数；

所述关键词集获取单元，包括：

第一处理单元，用于生成语料库词语统计结果集；

第二处理单元，用于获取预处理文本；

关键词集计算单元，用于将预处理文本及语料库词语统计结果集输入词语逆频率 TF-IDF 算法模型，得到待打标签文本的关键词集。

9. 根据权利要求 6 所述的自动打标签的装置，其特征在于，所述根据待打标签文本的关键词集得到初始化转移矩阵中，所述初始化转移矩阵为 n 维方阵， n 维方阵的维数与关键词集中关键词总个数相等；所述初始的关键词概率分布为每一行值均为 $1/n$ 的 n 维列向量；其中， n 为与关键词集中关键词总个数相等的正整数；

所述由初始化转移矩阵及初始的关键词概率分布经过多次迭代相乘运算记为 $V_m = MV_{m-1}$ ，其中， m 为正整数， V_0 为初始的关键词概率分布， M 为初始化转移矩阵。

10. 根据权利要求 9 所述的自动打标签的装置，其特征在于，还包括：

爬取单元，用于爬取待打标签文本，并存储至 MongoDB 数据库中。即原始数据从网上爬取，得到待打标签文本，存放到 MongoDB 数据库。通过爬取数据，可设置一筛选条件，即爬取未设置标签的文本从而进行打标签。

11. 一种计算机设备，包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，其特征在于，所述处理器执行所述计算机程序时实现以下步骤：

将待打标签文本进行分词预处理，得到预处理文本；

将预处理文本输入词语逆频率 TF-IDF 算法模型, 得到待打标签文本的关键词集;

根据待打标签文本的关键词集得到初始化转移矩阵, 由初始化转移矩阵及初始的关键词概率分布经过多次迭代相乘运算直至收敛后, 得到关键词最终概率分布;

获取关键词最终概率分布中概率最大值的对应行, 获取概率最大值的对应行所对应关键词, 并将所述关键词设置为待打标签文本的标签。

12. 根据权利要求 11 所述的计算机设备, 其特征在于, 所述将待打标签文本进行分词预处理, 得到预处理文本, 包括:

对待打标签文本进行分词, 得到分词后文本;

对分词后文本包括的分词一一设置加权值;

删除分词后文本中的停用词, 并统计各分词的词频, 得到第一三元组;

获取第一三元组中词频大于预设词频阈值所对分词之间的词语相似度;

若分词之间的词语相似度大于预设词语相似度阈值, 保留其中任意一个分词, 得到第二三元组, 并将第二三元组作为预处理文本。

13. 根据权利要求 12 所述的计算机设备, 其特征在于, 所述词语逆频率 TF-IDF 算法模型为:

$$TF\text{-}IWF_{i,j} \rightarrow TF_{i,j} \times IWF_i = \frac{n_{i,j}}{\sum_k n_{i,j}} \times \log \frac{\sum_{i=1}^m nt_i}{nt_i}$$

其中, TF 部分分子 $n_{i,j}$ 表示词语 t_i 在文本 j 中出现的次数, 分母表示文本 j 中所有的词语频词和, IWF 部分分子表示语料库中所有词语频数之和, nt_i 表示词语 t_i 在语料库中出现的总频数;

所述将预处理文本输入词语逆频率 TF-IDF 算法模型, 得到待打标签文本的关键词集, 包括:

生成语料库词语统计结果集;

获取预处理文本;

将预处理文本及语料库词语统计结果集输入词语逆频率 TF-IDF 算法模型, 得到待打标签文本的关键词集。

14. 根据权利要求 11 所述的计算机设备, 其特征在于, 所述根据待打标签文本的关键词集得到初始化转移矩阵中, 所述初始化转移矩阵为 n 维方阵, n 维

方阵的维数与关键词集中关键词总个数相等；所述初始的关键词概率分布为每一行值均为 $1/n$ 的 n 维列向量；其中， n 为与关键词集中关键词总个数相等的正整数；

所述由初始化转移矩阵及初始的关键词概率分布经过多次迭代相乘运算记为 $V_m = MV_{m-1}$ ，其中， m 为正整数， V_0 为初始的关键词概率分布， M 为初始化转移矩阵。

