(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 102521270 A
(43)申请公布日 2012.06.27

(21)申请号 201110373345.8

(22)申请日 2011.11.22

(30)优先权数据
12/951659 2010.11.22 US

(71)申请人 微软公司
地址 美国华盛顿州

(72)发明人 K.M. 里斯维克 M. 霍普克罗夫特
J.G. 贝内特 K. 卡尔亚纳拉曼
T. 基林比 V. 帕里克

(74)专利代理机构 中国专利代理(香港)有限公司 72001
代理人 孙之刚 刘鹏

(51)Int.Cl.
*G06F 17/30*(2006.01)

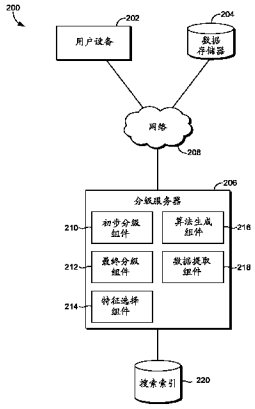
权利要求书 2 页 说明书 10 页 附图 5 页

(54)发明名称

用于有效预先计算的可分解的分级

(57)摘要

本发明涉及用于有效预先计算的可分解的分级。提供了方法和计算机存储介质以生成用于为候选文件提供初步分级的算法。为文件提供最终分级的最终分级功能被分析以识别潜在的初步分级特征，例如查询独立的静态分级特征和与单个原子相关的动态原子隔离的组件。基于多种因素从潜在的初步分级特征中选择初步分级特征。使用这些选择的特征，生成一种算法以在将最相关的文件传送至最终分级阶段之前为候选文件提供初步分级。




CN 102521270 A

1. 一种用于生成算法的方法，该算法被用于为多个文件提供初步分级，该方法包括：

分析（310）用于为多个文件计算最终分级的最终分级功能；

从最终分级功能中，识别（312）潜在的初步分级特征，包括一个或多个独立于查询的静态分级特征和一个或多个与单个原子相关的动态原子隔离的组件；

从潜在的初步分级特征中选择（314）一个或多个初步分级特征以用于初步分级功能；以及

使用至少一个或多个初步分级特征，生成（316）用于为多个文件提供初步分级的算法。

2. 如权利要求 1 所述的方法，其中一个或多个初步分级特征是手动选择的，因此需要用户交互作用。

3. 如权利要求 1 所述的方法，其中一个或多个初步分级特征是至少部分地通过机器学习工具选择的。

4. 如权利要求 1 所述的方法，其中一个或多个初步分级特征是基于以下的一个或多个而选择：

（1）将分级特征用于初步分级功能的容易度，

（2）由在该多个文件的初步分级和最终分级之间的逼真度测量确定的分级特征的有用性，

（3）当初步分级功能被改变时，分级特征的适应性，或

（4）相比于其它分级特征，计算该分级特征的低成本。

5. 如权利要求 1 所述的方法，其中初步分级功能识别多个文件的子集，其被发送至用于最终分级的最终分级功能。

6. 如权利要求 1 所述的方法，进一步包括：

接收搜索查询；

利用该算法，算法化地识别与该搜索查询最相关的多个文件的子集；和

将与该多个文件的子集关联的文件标识传送至用于最终分级的最终分级功能。

7. 如权利要求 1 所述的方法，其中提供初步分级的算法使用未在最终分级功能中使用的一个或多个分级特征。

8. 一种用于为文件计算初步分级的方法，该方法包括：

识别（410）独立于查询的静态分级特征；

识别（412）与单个原子相关的动态原子隔离的组件；

选择（414）初步分级特征的集合，其包括一个或多个静态分级特征和一个或多个动态原子隔离的组件；

对于第一文件，从搜索索引中提取（416）相应于该初步分级特征的集合的数据；和

基于搜索查询，利用（418）该提取出的数据计算第一文件的初步分级。

9. 如权利要求 8 所述的方法，进一步包括：

基于其它文件的初步分级，确定第一文件的相关度超过阈值；并

基于该第一文件的相关性超过阈值，将第一文件的标识发送至最终分级阶段，其为第一文件分配最终分级。

10. 如权利要求 9 所述的方法，其中最终分级阶段利用动态原子互相关组件以确定第一文件的最终分级。

11. 如权利要求 9 所述的方法，其中计算第一文件的初步分级的初步分级阶段利用在最终分级阶段中利用的一个或多个分级特征以计算第一文件的最终分级。

12. 如权利要求 8 所述的方法，其中提取出的数据包括与多个文件关联针对初步分级特征的集合预先计算的分数。

13. 如权利要求 12 所述的方法，其中预先计算的分数被存储在搜索索引中。

14. 如权利要求 8 所述的方法，其中动态原子的组件每个都与单个原子相关。

15. 一个或多个计算机存储介质，存储有计算机可用指令，当由计算设备使用时，该指令使得计算设备执行方法，该方法用于在初步分级阶段中利用来自最终分级阶段的分级特征来确定对于文件的初步分级，该方法包括：

分析（510）最终分级功能以识别分级特征的第一子集，其包括独立于查询的分级特征和单个原子分级特征；

选择（512）分级特征的第二子集，其未在最终分级功能中使用；

从分级特征的第一子集和第二子集中，选择（514）一个或多个初步分级特征以在使用初步分级功能计算多个文件的初步分级中使用，该初步分级功能限制使用最终分级功能分级的文件的数量；

至少基于从搜索索引中提取出的与分级特征的第一子集和第二子集关联的数据，使用初步分级功能算法化地识别（516）多个文件的子集；并

将相应于该多个文件的子集的文档标识传送（518）至最终分级阶段，最终分级阶段使用最终分级功能来计算该多个文件的子集中的每个文件的最终分级。

# 用于有效预先计算的可分解的分级

## 技术领域

[0001] 本申请涉及信息和文件搜索领域。

## 背景技术

[0002] 在互联网上可得到的信息和内容的数量非常快速地持续增长。给定巨大数量的信息，搜索引擎已经被开发以便于搜索电子文件。特别地，用户可以通过输入搜索查询以搜索信息和文件，该查询包括用户可能感兴趣的一个或多个术语 (term)。在从用户接收到搜索查询后，搜索引擎基于该搜索查询识别相关的文件和 / 或网页。因为其有用，网页搜索，即对于用户发出的搜索查询寻找相关网页和文件的过程，可以说已经成为互联网上当今最流行的业务。

[0003] 搜索引擎通过如下方式运行 :使用爬虫技术搜集 (crawling) 文件和在搜索索引中给关于该文件的信息编索引。当接收到搜索查询时，搜索引擎使用搜索索引以识别与该搜索查询相关的文件。例如，可以使用分级 (ranking) 功能以基于检索查询确定最相关的文件以呈现给用户。然而，分级功能已经变得日益复杂，这样数以百计的特征被用来分级文件。当单独使用时，复杂的分级功能由于成本和时间的约束是无效的。

[0004] 发明概述

提供这个概述以按照简化的形式介绍概念的选择，所述概念将在下面的详细描述部分被进一步描述。这个概述其意不在标识所请求保护的主题的关键特征或者必要特征，也不在被用来帮助确定所请求保护的主题的范围。

[0005] 本发明的实施方式涉及与整个分级过程的初步分级阶段结合使用的算法的生成。如下面进一步描述的，整个分级过程可以包括匹配阶段、初步分级阶段和最终分级阶段。可以给数以亿计或者甚至万亿计的文件编索引。因为最终分级功能通常比初步分级功能更加昂贵并且耗费时间，所以匹配阶段和初步分级阶段需要起作用以限制最终分级功能需要分级的候选文件的数量。通常，在初步分级阶段使用的初步分级功能是在最终分级阶段使用的最终分级功能的简化版本。这样，最终分级功能被分析以识别能够被预先计算的或者在接收查询后不容易被实时计算的分级特征 (例如文件分级特征) 以及容易被实时计算的分级特征。不在最终分级功能中使用的分级特征还可以在初步分级功能中被使用。一旦生成算法，其能够被用来计算文件的初步分级。

