(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 107220249 B
(45)授权公告日 2020.11.10

(21)申请号 201610162742.3

(22)申请日 2016.03.21

(65)同一申请的已公布的文献号
申请公布号 CN 107220249 A

(43)申请公布日 2017.09.29

(73)专利权人 伊姆西IP控股有限责任公司
地址 美国马萨诸塞州

(72)发明人 陈超 刘晶晶 张磊 薛丁萌 周旻弘 代洪涛

(74)专利代理机构 北京市金杜律师事务所 11256
代理人 王茂华

(51)Int.Cl.
G06F 16/31(2019.01)
G06F 16/33(2019.01)

(56)对比文件
US 2003101182 A1,2003.05.29
CN 102779185 A,2012.11.14

审查员 王国海

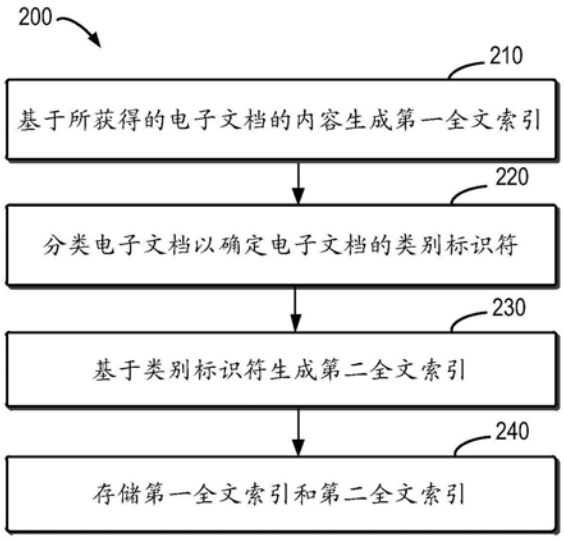
权利要求书4页 说明书9页 附图3页

(54)发明名称

基于分类的全文搜索

(57)摘要

本公开内容的各种实施例提供了一种基于分类的全文搜索的方案。在一些实施例中,提供了一种用于全文搜索的方法。该方法包括基于所获得的电子文档的内容生成第一全文索引。该方法还包括分类电子文档以确定电子文档的类别标识符,以及基于类别标识符生成第二全文索引。该方法进一步包括存储第一全文索引和第二全文索引。



CN 107220249 B

1.一种用于全文搜索的方法,包括:

基于所获得的电子文档的内容生成针对所述电子文档的第一全文索引;

分类所述电子文档以确定所述电子文档的类别标识符,所述类别标识符是分类结果,所述分类结果标识所述电子文档的文档类别;

至少部分通过以下来基于所述电子文档的所述类别标识符生成针对所述电子文档的第二全文索引:向所述第二全文索引添加与所述文档类别相关的前缀部分,所述前缀部分将针对所述电子文档的所述第二全文索引与针对所述电子文档的所述第一全文索引进行区别;以及

将所述第一全文索引和所述第二全文索引两者存储到全文索引库,其中所述第一全文索引和所述第二全文索引两者与所述电子文档的单个可访问地址相关联地被存储在所述全文索引库中,所述电子文档的所述单个可访问地址在搜索项的关键词匹配所述第一全文索引的情况下和在所述搜索项的所述关键词匹配所述第二全文索引的情况下被返回给用户。

2.根据权利要求1所述的方法,其中分类所述电子文档包括以下各项中的至少一项:

基于与所述电子文档相关联的元数据来分类所述电子文档;以及

通过分析所述电子文档中的内容的语义来分类所述电子文档。

3.根据权利要求1或2所述的方法,其中分类所述电子文档包括:

确定所述电子文档是否属于预定类别;

响应于确定所述电子文档属于所述预定类别,将与所述预定类别相关联的类别标识符确定为所述电子文档的类别标识符。

4.根据权利要求3所述的方法,其中分类所述电子文档进一步包括:

响应于确定所述电子文档属于所述预定类别,确定所述电子文档是否属于所述预定类别的子类别;以及

响应于确定所述电子文档属于所述子类别,将与所述子类别相关联的类别标识符确定为所述电子文档的类别标识符。

5.根据权利要求3所述的方法,其中所述预定类别与对应的子类别以树结构被存储,并且

其中分类所述电子文档包括:

遍历所述树结构,以确定所述电子文档的类别标识符。

6.根据权利要求1所述的方法,进一步包括:

基于与所述电子文档相关联的元数据生成针对所述电子文档的第三全文索引;以及

将所述第三全文索引与所述电子文档的所述可访问地址相关联地存储在所述全文索引库中。

7.根据权利要求1所述的方法,其中所述第二全文索引包括:

描述部分,所述描述部分描述多个预定文档类别中的预定文档标识符。

8.根据权利要求1所述的方法,其中生成所述第二全文索引还包括:

生成与所述类别标识符相关的描述部分。

9.根据权利要求6所述的方法,其中所述元数据包括以下至少一项:文档作者、文档创建时间、文档创建地点、修改时间、文档大小、文档格式、文档语言、文档主题和文档地址。

10.根据权利要求1所述的方法，还包括：

获取用户输入的搜索项，所述搜索项至少包括与待搜索的电子文档的类别标识符有关的类别关键词；

将所述搜索项与预定义的多个全文索引进行匹配，所述多个全文索引至少包括所述第一全文索引和所述第二全文索引；以及

基于所匹配的全文索引，确定相关联的电子文档。

11.根据权利要求10所述的方法，其中所述搜索项进一步包括与所述待搜索的电子文档的内容有关的内容关键词。

12.根据权利要求10或11所述的方法，其中获取用户输入的搜索项包括：

向所述用户提供与预定类别对应的第一选项；以及

响应于所述用户对所述第一选项的选择，基于所述预定类别的类别标识符确定所述类别关键词。

13.根据权利要求12所述的方法，其中获取用户输入的搜索项进一步包括：

向所述用户提供与所述预定类别的子类别对应的第二选项；以及

响应于所述用户对所述第二选项的选择，基于所述子类别的类别标识符确定所述类别关键词。

14.根据权利要求10所述的方法，还包括：

将所述搜索项中的每个关键词与所述第一全文索引和所述第二全文索引中的每个全文索引相比较；以及

响应于包括所述第一全文索引和所述第二全文索引的集合中的一个全文索引包括所述搜索项中的关键词之一，确定所述搜索项中的所述关键词之一与包括所述第一全文索引和所述第二全文索引的集合中的一个全文索引相匹配。

