

(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
2018 年 5 月 24 日 (24.05.2018)



(10) 国际公布号
WO 2018/090468 A1

(51) 国际专利分类号:
G06F 17/30 (2006.01)

(21) 国际申请号: PCT/CN2016/113642

(22) 国际申请日: 2016 年 12 月 30 日 (30.12.2016)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201611019485.4 2016 年 11 月 18 日 (18.11.2016) CN

(71) 申请人: 广州视源电子科技有限公司 (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, 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,

(54) Title: METHOD AND DEVICE FOR SEARCHING FOR VIDEO PROGRAM

(54) 发明名称: 视频节目的搜索方法和装置

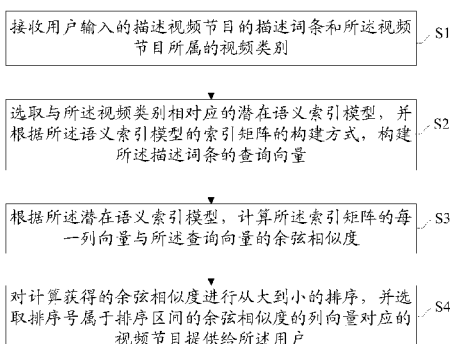


图 1

- S1 Receiving a description entry, input by a user, for describing a video program, and a video category to which the video program belongs
- S2 Selecting a potential semantic indexing model corresponding to the video category, and according to a construction mode of an index matrix of the semantic indexing model, constructing a query vector of the description entry
- S3 According to the potential semantic indexing model, calculating the cosine similarity between each column vector of the index matrix and the query vector
- S4 Sorting the cosine similarity obtained through calculation in a descending order, and selecting a video program corresponding to a column vector with the cosine similarity, of which a sorting number belongs to a sorting interval, and providing same to the user

(57) Abstract: Disclosed in the present invention is a method for searching for a video program. The method comprises: receiving a description entry, input by a user, for describing a video program, and a video category to which the video program belongs; selecting a potential semantic indexing model corresponding to the video category, and according to a construction mode of an index matrix of the semantic indexing model, constructing a query vector of the description entry; according to the potential semantic indexing model, calculating the cosine similarity between each column vector of the index matrix and the query vector; and sorting the cosine similarity obtained through calculation in a descending order, and selecting a video program corresponding to a column vector with the cosine similarity, of which a sorting number belongs to a sorting interval, and providing same to the user. Correspondingly, further disclosed in the present invention is a device for searching for a video program. By means of an embodiment of the present invention, a potential semantic meaning of a document can be mined, thereby improving the accuracy for searching for a video program and the search efficiency.

(57) 摘要: 本发明公开了一种视频节目的搜索方法, 包括: 接收用户输入的描述视频节目的描述词条和所述视频节目所属的视频类别; 选取与所述视频类别相对应的潜在语义索引模型, 并根据所述语义索引模型的索引矩阵的构建方式, 构建所述描述词条的查询向量; 根据所述潜在语义索引模型, 计算所述索引矩阵的每一列向量与所述查询向量的余弦相似度; 对计算获得的余弦相似度进行从大到小的排序, 并选取排序序号属于排序区间的余弦相似度的列向量对应的视频节目提供给所述用户。相应地, 本发明还公开了一种视频节目的搜索装置。采用本发明实施例, 能挖掘出文档的潜在语义, 提高搜索视频节目的准确度和搜索效率。

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))。

视频节目的搜索方法和装置

技术领域

本发明涉及计算机领域，尤其涉及视频节目的搜索方法和装置。

背景技术

在做综艺节目推荐时，ContentBase 方法是一种重要的策略，主要是通过综艺内容描述的相似度进行聚类推荐，这种方法将内容相近的文本进行了聚类，现有主要是基于 TF-IDF 的 Rocchio 算法，Rocchio 算法来源于向量空间模型理论，向量空间模型 Vector space model 的基本思想为采用向量来表示一个文本，之后的处理过程就可以转化为空间中向量的运算。Rocchio 算法训练的过程，其实就是建立类别特征向量的过程，对于给定的一个未知文本，生成该文本的向量，然后计算该向量与各类别特征向量的相似度，最后将该文本分到与其最相似的类别中去。

但是采用上述算法存在以缺点：Rocchio 算法无法挖掘文档的潜在语义。二、它假设训练数据是绝对正确的，因为它没有任何定量衡量样本是否含有噪声的机制，因而也就对错误数据毫无抵抗力。

发明内容

本发明实施例提出的一种视频节目的搜索方法和装置，能挖掘出文档的潜在语义，提高搜索视频节目的准确度和搜索效率。

本发明实施例提供的一种视频节目的搜索方法，包括：

接收用户输入的描述视频节目的描述词条和所述视频节目所属的视频类别；

选取与所述视频类别相对应的潜在语义索引模型，并根据所述语义索引模型的索引矩阵的构建方式，构建所述描述词条的查询向量；其中，所述潜在语义索引模型是对由描述同一视频类别的视频节目的描述文档所构建成的索引矩阵进行奇异值分解而获得的；

根据所述潜在语义索引模型，计算所述索引矩阵的每一列向量与所述查询向量的余弦相似度；

对计算获得的余弦相似度进行从大到小的排序，并选取排序号属于排序区间的余弦相似度的列向量对应的视频节目提供给所述用户。

进一步地, 由描述视频节目的描述文档构建索引矩阵的过程包括: 将第 i 个关键词在第 j 个视频节目的描述文档中出现的词频作为索引矩阵的第 j 列的第 i 个元素的数值;

构建所述描述词条的查询向量的过程包括: 设置所述查询向量的第 i 个元素代表的关键词与所述索引矩阵的第 i 行元素代表的关键词相同, 并将第 i 个元素对应的关键词在所述描述词条中出现的词频作为所述查询向量的第 i 个元素的数值; 其中, 所述查询向量为列向量。

