



(12)发明专利

(10)授权公告号 CN 106250552 B

(45)授权公告日 2020.02.07

(21)申请号 201610668619.9

(22)申请日 2012.07.31

(65)同一申请的已公布的文献号

申请公布号 CN 106250552 A

(43)申请公布日 2016.12.21

(30)优先权数据

13/205809 2011.08.09 US

(62)分案原申请数据

201280038545.0 2012.07.31

(73)专利权人 微软技术许可有限责任公司

地址 美国华盛顿州

(72)发明人 S·K·帕塔萨拉蒂 J·艾哈迈德

Y·萨拉夫 W·孙

(74)专利代理机构 永新专利商标代理有限公司
72002

代理人 张扬 王英

(51)Int.Cl.

G06F 16/953(2019.01)

G06F 16/35(2019.01)

(56)对比文件

US 2009070346 A1, 2009.03.12, 说明书第[0038]段, 第[0042]段, 第[0044]段, 第[0048]段, 第[0090]-[0098]段, 第[0101]-[0105]段, 第[0115]段, 附图1-2, 6-7, 8A, 9.

CN 101819573 A, 2010.09.01, 全文.

CN 102214241 A, 2011.10.12, 全文.

CN 101853261 A, 2010.10.06, 全文.

CN 101727500 A, 2010.06.09, 全文.

审查员 谢晶

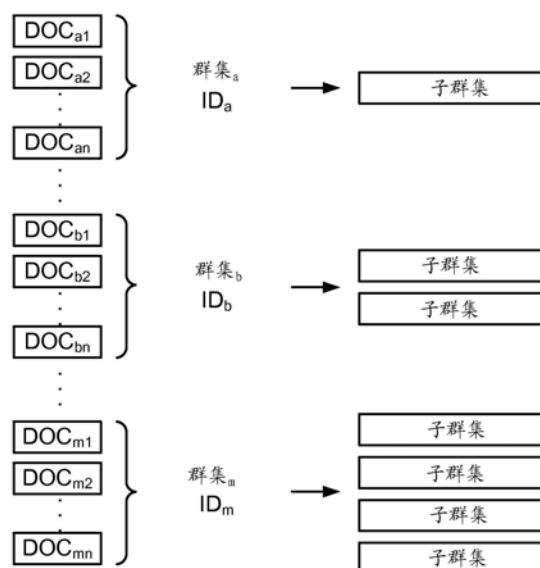
权利要求书2页 说明书7页 附图8页

(54)发明名称

在搜索引擎结果页面上聚集WEB页面

(57)摘要

提供了用于通过在经聚集的文档的“新鲜的”寿命外维持相应聚集的文档的标识(ID)号来递送针对新近的和非新近的事件的经聚集的搜索结果的方法、系统以及媒体。当群集根据相似内容被形成时, ID号和相关联的属性被分配给所述群集中的每一个。这提供了用来跟踪并且检索相应群集以用于搜索结果的后递送的机制。即使在所述文档不再被认为是“新鲜的”之后, 所述群集的相应ID号也被维持。这些相似内容的群集根据发表日期被进一步细分。这为以不同的时间跨度发生的相似内容的事件提供了单独的细分的群集, 其在SERP中连同单独的非聚集的搜索结果一起被递送。



1. 一种聚集搜索结果的文档的方法,所述方法包括:

访问与搜索引擎相关联的数据库,其中,所述数据库包括由所述搜索引擎可检索的多个存储的文档;

基于内容相似性将所述存储的文档中的一些文档聚集成一个或多个群集;

根据发表日期将所述一个或多个群集中的每一个群集细分为一个或多个细分的群集;

将标识符分配给所述存储的文档的群集中的每一个群集,其中,所述标识符在经聚集的存储的文档中的每一个文档的寿命期间被分配,并且其中,所述群集中的每一个群集的标识符贯穿经聚集的存储的文档中的每一个文档的寿命保持持久性;

响应于搜索查询,生成要在搜索结果页面上呈现的搜索结果,所述搜索结果包括所述细分的群集中的一个或多个;以及

在所述搜索结果页面上呈现所述搜索结果,其中,在所述搜索结果页面上呈现的每一个细分的群集包括所述细分的群集的摘要和至包含在所述细分的群集内的文档的链接。

2. 根据权利要求1所述的方法,其中,所述一个或多个群集中的一些包括检索到的为新鲜文档的文档。

3. 根据权利要求1所述的方法,其中,所述一个或多个群集中的一些包括检索到的不为新鲜文档的文档。

4. 根据权利要求1所述的方法,其中,所述一个或多个群集包括一个或多个被分组的统一资源定位符(URL)。

5. 根据权利要求1所述的方法,进一步包括:为所述一个或多个群集中的每一个提供简略摘要。

6. 根据权利要求5所述的方法,其中,所述简略摘要包括以下项中的一个或多个:文档的数目、主机域或针对所述一个或多个群集中的每一个群集的一个或多个显著特征。

7. 一种用于生成包括经聚集的文档的搜索结果的系统,包括:

一个或多个存储器存储设备,被配置为存储包括多个存储的文档的数据库;

一个或多个计算设备,被配置为:

访问包括所述多个存储的文档的所述数据库,

基于内容相似性将所述存储的文档中的一些文档聚集成一个或多个群集,

将标识符分配给所述存储的文档的群集中的每一个群集,其中,所述群集中的每一个群集的标识符是在经聚集的所述存储的文档中的每一个文档的寿命期间被分配的,并且贯穿经聚集的所述存储的文档中的每一个文档的寿命保持持久性,

根据发表日期将所述一个或多个群集中的每一个群集细分为一个或多个细分的群集,以及

响应于搜索查询,生成要在搜索结果页面上呈现的搜索结果,其中,所述搜索结果被组织为一个或多个细分的群集。

8. 根据权利要求7所述的系统,其中,所述一个或多个群集中的一些包括检索到的为新鲜文档的文档。

9. 根据权利要求7所述的系统,其中,所述一个或多个群集中的一些包括检索到的不为新鲜文档的文档。

10. 根据权利要求7所述的系统,其中,所述一个或多个群集包括一个或多个被分组的

统一资源定位符(URL)。

11. 根据权利要求7所述的系统,进一步包括:所述一个或多个计算设备被配置为:为所述一个或多个群集中的每一个提供简略提要。

12. 根据权利要求11所述的系统,其中,所述简略提要包括以下项中的一个或多个:文档的数目、主机域或针对所述一个或多个群集中的每一个群集的一个或多个显著特征。

13. 一种使用计算设备来生成包括经聚集的文档的搜索结果的计算机实施的方法,所述计算设备具有处理器、存储器和数据存储器,所述计算机实施的方法包括:

