

(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 5 月 18 日 (18.05.2017)



(10) 国际公布号
WO 2017/080320 A1

- (51) 国际专利分类号:
G06F 17/30 (2006.01)
- (21) 国际申请号: PCT/CN2016/099894
- (22) 国际申请日: 2016 年 9 月 23 日 (23.09.2016)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201510760110.2 2015 年 11 月 9 日 (09.11.2015) CN
- (71) 申请人: 北京奇虎科技有限公司 (BEIJING QIHOO TECHNOLOGY COMPANY LIMITED) [CN/CN]; 中国北京市西城区新街口外大街 28 号 D 座 112 室 (德胜园区), Beijing 100088 (CN)。奇智软件 (北京) 有限公司 (QIZHI SOFTWARE (BEIJING) COMPANY LIMITED) [CN/CN]; 中国北京市朝阳区酒仙桥路 6 号院 2 号楼 B 座 2 层、3 层 301-306 室, Beijing 100015 (CN)。
- (72) 发明人: 张超 (ZHANG, Chao); 中国北京市朝阳区酒仙桥路 6 号院 2 号楼, Beijing 100015 (CN)。
- (74) 代理人: 北京市隆安律师事务所 (BEIJING LONGAN LAW FIRM); 中国北京市朝阳区建国门

外大街 21 号北京国际俱乐部 188 室, Beijing 100020 (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, 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, 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: METHOD OF MINING AND CLEANING UP SIMILAR BOOKS IN BOOK DATABASE, AND DEVICE UTILIZING SAME

(54) 发明名称: 书籍库中相似书的挖掘、净化方法和装置

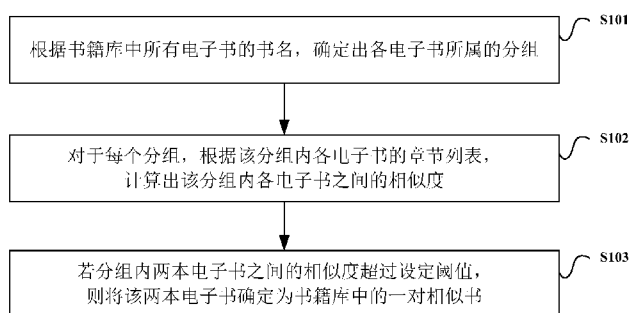


图 1

- S101 DETERMINE, ACCORDING TO BOOK TITLES OF ALL OF ELECTRONIC BOOKS IN A BOOK DATABASE, A GROUP TO WHICH EACH OF THE ELECTRONIC BOOKS BELONGS
- S102 WITH RESPECT TO EACH OF GROUPS, COMPUTE, ACCORDING TO A CHAPTER LIST OF EACH OF ELECTRONIC BOOKS IN EACH OF THE GROUPS, SIMILARITY BETWEEN ELECTRONIC BOOKS IN EACH OF THE GROUPS
- S103 IF SIMILARITY BETWEEN TWO ELECTRONIC BOOKS IN EACH OF THE GROUPS EXCEEDS A PRECONFIGURED THRESHOLD, DETERMINE THAT THE TWO ELECTRONIC BOOKS IS A PAIR OF SIMILAR BOOKS IN THE BOOK DATABASE

(57) Abstract: The method provides a method of mining and cleaning up similar books in a book database, and a device utilizing the same. The method comprises: determining, according to book titles of all of electronic books in a book database, a group to which each of the electronic books belongs; with respect to each of groups, computing, according to a chapter list of each of electronic books in each of the groups, similarity between electronic books in each of the groups; if similarity between two electronic books in each of the groups exceeds a preconfigured threshold, determining that the two electronic books is a pair of similar books in the book database; constructing, by employing similar books in all of the groups in the book database, a graph model of the book database, wherein each pair of similar books are two mutually connected vertices in the graph model; with respect to each of connected components in the graph model, selecting one electronic book to be kept from the connected components, and deleting the other electronic book. Applying the invention can increase a speed of mining similar books from a book database, and reduce a size of the book database.

(57) 摘要:

[见续页]



WO 2017/080320 A1

本发明提供了一种书籍库中相似书的挖掘、净化方法和装置，该净化方法包括：根据书籍库中所有电子书的书名，确定出各电子书所属的分组；对于每个分组，根据该分组内各电子书的章节列表，计算出该分组内各电子书之间的相似度，若该分组内两本电子书之间的相似度超过设定阈值，则将该两本电子书确定为书籍库中的一对相似书；利用书籍库中所有分组内的相似书，构建书籍库的图模型，其中，每对相似书为图模型中相连的两个端点；对于图模型中的每个连通分量，从该连通分量中选择一本电子书保留，删除其他电子书。应用本发明，可以提高书籍库中相似书的挖掘速度，降低书籍库的规模。

书籍库中相似书的挖掘、净化方法和装置

技术领域

5 本发明涉信息处理技术领域，具体而言，本发明涉及一种书籍库中相似书的挖掘、净化方法和装置。

背景技术

10 随着移动互联网的快速发展，传统的书籍逐渐被电子书所代替。通过 PC(Personal Computer, 个人计算机)、手机、阅读器等智能设备，用户可以上网搜索自己感兴趣的电子书，或者将自己喜欢的电子书存储到智能设备中。

目前，智能设备上可以支持各种阅读平台和阅读工具来满足用户的电子书阅读需求，比如，起点中文网、言情小说吧、360 小说、书旗小说、ireader 阅读器等等。而这些阅读平台和阅读工具通常会配有用于存储有大量的电子书的电子书数据库（本文中也可以称为书籍库）。

15 实际应用中，在大规模的书籍库（比如，百万级别）中，存在大量的书名、作者、章节列表及章节内容大致相同的相似书，这些相似书将占用大量的存储空间。而且，书籍库中电子书的信息不一定完整，例如，有的电子书的书名因为网页解析问题出现乱码或额外后缀，有的电子书缺少作者信息，有的电子书相对于完整的电子书来讲缺少一些章节列表，有的电子书的章节列表由于某些原因是错误的。

20 因此，有必要从书籍库中挖掘出这些相似书，并从中删除一些错误的或缺章少节的电子书，进行书籍库净化处理。现有可以针对相同书名的电子书，使用章节内容信息来进行匹配确定出相似书。虽然，通过章节内容信息进行匹配的方法来挖掘相似书的准确率高，但是，由于书籍库中电子书的章节内容存储量太大，会存在挖掘速度慢的问题。

因此，有必要提供一种快速挖掘出书籍库中相似书的方法，并对书籍库中的相似书进行净化处理，以此净化书籍库，降低书籍库的规模。

发明内容

25 鉴于上述问题，提出了本发明以便提供一种克服上述问题或者至少部分地解决上述问题的一种书籍库中相似书的挖掘、净化方法和装置。

本发明方案提供了一种书籍库中相似书的净化方法，包括：

根据所述书籍库中所有电子书的书名，确定出各电子书所属的分组；

30 对于每个分组，根据该分组内各电子书的章节列表，计算出该分组内各电子书之间的相似度，若该分组内两本电子书之间的相似度超过设定阈值，则将该两本电子书确定为所述书籍库中的一对相似书；

利用所述书籍库中所有分组内的相似书，构建所述书籍库的图模型，其中，每对相似书为所述图模型中相连的两个端点；

35 对于所述图模型中的每个连通分量，从该连通分量中选择一本电子书保留，删除其他电子书。

根据本发明的另一方面，还提供了一种书籍库中相似书的净化装置，包括：

书籍分组单元，用于根据所述书籍库中所有电子书的书名，确定出各电子书所属的分组；

40 相似书挖掘单元，用于对于每个分组，根据该分组内各电子书的章节列表，计算出该分组内各电子书之间的相似度；若该分组内两本电子书之间的相似度超过设定阈值，则将该两本电子书确定为所述书籍库中的一对相似书；

相似书净化单元，用于利用所述书籍库中所有分组内的相似书，构建所述书籍库的图模型，其中，每对相似书为所述图模型中相连的两个端点；对于所述书籍库的图模型中的每个连通分量，从连通分量中选择一本电子书保留，删除其他电子书。

45 本发明还提供了一种书籍库中相似书的挖掘方法，包括：

根据书籍库中所有电子书的书名，确定出各电子书所属的分组；

对于每个分组，根据该分组内各电子书的章节列表，计算出该分组内各电子书之间的相似度；

5 对于每个分组，若分组内两本电子书之间的相似度超过设定阈值，则将该两本电子书确定为所述书籍库中的一对相似书。

本发明还提供了一种书籍库中相似书的挖掘装置，包括：

书籍分组单元，用于根据所述书籍库中所有电子书的书名，确定出各电子书所属的分组；

相似度计算单元，用于对于每个分组，根据该分组内各电子书的章节列表，计算出该分组内各电子书之间的相似度；

10 相似书挖掘单元，用于对于每个分组，若该分组内两本电子书之间的相似度超过设定阈值，则将该两本电子书确定为所述书籍库中的一对相似书。

本发明还提供了一种基于净化的书籍库的电子书推送方法，包括：

接收包含电子书的书名的用户搜索请求；

15 从根据上述书籍库中相似书的净化方法所净化的书籍库中查询与所述书名相对应的电子书；

响应于所述用户搜索请求推送查询到的电子书的摘要信息页面。

本发明还提供了一种基于净化的书籍库的电子书推送装置，包括：

请求接收单元，用于接收包含电子书的书名的用户搜索请求；

20 电子书查询单元，用于获取所述用户搜索请求中的电子书的书名，从根据上述书籍库中相似书的净化方法所净化的书籍库中查询与所述书名相对应的电子书；

推送单元，用于响应于所述用户搜索请求推送查询到的电子书的摘要信息页面。

25 本发明还提供了一种计算机程序，包括计算机可读代码，当所述计算机可读代码在计算设备上运行时，导致所述计算设备执行如上所述的书籍库中相似书的净化方法，或者导致所述计算设备执行如上所述的书籍库中相似书的挖掘方法，或者导致所述计算设备执行如上所述的基于净化的书籍库的电子书推送方法。

本发明还提供了一种计算机可读介质，其中存储了如权利要求 33 所述的计算机程序。

30 本发明的技术方案中，可以根据电子书的书名，将书籍库中的电子书进行分组，即将书名相似的电子书归为一个分组；继而，对于每个分组，可以根据该分组内各电子书的章节列表，计算出该分组内各电子书之间的相似度，将相似度超过设定阈值的两本电子书确定为书籍库的一对相似书。这样，通过书名和章节列表的两次比对来挖掘相似书，可以避免大量的章节内容的比对，在保障了相似书挖掘的准确率的同时，可以减少挖掘工作量，提高了书籍库中相似书的挖掘速度。