进一步地, 由描述同一视频类别的视频节目的描述文档构建索引矩阵的过程, 具体为:

对于数据库存储的描述同一视频类别的视频节目的所有描述文档, 根据标准词条格式, 对所述所有描述文档包含的词条进行格式调整; 其中, 所述数据库存储有多种视频类别的描述文档, 一个描述文档描述一个视频节目, 不同的描述文档描述的视频节目互不相同;

调用分词工具;

利用所述分词工具对格式调整后的所述所有描述文档的词条进行分词, 获得第一词语集;

根据 TF-IDF 算法从所述第一词语集中提取关键词;

根据所提取的每一个关键词在每一个描述文档中出现的词频, 构建索引矩阵; 其中, 所述索引矩阵的行顺序是根据关键词在所述所有描述文档出现的总词频进行由高到低的排列, 所述索引矩阵的列顺序根据关键词在每一个描述文档中出现的词频进行由高到低的排列。

进一步地, 所述构建所述描述词条的查询向量, 具体为:

根据标准词条格式, 对所述描述词条进行格式调整;

调用分词工具;

利用所述分词工具对格式调整后的所述描述词条进行分词, 获得第二词语集;

根据 TF-IDF 算法从所述第二词语集中提取关键词;

根据所提取的每一个关键词在所述描述词条中出现的词频, 构建所述描述词条的查询向量。

进一步地, 所述索引矩阵为 H , 则对所述索引矩阵进行奇异值分解所获得的所述潜在语义索引模型为: $H = T * S * D^T$; 其中, T 为正交矩阵, 矩阵 T 的每一列是所述索引矩阵 H 的左奇异向量; S 为对角矩阵, 矩阵 S 的对角线元素是所述索引矩阵 H 的奇异值; D 为正交矩阵, 矩阵 D 的每一列为所述索引矩阵 H 的右奇异向量; 所述查询向量为 Q ;

所述根据所述潜在语义索引模型, 计算所述索引矩阵的每一列向量与所述查询向量的余弦相似度, 具体为:

选取 T_K 、 S_K 和 D_K 矩阵, 修订所述潜在语义索引模型为 $H_K = T_K * S_K * D_K^T$; 其中, T_K 为由矩阵 T 的前 K 列形成的矩阵, S_K 为由矩阵 S 的前 K 个对角线元素形成的对角矩阵, D_K 为由

矩阵 D 的前 K 列形成的矩阵； K 的数值大于所述排序区间包含的最大排序号；

对于修订后的所述潜在语义索引模型的索引矩阵 H_K ，计算所述查询向量的转置矩阵 Q^T 与所述矩阵 T_K 相乘所得的行向量和所述矩阵 D_K 与所述矩阵 S_K 相乘所得矩阵的第 j 行向量的两行向量之间的余弦相似度，作为所述索引矩阵 H_K 的第 j 列向量与所述查询向量 Q 的余弦相似度。

进一步地，所述搜索方法还包括：

当数据库增加描述新的视频节目的描述文档时，对与所述新的视频节目所属的视频类别相对应的潜在语义索引模型进行更新。

相应地，本发明实施例提供一种视频节目的搜索装置，包括：

用户信息接收模块，用于接收用户输入的描述视频节目的描述词条和所述视频节目所属的视频类别；

查询向量构建模块，用于选取与所述视频类别相对应的潜在语义索引模型，并根据所述语义索引模型的索引矩阵的构建方式，构建所述描述词条的查询向量；其中，所述潜在语义索引模型是对由描述同一视频类别的视频节目的描述文档所构建成的索引矩阵进行奇异值分解而获得的；

相似度计算模块，用于根据所述潜在语义索引模型，计算所述索引矩阵的每一列向量与所述查询向量的余弦相似度；

视频节目选取模块，用于对计算获得的余弦相似度进行从大到小的排序，并选取排序号属于排序区间的余弦相似度的列向量对应的视频节目提供给所述用户。

进一步地，所述查询向量构建模块包括的用于根据描述视频节目的描述文档构建索引矩阵的单元，具体用于：将第 i 个关键词在第 j 个视频节目的描述文档中出现的词频作为索引矩阵的第 j 列的第 i 个元素的数值；

所述查询向量构建模块包括的用于构建描述词条的查询向量的单元，具体用于：设置所述查询向量的第 i 个元素代表的关键词与所述索引矩阵的第 i 行元素代表的关键词相同，并将第 i 个元素对应的关键词在所述描述词条中出现的词频作为所述查询向量的第 i 个元素的数值；其中，所述查询向量为列向量。

进一步地，所述查询向量构建模块包括用于根据描述同一视频类别的视频节目的描述文档构建索引矩阵的单元，具体为：

第一格式调整单元，用于对于数据库存储的描述同一视频类别的视频节目的所有描述文档，根据标准词条格式，对所述所有描述文档包含的词条进行格式调整；其中，所述数据库

存储有多种视频类别的描述文档，一个描述文档描述一个视频节目，不同的描述文档描述的视频节目互不相同；

第一工具调用单元，用于调用分词工具；

第一分词单元，用于利用所述分词工具对格式调整后的所述所有描述文档的词条进行分词，获得第一词语集；

第一关键词提取单元，用于根据 TF-IDF 算法从所述第一词语集中提取关键词；

索引矩阵构建单元，用于根据所提取的每一个关键词在每一个描述文档中出现的词频，构建索引矩阵；其中，所述索引矩阵的行顺序是根据关键词在所述所有描述文档出现的总词频进行由高到低的排列，所述索引矩阵的列顺序根据关键词在每一个描述文档中出现的词频进行由高到低的排列。