15. 根据权利要求 11 所述的计算机设备，其特征在于，所述将待打标签文本进行分词预处理，得到预处理文本，之前包括：

爬取待打标签文本，并存储至 MongoDB 数据库中。

16. 一种存储介质，其特征在于，所述存储介质存储有计算机程序，所述计算机程序包括程序指令，所述程序指令当被处理器执行时使所述处理器执行以下操作：

将待打标签文本进行分词预处理，得到预处理文本；

将预处理文本输入词语逆频率 TF-IDF 算法模型，得到待打标签文本的关键词集；

根据待打标签文本的关键词集得到初始化转移矩阵，由初始化转移矩阵及初始的关键词概率分布经过多次迭代相乘运算直至收敛后，得到关键词最终概率分布；

获取关键词最终概率分布中概率最大值的对应行，获取概率最大值的对应行所对应关键词，并将所述关键词设置为待打标签文本的标签。

17. 根据权利要求 16 所述的存储介质，其特征在于，所述将待打标签文本进行分词预处理，得到预处理文本，包括：

对待打标签文本进行分词，得到分词后文本；

对分词后文本包括的分词一一设置加权值；

删除分词后文本中的停用词，并统计各分词的词频，得到第一三元组；

获取第一三元组中词频大于预设词频阈值所对分词之间的词语相似度；

若分词之间的词语相似度大于预设词语相似度阈值，保留其中任意一个分词，得到第二三元组，并将第二三元组作为预处理文本。

18. 根据权利要求 17 所述的存储介质，其特征在于，所述词语逆频率 TF-IDF 算法模型为：

$$\text{TF-IWF}_{i,j} \rightarrow \text{TF}_{i,j} \times \text{IWF}_i = \frac{n_{i,j}}{\sum_k n_{i,j}} \times \log \frac{\sum_{i=1}^m n_{t_i}}{n_{t_i}}$$

其中，TF 部分分子 $n_{i,j}$ 表示词语 t_i 在文本 j 中出现的次数，分母表示文本 j 中所有的词语频词和，IWF 部分分子表示语料库中所有词语频数之和， n_{t_i} 表示词语 t_i 在语料库中出现的总频数；

所述将预处理文本输入词语逆频率 TF-IDF 算法模型，得到待打标签文本的关键词集，包括：

生成语料库词语统计结果集；

获取预处理文本；

将预处理文本及语料库词语统计结果集输入词语逆频率 TF-IDF 算法模型，得到待打标签文本的关键词集。

19. 根据权利要求 16 所述的存储介质，其特征在于，所述根据待打标签文本的关键词集得到初始化转移矩阵中，所述初始化转移矩阵为 n 维方阵， n 维方阵的维数与关键词集中关键词总个数相等；所述初始的关键词概率分布为每一行值均为 $1/n$ 的 n 维列向量；其中， n 为与关键词集中关键词总个数相等的正整数；

所述由初始化转移矩阵及初始的关键词概率分布经过多次迭代相乘运算记为 $V_m = MV_{m-1}$ ，其中， m 为正整数， V_0 为初始的关键词概率分布， M 为初始化转移矩阵。

