(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 12 月 21 日 (21.12.2017)

WIPO | PCT

(10) 国际公布号
WO 2017/215244 A1

(51) 国际专利分类号:
G06F 17/30 (2006.01)

(21) 国际申请号: PCT/CN2016/113175

(22) 国际申请日: 2016 年 12 月 29 日 (29.12.2016)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201610445489.2 2016年6月17日 (17.06.2016) 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: 广州三环专利代理有限公司 (SCIHEAD PATENT AGENT CO., LTD.); 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, 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, 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,

(54) **Title:** METHOD AND DEVICE FOR PROVIDING RELEVANT WORDS

(54) **发明名称**: 提供相关词的方法和装置

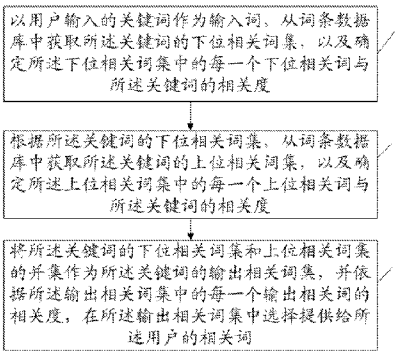


图 1

S1 Using user-inputted keywords as input words, obtaining from an entry database a word set subordinate to the keywords, and determining the degree of relevance to the keywords of each subordinate word in said subordinate word set

S2 On the basis of the word set subordinate to the keywords, obtaining from the entry database a set of words superordinate to the keywords, and determining the relevance to the keywords of each superordinate word in said superordinate word set

S3 Using the combined set of the subordinate word set and the superordinate word set as the relevant word set output with respect to the keywords, then, according to the degree of relevance of each relevant word in the output word set, selecting relevant words from said output word set to provide to the user

(57) Abstract: A method and a device for providing relevant words, the method comprising: using user-inputted keywords as input words, obtaining from an entry database a word set subordinate to the keywords, and determining the degree of relevance to the keywords of each subordinate word in said subordinate word set (S1); on the basis of the word set subordinate to the keywords, obtaining from the entry database a set of words superordinate to the keywords, and determining the relevance to the keywords of each superordinate word in said superordinate word set (S2); using the combined set of the subordinate word set and the superordinate word set as the relevant word set output with respect to the keywords, then, according to the degree of relevance of each relevant word in the output word set, selecting relevant words from said output word set to provide to the user (S3). Use of the present method makes possible provision of a greater number of, and more precise, relevant words.

(57) 摘要: 一种提供相关词的方法和装置，该方法包括：以用户输入的关键词作为输入词，从词条数据库中获取所述关键词的下位相关词集，以及确定所述下位相关词集中的每一个下位相关词与所述关键词的相关度（S1）；根据所述关键词的下位相关词集，从词条数据库中获取所述关键词的上位相关词集，以及确定所述上位相关词集中的每一个上位相关词与所述关键词的相关度（S2）；将所述关键词的下位相关词集和上位相关词集的并集作为所述关键词的输出相关词集，并依据所述输出相关词集中的每一个输出相关词的相关度，在所述输出相关词集中选择提供给所述用户的相关词（S3）。采用该方法，能够提供数量更多且更准确的相关词。

WO 2017/215244 A1

[见续页]

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))。

提供相关词的方法和装置

技术领域

本发明涉及计算机技术领域，尤其涉及一种提供相关词的方法和装置。

背景技术

目前，购物网站以及搜索引擎服务网站都提供的关键词搜索的功能，即用户输入想要搜索的商品或技术的关键词，服务器则根据该关键词搜索出相应的结果并返回给用户。服务器为了提供准确的搜索结果，服务器一般会对关键词进行扩展，即根据用户输入的关键词，查找出关键词对应的相关词，并提供查找到的相关词给用户，在用户通过关键词搜索而未能得到满意的搜索结果时，就根据相关词进行搜索。但现有的相关词扩展是通过已有词典进行扩展，例如 WordNet、《同义词林》，而这种方式所获得的相关词在数量上相当有限，而且所获得的相关词有可能跟不上语言的发展变化，不能满足相关词对时效性的要求。

发明内容

本发明实施例提出一种提供相关词的方法和装置，能够提供数量更多且更准确的相关词。

本发明实施例提出的一种提供相关词的方法，包括：

以用户输入的关键词作为输入词，从词条数据库中获取所述关键词的下位相关词集，以及确定所述下位相关词集中的每一个下位相关词与所述关键词的相关度；

根据所述关键词的下位相关词集，从词条数据库中获取所述关键词的上位相关词集，以及确定所述上位相关词集中的每一个上位相关词与所述关键词的相关度；

将所述关键词的下位相关词集和上位相关词集的并集作为所述关键词的输出相关词集，并依据所述输出相关词集中的每一个输出相关词的相关度，在所述输出相关词集中选择提供给所述用户的相关词。

作为本发明实施例的进一步改进，所述根据所述关键词的下位相关词集，从词条数据库中获取所述关键词的上位相关词集，以及确定所述上位相关词集中的每一个上位相关词与所述关键词的相关度，具体为：

针对所述下位相关词集中的每一个下位相关词，以该下位相关词来更新输入词，从词条数据库中获取更新后的输入词的下位相关词集；

判断下位相关词集的总数量是否大于预设阈值;

若是，则从下位相关词集中筛选出包含所述关键词的下位相关词集，并将所述包含所述关键词的下位相关词集对应的输入词作为上位相关词，获得所述关键词的上位相关词集；其中，在所述包含所述关键词的下位相关词集中的所述关键词与该下位相关词集对应的输入词的相关度，作为该输入词在作为上位相关词时与所述关键词的相关度;

若否，则继续执行以下操作：针对更新后的输入词的下位相关词集中的每一个下位相关词，以该下位相关词再次更新输入词，从词条数据库中获取再次更新后的输入词的下位相关词集，直至下位相关词集的总数量大于预设阈值。

进一步地，上述从词条数据库中获取下位相关词集的方式具体包括:

根据所述输入词，从词条数据库中获取包含所述输入词的词条，并对所述词条进行分词和筛选，获得待验相关词集;

对于所述待验相关词集中的每一个待验相关词，根据所述待验相关词，从所述词条数据库中获取包含所述待验相关词的词条，并对所述待验相关词的词条进行分词和筛选，获得所述待验相关词的对照词集;

当判定所述待验相关词的对照词集与所述待验相关词集的交集的绝对值大于筛选阈值时，所述待验相关词为所述输入词的下位相关词，获得下位相关词集；其中，所述绝对值作为所述下位相关词与所述关键词的相关度。

作为本发明的进一步改进，所述根据所述输入词，从词条数据库中获取包含所述输入词的词条，并对所述词条进行分词和筛选，获得待验相关词集，具体包括:

根据所述输入词，从词条数据库中获取包含所述输入词且排序在第 M 位前的词条;