15.一种用于全文搜索的系统，包括：

至少一个处理单元；以及

至少一个存储器，所述至少一个存储器耦合至所述至少一个处理单元并且在其上存储指令，所述指令在由所述至少一个处理单元执行时执行包括以下各项的动作：

基于所获得的电子文档的内容生成针对所述电子文档的第一全文索引；

分类所述电子文档以确定所述电子文档的类别标识符，所述类别标识符是分类结果，所述分类结果标识所述电子文档的文档类别；

至少部分通过以下来基于所述电子文档的所述类别标识符生成针对所述电子文档的第二全文索引：向所述第二全文索引添加与所述文档类别相关的前缀部分，所述前缀部分将针对所述电子文档的所述第二全文索引与针对所述电子文档的所述第一全文索引进行区别；以及

将所述第一全文索引和所述第二全文索引两者存储到全文索引库，其中所述第一全文索引和所述第二全文索引两者与所述电子文档的单个可访问地址相关联地被存储在所述全文索引库中，所述电子文档的所述单个可访问地址在搜索项的关键词匹配所述第一全文索引的情况下和在所述搜索项的所述关键词匹配所述第二全文索引的情况下被返回给用户。

16.根据权利要求15所述的系统，其中分类所述电子文档包括以下各项中的至少一项：

基于与所述电子文档相关联的元数据来分类所述电子文档;以及

通过分析所述电子文档中的内容的语义来分类所述电子文档。

17.根据权利要求15或16所述的系统,其中分类所述电子文档包括:

确定所述电子文档是否属于预定类别;

响应于确定所述电子文档属于所述预定类别,将与所述预定类别相关联的类别标识符确定为所述电子文档的类别标识符。

18.根据权利要求17所述的系统,其中分类所述电子文档进一步包括:

响应于确定所述电子文档属于所述预定类别,确定所述电子文档是否属于所述预定类别的子类别;以及

响应于确定所述电子文档属于所述子类别,将与所述子类别相关联的类别标识符确定为所述电子文档的类别标识符。

19.根据权利要求17所述的系统,其中所述预定类别与对应的子类别以树结构被存储,并且

其中分类所述电子文档包括:

遍历所述树结构,以确定所述电子文档的类别标识符。

20.根据权利要求15所述的系统,其中所述动作进一步包括:

基于与所述电子文档相关联的元数据生成针对所述电子文档的第三全文索引;以及

将所述第三全文索引与所述电子文档的所述可访问地址相关联地存储在所述全文索引库中。

21.根据权利要求15所述的系统,还包括:

获取用户输入的搜索项,所述搜索项至少包括与待搜索的电子文档的类别标识符有关的类别关键词;

将所述搜索项与预定义的多个全文索引进行匹配,所述多个全文索引至少包括所述第一全文索引和所述第二全文索引;以及

基于所匹配的全文索引,确定相关联的电子文档。

22.根据权利要求21所述的系统,其中所述搜索项进一步包括与所述待搜索的电子文档的内容有关的内容关键词。

23.根据权利要求21或22所述的系统,其中获取用户输入的搜索项包括:

向所述用户提供与预定类别对应的第一选项;以及

响应于所述用户对所述第一选项的选择,基于所述预定类别的类别标识符确定所述类别关键词。

24.根据权利要求23所述的系统,其中获取用户输入的搜索项进一步包括:

向所述用户提供与所述预定类别的子类别对应的第二选项;以及

响应于所述用户对所述第二选项的选择,基于所述子类别的类别标识符确定所述类别关键词。

25.根据权利要求15所述的系统,还包括

全文索引库,被配置为存储所述第一全文索引和所述第二全文索引。

26.一种计算机可读存储介质,所述计算机可读存储介质具有存储在其上的计算机可读程序指令,所述计算机可读程序指令用于执行根据权利要求1到14中任一项所述的方法

的步骤。

基于分类的全文搜索

技术领域

[0001] 本公开内容的各种实施方式涉及全文搜索领域，并且更具体地，涉及用于基于分类的全文搜索的方法、设备和系统。

背景技术

[0002] 随着互联网以及数据库技术的快速发展，对信息的搜索已经成为广泛存在的需求。全文搜索(full text search)是信息搜索领域中一种越来越受欢迎的搜索方法。

[0003] 通常在全文搜索系统中，搜索引擎将电子文档的内容解析成全文索引并且将全文索引存储在索引库中。每个全文索引可以包括电子文档的一个或多个字、词、符号或句子。在使用过程中，搜索引擎使用用户输入的关键字在索引库中进行搜索，并且返回与匹配的全文索引对应的电子文档。然而，这种搜索过程返回的搜索结果通常难以使得用户满意，特别是当索引库中存储有大量的电子文档的全文索引时。

发明内容

[0004] 本公开内容的多种实施方式提供了一种基于分类的全文搜索的方案。

[0005] 根据本公开内容的第一方面，提供了一种用于全文搜索的方法。该方法包括基于所获得的电子文档的内容生成第一全文索引。该方法还包括分类电子文档以确定电子文档的类别标识符，以及基于类别标识符生成第二全文索引。该方法进一步包括存储第一全文索引和第二全文索引。

[0006] 根据本公开内容的第二方面，提供了一种用于全文搜索的方法。该方法包括获取用户输入的搜索项，搜索项至少包括与待搜索的电子文档的类别标识符有关的类别关键词。该方法还包括将搜索项与预定义的多个全文索引进行匹配。多个全文索引至少包括第一全文索引，第一全文索引与通过分类至少一个电子文档而确定的类别标识符有关。该方法进一步包括基于所匹配的全文索引，确定相关联的电子文档。