进一步地，在挖掘出书籍库中的相似书之后，还可以基于挖掘出所有分组内的相似书，对书籍库进行相似书的净化处理，删除一些重复的内容，以此降低书籍库的规模。

35 本发明附加的方面和优点将在下面的描述中部分给出，这些将从下面的描述中变得明显，或通过本发明的实践了解到。

上述说明仅是本发明技术方案的概述，为了能够更清楚了解本发明的技术手段，而可依照说明书的内容予以实施，并且为了让本发明的上述和其它目的、特征和优点能够更明显易懂，以下特举本发明的具体实施方式。

40 附图说明

通过阅读下文优选实施方式的详细描述，各种其他的优点和益处对于本领域普通技术人员将变得清楚明了。附图仅用于示出优选实施方式的目的，而并不认为是对本发明的限制。而且在整个附图中，用相同的参考符号表示相同的部件。在附图中：

图 1 为本发明实施例的书籍库中相似书的挖掘方法的流程示意图；

45 图 2a、2b、2c 均为本发明实施例的书籍库中相似书的挖掘装置的内部结构示意图；

图 3 为本发明实施例的书籍库中相似书的净化方法的流程示意图；

图 4 为本发明实施例的书籍库中相似书的净化装置的内部结构示意图；

图 5 为本发明实施例的基于净化的书籍库的电子书推送方法的流程示意图；

图 6 为本发明实施例的基于净化的书籍库的电子书推送装置的内部结构示意图；

5 图 7 示意性地示出了用于执行根据本发明的方法的计算设备的框图；以及

图 8 示意性地示出了用于保持或者携带实现根据本发明的方法的程序代码的存储单元。

具体实施例

10 下面详细描述本发明的实施例，所述实施例的示例在附图中示出，其中自始至终相同或类似的标号表示相同或类似的元件或具有相同或类似功能的元件。下面通过参考附图描述的实施例是示例性的，仅用于解释本发明，而不能解释为对本发明的限制。

本技术领域技术人员可以理解，除非特意声明，这里使用的单数形式“一”、“一个”、“所述”和“该”也可包括复数形式。应该进一步理解的是，本发明的说明书中使用的措辞“包括”是指存在所述特征、整数、步骤、操作、元件和/或组件，但是并不排除存在或添加一个或多个其他特征、整数、步骤、操作、元件、组件和/或它们的组。应该理解，当我们称元件被“连接”或“耦接”到另一元件时，它可以直接连接或耦接到其他元件，或者也可以存在中间元件。此外，这里使用的“连接”或“耦接”可以包括无线连接或无线耦接。这里使用的措辞“和/或”包括一个或多个相关联的列出项的全部或任一单元和全部组合。

20 本技术领域技术人员可以理解，除非另外定义，这里使用的所有术语（包括技术术语和科学术语），具有与本发明所属领域中的普通技术人员的一般理解相同的意义。还应该理解的是，诸如通用字典中定义的那些术语，应该被理解为具有与现有技术的上下文中的意义一致的意义，并且除非像这里一样被特定定义，否则不会用理想化或过于正式的含义来解释。

本技术领域技术人员可以理解，这里所使用的“终端”、“终端设备”既包括无线信号接收器的设备，其仅具备无发射能力的无线信号接收器的设备，又包括接收和发射硬件的设备，其具有能够在双向通信链路上，进行双向通信的接收和发射硬件的设备。这种设备可以包括：蜂窝或其他通信设备，其具有单线路显示器或多线路显示器或多线路显示器的蜂窝或其他通信设备；PCS（Personal Communications Service，个人通信系统），其可以组合语音、数据处理、传真和/或数据通信能力；PDA（Personal Digital Assistant，个人数字助理），其可以包括射频接收器、寻呼机、互联网/内联网访问、网络浏览器、记事本、日历和/或 GPS（Global Positioning System，全球定位系统）接收器；常规膝上型和/或掌上型计算机或其他设备，其具有和/或包括射频接收器的常规膝上型和/或掌上型计算机或其他设备。这里所使用的“终端”、“终端设备”可以是便携式、可运输、安装在交通工具（航空、海运和/或陆地）中的，或者适合于和/或配置为在本地运行，和/或以分布形式，运行在地球和/或空间的任何其他位置运行。这里所使用的“终端”、“终端设备”还可以是通信终端、上网终端、音乐/视频播放终端，例如 35 可以是 PDA、MID（Mobile Internet Device，移动互联网设备）和/或具有音乐/视频播放功能的移动电话，也可以是智能电视、机顶盒等设备。

本发明的发明人发现，实际应用中，在两本电子书的作者、书名和章节列表的相似度比较高的情况下，就可以判定该两本电子书针对的是同一本书，即该两本电子书为一对相似书。

因此，本发明的发明人考虑，可以基于书籍库中的电子书的书名、以及电子书的章节列表，确定出书籍库中的相似书。具体地，可以根据电子书的书名，将书籍库中的电子书进行 40 分组，即将书名相似的电子书归为一个分组；继而，对于每个分组，可以根据该分组内各电子书的章节列表，计算出该分组内各电子书之间的相似度，将相似度超过设定阈值的两本电子书确定为书籍库的一对相似书。这样，通过书名和章节列表的两次比对来挖掘相似书，可以避免大量的章节内容的比对，在保障了相似书挖掘的准确率的同时，可以减少挖掘工作量， 45 提高了挖掘速度。

进一步地，在挖掘出书籍库中的相似书之后，还可以基于挖掘出所有分组内的相似书，对书籍库进行相似书的净化处理，删除一些重复的内容，以此降低书籍库的规模。

本发明所称的电子书，是指适于人们阅读的数字化出版物，区别于以纸张为载体的传统出版物。电子书通过数码方式记录在以光、电、磁为介质的设备中，必须借助于特定的电子阅读设备来读取、复制和传输。简言之，电子书通常以计算机文件的形式存在，其主要格式有 PDF、EXE、CHM、UMD、PDG、JAR、PDB、TXT、BRM、EPUB、MOBI 等等，进一步的，还可以扩展到泛指一切利用电子阅读设备上的电子阅读软件提供给人们进行阅读的形式，例如，在 APP 中提供一个网页浏览插件，通过预设的链接访问指向某个书城网站的一本电子书的网页，也同样能提供给用户相同或相近的阅读体验。因而，本领域技术人员不应对本发明的电子书做局限性理解，应以本发明的各个步骤所实现的功能的视角来确定。

下面结合附图详细说明本发明的技术方案。

本发明的方案中，在进行书籍库中的相似书的挖掘之前，可以对书籍库中的电子书进行预处理。

具体地，针对电子书的书名可以通过如下方式进行预处理：利用预设的书名过滤规则，对书籍库中所有电子书的书名进行过滤处理，得到电子书的书名字符串。

实际应用中，书籍库中电子书的书名通常由中文字符、阿拉伯数字、英文字母以及一些空字符组成，空格、制表符等。因此，为了方便后续的书名的比对，本发明的方案中，可以通过预先设置的书名过滤规则，将电子书的书名中的一些非必要的空字符过滤掉，保留英文字母、阿拉伯数据和中文字符，并将保留的字符组成该电子书的书名字符串。

针对电子书的章节列表可以通过如下方式进行预处理：利用预设的章节过滤规则，对书籍库中所有电子书的章节列表进行过滤处理，得到电子书的章节列表中的各章节标题。

实际应用中，书籍库中电子书的章节列表中通常包括一些常用前缀内容，比如“第 X 章”、“第 X 节”等，还包括一些常用的后缀内容，比如，页码、完/上/中/下等标识。

因此，本发明的方案中，章节过滤规则可以包括如下至少之一：前缀过滤规则、后缀过滤规则。

实际应用中，电子书的章节列表中“第一章 XXXX”中的“第一章”，“第一节 XXXX”中的“第一节”、“第一 XXXX”中的“第一”、“一 XXXX”中的“一”等等，均可以通过前缀过滤规则进行过滤。其中，“一”只是一个实例，可以替换为对应的阿拉伯数字，也可以替换为其他的数字或数字对应的中文。具体地，可以预先针对上述预过滤的前缀内容的数据格式，设置相应的前缀过滤规则及其对应的正则表达式。这样，在对书籍库中所有电子书的章节列表进行过滤处理过程中，可以通过前缀过滤规则的正则表达式，将电子书的章节列表中的前缀内容提取出。

而电子书的章节列表中“XXXX (1)”中的“(1)”、“XXXX (一)”中的“(一)”、“XXXX 1”中的“1”、“XXXX (完)”中的“(完)”、“XXXX (上)”中的“(上)”、“XXXX (下)”中的“(下)”等等，均可以通过后缀过滤规则进行过滤。其中，“一”只是一个实例，可以替换为对应的阿拉伯数字，也可以替换为其他的数字或数字对应的中文。具体地，可以预先针对上述预过滤的后缀内容的数据格式，设置相应的后缀过滤规则及其对应的正则表达式。这样，在对书籍库中所有电子书的章节列表进行后缀过滤处理过程中，可以通过后缀过滤规则的正则表达式，将电子书的章节列表中的后缀内容提取出。

更优地，考虑到章节列表中通常还包含了一些空字符，因此，在对电子书的章节列表进行预处理的过程中，还可以将章节列表中各章节标题的空字符删掉。其中，空字符可以包括：空格、制表符等。

进一步地，考虑到章节列表中通常还包含了标点符号，因此，在对电子书的章节列表进行预处理的过程中，还可以将章节列表中的标点符号去掉。

这样，通过上述空字符、标点符号、前缀过滤规则、后缀过滤规则对书籍库中电子书的

章节列表进行过滤处理，将可以得到电子书的章节列表中的各章节标题。

基于上述预处理之后的电子书，本发明提供了一种书籍库中相似书的挖掘方法，如图 1 所示，其流程具体包括如下步骤：

S101: 根据书籍库中所有电子书的书名，确定出各电子书所属的分组。

5 具体地，针对书籍库中的每本电子书，可以利用预设的向量化规则拆分该电子书的书名字符串，由拆分得到的各子字符串构成该电子书对应的书名向量。将拆分得到的所有子字符串进行合并去重，得到各分组的关键词。针对书籍库中的每本电子书，对于每个分组，若该电子书的书名向量中包含该分组的关键词，则将该电子书划分到该分组中。

10 其中，电子书的书名字符串可以通过预设的书名过滤规则对电子书的书名进行过滤处理得到。

实际应用中，通过向量化规则，可以将书名字符串中相邻的 N 个字符组成一个子字符串，之后将拆分得到的各子字符串构成该电子书对应的书名向量，其中，N 为预设的拆分基数，可以具体为 1、2、3……等正整数。