根据标准词条格式，对获取的词条进行格式调整;

调用分词工具;

利用所述分词工具对格式调整后的词条进行分词，获得第一词语集;

从所述第一词语集中提取属于用户词典中的核心词的词语作为待验相关词，获得待验相关词集；其中，所述用户词典是由所述分词工具提供的;

以及，所述根据所述待验相关词，从所述词条数据库中获取包含所述待验相关词的词条，并对所述待验相关词的词条进行分词和筛选，获得所述待验相关词的对照词集，具体包括:

根据所述待验相关词，从词条数据库中获取包含所述待验相关词且排序在第 M 位前的词条;

根据所述标准词条格式，对所述包含所述待验相关词且排序在第 M 位前的词条进行格式调整;

调用所述分词工具;

利用所述分词工具对格式调整后的包含所述待验相关词且排序在第 M 位前的词条进行分词，获得第二词语集;

从所述第二词语集中提取属于用户词典中的核心词的词语作为对照词，获得对照词集。

具体地，所述关键词的下位相关词集和所述上位相关词集的交集包含在所述关键词的输出相关词集中，则包含在所述交集中的每一个输出相关词的相关度为 T，T=(T1+T2)/2; 其中，T1 为在该输出相关词作为下位相关词时与所述关键词的相关度，T2 作为在该输出相关词作为上位相关词时与所述关键词的相关度。

作为本发明的进一步改，所述获取方法还包括:

将所述关键词的下位相关词集中的每一个下位相关词与所述关键词的相关度均减去所述筛选阈值;

将所述关键词的上位相关词集中的每一个上位相关词与所述关键词的相关度均减去所述筛选阈值，完成相关度的归一化。

相应地，本发明实施还提供一种提供相关词的装置，包括:

下位相关词集模块，用于以用户输入的关键词作为输入词，从词条数据库中获取所述关键词的下位相关词集，以及确定所述下位相关词集中的每一个下位相关词与所述关键词的相关度;

上位相关词集模块，用于根据所述关键词的下位相关词集，从词条数据库中获取所述关键词的上位相关词集，以及确定所述上位相关词集中的每一个上位相关词与所述关键词的相关度;

输出相关词集模块，用于将所述关键词的下位相关词集和上位相关词集的并集作为所述关键词的输出相关词集，并依据所述输出相关词集中的每一个输出相关词的相关度，在所述输出相关词集中选择提供给所述用户的相关词。

作为本发明实施例的进一步改进，所述上位相关词集模块具体包括: 下位词集获取单元、阈值判断单元和上位词集获取单元，其中，

所述下位词集获取单元，用于针对所述下位相关词集中的每一个下位相关词，以该下位相关词来更新输入词，从词条数据库中获取更新后的输入词的下位相关词集;

所述阈值判断单元，用于判断下位相关词集的总数量是否大于预设阈值;

所述上位词集获取单元，用于当判断下位相关词集的总数量大于预设阈值时，从下位相关词集中筛选出包含所述关键词的下位相关词集，并将所述包含所述关键词的下位相关词集

对应的输入词作为上位相关词，获得所述关键词的上位相关词集；其中，在所述包含所述关键词的下位相关词集中的所述关键词与该下位相关词集对应的输入词的相关度，作为该输入词在作为上位相关词时与所述关键词的相关度；

所述下位词获取单元，还用于当判断下位相关词集的总数量小于预设阈值时，继续执行以下操作：针对更新后的输入词的下位相关词集中的每一个下位相关词，以该下位相关词再次更新输入词，从词条数据库中获取再次更新后的输入词的下位相关词集，直至下位相关词集的总数量大于预设阈值。

进一步地，所述下位相关词集模块和所述下位词集获取单元还包括用于从词条数据库中获取下位相关词集的单元，具体为：

待验相关词集单元，用于根据所述输入词，从词条数据库中获取包含所述输入词的词条，并对所述词条进行分词和筛选，获得待验相关词集；

对照词集单元，用于对于所述待验相关词集中的每一个待验相关词，根据所述待验相关词，从所述词条数据库中获取包含所述待验相关词的词条，并对所述待验相关词的词条进行分词和筛选，获得所述待验相关词的对照词集；和

判断获取单元，用于当判定所述待验相关词的对照词集与所述待验相关词集的交集的绝对值大于筛选阈值时，所述待验相关词为所述输入词的下位相关词，获得下位相关词集；其中，所述绝对值作为所述下位相关词与所述关键词的相关度。

进一步地，所述待验相关词集单元，具体包括：

第一词条子单元，用于根据所述输入词，从词条数据库中获取包含所述输入词且排序在第 M 位前的词条；

第一调整子单元，用于根据标准词条格式，对获取的词条进行格式调整；

第一调用子单元，用于调用分词工具；

第一分词子单元，用于利用所述分词工具对格式调整后的词条进行分词，获得第一词语集；和，

第一提取子单元，用于从所述第一词语集中提取属于用户词中的核心词的词语作为待验相关词，获得待验相关词集；其中，所述用户词典是由所述分词工具提供的；

以及，所述对照词集单元具体包括：

第二词条子单元，用于根据所述待验相关词，从词条数据库中获取包含所述待验相关词且排序在第 M 位前的词条；

第二调整子单元，用于根据所述标准词条格式，对所述包含所述待验相关词且排序在第 M 位前的词条进行格式调整；

第二调用子单元，用于调用所述分词工具；

第二分词子单元，用于利用所述分词工具对格式调整后的包含所述待验相关词且排序在第 M 位前的词条进行分词，获得第二词语集；和，

第二提取子单元，用于根据从所述第二词语集中提取属于用户词典中的核心词的词语作为对照词，获得对照词集。

进一步地，所述提供相关词的装置还包括归一化模块：

所述归一化模块，用于将所述关键词的下位相关词集中的每一个下位相关词与所述关键词的相关度均减去所述筛选阈值；以及用于将所述关键词的上位相关词集中的每一个上位相关词与所述关键词的相关度均减去所述筛选阈值，完成相关度的归一化。

实施本发明实施例，具有如下有益效果：

本发明实施例提供的提供相关词的方法和装置，通过用户提供的关键词从词条数据库中获取所述关键词的下位相关词集，然后再根据该下位相关词集，求取出关键词的上位相关词集，最后该下位相关词集和该上位相关词集的并集作为所述关键词的输出相关词集，能扩展出大量的相关词提供给用户选择，另外，通过确定相关词的相关度，能准确地描述为相关词与关键词之间的相关程度，后续可依据相关词的相关度选择提供给所述用户的相关词，能通过相关词的相关度描述，准确地为用户提供相关词。