[0007] 根据本公开内容的第三方面，提供了一种用于全文搜索的设备。该设备包括至少一个处理单元；以及至少一个存储器。至少一个存储器耦合至至少一个处理单元并且在其上存储指令，指令在由至少一个处理单元执行时执行包括以下各项的动作：基于所获得的电子文档的内容生成第一全文索引；分类电子文档以确定电子文档的类别标识符；基于类别标识符生成第二全文索引；以及存储第一全文索引和第二全文索引。

[0008] 根据本公开内容的第四方面，提供了一种用于全文搜索的设备。该设备包括至少一个处理单元；以及至少一个存储器。至少一个存储器耦合至至少一个处理单元并且在其上存储指令，指令在由至少一个处理单元执行时执行包括以下各项的动作：获取用户输入的搜索项，搜索项至少包括与待搜索的电子文档的类别标识符有关的类别关键词；将搜索项与预定义的多个全文索引进行匹配，多个全文索引至少包括第一全文索引，第一全文索引与通过分类至少一个电子文档而确定的类别标识符有关；以及基于所匹配的全文索引，确定相关联的电子文档。

[0009] 根据本公开内容的第五方面，提供了一种用于全文搜索的系统。该系统包括根据第三方面描述的用于全文搜索的设备。该系统还包括根据第四方面描述的用于全文搜索的设备。该系统进一步包括全文索引库，被配置为存储第一全文索引和第二全文索引。

[0010] 根据本公开内容的第六方面，提供了一种计算机可读存储介质。该计算机可读存储介质具有存储在其上的计算机可读程序指令。这些计算机可读程序指令用于执行根据以上第一方面所描述的方法的步骤。

[0011] 根据本公开内容的第七方面，提供了一种计算机可读存储介质。该计算机可读存储介质具有存储在其上的计算机可读程序指令。这些计算机可读程序指令用于执行根据以上第二方面所描述的方法的步骤。

[0012] 提供发明内容部分是为了简化的形式来介绍对概念的选择，它们在下文的具体实施方式中将被进一步描述。发明内容部分无意标识本公开内容的关键特征或主要特征，也无意限制本公开内容的范围。

附图说明

[0013] 通过结合附图对本公开示例性实施例进行更详细的描述，本公开的上述以及其它目的、特征和优势将变得更加明显，其中，在本公开示例性实施例中，相同的参考标号通常代表相同部件。

[0014] 图1示出了可以在其中实施本公开内容的多个实施例的环境的示意图；

[0015] 图2示出了根据本公开内容的实施例的用于全文搜索的方法的流程图；

[0016] 图3A-3B示出了被存储为树结构的两个类别及其子类别的示意图；

[0017] 图4示出了根据本公开内容的实施例的用于全文搜索的方法的流程图；以及

[0018] 图5示出了可以用来实施本公开内容的实施例的示例设备的示意性框图。

具体实施例

[0019] 下面将参照附图更详细地描述本公开的优选实施例。虽然附图中显示了本公开的优选实施例，然而应该理解，可以以各种形式实现本公开而不应被这里阐述的实施例所限制。相反，提供这些实施例是为了使本公开更加透彻和完整，并且能够将本公开的范围完整地传达给本领域的技术人员。

[0020] 在本文中使用的术语“包括”及其变形表示开放性包括，即“包括但不限于”。除非特别申明，术语“或”表示“和/或”。术语“基于”表示“至少部分地基于”。术语“一个示例实施例”和“一个实施例”表示“至少一个示例实施例”。术语“另一实施例”表示“至少一个另外的实施例”。术语“第一”、“第二”等等可以指代不同的或相同的对象。下文还可能包括其他明确的和隐含的定义。

[0021] 图1示出了可以在其中实施本公开内容的多个实施例的环境100的示意图。环境100包括全文搜索系统110，可以用于对一个或多个电子文档进行索引化，并且可以向用户提供搜索服务。全文搜索系统110可以包括索引处理设备112，其被配置为针对获得的电子文档，生成全文索引。索引处理设备112还可以将生成的全文索引存储至全文索引库120中。如本文中所使用的，术语“电子文档”指的是任何机器可读格式的文件，包括但不限于pdf文件、txt文件、各种office文件、各种网页文件等等。全文搜索系统110可以从各种数据源获

得电子文档。例如，全文搜索系统110可以从各个网站(未示出)抓取网页文件。在某些示例中，还可以由用户终端、例如终端A 132和/或终端B 134向全文搜索系统110提供各种电子文档。

[0022] 除了向全文搜索系统110提供电子文档或者取而代之，终端A 132和/或终端B 134可以利用全文搜索系统110查询期望获得的电子文档。例如，终端A 132和/或终端B 134可以将用户输入的查询关键词发送给全文搜索系统110。全文搜索系统110的查询处理设备114可以使用查询关键词，在全文索引库120中查找匹配的全文索引，并且然后将匹配的全文索引对应的电子文档提供给相应的终端。在一些情况中，查询处理设备114可以将查找到的电子文档的地址提供给相应的终端，从而使得终端的用户可以根据该地址获得对应的电子文档。在一些实施例中，终端A 132和/或终端B 134可以经由有线和/或无线连接而连接至全文搜索系统110。终端A 132和/或终端B 134可以任意类型的移动终端、固定终端或便携式终端。

[0023] 应当认识到的是，虽然被示出为两个分离的设备，在一些实施例中，索引处理设备112和查询处理设备114可以由单个设备、例如服务器、计算设备等来实现。在另外一些实施例中，索引处理设备112或查询处理设备114还可以由多个设备、例如服务器、计算设备等来实现。全文搜索系统110有时也可以被称为搜索引擎。