[0006] 附图简述

下面参考附图详细描述本发明，其中 :

图 1 是适于用来实现本发明实施方式的示例性计算环境的框图 ;

图 2 是在其中可以使用本发明的实施方式的示例性系统的框图 ;

图 3 是示出根据本发明的实施方式生成用于为多个文件提供初步分级的算法的方法的流程图 ;

图 4 是示出根据本发明的实施方式计算文件的初步分级的方法的流程图 ;和

图 5 是示出根据本发明的实施方式在初步分级阶段中利用来自最终分级阶段的分级

特征以确定文件的初步分级方法的流程图。

[0007] 详细描述

在这里带有特殊性地描述了本发明的主题以满足法定的要求。然而，该描述其自身并不意图限制本专利的范围。相反地，发明人已经预期到还可以结合其它现有或未来的技术按照其它方式具体化所请求保护的主题，以包括与本文档中描述的内容类似的不同步骤或步骤组合。此外，虽然在此处可以使用术语“步骤”和 / 或“框”表示所使用的方法的不同元素，但是该术语不被解释为意味了在此处公开的多个步骤之中或步骤之间有任何特定的顺序，除非明确描述了个别步骤间的顺序。

[0008] 如上面提及的，本发明的实施方式提供用于生成在整个分级过程的初步分级阶段中使用的算法。实施方式还提供用于使用该算法计算文件的初步分级，这样发送至最终分级组件的文件数量被极大地减少。如所提到的，初步分级功能通常是快速并且低成本的计算，其是最终分级功能的有用估计。初步分级功能在识别减少的相关文件集方面是能够被信任的，其对于更多代价的最终分级阶段是有价值的。这样，能够被预先计算的（例如文件分级特征）或者在接收到查询后不容易被实时计算的分级特征，诸如能够被最终分级功能使用的静态特征和动态原子隔离组件（atom-isolated component），被识别为由初步分级功能使用的潜在分级特征。这些识别的分级特征包括以下的组合 :在查询匹配时刻容易被计算的那些特征、在查询匹配时刻不容易被计算的并且能够被预先计算的那些特征、对于根据在估计最终分级中的逼真度 (fidelity) 度量的测量有用的那些特征、和即使初步分级功能被更改时仍保持有用方面有适用性的那些特征。对于原子 (atom) / 文件对的预先计算的分数被存储在搜索索引中并且在初步分级的计算期间被提取。被发现是最相关的文件被发送至最终分级阶段。利用逼真度测量以确保最终分级功能和初步分级功能相类似地分级文件以确定在两个分级阶段之间的逼真度和低错误率。

[0009] 因此，在一个方面，本发明的实施方式指向一种生成算法的方法，该算法被用来提供对多个文件的初步分级。该方法包括分析最终分级功能，该最终分级功能用于为多个文件计算最终分级。从最终分级功能中，该方法进一步包括识别包括一个或多个静态分级特征和一个或多个动态原子隔离组件的潜在初步分级特征，该静态分级特征是独立于查询的，该动态原子隔离组件与单个原子相关。附加地，该方法包括从潜在初步分级特征中选择一个或多个初步分级特征以用于初步分级功能，并使用至少一个或多个初步分级特征以生成一种用于提供对多个文件的初步分级的算法。

[0010] 在另一个实施方式中，本发明的一个方面指向一种用于计算文件的初步分级的方法。该方法包括识别独立于查询的静态分级特征，并识别与单个原子相关的动态原子隔离组件。进一步地，该方法包括选择一组初步分级特征，包括一个或多个静态分级特征和一个或多个动态原子隔离组件。对于第一文件，该方法从搜索索引中提取相应于初步分级特征的集合的数据。基于搜索查询，该方法使用该提取出的数据以计算第一文件的初步分级。

[0011] 本发明的另一个实施方式指向一个或多个计算机存储介质，存储有计算机可使用指令，当由计算设备使用时，其使得该计算设备执行用于在初步分级阶段中利用来自最终分级阶段的分级特征来确定文件的初步分级的方法。该方法包括分析最终分级功能以识别分级特征的第一子集，包括独立于查询的分级特征和单个原子分级特征，并选择在最终分级功能中未使用的分级特征的第二子集。进一步地，该方法包括从分级特征的第一子集和

第二子集中选择一个或多个初步分级特征以供使用初步分级功能计算多个文件的初步分级时使用，该初步分级功能限制了使用最终分级功能分级的文件的数量。至少基于从搜索索引中提取的与分级特征的第一子集和第二子集关联的数据，该方法使用初步分级功能算法化地识别多个文件的子集。该方法附加地包括将相应于多个文件的子集的文件标识通信至最终分级阶段，其使用最终分级功能以计算该多个文件的子集中的每个文件的最终分级。

[0012] 已经简要地描述了本发明实施方式的概况，在下面描述一种可以在其中实现本发明实施方式的示例性操作环境，以便为本发明的各个方面提供一种一般的背景。特别地，初步参考图 1，示出了一种用于实现本发明实施方式的示例性操作环境，一般地指定为计算设备 100。计算设备 100 仅是合适的计算环境的一个例子，并不意图暗示对本发明用途或功能性的范围的限制。计算设备 100 也不应被解释为对于所图示的组件中任意一个或其组合有任何依赖或要求。

[0013] 本发明可以在由计算机或其它机器（诸如个人数据助理或者其它手持设备）执行的计算机代码或包括计算机可执行指令（诸如程序模块）的机器可使用指令的一般背景下描述。通常，包括例程、程序、对象、组件、数据结构等的程序模块指的是执行特定任务或者实现特定抽象数据类型的代码。本发明可以在多种系统配置中实行，包括手持设备、消费电子、通用计算机、更专用的计算设备等。本发明还可以在分布式计算环境中实行，其中任务由通过通信网络链接的远程处理设备执行。

[0014] 参考图 1，计算设备 100 包括总线 110，其直接或间接地耦合下述设备：存储器 112、一个或多个处理器 114、一个或多个呈现组件 116、输入 / 输出（I/O）端口 118、输入 / 输出组件 120 和图示的电源 122。总线 110 代表一条或多条总线（诸如地址总线、数据总线或其组合）。虽然为了清楚的目的使用线条示出了图 1 中的多个框，但是事实上，描绘各个组件不是那么清楚，打个比方，线条更精确地将是灰色和模糊的。例如，可以认为诸如显示设备的呈现组件是 I/O 组件。而且，处理器具有存储器。发明人认识到这是现有技术的特征，并且重申图 1 的图仅是能够结合本发明的一个或多个实施方式使用的示例性计算设备的图示。在诸如“工作站”、“服务器”、“膝上型电脑”、“手持设备”等种类之间并不进行区分，因为所有这些都预期在图 1 的范围中并都被称为“计算设备”。