附图说明

图 1 是本发明提供的提供相关词的方法的一个实施例的流程示意图；

图 2 是图 1 提供的提供相关词的方法的步骤 S2 的一个实施例的流程示意图；

图 3 是图 1 提供的提供相关词的方法的步骤 S3 的一个实施的流程示意图；

图 4 是本发明提供的提供相关词的方法的另一个实施例的流程示意图；

图 5 是本发明提供的提供相关词的装置的一个实施例的结构示意图；

图 6 是本发明提供的提供相关词的装置的上位相关词集模块的一个实施例的结构示意图；

图 7 是本发明提供的提供相关词的装置的用于获取下位相关词集的单元的一个实施例的结构示意图；

图 8 是本发明提供的提供相关词的装置的待验相关词集单元的一个实施例的结构示意图；

图 9 是本发明提供的提供相关词的装置的对照词集单元的一个实施例的结构示意图。

具体实施方式

下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

参见图 1、图 2 和图 3，图 1 是本发明提供的提供相关词的方法的一个实施例的流程示意图，图 2 是图 1 提供的提供相关词的方法的步骤 S2 的一个实施例的流程示意图，图 3 是图 1 提供的提供相关词的方法的步骤 S3 的一个实施的流程示意图。下面将结合这三个流程图，以论文数据库（例如中国知网）作为词条数据库，从中获取关键词 Java 的相关词为例，详细说明本实施例的提供相关词的方法，该方法包括以下步骤:

S1，以用户输入的关键词 Java 为输入词，从词条数据库中获取关键词 Java 的下位相关词集，以及确定所述下位相关词集中的每一个下位相关词与所述关键词的相关度。步骤 S1 包括步骤 S11 至 S13，具体如下:

S11，根据所述输入词 Java 从论文数据库中获取包含所述输入词 Java 的词条，并对所述词条进行分词和筛选，获得待验相关词集 $A=\{a_1,\ldots,a_n\}$；此步骤的具体实施过程如下:

利用搜索引擎根据所述输入词 Java 从论文数据库中获取包含所述输入词 Java 且排序在第 M 位前的词条，例如，前 50 页论文摘要作为词条，或者，在维基中搜索关键词 Java 的前 500 条摘要;

根据标准词条格式对所述词条进行格式调整；例如，将词条中的小写统一成大写、对词条中多余的空格删除、统一词条中的标点符号、将词条的全角格式或半角格式统一为一种等。

调用分词工具；优选地，所述分词工具为 jieba 分词工具，但不限于为此分词工具。

利用所述分词工具对格式调整后的词条进行分词，获得第一词语集;

根据关键词提取算法，从所述第一词语集中提取与所述输入词相关的词语作为待验相关词$\{a_1,\ldots,a_n\}$，获得待验相关词集 $A=\{a_1,\ldots,a_n\}$。需要说明的是，可通过分词工具或通过本提供相关词的装置添加词典，利用词典提供的核心词，从所述第一词语集中提取核心词作为待验相关词。

S12，对于所述待验相关词集 $A=\{a_1,\ldots,a_n\}$中的每一个待验相关词，根据所述待验相关词从所述词条数据库中获取包含所述待验相关词的词条，并对所述待验相关词的词条进行分词和筛选，获得所述待验相关词的对照词集。需要说明的是，此步骤 S22 与上一个步骤 S21 的具体实施过程相同，只是区别在于步骤 S21 中的输入词变为待验相关词$\{a_1,\ldots,a_n\}$，然后将所获得待验相关词 a_i 的待验相关词集 $B_{ai}=\{b_{i1},\ldots,b_{in}\}$作为待验相关词 a_i 的对照词集，因而在此不

再赘述。

S13，当判定所述待验相关词 a_i 的对照词集 $B_{ai}=\{b_{i1},\ldots,b_{in}\}$ 与所述待验相关词集 $A=\{a_1,\ldots,a_n\}$ 的交集的绝对值 r 大于筛选阈值 p 时，即 B_{ai} 集合与 A 集合中相同元素的数量大于筛选阈值 p 时，所述待验相关词 a_i 为所述输入词 Java 的下位相关词，获得所述关键词的下位相关词集 $A'=\{a_j\}$，且 $j\in\{1,\ldots,n\}$、$|A'|\leq n$、$|A\cap B_{aj}|>p$；其中，所述交集的绝对值 r 为所述下位相关词在所述下位相关词集中的相关度。需要说明的是，所述相关度表示为相关词集中的相关词与该相关词集的输入词之间的相关程度。

通过上述步骤 S11、S12 和 S13 来获取输入词的下位相关词集，能滤除噪音词，提高获取下位相关词的效率。

根据所述关键词的下位相关词集，从词条数据库中获取所述关键词的上位相关词集，以及确定所述上位相关词集中的每一个上位相关词与所述关键词的相关度

S2，根据所述关键词的下位相关词集 $A'=\{a_j\}$，从词条数据库中获取所述关键词的上位相关词集，以及确定所述上位相关词集中的每一个上位相关词与所述关键词的相关度。该步骤的具体包括如下步骤 S21 至 S24:

S21，针对所述下位相关词集 $A'=\{a_j\}$ 中的每一个下位相关词，以该下位相关词 a_j 来更新输入词，即作为输入词，从论文数据库中获取更新后的输入词 a_j 的下位相关词集 A''；需要说明的，在本实施例中，优选地，步骤 S31 中获取下位相关词集的方式与上述步骤 S2 中获取下位相关词集的方式一致，在此不再赘述。

S22，判断当前下位相关词集的总数量 N 是否大于预设阈值 S;

S23，若是，则从下位相关词集中筛选出所有包含所述关键词的下位相关词集，并将所述包含所述关键词的下位相关词集对应的输入词作为上位相关词，获得所述关键词的上位相关词集 C；其中，在所述包含所述关键词的下位相关词集中的所述关键词与该下位相关词集对应的输入词的相关度，作为该输入词在作为上位相关词时与所述关键词的相关度

S24，若否，则继续执行以下操作：针对更新后的输入词的下位相关词集中的每一个下位相关词，以该下位相关词再次更新输入词，从词条数据库中获取再次更新后的输入词的下位相关词集，直至下位相关词集的总数量 N 大于预设阈值 S；需要说明的是，在步骤 S21 和 S23 中获取下位相关词集的方式与在上述步骤 S1 中获取下位相关词集的方式也是一致的，在此不再赘述。

也就是说，例如对于关键词 Java 来说，Java 的上位相关词集是一个集合中的元素为输入词的集合，即该集合中的每一个元素的下位相关词集中都有相同的元素为 Java。通过采用与获取下位相关词集相同的方式逆求取关键词的上位相关词集，能从多个维度为用户提供相关

词。

S3，将所述关键词的下位相关词集和上位相关词集的并集作为所述关键词的输出相关词集，并依据所述输出相关词集中的每一个输出相关词的相关度，在所述输出相关词集中选择提供给所述用户的相关词。

具体地所述关键词的下位相关词集和所述上位相关词集的交集包含在所述关键词的输出相关词集中，则包含在所述交集中的每一个输出相关词的相关度为 T，T=(T1+T2)/2；其中，T1 为在该输出相关词作为下位相关词时与所述关键词的相关度，T2 作为在该输出相关词作为上位相关词时与所述关键词的相关度。也就是说，并集后，上位相关词集和下位相关词集中相同的相关词的相关度的取值为该相关词在这两个集合中的相关度的均值。