15 例如，以 3 作为拆分基数，电子书的书名字符串“异世之我是传奇”的书名向量为“异世之世之我 之我是 我是传 是传奇”。其中，书名向量中的向量元素“异世之”、“世之我”、“之我是”、“我是传”、“是传奇”均是电子书的书名字符串拆分得到的子字符串。实际应用中，若电子书的书名字符串的书名长度小于拆分基数（比如 3），则可以不进行拆分，而是将该书名字符串直接构成该电子书的书名向量。

20 本发明的方案中，在得到书籍库中所有电子书各自对应的书名向量之后，可以将所有电子书所对应的书名向量合并成一个向量。之后，根据合并成的向量中的向量元素进行分组，并确定各分组的关键词。

具体地，合并成的向量中的每个向量元素，可以对应一个分组，该分组的关键词即为该向量元素。例如，将三百万本书的所有书名向量合并成一个向量，合并成的向量中向量元素可能只有一百万，即合并成的向量的长度为一百万。

25 实际应用中，在对合并成的向量进行分组的过程中，分组的个数可以等于该合并成的向量的长度，也可以小于该合并成的向量的长度。

在完成合并成的向量的分组之后，可以将书籍库中所有的电子书进行分组划分。具体地，可以针对每本电子书，将该电子书的书名向量与各个分组的关键词进行比对，若该书名向量中包含了某个分组的关键词，则可以将该电子书划分到该分组中。实际应用中，若电子书的书名向量中包含了多个分组的关键词，因此，该电子书将被划分到多个分组中。

30 S102: 对于每个分组，根据该分组内各电子书的章节列表，计算出该分组内各电子书之间的相似度。

本发明的方案中，两本电子书之间的相似度是通过两本书的章节列表之间的相似度来体现的。因此，对于每个分组，可以根据该分组内各电子书的章节列表，计算出该分组内各电子书之间的相似度。

35 实际应用中，可以以二次规划的最优化问题来计算两个章节列表之间的相似度，但是，由于二次规划的最优化问题计算过程中需要生成一个矩阵，而在大规模的书籍库中电子书的数量庞大，因此，矩阵的生成过程就会消耗很多时间，不利于挖掘速度的提高。

因此，本发明的方案中可以采用近似最优的方法，将分组内的电子书两两进行比对，得到比对的兩本电子书之间的相似度，提高相似度的计算速度，从而提高相似书的挖掘速度。具体地，对于每个分组，可以针对该分组内的任意两本电子书，将该两本电子书各自的章节列表中章节标题进行比对，计算出该两本电子书之间的 jaccard（杰卡德）系数，并作为两本电子书之间的相似度。

45 其中，章节列表的章节标题可以通过预设的章节过滤规则对电子书的章节列表进行过滤处理得到。关于 jaccard 系数的计算方法可以采用本领域技术人员常用的技术手段，在此不再

赘述。

S103: 若分组内两本电子书之间的相似度超过设定阈值, 则将该两本电子书确定为书籍库中的一对相似书。

5 具体地, 对于每个分组, 通过步骤 S102 计算出该分组内各电子书之间的相似度之后, 若分组内两本电子书之间的相似度超过设定阈值, 则将该两本电子书确定为书籍库中的一对相似书。

其中, 设定阈值是由本领域技术人员预先设置的, 比如, 该设定阈值可以设置为 0.8。这样, 在将分组内的电子书的章节列表进行两两比对之后, 若比对的两本电子书之间的相似度超过 0.8, 则说明该两本电子书的章节列表相似, 可以确定为该分组内的一对相似书。

10 本发明的方案中, 在通过步骤 103 确定出各分组内相似度超过设定阈值的各对相似书之后, 即可挖掘出了书籍库中所有的相似书。相比现有基于电子书的章节内容进行比对来挖掘相似书, 本发明的方案中通过书名和章节列表的比对来挖掘相似书, 在保障了相似书挖掘的准确率的同时, 还可以减少挖掘工作量, 提高了挖掘速度。

15 进一步地, 本发明的方案中, 在挖掘出书籍库中的相似书之后, 可以基于挖掘出的相似书对书籍库进行净化处理。因此, 本发明提供的书籍库中相似书的挖掘方法中, 在完成步骤 S103 之后, 还可以利用书籍库中所有分组内的相似书, 构建书籍库的图模型。

其中, 每对相似书为图模型中相连的两个端点。

为了便于描述, 本发明的方案中, 可以将书籍库中的一对相似书称为书籍库中的一个相似对。

20 实际应用中, 在挖掘出书籍库中所有分组内的相似书之后, 可以利用所有的相似对, 构建图模型。在该图模型中, 每个相似对将作为一条边的两个端点。这样, 通过构建的图模型可以将书籍库中所有的相似对中相关联的相似对连接起来。

进而, 对于构建的图模型中的每个连通分量, 可以从该连通分量中选择一本电子书保留, 删除其他电子书。

25 实际应用中, 如果从一个端点到另一个端点有路径, 则称这两个端点连通。而如果图中任意两个端点之间都连通, 则可以称该图为连通图, 否则, 称该图为非连通图, 则其中的极大连通子图称为连通分量, 这里所谓的极大是指子图中包含的端点个数极大。

因此, 本发明的方案中, 在构建了图模型之后, 可以确定出该图模型中存在的连通分量。实际应用中, 每个连通分量其所包含的所有电子书, 其针对的往往是同一本书。因此, 为了降低书籍库的规模, 对于图模型中确定出的每个连通分量, 可以从该连通分量的内部寻找最好最完整的电子书进行保留而删除其他电子书。

30 例如, 书籍库中挖掘出了 1 万个相似对, 利用这 1 万个相似对构建出图模型之后, 若从该图模型中寻找出了 1000 个连通分量, 而每个连通分量中平均有 10 本可以互相成为一对相似书的电子书。为了降低书籍库的规模, 可以从该 10 本电子书中寻找出最好的一本进行保留, 而删除其他 9 本电子书。

具体地, 对于每个连通分量, 可以针对该连通分量中的每本电子书, 将该连通分量中与该电子书相连的电子书的个数, 作为该电子书的关联度; 之后, 选取关联度最高的电子书保留, 而删除该连通分量中其他的电子书。

40 或者, 对于每个连通分量, 可以针对该连通分量中的每本电子书, 将该连通分量中该电子书相连的各本电子书与该电子书之间的相似度进行累加, 得到该电子书的关联度; 选取关联度最高的电子书保留。

45 进一步地, 针对该连通分量中的每本电子书, 在确定出该连通分量中与该电子书相连的电子书之后, 将各本电子书与该电子书之间的相似度进行累加, 并将累加的相似度除以与该电子书相连的电子书的数量, 得到该电子书的平均相似度, 并作为该电子书的关联度; 选取关联度最高的电子书保留。

本发明的方案中，在挖掘出书籍库中的相似书之后，还可以基于挖掘出所有分组内的相似书，构建图模型；进而针对图模型中每个连通分量，从该连通分量中选择一个电子书保留，而删除一些相似书，从而实现对书籍库进行相似书的净化处理，以此降低书籍库的规模。

5 基于上述书籍库中相似书的挖掘方法，本发明还提供了一种书籍库中相似书的挖掘装置，如图 2a 所示，该挖掘装置包括：书籍分组单元 201、相似度计算单元 202、以及相似书挖掘单元 203。

其中，书籍分组单元 201 用于根据书籍库中所有电子书的书名，确定出各电子书所属的分组。

10 相似度计算单元 202 用于对于每个分组，根据该分组内各电子书的章节列表，计算出该分组内各电子书之间的相似度。

相似书挖掘单元 203 用于对于每个分组，若该分组内两本电子书之间的相似度超过设定阈值，则将该两本电子书确定为书籍库中的一对相似书。

进一步地，如图 2b 所示，本发明提供的书籍库中相似书的挖掘装置中除了包括：书籍分组单元 201、相似度计算单元 202、以及相似书挖掘单元 203，还可以包括：预处理单元 204。

15 预处理单元 204 用于利用预设的书名过滤规则，对书籍库中所有电子书的书名进行过滤处理，得到电子书的书名字符串；利用预设的章节过滤规则，对书籍库中所有电子书的章节列表进行过滤处理，得到电子书的章节列表中的各章节标题。

其中，章节过滤规则包括如下至少之一：前缀过滤规则、后缀过滤规则。

20 相应地，书籍分组单元 201 针对书籍库中的每本电子书，可以利用预设的向量化规则拆分该电子书的书名字符串，由拆分得到的各子字符串构成该电子书对应的书名向量；将拆分得到的所有子字符串进行合并去重，得到各分组的关键词；针对书籍库中的每本电子书，对于每个分组，若该电子书的书名向量中包含该分组的关键词，则将该电子书划分到该分组中。

25 相似度计算单元 202 对于每个分组，可以针对该分组内的任意两本电子书，将该两本电子书各自的章节列表中章节标题进行比对，计算出该两本电子书之间的 jaccard 系数，并作为两本电子书之间的相似度。

更优地，如图 2c 所示，本发明提供的书籍库中相似书的挖掘装置中除了包括：书籍分组单元 201、相似度计算单元 202、相似书挖掘单元 203、以及预处理单元 204，还可以进一步包括：书籍净化单元 205。

30 书籍净化单元 205 用于利用相似书挖掘单元 203 确定出的书籍库中所有分组内的相似书，构建书籍库的图模型；对于图模型中的每个连通分量，从连通分量中选择一本电子书保留，删除其他电子书。其中，每对相似书为图模型中相连的两个端点。

具体地，书籍净化单元 205 可以利用书籍库中所有分组内的相似书，构建书籍库的图模型；对于图模型中的每个连通分量，针对连通分量中的每本电子书，将连通分量中与该电子书相连的电子书的个数，作为该电子书的关联度；选取关联度最高的电子书保留，删除其他

35 电子书。

或者，书籍净化单元 205 可以利用书籍库中所有分组内的相似书，构建书籍库的图模型；对于图模型中的每个连通分量，针对连通分量中的每本电子书，将连通分量中该电子书相连的各本电子书与该电子书之间的相似度进行累加，得到该电子书的关联度；选取关联度最高的电子书保留，删除其他电子书。

40 进一步地，书籍净化单元 205 可以针对该连通分量中的每本电子书，在确定出该连通分量中与该电子书相连的电子书之后，将各本电子书与该电子书之间的相似度进行累加，并将累加的相似度除以与该电子书相连的电子书的数量，得到该电子书的平均相似度，并作为该电子书的关联度；选取关联度最高的电子书保留。

