

(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
2017 年 6 月 29 日 (29.06.2017)



(10) 国际公布号
WO 2017/107566 A1

- (51) 国际专利分类号:
G06F 17/30 (2006.01)
- (21) 国际申请号: PCT/CN2016/098234
- (22) 国际申请日: 2016 年 9 月 6 日 (06.09.2016)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201511003865.4 2015 年 12 月 25 日 (25.12.2015) CN
- (71) 申请人: 广州视源电子科技股份有限公司 (GUANGZHOU SHIYUAN ELECTRONICS CO., LTD.) [CN/CN]; 中国广东省广州市黄埔区云埔四路 6 号, Guangdong 510530 (CN)。
- (72) 发明人: 李贤 (LI, Xian); 中国广东省广州市黄埔区云埔四路 6 号, Guangdong 510530 (CN)。
- (74) 代理人: 广州华进联合专利商标代理有限公司 (ADVANCE CHINA IP LAW OFFICE); 中国广东省广州市天河区花城大道 85 号 3901 房, Guangdong 510623 (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, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, 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, 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))。

(54) Title: RETRIEVAL METHOD AND SYSTEM BASED ON WORD VECTOR SIMILARITY

(54) 发明名称: 基于词向量相似度的检索方法和系统

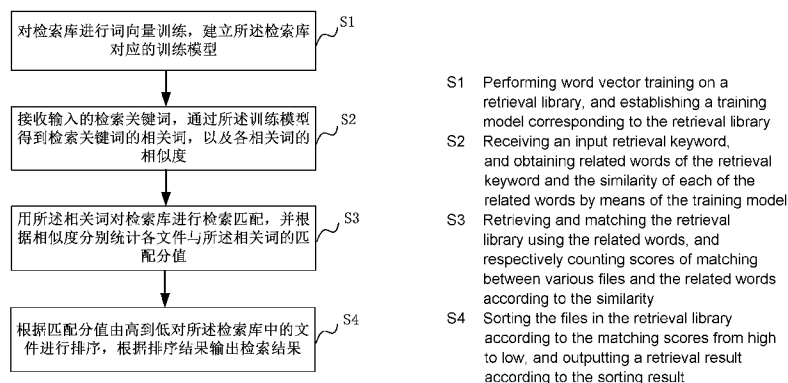


图 1

(57) Abstract: A retrieval method and system based on a word vector similarity. The method comprises: performing word vector training on a retrieval library, and establishing a training model corresponding to the retrieval library (S1); receiving an input retrieval keyword, and obtaining related words of the retrieval keyword and the similarity between each of the related words and the retrieval keyword by means of the training model (S2); retrieving and matching the retrieval library using the related words, and respectively counting scores of matching between various files in the retrieval library and the related words according to the similarity (S3); and sorting the files in the retrieval library according to the matching scores from high to low, and outputting a retrieval result according to the sorting result (S4). By means of the method, the capabilities of retrieving and matching related words can be enhanced in combination with the lexical characteristics in various retrieval libraries, thereby improving the accuracy rate and the robustness of retrieval.

(57) 摘要:

[见续页]



WO 2017/107566 A1

一种基于词向量相似度的检索方法和系统。所述方法包括：对检索库进行词向量训练，建立所述检索库对应的训练模型（S1）；接收输入的检索关键词，通过所述训练模型得到所述检索关键词的相关词，以及各相关词与所述检索关键词的相似度（S2）；用所述相关词对所述检索库进行检索匹配，并根据所述相似度分别统计所述检索库中各文件与所述相关词的匹配分值（S3）；根据所述匹配分值由高到低对所述检索库中的文件进行排序，根据排序结果输出检索结果（S4）。通过本方法，能够结合各检索库中的用词特点，增强对相关词的检索匹配能力，从而提高检索准确率和鲁棒性。