基于页面内容的相似性将数据库中存储的多个文档分组以形成一个或多个群集;

向所述一个或多个群集中的每一个群集分配标识符和一个或多个相应的相关属性;

维持所述一个或多个群集中的每一个群集的所分配的标识符和所述相应的相关属性,其中,所述群集中的每一个群集的标识符贯穿经聚集的存储的文档中的每一个文档的整个寿命保持持久性;

根据发表日期将所述一个或多个群集中的每一个群集细分成一个或多个细分的群集;以及

响应于搜索查询,生成要在搜索结果页面上呈现的搜索结果,所述搜索结果包括所述细分的群集中的一个或多个。

14. 根据权利要求13所述的计算机实施的方法,其中,将多个文档分组包括将多个新鲜文档分组。

15. 根据权利要求13所述的计算机实施的方法,其中,将多个文档分组包括将多个非新近的事件文档分组。

16. 根据权利要求13所述的计算机实施的方法,其中,所分配的标识符贯穿每个相应文档的生命期保持持久性。

17. 根据权利要求13所述的计算机实施的方法,其中,所述多个文档中的每个文档被认为是一个月寿命的新鲜文档。

18. 根据权利要求13所述的计算机实施的方法,进一步包括:响应于用户搜索查询,向所述计算设备的用户接口显示所述一个或多个群集中的一个群集的、按发表日期的所述一个或多个细分的群集。

19. 根据权利要求18所述的计算机实施的方法,其中,显示所述一个或多个细分的群集中的每个细分的群集包括显示以下项中的相应一个或多个:显著标题、显著图像、或显著新闻概要。

20. 根据权利要求13所述的计算机实施的方法,其中,所述一个或多个细分的群集包括根据所述一个或多个细分的群集的相应ID号被分组的统一资源定位符(URL)。

在搜索引擎结果页面上聚集WEB页面

[0001] 本申请是申请日为2012年7月31日、申请号为201280038545.0、发明名称为“在搜索引擎结果页面上聚集WEB页面”的中国专利申请的分案申请。

背景技术

[0002] 数据可以被以电子形式存储以用于以计算机化的技术使用。结合各种不同的应用使用的大量计算机化的数据对于如何定位并且组织相关信息提出了挑战。聚集指的是将诸如在计算机化的数据中包括的文档之类的一组数据对象分类成组以便每个组包括相似对象并且属于其它组的对象不相似的过程。

[0003] 聚集提供了当存在关于特定话题的大量的信息或新闻时将新鲜文档分组在一起的手段。群集的概要或抽象连同到群集内的文档的链接和其它相干信息一起被显示在搜索引擎结果页面(SERP)中。新鲜文档是关于新近话题或感兴趣主题的文档。在短时间段之后,文档不再被认为是新鲜文档。文档在它们是新鲜的时被聚集,并且它们被提供有标识号。这个信息与文档一起保持以便帮助辨别这个群集和关于相似话题的新的群集。

[0004] 用于聚集对象的技术包括但不限于层次聚集方法或分割方法。层次算法通过要么将较小群集合并成较大群集要么通过将较大群集分成较小群集来接连地进行。相比之下,分割算法通过将数据集分解成一组不相交的群集立刻确定所有群集。层次聚集算法能够被进一步描述为分裂方法(即,自顶向下)或凝聚方法(即,自底向上)。分裂算法始于整个集合并且递归地将该数据集分割成两个(或更多个)片,从而形成树。凝聚算法从它自己的群集中的每个对象开始并且迭代地合并群集。

发明内容

[0005] 本发明的实施例被以下权利要求限定。各种实施例的高级综述被提供来介绍下面在具体实施方式部分中被进一步描述的系统、方法以及媒体的概述。本概述既不旨在标识所要求保护的主题的关键特征或必要特征,也不旨在被用来孤立地帮助确定所要求保护的主题的范围。

[0006] 用于递送针对新近的和非新近的事件的聚集的搜索结果的系统、方法以及计算机可读存储媒体被描述。相应群集的标识号贯穿相应文档的生命或者持续达任何其它指定的时间周期在经聚集的文档的“新鲜的”寿命外被维持。这些长期群集根据发表日期被进一步划分成一个或多个群集。结果,非新鲜文档继续被聚集,但同样被根据不同的时间线按发表日期划分成单独的相似事件。

附图说明

[0007] 参考附图,在下面详细地描述本发明的说明性实施例,附图通过引用被并入这里,并且其中:

[0008] 图1是依据本发明的实施例使用的示范性计算机操作系统的示意表示;

[0009] 图2是依据本发明的实施例使用的查询聚集系统的示意表示;

- [0010] 图3是依据本发明的实施例使用的代表性查询聚集过程的流程图；
- [0011] 图4是依据本发明的实施例使用的群集和子群集信息的说明；
- [0012] 图5是依据本发明的实施例使用的递送搜索结果的计算机实施的方法的流程图；
- [0013] 图6是依据本发明的实施例使用的搜索引擎结果页面的说明；
- [0014] 图7是递送持久性群集的方法的流程图，指令被存储在依据本发明的实施例使用的一个或多个计算机可读存储媒体上；以及
- [0015] 图8是在SERP中提供经聚集的非唯一结果的方法的流程图，指令被存储在依据本发明的实施例使用的一个或多个计算机可读存储媒体上。

具体实施方式

[0016] 许多新闻故事在初始发表之后不久(诸如一个或两个星期以后)对于一般公众而言便不是有趣的。此外，如果新闻故事仍然是有趣的，则它可能已经某种程度上从初始新闻推力(thrust)改变了其焦点。然而，存在其中“旧”新闻故事诸如鉴于相似的新近的新闻故事收回兴趣的数个实例。例如，新近的大地震或海啸被频繁地与在新近事件之前数个星期、月、年发生的在先的地震或海啸相比较。因此，维持以前事件的群集达较长的时间段(诸如贯穿组成群集的单独的文档的生命)将是有利的。

[0017] 当群集根据相似内容被形成时，标识(ID)号和相关联的属性被分配给群集中的每一个。这提供了一种机制来跟踪并且检索相应群集以用于搜索结果的后续递送。即使在文档不再被认为是“新鲜的”之后，群集的相应ID号也被维持。这些相似内容的群集根据发表日期被进一步细分。这为以不同的时间跨度发生的相似内容事件提供了单独的细分的群集。作为例子，针对在中国在2002、2006以及2010年发生的三次不同的飞机事故的搜索结果将在SERP中连同单独的相关搜索结果一起作为三个不同的群集被递送。

[0018] 本发明的实施例提供了用于递送针对新近的和非新近的事件的聚集的搜索结果的系统、方法以及计算机可读存储媒体。这个具体描述以下权利要求满足可适用的法定要求。