[0024] 在已有的全文搜索系统中，电子文档的内容被解析为一个或多个全文索引，其中每个全文索引可以包括电子文档的一个或多个字、词、符号或句子。用户输入的关键词被用于与全文索引相匹配，以便查询电子文档。如先前提及的，这种全文搜索方法难以返回用户期望的电子文档。在一些情况下，通过利用关键词来匹配全文索引，将会返回大量的电子文档，使得用户难以从中准确地获得期望的内容。例如，如果用户期望查询由“Tom”撰写的与“数据存储”领域中的“备份恢复”有关的电子文档，他可能会尝试输入关键词“数据存储备份修复Tom”。全文搜索系统根据用户输入的关键词，可能会返回大量的与其他数据存储领域的其他方面或者由其他作者撰写的电子文档。这样的搜索结果是不准确的，严重影响用户体验。

[0025] 根据本公开内容的实施例，提供了一种用于全文搜索的方案。在创建全文索引时，除了基于电子文档的内容生成全文索引之外，还对电子文档进行分类，以基于分类结果生成另外的全文索引。与文档内容有关的全文索引和与文档类别有关的全文索引均被存储到例如全文索引库中。在用户使用时，用户可以选择期望的文档类别。与文档类别有关的信息被用作搜索关键词可以与用户输入的其他与文档内容有关的关键词一起，被用于查询全文索引库。通过这种方式，可以从全文索引库中查找到与文档类别并且还与文档内容对应的电子文档，从而减少了搜索结果的范围并且提高了搜索结果的准确度。

[0026] 现在参照图2，其中示出了根据本公开内容的实施例的用于全文搜索的方法200的流程图。方法200可以用于创建全文索引，并且可以被实施在例如全文搜索系统110的索引处理设备112处。理解的是，方法200还可以包括附加的步骤和/或省略执行示出的步骤。本公开内容的范围在此方面不受限制。

[0027] 在步骤210处，基于所获得的电子文档的内容生成第一全文索引。第一全文索引是与文档内容有关的全文索引。在一些实施例中，全文搜索系统110例如可以主动从各种数据源获取新创建或者更新后的电子文档。备选地或附加地，各种数据源可以主动向全文搜索

系统110传输新创建或者更新后的电子文档。电子文档可以是任何机器可读格式的文件并且可以包括任何人类或机器语言的内容。全文搜索系统110的索引处理设备112可以例如提取电子文档的内容并且将电子文档的内容划分成一个或多个全文索引，每个全文索引可以包括一个或多个字、词、符号或句子。应当认识到，可以采用目前已知的或者未来开发的各种技术来将电子文档的内容分解成全文索引。

[0028] 接下来，方法200行进至步骤220，其中电子文档被分类以确定电子文档的类别标识符。在一些实施例中，可以预先设置一个或多个文档类别。这些文档类别可以基于对所获得的电子文档的分析来设置。备选地或附加地，还可以由全文搜索系统110的用户或管理者来设置。应当理解的是，本公开内容对于文档的分类没有具体的限定，可以从各个方面对文档进行分类。作为示例但并非限制性地，可以依据文档的作者、创建时间、创建地点、修改时间、文档大小、文档格式、文档的语言、文档主题和文档的可访问地址中的一项或者多项来确定电子文档属于一个或者多个类别。

[0029] 在一些实施例中，可以获得电子文档的元数据，并且可以基于与电子文档相关联的元数据来分类电子文档。电子文档的元数据可以包括与电子文档有关的各种说明性信息。电子文档的元数据可以包括但不限于文档的作者、创建时间、创建地点、修改时间、文档大小、文档格式、文档的语言、文档主题、文档的可访问地址等等。电子文档的元数据可以变化，并且每个电子文档的元数据的信息的类型也可以不同。在一些实施例中，元数据可以从文档的数据源获得。文档的创建者也可以规定该文档的元数据中的一项或多项。

[0030] 备选地或附加地，可以通过分析电子文档中的内容的语义来分类电子文档。可以利用各种目前已知的或者将来要开发的技术来从文档的内容的语义中确定文档的类别。作为示例，可以通过分析文档内容来确定文档主题，诸如确定文档属于哪个知识领域。在另一个示例中，还可以通过语义分析确定文档的语言，例如属于中文、英文、或者其他人类或机器语言。在其他实施例中，还可以由全文搜索系统110的用户或管理者手动划分电子文档的类别。

[0031] 因此，在一些实施例中，基于所获得的电子文档相关联的元数据或者语义分析的结果，可以将该文档划分到相应的预定类别中。例如，如果预先设置了与文档的作者、创建时间、创建地点、修改时间、文档大小、文档格式和/或文档主题有关的类别，则可以依据当前文档中的元数据所包含的信息，确定文档的相应类别。在一些实施例中，可以仅预先设置文档的分类规则，并且然后依据所获得的文档相关联的元数据来创建对应的文档。例如，可以设置对文档作者进行分类的规则。如果新获得的文档的作者属于之前创建的某个作者有关的类别，则将该文档划分到已有的类别中。如果新获得的文档的作者不存在，则可以创建新的作者分类，并且将该文档划分到该新创建的类别中。在一些实施例中，还可以预先设置多个类别的划分准则，并且然后依据该准则来分类电子文档。例如，可以将文档大小划分为巨大、大、中、小和空五个类别。依据新获得的文档的大小，将该文档与五个类别之一相关联。

[0032] 在一些实施例中，还可以预先定义多个类别，并且可以确定电子文档是否属于这些类别中的一个或多个类别。通常多个类别可以从多个方面来分类该电子文档。在一些实施例中，电子文档可以以更精细的方式来划分。对于预定的类别中的一个或多个类别，还可以继续划分成一个或多个子类别。因此，在确定新获得的电子文档属于某个大的类别时，如

果该类别还存在一个或多个子类别，则可以继续确定该电子文档是否属于某个子类别。例如，对于某个文档主题的类别，还可以继续定义该主题下的多个更细的主题。应当理解的是，还可以将一个或者多个子类别继续细分，并且本公开内容的范围在此方面不受限制。