进一步地，所述查询向量构建模块还包括用于构建所述描述词条的查询向量的单元，具体为：

第二格式调整单元，用于根据标准词条格式，对所述描述词条进行格式调整；

第二工具调用单元，用于调用分词工具；

第二分词单元，用于利用所述分词工具对格式调整后的所述描述词条进行分词，获得第二词语集；

第二关键词提取单元，用于根据 TF-IDF 算法从所述第二词语集中提取关键词；

查询向量构建单元，用于根据所提取的每一个关键词在所述描述词条中出现的词频，构建所述描述词条的查询向量。

进一步地，所述索引矩阵为 H ，则对所述索引矩阵进行奇异值分解所获得的所述潜在语义索引模型为： $H=T*S*D^T$ ；其中， T 为正交矩阵，矩阵 T 的每一列是所述索引矩阵 H 的左奇异向量； S 为对角矩阵，矩阵 S 的对角线元素是所述索引矩阵 H 的奇异值； D 为正交矩阵，矩阵 D 的每一列为所述索引矩阵 H 的右奇异向量；所述查询向量为 Q ；

所述相似度计算模块具体包括：

模型修订单元，用于选取 T_K 、 S_K 和 D_K 矩阵，修订所述潜在语义索引模型为 $H_K=T_K*S_K*D_K^T$ ；其中， T_K 为由矩阵 T 的前 K 列形成的矩阵， S_K 为由矩阵 S 的前 K 个对角线元素形成的对角矩阵， D_K 为由矩阵 D 的前 K 列形成的矩阵； K 的数值大于所述排序区间包含的最大排序号；

计算单元，用于对于修订后的所述潜在语义索引模型的索引矩阵 H_K ，计算所述查询向量的转置矩阵 Q^T 与所述矩阵 T_K 相乘所得的行向量和所述矩阵 D_K 与所述矩阵 S_K 相乘所得矩阵的第 j 行向量的两行向量之间的余弦相似度，作为所述索引矩阵 H_K 的第 j 列向量与所述查询

向量 Q 的余弦相似度。

进一步地，所述搜索装置还包括：

模型更新模块，用于当数据库增加描述新的视频节目的描述文档时，对与所述新的视频节目所属的视频类别相对应的潜在语义索引模型进行更新。

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视频节目的搜索方法和装置，通过计算要搜索视频的查询向量与潜在语义索引模型的索引矩阵的每一列向量的余弦相似度，可获得要搜索视频的描述词条与索引矩阵的每一列向量代表的描述文档之间的相关程度，数值越高，则相关程度越高，进而将与该描述词条相关程度高的描述文档所对应的视频节目推荐给用户，并由于潜在语义索引模型是根据描述视频节目的描述文档构建（训练）成的，能挖掘出文档的潜在语义，提高搜索视频节目的准确度。另外，通过用户输入的所述视频节目所属的视频类别，选择与该视频类别对应的潜在语义索引模型来进行计算，能进一步提高搜索视频节目的效率。

附图说明

图 1 是本发明提供的视频节目的搜索方法的一个实施例的流程示意图；

图 2 是本发明提供的视频节目的搜索装置的一个实施例的结构示意图；

图 3 是本发明提供的视频节目的搜索装置的查询向量构建模块的一个实施例的结构示意图；

图 4 是本发明提供的视频节目的搜索装置的相似度计算模块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，是本发明提供的视频节目的搜索方法的一个实施例的流程示意图；该搜索方法，包括步骤 S1 至 S4，具体为：

S1，接收用户输入的描述视频节目的描述词条和所述视频节目所属的视频类别；

S2，选取与所述视频类别相对应的潜在语义索引模型，并根据所述语义索引模型的索引矩阵的构建方式，构建所述描述词条的查询向量；其中，所述潜在语义索引模型是对由描述

同一视频类别的视频节目的描述文档所构建成的索引矩阵进行奇异值分解而获得的；所述索引矩阵的第 j 列的第 i 个元素的数值代表第 i 个关键词在第 j 个视频节目的描述文档中出现的词频；所述查询向量为列向量，所述查询向量的第 i 个元素代表的关键词与所述索引矩阵的第 i 行元素代表的关键词相同，且所述查询向量的第 i 个元素的数值代表所述第 i 个元素对应的关键词在所述描述词条中出现的词频；

S3，根据所述潜在语义索引模型，计算所述索引矩阵的每一列向量与所述查询向量的余弦相似度；

S4，对计算获得的余弦相似度进行从大到小的排序，并选取排序号属于排序区间的余弦相似度的列向量对应的视频节目提供给所述用户。

需要说明的是，通过计算要搜索视频的查询向量与潜在语义索引模型的索引矩阵的每一列向量的余弦相似度，可获得要搜索视频的描述词条与索引矩阵的每一列向量代表的描述文档之间的相关程度，数值越高，则相关程度越高，进而将与该描述词条相关程度高的描述文档所对应的视频节目推荐给用户，并由于潜在语义索引模型是根据描述视频节目的描述文档构建（训练）成的，能挖掘出文档的潜在语义，提高搜索视频节目的准确度。另外，通过用户输入的所述视频节目所属的视频类别，选择与该视频类别对应的潜在语义索引模型来进行计算，能进一步提高搜索视频节目的效率。其中，上述的排序区间一般优选为排列在前的 10 个排序号。

进一步地，上述步骤 S2 中的根据描述同一视频类别的视频节目的描述文档构建索引矩阵的过程，具体为：

对于数据库存储的描述同一视频类别的视频节目的所有描述文档，根据标准词条格式，对所述所有描述文档包含的词条进行格式调整；其中，所述数据库存储有多种视频类别的描述文档，一个描述文档描述一个视频节目，不同的描述文档描述的视频节目互不相同；对于对词条的格式调整，可以但不限于，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所述所有描述文档的词条进行分词，获得第一词语集；分词工具对描述词条进行分词的模式有多种，除了按正常分词模式切分外，还可以继续长词进行切分，提高召回率，尤其对短文本，可以切出比正常切分出更多的词，对后续的输出视频节目的准确度有提升效果。