作为本发明的进一步改进，所述获取方法还包括对相关度进行归一化:

将所述关键词的下位相关词集中的每一个下位相关词与所述关键词的相关度均减去所述筛选阈值;

将所述关键词的上位相关词集中的每一个上位相关词与所述关键词的相关度均减去所述筛选阈值，完成相关度的归一化。

需要说明的是，归一化的目的是让关键词的输出相关词集中的相关词与该关键词的相关程度的相关度的数值能以 0 为基准，数值越高，相关词与关键词的相关程度就越高，方便在步骤 S4 中在输出相关词集中选择提供给用户的相关词。

实施本发明实施例的提供相关词的方法，通过对获取的待验下位相关词进行对照验证后的相关词作为下位相关词，能滤除噪声词的影响，提高获取到的相关词的质量，也就是说，能确保提供给用户的相关词的准确性。另一方面，在获取到关键词的下位相关词集后，继续通过下位相关词集时进行逆求取关键词的上位相关词，能大量扩展为用户提供的相关词的数量，并能确上位相关词的质量。

参见图 4，是本发明提供的提供相关词的方法的另一个实施例的流程示意图；本实施例的提供相关词的方法为：分别以论文数据库和维基百科数据库作为词条数据库，从中获取相应的第一输出相关词集和第二输出相关词集，然后将第一输出相关词集和第二输出相关词集的并集作为关键词的最终的输出相关词集；其中，在维基百科数据库中获取第二输出相关词集的方式与上一实施例中在论文数据库中获取输出相关词集的方式相同。本实施例采用两种不同的词条数据库且词条数据库为论文数据库和维基百科数据库，来进行相关词的挖掘，一方面对于相关词的扩展针对性强，且能避免语料单一，而导致的为用户提供的相关词获取过于片面。

相应地，参见图 5，是本发明提供的提供相关词的装置的一个实施例的结构示意图，能实现上述两种实施例的全部流程，该提供相关词的装置包括：

下位相关词集模块 10，用于以用户输入的关键词作为输入词，从词条数据库中获取所述关键词的下位相关词集，以及确定所述下位相关词集中的每一个下位相关词与所述关键词的相关度；

上位相关词集模块 20，用于根据所述关键词的下位相关词集，从词条数据库中获取所述关键词的上位相关词集，以及确定所述上位相关词集中的每一个上位相关词与所述关键词的相关度；

输出相关词集模块 30，用于将所述关键词的下位相关词集和上位相关词集的并集作为所述关键词的输出相关词集，并依据所述输出相关词集中的每一个输出相关词的相关度，在所述输出相关词集中选择提供给所述用户的相关词。

作为本发明实施例的进一步改进，如图 6 所示，图 6 是本发明提供的提供相关词的装置的上位相关词集模块的一个实施例的结构示意图；该上位相关词集模块 30 具体包括：下位词集获取单元 31、阈值判断单元 32 和上位词集获取单元 33，其中，

所述下位词集获取单元 31，用于针对所述下位相关词集中的每一个下位相关词，以该下位相关词来更新输入词，从词条数据库中获取更新后的输入词的下位相关词集；

所述阈值判断单元 32，用于判断下位相关词集的总数量是否大于预设阈值；

所述上位词集获取单元 33，用于当判断下位相关词集的总数量大于预设阈值时，从下位相关词集中筛选出包含所述关键词的下位相关词集，并将所述包含所述关键词的下位相关词集对应的输入词作为上位相关词，获得所述关键词的上位相关词集；其中，在所述包含所述关键词的下位相关词集中的所述关键词与该下位相关词集对应的输入词的相关度，作为该输入词在作为上位相关词时与所述关键词的相关度；

所述下位词集获取单元 31，还用于当判断下位相关词集的总数量小于预设阈值时，继续执行以下操作：针对更新后的输入词的下位相关词集中的每一个下位相关词，以该下位相关词再次更新输入词，从词条数据库中获取再次更新后的输入词的下位相关词集，直至下位相关词集的总数量大于预设阈值。

进一步地，所述下位相关词集模块 20 和所述下位词集获取单元 31 均还包括用于从词条数据库中获取下位相关词集的单元，如图 7 所示，图 7 是本发明提供的提供相关词的装置的用于获取下位相关词集的单元的一个实施例的结构示意图，具体包括：

待验相关词集单元 1，用于根据所述输入词，从词条数据库中获取包含所述输入词的词

条，并对所述词条进行分词和筛选，获得待验相关词集；

对照词集单元 2，用于对于所述待验相关词集中的每一个待验相关词，根据所述待验相关词，从所述词条数据库中获取包含所述待验相关词的词条，并对所述待验相关词的词条进行分词和筛选，获得所述待验相关词的对照词集；和

判断获取单元 3，用于当判定所述待验相关词的对照词集与所述待验相关词集的交集的绝对值大于筛选阈值时，所述待验相关词为所述输入词的下位相关词，获得下位相关词集；其中，所述绝对值作为所述下位相关词在与所述关键词的相关度。

进一步地，如图 8 所示，图 8 是本发明提供的提供相关词的装置的待验相关词集单元的一个实施例的结构示意图；所述待验相关词集单元 1，具体包括：

第一词条子单元 11，用于根据所述输入词，从词条数据库中获取包含所述输入词且排序在第 M 位前的词条；

第一调整子单元 12，用于根据标准词条格式，对获取的词条进行格式调整；

第一调用子单元 13，用于调用分词工具；

第一分词子单元 14，用于利用所述分词工具对格式调整后的词条进行分词，获得第一词语集；和，

第一提取子单元 15，用于根从所述第一词语集中提取属于用户词中的核心词的词语作为待验相关词，获得待验相关词集；其中，所述用户词典是由所述分词工具提供的；

以及，如图 9 所示，图 9 是本发明提供的提供相关词的装置的对照词集单元的一个实施例的结构示意图；所述对照词集单元 2 具体包括：

第二词条子单元 21，用于根据所述待验相关词，从词条数据库中获取包含所述待验相关词且排序在第 M 位前的词条；

第二调整子单元 22，用于根据所述标准词条格式，对所述包含所述待验相关词且排序在第 M 位前的词条进行格式调整；

第二调用子单元 23，用于调用所述分词工具；

第二分词子单元 24，用于利用所述分词工具对格式调整后的包含所述待验相关词且排序在第 M 位前的词条进行分词，获得第二词语集；和，

第二提取子单元 25，用于从所述第二词语集中属于用户词典中的核心词的词语作为对照词，获得对照词集。

具体地，所述关键词的下位相关词集和所述上位相关词集的交集包含在所述关键词的输出相关词集中，则包含在所述交集中的每一个输出相关词的相关度为 T，T=(T1+T2)/2；其中，T1 为在该输出相关词作为下位相关词时与所述关键词的相关度，T2 作为在该输出相关词作