[0033] 在一些实施例中，每个类别和子类别可以具有相关联的类别标识符，用以在这些类别和子类别之间进行区别。例如，对于文档作者的类别，可以将作者的名字作为每个类别的标识符。对于其他类别，也可以类似地进行分配类别标识符。在一些实施例中，当确定所获得的电子文档属于一个或多个类别之后，将这些类别的标识符确定为电子文档的标识符。如果电子文档即属于某个大的类别，又属于该类别下的某个子类别，则可以将该类别和子类别的标识符均确定为该电子文档的标识符。

[0034] 在一些实施例中，每个预定类别及其子类别可以以树结构被存储。树结构的根节点可以描述该类别，并且每个预定类别及其子类别可以被认为是树结构中的子节点。这个树结构也可以被称为决策树。当获得新的电子文档时，通过遍历每个树结构、例如遍历树结构中的每个节点，可以方便地确定该电子文档是否属于该类别或者子类别。在一些实施例中，每个树结构可以被存储为一个文件。其他实施例中，多个树结构还可以被存储为一个文件。

[0035] 图3A-3B图示了被存储为树结构310-320的两个类别及其子类别的示意图。在图3A中，树结构310与文档作者的类别有关，其中根节点312描述该树结构，并且子节点314和316指示两个类别。在图3B中，树结构320与文档主题的类别有关，其中根节点322描述该树结构，并且子节点324指示一个类别。通过遍历树结构310和320，可以确定电子文档是否属于与某个作者有关的类别，或者其包括的内容是否属于某个主题以及该主题下的子主题。

[0036] 在一些实施例中，树结构还可以动态地增加。例如，如果确定电子文档的作者不属于已有的作者类别中的任何一个，则可以一个节点，该节点与该作者的类别有关。然后还可以将该电子文档划分到该类别中。

[0037] 继续参考图2，方法200行进至步骤230，其中基于类别标识符生成第二全文索引。第二全文索引是与文档类别有关的索引。在一些实施例中，可以避免第二全文索引与第一全文索引相同。例如，在一些示例中，通过电子文档的内容得到的第一全文索引可能包括与文档作者的名字有关的词语。为了避免后续可能的搜索错误，可以将与文档类别有关的第二全文索引确定为能够区别于第一全文索引。例如，可以为第二全文索引增添前缀，用于区别于文档内容有关的第一全文索引。

[0038] 在一个实施例中，第二全文索引可以包括前缀部分和描述部分，其中前缀部分可以是用于区别文档内容有关的索引和文档类别有关的索引，并且描述部分用于描述文档的类别标识符。举例而言，如果确定电子文档属于作者为“Tom”的类别，则可以生成与文档作者的类别有关的前缀部分“DT_AUTHOR”和与该类别的标识符有关的描述部分“Tom”。在一些实施例中，也可以将预定的类别或子类别的标识符确定为能够与第一全文索引区别，并且因此可以将类别标识符直接确定为第二全文索引。例如，可以将“DT_AUTHOR_Tom”作为作者为“Tom”的类别的标识符并且因此可以将它直接用作第二全文索引。

[0039] 应当认识到的是，如果在步骤220中确定电子文档属于多个类别或者一个或多个子类别，则还可以以类似的方式基于每个类别或子类别的标识符来生成相应的第二全文索引。

[0040] 在方法200的步骤240中，可以存储第一全文索引和第二全文索引。例如，全文搜索系统110的索引处理设备112可以将第一和第二全文索引存储至全文索引库120中。在一些实施例中，还可以将电子文档的可访问地址与第一和第二全文索引相关联地存储。在另外一些实施例中，还可以将电子文档的原始内容与第一和第二全文索引相关联地存储。通过这样的方式，当依据第一或第二全文索引搜索到该电子文档时，可以将该电子文档的地址或者内容呈现给用户以供用户访问。

[0041] 在一些实施例中，还可以基于与电子文档相关联的元数据生成第三全文索引，并且存储第三全文索引。例如，第三全文索引可以与第一和第二全文索引一起被存储至全文索引库120中。应当理解的是，第三全文索引可以包括元数据所包括的内容中的一个或多个字、词、字符或句子。

[0042] 以上参照图2描述了创建全文索引的过程。每当接收到新的电子文档时，均可以根据图2的方法200为该电子文档创建全文索引。接下来将参照图4描述基于建立的全文索引进行搜索的方法400。方法400可以被实施在例如全文搜索系统110的查询处理设备114处。理解的是，方法400还可以包括附加的步骤和/或省略执行示出的步骤。本公开内容的范围在此方面不受限制。

[0043] 在步骤410处，获取用户输入的搜索项。用户可以经由终端发出查询请求，并且给出相应的搜索项。在一些实施例中，搜索项可以包括待搜索的电子文档的内容有关的内容关键词，指示用户期望获得其内容中包括所指定的关键词的电子文档。在一些实施例中，该搜索项还包括与待搜索的电子文档的类别标识符有关的类别关键词。当基于文档分类来创建全文索引之后，可以为用户提供用户接口，以使得用户可以选择相应的类别。在一些实施例中，可以例如经由用户所使用的终端中的用户界面接口提供与预定的一个或多个类别对应的选项。用户可以通过选择这些选项来确定期望获取的电子文档的类别。在一些实施例中，对于包括一个或多个子类别的大类别，还可以继续向用户提供与子类别对应的选项以供选择。所提供的选项可以由对应的类别或子类别的标识符来指示。

[0044] 在一些实施例中，响应于用户对一个或多个选项的选择，可以基于对应的类别或子类别的标识符来确定类别关键词。不同于与文档的内容有关的内容关键词(其可能是用户直接输入的关键词)，类别关键词可以通过用户对于文档的类别或子类别的选择来生成。例如，如果用户期望获得作者“Tom”撰写的文档并且选择了与该作者类别对应的选项，则可以生成“DT_AUTHOR_Tom”作为类别关键词。除了向用户提供选项以供选择之外或者备选地，用户还可以直接输入与所创建的文档类别全文索引的类型相似的关键词，例如“DT_AUTHOR_Tom”，以便于获得在该类别中的电子文档。