45 本发明的方案中，书籍库中相似书的挖掘装置中的各单元的具体实现可参照上述书籍库中相似书的挖掘方法中的各步骤，在此不再赘述。

基于上述书籍库中相似书的挖掘方法和挖掘装置，本发明还提供了一种书籍库中相似书的净化方法，如图3所示，其流程具体包括如下步骤：

S301：根据书籍库中所有电子书的书名，确定出各电子书所属的分组。

具体地，针对书籍库中的每本电子书，利用预设的向量化规则拆分该电子书的书名字符串，由拆分得到的各子字符串构成该电子书对应的书名向量。之后，将拆分得到的所有子字符串进行合并去重，得到各分组的关键词。针对书籍库中的每本电子书，对于每个分组，若该电子书的书名向量中包含该分组的关键词，则将该电子书划分到该分组中。

其中，电子书的书名字符串可以通过预设的书名过滤规则对电子书的书名进行过滤处理得到。实际应用中，书籍库中电子书的书名通常由中文字符、阿拉伯数字、英文字母以及一些空字符组成，空格、制表符等。因此，为了方便电子书的书名比对，本发明的方案中，在进行书籍库中相似书的净化之前，可以利用预设的书名过滤规则，对书籍库中所有电子书的书名进行过滤处理，得到电子书的书名字符串。例如，将电子书的书名中的一些非必要的空字符过滤掉，保留英文字母、阿拉伯数据和中文字符，并将保留的字符组成该电子书的书名字符串。

本发明的方案中，可以通过向量化规则，将书名字符串中相邻的N个字符组成一个子字符串，之后将拆分得到的各子字符串构成该电子书对应的书名向量，其中，N为预设的拆分基数，可以具体为1、2、3……等正整数。实际应用中，若电子书的书名字符串的书名长度小于拆分基数（比如3），则可以不进行拆分，而是将该书名字符串直接构成该电子书的书名向量。

在得到书籍库中所有电子书各自对应的书名向量之后，可以将所有电子书所对应的书名向量合并成一个向量；并根据合并成的向量中的向量元素进行分组，确定出各分组的关键词。

具体地，合并成的向量中的每个向量元素，可以对应一个分组，该分组的关键词即为该向量元素。例如，将三百万本书的所有书名向量合并成一个向量，合并成的向量中向量元素可能只有一百万，即合并成的向量的长度为一百万。

实际应用中，在对合并成的向量进行分组的过程中，分组的个数可以等于该合并成的向量的长度，也可以小于该合并成的向量的长度。

在完成合并成的向量的分组之后，可以将书籍库中所有的电子书进行分组划分。具体地，针对每本电子书，将该电子书的书名向量与各个分组的关键词进行比对，若该书名向量中包含了某个分组的关键词，则可以将该电子书划分到该分组中。实际应用中，若电子书的书名向量中包含了多个分组的关键词，因此，该电子书将被划分到多个分组中。

S302：对于每个分组，根据该分组内各电子书的章节列表，计算出该分组内各电子书之间的相似度。

本发明的方案中，两本电子书之间的相似度是通过两本书的章节列表之间的相似度来体现的。因此，对于每个分组，可以根据该分组内各电子书的章节列表，计算出该分组内各电子书之间的相似度。

实际应用中，可以以二次规划的最优化问题来计算两个章节列表之间的相似度，但是，由于二次规划的最优化问题计算过程中需要生成一个矩阵，而在大规模的书籍库中电子书的数量庞大，因此，矩阵的生成过程就会消耗很多时间，不利于挖掘速度的提高。

因此，本发明的方案中可以采用近似最优的方法，将分组内的电子书两两进行比对，得到比对的两本电子书之间的相似度，提高相似度的计算速度，从而提高相似书的挖掘速度。

具体地，对于每个分组，可以针对该分组内的任意两本电子书，将该两本电子书各自的章节列表中章节标题进行比对，计算出该两本电子书之间的jaccard系数，并作为两本电子书之间的相似度。关于jaccard系数的计算方法可以采用本领域技术人员常用的技术手段，在此不再赘述。

其中，章节列表的章节标题可以通过预设的章节过滤规则对电子书的章节列表进行过滤处理得到。具体地，在进行书籍库中相似书的净化之前，可以利用预设的章节过滤规则，

对书籍库中所有电子书的章节列表进行过滤处理，得到电子书的章节列表中的各章节标题。其中，章节过滤规则包括如下至少之一：前缀过滤规则、后缀过滤规则。

更优地，考虑到章节列表中通常还包含了一些空字符，因此，在对电子书的章节列表进行预处理的过程中，还可以将章节列表中各章节标题的空字符删掉。其中，空字符可以包括：5 空格、制表符等。

进一步地，考虑到章节列表中通常还包含了标点符号，因此，在对电子书的章节列表进行预处理的过程中，还可以将章节列表中的标点符号去掉。

这样，通过上述空字符、标点符号、前缀过滤规则、后缀过滤规则对书籍库中电子书的章节列表进行过滤处理，将可以得到电子书的章节列表中的各章节标题。

10 S303：对于每个分组，若该分组内两本电子书之间的相似度超过设定阈值，则将该两本电子书确定为书籍库中的一对相似书。

具体地，对于每个分组，通过步骤 302 计算出该分组内各电子书之间的相似度之后，若分组内两本电子书之间的相似度超过设定阈值，则将该两本电子书确定为书籍库中的一对相似书。

15 其中，设定阈值是由本领域技术人员预先设置的，比如，该设定阈值可以设置为 0.8。这样，在将分组内的电子书的章节列表进行两两比对之后，若比对的两本电子书之间的相似度超过 0.8，则说明该两本电子书的章节列表相似，可以确定为该分组内的一对相似书。

S304：利用书籍库中所有分组内的相似书，构建书籍库的图模型；其中，每对相似书为图模型中相连的两个端点。

20 为了便于描述，本发明的方案中，可以将书籍库中的一对相似书称为书籍库中的一个相似对。

实际应用中，在挖掘出书籍库中所有分组内的相似书之后，可以利用所有的相似对，构建图模型。在该图模型中，每个相似对将作为一条边的两个端点。这样，通过构建的图模型可以将书籍库中所有的相似对中相关联的相似对连接起来。

25 S305：对于图模型中的每个连通分量，从该连通分量中选择一本电子书保留，删除其他电子书。

实际应用中，如果从一个端点到另一个端点有路径，则称这两个端点连通。而如果图中任意两个端点之间都连通，则可以称该图为连通图，否则，称该图为非连通图，则其中的极大连通子图称为连通分量，这里所谓的极大是指子图中包含的端点个数极大。

30 因此，本发明的方案中，在构建了图模型之后，可以确定出该图模型中存在的连通分量。实际应用中，每个连通分量其所包含的所有电子书，其针对的往往是同一本书。因此，为了降低书籍库的规模，对于图模型中确定出的每个连通分量，可以从该连通分量的内部寻找最好最完整的电子书进行保留而删除其他电子书。

35 例如，书籍库中挖掘出了 1 万个相似对，利用这 1 万个相似对构建出图模型之后，若从该图模型中寻找出了 1000 个连通分量，而每个连通分量中平均有 10 本可以互相成为一对相似书的电子书。为了降低书籍库的规模，可以从该 10 本电子书中寻找出最好的一本进行保留，而删除其他 9 本电子书。

40 具体地，对于每个连通分量，可以针对该连通分量中的每本电子书，将该连通分量中与该电子书相连的电子书的个数，作为该电子书的关联度；之后，选取关联度最高的电子书保留，而删除该连通分量中其他的电子书。

或者，对于每个连通分量，可以针对该连通分量中的每本电子书，将该连通分量中该电子书相连的各本电子书与该电子书之间的相似度进行累加，得到该电子书的关联度；选取关联度最高的电子书保留。

45 进一步地，针对该连通分量中的每本电子书，在确定出该连通分量中与该电子书相连的电子书之后，将各本电子书与该电子书之间的相似度进行累加，并将累加的相似度除以与该

电子书相连的电子书的数量，得到该电子书的平均相似度，并作为该电子书的关联度；选取关联度最高的电子书保留。

本发明的方案中，在挖掘出书籍库中的相似书之后，基于挖掘出所有分组内的相似书，构建图模型；进而针对图模型中每个连通分量，从该连通分量中选择一个电子书保留，而删除一些相似书，从而实现对书籍库进行相似书的净化处理，以此降低书籍库的规模。

基于上述书籍库中相似书的净化方法，本发明还提供了一种书籍库中相似书的净化装置，如图4所示，该净化装置可以包括：书籍分组单元401、相似书挖掘单元402、以及相似书净化单元403。

其中，书籍分组单元401用于根据书籍库中所有电子书的书名，确定出各电子书所属的分组。

相似书挖掘单元402用于对于每个分组，根据该分组内各电子书的章节列表，计算出该分组内各电子书之间的相似度；若该分组内两本电子书之间的相似度超过设定阈值，则将该两本电子书确定为书籍库中的一对相似书。

相似书净化单元403用于利用书籍库中所有分组内的相似书，构建书籍库的图模型，其中，每对相似书为图模型中相连的两个端点；对于书籍库的图模型中的每个连通分量，从连通分量中选择一本电子书保留，删除该连通分量中的其他电子书。

具体地，相似书净化单元403可以利用书籍库中所有分组内的相似书，构建书籍库的图模型；对于图模型中的每个连通分量，针对连通分量中的每本电子书，将连通分量中与该电子书相连的电子书的个数，作为该电子书的关联度；选取关联度最高的电子书保留，删除该连通分量中的其他电子书。

或者，相似书净化单元403可以利用书籍库中所有分组内的相似书，构建书籍库的图模型；对于图模型中的每个连通分量，针对连通分量中的每本电子书，将连通分量中该电子书相连的各本电子书与该电子书之间的相似度进行累加，得到该电子书的关联度；选取关联度最高的电子书保留，删除其他电子书。

进一步地，相似书净化单元403可以针对该连通分量中的每本电子书，在确定出该连通分量中与该电子书相连的电子书之后，将各本电子书与该电子书之间的相似度进行累加，并将累加的相似度除以与该电子书相连的电子书的数量，得到该电子书的平均相似度，并作为该电子书的关联度；选取关联度最高的电子书保留，删除其他电子书。

更优地，如图4所示，书籍库中相似书的净化装置还包括：预处理单元404。

预处理单元404用于利用预设的书名过滤规则，对书籍库中所有电子书的书名进行过滤处理，得到电子书的书名字符串；利用预设的章节过滤规则，对书籍库中所有电子书的章节列表进行过滤处理，得到电子书的章节列表中的各章节标题。