为上位相关词时与所述关键词的相关度。

进一步地，如图 5 所示，所述提供相关词的装置还包括归一化模块 40:

所述归一化模块，用于将所述关键词的下位相关词集中的每一个下位相关词与所述关键词的相关度均减去所述筛选阈值；以及用于将所述关键词的上位相关词集中的每一个上位相关词与所述关键词的相关度均减去所述筛选阈值，完成相关度的归一化。

本发明实施例提供的提供相关词的装置，通过对获取的待验下位相关词进行对照验证后的相关词作为下位相关词，能滤除噪声词的影响，提高获取到的相关词的质量，也就是说，能确保提供给用户的相关词的准确性。另一方面，在获取到关键词的下位相关词集后，继续通过下位相关词集时进行逆求取关键词的上位相关词，能大量扩展为用户提供的相关词的数量，并能确上位相关词的质量。

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，是可以通过计算机程序来指令相关的硬件来完成，所述的程序可存储于一计算机可读取存储介质中，该程序在执行时，可包括如上述各方法的实施例的流程。其中，所述的存储介质可为磁碟、光盘、只读存储记忆体（Read-Only Memory，ROM）或随机存储记忆体（Random Access Memory，RAM）等。

以上所述是本发明的优选实施方式，应当指出，对于本技术领域的普通技术人员来说，在不脱离本发明原理的前提下，还可以做出若干改进和润饰，这些改进和润饰也视为本发明的保护范围。

权利要求书

1、一种提供相关词的方法，其特征在于，包括:

以用户输入的关键词作为输入词，从词条数据库中获取所述关键词的下位相关词集，以及确定所述下位相关词集中的每一个下位相关词与所述关键词的相关度;

根据所述关键词的下位相关词集，从词条数据库中获取所述关键词的上位相关词集，以及确定所述上位相关词集中的每一个上位相关词与所述关键词的相关度;

将所述关键词的下位相关词集和上位相关词集的并集作为所述关键词的输出相关词集，并依据所述输出相关词集中的每一个输出相关词的相关度，在所述输出相关词集中选择提供给所述用户的相关词。

2、如权利要求 1 所述的提供相关词的方法，其特征在于，所述根据所述关键词的下位相关词集，从词条数据库中获取所述关键词的上位相关词集，以及确定所述上位相关词集中的每一个上位相关词与所述关键词的相关度，具体为:

针对所述下位相关词集中的每一个下位相关词，以该下位相关词来更新输入词，从词条数据库中获取更新后的输入词的下位相关词集;

判断下位相关词集的总数量是否大于预设阈值;

若是，则从下位相关词集中筛选出包含所述关键词的下位相关词集，并将所述包含所述关键词的下位相关词集对应的输入词作为上位相关词，获得所述关键词的上位相关词集；其中，在所述包含所述关键词的下位相关词集中的所述关键词与该下位相关词集对应的输入词的相关度，作为该输入词在作为上位相关词时与所述关键词的相关度;

若否，则继续执行以下操作：针对更新后的输入词的下位相关词集中的每一个下位相关词，以该下位相关词再次更新输入词，从词条数据库中获取再次更新后的输入词的下位相关词集，直至下位相关词集的总数量大于预设阈值。

3、如权利要求 2 所述的提供相关词的方法，其特征在于，从词条数据库中获取下位相关词集的方式具体包括:

根据所述输入词，从词条数据库中获取包含所述输入词的词条，并对所述词条进行分词和筛选，获得待验相关词集;

对于所述待验相关词集中的每一个待验相关词，根据所述待验相关词，从所述词条数据库中获取包含所述待验相关词的词条，并对所述待验相关词的词条进行分词和筛选，获得所

述待验相关词的对照词集;

当判定所述待验相关词的对照词集与所述待验相关词集的交集的绝对值大于筛选阈值时，所述待验相关词为所述输入词的下位相关词，获得下位相关词集；其中，所述绝对值作为所述下位相关词与所述关键词的相关度。

4、如权利要求 3 所述的提供相关词的方法，其特征在于，所述根据所述输入词，从词条数据库中获取包含所述输入词的词条，并对所述词条进行分词和筛选，获得待验相关词集，具体包括:

根据所述输入词，从词条数据库中获取包含所述输入词且排序在第 M 位前的词条;

根据标准词条格式，对获取的词条进行格式调整;

调用分词工具;

利用所述分词工具对格式调整后的词条进行分词，获得第一词语集;

从所述第一词语集中提取属于用户词典中的核心词的词语作为待验相关词，获得待验相关词集；其中，所述用户词典是由所述分词工具提供的;

以及，所述根据所述待验相关词，从所述词条数据库中获取包含所述待验相关词的词条，并对所述待验相关词的词条进行分词和筛选，获得所述待验相关词的对照词集，具体包括:

根据所述待验相关词，从词条数据库中获取包含所述待验相关词且排序在第 M 位前的词条;

根据所述标准词条格式，对所述包含所述待验相关词且排序在第 M 位前的词条进行格式调整;

调用所述分词工具;

利用所述分词工具对格式调整后的包含所述待验相关词且排序在第 M 位前的词条进行分词，获得第二词语集;

从所述第二词语集中提取属于用户词典中的核心词的词语作为对照词，获得对照词集。

5、如权利要求 1 所述的提供相关词的方法，其特征在于，所述关键词的下位相关词集和所述上位相关词集的交集包含在所述关键词的输出相关词集中，则包含在所述交集中的每一个输出相关词的相关度为 T，T=(T1+T2)/2；其中，T1 为在该输出相关词作为下位相关词时与所述关键词的相关度，T2 作为在该输出相关词作为上位相关词时与所述关键词的相关度。

6、如权利要求 3 所述的提供相关词的方法，其特征在于，所述获取方法还包括:

将所述关键词的下位相关词集中的每一个下位相关词与所述关键词的相关度均减去所述筛选阈值;

将所述关键词的上位相关词集中的每一个上位相关词与所述关键词的相关度均减去所述筛选阈值，完成相关度的归一化。

7、一种提供相关词的装置，其特征在于，包括:

下位相关词集模块，用于以用户输入的关键词作为输入词，从词条数据库中获取所述关键词的下位相关词集，以及确定所述下位相关词集中的每一个下位相关词与所述关键词的相关度;

上位相关词集模块，用于根据所述关键词的下位相关词集，从词条数据库中获取所述关键词的上位相关词集，以及确定所述上位相关词集中的每一个上位相关词与所述关键词的相关度;

输出相关词集模块，用于将所述关键词的下位相关词集和上位相关词集的并集作为所述关键词的输出相关词集，并依据所述输出相关词集中的每一个输出相关词的相关度，在所述输出相关词集中选择提供给所述用户的相关词。