[0045] 接下来，在方法400的步骤420中，将搜索项与预定义的多个全文索引进行匹配。如以上关于方法200所描述的，多个全文索引可以包括与文档内容有关的第一全文索引和与文档类别有关的第二内容索引。在一些实施例中，可以将搜索项中的每个关键词、包括文档内容关键词和类别关键词与每个全文索引进行比较。如果该全文索引中包括一个或多个关键词，则可以确定该全文索引与该关键词相匹配。

[0046] 在一些实施例中，可以设置搜索项的关键词之间的约束关系。例如，多个文档内容的关键词之间可以是“和”或者“或”的关系。多个文档类别的关键词之间可以是“和”或者“或”的关系，并且基于每个类别之下的子类别确定的关键词可以是“或”的关系。在一些实

施例中，可以基于这些约束关系在每个电子文档相关联的多个全文索引中进行匹配。作为一个示例，假设用户输入文档内容关键词“速度提高”和“存储空间有效”，并且用户还选择了作者类别“Tom”和文档主题类别“数据存储”以及该类别之下的子类别“备份恢复”和“性能提升”。在获取相应的类别关键词之后，可以为在每个电子文档对应的全文索引中查找与“Tom”和“数据存储”和“备份恢复”或“性能提升”有关的关键词匹配的全文索引，此外还要确定该电子文档的全文索引是否还包括与“速度提高”和“存储空间有效”这两个内容关键词匹配的全文索引。如果在某个电子文档的全文索引中能够查找到与搜索项的类别关键词以及文档内容关键词匹配时，则可以确定这些全文索引对应的电子文档与用户的搜索项匹配。在一些对于搜索精度要求不高的情况中，如果某个电子文档的全文索引与一个或多个关键词匹配，也可以确定结果是匹配的。

[0047] 在方法400的步骤430中，基于所匹配的全文索引，确定相关联的电子文档。通过使用搜索项来查找全文索引，如果查找到满足条件的全文索引，则可以将该索引对应的电子文档作为搜索结果返回给用户。在一些实施例中，可以将电子文档的可访问地址返回给用户。在一些实施例中，可以按照匹配度向用户提供搜索结果。匹配度可以根据电子文档相关联的全文索引与搜索项中的关键词匹配的数目来确定。

[0048] 以上参照图2和4描述了本公开内容的各种实施例。通过本公开内容的全文搜索方法，可以为用户提供更准确的搜索结果。在一些实施例中，由于为电子文档进行分类，可以检索出文档内容是空白的电子文档。因为虽然由于文档内容是空白的而无法生成文档内容有关的全文索引，但是可以依据分类结果为该电子文档生成对应的文档类别全文索引。用户在搜索时可以通过定义相应的类别来查询该文档。

[0049] 图5示出了可以用来实施本公开内容的实施例的示例设备500的示意性框图。设备500可以用于实现图1的索引处理设备112和/或查询处理设备114。如图所示，设备500包括中央处理单元(CPU)501，其可以根据存储在只读存储器(ROM)502中的计算机程序指令或者从存储单元508加载到随机访问存储器(RAM)503中的计算机程序指令，来执行各种适当的动作和处理。在RAM 503中，还可存储设备500操作所需的各种程序和数据。CPU 501、ROM 502以及RAM 503通过总线504彼此相连。输入/输出(I/O)接口505也连接至总线504。

[0050] 设备500中的多个部件连接至I/O接口505，包括：输入单元506，例如键盘、鼠标等；输出单元507，例如各种类型的显示器、扬声器等；存储单元508，例如磁盘、光盘等；以及通信单元509，例如网卡、调制解调器、无线通信收发机等。通信单元509允许设备500通过诸如因特网的计算机网络和/或各种电信网络与其他设备交换信息/数据。

[0051] 上文所描述的各个方法和处理，例如方法200和/或400，可由处理单元501执行。例如，在一些实施例中，方法200和/或400可被实现为计算机软件程序，其被有形地包含于机器可读介质，例如存储单元508。在一些实施例中，计算机程序的部分或者全部可以经由ROM 502和/或通信单元509而被载入和/或安装到设备500上。当计算机程序被加载到RAM 503并由CPU 501执行时，可以执行上文描述的方法200和/或400的一个或多个步骤。

[0052] 本公开内容可以是系统、方法和/或计算机程序产品。计算机程序产品可以包括计算机可读存储介质，其上载有用于执行本公开内容的各个方面的计算机可读程序指令。

[0053] 计算机可读存储介质可以是可以保持和存储由指令执行设备使用的指令的有形设备。计算机可读存储介质例如可以是——但不限于——电存储设备、磁存储设备、光存储

设备、电磁存储设备、半导体存储设备或者上述的任意合适的组合。计算机可读存储介质的更具体的例子(非穷举的列表)包括:便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、静态随机存取存储器(SRAM)、便携式压缩盘只读存储器(CD-ROM)、数字多功能盘(DVD)、记忆棒、软盘、机械编码设备、例如其上存储有指令的打孔卡或凹槽内凸起结构、以及上述的任意合适的组合。这里所使用的计算机可读存储介质不被解释为瞬时信号本身,诸如无线电波或者其他自由传播的电磁波、通过波导或其他传输媒介传播的电磁波(例如,通过光纤电缆的光脉冲)、或者通过电线传输的电信号。

[0054] 这里所描述的计算机可读程序指令可以从计算机可读存储介质下载到各个计算/处理设备,或者通过网络、例如因特网、局域网、广域网和/或无线网下载到外部计算机或外部存储设备。网络可以包括铜传输电缆、光纤传输、无线传输、路由器、防火墙、交换机、网关计算机和/或边缘服务器。每个计算/处理设备中的网络适配卡或者网络接口从网络接收计算机可读程序指令,并转发该计算机可读程序指令,以供存储在各个计算/处理设备中的计算机可读存储介质中。