其中，章节过滤规则包括如下至少之一：前缀过滤规则、后缀过滤规则。

相应地，书籍分组单元401可以针对书籍库中的每本电子书，利用预设的向量化规则拆分该电子书的书名字符串，由拆分得到的各子字符串构成该电子书对应的书名向量；将拆分得到的所有子字符串进行合并去重，得到各分组的关键词；针对书籍库中的每本电子书，对于每个分组，若该电子书的书名向量中包含该分组的关键词，则将该电子书划分到该分组中。

相似书挖掘单元402可以对于每个分组，针对该分组内的任意两本电子书，将该两本电子书各自的章节列表中章节标题进行比对，计算出该两本电子书之间的jaccard系数，并作为两本电子书之间的相似度；若该分组内两本电子书之间的相似度超过设定阈值，则将该两本电子书确定为书籍库中的一对相似书。

本发明的方案中，书籍库中相似书的净化装置中的各单元的具体实现可参照上述书籍库中相似书的净化方法中的各步骤，在此不再赘述。

实际应用中，在通过上述书籍库中相似书的净化方法对书籍库中的相似书进行进化之后，若接收到用户针对某一电子书的搜索请求之后，可以将净化后的书籍库中对应的电子书推送

给用户。

因此，本发明还提供了一种基于净化的书籍库的电子书推送方法，如图 5 所示，该方法包括如下步骤：

S501：接收包含电子书的书名的用户搜索请求。

5 S502：从净化后的书籍库中查询与书名相对应的电子书。

其中，净化后的书籍库具体是指根据上述书籍库中相似书的净化方法所净化的书籍库。

本发明的方案中，获取接收的用户搜索请求中包含的电子书的书名，根据获取的书名从净化后的书籍库中查询与该书名相对应的电子书。

S503：响应于该用户搜索请求推送查询到的电子书的摘要信息页面。

10 具体地，通过步骤 S502 查询到与用户搜索请求中包含的书名相对应的电子书之后，可以响应于该用户搜索请求，向用户推送该电子书的摘要信息页面。

基于上述基于净化的书籍库的电子书推送方法，本发明还提供了一种基于净化的书籍库的电子书推送装置，如图 6 所示，该推送装置可以包括：请求接收单元 601、电子书查询单元 602、以及推送单元 603。

15 其中，请求接收单元 601 用于接收包含电子书的书名的用户搜索请求。

电子书查询单元 602 用于获取用户搜索请求中的电子书的书名，从净化后的书籍库中查询与该书名相对应的电子书。

其中，净化后的书籍库具体是指根据上述书籍库中相似书的净化方法所净化的书籍库。

推送单元 603 用于响应于用户搜索请求推送查询到的电子书的摘要信息页面。

20 本发明的方案中，基于净化的书籍库的电子书推送装置中的各单元的具体实现可参照上述基于净化的书籍库的电子书推送方法中的各步骤，在此不再赘述。

本发明的技术方案中，可以根据电子书的书名，将书籍库中的电子书进行分组，即将书名相似的电子书归为一个分组；继而，对于每个分组，可以根据该分组内各电子书的章节列表，计算出该分组内各电子书之间的相似度，将相似度超过设定阈值的两本电子书确定为书籍库的一对相似书。这样，通过书名和章节列表的两次比对来挖掘相似书，可以避免大量的章节内容的比对，在保障了相似书挖掘的准确率的同时，可以减少挖掘工作量，提高了书籍库中相似书的挖掘速度。

进一步地，在挖掘出书籍库中的相似书之后，还可以基于挖掘出所有分组内的相似书，对书籍库进行相似书的净化处理，删除一些重复的内容，以此降低书籍库的规模。

30 本技术领域技术人员可以理解，本发明包括涉及用于执行本申请中所述操作中的一项或多项的设备。这些设备可以为所需的目的而专门设计和制造，或者也可以包括通用计算机中的已知设备。这些设备具有存储在其内的计算机程序，这些计算机程序选择性地激活或重构。这样的计算机程序可以被存储在设备（例如，计算机）可读介质中或者存储在适于存储电子指令并分别耦联到总线的任何类型的介质中，所述计算机可读介质包括但不限于任何类型的盘（包括软盘、硬盘、光盘、CD-ROM、和磁光盘）、ROM（Read-Only Memory，只读存储器）、RAM（Random Access Memory，随机存储器）、EPROM（Erasable Programmable Read-Only Memory，可擦写可编程只读存储器）、EEPROM（Electrically Erasable Programmable Read-Only Memory，电可擦可编程只读存储器）、闪存、磁性卡片或光线卡片。也就是，可读介质包括由设备（例如，计算机）以能够读的形式存储或传输信息的任何介质。

40 本技术领域技术人员可以理解，可以用计算机程序指令来实现这些结构图和/或框图和/或流程图中的每个框以及这些结构图和/或框图和/或流程图中的框的组合。本技术领域技术人员可以理解，可以将这些计算机程序指令提供给通用计算机、专业计算机或其他可编程数据处理方法的处理器来实现，从而通过计算机或其他可编程数据处理方法的处理器来执行本发明公开的结构图和/或框图和/或流程图的框或多个框中指定的方案。

45 本技术领域技术人员可以理解，本发明中已经讨论过的各种操作、方法、流程中的步骤、

措施、方案可以被交替、更改、组合或删除。进一步地，具有本发明中已经讨论过的各种操作、方法、流程中的其他步骤、措施、方案也可以被交替、更改、重排、分解、组合或删除。进一步地，现有技术中的具有与本发明中公开的各种操作、方法、流程中的步骤、措施、方案也可以被交替、更改、重排、分解、组合或删除。

5 以上所述仅是本发明的部分实施方式，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，这些改进和润饰也应视为本发明的保护范围。

10 本发明的各个部件实施例可以以硬件实现，或者以在一个或者多个处理器上运行的软件模块实现，或者以它们的组合实现。本领域的技术人员应当理解，可以在实践中使用微处理器或者数字信号处理器（DSP）来实现根据本发明实施例的书籍库中相似书的挖掘、净化装置中的一些或者全部部件的一些或者全部功能。本发明还可以实现为用于执行这里所描述的方法的一部分或者全部的设备或者装置程序（例如，计算机程序和计算机程序产品）。这样的实现本发明的程序可以存储在计算机可读介质上，或者可以具有一个或者多个信号的形式。这样的信号可以从因特网网站上下载得到，或者在载体信号上提供，或者以任何其他形式提供。

15 例如，图 7 示出了用于执行根据本发明的方法的计算设备。该计算设备传统上包括处理器 710 和以存储器 720 形式的计算机程序产品或者计算机可读介质。存储器 720 可以是诸如闪存、EEPROM（电可擦除可编程只读存储器）、EPROM、硬盘或者 ROM 之类的电子存储器。存储器 720 具有用于执行上述方法中的任何方法步骤的程序代码 731 的存储空间 730。例如，用于程序代码的存储空间 730 可以包括分别用于实现上面的方法中的各种步骤的各个程序代码 731。这些程序代码可以从一个或者多个计算机程序产品中读出或者写入到这一个或者多个计算机程序产品中。这些计算机程序产品包括诸如硬盘，紧致盘（CD）、存储卡或者软盘之类的程序代码载体。这样的计算机程序产品通常为如参考图 8 所述的便携式或者固定存储单元。该存储单元可以具有与图 7 的计算设备中的存储器 720 类似布置的存储段、存储空间等。程序代码可以例如以适当形式进行压缩。通常，存储单元包括计算机可读代码 731'，即可以由例如诸如 710 之类的处理器读取的代码，这些代码当由计算设备运行时，导致该计算设备执行上面所描述的方法中的各个步骤。

20 本文中所述的“一个实施例”、“实施例”或者“一个或者多个实施例”意味着，结合实施例描述的特定特征、结构或者特性包括在本发明的至少一个实施例中。此外，请注意，这里“在一个实施例中”的词语例子不一定全指同一个实施例。

30 在此处所提供的说明书中，说明了大量具体细节。然而，能够理解，本发明的实施例可以在没有这些具体细节的情况下被实践。在一些实例中，并未详细示出公知的方法、结构和技术，以便不模糊对本说明书的理解。

35 应该注意的是上述实施例对本发明进行说明而不是对本发明进行限制，并且本领域技术人员在不脱离所附权利要求的范围的情况下可设计出替换实施例。在权利要求中，不应将位于括号之间的任何参考符号构造成对权利要求的限制。单词“包含”不排除存在未列在权利要求中的元件或步骤。位于元件之前的单词“一”或“一个”不排除存在多个这样的元件。本发明可以借助于包括有若干不同元件的硬件以及借助于适当编程的计算机来实现。在列举了若干装置的单元权利要求中，这些装置中的若干个可以是通过同一个硬件项来具体体现。单词第一、第二、以及第三等的使用不表示任何顺序。可将这些单词解释为名称。

40 此外，还应当注意，本说明书中使用的语言主要是为了可读性和教导的目的而选择的，而不是为了解释或者限定本发明的主题而选择的。因此，在不偏离所附权利要求书的范围和精神的情况下，对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。对于本发明的范围，对本发明所做的公开是说明性的，而非限制性的，本发明的范围由所附权利要求书限定

权 利 要 求

1. 一种书籍库中相似书的净化方法，其中，包括：

根据所述书籍库中所有电子书的书名，确定出各电子书所属的分组；

5 对于每个分组，根据该分组内各电子书的章节列表，计算出该分组内各电子书之间的相似度，若该分组内两本电子书之间的相似度超过设定阈值，则将该两本电子书确定为所述书籍库中的一对相似书；

利用所述书籍库中所有分组内的相似书，构建所述书籍库的图模型，其中，每对相似书为所述图模型中相连的两个端点；

10 对于所述图模型中的每个连通分量，从该连通分量中选择一本电子书保留，删除其他电子书。

2. 根据权利要求1所述的方法，其中，所述根据所述书籍库中所有电子书的书名，确定出各书所属的分组之前，还包括：

15 利用预设的书名过滤规则，对所述书籍库中所有电子书的书名进行过滤处理，得到电子书的书名字符串；

利用预设的章节过滤规则，对所述书籍库中所有电子书的章节列表进行过滤处理，得到电子书的章节列表中的各章节标题。

3. 根据权利要求2所述的方法，其中，所述章节过滤规则包括如下至少之一：前缀过滤规则、后缀过滤规则。

4. 根据权利要求1所述的方法，其中，所述根据所述书籍库中所有电子书的书名，确定出各电子书所属的分组，具体包括：

针对书籍库中的每本电子书，利用预设的向量化规则拆分该电子书的书名字符串，由拆分得到的各子字符串构成该电子书对应的书名向量；

25 将拆分得到的所有子字符串进行合并去重，得到各分组的关键词；

针对书籍库中的每本电子书，对于每个分组，若该电子书的书名向量中包含该分组的关键词，则将该电子书划分到该分组中。

5. 根据权利要求1所述的方法，其中，所述根据该分组内各电子书的章节列表，计算出该分组内各电子书之间的相似度，具体包括：

30 针对该分组内的任意两本电子书，将该两本电子书各自的章节列表中章节标题进行比对，计算出该两本电子书之间的杰卡德jaccard系数，并作为两本电子书之间的相似度。