根据 TF-IDF 算法从所述第一词语集中提取关键词；

根据所提取的每一个关键词在每一个描述文档中出现的词频，构建索引矩阵；其中，所

述索引矩阵的行顺序是根据关键词在所述所有描述文档出现的总词频进行由高到低的排列，所述索引矩阵的列顺序根据关键词在每一个描述文档中出现的词频进行由高到低的排列。

需要说明的是，构建上述索引矩阵是预先根据数据库存储的描述文档构建而成的，构建过程需遵循：索引矩阵的第 j 列的第 i 个元素的数值代表第 i 个关键词在第 j 个视频节目的描述文档中出现的词频。其中，索引矩阵的第 i 行的所有元素所代表的同一个关键词，且不同行的元素所代表的关键词不相同。例如，假设索引矩阵的第 1 行的所有元素代表关键词 A，索引矩阵的第 1 列的元素代表描述文档 B，则该索引矩阵的第 1 行第 1 列的元素的数值代表关键词 A 在描述文档 B 出现的概率。

进一步地，上述步骤 S2 中的构建所述描述词条的查询向量，具体为：

根据标准词条格式，对所述描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所述描述词条进行分词，获得第二词语集；分词工具对描述词条进行分词的模式有多种，除了按正常分词模式切分外，还可以继续长词进行切分，提高召回率，尤其对短文本，可以切出比正常切分出更多的词，对后续的输出视频节目的准确度有提升效果。

根据 TF-IDF 算法从所述第二词语集中提取关键词；

根据所提取的每一个关键词在所述描述词条中出现的词频，构建所述描述词条的查询向量。

需要说明的是，构建所述描述词条的查询向量时，要确保所述查询向量的第 i 个元素代表的关键词与上述潜在语义索引模型的索引矩阵的第 i 行元素代表的关键词相同，使得比较查询向量与索引矩阵的每一列向量的余弦相似度具有意义。

另外，构建向量的过程还需遵循以下原则：所述查询向量的第 i 个元素代表的关键词与所述索引矩阵的第 i 行元素代表的关键词相同，且所述查询向量的第 i 个元素的数值代表所述第 i 个元素对应的关键词在所述描述词条中出现的词频；例如，在假设索引矩阵的第 1 行的所有元素代表关键词 A，则查询向量的第 1 行的元素代表的关键词为关键词 A，则查询向量的第 1 行的元素的数值代表关键词 A 在描述词条中出现的词频。

进一步地，所述索引矩阵为 H ，则对所述索引矩阵进行奇异值分解所获得的所述潜在语义索引模型为： $H=U*S*V^T$ ；其中， U 为正交矩阵，矩阵 U 的每一列是所述索引矩阵 H 的左奇异向量； S 为对角矩阵，矩阵 S 的对角线元素是所述索引矩阵 H 的奇异值； V 为正交矩阵，

矩阵 D 的每一列为所述索引矩阵 H 的右奇异向量；所述查询向量为 Q；

上述步骤 S3 的具体实施过程具体为：

选取 T_K 、 S_K 和 D_K 矩阵，修订所述潜在语义索引模型为 $H_K = T_K * S_K * D_K^T$ ；其中， T_K 为由矩阵 T 的前 K 列形成的矩阵， S_K 为由矩阵 S 的前 K 个对角线元素形成的对角矩阵， D_K 为由矩阵 D 的前 K 列形成的矩阵；K 的数值大于所述排序区间包含的最大排序号；

对于修订后的所述潜在语义索引模型的索引矩阵 H_K ，计算所述查询向量的转置矩阵 Q^T 与所述矩阵 T_K 相乘所得的行向量和所述矩阵 D_K 与所述矩阵 S_K 相乘所得矩阵的第 j 行向量的两行向量之间的余弦相似度，作为所述索引矩阵 H_K 的第 j 列向量与所述查询向量 Q 的余弦相似度。

需要说明的是，此处的 K 值是个阈值选择，可以根据实际情况选择，分解过程采用 H 的 K 秩，是让索引矩阵 H 的前 K 个最大奇异值以后的奇异值都为零。上述对潜在语义索引模型的修订，能够提高检索效率。

进一步地，所述搜索方法还包括：

当数据库增加描述新的视频节目的描述文档时，对与所述新的视频节目所属的视频类别相对应的潜在语义索引模型进行更新。

需要说明的是，由于视频节目会不断增加的，而对于描述新增加的视频节目的描述文档也会不断添加到数据库当中，因此需要对潜在语义索引模型进行更新。

本发明实施例提供的视频节目的搜索方法，通过计算要搜索视频的查询向量与潜在语义索引模型的索引矩阵的每一列向量的余弦相似度，可获得要搜索视频的描述词条与索引矩阵的每一列向量代表的描述文档之间的相关程度，数值越高，则相关程度越高，进而将与该描述词条相关程度高的描述文档所对应的视频节目推荐给用户，并由于潜在语义索引模型是根据描述视频节目的描述文档构建（训练）成的，能挖掘出文档的潜在语义，提高搜索视频节目的准确度。另外，通过用户输入的所述视频节目所属的视频类别，选择与该视频类别对应的潜在语义索引模型来进行计算，能进一步提高搜索视频节目的效率。

参阅图 2，是本发明提供的视频节目的搜索装置的一个实施例的结构示意图。该搜索装置能够执行上述实施例提供的视频节目的搜索方法的全部流程，该搜索装置，包括：

用户信息接收模块 10，用于接收用户输入的描述视频节目的描述词条和所述视频节目所属的视频类别；

查询向量构建模块 20，用于选取与所述视频类别相对应的潜在语义索引模型，并根据所述潜在语义索引模型的索引矩阵的构建方式，构建所述描述词条的查询向量；其中，所述潜在语

义索引模型是对由描述同一视频类别的视频节目的描述文档所构建成的索引矩阵进行奇异值分解而获得的;

相似度计算模块 30, 用于根据所述潜在语义索引模型, 计算所述索引矩阵的每一列向量与所述查询向量的余弦相似度;