[0019] 术语“步骤”、“框”等在本文中可能被用来暗示被采用方法的不同动作，但术语不应该被解释为隐含任何特定顺序，除非单独的步骤、框等的顺序被明确地描述了。同样地，术语“模块”等在本文中可能被用来暗示被采用系统的不同构件，但术语不应该被解释为暗含任何特定顺序，除非单独的模块等的顺序被明确地描述了。

[0020] 本发明的实施例包括而不限于方法、系统以及在一个或多个计算机可读媒体上体现的计算机可执行指令集。计算机可读媒体包括易失性和非易失性媒体、可拆卸和非可拆卸媒体、以及可由数据库和各种其它网络设备读取的媒体。通过例子和非限制的方式，计算机可读存储媒体包括用任何方法或技术实施以用于存储信息的媒体。存储的信息的例子包括计算机可用指令、数据结构、程序模块以及其它数据表示。媒体例子包括但不限于信息递送媒体、随机存取存储器(RAM)、只读存储器(ROM)、电可擦可编程只读存储器(EEPROM)、闪存存储器或其它存储器技术、光盘只读存储器(CD-ROM)、数字多功能盘(DVD)、蓝光盘、全息媒体或其它光盘存储器、磁盒、磁带、磁盘存储器和其它磁存储设备。媒体的这些例子能够被配置成暂时地、临时地或者永久地存储数据。计算机可读媒体包括协作或互连的计算机可读媒体，其排他地存在于处理系统上或者分布在对于该处理系统可以是本地的或者远离

该处理系统的多个互连的处理系统之中。

[0021] 本发明的实施例可以在计算机代码或机器可用指令的一般上下文中被描述,所述计算机代码或机器可用指令包括被计算系统或其它机器执行的诸如程序模块之类的计算机可执行指令。一般地,包括例程序、程序、对象、构件、数据结构等等的程序模块指的是执行特定任务或者实施特定数据类型的代码。本文中所描述的实施例可以使用各种系统配置而被实施,所述各种系统配置包括手持式设备、消费电子装置、通用计算机、更专用计算设备等。本文中所描述的实施例还可以使用远程处理设备而被实施在分布式计算环境中,所述远程设备通过诸如因特网之类的通信网络被链接。

[0022] 在一些实施例中,递送一个或多个事件的搜索结果的计算机实施的方法被描述。基于页面内容相似性的多个文档被分组以便形成一个或多个群集。标识(ID)号和相应的相关属性被分配给一个或多个群集。在所述多个文档不再被认为是新鲜文档之后,群集的所分配的ID号和相应的相关属性被维持。所形成的群集根据发表日期被细分成一个或多个细分的群集。

[0023] 在其它实施例中,包含体现在其上的计算机可读指令的一个或多个计算机可读存储媒体被描述,所述计算机可读指令当被计算设备执行时,执行在搜索引擎结果页面中递送持久性群集的方法。文档根据接收到的搜索查询被从数据库中检索。所检索到的文档中的一些基于内容相似性和发表日期被聚集成一个或多个群集。标识(ID)号被分配给检索到的文档的群集中的每一个,其中群集中的每一个的ID号贯穿经聚集的检索到的文档中的每一个的寿命保持持久性。响应于所接收到的搜索查询,群集在搜索引擎结果页面中连同其它单独的结果一起被递送到用户接口。

[0024] 在又一些实施例中,包含体现在其上的计算机可读指令的一个或多个计算机可读存储媒体被描述,所述计算机可读指令当被计算设备执行时,执行在搜索引擎结果页面中提供聚集的非唯一结果的方法。响应于接收到的用户搜索查询,多个文档被检索。一定数目的顶部结果从所检索到的文档被选择。顶部结果根据发表日期或内容相似性使用相应的一个或多个检索到的群集的一个或多个标识(ID)号被分组。响应于所接收到的用户搜索查询,搜索结果被递送到用户接口,其中搜索引擎结果页面包括经分组的顶部结果。

[0025] 已经在本文中简要地描述了实施例的一般综述,示范性计算系统在下面被描述。参考图1,用于实施本发明的实施例的示范性操作环境被示出并且被一般地指定为计算设备100。计算设备100只是适合的计算系统的一个例子,并且不旨在关于本发明的实施例的用途或功能性的范围建议任何限制。计算设备100也不应该被解释为具有与被说明构件中的任何一个或组合相关的任何依赖或要求。在一个实施例中,计算设备100是具有处理器、存储器以及数据存储器子系统的常规计算机(例如,个人计算机或膝上型电脑)。本发明的实施例同样适用于多个互连的计算设备,诸如计算设备100(例如,无线电话、个人数字助理或其它手持式设备)。

[0026] 计算设备110包括总线110,其直接地或者间接地耦合以下设备:存储器112、一个或多个处理器114、一个或多个呈现构件116、输入/输出(I/O)端口118、输入/输出构件120以及说明性电源122。总线110表示可以为一条或多条总线的东西(诸如地址总线、数据总线或其组合)。尽管为了清楚起见,图1的各种框被用线示出,但是刻划各种构件实际上不是如此清楚的,并且比喻地,线更准确地说将是灰色的且模糊的。例如,一个人可以将诸如显示

设备之类的呈现构件116认为是I/O构件120。同样,处理器114具有存储器112。本领域的技术人员应理解,这是本领域的本性,并且如先前提到的那样,图1的图仅仅说明了能够结合本发明的一个或多个实施例被使用的示范性计算设备。在如“工作站”、“服务器”、“膝上型电脑”、“手持式设备”等这样的类之间未做区分,因为全部都被设想在图1的范围内,并且作为“计算设备”或“计算系统”被引用。

[0027] 上面关于计算设备100所描述的构件还可以被包括在无线设备中。如本文中所描述的无线设备指的是无线地通信的无线电话、手持式设备、个人数字助理(PDA)、黑莓(BlackBerry)®、智能电话、数字相机或的其它移动设备(除了膝上型电脑以外)中的任何类型。本领域的技术人员将领会,无线设备还将包括执行各种功能的处理器和计算机存储媒体。本文中所描述的实施例适用于计算设备和无线设备两者。在实施例中,计算设备还能够指的是运行这样的应用的设备,所述应用的图像被无线设备中的相机捕获。

[0028] 上面所描述的计算系统被配置成与上面一般地描述并且在下文中更详细地描述的用于在搜索引擎结果页面中递送持久性群集的数个计算机实施的方法、系统以及媒体一起被使用。