[0015] 计算设备 100 典型地包括多种计算机可读介质。计算机可读介质能够是可以由计算设备 100 存取的任何可用介质，并且包括易失和非易失的介质、可移除和非可移除的介质。通过示例和非限制的，计算机可读介质可以包括计算机存储介质和通信介质。计算机存储介质包括以任意方法或技术实现的易失和非易失、可移除和非可移除的介质，用于存储诸如计算机可读指令、数据结构、程序模块或其它数据之类的信息。计算机存储介质包括但不限于 RAM、ROM、EEPROM、闪存或其它存储器技术、CD-ROM、数字多功能光盘 DVD 或其它光盘存储、盒式磁带、磁带、磁盘存储或其它磁存储设备或者能够用来存储所需要的信息并且能够由计算设备 100 存取的任意其它介质。通信介质典型地具体化为在已调制数据信号中的计算机可读指令、数据结构、程序模块或其它数据，该已调制数据信号例如是载波或其它传输机制，并且包括任意信息传递介质。术语“已调制数据信号”意指这样的信号，该信号的特性中的一个或多个以将信息编码在该信号中的方式被设置或者改变。通过示例而非限制地，通信介质包括诸如有线网络或者直接有线连接的有线介质和诸如声音、RF、红外和其

它无线介质的无线介质。以上的任意组合也将被包括在计算机可读介质的范围内。

[0016] 存储器 112 包括以易失和 / 或非易失形式的计算机存储介质。存储器可以是可移除的、非可移除的或者其组合。示例性的硬件设备包括固态存储器、硬盘驱动器、光盘驱动器等。计算设备 100 包括一个或多个处理器，从诸如存储器 112 或 I/O 组件 120 的多种实体中读取数据。(多个)呈现组件 116 将数据指示呈现给用户或其它设备。示例性呈现组件包括显示设备、扬声器、打印组件、振动组件等。

[0017] I/O 端口 118 允许计算设备 100 逻辑地耦合至包括 I/O 组件 120 的其它设备，I/O 组件 120 中的一些可以是内建的。图示的组件包括麦克风、操纵杆、游戏手柄、碟形卫星天线、扫描仪、打印机、无线设备等。

[0018] 现在参考图 2，提供一种框图以图示其中可以使用本发明实施方式的示例性系统 200。应当理解此处描述的这些和其它设置只是作为例子而提出。在所示出的那些之外或代替示出的那些，能够使用其它设置和元素(例如机器、接口、功能、次序和功能组合等)，并且一些元素能够被一起省略。进一步地，此处描述的许多元素是功能实体，其能够实现为分离的或者分布式组件或者与其它组件相结合，并在任何合适的组合和位置中。此处描述的由一个或多个实体执行的各种功能可以由硬件、固件和 / 或软件完成。例如，多种功能可以通过处理器执行存储在存储器中的指令而完成。

[0019] 除了未示出的其它组件，系统 200 包括用户设备 202、数据存储器 204、分级服务器 206 和搜索索引 220。在图 2 中示出的组件的每一个可以是任意类型的计算设备，诸如例如参考图 1 描述的计算设备 100。组件可以经由网络 208 彼此通信，网络 208 可以包括但不限于一个或多个局域网(LAN)和 / 或广域网(WAN)。在办公室、企业计算机网络、内联网和互联网中，该联网环境是普通的。应当理解在本发明的范围内，在系统 200 中可以使用任意数量的用户设备、分级服务器、分级生成器、数据存储器和搜索索引。每个可以包括单个设备或者在分布式环境中协作的多个设备。例如，分级服务器 206 可以包括布置在分布式环境中的多个设备，其共同地提供此处描述的分级服务器 206 的功能。另外，未示出的其它组件也可以被包括在系统 200 中，同时在图 2 中示出的组件在一些实施方式中可以被省略。

[0020] 本发明实施方式使用的搜索索引 220 给文件中更高阶的原语 (higher order primitive) 或“原子”编索引，与简单地给单个术语编索引成对比。如此处使用，“原子”可以指查询或文件的多种单元。这些单元可以包括例如术语、n-gram(n 元语法)，n 元组，k-near n 元组等。术语向下映射至通过所使用的特定记号化装置(tokenizer)技术定义的单个符号或单词。术语在一个实施方式中是单个字符。在另一个实施方式中，术语是单个单词或一组单词。n-gram 是可以从文件中提取出来的“n”个连续的或者几乎连续的术语的序列。如果它相应于连续术语的游程 (run)，n-gram 被称为“紧的(tight)”，如果它以术语在文件中出现的顺序包含术语，但该术语不是必须连续的，则被称为“松的(loose)”。松 n-gram 典型地被用于表示等同的短语的类，等同的短语其差别在于无意义的词语(例如“if it rains I’ll get wet(如果下雨我将被淋湿)”和“if it rains then I’ll get wet(如果下雨则我将被淋湿)”)。如此处使用的，n 元组是文件中同时出现(独立于顺序或者依赖于顺序)的“n”个术语的集合。进一步地，如此处使用的，k-near n 元组指的是在文件中“k”个术语的窗口之内同时出现的“n”个术语的集合。因此，原子通常被定义为以上所有的概括。实现本发明的实施方式可以使用不同种类的原子，但是如此处使用的，原子通

常描述上述种类中的每一个。

[0021] 用户设备 202 可以是能够接入网络 208 的终端用户拥有和 / 或操作的任意类型的计算设备。例如，用户设备 202 可以是桌面计算机、膝上型电脑、平板电脑、移动设备或具有网络接入的任意其它设备。通常，除了别的之外，终端用户还可以使用用户设备 202 以访问系统维持的电子文件，系统例如是分级服务器 206 或类似的。例如，终端用户可以使用用户设备 202 上的网页浏览器以访问并观看来自分级服务器 206 的电子文件。在其它实施方式中，文件没有被存储在分级服务器 206 上，而是可以被存储在数据存储器 204 中。

[0022] 分级服务器 206 通常负责选择分级特征以用于整个分级过程的初步分级阶段。通常，整个分级过程包括两个或多个分级阶段，诸如初步分级阶段和最终分级阶段。在此处描述的实施方式中，初步分级阶段利用在最终分级阶段中使用的一个或多个分级特征，诸如不具有原子相互依赖性的那些分级特征。而第二分级阶段被命名为“最终分级过程”的“最终分级阶段”，在这个阶段之后还可能有其它分级阶段，这样词语“最终”的使用并不意味着暗示了它是最后的阶段。例如，初步分级阶段可以是第一分级阶段，最终分级阶段可以是第二分级阶段。在特定实施方式中可以使用特殊化的第三分级阶段，这预期也包括在本发明的范围内。

[0023] 如上面简要提及的，当接收到搜索查询时，采用整个分级过程以将匹配文件的数量降低到易处理的规模。在一些实施方式中，搜索引擎可以采用阶段化的过程以选择搜索查询的搜索结果。

[0024] 当接收到搜索查询时，分析该搜索查询以识别原子。然后在整个分级过程的各个阶段期间使用该原子。这些阶段可以被称为 L0 阶段（匹配阶段）以查询搜索索引并识别包含来自搜索查询的该原子或原子中至少一些的匹配文件的初步集合。这个初步的过程可以将候选文件的数量从搜索索引中编了索引的全部文件减少到与来自搜索查询的原子相匹配的那些文件。例如，搜索引擎可以在百万甚至数兆文件中搜索以确定与特定搜索查询最相关的那些。一旦 L0 匹配阶段完成，候选文件的数量被大大地减少。然而，用于定位最相关文件的许多算法成本高并且浪费时间。这样，也可以采用两个或更多阶段（N 阶段），包括初步分级阶段和最终分级阶段。初步分级阶段常常比使用最终分级阶段成本有效地深度分析的能识别更多候选文件。当在整个分级过程中采用 N 个阶段时，每个早期的阶段可以利用在后期阶段中使用的特征的子集，并且还可以使用在后期阶段中未使用的特征。这样，每个早期阶段基本是后期阶段提供的分级的近似，但是较不昂贵并且可能是简化的。