视频节目选取模块 40, 用于对计算获得的余弦相似度进行从大到小的排序, 并选取排序号属于排序区间的余弦相似度的列向量对应的视频节目提供给所述用户。

进一步地, 所述查询向量构建模块包括的用于根据描述视频节目的描述文档构建索引矩阵的单元, 具体用于: 将第 i 个关键词在第 j 个视频节目的描述文档中出现的词频作为索引矩阵的第 j 列的第 i 个元素的数值;

所述查询向量构建模块包括的用于构建描述词条的查询向量的单元, 具体用于: 设置所述查询向量的第 i 个元素代表的关键词与所述索引矩阵的第 i 行元素代表的关键词相同, 并将第 i 个元素对应的关键词在所述描述词条中出现的词频作为所述查询向量的第 i 个元素的数值; 其中, 所述查询向量为列向量。

进一步地, 参见图 3, 是本发明提供的视频节目的搜索装置的查询向量构建模块的一个实施例的结构示意图, 所述查询向量构建模块 20 包括用于根据描述同一视频类别的视频节目的描述文档构建索引矩阵的单元, 具体为:

第一格式调整单元 21, 用于对于数据库存储的描述同一视频类别的视频节目的所有描述文档, 根据标准词条格式, 对所述所有描述文档包含的词条进行格式调整; 其中, 所述数据库存储有多种视频类别的描述文档, 一个描述文档描述一个视频节目, 不同的描述文档描述的视频节目互不相同;

第一工具调用单元 22, 用于调用分词工具;

第一分词单元 23, 用于利用所述分词工具对格式调整后的所述所有描述文档的词条进行分词, 获得第一词语集;

第一关键词提取单元 34, 用于根据 TF-IDF 算法从所述第一词语集中提取关键词;

索引矩阵构建单元 25, 用于根据所提取的每一个关键词在每一个描述文档中出现的词频, 构建索引矩阵; 其中, 所述索引矩阵的行顺序是根据关键词在所述所有描述文档出现的总词频进行由高到低的排列, 所述索引矩阵的列顺序根据关键词在每一个描述文档中出现的词频进行由高到低的排列。

进一步地, 所述查询向量构建模块 20 还包括用于构建所述描述词条的查询向量的单元, 具体为:

第二格式调整单元 26, 用于根据标准词条格式, 对所述描述词条进行格式调整;

第二工具调用单元 27, 用于调用分词工具;

第二分词单元 28, 用于利用所述分词工具对格式调整后的所述描述词条进行分词, 获得第二词语集;

第二关键词提取单元 29, 用于根据 TF-IDF 算法从所述第二词语集中提取关键词;

查询向量构建单元 31, 用于根据所提取的每一个关键词在所述描述词条中出现的词频, 构建所述描述词条的查询向量。

进一步地, 参见图 4, 是本发明提供的视频节目的搜索装置的相似度计算模块的一个实施例的结构示意图, 所述索引矩阵为 H , 则对所述索引矩阵进行奇异值分解所获得的所述潜在语义索引模型为: $H=T*S*D^T$; 其中, T 为正交矩阵, 矩阵 T 的每一列是所述索引矩阵 H 的左奇异向量; S 为对角矩阵, 矩阵 S 的对角线元素是所述索引矩阵 H 的奇异值; D 为正交矩阵, 矩阵 D 的每一列为所述索引矩阵 H 的右奇异向量; 所述查询向量为 Q ;

所述相似度计算模块 30 具体包括:

模型修订单元 32, 用于选取 T_K 、 S_K 和 D_K 矩阵, 修订所述潜在语义索引模型为 $H_K=T_K*S_K*D_K^T$; 其中, T_K 为由矩阵 T 的前 K 列形成的矩阵, S_K 为由矩阵 S 的前 K 个对角线元素形成的对角矩阵, D_K 为由矩阵 D 的前 K 列形成的矩阵; K 的数值大于所述排序区间包含的最大排序号;

计算单元 33, 用于对于修订后的所述潜在语义索引模型的索引矩阵 H_K , 计算所述查询向量的转置矩阵 Q^T 与所述矩阵 T_K 相乘所得的行向量和所述矩阵 D_K 与所述矩阵 S_K 相乘所得矩阵的第 j 行向量的两行向量之间的余弦相似度, 作为所述索引矩阵 H_K 的第 j 列向量与所述查询向量 Q 的余弦相似度。

进一步地, 所述搜索装置还包括:

模型更新模块 50, 用于当数据库增加描述新的视频节目的描述文档时, 对与所述新的视频节目所属的视频类别相对应的潜在语义索引模型进行更新。

本发明实施例提供的视频节目的搜索装置, 通过计算要搜索视频的查询向量与潜在语义索引模型的索引矩阵的每一列向量的余弦相似度, 可获得要搜索视频的描述词条与索引矩阵的每一列向量代表的描述文档之间的相关程度, 数值越高, 则相关程度越高, 进而将与该描述词条相关程度高的描述文档所对应的视频节目推荐给用户, 并由于潜在语义索引模型是根据描述视频节目的描述文档构建 (训练) 成的, 能挖掘出文档的潜在语义, 提高搜索视频节目的准确度。另外, 通过用户输入的所述视频节目所属的视频类别, 选择与该视频类别对应的潜在语义索引模型来进行计算, 能进一步提高搜索视频节目的效率。

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、一种视频节目的搜索方法，其特征在于，包括：

接收用户输入的描述视频节目的描述词条和所述视频节目所属的视频类别；

选取与所述视频类别相对应的潜在语义索引模型，并根据所述语义索引模型的索引矩阵的构建方式，构建所述描述词条的查询向量；其中，所述潜在语义索引模型是对由描述同一视频类别的视频节目的描述文档所构建成的索引矩阵进行奇异值分解而获得的；