6. 根据权利要求1所述的方法，其中，所述从该连通分量中选择一本电子书保留，具体包括：

35 针对该连通分量中的每本电子书，将该连通分量中与该电子书相连的电子书的个数，作为该电子书的关联度；

选取关联度最高的电子书保留。

7. 根据权利要求1所述的方法，其中，所述从该连通分量中选择一本电子书保留，具体包括：

40 针对该连通分量中的每本电子书，将该连通分量中该电子书相连的各本电子书与该电子书之间的相似度进行累加，得到该电子书的关联度；

选取关联度最高的电子书保留。

8. 一种书籍库中相似书的净化装置，其中，包括：

书籍分组单元，用于根据所述书籍库中所有电子书的书名，确定出各电子书所属的分

组;

相似书挖掘单元,用于对于每个分组,根据该分组内各电子书的章节列表,计算出该分组内各电子书之间的相似度;若该分组内两本电子书之间的相似度超过设定阈值,则将该两本电子书确定为所述书籍库中的一对相似书;

- 5 相似书净化单元,用于利用所述书籍库中所有分组内的相似书,构建所述书籍库的图模型,其中,每对相似书为所述图模型中相连的两个端点;对于所述书籍库的图模型中的每个连通分量,从连通分量中选择一本电子书保留,删除其他电子书。

9. 根据权利要求8所述的装置,其中,该装置还包括:

- 10 预处理单元,用于利用预设的书名过滤规则,对所述书籍库中所有电子书的书名进行过滤处理,得到电子书的书名字符串;利用预设的章节过滤规则,对所述书籍库中所有电子书的章节列表进行过滤处理,得到电子书的章节列表中的各章节标题。

10. 根据权利要求9所述的装置,其中,所述章节过滤规则包括如下至少之一:前缀过滤规则、后缀过滤规则。

- 15 11. 根据权利要求8所述的装置,其中,所述书籍分组单元具体用于针对书籍库中的每本电子书,利用预设的向量化规则拆分该电子书的书名字符串,由拆分得到的各子字符串构成该电子书对应的书名向量;将拆分得到的所有子字符串进行合并去重,得到各分组的关键词;针对书籍库中的每本电子书,对于每个分组,若该电子书的书名向量中包含该分组的关键词,则将该电子书划分到该分组中。

- 20 12. 根据权利要求8所述的装置,其中,所述相似书挖掘单元具体用于对于每个分组,针对该分组内的任意两本电子书,将该两本电子书各自的章节列表中章节标题进行比对,计算出该两本电子书之间的杰卡德jaccard系数,并作为两本电子书之间的相似度;若该分组内两本电子书之间的相似度超过设定阈值,则将该两本电子书确定为所述书籍库中的一对相似书。

- 25 13. 根据权利要求8所述的装置,其中,所述相似书净化单元具体用于利用所述书籍库中所有分组内的相似书,构建所述书籍库的图模型;对于所述图模型中的每个连通分量,针对连通分量中的每本电子书,将连通分量中与该电子书相连的电子书的个数,作为该电子书的关联度;选取关联度最高的电子书保留,删除其他电子书。

- 30 14. 根据权利要求8所述的装置,其中,所述相似书净化单元具体用于利用所述书籍库中所有分组内的相似书,构建所述书籍库的图模型;对于所述图模型中的每个连通分量,针对连通分量中的每本电子书,将连通分量中该电子书相连的各本电子书与该电子书之间的相似度进行累加,得到该电子书的关联度;选取关联度最高的电子书保留,删除其他电子书。

15. 一种书籍库中相似书的挖掘方法,其特征在于,包括:

根据书籍库中所有电子书的书名,确定出各电子书所属的分组;

- 35 对于每个分组,根据该分组内各电子书的章节列表,计算出该分组内各电子书之间的相似度;

对于每个分组,若分组内两本电子书之间的相似度超过设定阈值,则将该两本电子书确定为所述书籍库中的一对相似书。

- 40 16. 根据权利要求15所述的方法,其中,所述若分组内两本电子书之间的相似度超过设定阈值,则将该两本电子书确定为所述书籍库中的一对相似书之后,还包括:

利用所述书籍库中所有分组内的相似书,构建所述书籍库的图模型,其中,每对相似书为所述图模型中相连的两个端点;

对于所述图模型中的每个连通分量,从该连通分量中选择一本电子书保留,删除其他

电子书。

17. 根据权利要求 16 所述的方法, 其中, 所述从该连通分量中选择一本电子书保留, 具体包括:

5 针对该连通分量中的每本电子书, 将该连通分量中与该电子书相连的电子书的个数, 作为该电子书的关联度;

选取关联度最高的电子书保留。

18. 根据权利要求 16 所述的方法, 其中, 所述从该连通分量中选择一本电子书保留, 具体包括:

10 针对该连通分量中的每本电子书, 将该连通分量中该电子书相连的各本电子书与该电子书之间的相似度进行累加, 得到该电子书的关联度;

选取关联度最高的电子书保留。

19. 根据权利要求 15 所述的方法, 其中, 所述根据书籍库中所有电子书的书名, 确定出各书所属的分组之前, 还包括:

15 利用预设的书名过滤规则, 对所述书籍库中所有电子书的书名进行过滤处理, 得到电子书的书名字符串;

利用预设的章节过滤规则, 对所述书籍库中所有电子书的章节列表进行过滤处理, 得到电子书的章节列表中的各章节标题。

20. 根据权利要求 19 所述的方法, 其中, 所述章节过滤规则包括如下至少之一: 前缀过滤规则、后缀过滤规则。

20 21. 根据权利要求 15 所述的方法, 其中, 所述根据书籍库中所有电子书的书名, 确定出各电子书所属的分组, 具体包括:

针对书籍库中的每本电子书, 利用预设的向量化规则拆分该电子书的书名字符串, 由拆分得到的各子字符串构成该电子书对应的书名向量;

将拆分得到的所有子字符串进行合并去重, 得到各分组的关键词;

25 针对书籍库中的每本电子书, 对于每个分组, 若该电子书的书名向量中包含该分组的关键词, 则将该电子书划分到该分组中。

22. 根据权利要求 15 所述的方法, 其中, 所述根据该分组内各电子书的章节列表, 计算出该分组内各电子书之间的相似度, 具体包括:

30 针对该分组内的任意两本电子书, 将该两本电子书各自的章节列表中章节标题进行比对, 计算出该两本电子书之间的杰卡德 jaccard 系数, 并作为两本电子书之间的相似度。

23. 一种书籍库中相似书的挖掘装置, 其中, 包括:

书籍分组单元, 用于根据所述书籍库中所有电子书的书名, 确定出各电子书所属的分组;

35 相似度计算单元, 用于对于每个分组, 根据该分组内各电子书的章节列表, 计算出该分组内各电子书之间的相似度;

相似书挖掘单元, 用于对于每个分组, 若该分组内两本电子书之间的相似度超过设定阈值, 则将该两本电子书确定为所述书籍库中的一对相似书。

24. 根据权利要求 23 所述的装置, 其中, 该装置还包括:

40 书籍净化单元, 用于利用所述相似书挖掘单元确定出的书籍库中所有分组内的相似书, 构建所述书籍库的图模型; 对于所述图模型中的每个连通分量, 从连通分量中选择一本电子书保留, 删除其他电子书;

其中, 每对相似书为所述图模型中相连的两个端点。

25. 根据权利要求 24 所述的装置, 其中, 所述书籍净化单元具体用于利用所述书籍

库中所有分组内的相似书,构建所述书籍库的图模型;对于所述图模型中的每个连通分量,针对连通分量中的每本电子书,将连通分量中与该电子书相连的电子书的个数,作为该电子书的关联度;选取关联度最高的电子书保留,删除其他电子书。

26. 根据权利要求 24 所述的装置,其中,所述书籍净化单元具体用于利用所述书籍库中所有分组内的相似书,构建所述书籍库的图模型;对于所述图模型中的每个连通分量,针对连通分量中的每本电子书,将连通分量中该电子书相连的各本电子书与该电子书之间的相似度进行累加,得到该电子书的关联度;选取关联度最高的电子书保留,删除其他电子书。

27. 根据权利要求 23 所述的装置,其中,该装置还包括:
10 预处理单元,用于利用预设的书名过滤规则,对所述书籍库中所有电子书的书名进行过滤处理,得到电子书的书名字符串;利用预设的章节过滤规则,对所述书籍库中所有电子书的章节列表进行过滤处理,得到电子书的章节列表中的各章节标题。

28. 根据权利要求 27 所述的装置,其中,所述章节过滤规则包括如下至少之一:前缀过滤规则、后缀过滤规则。

15 29. 根据权利要求 23 所述的装置,其中,所述书籍分组单元具体用于针对书籍库中的每本电子书,利用预设的向量化规则拆分该电子书的书名字符串,由拆分得到的各子字符串构成该电子书对应的书名向量;将拆分得到的所有子字符串进行合并去重,得到各分组的关键词;针对书籍库中的每本电子书,对于每个分组,若该电子书的书名向量中包含该分组的关键词,则将该电子书划分到该分组中。

20 30. 根据权利要求 23 所述的装置,其中,所述相似度计算单元具体用于对于每个分组,针对该分组内的任意两本电子书,将该两本电子书各自的章节列表中章节标题进行对比,计算出该两本电子书之间的杰卡德 jaccard 系数,并作为两本电子书之间的相似度。

31. 一种基于净化的书籍库的电子书推送方法,其特征在于,包括:

接收包含电子书的书名的用户搜索请求;

25 从根据如权利要求 1 至 7 任一所述的方法所净化的书籍库中查询与所述书名相对应的电子书;

响应于所述用户搜索请求推送查询到的电子书的摘要信息页面。

32. 一种基于净化的书籍库的电子书推送装置,其特征在于,包括:

请求接收单元,用于接收包含电子书的书名的用户搜索请求;

30 电子书查询单元,用于获取所述用户搜索请求中的电子书的书名,从根据如权利要求 1 至 7 任一所述的方法所净化的书籍库中查询与所述书名相对应的电子书;

