



(12) 发明专利

(10) 授权公告号 CN 101911061 B

(45) 授权公告日 2013. 04. 10

(21) 申请号 200880123231. 4

(22) 申请日 2008. 11. 10

(30) 优先权数据

11/964, 729 2007. 12. 27 US

(85) PCT申请进入国家阶段日

2010. 06. 25

(86) PCT申请的申请数据

PCT/US2008/082983 2008. 11. 10

(87) PCT申请的公布数据

W02009/085410 EN 2009. 07. 09

(73) 专利权人 微软公司

地址 美国华盛顿州

(72) 发明人 M·比连科 M·A·芒森

(74) 专利代理机构 上海专利商标事务所有限公

司 31100

代理人 顾嘉运 钱静芳

(51) Int. Cl.

G06F 17/30 (2006. 01)

G06F 17/00 (2006. 01)

(56) 对比文件

US 20070185934 A1, 2007. 08. 09, 说明书
[0005]–[0057].

US 2005/0021530 A1, 2005. 01. 27, 全文.
Jaime Teevan 等. Personalizing Search
via Automated Analysis of Interests and
Activities. 《ACM. SIGIR》. 2005, 摘要,
1. Introduction–4. Evaluation Framework.

审查员 李佳

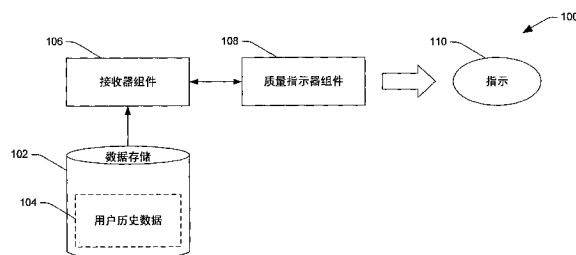
权利要求书 2 页 说明书 8 页 附图 9 页

(54) 发明名称

确定层分配的质量

(57) 摘要

本文描述了一种方法, 包括接收用户历史数据和生成用于将可搜索数字项存储在分层存储系统中的层分配的质量的指示, 其中该指示至少部分地基于用户历史数据的子集。本文还描述了一种系统, 包括接收用户历史数据的接收器组件。该系统还包括确定用来存储可通过查询检索的数字项的层分配的的质量的指示的质量指示器组件, 其中该质量指示器组件至少部分地基于用户历史数据的子集生成该指示并且该层分配指示数字项要被存储在分层存储系统中的哪里。



1. 一种用于确定层分配的的质量的方法,包括 :
接收用户历史数据 ;以及
生成用来将可搜索数字项存储在分层存储系统中的层分配的的质量的指示,其中所述指示至少部分地基于所述用户历史数据的子集 ;
其中生成所述指示包括 :
确定被分配给查询的权重,其中所述权重指示发起所述查询的频率 ;
为所述查询确定系统负载背景 ;以及
确定特定层是当用来执行所述查询的搜索引擎处于所述系统负载下时使用所述查询进行搜索时所访问的最低层的概率。
2. 如权利要求 1 所述的方法,其特征在于,所述用户历史数据包括 :
提供给搜索引擎的查询 ;
所述搜索引擎相对于所述查询返回的结果 ;
所述搜索结果中的被一个或多个用户选择的数字项 ;以及
在用户提交所述查询后被所述用户访问的一系列数字项。
3. 如权利要求 1 所述的方法,其特征在于,所述用户历史数据包括 :
输入到数据库管理系统的查询 ;
响应于所述查询返回的结果 ;以及
在用户提交所述查询后被该用户访问的一系列项。
4. 如权利要求 1 所述的方法,其特征在于,所述分层存储系统是搜索引擎索引。
5. 如权利要求 1 所述的方法,其特征在于,还包括至少部分地基于所述用户历史数据来生成经改进的层分配。
6. 如权利要求 5 所述的方法,其特征在于,生成所述经改进的层分配包括相对于层分配质量度量来最优化或基本上最优化所述经改进的层分配。
7. 如权利要求 5 所述的方法,其特征在于,还包括至少部分地基于所述经改进的层分配来更新用来将数字项分配给所述分层存储系统的各层的分层策略。
8. 如权利要求 7 所述的方法,其特征在于,还包括使用机器学习技术来参数化所述分层策略。
9. 如权利要求 1 所述的方法,其特征在于,生成所述指示包括使用以下算法 :

$$TQ(T(D), L, Q') = \sum_{q \in Q'} w(q) \sum_{t=1 \rightarrow k} P(t|q, T(D), L) \times Utility(t, R(q), T(D))$$
, 其 中
 $TQ(T(D), L, Q')$ 是所述层分配的的质量的量度, T 表示 k 个可用层的集合, D 是存储在所述 k 个层中的数字项的集合, L 是在处理查询集合 Q' 中的查询 q 期间观察到的系统负载, w 是分配给所述查询 q 的权重, $P(t|q, T(D), L)$ 是在所述系统负载 L 下所述 k 个层的第 t 层是检索期间访问的最低层的概率,效用 $(t, R(q), T(D))$ 所述查询 q 的质量量度,并且 $R(q)$ 是与所述查询 q 相对应的数字项的结果集。
10. 一种用于确定层分配的的质量的系统,包括 :
接收器组件(106),所述接收器组件接收用户历史数据(104) ;以及
质量指示器组件(108),所述质量指示器组件确定用来存储可通过查询检索的数字项的层分配的的质量的指示,其中所述质量指示器组件(108)至少部分地基于用户历史数据

(104) 的子集生成所述指示并且所述层分配指示数字项要被存储在分层存储系统中的哪里；

权重确定器组件,所述权重确定器组件确定分配给一个或多个查询的权重,其中所述权重指示发起所述一个或多个查询的频率；

负载确定器组件,所述负载确定器组件确定搜索引擎上的系统负载；以及

层确定器组件,所述层确定器组件确定特定层是当所述搜索引擎处于所述系统负载下时使用所述一个或多个查询时在所述分层存储系统中所访问的最低层的概率。

11. 如权利要求 10 所述的系统,其特征在于,所述指示是搜索引擎索引的层分配的质量的指示。

12. 如权利要求 10 所述的系统,其特征在于,还包括更新组件,所述更新组件至少部分地基于所述用户历史数据来生成经改进的层分配。

13. 如权利要求 12 所述的系统,其特征在于,所述更新组件组合多个层分配以生成所述经改进的层分配。

14. 如权利要求 12 所述的系统,其特征在于,所述更新组件至少部分地基于所述经改进的层分配来更新用来将数字项分配给所述分层存储系统的各层的分层策略。

15. 如权利要求 14 所述的系统,其特征在于,所述更新组件使用机器学习技术来更新所述分层策略。

16. 如权利要求 12 所述的系统,其特征在于,所述更新组件相对于层质量度量来最优化或基本上最优化层分配以生成所述经改进的层分配。

17. 如权利要求 10 所述的系统,其特征在于,还包括确定当搜索检索在特定层结束时搜索质量的测量的效用确定器组件。

18. 一种通过创建当前分层策略来将数字项分配给搜索引擎索引的各层的方法,所述方法包括：

接收用户历史数据；

输出层分配的质量的指示；

至少部分地基于接收到的用户历史数据和所述层分配的质量的指示来确定经改进的层分配；以及

更新分层策略以创建所述当前分层策略,其中所述更新至少部分地基于所述经改进的层分配；

基于所述当前分层策略将数字项分配给所述搜索引擎索引的各层；

其中输出所述指示包括：

确定被分配给查询的权重,其中所述权重指示发起所述查询的频率；

为所述查询确定系统负载背景；以及

确定特定层是当用来执行所述查询的搜索引擎处于所述系统负载下时使用所述查询进行搜索时所访问的最低层的概率。

确定层分配的质量

[0001] 背景

[0002] 搜索引擎使用户能够通过因特网快速访问信息。具体而言,用户可以向搜索引擎提交查询并细察该搜索引擎所返回的排名结果。例如,用户可以向搜索引擎提供查询“Spider(蜘蛛)”,并且被提供了与各种蛛形纲动物相关的网页、与汽车相关的网页、与电影相关的网页、与 web 爬行器相关的网页、以及其他网页。搜索引擎可被用来向查询的发起者返回图像、学术论文、视频、以及其他信息。