20. 根据权利要求 16 所述的存储介质，其特征在于，所述将待打标签文本进行分词预处理，得到预处理文本，之前包括：

爬取待打标签文本，并存储至 MongoDB 数据库中。

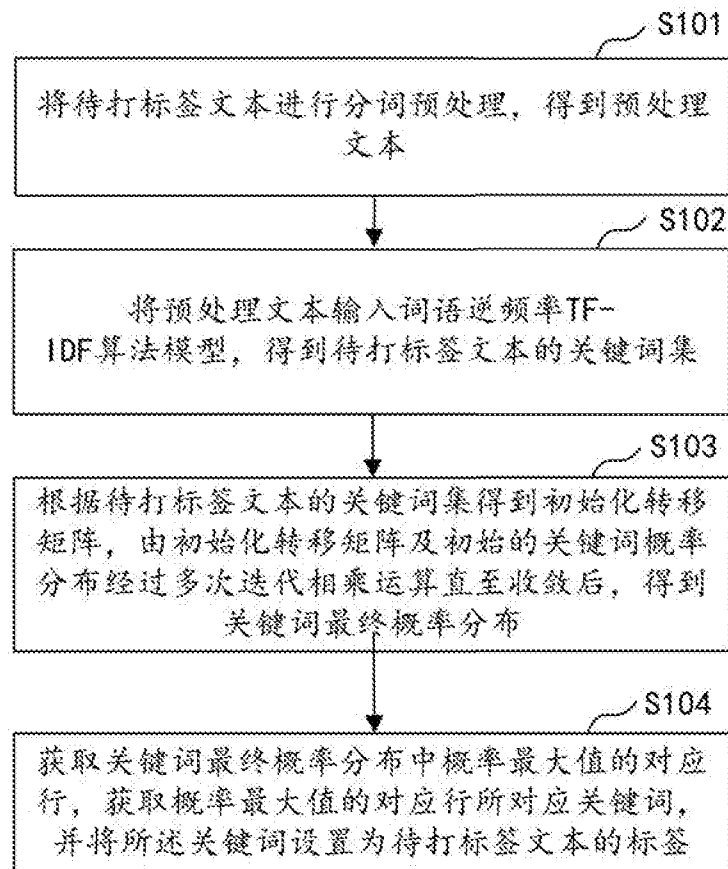


图 1

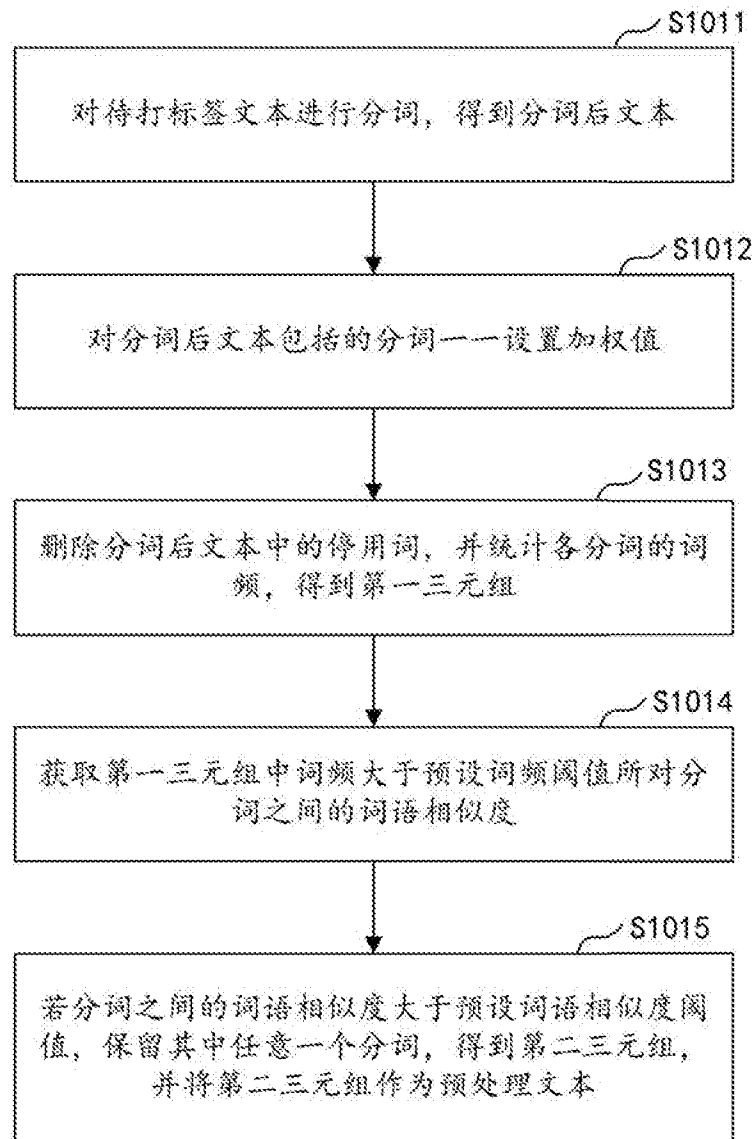


图 2