8、如权利要求 7 所述的提供相关词的装置，其特征在于，所述上位相关词集模块具体包括：下位词集获取单元、阈值判断单元和上位词集获取单元，其中，

所述下位词集获取单元，用于针对所述下位相关词集中的每一个下位相关词，以该下位相关词来更新输入词，从词条数据库中获取更新后的输入词的下位相关词集;

所述阈值判断单元，用于判断下位相关词集的总数量是否大于预设阈值;

所述上位词集获取单元，用于当判断下位相关词集的总数量大于预设阈值时，从下位相关词集中筛选出包含所述关键词的下位相关词集，并将所述包含所述关键词的下位相关词集对应的输入词作为上位相关词，获得所述关键词的上位相关词集；其中，在所述包含所述关键词的下位相关词集中的所述关键词与该下位相关词集对应的输入词的相关度，作为该输入词在作为上位相关词时与所述关键词的相关度;

所述下位词集获取单元，还用于当判断下位相关词集的总数量小于预设阈值时，继续执行以下操作：针对更新后的输入词的下位相关词集中的每一个下位相关词，以该下位相关词再次更新输入词，从词条数据库中获取再次更新后的输入词的下位相关词集，直至下位相关词集的总数量大于预设阈值。

9、如权利要求 8 所述的提供相关词的装置，其特征在于，所述下位相关词集模块和所述下位词集获取单元还包括用于从词条数据库中获取下位相关词集的单元，具体为：

待验相关词集单元，用于根据所述输入词，从词条数据库中获取包含所述输入词的词条，并对所述词条进行分词和筛选，获得待验相关词集；

对照词集单元，用于对于所述待验相关词集中的每一个待验相关词，根据所述待验相关词，从所述词条数据库中获取包含所述待验相关词的词条，并对所述待验相关词的词条进行分词和筛选，获得所述待验相关词的对照词集；和

判断获取单元，用于当判定所述待验相关词的对照词集与所述待验相关词集的交集的绝对值大于筛选阈值时，所述待验相关词为所述输入词的下位相关词，获得下位相关词集；其中，所述绝对值作为所述下位相关词与所述关键词的相关度。

10、如权利要求 9 所述的提供相关词的装置，其特征在于，所述待验相关词集单元，具体包括：

第一词条子单元，用于根据所述输入词，从词条数据库中获取包含所述输入词且排序在第 M 位前的词条；

第一调整子单元，用于根据标准词条格式，对获取的词条进行格式调整；

第一调用子单元，用于调用分词工具；

第一分词子单元，用于利用所述分词工具对格式调整后的词条进行分词，获得第一词语集；和，

第一提取子单元，用于从所述第一词语集中提取属于用户词中的核心词的词语作为待验相关词，获得待验相关词集；其中，所述用户词典是由所述分词工具提供的；

以及，所述对照词集单元具体包括：

第二词条子单元，用于根据所述待验相关词，从词条数据库中获取包含所述待验相关词且排序在第 M 位前的词条；

第二调整子单元，用于根据所述标准词条格式，对所述包含所述待验相关词且排序在第 M 位前的词条进行格式调整；

第二调用子单元，用于调用所述分词工具；

第二分词子单元，用于利用所述分词工具对格式调整后的包含所述待验相关词且排序在第 M 位前的词条进行分词，获得第二词语集；和，

第二提取子单元，用于根据从所述第二词语集中提取属于用户词典中的核心词的词语作为对照词，获得对照词集。

11、如权利要求 10 所述的提供相关词的装置，其特征在于，所述提供相关词的装置还包括归一化模块：

所述归一化模块，用于将所述关键词的下位相关词集中的每一个下位相关词与所述关键词的相关度均减去所述筛选阈值；以及用于将所述关键词的上位相关词集中的每一个上位相关词与所述关键词的相关度均减去所述筛选阈值，完成相关度的归一化。

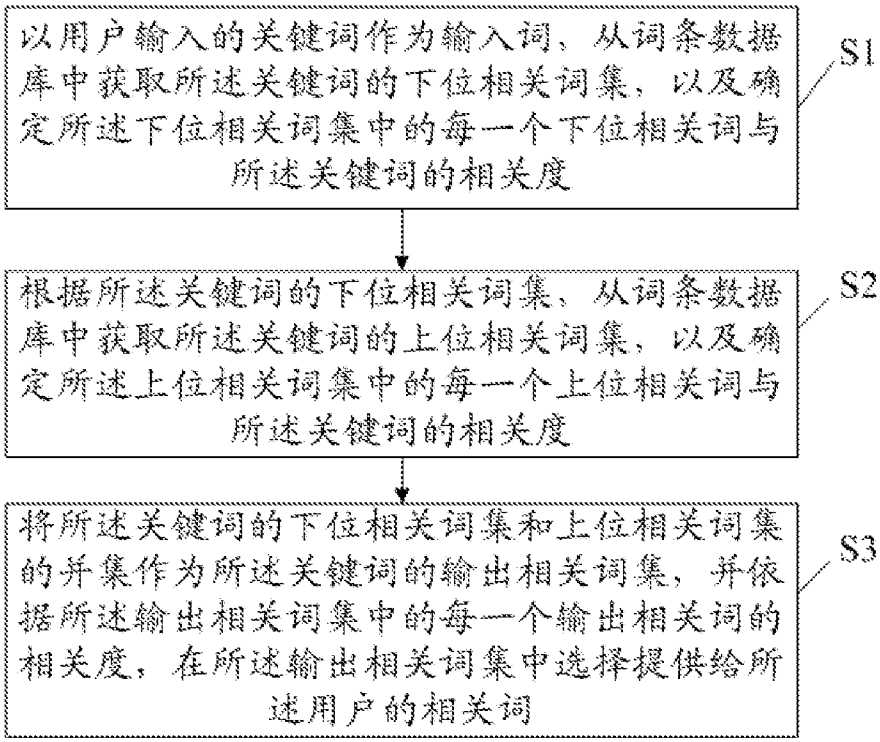

以用户输入的关键词作为输入词，从词条数据库中获取所述关键词的下位相关词集，以及确定所述下位相关词集中的每一个下位相关词与所述关键词的相关度
S1
根据所述关键词的下位相关词集，从词条数据库中获取所述关键词的上位相关词集，以及确定所述上位相关词集中的每一个上位相关词与所述关键词的相关度
S2
将所述关键词的下位相关词集和上位相关词集的并集作为所述关键词的输出相关词集，并依据所述输出相关词集中的每一个输出相关词的相关度，在所述输出相关词集中选择提供给所述用户的相关词
S3