基于词向量相似度的检索方法和系统

技术领域

本发明涉及信息检索技术领域，特别是涉及基于词向量相似度的检索方法
5 和基于词向量相似度的检索系统。

背景技术

现有的对简历搜索匹配过程的技术，通常是通过多个关键词进行检索。通过用户提供一组关键词在检索库中进行检索，以匹配词命中的数量作为匹配分
10 值，根据匹配分值由高到低的排列输出检索结果，默认排在前的结果更符合用户要求。然而，这种检索方式存在以下缺点：

（1）没能考虑到不同检索库的用词特点，例如英文的大小写，字符的全角半角等；

（2）不能考虑到词与词之间的关系，导致检索过程中，对与关键词存在
15 很强联系的其它词缺乏信息匹配能力；例如关键词设为“程序”，却无法对检索库中“软件”的信息进行检索匹配；

（3）对关键词选取的要求高，检索鲁棒性差；如果关键词遗漏或者输错，对最终检索结果会产生很大影响。

综上所述，现有的基于关键词的检索方法，其检索召回率和检索结果准确
20 率都不够理想，同时存在鲁棒性和适应性较差的问题。

发明内容

基于此，本发明提供一种基于词向量相似度的检索方法和系统，能够提高检索准确率和鲁棒性。

本发明一方面提供一种基于词向量相似度的检索方法，包括：

对检索库进行词向量训练，建立所述检索库对应的训练模型；

接收输入的检索关键词，通过所述训练模型得到所述检索关键词的相关词，以及各相关词与所述检索关键词的相似度；

用所述相关词对所述检索库进行检索匹配，并根据所述相似度分别统计所述检索库中各文件与所述相关词的匹配分值；

根据所述匹配分值由高到低对所述检索库中的文件进行排序，根据排序结果输出检索结果。

优选的，所述对检索库进行词向量训练，之前包括：

对检索库中各文件分别进行预处理，将各文件预处理后的数据存储到一对应的训练样本文件中；所述预处理包括数据清洗和提取数据描述；

所述对检索库进行词向量训练包括：

基于所述训练样本文件对所述检索库进行词向量训练。

优选的，所述数据清洗包括统一大小写、消除多余空格、统一标点符号、统一全半角格式中至少一种；

所述提取数据描述包括通过添加用户词典进行分词。

优选的，所述对检索库进行词向量训练包括：

通过 word2vec 对所述训练样本文件进行词向量训练。

优选的，用所述相关词对所述检索库进行检索匹配，并根据所述相似度分别统计所述检索库中各文件与所述相关词的匹配分值，包括：

用所述相关词分别对所述检索库中各文件进行检索匹配，得到各文件与所述相关词的匹配结果；

将各相关词对应的相似度作为累加权值，结合所述匹配结果分别得出各文件与所述相关词的匹配分值。

本发明另一方面提供一种基于词向量相似度的检索系统，包括：

模型训练单元，用于对检索库进行词向量训练，建立所述检索库对应的训练模型；

生成相关词单元，用于接收输入的检索关键词，通过所述训练模型得到所述检索关键词的相关词，以及各相关词与所述检索关键词的相似度；

检索匹配单元，用于用所述相关词对所述检索库进行检索匹配，并根据所述相似度分别统计所述检索库中各文件与所述相关词的匹配分值；

结果输出单元，用于根据所述匹配分值由高到低对所述检索库中的文件进行排序，根据排序结果输出检索结果。

10 优选的，所述模型训练单元，还用于对检索库进行词向量训练之前，对检索库中各文件分别进行预处理，将各文件预处理后的数据存储在对应的训练样本文件中；所述预处理包括数据清洗和提取数据描述；