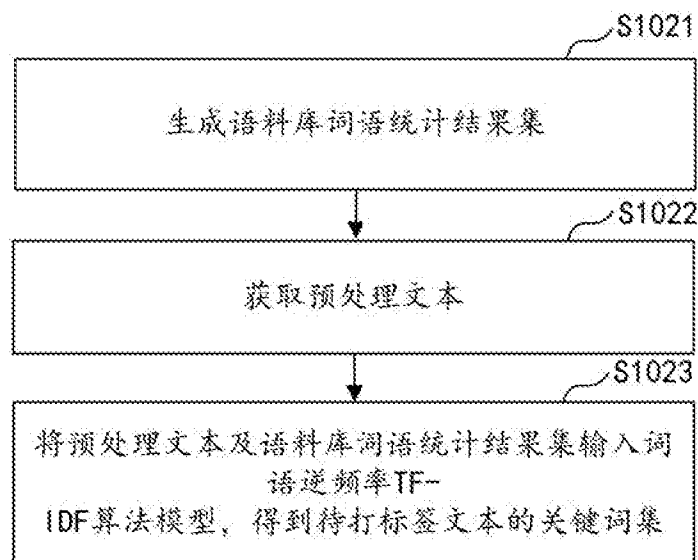


图 3

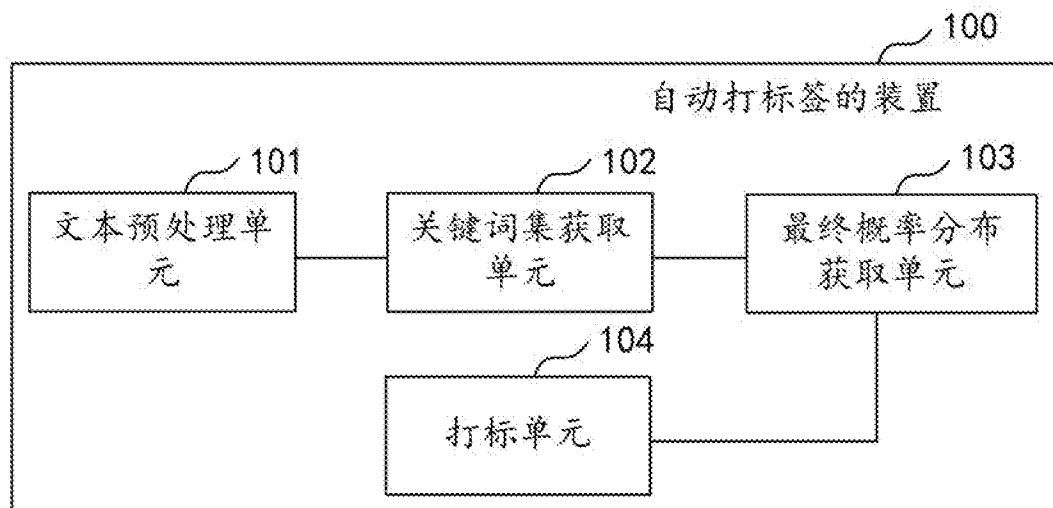


图 4

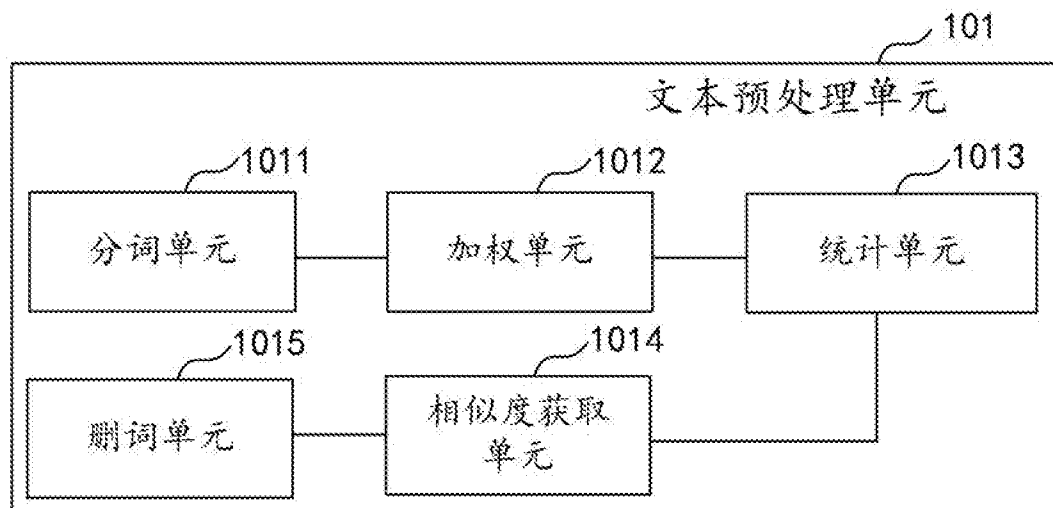


图 5

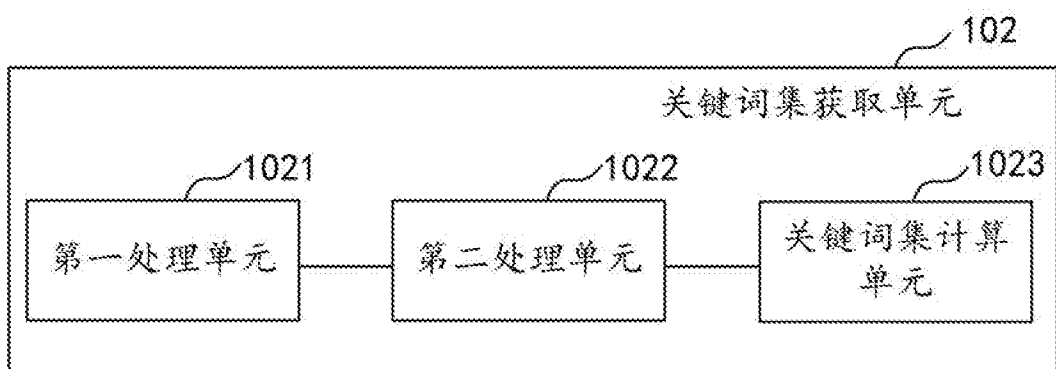


图 6

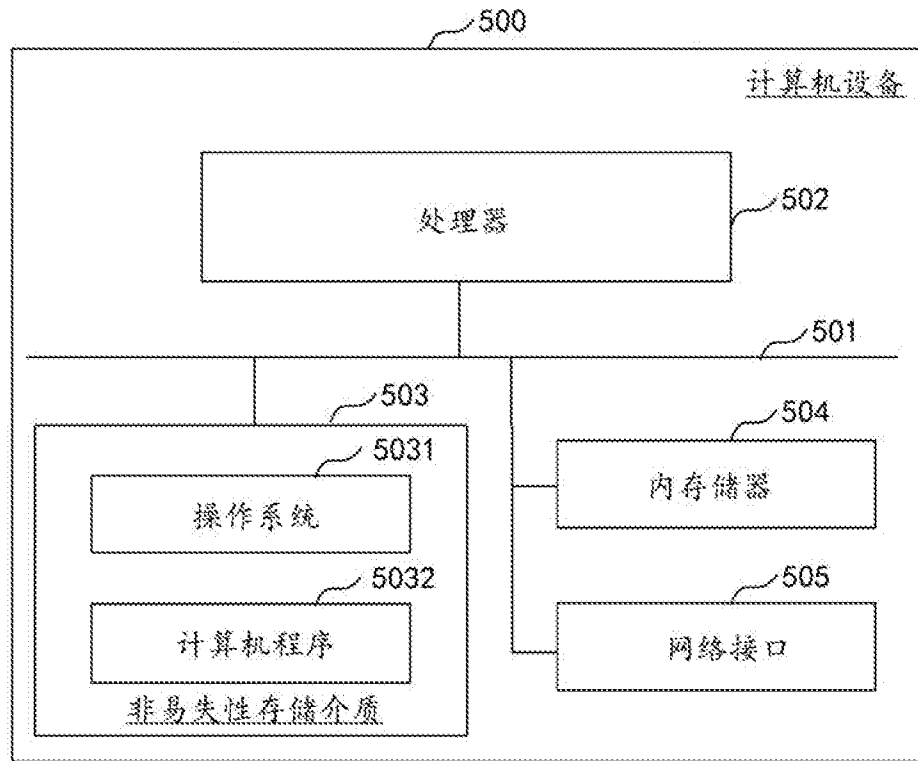


图 7

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2018/085348

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30(2006.01)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)