[0025] 初步分级阶段，也被称为 L1 阶段，采用简化的计分功能，用于为上面描述的 L0 匹配阶段中保留的候选文件计算初步得分或分级。这样，初步分级组件 210 负责为 L0 匹配阶段中保留的每个候选文件提供初步分级。与最终分级阶段相比较，初步分级阶段是简化的，因为它只采用最终分级阶段使用的分级特征的子集。例如，在最终分级阶段中使用分级特征中的一个或多个，但很可能不是全部，被初步分级阶段采用。另外，初步分级阶段也可以采用最终分级阶段不采用的特征。在本发明的实施方式中，初步分级阶段使用的分级特征不具有原子相互依赖性，诸如术语紧密和术语共存。例如，在初步分级阶段中使用的分级特征可以包括，仅为了示例的目的，静态特征和动态原子隔离的组件。静态特征通常是那些只针对独立于查询的特征的组件。静态特征的例子包括页面分级、特定网页的垃圾邮件评分等。动态原子隔离的组件是一次只关注与单个原子相关的特征的组件。例子包括例如

BM25f，文件中特定原子的频率，文件中原子的位置（上下文）（例如标题、URL、作者、头部、主体、业务、类、属性）等。

[0026] 一旦候选文件的数量已经被初步分级阶段再次地减少，最终分级阶段，也被称作L2阶段，分级由初步分级阶段提供给它的候选文件。与最终分级阶段共同使用的算法是更昂贵的操作，与初步分级阶段使用的分级特征相比其使用更大数量的分级特征。然而，最终分级算法被应用于数量少得多的候选文件。最终分级算法提供了已分级文件的集合，基于该已分级文件的集合，响应于原始搜索查询提供了搜索结果。

[0027] 回到分级服务器206，分级服务器206包括多种组件，其中每个为计算候选文件的初步分级和通过分级和削减只选择与搜索查询相关的那些文件传送给最终分级阶段的过程提供了功能性。这些组件包括初步分级组件210、最终分级组件212、特征选择组件214、算法生成组件216和数据提取组件218。在图2中未图示的可以被用于为文件提供初步分级并将保留的文件数量削减到可管理的规模的组件也预想被包括在本发明的范围内。进一步地，并不是与分级服务器206相关示出的所有组件都被使用，或者在一些实施方式中，可以与其它组件结合。

[0028] 如上面简要描述的，初步分级组件210负责分级一个候选文件集合并且因此减少将被传送至最终分级阶段的候选文件的数量，该最终分级阶段利用最终分级功能212以分级该较小的候选文件集合。例如，在L0匹配阶段搜索到上亿或者甚至上兆的文件。在初步分级阶段之后，相关文件的数量可以被削减到数千文件，并且在最终分级阶段之后被进一步削减至几十个文件。然后这些文件可以在搜索结果页面上被呈现给用户。在一些实施方式中，简化的计分功能可以作为将最终被用来分级文件的最终分级算法的近似而起作用。然而，该简化的计分功能提供了比最终分级算法较不昂贵的操作，允许更大数量的候选文件被快速地处理。基于该初步得分，候选文件被削减。例如，只有具有最高初步得分的前N个文件可以被保留。

[0029] 为了计算文件的分级，初步分级组件210利用初步分级特征，其中一些也在最终分级阶段中被使用。除了别的之外，初步计分在存储在搜索索引中为文件/原子对预先计算的得分上操作。如所提及的，静态特征和诸如动态原子隔离组件之类隔离的特征可以被初步分级组件210使用。在实施方式中，初步分级组件210通过数据提取组件218访问搜索索引220或诸如在数据存储204中存储的数据之类的其它数据，以提取与初步分级组件210使用的初步特征相关联的数据。在一些实例中，该数据可以以预先计算的分数的形式被存储。例如，特定原子可以具有与其关联的一个或多个预先计算的得分，与对应于特定文件的各种属性相关。例如，在特定文件中，第一原子可能被重复10次，在那个同样的文件中，第二原子可能被重复55次。该第二原子可以具有比第一原子更高的得分，因为在那个文件中它被更多次地找到。或者，系统可以被这样设置，在标题中找到的原子被给出比只在URL中找到的原子更高的分数。多种规则可以被结合到初步计分功能。这样，预先计算的分数被存储在搜索索引中或者其它数据存储器中，在初步计分功能中这个数据能够被提取出来并使用。

[0030] 在实施方式中，虽然第一轮（L1）特征或者初步分级特征是独立于查询的，L1或者初步分级功能可以不是独立于查询的。例如初步分级功能依赖于在特定查询中有多少原子，是否有可替代的解释或拼写，我们对那些因素有多么肯定，查询看起来是来自什么语言

和国度等。所以，初步分级功能非常依赖预先计算的、作为每个原子的概要分级而传递的独立于查询的特征，初步分级功能也可以按照依赖于查询的方式组合它们。

[0031] 在一个实施方式中，对于每个原子 / 文件对只有一个预先计算的分数，这样对于特定的原子预先计算的分数考虑多个特征。例如，预先计算的分数可以考虑文件中特定原子的频率、原子的多个实例彼此之间有多靠近、原子的上下文，诸如它在文件中所处的位置，等等。在可替代的实施方式中，原子具有与特定文件关联的多于一个预先计算的分数。例如，原子可以具有一个考虑文件中原子的频率的预先计算的分数，和对于文件中找到原子的部分具有另一个预先计算的分数。

[0032] 在初步分级功能计算文件的初步分数之前，确定初步分级特征。如在前提及的，初步分级特征可以来自多种源。在一个实例中，最终分级组件 212 使用的分级特征被分析。最终分级组件 212 使用的分级特征通常被分为三个主要的种类。这些种类至少包括静态特征、动态原子隔离的组件和动态原子互相关组件，或者具有原子相互依赖性的那些。在一个实施方式中，特征选择组件 214 执行将特征分为这些种类的功能。特征选择组件 214 可以选择属于静态特征或者动态原子隔离组件的那些特征作为潜在的初步分级特征。这些特征甚至被进一步分析，因为不是所有这些特征都可以被选择在初步分级功能中使用。最后被选择的那些特征可以是易于计算的（例如易于在初步分级功能中使用）、由在初步分级和最终分级之间的逼真度测量而确定有用的、以及在初步分级功能被更改时分级特征执行方面是可适应的、等等。进一步地，相比于其它分级特征，所选择的特征可以被低成本地计算。虽然这些特征可以是易于计算的，当接收到查询时一些特征可能是难以实时计算的，因此可以在初步分级功能中使用，这样它们能够被预先计算并存储在搜索索引中作为预先计算的分数。

[0033] 在一个实施方式中，在初步分级功能中使用的一个或多个初步特征是手动选择的。这样，这个选择过程至少需要一些用户交互。可替代地或者与在前的实施方式相结合地，至少一些初步特征被自动地选择，诸如通过机器学习工具。在一个实施方式中，机器学习工具可以被结合到特征选择组件 214 中。手动选择和机器学习工具可以彼此结合地使用以选择初步特征。或者，特征可以被手动地选择，然后机器学习工具可以确定那些特征在计算文件分级中是否有帮助。这种机器学习环境虑及初步分级功能中每个特征的有用性。如果特定特征被发现不是特别有用，它可以从初步分级功能中被移除。

[0034] 一旦已经由特征选择组件 214 选择了特征，算法生成组件 216 生成为每个文件计算分级的算法。使用所识别的初步分级特征，诸如由于易于计算和有用而从最终分级功能中选择的那些，生成该算法。在一个实施方式中，在最终分级功能中未必使用但是被证明在初步分级功能中有用的特征也被使用。