[0029] 图2是依据本发明的实施例的查询聚集系统200的说明。系统200包括数据储存器210,其存储要在用户接口(未示出)处被显示的多个数据项220。这样的数据项220能够包括能够在用户接口处在各种可查看状态下出现的文档、文件、文件夹、图像、音频文件、源代码等等。数据项220同样与各种属性相关联,其诸如描述如项的类型(例如图像、文档、电子表格、二进制等等)、创建的日期、与项相关联的人、位置、类、用户定义的属性等等这样的方面的元数据。聚合器230收集数据项220和相关联的属性并且将项呈现给属性分析器240,所述属性分析器240执行相应项和属性的分析。例如,这样的分析能够包括自动地为各种可能的聚集场景确定分数或者为项确定潜在的分组。

[0030] 基于通过分析器240的分析,群集组织器250将新的群集260的优化分组呈现给用户。群集260的优化分组便于从数据储存器210中查找并且检索期望的信息,所述数据储存器210能够包括本地存储介质、远程存储介质或本地储存器和远程储存器的组合。

[0031] 图3是用于使用诸如上面参考图2所描述的系统200之类的系统来聚集搜索结果的示范性过程的流程图。搜索结果被用文档检索系统来生成。在步骤310中,与其索引关键词中的每一个相关联的每个索引的文档的类别被记录。所述类别可以包括文档在它被搜索和用每个特定索引关键词索引时的所有可能的类别(或最相关的或最常用的类别)。在步骤320中,处理响应于搜索查询通过选择与搜索查询相关的一组文档并且对其进行排名来生成搜索结果。作为例子,步骤320能够通过使用倒排索引被执行。搜索查询通常包含特定数目的关键词,所述关键词使用诸如上面参考图1所描述的计算设备之类的计算设备或计算机网络从搜索器与搜索请求一起被提交。

[0032] 在步骤330中,搜索结果被分组成特定数目的文档群集。作为例子,文档群集能够被按关键词关联聚集类别(KWAC)分组。每个结果文档被放入它的与搜索关键词中的每一个相关联的类别中的每一个,并且所有的结果文档的类别的联合可以被用来构建针对搜索结果的最后的文档群集。群集根据在每个群集中包括的文档的排名和经聚集的文档与对应群集的关联权重被排名,使得具有较高排名的群集和在群集中的每一个中具有较高排名的文档被首先标识。在步骤340中,经聚集的搜索结果被组织以用于显示并且递送给用户。

[0033] 图2中所说明的查询聚集系统和图3中所说明的用于聚集搜索结果的过程仅仅是能够被与本发明的实施例一起使用的一个示范性系统和方法。聚集查询结果的数个其它系统和方法被设想在本发明的范围内。

[0034] 图4说明了其中多个文档被分组成群集的系统 and 过程,其能够使用分别参考图2和3上面所描述的聚集系统和方法而被实施。在图4的顶部,标记为 $a_1, a_2 \dots a_n$ 的多个文档根据内容被分组以便形成群集a。群集a被用标识号 ID_a 标识。群集a中的所述多个文档然后根据发表日期被进一步分组或者划分,以便形成一个或多个细分的群集。

[0035] 以下例子被给出来说明形成群集和细分的群集的用途和优点。假设接收到的用户查询包含单词[royal wedding (皇家婚礼)]。数个结果将被返回,其中大部分结果将与威廉(William)和凯特(Kate)的新近的2011皇家婚礼有关,但还能够存在从三十年前发生的Charles (查尔斯)和Diana (戴安娜)的皇家婚礼返回的结果。例如,威廉和凯特的皇家婚礼将在事件之后不久已经被聚集了,并且传统上,将保持被聚集达2-4个星期的周期或者达近似一个月。新近产生的文档(诸如在最后2-4个星期内或在最后一个月份内产生的文档)被称为新鲜文档。使用传统方法,在文档不再被认为是新鲜文档之后,相关联的群集停止存在,并且包含针对两个皇家婚礼的那些非新鲜文档的结果将作为单独的非聚集的结果被互相交错和呈现。然而,通过使用依据本发明的实施例的持久性ID号,内容相关的群集(例如[royal wedding]搜索结果)能够被维持并且进一步划分成附加的类,诸如发表日期。持久性群集附加于任何新近形成的群集。在刚刚说明的例子中,在[royal wedding]搜索结果中将很可能存在两个细分的群集——一个针对威廉和凯特的皇家婚礼,以及第二个细分的群集针对查尔斯和戴安娜的皇家婚礼。

[0036] 图4进一步说明了接收到的搜索查询能够导致返回说明为 $b_1, b_2 \dots b_n$ 的文档的第二群集,以便形成具有标识号 ID_b 的群集b。群集b根据文档 $b_1, b_2 \dots b_n$ 内的相似内容被形成。在群集b中包含的文档然后根据发表日期被分组或者划分以便形成一个或多个细分的群集。

[0037] 群集m同样根据检索到的文档 $m_1, m_2 \dots m_n$ 的相似内容形成,其具有标识号 ID_m ,所述检索到的文档然后根据发表日期被划分成一个或多个细分的群集。取决于所接收到的搜索查询的内容,任何数目的群集或细分的群集能够被形成。图4同样说明了特定群集可以包含仅一个细分的群集,其中经聚集的文档中的全部都和相同的时间跨度有关,然而其它群集能够被划分成各种多个细分的群集。

[0038] 图5说明了用于递送一个或多个事件的搜索结果的计算机实施的方法的流程图。响应于接收到的搜索查询,多个文档被检索。文档能够包括新鲜文档、诸如非新近的文档之类的非新鲜文档或两者的组合。在步骤510中,包含相似或相关内容的那些检索到的文档在搜索查询的时候被分组成一个或多个群集。替代实施例提供在文档被索引的时间发生的聚集,其独立于搜索查询的时间。任何数目的群集能够被形成,但所检索到的文档中的许多可以保持为未被与其它文档一起聚集的单独的文档。在步骤520中,所形成的群集中的每一个都分配了标识(ID)号。每个ID号可以具有与其相关联的属性,这使能实现对特定群集在它先前已被形成了之后的跟踪。

[0039] 在步骤530中,所分配的针对群集的ID号被维持并且因此,贯穿每个特定文档的生命期保持持久性,或者保持持久性达任何其它指定的时间周期。在聚集搜索结果的传统方