所述对检索库进行词向量训练包括：

基于所述训练样本文件对所述检索库进行词向量训练。

15 优选的，所述数据清洗包括统一大小写、消除多余空格、统一标点符号、统一全半角格式中至少一种；

所述提取数据描述包括通过添加用户词典进行分词。

优选的，所述对检索库进行词向量训练包括：

通过 word2vec 对所述训练样本文件进行词向量训练。

20 优选的，所述检索匹配单元包括：

匹配模块，用于用所述相关词分别对所述检索库中各文件进行检索匹配，得到各文件与所述相关词的匹配结果；

统计模块，用于将各相关词对应的相似度作为累加权值，结合所述匹配结

果分别得出各文件与所述相关词的匹配分值。

上述技术方案的基于词向量相似度的检索方法和系统，通过对检索库进行词向量训练，建立所述检索库对应的训练模型；接收输入的检索关键词，通过所述训练模型得到所述检索关键词的相关词，以及各相关词与所述检索关键词的相似度；用所述相关词对所述检索库进行检索匹配，并根据所述相似度分别统计所述检索库中各文件与所述相关词的匹配分值；根据所述匹配分值由高到低对所述检索库中的文件进行排序，根据排序结果输出检索结果。首先由于训练模型是基于检索库训练得到的，因此能很好反映检索库的用词特点，有利于提高检索准确率；其次将关键词以词向量的形式进行表示，检索时根据关键词的相关词进行检索匹配，增加了对相关词的检索匹配能力，从而提高了检索鲁棒性。

附图说明

图 1 为本发明实施例的基于词向量相似度的检索方法的示意性流程图；

图 2 为本发明实施例的基于词向量相似度的检索系统的示意性结构图。

具体实施方式

为了使本发明的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本发明进行进一步详细说明。应当理解，此处所描述的具体实施例仅仅用以解释本发明，并不用于限定本发明。

本发明提供的实施例包括基于词向量相似度的检索方法实施例，还包括相应的基于词向量相似度的检索系统实施例。以下分别进行详细说明。

图 1 为本发明实施例的基于词向量相似度的检索方法的示意性流程图；如

图 1 所示，本实施例的基于词向量相似度的检索方法包括如下步骤 S1 至 S4，各步骤详述如下：

S1，对检索库进行词向量训练，建立所述检索库对应的训练模型；

自然语言理解的问题要转化为机器学习的问题，第一步需要找一种方法把这些符号数学化，例如把每个词都表示为一个特有的向量。词向量是“Word Representation”或“Word Embedding”的中文俗称。

本实施例中的词向量应当具有的特点包括：让相关或者相似的词，在距离上更接近，例如“麦克”和“话筒”的距离会远小于“麦克”和“天气”的距离。向量的距离可以用传统的欧氏距离来衡量，也可以用 $\cos$ 夹角来衡量。

优选的，所述词向量可为用 Distributed Representation 表示的词向量。Distributed Representation 表示的词向量为一种低维实数向量，这种向量一般形式为 $[0.792, -0.177, -0.107, 0.109, -0.542, \dots]$ ，维度以 50 维和 100 维比较常见。

作为一优选实施方式，在对检索库进行词向量训练之前，还可对检索库中各文件分别进行预处理，将各文件预处理后的数据存储到一对应的训练样本文件中。

优选的，其中所述预处理包括数据清洗和提取数据描述。其中数据清洗主要用于实现检索库中数据的一致性，具体可包括统一大小写、消除多余空格、统一标点符号、统一全半角格式中至少一种；所述提取数据描述包括通过添加用户词典进行分词，具体可为添加用户词典并通过 NLPIR（又名 ICTCLAS2013，汉语分词系统）进行分词。

进一步的，基于所述训练样本文件对所述检索库进行词向量训练，以建立所述检索库对应的训练模型。具体方式可为：通过 word2vec 对所述训练样本

文件进行词向量训练，训练设置如下：

```
./word2vec -train result_cropus.txt -output vectors.bin -cbow 0 -size 50  
-window 5 -negative 0 -hs 1 -sample 1e-3 -threads 4 -binary 1 -min_count 3 ;
```