根据所述潜在语义索引模型，计算所述索引矩阵的每一列向量与所述查询向量的余弦相似度；

对计算获得的余弦相似度进行从大到小的排序，并选取排序号属于排序区间的余弦相似度的列向量对应的视频节目提供给所述用户。

2、如权利要求1所述的视频节目的搜索方法，其特征在于，

由描述视频节目的描述文档构建索引矩阵的过程包括：将第*i*个关键词在第*j*个视频节目的描述文档中出现的词频作为索引矩阵的第*j*列的第*i*个元素的数值；

构建所述描述词条的查询向量的过程包括：设置所述查询向量的第*i*个元素代表的关键词与所述索引矩阵的第*i*行元素代表的关键词相同，并将第*i*个元素对应的关键词在所述描述词条中出现的词频作为所述查询向量的第*i*个元素的数值；其中，所述查询向量为列向量。

3、如权利要求1或2所述的视频节目的搜索方法，其特征在于，由描述同一视频类别的视频节目的描述文档构建索引矩阵的过程，具体为：

对于数据库存储的描述同一视频类别的视频节目的所有描述文档，根据标准词条格式，对所述所有描述文档包含的词条进行格式调整；其中，所述数据库存储有多种视频类别的描述文档，一个描述文档描述一个视频节目，不同的描述文档描述的视频节目互不相同；

调用分词工具；

利用所述分词工具对格式调整后的所述所有描述文档的词条进行分词，获得第一词语集；

根据 TF-IDF 算法从所述第一词语集中提取关键词；

根据所提取的每一个关键词在每一个描述文档中出现的词频，构建索引矩阵；其中，所述索引矩阵的行顺序是根据关键词在所述所有描述文档出现的总词频进行由高到低的排列，所述索引矩阵的列顺序根据关键词在每一个描述文档中出现的词频进行由高到低的排列。

4、如权利要求 1 或 2 所述的视频节目的搜索方法，其特征在于，所述构建所述描述词条的查询向量，具体为：

根据标准词条格式，对所述描述词条进行格式调整；

调用分词工具；

利用所述分词工具对格式调整后的所述描述词条进行分词，获得第二词语集；

根据 TF-IDF 算法从所述第二词语集中提取关键词；

根据所提取的每一个关键词在所述描述词条中出现的词频，构建所述描述词条的查询向量。

5、如权利要求 3 所述的视频节目的搜索方法，其特征在于，所述索引矩阵为 H ，则对所述索引矩阵进行奇异值分解所获得的所述潜在语义索引模型为： $H=T*S*D^T$ ；其中， T 为正交矩阵，矩阵 T 的每一列是所述索引矩阵 H 的左奇异向量； S 为对角矩阵，矩阵 S 的对角线元素是所述索引矩阵 H 的奇异值； D 为正交矩阵，矩阵 D 的每一列为所述索引矩阵 H 的右奇异向量；所述查询向量为 Q ；

所述根据所述潜在语义索引模型，计算所述索引矩阵的每一列向量与所述查询向量的余弦相似度，具体为：

选取 T_K 、 S_K 和 D_K 矩阵，修订所述潜在语义索引模型为 $H_K=T_K*S_K*D_K^T$ ；其中， T_K 为由矩阵 T 的前 K 列形成的矩阵， S_K 为由矩阵 S 的前 K 个对角线元素形成的对角矩阵， D_K 为由矩阵 D 的前 K 列形成的矩阵； K 的数值大于所述排序区间包含的最大排序号；

对于修订后的所述潜在语义索引模型的索引矩阵 H_K ，计算所述查询向量的转置矩阵 Q^T 与所述矩阵 T_K 相乘所得的行向量和所述矩阵 D_K 与所述矩阵 S_K 相乘所得矩阵的第 j 行向量的两行向量之间的余弦相似度，作为所述索引矩阵 H_K 的第 j 列向量与所述查询向量 Q 的余弦相似度。

6、如权利要求 1 所述的视频节目的搜索方法，其特征在于，所述搜索方法还包括：

当数据库增加描述新的视频节目的描述文档时，对与所述新的视频节目所属的视频类别相对应的潜在语义索引模型进行更新。

7、一种视频节目的搜索装置，其特征在于，包括：

用户信息接收模块，用于接收用户输入的描述视频节目的描述词条和所述视频节目所属的视频类别；

查询向量构建模块，用于选取与所述视频类别相对应的潜在语义索引模型，并根据所述语义索引模型的索引矩阵的构建方式，构建所述描述词条的查询向量；其中，所述潜在语义索引模型是对由描述同一视频类别的视频节目的描述文档所构建成的索引矩阵进行奇异值分解而获得的；

相似度计算模块，用于根据所述潜在语义索引模型，计算所述索引矩阵的每一列向量与所述查询向量的余弦相似度；

视频节目选取模块，用于对计算获得的余弦相似度进行从大到小的排序，并选取排序号属于排序区间的余弦相似度的列向量对应的视频节目提供给所述用户。

8、如权利要求 7 所述的视频节目的搜索装置，其特征在于，

所述查询向量构建模块包括的用于根据描述视频节目的描述文档构建索引矩阵的单元，具体用于：将第 i 个关键词在第 j 个视频节目的描述文档中出现的词频作为索引矩阵的第 j 列的第 i 个元素的数值；

所述查询向量构建模块包括的用于构建描述词条的查询向量的单元，具体用于：设置所述查询向量的第 i 个元素代表的关键词与所述索引矩阵的第 i 行元素代表的关键词相同，并将第 i 个元素对应的关键词在所述描述词条中出现的词频作为所述查询向量的第 i 个元素的数值；其中，所述查询向量为列向量。

9、如权利要求 7 或 8 所述的视频节目的搜索装置，其特征在于，所述查询向量构建模块包括用于根据描述同一视频类别的视频节目的描述文档构建索引矩阵的单元，具体为：