[0035] 为了确定特征在初步分级功能中是多么有用，可以使用一些类型的逼真度测量。该逼真度测量可以比较与特定文件关联的最终分级和初步分级以确定所述分级有多接近。在实施方式中，第一或初步分级阶段操作作为第二或者最终分级阶段的估计。在最理想的情况中，初步分级将一直匹配最终分级。然而，这通常不是真实情况。能够以许多方式测量逼真度，此处没有描述其全部，但是预期都包括在本发明的范围内。例如，对于一些有效的最前数量（例如，10、100），可以通过以下来定义逼真度：初步分级功能将建议与最终分级功能被用于对初步分级功能分级的所有同样的文件进行分级所找到的元素或文件相同的元

素或文件。这样，可以通过取由最终分级功能分级的最前十个已分级的文件，并确定这些文件中有多少个被分级在由初步分级功能而分级的最前十个已分级的文件中，从而测量逼真度。因此，如果初步分级功能将最终分级功能的最前十个文件中的八个分级在它的最前十个文件中，则逼真度可以被计算为80%(8/10)。由初步分级功能产生的最前十个结果中的八个可以不是以与最终分级功能产生的最前十个结果同样的顺序。在一些实例中，关于逼真度分级而考虑这个，但是在其它实例中不考虑。

[0036] 类似地，逼真度测量可以确定有多少候选文件需要从初步分级计算中被返回，从而足够数量的结果(例如候选文件)被返回作为最终分级计算的结果。可替换地，逼真度测量可以被用来确定从初步分级功能返回的候选文件的数量以保证最终分级功能返回初步分级功能的最前十个结果的全部。再另一种利用逼真度测量的方式是设置阈值，诸如99%。因此，例如，目标可以是99%的时间，最终分级功能返回的最前十个结果位于初步分级功能返回的最前50个文件中。当然，这些数字能够变化并且只是为了阐释的目的而给定。

[0037] 此处描述的实施方式使得最终分级功能能够具有灵活性，不需要全部重建用于初步分级阶段的预先计算的数据。通过测量在已有的初步阶段计算和最终分级功能的新的候选之间的逼真度，可以对于期望的误差范围确定初步分级阶段的新的削减阈值。例如，99%的时间，初步分级阶段在其最前50个文件中得到最终分级阶段的最前十个被分级的文件。可以确定新的分级特征以便增加最终分级功能的精确度。在新的最终分级功能和初步分级功能的结果之间的任何不一致可以被很好地调整，不需要重新计算预先计算的分数。只要初步分级阶段已经通过旧的标准实现很好的工作，很可能通过新的标准或者新的/更新的最终分级功能就也会完成很好的工作。

[0038] 现在转向图3，示出了用于生成算法的方法300的流程图，该算法用于提供多个文件的初步分级。开始，在步骤310分析最终分级功能。如上面描述的，该最终分级功能被用于计算多个文件的最终分级。在实施方式中，最终分级功能的执行是昂贵的并且因此对于有限数量的候选文件而使用，诸如从初步分级功能返回的那些文件。在一个实施方式中，虽然最终分级功能被称作“最终”，但是在最终分级阶段之后可以采用一个或多个分级阶段。其被命名为“最终”是因为其是此处涉及的最后的阶段。在步骤312，从最终分级功能中识别潜在的初步分级特征。这些识别的特征可以包括独立于查询的静态分级特征和只与单个原子相关的动态原子隔离的组件。没有被识别作为潜在初步分级特征的那些分级特征可以是具有原子相互依赖性(例如术语接近、术语共存等)的动态原子互相关组件的那些分级特征。如所提及的，静态分级特征(例如页面分级、垃圾邮件打分)是独立于查询的分级特征和不依赖查询和甚至可以在接收查询之前被计算的那些特征。动态原子隔离组件只考虑一次与单个原子相关的那些特征(例如频率、上下文)。

[0039] 在步骤314，从步骤312中识别的潜在初步分级特征中选择初步分级特征。初步分级特征被使用在初步分级功能中以计算候选文件的分级。在一个实施方式中，初步分级特征包括在步骤314中识别的分级特征中的一些，但是也包括在最终分级功能中未使用，但是已经被证明在初步分级功能中是有用的和精确的一些特征。进一步地，初步分级特征可以是手动识别的，因此需要用户交互(例如，人机工程)。可替代地或者与上述组合地，可以在机器学习工具的辅助下选择初步分级特征，该机器学习工具估计特定特征的计算容易度、有用性、适应性等，并然后确定该特征是否应当在初步分级功能中被使用。在一个实施

方式中，人工选择和机器学习工具的组合被利用来选择初步分级特征。如上面简要提及的，基于许多因素选择初步分级特征。这些因素可以包括(仅为了示例的目的)在初步分级功能中分级特征使用的容易度、通过在文件的初步分级和最终分级之间的逼真度测量而确定的分级特征的有用性，当初步分级功能更改时分级特征的适应性，计算特征的成本等等。在一些实例中，虽然初步分级特征容易被计算，但是它们也可能是在接收到查询时难以实时计算的，并因此可以用于初步分级，因为它们能够被预先计算，消除对它们实时计算的需要。这些因素的组合可以被考虑。在步骤 316，从初步分级特征生成算法以计算文件的初步分级。一旦分级被分配给每个文件，最高分级的文件(例如最前 100 个、1000 个、2000 个)被发送给最终分级功能用于最终分级。来自最终分级功能的最高分级的文件是响应于用户的搜索查询而呈现给用户的那些。

[0040] 在一个实施方式中，从用户接收搜索查询。在步骤 316 生成的用于初步分级功能的算法被用于算法化地识别与搜索查询最相关的文件的子集。这些候选文件被传送(例如通过文件标识)至最终分级阶段，这样最终分级功能能够给候选文件分配最终分级并确定与搜索查询最相关的那些。这些结果被呈现给用户。

[0041] 图 4 是示出了用于为文件计算初步分级的方法 400 的流程图。开始，在步骤 410 识别静态分级特征。静态分级特征是独立于查询的那些，并且在一些实例中，可以是与搜索查询完全不相关的。例如，静态特征可以包括页面分级、垃圾邮件评分、页面语言等。在步骤 412，识别动态原子隔离的组件。动态原子隔离的组件是一次与单个原子相关以及与在特定文件的上下文中原子如何出现相关的分级特征，这样在接收搜索查询之前，预先计算的分数能够被分配给原子 / 文件对，并且能够被存储在例如搜索索引中。在一个实施方式中，静态特征和动态原子隔离的组件至少部分地从最终分级功能中被识别，这样初步分级功能基本是最终分级功能的简化形式。最终分级功能未使用的特征也可以被用于初步分级功能。

[0042] 在步骤 414 选择初步分级特征的集合。这些初步分级特征可以是静态分级特征和 / 或动态原子隔离的组件。对于第一文件，在步骤 416 提取相应于初步分级特征的集合的数据。该数据例如可以从搜索索引中被提取出。进一步地，提取出的数据可以包括与多个文件关联的对于该初步分级特征的集合的预先计算的分数。预先计算的分数可以是针对特定原子 / 文件对，这样预先计算的分数考虑多种因素，或者预先计算的分数可以是仅仅对于特定特征针对一种原子 / 文件对，该特定特征例如是原子在特定文件中出现了多少次。预先计算的分数可以被存储在搜索索引中。基于搜索查询，在步骤 418 利用提取出的数据以计算第一文件的初步分级。例如，如前面描述的，初步分级功能可以利用用于计算文件的初步分级的算法。