其中，各参数的含义为：

- 5 -train 后面表示参与训练的训练样本文件名，
- cbow 表示采用跳空词袋模型，
- size 表示词向量采用的维度，
- window 表示上下文窗口长度，
- negative 表示是否采用负采样方法，0 表示不使用，1 表示使用，
- 10 -hs 表示是否使用 HS 方法，0 表示不使用，1 表示使用，
- sample 1e-3 表示采样的阈值为 10^{-3} ，如果一个词在训练样本中出现的
 频率越大，那么就越会被采样；
- thread 表示开启线程数，
- binary 表示输出是否为二进制文件，0 表示不使用，1 表示使用，
- 15 -min_count 表示设置的最低频率，默认为 5，如果一个词语在文档中
 出现的次数小于该阈值，那么该词就会被舍弃。

至此，可得到与所述检索库对应的训练模型。

S2，接收输入的检索关键词，通过所述训练模型得到所述检索关键词的相关词，以及各相关词与所述检索关键词的相似度；

- 20 本实施例中，两个词向量的相似度指的是余弦相似度，最高可为 1，最低
 可为 0。由于训练模型是基于检索库训练得到的，因此基于该训练模型得到的
 相关词能很好反映检索库的用词特点。具体可通过 `./distance vectors.bin` 命令产
 生相关词及相似度，并通过 `sh` 脚本和 `expect` 脚本自动生成。

例如：需要在简历库中检索 C++软件开发工程师，输入关键词为 C++、软件、MFC、数据结构，基于该简历库的训练模型可以得到以下相关词词表和相似度，详见下表：

关键词	相关词及相似度	
C++	C/C++	0.825696
	JAVA	0.663165
	IDL	0.645119
	PYTHON	0.603219
软件	LABVIEW	0.588997
	CCS3.3	0.588270
	程序	0.585543
	人机界面	0.581459
MFC	QT	0.851808
	LABVIEW	0.750083
	DELPHI	0.736733
	VS2010	0.700681
数据结构	导论	0.632002
	素养	0.524529
	PRIMER	0.509389
	建模	0.505594

S3，用所述相关词对所述检索库进行检索匹配，并根据所述相似度分别统计所述检索库中各文件与所述相关词的匹配分值；

本实施例中，用上述步骤得出的相关词分别对所述检索库中各文件进行检索匹配，得到各文件与所述相关词的匹配结果；将各相关词对应的相似度作为累加权值，结合所述匹配结果分别可得出各文件与所述相关词的匹配分值。

S4，根据所述匹配分值由高到低对所述检索库中的文件进行排序，根据排序结果输出检索结果。

优选的，可设定分值门限，仅对匹配分值高于所述分值门限的检索结果进行排序，并按照匹配分值由高到低的排序输出。通过设定分值门限对检索结果进一步筛选，有利于用户查阅检索结果。

通过上述实施例的基于词向量相似度的检索方法，通过对检索库进行词向

量训练，建立所述检索库对应的训练模型；接收输入的检索关键词，通过所述训练模型得到所述检索关键词的相关词，以及各相关词与所述检索关键词的相似度；用所述相关词对所述检索库进行检索匹配，并根据所述相似度分别统计所述检索库中各文件与所述相关词的匹配分值；根据所述匹配分值由高到低对所述检索库中的文件进行排序，根据排序结果输出检索结果。首先由于训练模型是基于检索库训练得到的，因此能很好反映检索库的用词特点，有利于提高检索准确率；其次将关键词以词向量的形式进行表示，检索时根据关键词的相关词进行检索匹配，增加了对相关词的检索匹配能力，从而提高了检索鲁棒性。

需要说明的是，对于前述的方法实施例，为了简便描述，将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本发明并不受所描述的动作顺序的限制，因为依据本发明，某些步骤可以采用其它顺序或者同时进行。

以下对可用于执行上述基于词向量相似度的检索方法的基于词向量相似度的检索系统实施例进行说明。为了便于说明，基于词向量相似度的检索系统实施例的结构示意图中，仅仅示出了与本发明实施例相关的部分，本领域技术人员可以理解，图中示出的系统结构并不构成对系统的限定，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。