推送单元,用于响应于所述用户搜索请求推送查询到的电子书的摘要信息页面。

33. 一种计算机程序,包括计算机可读代码,当所述计算机可读代码在计算设备上运行时,导致所述计算设备执行根据权利要求 1-7 中的任一个所述的书籍库中相似书的净化方法,或者导致所述计算设备执行根据权利要求 15-22 中任一个所述的书籍库中相似书的挖掘方法,或者导致所述计算设备执行根据权利要求 31 所述的基于净化的书籍库的电子书推送方法。

34. 一种计算机可读介质,其中存储了如权利要求 33 所述的计算机程序。

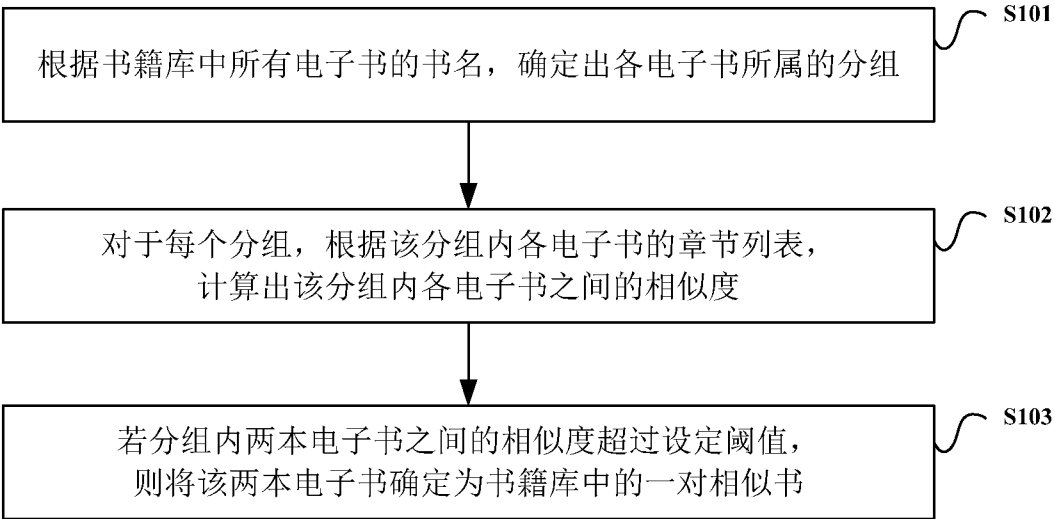


图 1

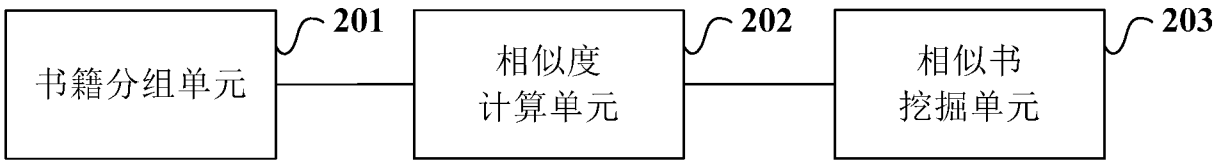


图 2a

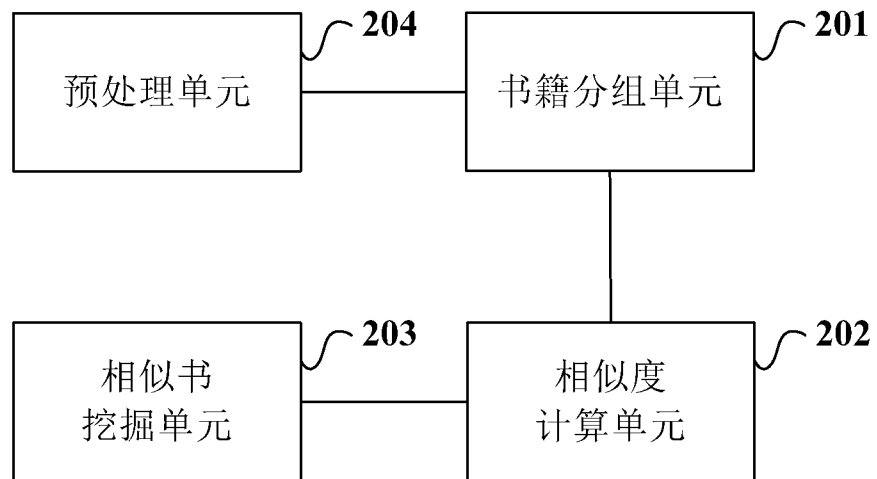


图 2b

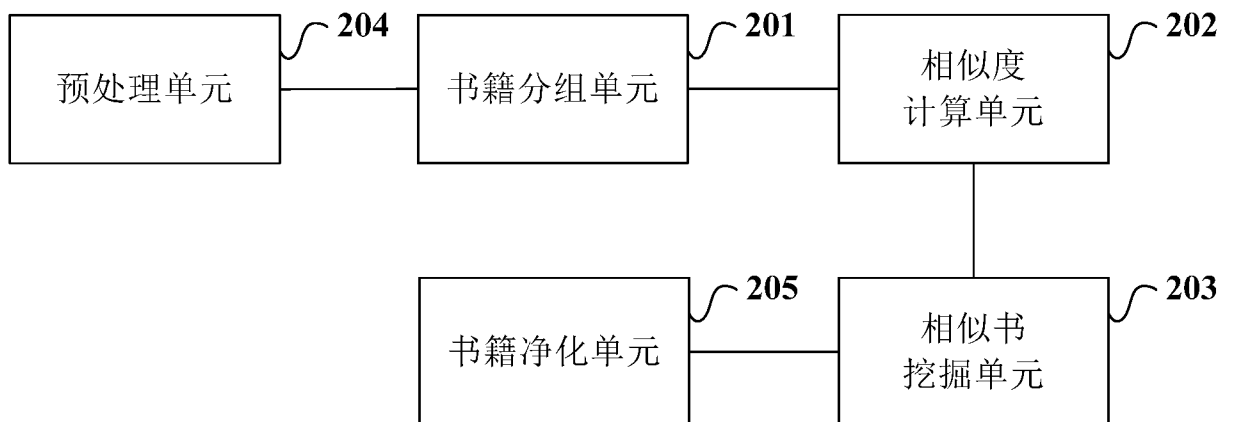


图 2c

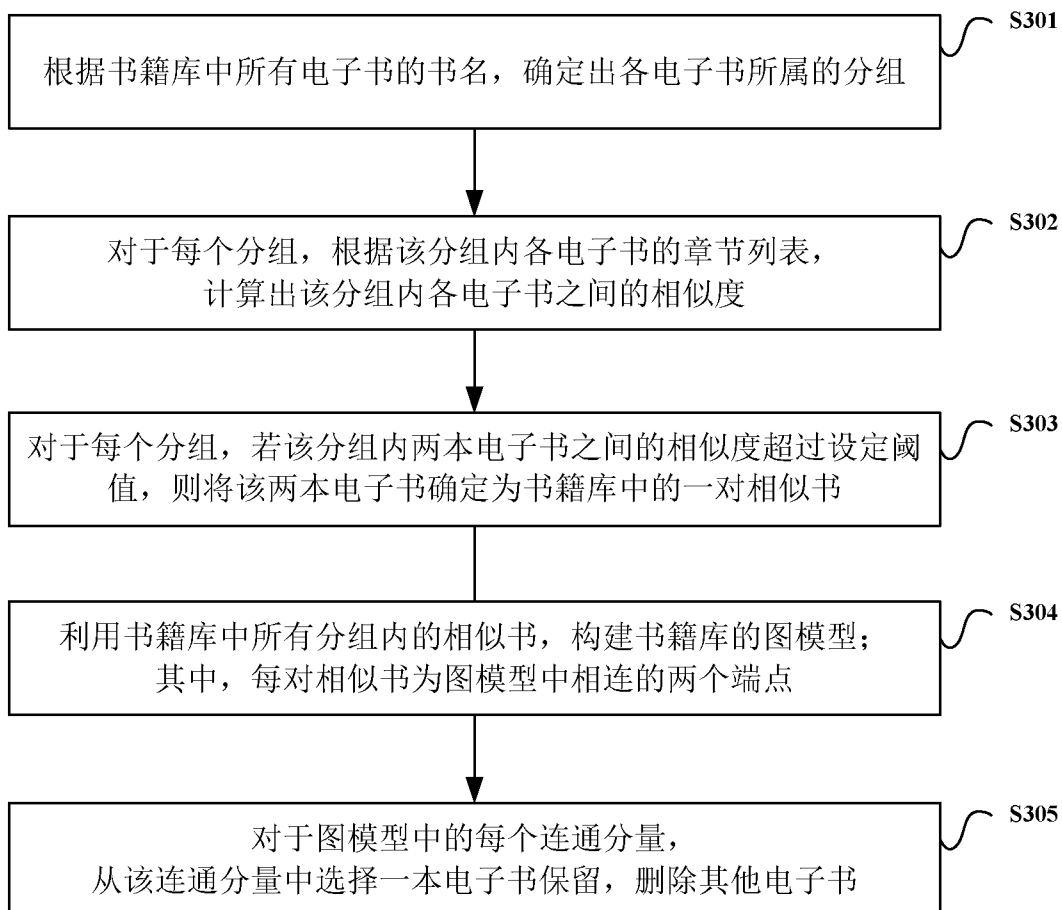


图 3

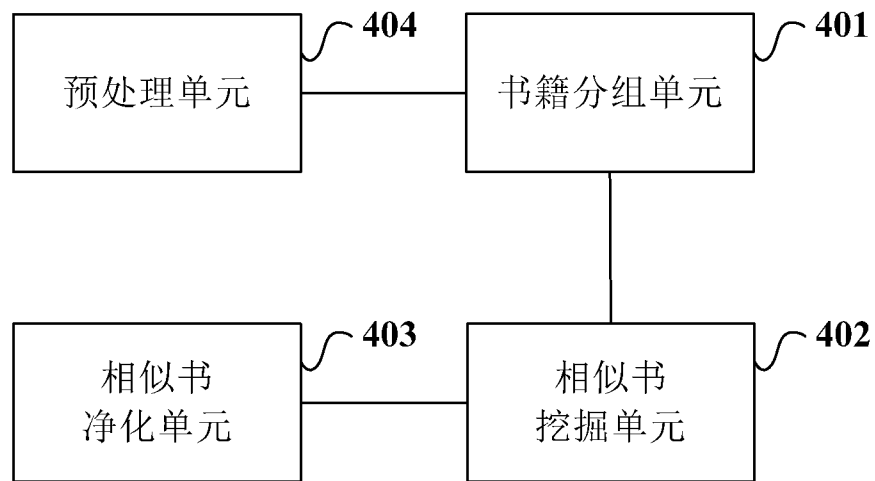


图 4

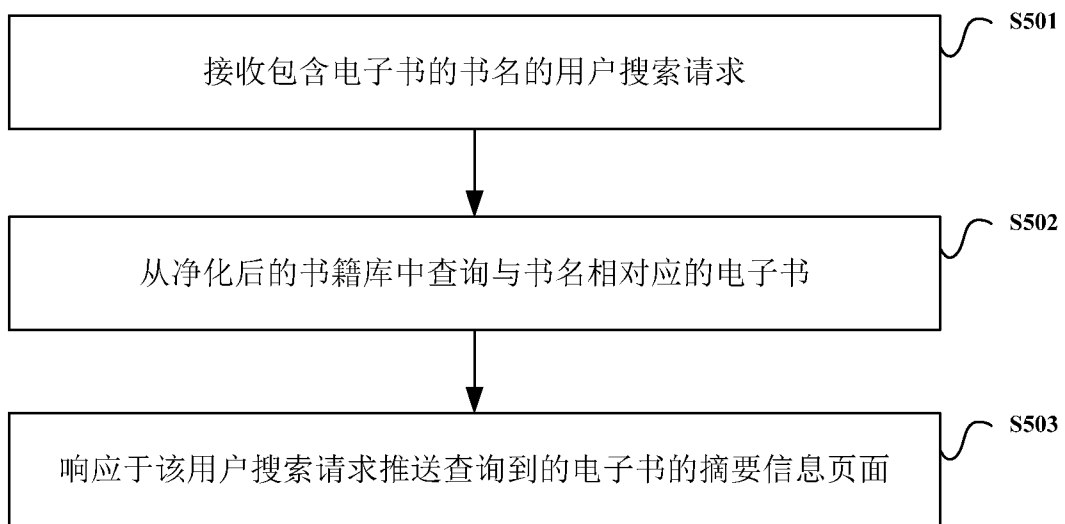


图 5

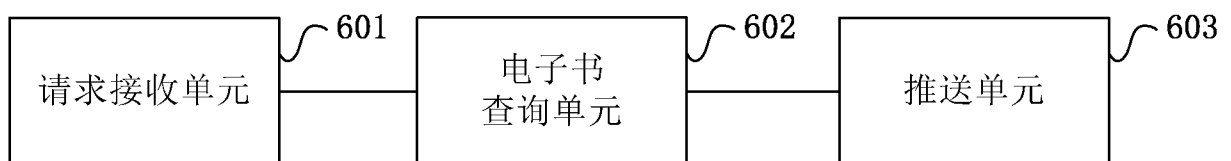


图 6

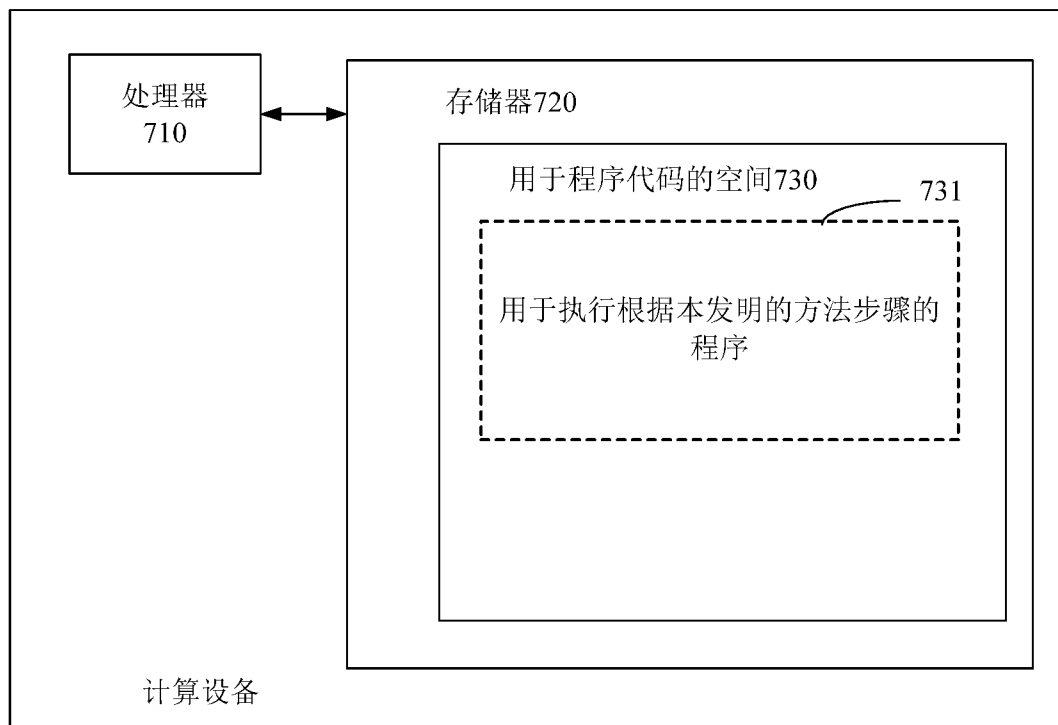


图 7

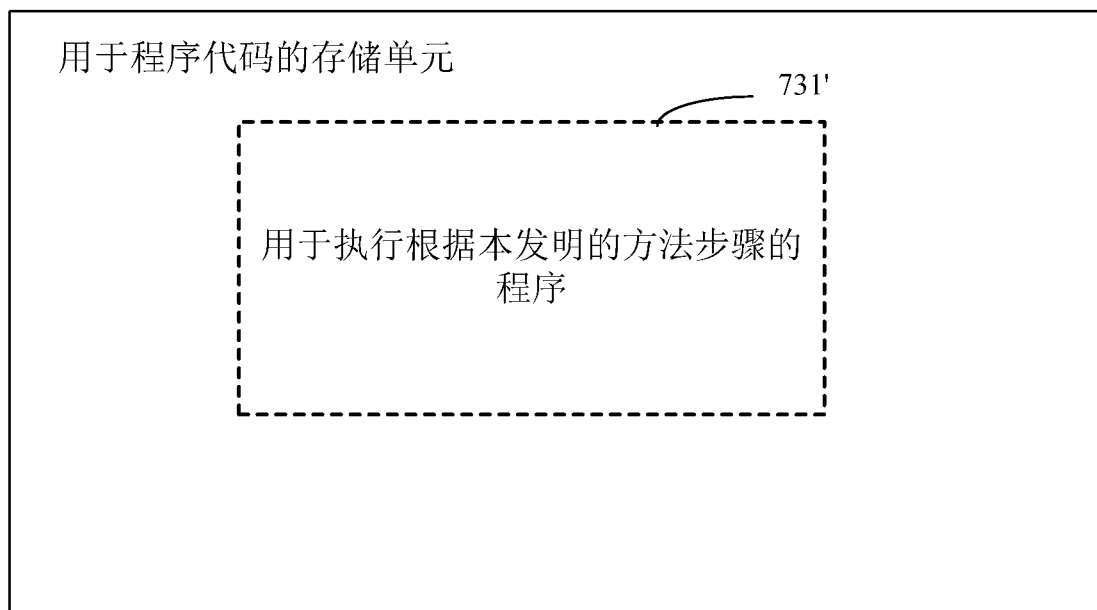


图 8

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2016/099894

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 17/30

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)

CNABS, CNKI, DWPI, SIPOABS: QIHOO; QIZHI; ZHANG, Chao; book, name, chapter, purify+, min+, threshold, similar+, construction model, graph model, jaccard

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 102024065 A (CENTRAL SOUTH UNIVERSITY), 20 April 2011 (20.04.2011), description, paragraphs [0078]-[0148], and figures 1-5	1-34
PX	CN 105373604 A (BEIJING QIHOO TECHNOLOGY CO., LTD. et al.), 02 March 2016 (02.03.2016), description, paragraphs [0095]-[0224], and figures 1-6	1-34
X	US 2013332466 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION), 12 December 2013 (12.12.2013), description, paragraphs [0022]-[0048], and figures 1 and 2	1-34

☐ 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 22 December 2016 (22.12.2016)	Date of mailing of the international search report 03 January 2017 (03.01.2017)
Name and mailing address of the ISA/CN: 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 WANG, Kun Telephone No.: (86-10) 62412187

International application No.
PCT/CN2016/099894