法中,聚集典型地仅针对新鲜文档被执行。因此,在文档不再被认为是在流行的巅峰(即“新鲜的”)之后,群集不再存在。这个时间周期能够变化,但新鲜文档寿命的例子是约2-4个星期或近似一个月。针对特定群集的ID号将典型地在“新鲜的”寿命的结束时期满,并且群集将停止存在。然而,通过依据本发明的实施例来维持ID号及其相关联的属性,群集同样在“新鲜的”寿命外被维持。本发明的实施例维持群集的ID号达其文档的整个正常寿命。作为仅用于说明性目的的例子,考虑在夏威夷群岛(Hawaiian Islands)发生的许多火山爆发(volcanoes)。示范性查询可能包括两个常见关键词或短语“夏威夷群岛”和“火山爆发”。使用本发明的实施例,相关的查询结果将被进一步划分到每个事件在其期间发生的时间周期。附加的实施例将贯穿相应文档的寿命维持这些时间相关的群集。

[0040] 本发明的实施例同样使能实现目前未被聚集的现有“旧”文档的持久性聚集。这可能是未实施先前聚集被或者期满的群集的结果。

[0041] 在步骤540中,来自检索到的文档的任何先前形成的群集以及新近形成的群集根据发表日期被划分成细分的群集。这个细分允许临时相关的事件的在特定群集上的更细分组。聚集算法按ID号并且然后按发表日期对文档进行分组。在步骤550中,经细分的群集然后被递送到用户接口以便将更加一致的体验提供给用户并且以便通过其它单独的非聚集的结果的显示来允许更多多样性。

[0042] 图6说明了在搜索引擎结果页面(SERP)的用户接口上的显示,其包含两个细分的群集。在所显示的例子中,包含关键词China(中国)、plane(飞机)以及crash(事故)的搜索查询被实施。因为所检索到的文档包含涵盖在中国的两次不同飞机事故的多个文档,所以经返回的结果包括针对两次不同飞机事故的两个细分的群集,而不是因为这些是“旧”文档而被组合成仅一个群集或者根本没有群集的相关文档中的全部。每个细分的群集显示给出特定细分的群集的简略提要(thumbnail synopsis),到文档的链接被包含在经细分的群集内。简略提要提供在经细分的群集内的内容的概要。细分的群集提要的本发明的实施例包含一个或多个特征,诸如显著标题、显著图像、显著新闻概要或描绘经细分的群集事件的图像。经细分的群集提要还可以包含在经细分的群集内的文档的数目、主机域或经细分的群集的一个或多个显著特征。

[0043] 图7说明了用于在搜索引擎结果页面(SERP)中递送持久性群集的方法的流程图,针对其的计算机可读指令被体现在一个或多个计算机可读存储媒体上。在步骤710中,在搜索查询的时候,多个文档响应于搜索查询被从数据库中检索。替代实施例提供在文档被索引的时间发生的聚集,其独立于搜索查询的时间。在步骤720中,每个检索到的文档的内容被分析以便确定所检索到的文档中的一些是否应该基于内容相似性或发表日期或两者被聚集。每个群集将包含针对每个检索到的文档的所关联的统一资源定位符(URL)。具有相似内容的文档的数目将变化,其将保证形成群集。文档的门限数目能够被先前建立,所述门限数目将触发群集的形成。在步骤730中,标识(ID)号被分配给每个群集。ID号具有与它相关联的群集属性以用于该群集的后续跟踪和检索。在步骤740中,所形成的群集和单独的非聚集的检索到的文档被在SERP中递送到用户接口。在步骤750中,每个已递送群集的简略提要同样被显示。每个群集的简略提要将给出群集内容的总体概要或视图。

[0044] 图8说明了用于在搜索引擎结果页面(SERP)中提供聚集的非唯一结果的方法的流程图,针对其的计算机可读指令被体现在一个或多个计算机可读存储媒体上。在步骤810

中,在搜索查询的时候,响应于输入的搜索查询,文档被搜索引擎检索到。替代实施例提供在文档被索引的时间发生的聚集,其独立于搜索查询的时间。在步骤820中,一定数目的顶部结果被选择。一定数目可以是可配置的数,诸如十个以上的文档。一定数目能够基于诸如系统资源之类的特定考虑事项而变化。在步骤830中,所选顶部结果通过跟踪先前形成的群集的ID号根据发表日期或内容相似性而被分组。在步骤840中,所选顶部结果的检索到的单独的文档和检索到的群集的结果被递送到用户接口。结果还可以包含新近形成的群集。经聚集的结果能够包含新鲜文档,以及不再被认为是新鲜文档的文档。

[0045] 按发表日期形成细分的群集能够具有在主要群集内形成自然划分的优点。图6说明了这点,其中不同的时间线自然地将[China,plane,crash]搜索查询结果分离成两个不同的飞机事故结果。另一例子能够针对特定的事件链而发生,诸如大地震或飓风。初始群集可能是实际破坏,第二群集可能是搜索和救援任务,第三群集可能是临时性房屋搭建,第四群集可能是残余物清除,以及第五群集可能是重建努力。此外,初始群集可以变得非常大,并且能够被随后细分成较小群集。如果以前的新闻文档贯穿文档的生命保持被聚集而不是保持完整(intact)仅达短的时间段,则新闻研究努力无疑将被加快(expedited)。

[0046] 在不背离本发明的精神和范围的情况下,所描绘的各种构件的许多不同布置以及未示出的实施例是可能的。本发明的实施例已经以说明性而不是约束性的目的进行了描述。

[0047] 应理解,特定特征和子组合是实用的,以及可以在不参考其它特征和子组合的情况下被采用,并且被设想在权利要求的范围内。并非在各种图中列举的所有步骤需要被以所描述的特定顺序执行。