图 2 为本发明实施例的基于词向量相似度的检索系统的示意性结构图；如图 2 所示，本实施例的基于词向量相似度的检索系统包括：模型训练单元 210、生成相关词单元 220、检索匹配单元 230 以及结果输出单元 240，各单元详述如下：

所述模型训练单元 210，用于对检索库进行词向量训练，建立所述检索库对应的训练模型；

本实施例中的词向量应当具有的特点包括：让相关或者相似的词，在距离

上更接近，例如“麦克”和“话筒”的距离会远小于“麦克”和“天气”的距离。向量的距离可以用传统的欧氏距离来衡量，也可以用 $\cos$ 夹角来衡量。

优选的，所述词向量可为用 Distributed Representation 表示的词向量。Distributed Representation 表示的词向量为一种低维实数向量，这种向量一般形式为 $[0.792, -0.177, -0.107, 0.109, -0.542, \dots]$ ，维度以 50 维和 100 维比较常见。

作为一优选实施方式，所述模型训练单元 210，还用于对检索库进行词向量训练之前，对检索库中各文件分别进行预处理，将各文件预处理后的数据存储到一对应的训练样本文件中，以基于所述训练样本文件对所述检索库进行词向量训练。其中，所述预处理包括数据清洗和提取数据描述。所述数据清洗包括统一大小写、消除多余空格、统一标点符号、统一全半角格式中至少一种；所述提取数据描述包括通过添加用户词典进行分词，具体方式可为添加用户词典并通过 NLPIR（又名 ICTCLAS2013，汉语分词系统）进行分词。

优选的，可通过 word2vec 对所述训练样本文件进行词向量训练，训练设置如下：

```
./word2vec -train result_cropus.txt -output vectors.bin -cbow 0 -size 50  
-window 5 -negative 0 -hs 1 -sample 1e-3 -threads 4 -binary 1 -min_count 3 ;
```

其中，各参数的含义为：

-train 后面表示参与训练的训练样本文件名，
-cbow 表示采用跳空词袋模型，
-size 表示词向量采用的维度，
-window 表示上下文窗口长度，
-negative 表示是否采用负采样方法，0 表示不使用，1 表示使用，

-hs 表示是否使用 HS 方法，0 表示不使用，1 表示使用，

-sample 1e-3 表示采样的阈值为 10^{-3} ，

-thread 表示开启线程数，

-binary 表示输出是否为二进制文件，0 表示不使用，1 表示使用，

5 -min_count 表示设置的最低频率，默认为 5。

进一步的，所述生成相关词单元 220，用于接收输入的检索关键词，通过所述训练模型得到所述检索关键词的相关词，以及各相关词与所述检索关键词的相似度；

本实施例中，两个词向量的相似度指的是余弦相似度，最高可为 1，最低
10 可为 0。由于训练模型是基于检索库训练得到的，因此基于该训练模型得到的相关词能很好反映检索库的用词特点。

所述检索匹配单元 230，用于用所述相关词对所述检索库进行检索匹配，并根据所述相似度分别统计所述检索库中各文件与所述相关词的匹配分值；

优选的，所述检索匹配单元 230 可具体包括：匹配模块，用于用所述相关
15 词分别对所述检索库中各文件进行检索匹配，得到各文件与所述相关词的匹配结果；统计模块，用于将各相关词对应的相似度作为累加权值，结合所述匹配结果分别得出各文件与所述相关词的匹配分值。

所述结果输出单元 240，用于根据所述匹配分值由高到低对所述检索库中的文件进行排序，根据排序结果输出检索结果。

20 本实施例中，还可设定一个分值门限，仅对匹配分值高于所述分值门限的检索结果进行排序，并按照匹配分值由高到低的排序输出。通过设定分值门限对检索结果进一步筛选，有利于用户查阅检索结果。