[0003] 搜索引擎的操作可包括采用 web 爬行器来定位和存储可在万维网上获得的大量信息(例如,网页)。例如,与其相关的网页或信息可以存储在搜索引擎索引中,该索引在接收到查询时使用(结合一个或多个搜索算法)。

[0004] 常规上,搜索引擎索引存储在若干层中,其中不同的层提供不同级别的性能。搜索引擎索引的分层与在计算机体系结构中使用的存储器分层结构类似:索引的总体存储容量在不同的级别之间划分,这些级别在大小、速度、等待时间、以及成本方面各异。索引的较高层通常具有较高速度但具有较小容量和较高成本。因此,仔细地索引网页以最大化搜索引擎的效率是合乎需要的。

[0005] 一种已经使用的用于将网页分层的方式是根据网页的由某一度量(如该网页的静态排名)确定的相对重要性来选择在其中放置该网页的索引的一层。具体而言,到一网页的链接的数量可被用来选择在其中定位该网页的索引的一层。然而,页面的相对重要性不一定指示该页面是否被经常访问,并且因而对于在搜索引擎索引中索引网页而言不是最优的。然而,评估层分配是困难的问题,因为不清楚哪些度量捕捉到了网页到这些层的特定分配的质量。

[0006] 概述

[0007] 以下是在本文详细描述的主题的简要概述。本概述不旨在权利要求的范围。

[0008] 本文描述了关于将数字项(如网页)分层的各种技术。可以监视用户与搜索引擎、数据库管理系统等的交互并且可以收集与这样的用户交互相关的数据。例如,可以监视并保留用户提交的查询、响应于这些查询提供的搜索结果(例如,数字项)、以及用户相对于这些搜索结果的动作。在一特定示例中,可以使用浏览器上的工具栏来收集用户历史数据。至少部分地基于该用户历史数据,可以生成可搜索数字项的层分配的质量的指示,其中层分配指示可搜索数字项被分配给若干层中的哪一层。层的质量的指示可以是符合定义的分层分配质量度量的值,该度量在本文详细描述。

[0009] 在一示例中,质量的指示可以通过查明若干参数来确定。例如,层分配的质量的指示可以至少部分地基于被分配给观察到的查询的权重。在一示例中,这些权重可以指示这些查询的相对重要性,并可至少部分地基于发起这些查询的频率。在另一示例中,对于特定查询和所确定的系统负载(例如,在接收到该查询时系统有多忙碌),层分配的质量的指示可以至少部分地基于数字项的检索将在指定层结束的概率。可为多个层确定该概率。在又一示例中,层分配的质量的指示可以至少部分地基于当检索在特定层中结束时获得的搜索结果质量的量度。可以结合确定分层质量的量度来使用归一化贴现累计收益(NDCG)、平均

精确度、Q- 测量或用于测量信息检索损失或搜索结果质量的其他合适机制。

[0010] 另外,可以至少部分地基于层分配的质量的指示和 / 或用户历史数据来生成经改进的层分配。例如,层分配的质量的指示可以符合定义的层分配质量度量,并且可以相对于该度量来最优化或基本上最优化经改进的层分配。此外,可以至少部分地基于经改进的层分配来更新分层策略。分层策略是用来的将数字项分配给各层的策略,并且可以考虑与数字项相对应的各种特征,如数字项被用户访问的次数、数字项的大小等。例如,分层策略可以通过使用机器学习技术来更新。

[0011] 在阅读并理解了附图和描述后,可以明白本申请的其他方面。

[0012] 附图简述

[0013] 图 1 是便于确定层分配的质量的指示的示例系统的功能框图。

[0014] 图 2 是生成层分配的质量的指示的示例组件的功能框图。

[0015] 图 3 是便于生成经改进的层分配的示例系统的功能框图。

[0016] 图 4 是便于生成经改进的层分配的示例系统的功能框图。

[0017] 图 5 是示出用于生成层分配的质量的指示的示例方法的流程图。

[0018] 图 6 是示出用于生成层分配的质量的指示的示例方法的流程图。

[0019] 图 7 是示出用于输出相对于层分配质量度量来最优化或基本上最优化的层分配的示例方法的流程图。

[0020] 图 8 是示出用于更新分层策略的示例方法的流程图。

[0021] 图 9 是示例计算系统。

[0022] 详细描述

[0023] 现在将参考附图来描述与确定层分配的质量、生成经改进的层分配、以及自动地更新分层策略有关的各种技术,全部附图中相同的附图标记表示相同的元素。另外,本文出于解释的目的示出并描述了各示例系统的若干功能框图;然而可以理解,被描述为由特定系统组件执行的功能可以由多个组件来执行。类似地,例如,单个组件可被配置成执行被描述为由多个组件执行的功能。

[0024] 参考图 1,示出了便于输出关于分层存储系统(未示出)的层分配的质量的指示的示例系统 100。按照一示例,分层存储系统可以是具有多个层的搜索引擎索引,其中第一(最高)层成本较高并且具有相对少量的存储空间,但从第一层检索的数字项的检索时间小于从其他层(较低层)检索的数字项的检索时间。第二层较不昂贵并且比第一层具有更多存储空间,但在与同第一层相对应的检索时间相比时检索时间较大。在另一示例中,分层存储系统可包括结合数据库管理系统使用的各层存储。例如,数据库管理系统中使用的服务器可具有硬盘驱动器、随机存取存储器、以及高速随机存取存储器,它们可以各自作为一层。

[0025] 系统 100 包括包含用户历史数据 104 的数据存储 102。用户历史数据 104 可包括例如用户发起的查询、响应于这些查询提供给用户的搜索结果、用户响应于被提供了各搜索结果而选择的一搜索结果、和 / 或其他合适的信息。在一示例中,用户历史数据 104 可以通过监视用户相对于搜索引擎的交互来累积。例如,工具栏插件可以安装在浏览器中,并且输入到浏览器的查询以及响应于这些查询返回的搜索结果、用户对特定搜索结果的选择、以及用户在提交查询后查看的一系列页面都可以由该工具栏插件来收集。

[0026] 接收器组件 106 接收用户历史数据 104 的子集。质量指示器组件 108 与接收器组件 106 通信并且从接收器子集 106 接收用户历史数据 104 的子集。质量指示器组件 108 可以生成层分配的质量的指示 110, 其中该层分配指示数字项要被分配给分层存储系统中的哪一层。例如, 质量的指示可符合层分配质量度量, 该度量在以下详细描述。另外, 以下更详细地描述质量指示器组件 108 的操作。

[0027] 现在参考图 2, 示出了质量指示器组件 108 的示例。质量指示器组件 108 包括权重确定器组件 202、负载确定器组件 204、层确定器组件 206、以及效用确定器组件 208。权重确定器组件 202 确定被分配给每一查询的权重, 质量指示器组件 108 使用该权重来生成与分层存储系统相对应的层分配的质量的指示。在一示例中, 权重确定器组件 202 所确定的权重可以至少部分地基于发起该查询的频率 (例如, 从查询日志查明的)。

[0028] 负载确定器组件 204 确定在搜索组件 (例如, 搜索引擎、数据库系统……) 执行特定查询时观察到的系统负载。系统负载可以至少部分地基于在处理特定查询时搜索组件所处理的查询的数量、在执行特定查询时检索搜索结果所专用的处理周期的数量、或搜索组件一般有多“忙碌”。

[0029] 层确定器组件 206 可以确定特定层是在负载确定器组件 204 所确定的系统负载下针对 (关于特定查询的) 数字项来搜索的最后一层的概率。一般而言, 在查询被输入搜索组件 (例如, 搜索引擎) 时, 首先在与较低层相比通常较小但具有较快访问和检索时间的较高层执行检索。取决于在较高层获得的结果的数量和质量以及当前系统负载, 可以在较低层执行或不执行检索。因此, 如上所述, 层确定器组件 206 可以确定特定层是针对 (关于特定查询并在所确定的系统负载下) 数字项来搜索的最后一层的概率。可为分层存储系统中的每一层确定该概率。

[0030] 效用确定器组件 208 确定当检索在特定层结束时搜索结果质量 (相对于特定查询) 的指示, 其中搜索结果质量的指示可以使用任何合适的度量来计算。在一示例中, 可以使用归一化贴现累计收益 (NDCG) 来确定搜索结果质量的指示。在另一示例中, 可以使用平均精确度 (MAP) 来确定搜索结果质量的指示。在又一示例中, 可以使用 Q- 测量来确定搜索结果质量的指示。因此, 可以理解, 效用确定器组件 208 可以利用任何合适的机制 / 度量来确定当检索在特定层结束时相对于特定查询的搜索结果质量的指示。