[0043] 在一个实施方式中，一旦已经为候选文件计算了初步分级，最前 N 个最高分级的文件能够被识别，并且被发送至最终分级阶段，其中 N 能够是任何数量并可以改变。例如，可以确定由初步分级确定的第一文件的相关性是否超过阈值。基于第一文件超过阈值的相关性，第一文件的文件标识可以被发送至最终分级阶段，其为第一文件分配最终分级。如所提及的，除了静态特征和动态原子隔离的组件，最终分级阶段利用依赖查询的动态原子互相关组件以确定文件的最终分级。动态原子互相关组件可以是文件中特定原子的频率或者文件中特定原子的上下文位置。例如，上下文位置包括文件的标题、作者、头部、主体、业务分类、属性和统一资源定位符(URL)。

[0044] 参考图 5，流程图图示了用于在初步分级阶段利用来自最终分级阶段的分级特征以为文件确定初步分级的方法 500。在步骤 510，分析最终分级功能。识别分级特征的第一子集，包括独立于查询的分级特征和单个原子分级特征。在步骤 512，选择分级特征的第二子集。这些分级特征在最终分级功能中不使用。在步骤 514，从分级特征的第一和第二子集中选择初步分级特征。这些被选择的初步分级特征被用于使用初步分级功能计算文件的初步分级，初步分级功能限制将使用最终分级功能最终分级的文件数量。基于初步分级功能，在步骤 516 算法化地识别文件的子集。初步分级功能利用与分级特征的第一和第二子集关联的数据，诸如预先计算的原子 / 文件对的分数，包括与特定文件关联的关于独立于查询的分级特征（例如静态特征）的分数。数据可以从搜索索引中被提取，诸如正向索引（forward index）（例如根据文件标识编索引）或者反向索引（例如根据原子编索引）。

[0045] 在步骤 518，相应于初步分级阶段产生的文件子集的文件标识被传送至最终分级阶段，其计算文件子集的最终分级，这样基于用户的搜索查询，来自最终分级阶段的最高分级的文件被呈现给用户。在实施方式中，对于一组文件，在初步分级和最终分级之间的逼真度度量被计算以确定初步分级阶段的精度，该初步分级阶段通常是最终分级阶段的简化版本。在上面更详细描述了逼真度测量。

[0046] 已经关于特定实施方式描述了本发明，特定实施方式在所有方面都意图作为说明性而非限制性的。对于本发明所属领域的技术人员，可替代的实施方式是明显的，而不脱离本发明的范围。

[0047] 根据前述内容，将看出本发明很好地适用于获得上面提出的所有目的和目标，以及其它优势，其对于该系统和方法是明显并固有的。可以理解特定特征和子组合是有用的并且可以被采用而不涉及其它特征和子组合。这被权利要求所预期并且包括在权利要求的范围内。