需要说明的是，上述示例的基于词向量相似度的检索系统的实施方式中，

各模块/单元之间的信息交互、执行过程等内容，由于与本发明前述方法实施例基于同一构思，其带来的技术效果与本发明前述方法实施例相同，具体内容可参见本发明方法实施例中的叙述，此处不再赘述。

此外，上述示例的基于词向量相似度的检索系统的实施方式中，各功能模块的逻辑划分仅是举例说明，实际应用中可以根据需要，例如出于相应硬件的配置要求或者软件的实现的便利考虑，将上述功能分配由不同的功能模块完成，即将所述基于词向量相似度的检索系统的内部结构划分成不同的功能模块，以完成以上描述的全部或者部分功能。

另外，上述示例的基于词向量相似度的检索系统的实施方式中，各功能模块可以集成在一个处理模块中，也可以是各个模块单独物理存在，也可以两个或两个以上模块集成在一个模块中。上述集成的模块既可以采用硬件的形式实现，也可以采用软件功能模块的形式实现。

所述集成的模块如果以软件功能模块的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。本领域普通技术人员可以理解本发明的任意实施例指定的方法的全部或部分步骤是可以通程序来指令相关的硬件(个人计算机、服务器、或者网络设备等)来完成。该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，可执行上述任意实施例指定的方法的全部或部分步骤。前述存储介质可以包括任何可以存储程序代码的介质，例如只读存储器(Read-Only Memory, ROM)、随机存取器(Random Access Memory, RAM)、磁盘或光盘等。

在上述实施例中，对各个实施例的描述都各有侧重，某个实施例中沒有详述的部分，可以参见其它实施例的相关描述。

以上所述实施例仅表达了本发明的几种实施方式，不能理解为对本发明专

利范围的限制。应当指出的是，对于本领域的普通技术人员来说，在不脱离本发明构思的前提下，还可以做出若干变形和改进，这些都属于本发明的保护范围。因此，本发明专利的保护范围应以所附权利要求为准。