Form PCT/ISA/210 (patent family annex) (July 2009)

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 17/30 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)) CNABS, CNKI, DWPI, SIPOABS 奇虎, 奇智, 张超, 书, 名, 章节, 净化, 挖掘, 阈值, 相似, 图模型, QIHOO, QIZHI, ZHANG CHAO, book, name, chapter, purify+, min+, threshold, similar+, construction model, graph model, jaccard														
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>X</td> <td>CN 102024065 A (中南大学) 2011年 4月 20日 (2011 - 04 - 20) 说明书第[0078]-[0148]段, 图1-5</td> <td>1-34</td> </tr> <tr> <td>PX</td> <td>CN 105373604 A (北京奇虎科技有限公司等) 2016年 3月 2日 (2016 - 03 - 02) 说明书第[0095]-[0224]段, 图1-6</td> <td>1-34</td> </tr> <tr> <td>X</td> <td>US 2013332466 A1 (国际商业机器公司) 2013年 12月 12日 (2013 - 12 - 12) 说明书第[0022]-[0048]段, 图1、2</td> <td>1-34</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	X	CN 102024065 A (中南大学) 2011年 4月 20日 (2011 - 04 - 20) 说明书第[0078]-[0148]段, 图1-5	1-34	PX	CN 105373604 A (北京奇虎科技有限公司等) 2016年 3月 2日 (2016 - 03 - 02) 说明书第[0095]-[0224]段, 图1-6	1-34	X	US 2013332466 A1 (国际商业机器公司) 2013年 12月 12日 (2013 - 12 - 12) 说明书第[0022]-[0048]段, 图1、2	1-34
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求												
X	CN 102024065 A (中南大学) 2011年 4月 20日 (2011 - 04 - 20) 说明书第[0078]-[0148]段, 图1-5	1-34												
PX	CN 105373604 A (北京奇虎科技有限公司等) 2016年 3月 2日 (2016 - 03 - 02) 说明书第[0095]-[0224]段, 图1-6	1-34												
X	US 2013332466 A1 (国际商业机器公司) 2013年 12月 12日 (2013 - 12 - 12) 说明书第[0022]-[0048]段, 图1、2	1-34												
☐ 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 22日		国际检索报告邮寄日期 2017年 1月 3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 王昆 电话号码 (86-10) 62412187												

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2016/099894

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	102024065	A	2011年 4月 20日	CN 102024065 B	2013年 1月 2日
CN	105373604	A	2016年 3月 2日	无	
US	2013332466	A1	2013年 12月 12日	US 2013332467 A1	2013年 12月 12日