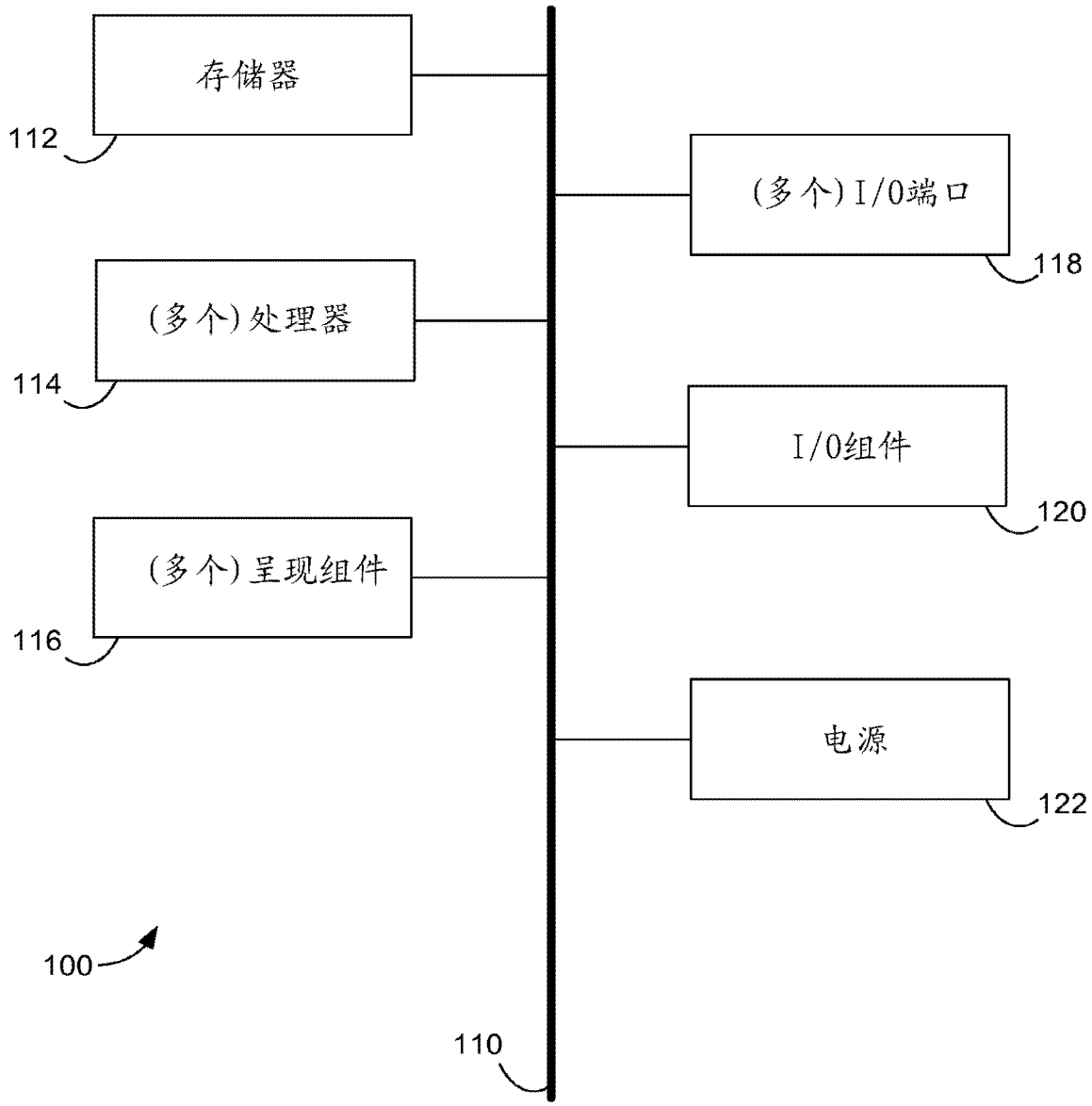

存储器
112
(多个)处理器
114
(多个)呈现组件
116
(多个)I/O端口
118
I/O组件
120
电源
122
100
110

图 1

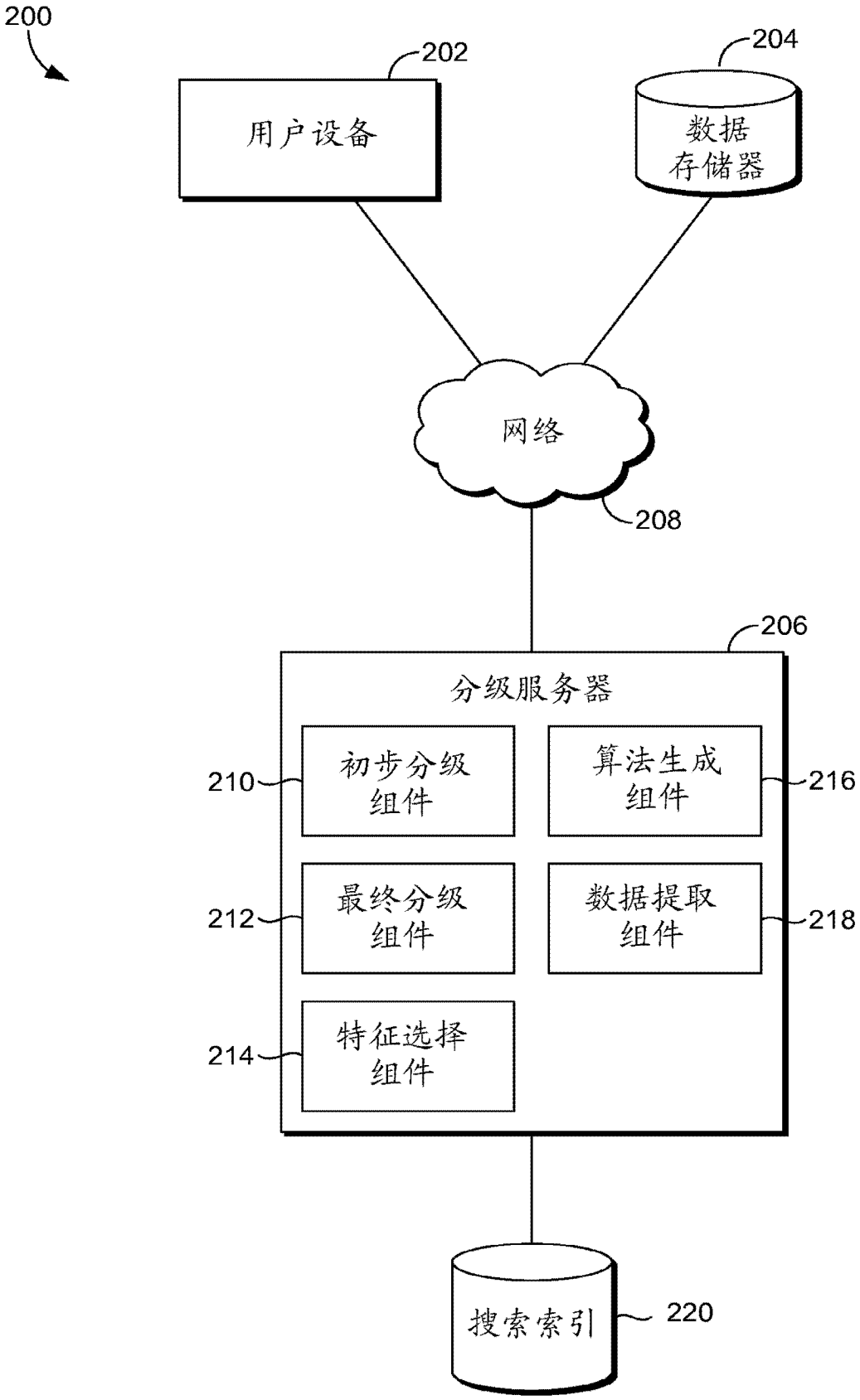

200
202
用户设备
204
数据
存储器
网络
208
206
分级服务器
210
初步分级
组件
算法生成
组件
216
212
最终分级
组件
数据提取
组件
218
214
特征选择
组件
搜索索引
220

图 2

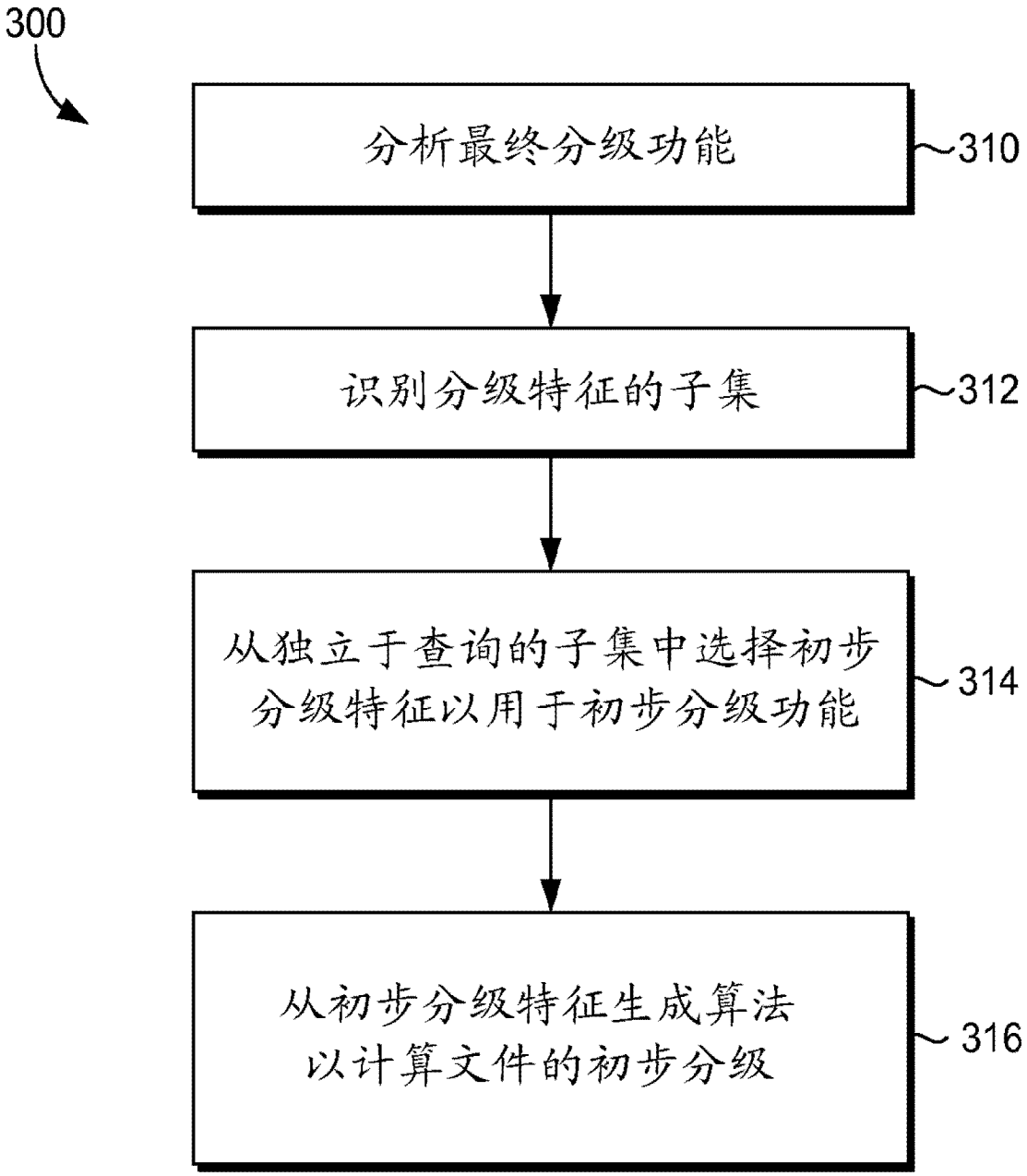

300
分析最终分级功能
310
识别分级特征的子集
312
从独立于查询的子集中选择初步分级特征以用于初步分级功能
314
从初步分级特征生成算法以计算文件的初步分级
316