权利要求书

1、一种基于词向量相似度的检索方法，其特征在于，包括：

对检索库进行词向量训练，建立所述检索库对应的训练模型；

接收输入的检索关键词，通过所述训练模型得到所述检索关键词的相关

5 词，以及各相关词与所述检索关键词的相似度；

用所述相关词对所述检索库进行检索匹配，并根据所述相似度分别统计所述检索库中各文件与所述相关词的匹配分值；

根据所述匹配分值由高到低对所述检索库中的文件进行排序，根据排序结果输出检索结果。

10 2、根据权利要求 1 所述的基于词向量相似度的检索方法，其特征在于，所述对检索库进行词向量训练之前包括：

对检索库中各文件分别进行预处理，将各文件预处理后的数据存储到一对应的训练样本文件中；所述预处理包括数据清洗和提取数据描述；

所述对检索库进行词向量训练包括：

15 基于所述训练样本文件对所述检索库进行词向量训练。

3、根据权利要求 2 所述的基于词向量相似度的检索方法，其特征在于，所述数据清洗包括统一大小写、消除多余空格、统一标点符号、统一全半角格式中至少一种；

所述提取数据描述包括通过添加用户词典进行分词。

20 4、根据权利要求 2 所述的基于词向量相似度的检索方法，其特征在于，所述对检索库进行词向量训练包括：

通过 word2vec 对所述训练样本文件进行词向量训练。

5、根据权利要求 1 所述的基于词向量相似度的检索方法，其特征在于，

用所述相关词对所述检索库进行检索匹配，并根据所述相似度分别统计所述检索库中各文件与所述相关词的匹配分值包括：

用所述相关词分别对所述检索库中各文件进行检索匹配，得到各文件与所述相关词的匹配结果；

- 5 将各相关词对应的相似度作为累加权值，结合所述匹配结果分别得出各文件与所述相关词的匹配分值。

6、一种基于词向量相似度的检索系统，其特征在于，包括：

模型训练单元，用于对检索库进行词向量训练，建立所述检索库对应的训练模型；

- 10 生成相关词单元，用于接收输入的检索关键词，通过所述训练模型得到所述检索关键词的相关词，以及各相关词与所述检索关键词的相似度；

检索匹配单元，用于用所述相关词对所述检索库进行检索匹配，并根据所述相似度分别统计所述检索库中各文件与所述相关词的匹配分值；

- 15 结果输出单元，用于根据所述匹配分值由高到低对所述检索库中的文件进行排序，根据排序结果输出检索结果。

7、根据权利要求 6 所述的基于词向量相似度的检索系统，其特征在于，所述模型训练单元，还用于对检索库进行词向量训练之前，对检索库中各文件分别进行预处理，将各文件预处理后的数据存储到一对应的训练样本文件中；所述预处理包括数据清洗和提取数据描述；

- 20 所述对检索库进行词向量训练包括：

基于所述训练样本文件对所述检索库进行词向量训练。

8、根据权利要求 7 所述的基于词向量相似度的检索系统，其特征在于，所述数据清洗包括统一大小写、消除多余空格、统一标点符号、统一全半角格

式中至少一种；

所述提取数据描述包括通过添加用户词典进行分词。

9、根据权利要求 7 所述的基于词向量相似度的检索系统，其特征在于，所述对检索库进行词向量训练包括：

5 通过 word2vec 对所述训练样本文件进行词向量训练。

10、根据权利要求 6 所述的基于词向量相似度的检索系统，其特征在于，所述检索匹配单元包括：

匹配模块，用于用所述相关词分别对所述检索库中各文件进行检索匹配，得到各文件与所述相关词的匹配结果；

10 统计模块，用于将各相关词对应的相似度作为累加权值，结合所述匹配结果分别得出各文件与所述相关词的匹配分值。

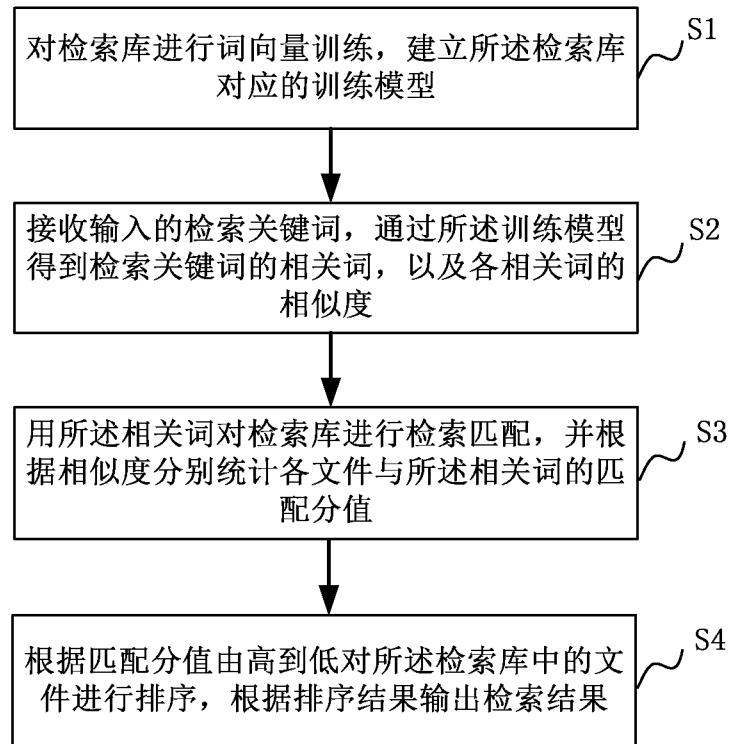


图 1

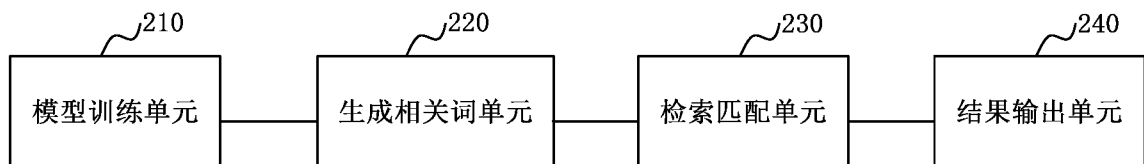


图 2

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2016/098234

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)