第一格式调整单元，用于对于数据库存储的描述同一视频类别的视频节目的所有描述文档，根据标准词条格式，对所述所有描述文档包含的词条进行格式调整；其中，所述数据库存储有多种视频类别的描述文档，一个描述文档描述一个视频节目，不同的描述文档描述的 video 节目互不相同；

第一工具调用单元，用于调用分词工具；

第一分词单元，用于利用所述分词工具对格式调整后的所述所有描述文档的词条进行分词，获得第一词语集；

第一关键词提取单元，用于根据 TF-IDF 算法从所述第一词语集中提取关键词；

索引矩阵构建单元，用于根据所提取的每一个关键词在每一个描述文档中出现的词频，构建索引矩阵；其中，所述索引矩阵的行顺序是根据关键词在所述所有描述文档出现的总词频进行由高到低的排列，所述索引矩阵的列顺序根据关键词在每一个描述文档中出现的词频

进行由高到低的排列。

10、如权利要求 7 或 8 所述的视频节目的搜索装置，其特征在于，所述查询向量构建模块还包括用于构建所述描述词条的查询向量的单元，具体为：

第二格式调整单元，用于根据标准词条格式，对所述描述词条进行格式调整；

第二工具调用单元，用于调用分词工具；

第二分词单元，用于利用所述分词工具对格式调整后的所述描述词条进行分词，获得第二词语集；

第二关键词提取单元，用于根据 TF-IDF 算法从所述第二词语集中提取关键词；

查询向量构建单元，用于根据所提取的每一个关键词在所述描述词条中出现的词频，构建所述描述词条的查询向量。

11、如权利要求 9 所述的视频节目的搜索装置，其特征在于，所述索引矩阵为 H ，则对所述索引矩阵进行奇异值分解所获得的所述潜在语义索引模型为： $H=T*S*D^T$ ；其中， T 为正交矩阵，矩阵 T 的每一列是所述索引矩阵 H 的左奇异向量； S 为对角矩阵，矩阵 S 的对角线元素是所述索引矩阵 H 的奇异值； D 为正交矩阵，矩阵 D 的每一列为所述索引矩阵 H 的右奇异向量；所述查询向量为 Q ；

所述相似度计算模块具体包括：

模型修订单元，用于选取 T_K 、 S_K 和 D_K 矩阵，修订所述潜在语义索引模型为 $H_K=T_K*S_K*D_K^T$ ；其中， T_K 为由矩阵 T 的前 K 列形成的矩阵， S_K 为由矩阵 S 的前 K 个对角线元素形成的对角矩阵， D_K 为由矩阵 D 的前 K 列形成的矩阵； K 的数值大于所述排序区间包含的最大排序号；

计算单元，用于对于修订后的所述潜在语义索引模型的索引矩阵 H_K ，计算所述查询向量的转置矩阵 Q^T 与所述矩阵 T_K 相乘所得的行向量和所述矩阵 D_K 与所述矩阵 S_K 相乘所得矩阵的第 j 行向量的两行向量之间的余弦相似度，作为所述索引矩阵 H_K 的第 j 列向量与所述查询向量 Q 的余弦相似度。