[0031] 权重确定器组件 202 所确定的权重、负载确定器组件 204 所确定的系统负载、层确定器组件所确定的概率、以及效用确定器组件 208 所确定的搜索结果质量的指示可由质量指示器组件 108 用来确定层分配的质量的指示。

[0032] 根据一示例, 以下算法可被用来定义层分配质量的度量, 并且可由质量指示器组件 108 用来确定层分配的质量的指示:

[0033]

$$(1) \quad TQ(T(D), L) = \sum_{q \in Q} w(q) \sum_{t=1 \rightarrow k} P(t|q, T(D), L) \times \text{效用}(t, q, T(D)),$$

[0034] 其中 $D = \{d_1, \dots, d_{|D|}\}$ 是要存储在具有对应的容量 $|T_1|, \dots, |T_k|$ 的 k 层 $T_1, \dots, T_k$ 中的所有数字项 (d_i) 的集合; $t(d_i)$ 是数字项集合 D 中的每一项的层分配, 其中 $t(d_i)$ 可具有值 $1, \dots, k$; $T(D) = \{t(d_1), \dots, t(d_i)\}$ 是层分配的总体集合; $TQ(T(D), L)$ 是针对当前系统负载 L 的层分配质量的量度; Q 是所有可能查询的集合; $w(q)$ 是查询 q 的权重 (例

如,相对重要性); $P(t|q, T(D), L)$ 是在当前系统负载 L 下第 t 层是检索期间访问的最低层的概率;以及效用 $(t, q, T(D))$ 是当检索在第 t 层结束时获得的搜索结果质量的量度。算法(1)因而计算跨给定层分配的所有可能查询跨每一层中的结束检索的概率分布的总体层分配质量的期望。

[0035] 然而,可以理解,所有可能查询的数量是无限的。因此,观察到的查询的集合 Q' 可由质量指示器组件 108 用作所有可能查询的分布的近似。在一示例中,这些观察到的查询 Q' 可以从包括多个观察到的查询(例如,用户历史数据 104)的数据储存库中随机选择,其中选择任何查询 $q \in Q'$ 的概率可以作为选择搜索组件(例如,搜索引擎、数据库管理系统……)接收到的随机查询的可能性来计算。在另一示例中,可以选择观察到的查询的集合 Q' 以使得它们表示所有可能的查询。例如,可以选择查询 Q' 以使得具有特定长度(以单词、字符等来测量的)的查询的数量不超过阈值。另外,可以选择针对不同主题的查询。在又一示例中,可以基于与这些查询相关联的用户数据的量来选择查询 Q' 。例如,查询 Q' 可被限于具有与其相关联的相继用户数据的查询,如用户点击响应于这些查询提供的一个或多个搜索结果和/或广告。可以理解,构想了用于选择观察到的查询的子集的任何合适的方式,并且这些方式旨在落入所附权利要求书的范围内。

[0036] 对于 Q' 中的每一所选查询 q ,质量指示器组件 108 可以构造包括只有 M 项的相关结果集 $R(q) = \{d_{q,1}, \dots, d_{q,M}\}$,其中这些项可部分地从最相关到最不相关来排序。在一示例中,结果集可包括用户在向搜索组件提交查询后频繁选择/访问的数字项,其中选择/访问的频率可以与用户查看这些数字项所花费的时间相组合;和/或可包括搜索组件跨分层存储系统的所有层返回的作为查询的相关结果的数字项。

[0037] 使用查询 Q' 和对应的结果集,以下算法可被用来定义层分配质量的度量,并且可由质量指示器组件 108 用来确定层分配的质量的指示:

[0038]

$$(2) \quad TQ(T(D), L, Q') = \sum_{q \in Q'} w(q) \sum_{t=1 \rightarrow k} P(t|q, T(D), L) \times \text{效用}(t, R(q), T(D)),$$

[0039] 其中 $TQ(T(D), L, Q')$ 是相对于查询集合 Q' 在当前系统负载 L 下层分配质量的量度;并且效用 $(t, R(q), T(D))$ 是当检索在第 t 层结束时获得的搜索结果质量的量度。

[0040] 如上所述,质量指示器组件 108 可以确定层分配的质量的指示。更具体地,权重确定器组件 202 可以确定查询集合 Q' 中的每一查询的权重(w)。负载确定器组件 204 可以确定查询集合 Q' 中的每一查询所造成的系统负载 L 。层确定器组件 206 可以确定 $P(t|q, T(D), L)$,并且效用确定器组件 208 可以确定效用 $(t, R(q), T(D))$ 。在一示例中,效用确定器组件 208 可以使用归一化贴现累计收益(NDCG)来确定效用 $(t, R(q), T(D))$ 。效用确定器组件 208 可以采用其他机制来测量效用;各示例包括平均精确度(MAP)以及 Q -测量。这些示例不旨在是限制性的,因为可以采用并构想了用于测量效用的其他机制。

[0041] 在一特定示例中,效用确定器组件 208 可以利用以下算法来确定当检索在第 t 层结束时的搜索结果质量的测量,其中该算法是 NDCG 的修改:

[0042]

$$\text{效用}_{NDCG}(t, R(q), T(D)) = N \sum_{d \in R_t(q)} \frac{2^{\text{相关性}(d)} - 1}{\log(\text{排名}(d) + 1)} \quad (3)$$

[0043] 其中 N 是归一化因子, $R_t(q)$ 是存储在 1 到 t 层的 $R(q)$ 中的数字项的排序子集, 相关性 (d) 是数字项 d 的相关性分数, 并且是排名 (d) 是该数字项在 $R_t(q)$ 中的排名位置。注意, 在更多相关数字项处于较低 (较深) 层中的情况下, 排名 (d) 依赖于 t ; 在检索不超过 t 层的情况下, 这些未被检索。如上所述, 使用 NDCG 的修改仅是在给定当前层分配的情况下特定查询的搜索结果质量的一个可能量度, 并且可以利用其他量度, 如接收到的相关结果的比例等。

[0044] 如可从上文理解的, 用户历史数据 104 (图 1) 可被用来构造可被用来评估层分配的查询集合 Q' 和对应的结果集 $R(q)$ 。假设当前层分配是 $T(D)$, 则 $P(t|q, T(D), L)$ 可以针对特定系统来实例化以反映在分层存储系统中用来在观察到的负载 L 下将查询转发到第 t 层的分层策略。随后, 给定替换层分配 (例如, $T_1(D)$ 和 $T_2(D)$), 可以通过计算 TQ 来选择优选分配。另外, 质量指示器组件 108 可以使用 TQ 在变化的负载下 (并且因而调查在不同的负载下层分配的质量) 以及对于用于将查询转发到不同层的分层策略的不同实例, 来调查搜索结果的预期质量如以下详细描述。

[0045] 现在参考图 3, 示出了便于自动地更新相对于分层存储系统的层分配的示例系统 300。系统 300 包括可包含多个层的分层存储系统 302, 其中这些层中的每一层都可被用来存储一个或多个数字项, 如网页、图像、文档等。搜索组件 304 至少部分地基于接收到的查询来执行对存储在分层存储系统 302 中的数字项的搜索。例如, 搜索组件 304 可以是被配置成响应于接收到查询而搜索通过分层搜索索引的搜索引擎。在另一示例中, 搜索组件 204 可以是用于响应于接收到查询而搜索存储的各层 (例如, 存储器、硬盘驱动器……) 的数据管理系统的一部分。在又一示例中, 搜索组件 304 可以是用于搜索计算机上的项的桌面搜索模块。还构想了其他搜索组件。

[0046] 数据存储 102 保留可从搜索组件 304 接收到的用户历史数据 104。例如, 提供给搜索组件 304 的查询、被提供了搜索结果后的用户动作、以及响应于该查询而提供给用户的搜索结果集合都可以存储在用户历史数据 104 中。接收器组件 106 接收用户历史数据 104 的子集。如上所述, 质量指示器组件 108 可以生成层分配的质量的指示 110。在一示例中, 在由质量指示器组件 108 生成指示 110 后, 指示 110 可被存储在计算机可读介质中。