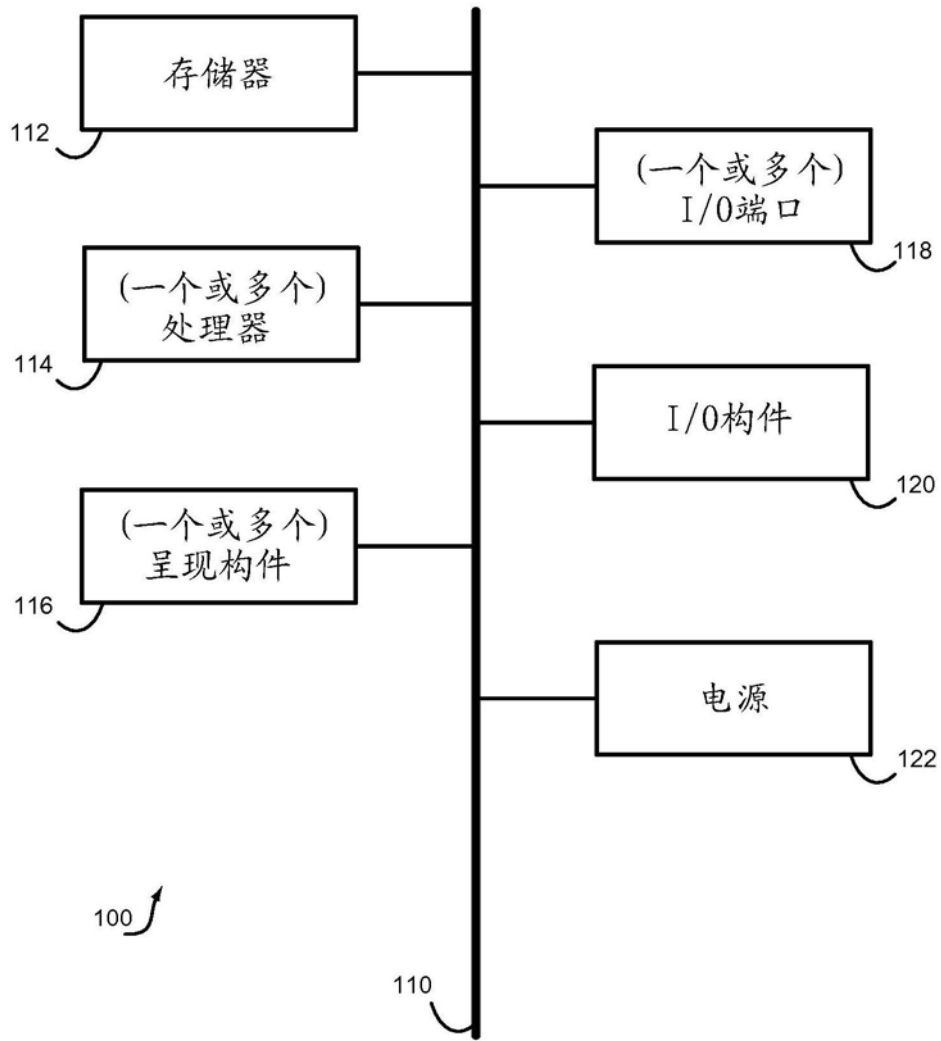


图1

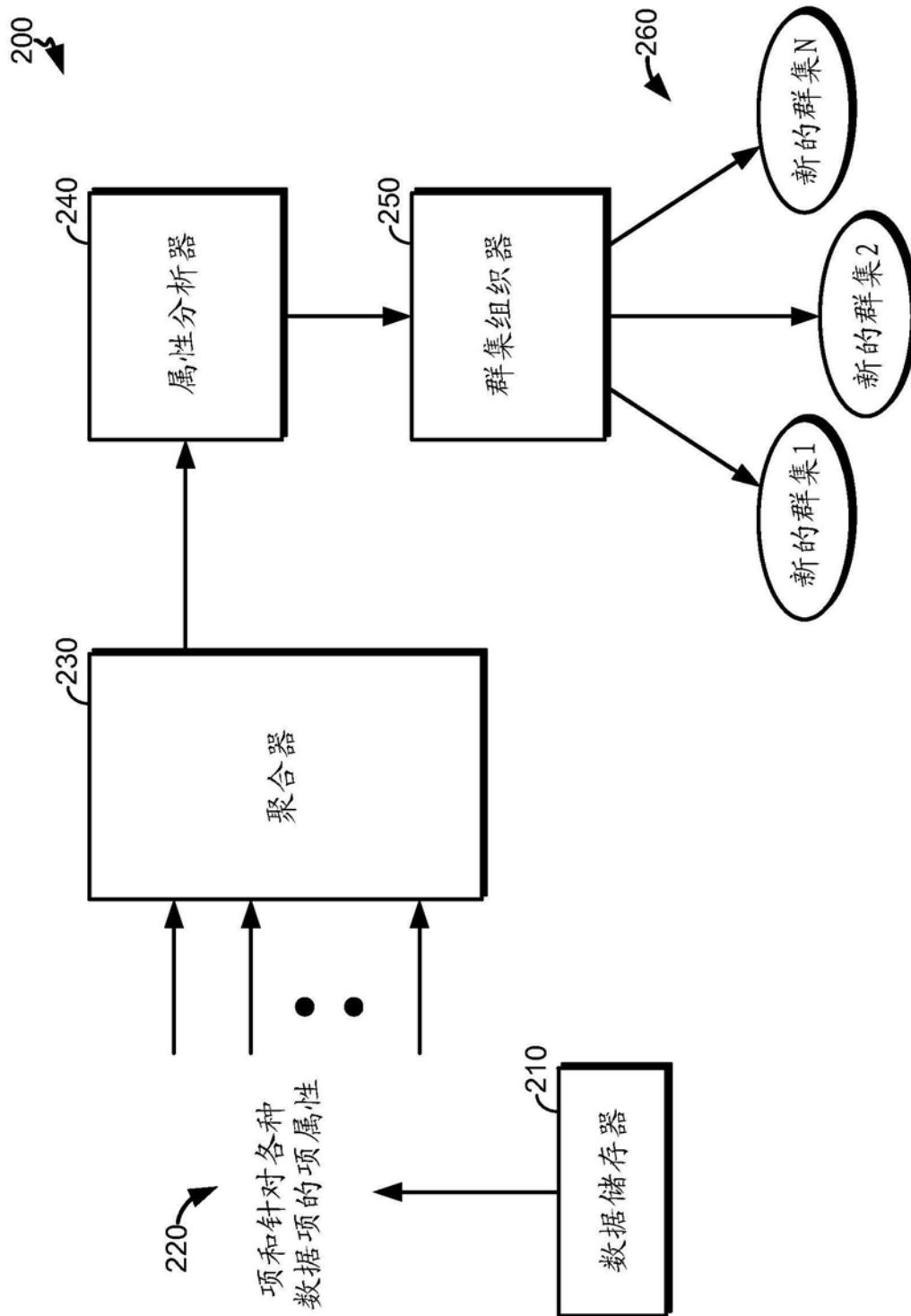


图2

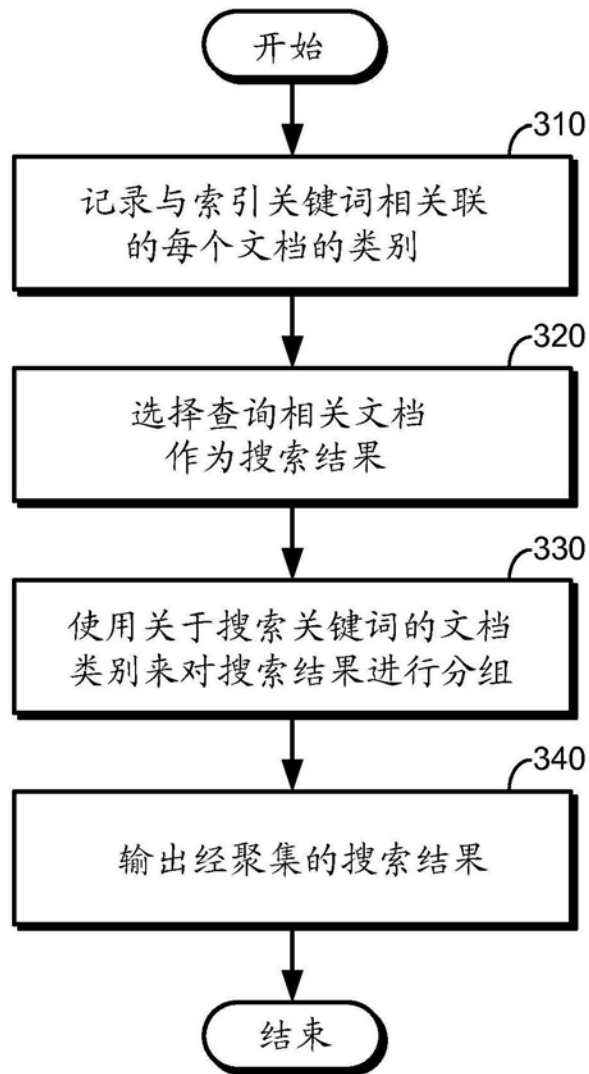


图3

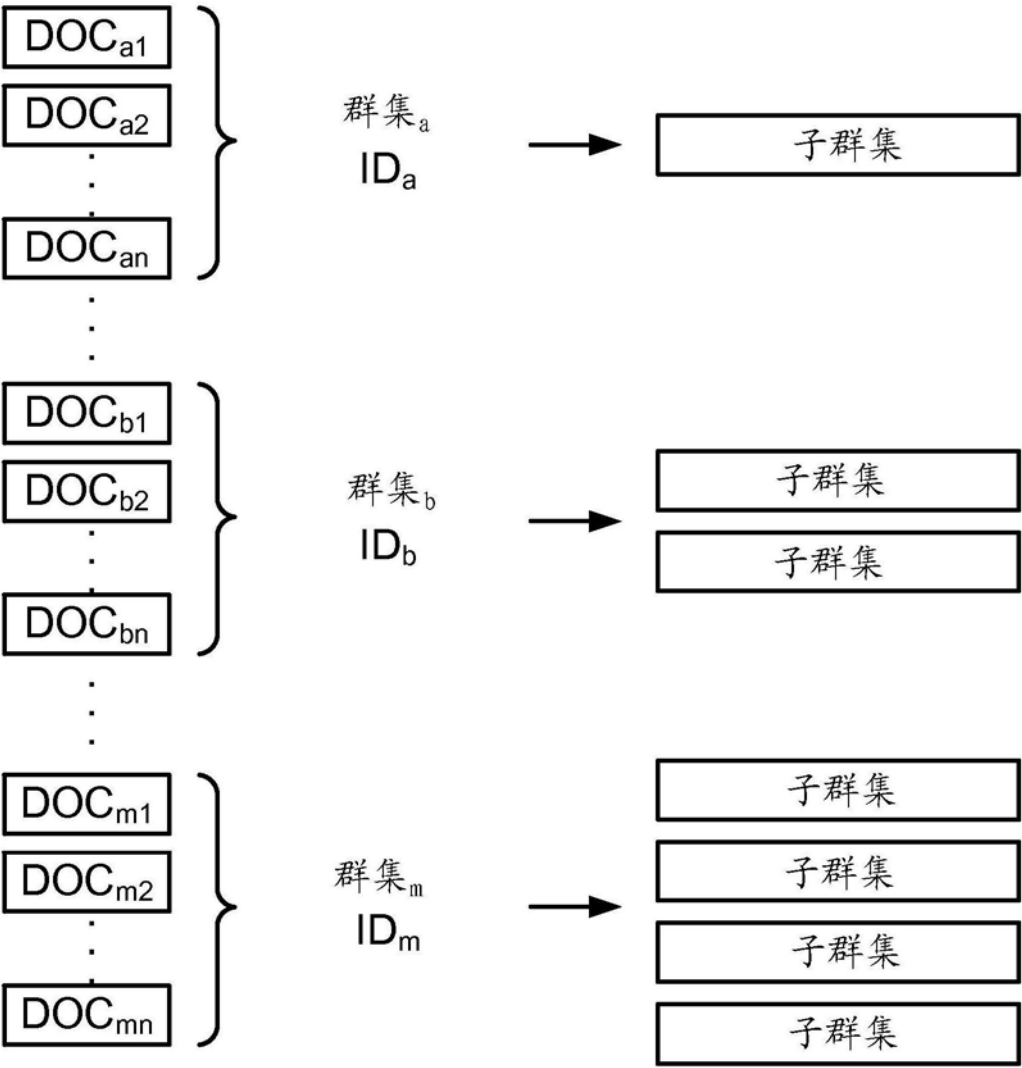


图4

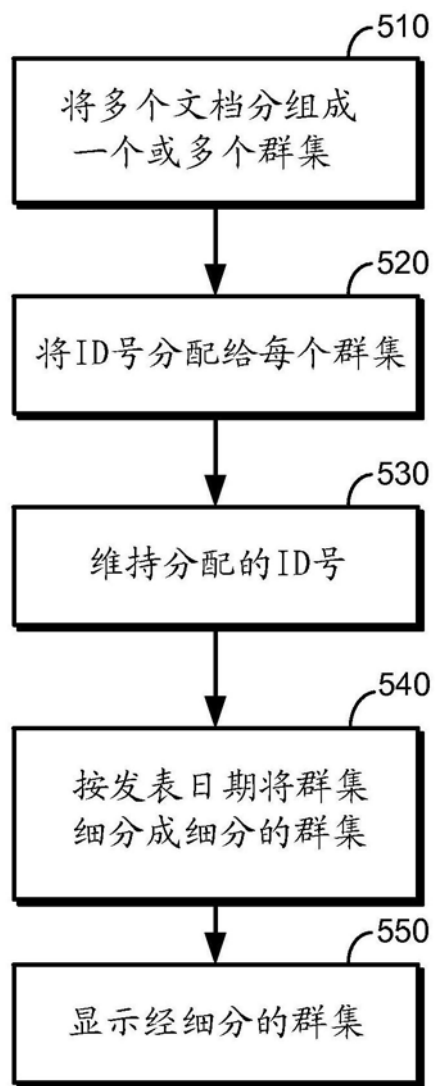


图5

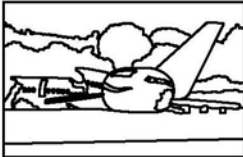
SERP – CARD TOP	
CHINA PLANE CRASH	
ALL RESULTS	1-10 OF 24,500 RESULTS - ADVANCED
ADVERTISEMENT TITLE	SPONSORED SITE
ATTRIBUTION.COM – LUCTUS ID HENDRERIT UT, AUCTOR EGET MET	
PLANE CRASH IN TIBET IS SHROUDED IN MYSTERY	
	LOREM IPSUO ALIQUET LOBORM DOLOR SIT AMET .ADIPISCING ELIT PRAESM POSUERE LEO ALIQUET ENT PLACERAT TINDICUNT ENIM. IN VEL QUALOBORTIS ET IN BASKH GRALE ET AL SIGM LIBERO TINDICUNT ENIM. AMESO IPSUM NEW YORK TIMES – 1 DAY AGO REPORT: PASSENGER PLANE CRASHES IN TIBET LA TIMES – 1 DAY AGO PASSENGER PLANE CRASHES IN TIBET PHILADELPHIA LOCAL NEWS – 1 DAY AGO SEE ALSO: 4 RELATED ARTICLES