12、如权利要求 7 所述的视频节目的搜索装置，其特征在于，所述搜索装置还包括：

模型更新模块，用于当数据库增加描述新的视频节目的描述文档时，对与所述新的视频节目所属的视频类别相对应的潜在语义索引模型进行更新。

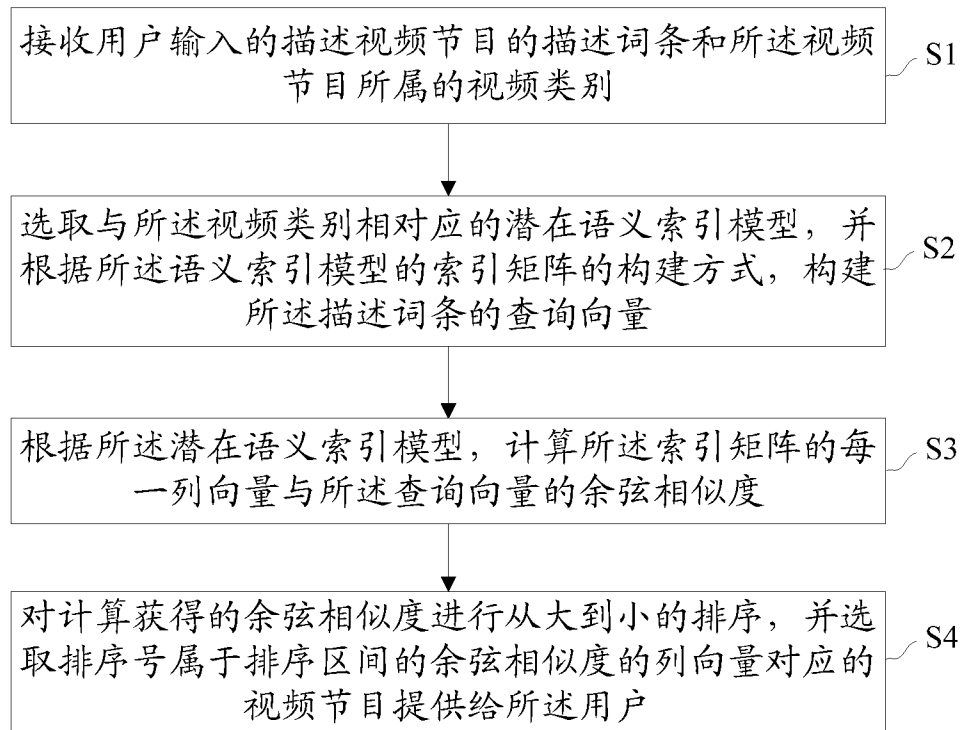


图 1



图 2

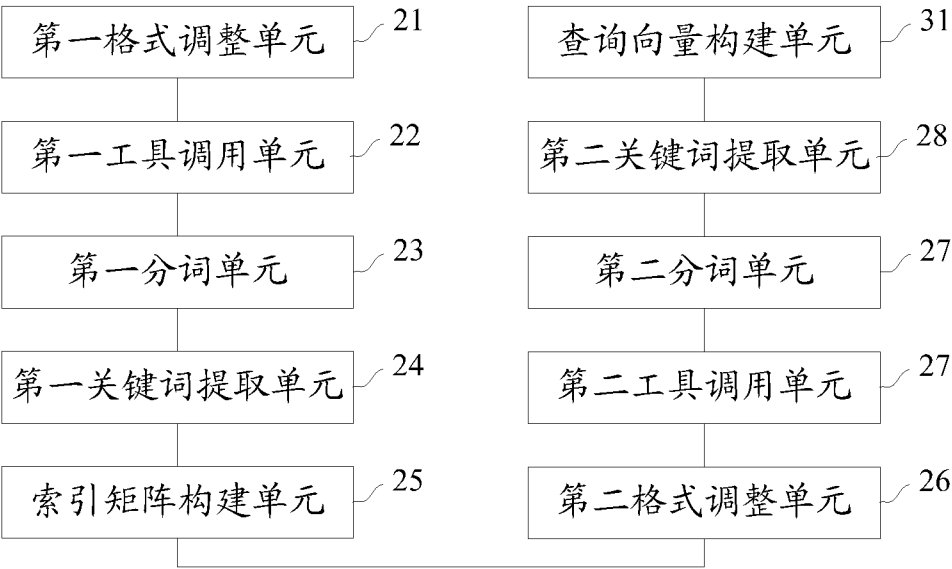


图 3

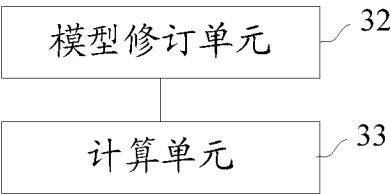


图 4

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2016/113642

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)

CNPAT, CNKI: 视频, 节目, 图像, 检索, 搜索, 矩阵, 奇异值, 余弦, 相似度;

WPI, EPODOC, IEEE: video, program, image, picture, search, matrix, singular value, SVD, similarity

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 103559196 A (ZHEJIANG UNIVERSITY), 05 February 2014 (05.02.2014), description, paragraphs 0037-0065	1-12
Y	伍春江, “基于聚类分析的潜在语义文献检索”, 中国优秀硕士学位论文全文数据库(电子期刊), 30 November 2013 (30.11.2013), the main body, pages 5-6 and 9-17, (WU, Chunjinag, Latent Semantic Retrieval Based on Document Clustering Analysis, Electronic Technology & Information Science, China Master's Theses Full-Text Database (Electronic Journals))	1-12
Y	王振风, “基于 Lucene 的分布式全文检索技术的研究与应用”, 中国优秀硕士学位论文全文数据库(电子期刊), 31 July 2015 (31.07.2015), the main body, pages 25 and 29, (WANG, Zhenfeng, The Research and Application of Full-text Retrieval Technology Based on Lucene, Electronic Technology & Information Science, China Master's Theses Full-Text Database (Electronic Journals))	3-5, 9-11
A	CN 103152618 A (BEIJING STARTIMES SOFTWARE TECHNOLOGY CO., LTD.), 12 June 2013 (12.06.2013), the whole document	1-12
A	US 6189002 B1 (DOLPHIN SEARCH), 13 February 2001 (13.02.2001), the whole document	1-12

☐ 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 24 February 2017	Date of mailing of the international search report 30 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 LIU, Chang Telephone No. (86-10) 62414452

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2016/113642

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 103559196 A	05 February 2014	None	
CN 103152618 A	12 June 2013	None	
US 6189002 B1	13 February 2001	AU 782200 B2	07 July 2005
		CA 2393480 A1	14 June 2001
		WO 0142984 A1	14 June 2001
		AU 7089200 A	18 June 2001
		EP 1247207 A1	09 October 2002

国际检索报告

国际申请号

PCT/CN2016/113642

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))

CNPAT, CNKI: 视频, 节目, 图像, 检索, 搜索, 矩阵, 奇异值, 余弦, 相似度;
program, image, picture, search, matrix, singular value, SVD, similarity

WPI, EPDOC, IEEE: video,

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
Y	CN 103559196 A (浙江大学) 2014年 2月 5日 (2014 - 02 - 05) 说明书第0037-0065段	1-12
Y	伍春江, “基于聚类分析的潜在语义文献检索” 中国优秀硕士学位论文全文数据库(电子期刊), 2013年 11月 30日 (2013 - 11 - 30), 正文第5-6、9-17页	1-12
Y	王振风, “基于 Lucene 的分布式全文检索技术的研究与应用” 中国优秀硕士学位论文全文数据库(电子期刊), 2015年 7月 31日 (2015 - 07 - 31), 正文第25和29页	3-5、9-11
A	CN 103152618 A (北京四达时代软件技术股份有限公司) 2013年 6月 12日 (2013 - 06 - 12) 全文	1-12
A	US 6189002 B1 (DOLPHIN SEARCH) 2001年 2月 13日 (2001 - 02 - 13) 全文	1-12

☐ 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

2017年 2月 24日

国际检索报告邮寄日期

2017年 3月 30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

受权官员

刘昶

传真号 (86-10)62019451

电话号码 (86-10)62414452

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2016/113642

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	103559196	A	2014年 2月 5日	无			
CN	103152618	A	2013年 6月 12日	无			
US	6189002	B1	2001年 2月 13日	AU	782200	B2	2005年 7月 7日
				CA	2393480	A1	2001年 6月 14日
				WO	0142984	A1	2001年 6月 14日
				AU	7089200	A	2001年 6月 18日
				EP	1247207	A1	2002年 10月 9日

表 PCT/ISA/210 (同族专利附件) (2009年7月)