[0047] 更新组件 306 可以接收指示 110 并至少部分地基于指示 110 输出经改进的层分配 308。例如, 更新组件 306 可以接收其他可能的层分配和对应的质量指示, 并选择与最高质量指示相对应的层分配。例如, 更新组件 306 可以使用试探法来确定最优或基本上最优的层分配 (相对于对应的层分配质量度量)。在另一示例中, 如以下更详细地描述的, 更新组件 306 可以利用机器学习技术来输出经改进的层分配 308。随后可至少部分地基于经改进的分层分配 308 来将数字项 310 分配给分层存储系统 302。

[0048] 有了关于更新组件 306 的更多细节, 初始层分配的质量的指示 110 可以提供用于开发用于标识数字项的经改进的层分配的算法 / 技术的基础。给定可能的层分配的空间 $T = \{T^{(1)}(D), \dots, T^{(N)}(D)\}$, 标识如由等式 (2) 输出的具有最优或基本上最优的层质量指示的层分配 $T^*(D)$ 可以如下定义:

$$[0049] \quad (4) \quad T^*(D) = \arg \max_{T^{(i)}(D) \in T} TQ(T^{(i)}(D), L, Q')$$

[0050] 可能的层分配 T 的集合可被定义为由一些变量 (如静态排名方案的参数) 参数化

的替换分配或分配组的集合。随后更新组件 306 可以使用机器学习技术来搜索替换分配的集合以将这样的分配之一标识为最优或基本上最优。例如,更新组件 306 可以使用神经网络、回归树、贝叶斯网络、或任何其他机器学习技术来确定最优化或基本上最优化指示 110 的分层分配。

[0051] 此外,更新组件 306 可以确定被用来至少部分地基于经改进的层分配 308 和 / 或用户历史数据 104 的子集将数字项 310 分配给分层存储系统 302 中的特定层的分层策略 312。分层策略可被用来确定在存储数字项时要使用分层存储系统 302 的哪些层。例如,分层策略 312 可以考虑可响应于一个或多个查询来返回的可搜索数字项的各个特征。这些特征可包括从链结构导出的静态排名(例如,数字项的页面排名)、包括该数字项的域的排名、数字项在搜索引擎结果之间的流行性、数字项中单词的数量、数字项中图像的色谱,等等。这些特征中的每一个可由更新组件 306 来参数化。换言之,可以向这些特征分配权重,分层策略 312 使用这些权重来将对应的数字项分配给分层存储系统 302 中的一层。更新组件 306 可以使用机器学习技术来学习要被分配给这些特征的权重,并且分层策略可以被用来将数字项分配给分层存储系统 302 的各层。

[0052] 现在参考图 4,示出了便于基于多个可能的层分配来更新层分配的示例系统 400。系统 400 包括可以生成层分配的的质量的指示的质量指示器组件 108。更具体地,质量指示器组件 108 可以至少部分地基于用户历史数据 104 生成第一层分配 402 到第 N 层分配 404 的质量的指示。更新组件 306 可以接收质量的指示(其可以是与定义的层分配质量度量相对应的值)并且将若干不同的层分配以如下方式组合:所得的经改进的层分配 406 与任何单独的层分配相比具有更高的质量(由质量指示器组件 108 确定)。更新组件 306 可以至少部分地基于与层分配 402-404 和 / 或用户历史数据 104 的子集相对应的的质量的指示来组合不同的层分配。

[0053] 更具体地,组合层分配可以是算法(4)的特定实例,其中可能的分配的集合 T 可以是各单独的层分配的可能组合的集合。可能组合的集合可以由一些变量(如静态排名方案的参数)来参数化。更新组件 306 可以使用机器学习技术来确定相对于定义的层分配质量度量最优或基本上最优的各单独的层分配的组合。另外,如上所述,更新组件 306 可以至少部分地基于经改进的层分配 406 生成或更新被用来将数字项分配给分层存储系统的各层的分层策略 312。

[0054] 现在参考图 5-8,示出并描述了各示例方法。尽管各方法被描述为顺序地执行的一系列动作,但可以理解,这些方法不受该顺序的次序的限制。例如,一些动作能以与本文描述的不同的次序发生。另外,动作可以与另一动作同时发生。此外,在一些情况下,实现本文描述的方法并不需要所有动作。

[0055] 此外,本文描述的动作可以是可由一个或多个处理器实现的和 / 或存储在一个或多个计算机可读介质上的计算机可执行指令。计算机可执行指令可包括例程、子例程、程序、执行的线程等。另外,搜索引擎和 / 或数据库管理系统中的层分配可以至少部分地基于本文描述的方法来确定。另外,这些方法的动作的结果可以存储在计算机可读介质中,显示在显示设备上,等等。

[0056] 具体参考图 5,示出了用于确定层分配的的质量的指示的示例方法 500。方法 500 在 502 开始,并在 504 接收用户历史数据。例如,用户历史数据可包括由用户发起的查询、响应

于查询而提供给用户的搜索结果、用户对搜索结果的选择、以及在发起查询后用户所查看的一系列页面。用户历史数据还可包括标记数据,其中搜索结果对查询的相关性由用户明确定义。

[0057] 在 506,至少部分地基于用户历史数据的子集来生成层分配的的质量的指示。方法 500 在 508 完成。

[0058] 现在转向图 6,示出了便于确定相对于分层存储系统的层分配的的质量的指示的方法 600。方法 600 在 602 开始,并且在 604 确定分配给查询的权重。例如,该权重可取决于发起该查询的频率。在另一示例中,一个或多个用户可以明确地向查询分配权重以指示该查询的相对重要性。

[0059] 在 606,确定该查询的系统负载背景。如上所述,系统负载可以与由诸如搜索引擎或数据库管理系统的搜索组件在处理该查询时正在处理的查询的数量相关。

[0060] 在 608,确定特定层是当搜索引擎在该系统负载下时所访问的最低层的概率。例如,可为用来存储可搜索数字项的每一层确定这一概率。

[0061] 在 610,确定层分配的的质量的指示,其中使用层分配来在分层存储系统中存储与查询相对应的数字项。质量的指示是至少部分地基于权重、系统负载、以及所确定的概率来确定的。在一示例中,可以将所确定的质量的指示至少临时地存储在计算机可读介质中。方法 600 在 612 结束。

[0062] 现在参考图 7,示出了用于确定最优或基本上最优的层分配(例如,针对定义的层分配质量度量来最优化或基本上最优化)的方法 700。方法 700 在 702 开始,并且在 704 接收多个不同的层分配。在 706,接收用户历史数据。如上所述,用户历史数据可包括查询、响应于查询而提供的搜索结果、和 / 或用户对响应于查询而提供的搜索结果的选择。

[0063] 在 708,为多个不同的层分配的子集来确定质量的指示。在 710,组合层分配以使得所得组合与任何单独的层分配相比具有更高的质量指示。方法 700 在 712 结束。

[0064] 现在参考图 8,示出了便于更新分层策略的方法 800。在一示例中,使用分层策略来将数字项分配给搜索引擎索引的各层的搜索引擎可以使用方法 800 的动作来更新分层策略。方法 800 在 802 开始,并在 804 接收用户历史数据。在 806,确定层分配的的质量的指示。在 808,至少部分地基于用户历史数据和 / 或在 806 确定的质量的指示来确定经改进的层分配。在 810,至少部分地基于用户历史数据和经改进的层分配来更新分层策略。例如,经改进的层分配可以考虑与用户历史数据相关的数字项,并且分层策略可以用来将在经改进的层分配中未考虑的数字项分配给特定层。方法 800 在 812 结束。

[0065] 现在参考图 9,示出了可以根据本文公开的系统和方法使用的示例计算设备 900 的高级图示。例如,计算设备 900 可以用于搜索引擎系统中。在另一示例中,计算设备 900 可以用于数据库管理系统中。计算设备 900 可以是服务器,或可以在诸如个人计算机、个人数字助理等常规上被认为客户机设备的设备中使用。计算设备 900 包括执行存储在存储器 904 中的指令的至少一个处理器 902。这些指令可以是例如用于实现被描述为由上述一个或多个组件执行的功能的指令或用于实现上述方法中的一个或多个的指令。处理器 902 可以通过系统总线 906 访问存储器。除存储可执行指令之外,存储器 904 还可以存储数字项、层分配的至少一部分、一个或多个层分配的的质量的指示等。