PLANE CRASH IN CHINA KILLS 3 AMERICANS	
	LOREM IPSUO ALIQUET LOBORM DOLOR SIT AMET.ADIPISCING ELIT PRAESM POSUERE LEO ALIQUET ENT PLACERAT TINDICUNT ENIM. IN VEL QUALOBORTIS ET IN BASKH GRALE ET AL SIGM LIBERO TINDICUNT ENIM. AMESO IPSUM NEW YORK TIMES – 1 DAY AGO THRE AMERICANS KILLED IN A PLANE CRASH IN CHINA LA TIMES – 1 MONTH AGO CHINA PLANE CRASH: 42 KILLED, 3 AMERICANS NEW YORK TIMES - 1 MONTH AGO 3 AMERICANS DEAD IN CHINA PLANE CRASH PHILADELPHIA NEWS – 1 MONTH AGO CHINA: 42 DEAD FROM PLANE CRASH: 49 RESCUED CBS NEWS – 1 MONTH AGO
ALGO TITLE LOREM IPSUM DOLOR	
ALGO TITLE LOREM IPSUM DOLOR SIT AMET.ADIPISCING ELIT PRAESNT PLACERAT TINDICUNT ENIM. IN VE QUAM POSUERE LEO ALIQUET LOBORTIS ET IN LIBERO. MAURIS AT FAUCIBUS ENIM. NEW YORK TIMES	