[0055] 用于执行本公开内容操作的计算机程序指令可以是汇编指令、指令集架构(ISA)指令、机器指令、机器相关指令、微代码、固件指令、状态设置数据、或者以一种或多种编程语言的任意组合编写的源代码或目标代码,所述编程语言包括面向对象的编程语言—诸如Smalltalk、C++等,以及常规的过程式编程语言—诸如“C”语言或类似的编程语言。计算机可读程序指令可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中,远程计算机可以通过任意种类的网络—包括局域网(LAN)或广域网(WAN)—连接到用户计算机,或者,可以连接到外部计算机(例如利用因特网服务提供商来通过因特网连接)。在一些实施例中,通过利用计算机可读程序指令的状态信息来个性化定制电子电路,例如可编程逻辑电路、现场可编程门阵列(FPGA)或可编程逻辑阵列(PLA),该电子电路可以执行计算机可读程序指令,从而实现本公开内容的各个方面。

[0056] 这里参照根据本公开内容实施例的方法、装置(系统)和计算机程序产品的流程图和/或框图描述了本公开内容的各个方面。应当理解,流程图和/或框图的每个方框以及流程图和/或框图中各方框的组合,都可以由计算机可读程序指令实现。

[0057] 这些计算机可读程序指令可以提供给通用计算机、专用计算机或其它可编程数据处理装置的处理器,从而生产出一种机器,使得这些指令在通过计算机或其它可编程数据处理装置的处理器执行时,产生了实现流程图和/或框图中的一个或多个方框中规定的功能/动作的装置。也可以把这些计算机可读程序指令存储在计算机可读存储介质中,这些指令使得计算机、可编程数据处理装置和/或其他设备以特定方式工作,从而,存储有指令的计算机可读介质则包括一个制造品,其包括实现流程图和/或框图中的一个或多个方框中规定的功能/动作的各个方面的指令。

[0058] 也可以把计算机可读程序指令加载到计算机、其它可编程数据处理装置、或其它设备上,使得在计算机、其它可编程数据处理装置或其它设备上执行一系列操作步骤,以产生计算机实现的过程,从而使得在计算机、其它可编程数据处理装置、或其它设备上执行的

指令实现流程图和/或框图中的一个或多个方框中规定的功能/动作。

[0059] 附图中的流程图和框图显示了根据本公开内容的多个实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段或指令的一部分，所述模块、程序段或指令的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。在有些作为替换的实现中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个连续的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的是，框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或动作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

[0060] 以上已经描述了本公开内容的各实施例，上述说明是示例性的，并非穷尽性的，并且也不限于所披露的各实施例。在不偏离所说明的各实施例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。本文中所用术语的选择，旨在最好地解释各实施例的原理、实际应用或对市场中的技术的技术改进，或者使本技术领域的其它普通技术人员能理解本文披露的各实施例。

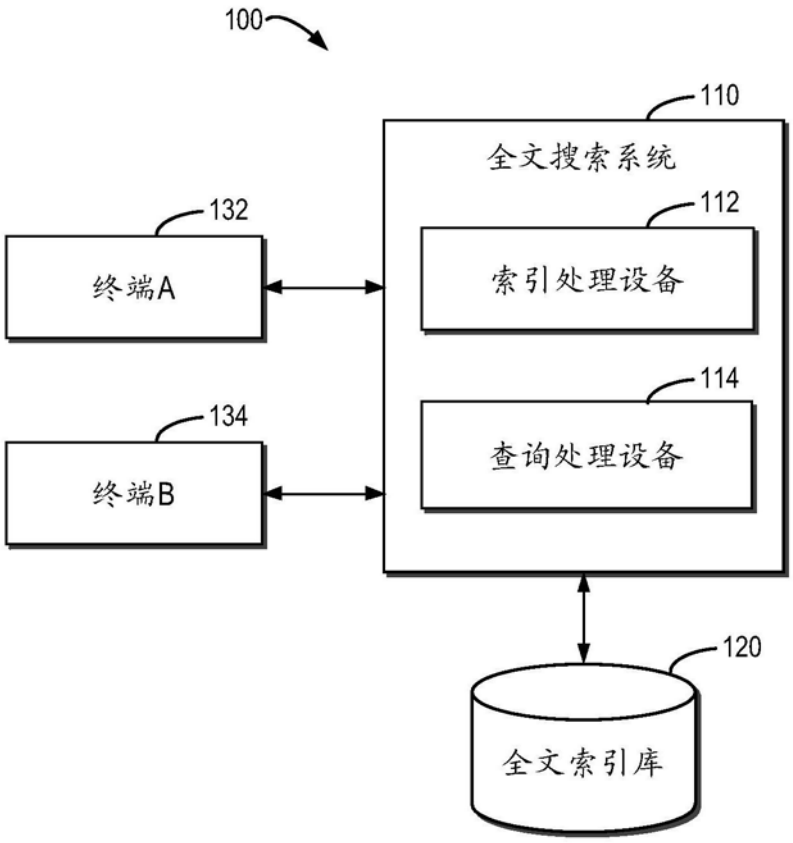

100
110
全文搜索系统
132
终端A
112
索引处理设备
114
查询处理设备
134
终端B
120
全文索引库

图1

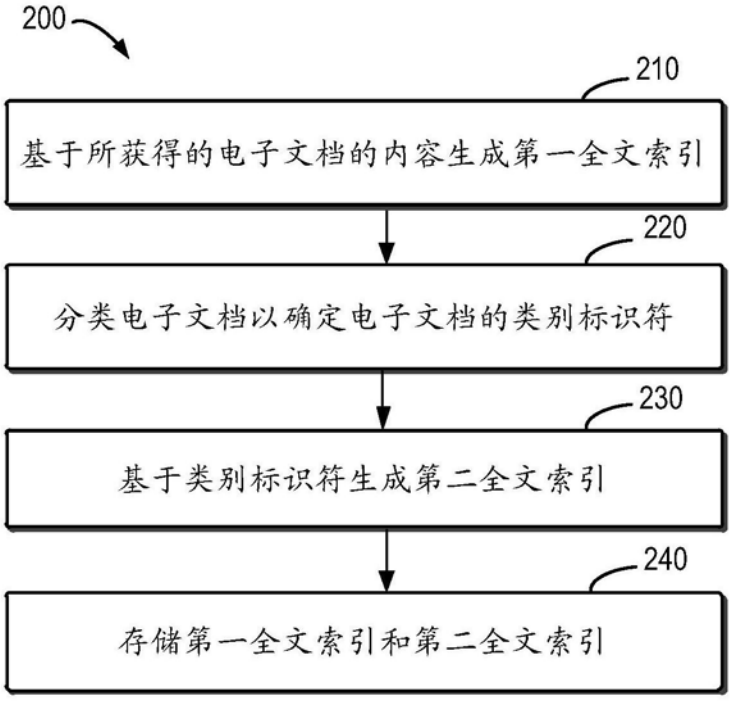

200
210
基于所获得的电子文档的内容生成第一全文索引
220
分类电子文档以确定电子文档的类别标识符
230
基于类别标识符生成第二全文索引
240
存储第一全文索引和第二全文索引