CNKI, CNPAT, WPI, EPODOC, GOOGLE, IEEE: 标签, 打标, 分词, 转移, 矩阵, 概率, 关键词, 迭代, 收敛, tag, keyword?, TF-IDF, cluster?, tab, label, probability, matrix

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 107301199 A (BEIJING FUSIO TECHNOLOGY CO., LTD.) 27 October 2017 (2017-10-27) claims 1 and 4, and figures 1 and 2	1-20
A	CN 104077417 A (INSTITUTE OF COMPUTING TECHNOLOGY CHINESE ACADEMY OF SCIENCES) 01 October 2014 (2014-10-01) entire document	1-20
A	CN 103778215 A (BEIJING INSTITUTE OF TECHNOLOGY) 07 May 2014 (2014-05-07) entire document	1-20
A	US 5640488 A (PANASONIC TECHNOLOGIES, INC.) 17 June 1997 (1997-06-17) 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

11 September 2018

Date of mailing of the international search report

25 September 2018

Name and mailing address of the ISA/CN

State Intellectual Property Office of the P. R. China (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2018/085348

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	107301199	A	27 October 2017	None	
CN	104077417	A	01 October 2014	None	
CN	103778215	A	07 May 2014	None	
US	5640488	A	17 June 1997	None	

国际检索报告

国际申请号

PCT/CN2018/085348

A. 主题的分类 G06F 17/30 (2006.01) 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)) CNKI, CNPAT, WPI, EPODOC, GOOGLE, IEEE: 标签, 打标, 分词, 转移, 矩阵, 概率, 关键词, 迭代, 收敛, tag, keyword?, TF-IDF, cluster?, tab, label, probability, matrix																	
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>A</td> <td>CN 107301199 A (北京融数云途科技有限公司) 2017年 10月 27日 (2017 - 10 - 27) 权利要求1、4, 附图1、2</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>CN 104077417 A (中国科学院计算技术研究所) 2014年 10月 1日 (2014 - 10 - 01) 全文</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>CN 103778215 A (北京理工大学) 2014年 5月 7日 (2014 - 05 - 07) 全文</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>US 5640488 A (PANASONIC TECHNOLOGIES, INC.) 1997年 6月 17日 (1997 - 06 - 17) 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	A	CN 107301199 A (北京融数云途科技有限公司) 2017年 10月 27日 (2017 - 10 - 27) 权利要求1、4, 附图1、2	1-20	A	CN 104077417 A (中国科学院计算技术研究所) 2014年 10月 1日 (2014 - 10 - 01) 全文	1-20	A	CN 103778215 A (北京理工大学) 2014年 5月 7日 (2014 - 05 - 07) 全文	1-20	A	US 5640488 A (PANASONIC TECHNOLOGIES, INC.) 1997年 6月 17日 (1997 - 06 - 17) 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															
A	CN 107301199 A (北京融数云途科技有限公司) 2017年 10月 27日 (2017 - 10 - 27) 权利要求1、4, 附图1、2	1-20															
A	CN 104077417 A (中国科学院计算技术研究所) 2014年 10月 1日 (2014 - 10 - 01) 全文	1-20															
A	CN 103778215 A (北京理工大学) 2014年 5月 7日 (2014 - 05 - 07) 全文	1-20															
A	US 5640488 A (PANASONIC TECHNOLOGIES, INC.) 1997年 6月 17日 (1997 - 06 - 17) 全文	1-20															
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																	
* 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 2018年 9月 11日		国际检索报告邮寄日期 2018年 9月 25日															
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局 (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		授权官员 吴瑶 电话号码 86-10-53961339															

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2018/085348

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	107301199	A	2017年 10月 27日	无	
CN	104077417	A	2014年 10月 1日	无	
CN	103778215	A	2014年 5月 7日	无	
US	5640488	A	1997年 6月 17日	无	