ALGO TITLE LOREM IPSUM	
ALGO TITLE LOREM IPSUM DOLOR SIT AMET.ADIPISCING ELIT PRAESNT PLACERAT TINDICUNT ENIM. BBC NEWS	
ALGO TITLE LOREM IPSUM DOLOR	
ALGO TITLE LOREM IPSUM DOLOR SIT AMET.ADIPISCING ELIT PRAESNT PLACERAT TINDICUNT ENIM. IN VE QUAM POSUERE LEO ALIQUET LOBORTIS ET IN LIBERO. MAURIS AT FAUCIBUS ENIM. HUFFINGTON POST	
ALGO TITLE LOREM IPSUM	
ALGO TITLE LOREM IPSUM DOLOR SIT AMET.ADIPISCING ELIT PRAESNT PLACERAT TINDICUNT ENIM. LA TIMES	
ALGO TITLE LOREM IPSUM DOLOR	
ALGO TITLE LOREM IPSUM DOLOR SIT AMET.ADIPISCING ELIT PRAESNT PLACERAT TINDICUNT ENIM. IN VE QUAM POSUERE LEO ALIQUET LOBORTIS ET IN LIBERO. MAURIS AT FAUCIBUS ENIM. NEW YORK TIMES	

图6

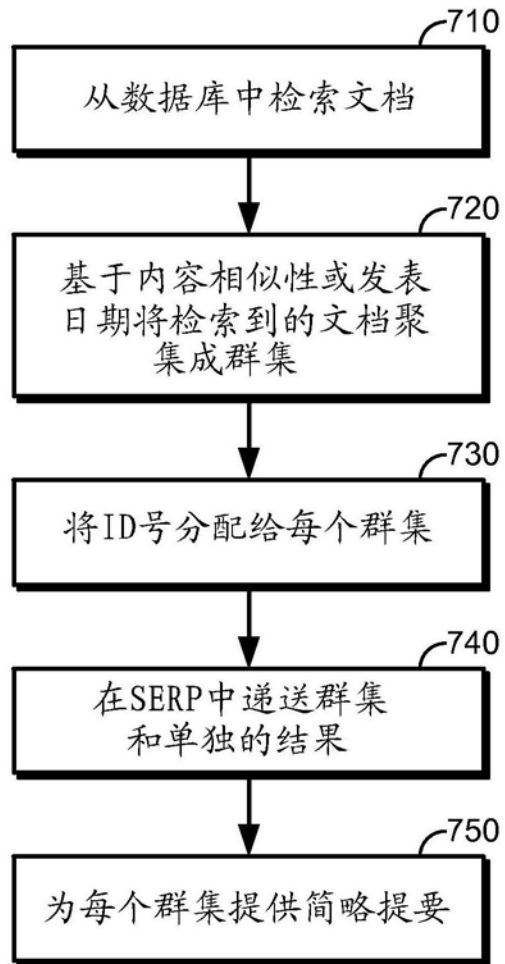


图7

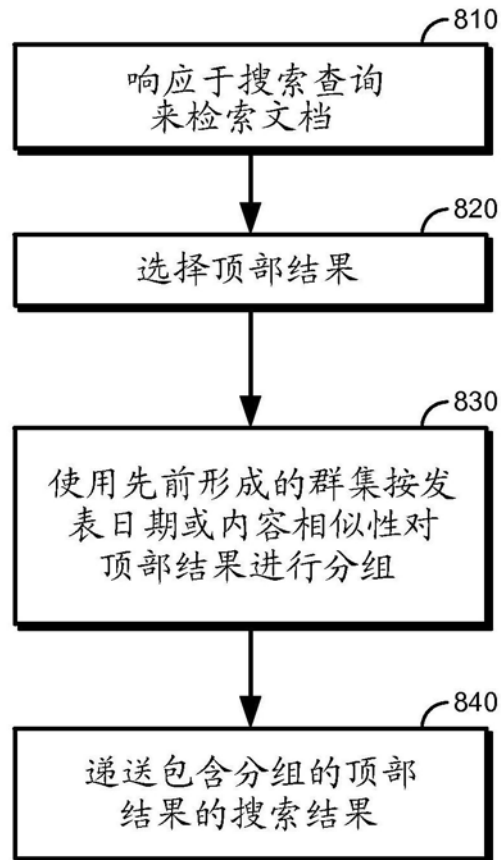


图8