[0066] 计算设备 900 另外包括可由处理器 902 通过系统总线 906 访问的数据存储 908。数

据存储 908 可包括可执行指令、一个或多个层分配、层分配的质量的指示、用户历史数据、标记数据等。计算设备 900 还包括允许外部设备与计算设备 900 进行通信的输入接口 910。例如,输入接口 910 可被用来通过网络接收来自用户的查询。计算设备 900 还包括将计算设备 900 与一个或多个外部设备进行接口的输出接口 912。例如,计算设备 900 可以通过输出接口 912 显示搜索结果。

[0067] 另外,尽管被示为单个系统,但可以理解,计算设备 900 可以是分布式系统环境。因此,例如,若干设备可以通过网络连接进行通信并且可共同执行被描述为由计算设备 900 执行的任务。

[0068] 如此处所使用的,术语“组件”和“系统”旨在涵盖硬件、软件、或硬件和软件的组合。因此,例如,系统或组件可以是进程、在处理器上执行的进程、或处理器。另外,组件或系统可以位于单个设备上或分布在若干设备之间。

[0069] 注意,出于解释目的提供了若干示例。这些示例不应被解释为限制所附权利要求书。另外,可以认识到,本文提供的示例可被改变而仍然落入权利要求的范围内。

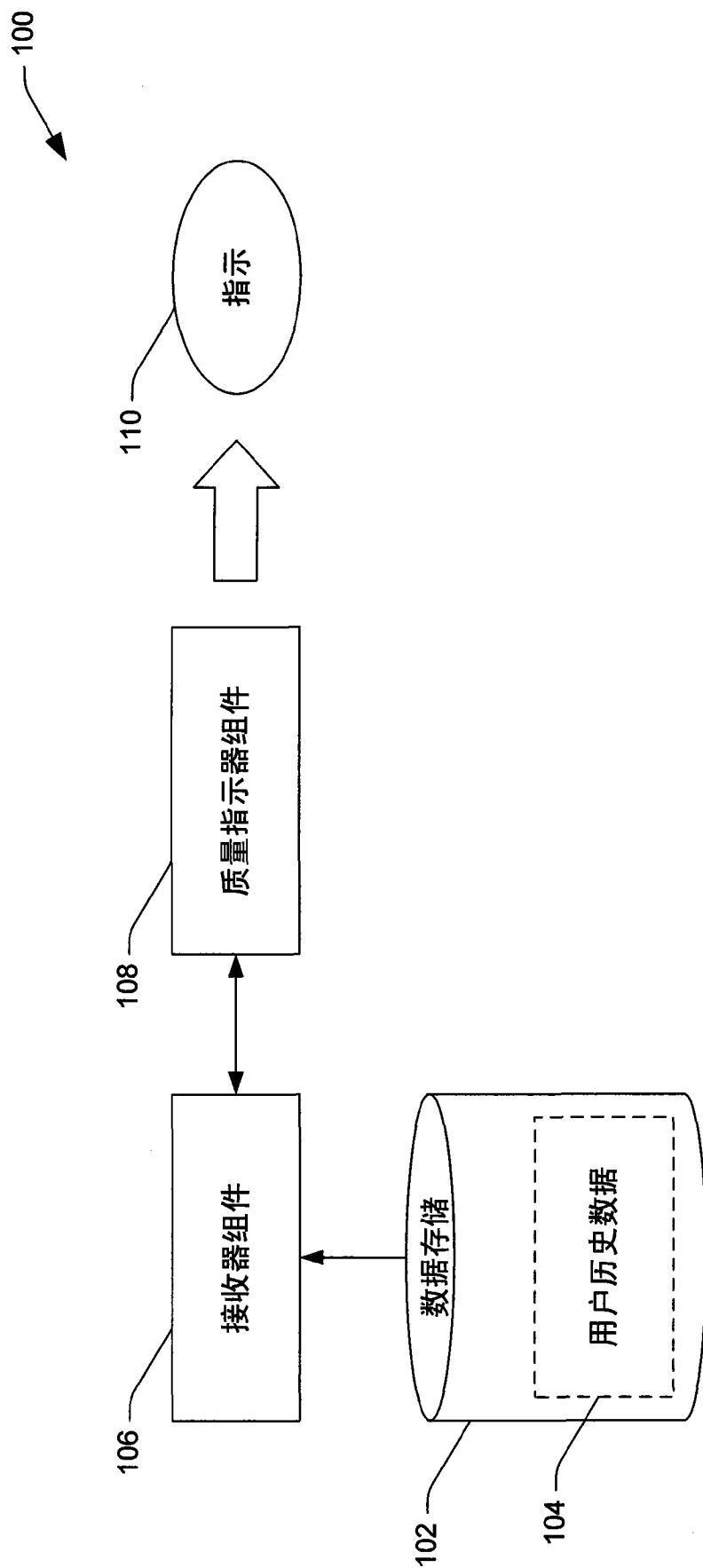


图 1

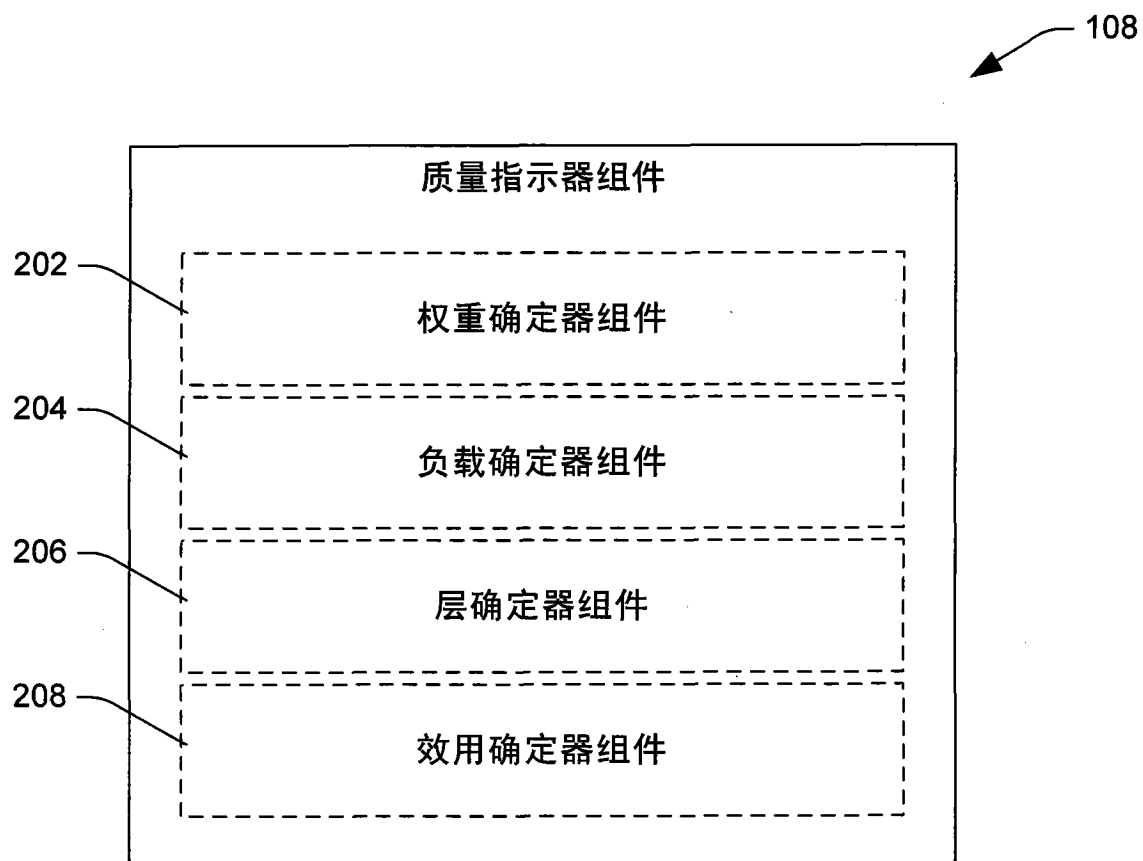


图 2

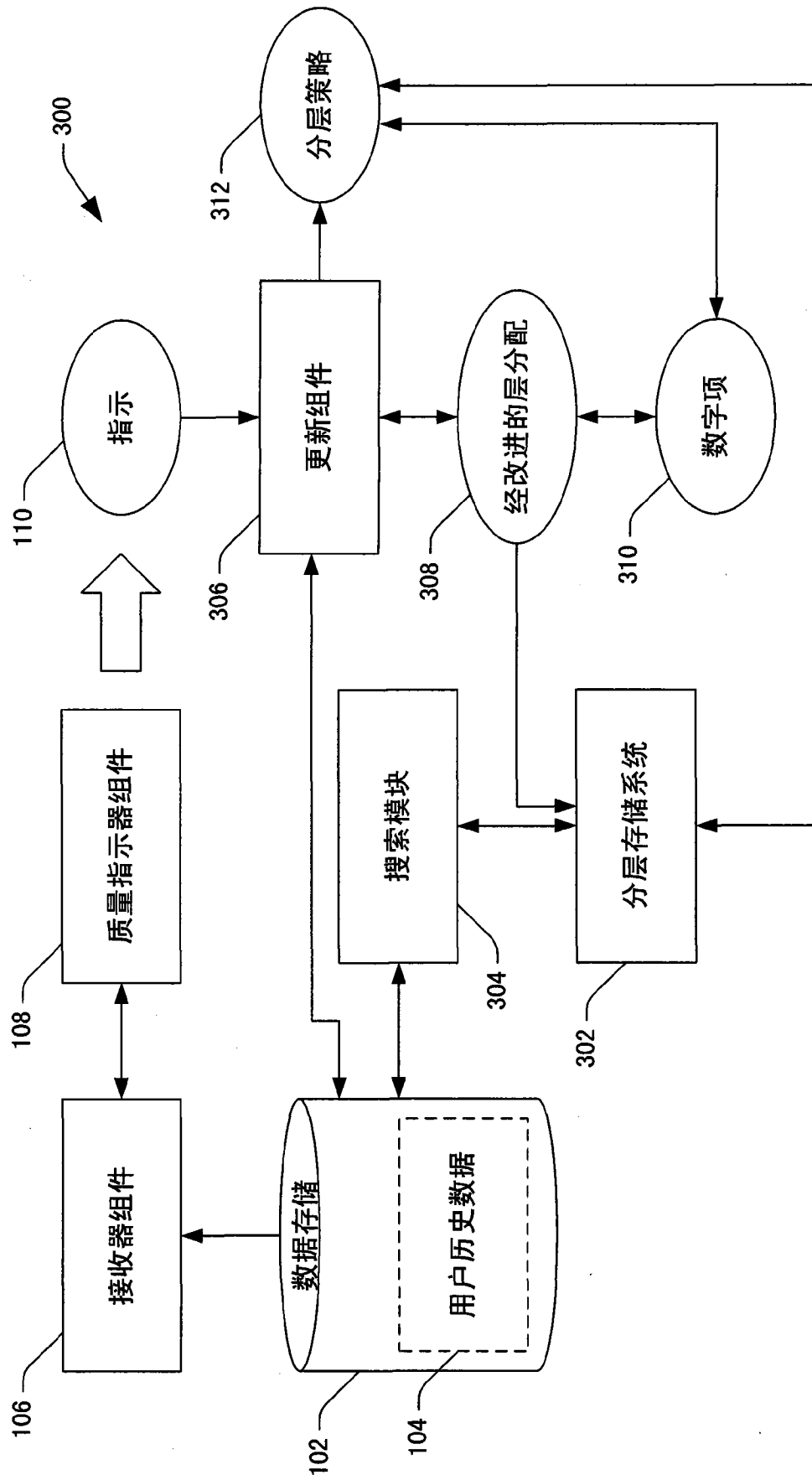


图 3

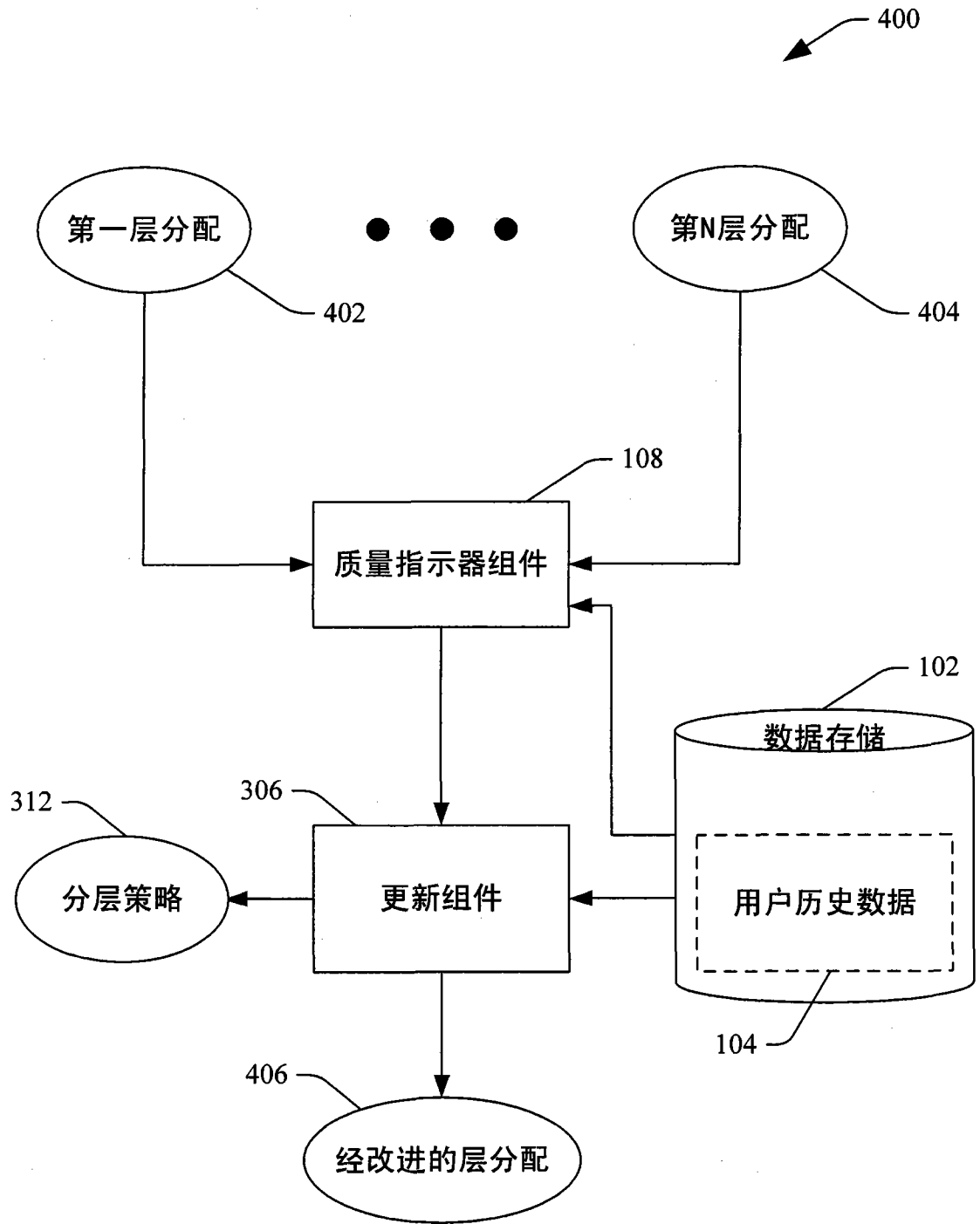