图 1

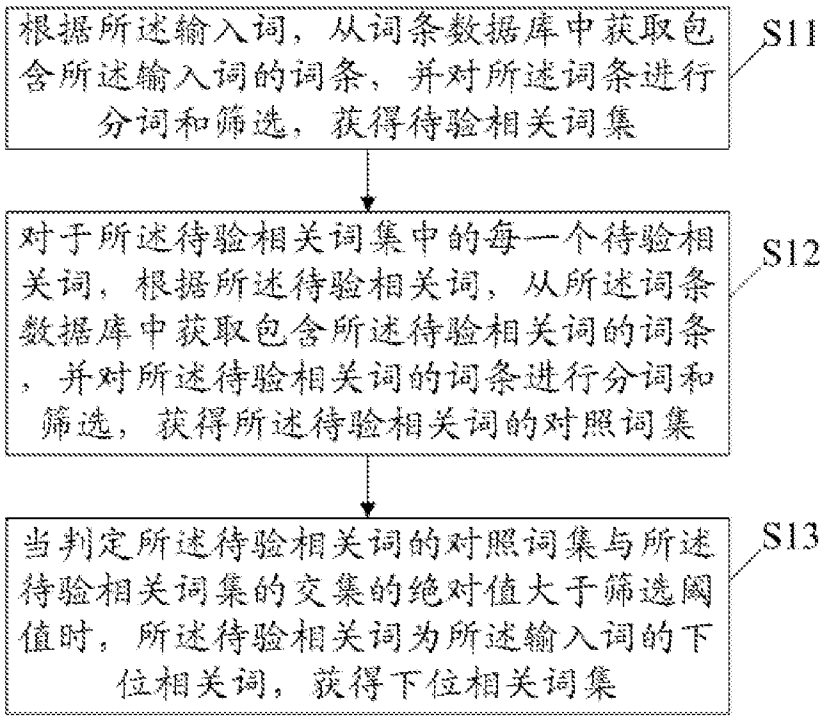

根据所述输入词，从词条数据库中获取包含所述输入词的词条，并对所述词条进行分词和筛选，获得待验相关词集
S11
对于所述待验相关词集中的每一个待验相关词，根据所述待验相关词，从所述词条数据库中获取包含所述待验相关词的词条，并对所述待验相关词的词条进行分词和筛选，获得所述待验相关词的对照词集
S12
当判定所述待验相关词的对照词集与所述待验相关词集的交集的绝对值大于筛选阈值时，所述待验相关词为所述输入词的下位相关词，获得下位相关词集
S13

图 2

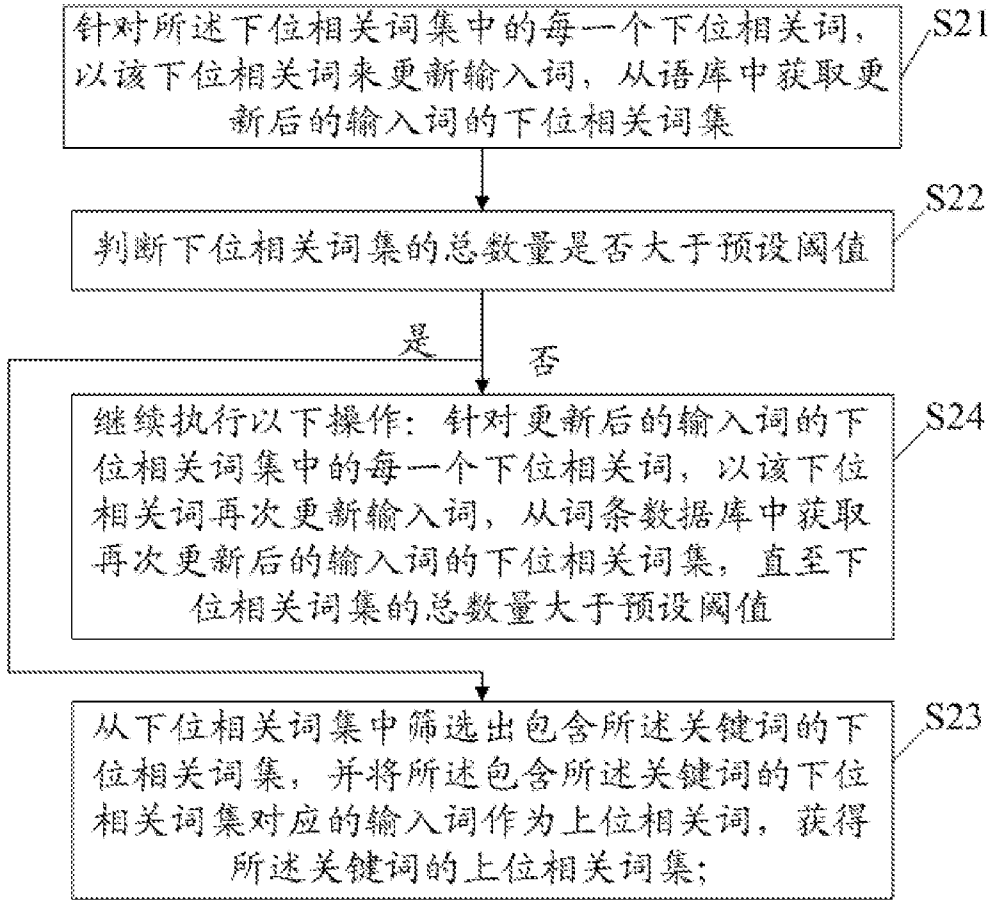

针对所述下位相关词集中的每一个下位相关词，以该下位相关词来更新输入词，从语库中获取更新后的输入词的下位相关词集
S21
判断下位相关词集的总数量是否大于预设阈值
S22
是
否
继续执行以下操作：针对更新后的输入词的下位相关词集中的每一个下位相关词，以该下位相关词再次更新输入词，从词条数据库中获取再次更新后的输入词的下位相关词集，直至下位相关词集的总数量大于预设阈值
S24
从下位相关词集中筛选出包含所述关键词的下位相关词集，并将所述包含所述关键词的下位相关词集对应的输入词作为上位相关词，获得所述关键词的上位相关词集；
S23

图 3

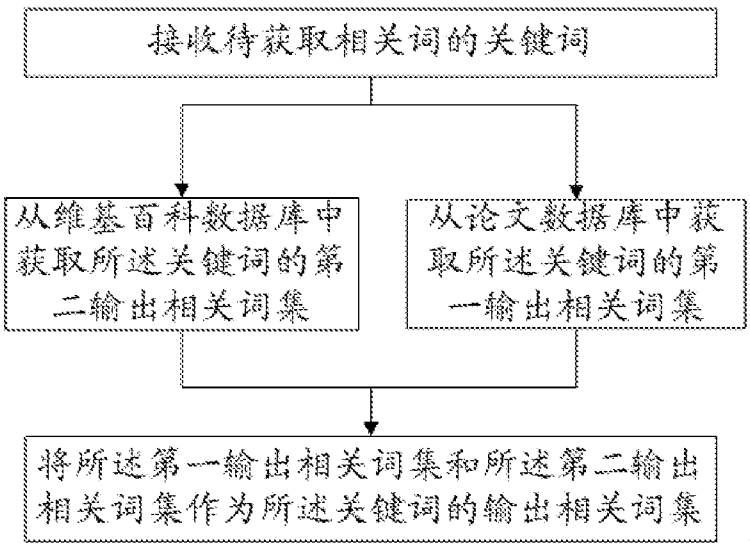

接收待获取相关词的关键词
从维基百科数据库中获取所述关键词的第二输出相关词集
从论文数据库中获取所述关键词的第一输出相关词集
将所述第一输出相关词集和所述第二输出相关词集作为所述关键词的输出相关词集