图2

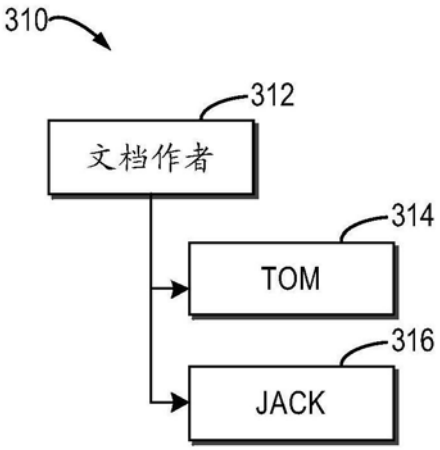

310
312
文档作者
314
TOM
316
JACK

图3A

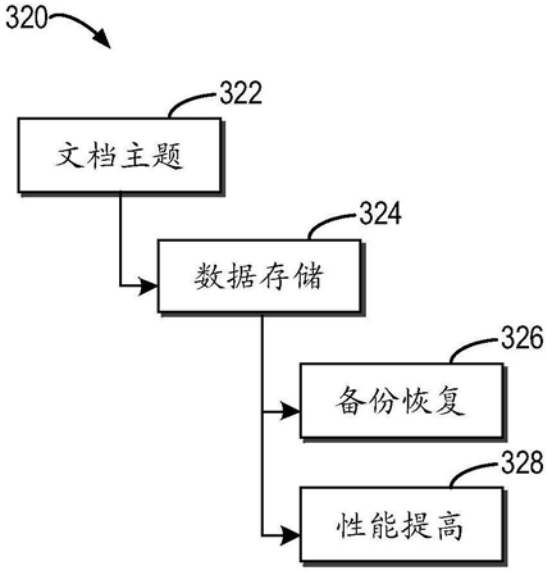

320
322
文档主题
324
数据存储
326
备份恢复
328
性能提高

图3B

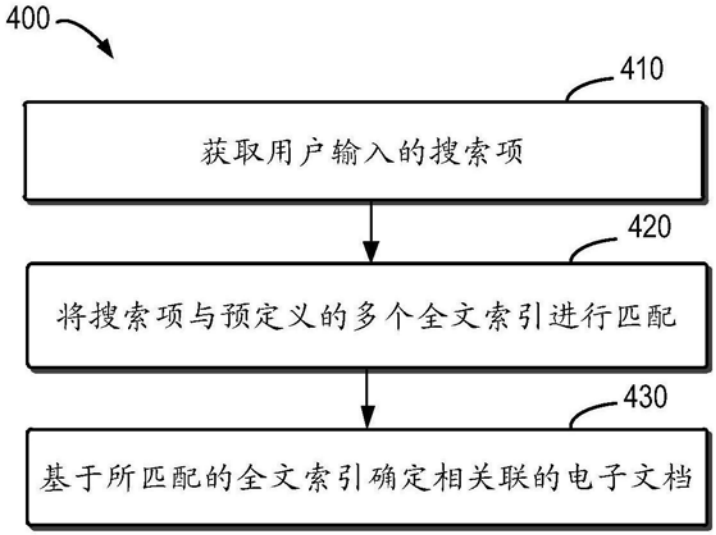

400
410
获取用户输入的搜索项
420
将搜索项与预定义的多个全文索引进行匹配
430
基于所匹配的全文索引确定相关联的电子文档

图4

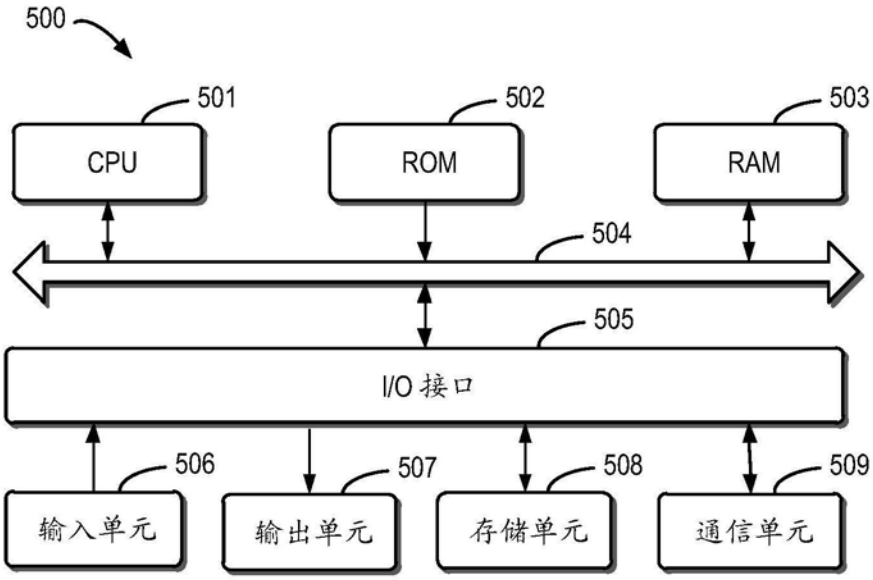

500
501
CPU
502
ROM
503
RAM
504
505
I/O 接口
506
输入单元
507
输出单元
508
存储单元
509
通信单元

图5