图 4

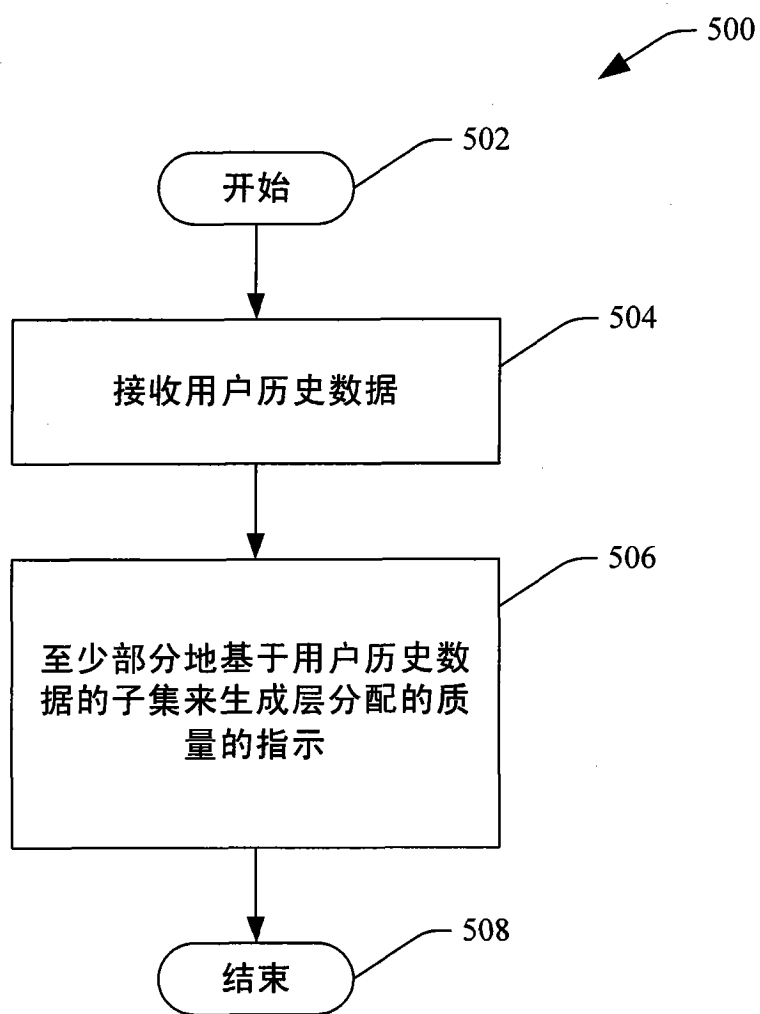


图 5

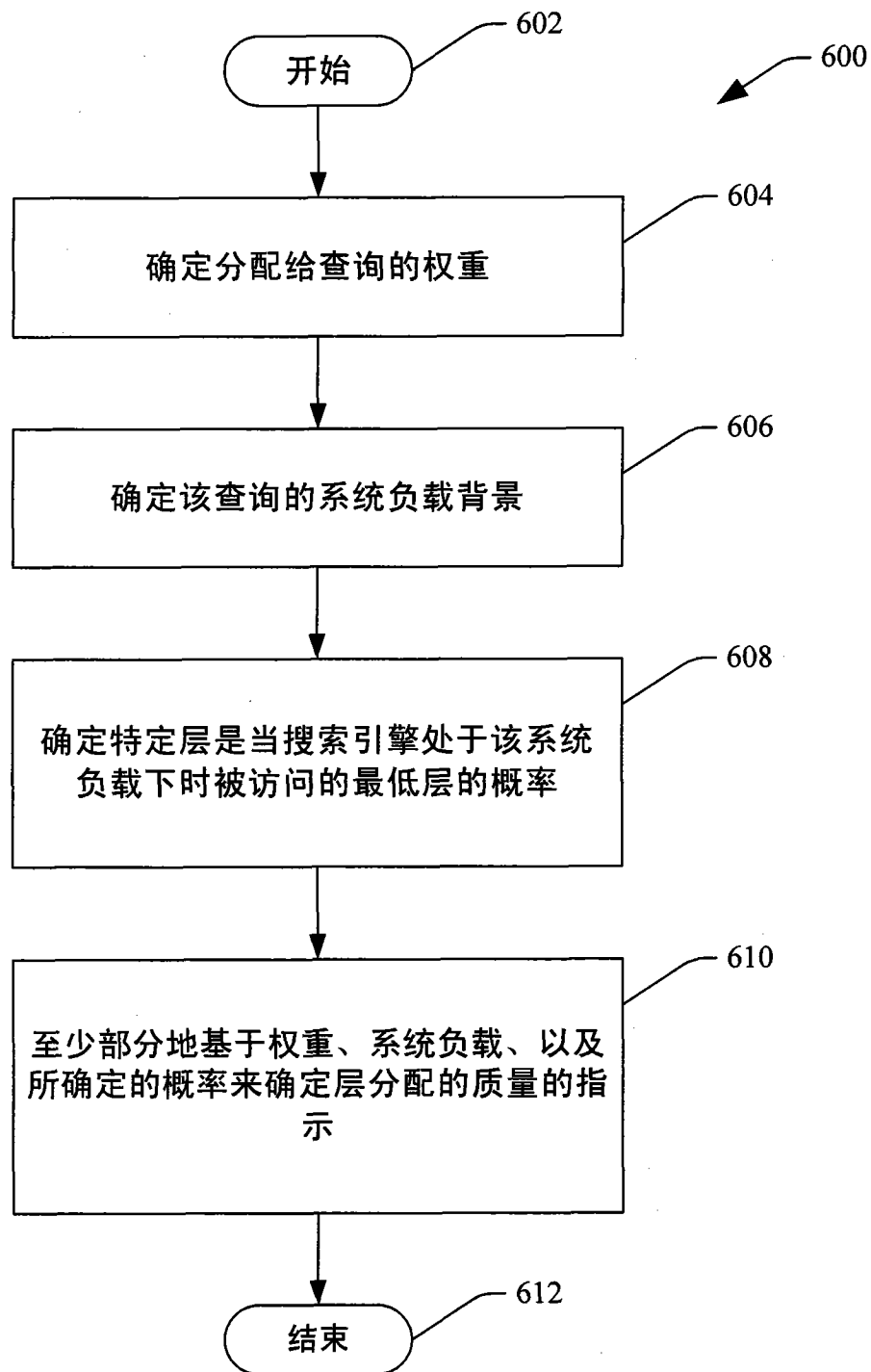


图 6

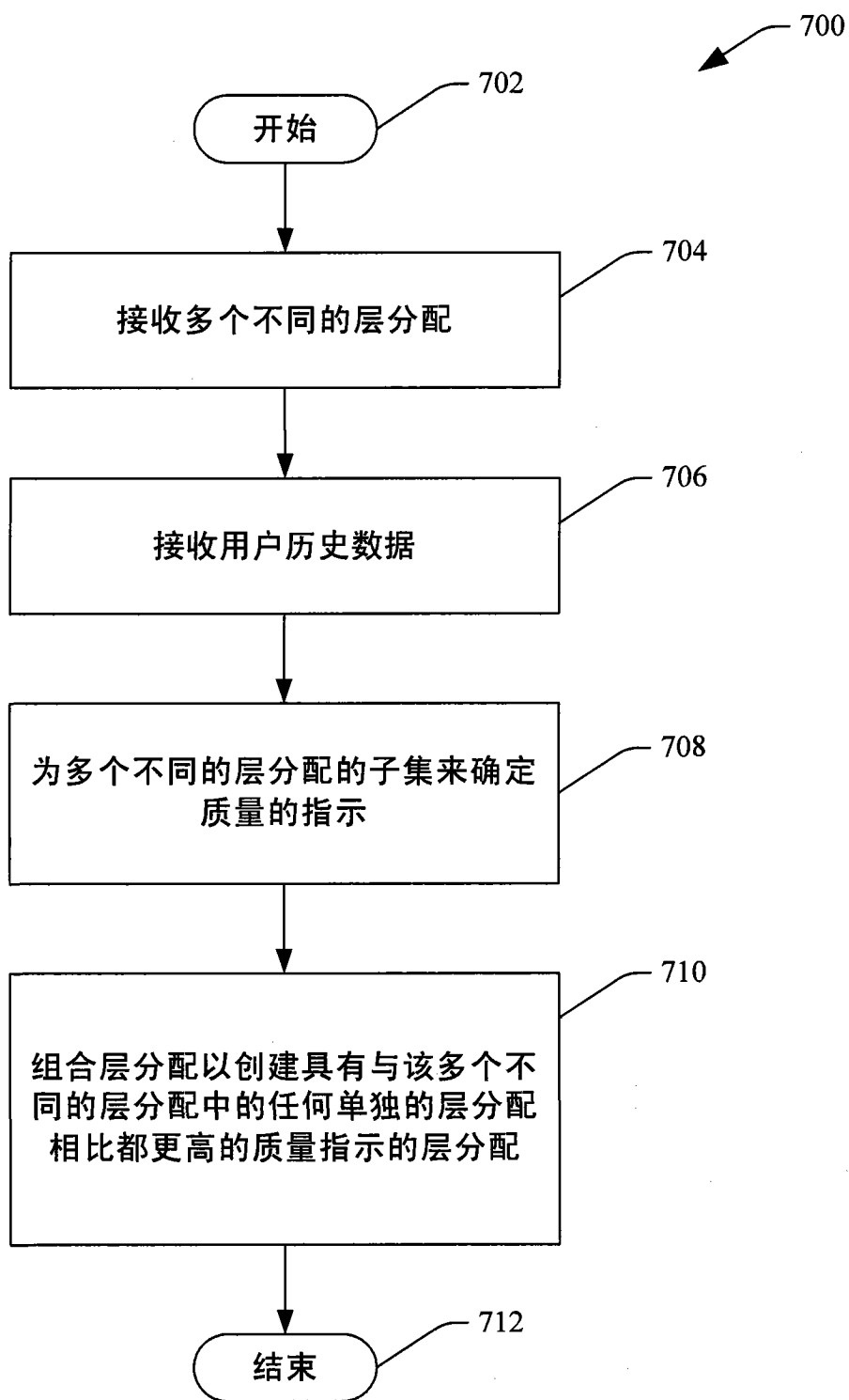


图 7

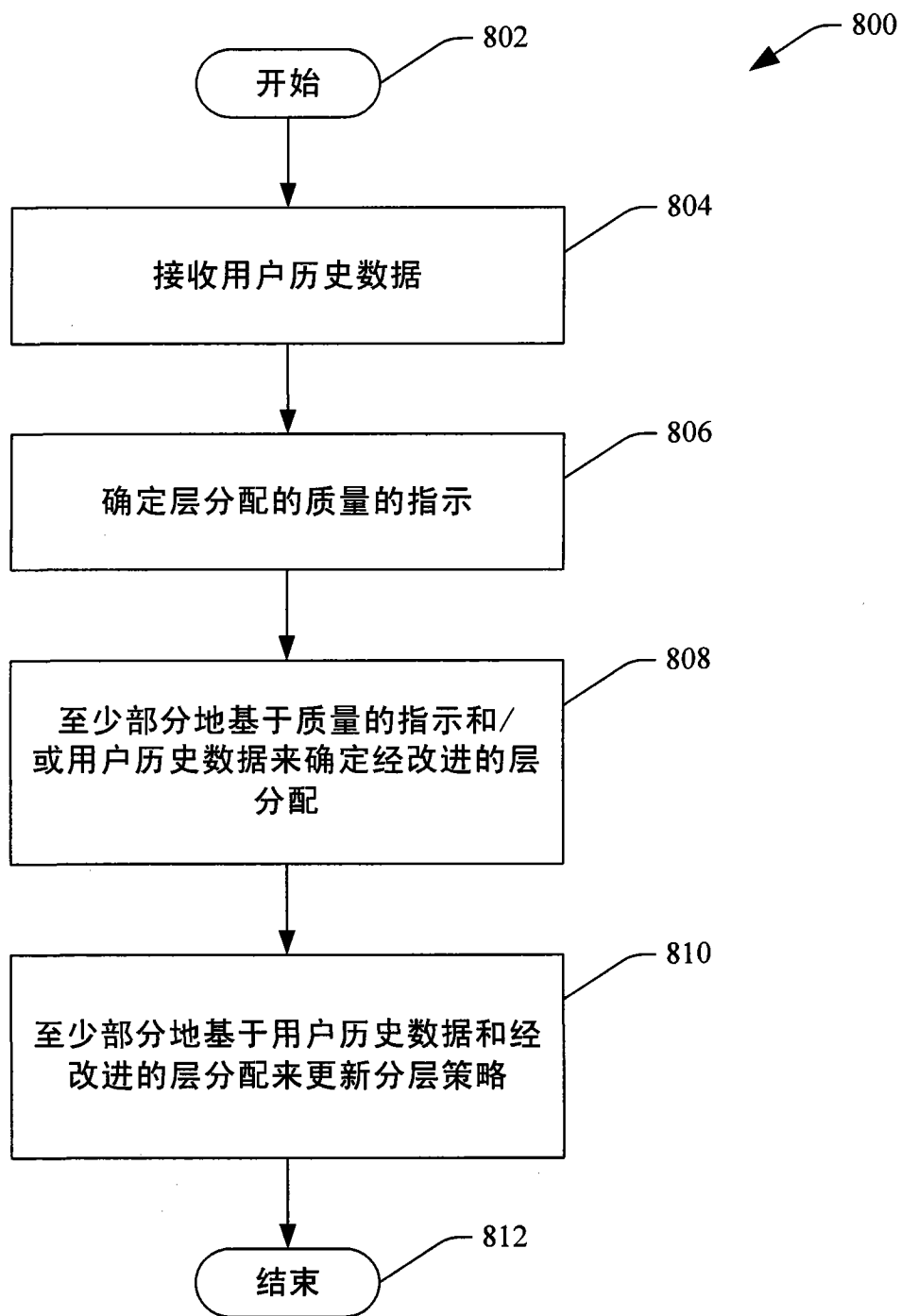


图 8

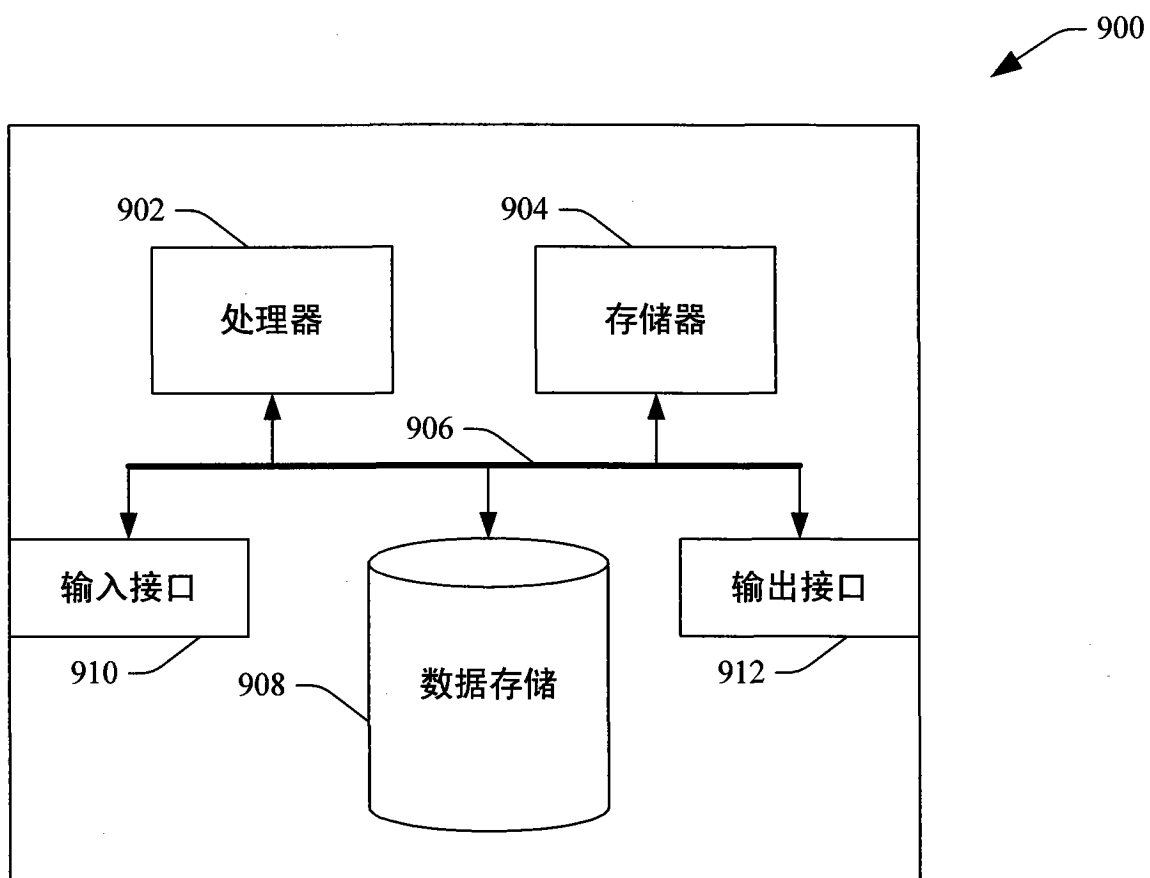


图 9