图 4

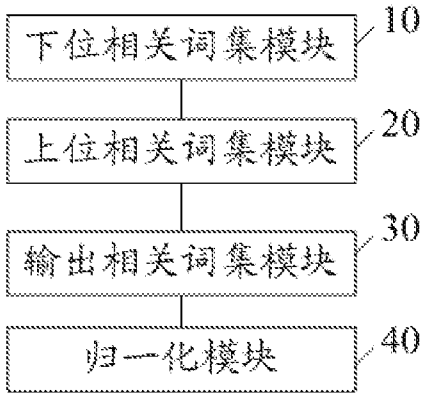

下位相关词集模块
10
上位相关词集模块
20
输出相关词集模块
30
归一化模块
40

图 5

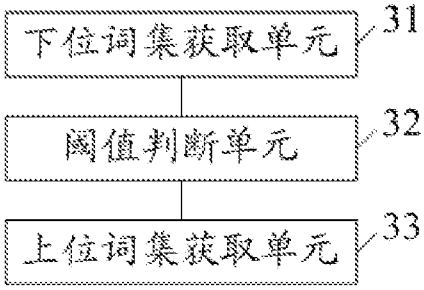

下位词集获取单元
31
阈值判断单元
32
上位词集获取单元
33

图 6


待验相关词集单元
1
对照词集单元
2
判断获取单元
3

图 7

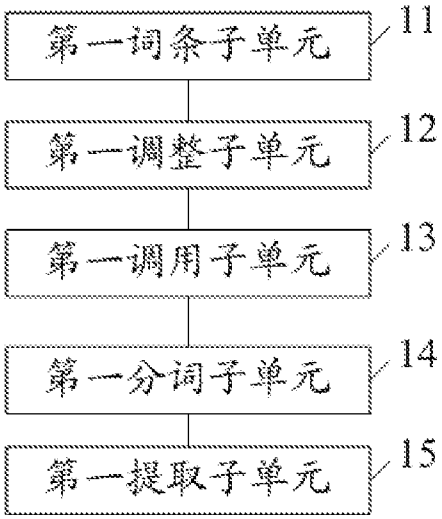

第一词条子单元
11
第一调整子单元
12
第一调用子单元
13
第一分词子单元
14
第一提取子单元
15

图 8

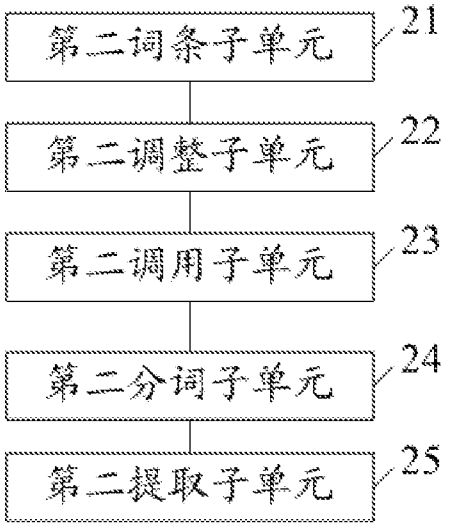

第二词条子单元
21
第二调整子单元
22
第二调用子单元
23
第二分词子单元
24
第二提取子单元
25

图 9

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2016/113175

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)

EPODOC, WPI, CNPAT, CNKI, IEEE, GOOGLE: key, correlat+, relat+, associat+, word?, recommend, expand, provid+, upper, lower, relevance, excavate

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 106126588 A (GUANGZHOU CVTE ELECTRONIC TECHNOLOGY CO., LTD.) 16 November 2016 (16.11.2016) c1aims 1-11	1-11
X	CN 103778262 A (BEIJING FORESTRY UNIVERSITY) 07 May 2014 (07.05.2014) c1aims 1-10	1, 5, 7
A	CN 104008097 A (NEC CHINA CO., LTD.) 27 August 2014 (27.08.2014) the who1e document	1-11
A	CN 101251844 A (FUJI FILM CO., LTD.) 27 August 2008 (27.08.2008) the who1e document	1-11
A	US 2012072443 A1 (INVENTEC CORPORATION) 22 March 2012 (22.03.2012) the who1e document	1-11

☐ 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	Date of mailing of the international search report
02 March 2017	29 March 2017
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 LIN, Wanjuan Telephone No. (86-10) 010-82246854

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2016/113175

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 106126588 A	16 November 2016	None	
CN 103778262 A	07 May 2014	None	
CN 104008097 A	27 August 2014	None	
CN 101251844 A	27 August 2008	JP 2008192110 A	21 August 2008
		US 2008201322 A1	21 August 2008
US 2012072443 A1	22 March 2012	TW 201214163 A	01 April 2012

国际检索报告

国际申请号
PCT/CN2016/113175

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））

EPODOC,WPI,CNPAT,CNKI,IEEE,GOOGLE:关键，相关，关联，联想，词，推荐，扩展，提供，挖掘，上位，下位，相关度，相似度，key, correlat+, relat+, associat+, word?, recommend, expand, provid+, upper, lower

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
PX	CN 106126588 A (广州视源电子科技股份有限公司) 2016年 11月 16日 (2016-11-16) 权利要求1-11	1-11
X	CN 103778262 A (北京林业大学) 2014年 5月 7日 (2014-05-07) 权利要求1-10	1,5,7
A	CN 104008097 A (日电中国有限公司) 2014年 8月 27日 (2014-08-27) 全文	1-11
A	CN 101251844 A (富士胶片株式会社) 2008年 8月 27日 (2008-08-27) 全文	1-11
A	US 2012072443 A1 (INVENTEC CORPORATION) 2012年 3月 22日 (2012-03-22) 全文	1-11

☐ 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	国际检索报告邮寄日期
2017年 3月 2日	2017年 3月 29日

ISA/CN的名称和邮寄地址	受权官员
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	林婉娟
传真号 (86-10)62019451	电话号码 (86-10)010-82246854

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2016/113175

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	106126588	A	2016年 11月 16日	无			
CN	103778262	A	2014年 5月 7日	无			
CN	104008097	A	2014年 8月 27日	无			
CN	101251844	A	2008年 8月 27日	JP	2008192110	A	2008年 8月 21日
				US	2008201322	A1	2008年 8月 21日
US	2012072443	A1	2012年 3月 22日	TW	201214163	A	2012年 4月 1日