图 3

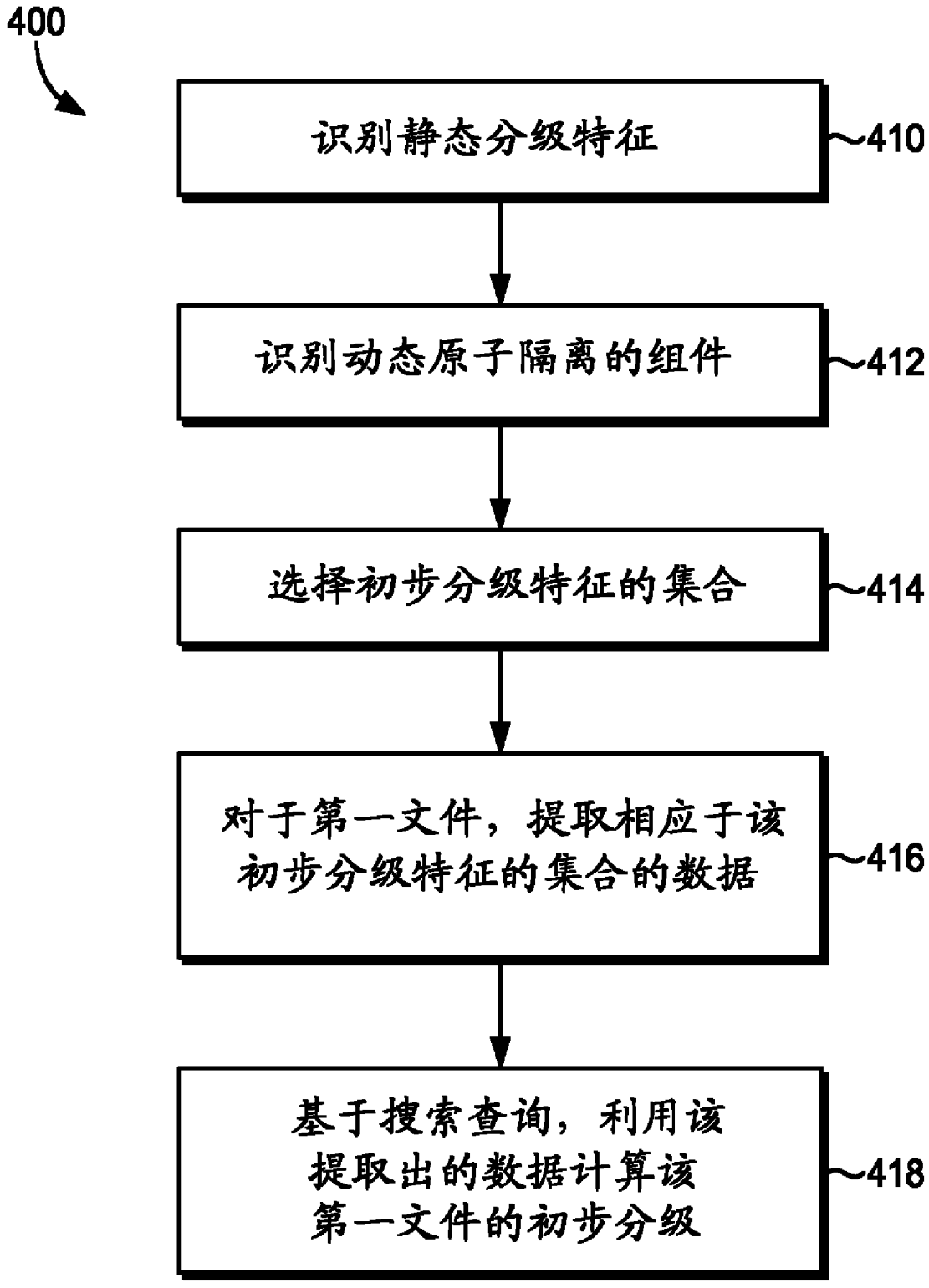

400
识别静态分级特征
410
识别动态原子隔离的组件
412
选择初步分级特征的集合
414
对于第一文件，提取相应于该初步分级特征的集合的数据
416
基于搜索查询，利用该提取出的数据计算该第一文件的初步分级
418

图 4

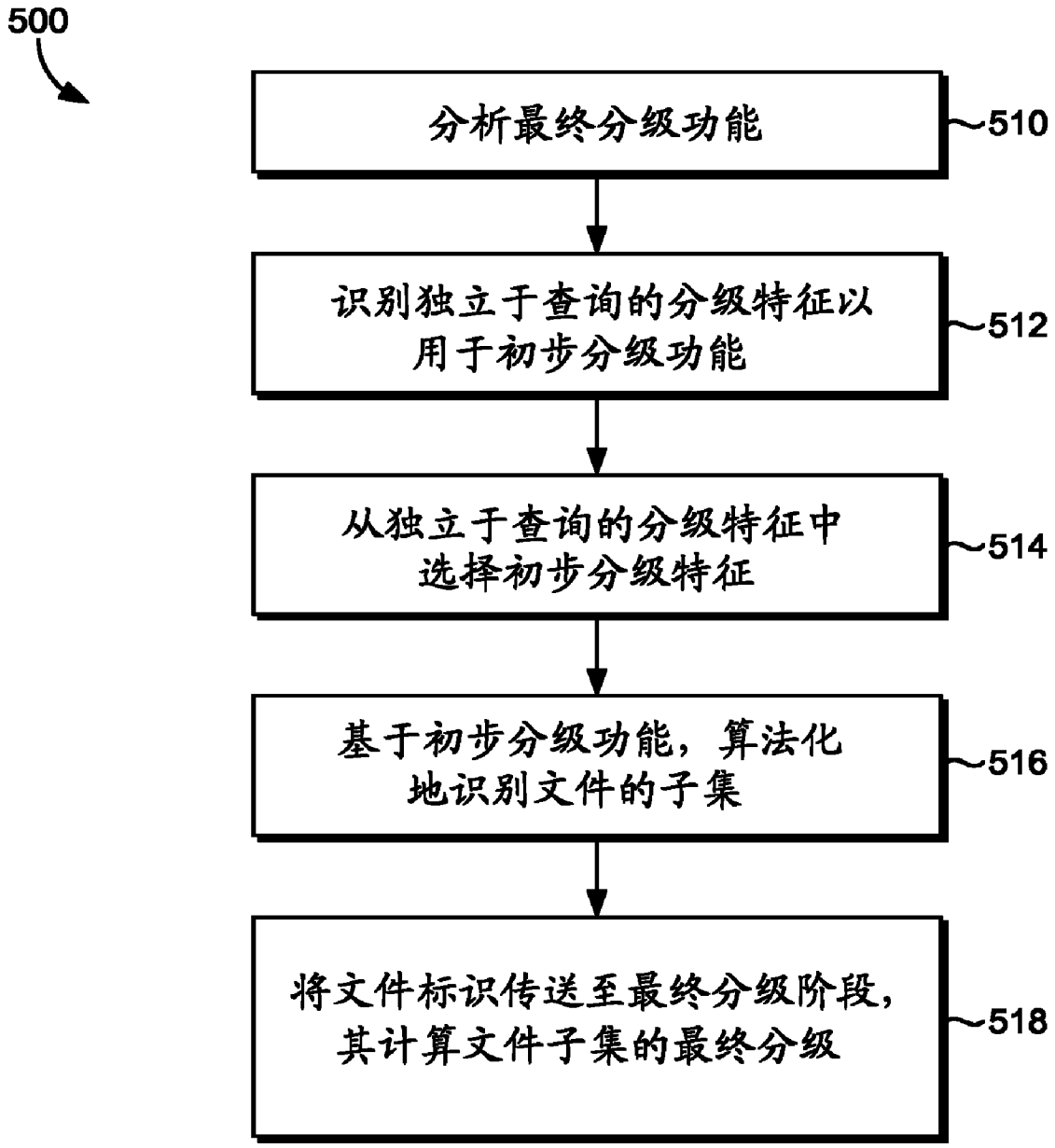

500
分析最终分级功能
510
识别独立于查询的分级特征以用于初步分级功能
512
从独立于查询的分级特征中选择初步分级特征
514
基于初步分级功能，算法化地识别文件的子集
516
将文件标识传送至最终分级阶段，其计算文件子集的最终分级
518

图 5