WPI, EPODOC, CNPAT, CNKI, IEEE: relative, similarity, weight, word vector, word2vec, google, gensim, score, expand, retrieval, query

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 104933183 A (CHONGQING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS) 23 September 2015 (23.09.2015) description, paragraphs [0013], [0018]-[0021] and [0030]-[0039]	1-10
PX	CN 105631009 A (GUANGZHOU CVTE ELECTRONIC TECHNOLOGY CO., LTD.) 01 June 2016 (01.06.2016) claims 1-10	1-10
A	CN 104778161 A (CHEZHI HULIAN (BEIJING) SCIENCE & TECHNOLOGY CO.,LTD.) 15 July 2015 (15.07.2015) the whole document	1-10
A	CN 105005589 A (TENCENT TECHNOLOGY (SHENZHEN) CO., LTD.) 28 October 2015 (28.10.2015) the whole document	1-10

☒ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	“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
“A” document defining the general state of the art which is not considered to be of particular relevance	“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
“E” earlier application or patent but published on or after the international filing date	“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
“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)	“&” document member of the same patent family
“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	

Date of the actual completion of the international search 01 December 2016	Date of mailing of the international search report 09 December 2016
Name and mailing address of the ISA State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No. (86-10) 62019451	Authorized officer YANG, Dong Telephone No. (86-10) 62413693

INTERNATIONAL SEARCH REPORTInternational application No.
PCT/CN2016/098659

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2015248608 A1 (EDUCATIONAL TESTING SERVICE) 03 September 2015 (03.09.2015) the whole document	1-10

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2016/098234

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 104933183 A	23 September 2015	None	
CN 105631009 A	01 June 2016	None	
CN 104778161 A	15 July 2015	None	
CN 105005589 A	28 October 2015	None	
US 2015248608 A1	03 September 2015	None	

国际检索报告

国际申请号

PCT/CN2016/098234

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))

WPI, EPODOC, CNPAT, CNKI, IEEE: 相关, relative, 相似, similarity, 权重, weight, 词向量, word vector, word2vec, 谷歌, google, gensim, 分值, score, 扩展, expand, 检索, retrieval, 查询, query

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
X	CN 104933183 A (重庆邮电大学) 2015年 9月 23日 (2015 - 09 - 23) 说明书第[0013], [0018]-[0021], [0030]-[0039]段	1-10
PX	CN 105631009 A (广州视源电子科技有限公司) 2016年 6月 1日 (2016 - 06 - 01) 权利要求1-10	1-10
A	CN 104778161 A (车智互联北京科技有限公司) 2015年 7月 15日 (2015 - 07 - 15) 全文	1-10
A	CN 105005589 A (腾讯科技深圳有限公司) 2015年 10月 28日 (2015 - 10 - 28) 全文	1-10
A	US 2015248608 A1 (EDUCATIONAL TESTING SERVICE) 2015年 9月 3日 (2015 - 09 - 03) 全文	1-10

☐ 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

2016年 12月 1日

国际检索报告邮寄日期

2016年 12月 9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

受权官员

杨栋

电话号码 (86-10)010-62413693

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2016/098234

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	104933183	A	2015年 9月 23日	无	
CN	105631009	A	2016年 6月 1日	无	
CN	104778161	A	2015年 7月 15日	无	
CN	105005589	A	2015年 10月 28日	无	
US	2015248608	A1	2015年 9月 3日	无	