(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 102682059 B
(45)授权公告日 2014.11.12

(21)申请号 201210016687.9
(22)申请日 2006.08.15
(30)优先权数据
11/204,922 2005.08.15 US
(62)分案原申请数据
200680038100.7 2006.08.15
(73)专利权人 谷歌公司
地址 美国加利福尼亚州
(72)发明人 马尤尔·达塔尔 阿舒托什·加尔格
(74)专利代理机构 中原信达知识产权代理有限责任公司 11219
代理人 周亚荣 安翔
(51)Int.Cl.
G06F 17/30(2006.01)
(56)对比文件
US 6134532 A,2000.10.17, 全文.
CN 1578952 A,2005.02.09, 全文.
CN 1476568 A,2004.02.18, 全文.
CN 1462004 A,2003.12.17, 全文.
CN 1648903 A,2005.08.03, 全文.
审查员 张伯

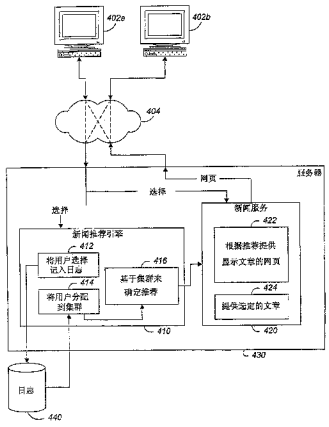
权利要求书3页 说明书13页 附图3页

(54)发明名称

用于将用户分配到集群的方法和系统

(57)摘要

本发明涉及用于将用户分配到集群的方法和系统。一方面,运行程序以获得多个用户之中的每个用户各自的兴趣集,每个兴趣集代表在其中各个用户表达过兴趣的项目;为每个用户,确定各个兴趣集的k个散列值,其中第i个散列值是在对应的第i个散列函数之下的最小值;并将多个用户之中的每一个用户分配到为各个用户建立的各个k个集群之中的每一个集群,第i个集群由第i个散列值所代表。每个用户到k个集群的分配的完成不考虑任何其它用户到k个集群的分配。



CN 102682059 B

1. 一种用于将用户分配到集群的方法，所述方法包括：

为多个用户之中的每一个用户获得各自的兴趣集，每个兴趣集是元素集，每个元素代表所述各个用户已经通过与数据处理系统进行交互而表达了兴趣的各个项目；

对所述多个用户之中的每一个用户，对于在 1 和 k 之间的每个整数 i，向用户的兴趣集的每个元素应用第 i 个散列函数以获得与各个元素相对应的各个函数值，k 个散列函数彼此不同，并且根据从所述 k 个散列函数获得的函数值来确定所述各自的兴趣集的 k 个散列值，其中所述各自的兴趣集的第 i 个散列值是通过向用户的兴趣集的元素应用第 i 个散列函数而获得的函数值中的最小值，其中 k 是大于或等于 1 的整数；以及

将所述多个用户之中的每一个用户分配到为所述各个用户建立的 k 个集群之中的每一个集群，第 i 个集群由所述各个用户的各自的兴趣集的第 i 个散列值所代表，其中所述将多个用户之中的每一个用户分配到 k 个集群的完成不考虑任何其他用户到 k 个集群的分配。

2. 如权利要求 1 所述的方法，还包括：

将表达用户兴趣的行为记录在日志中；以及

使用所述日志来为所述多个用户生成所述兴趣集。

3. 如权利要求 1 所述的方法，还包括：

执行协同过滤计算机程序应用，以基于所述多个用户的第一用户到 k 个集群之中的一个或多个集群的分配而向所述第一用户提供信息。

4. 如权利要求 1 所述的方法，还包括：

为所述多个用户的第一用户获得已改变的兴趣集；

使用所述已改变的兴趣集，为所述第一用户确定 k 个散列值；以及

将所述第一用户仅分配到由通过使用所述已改变的兴趣集所确定的 k 个散列值所代表的各个 k 个集群之中的每一个集群，而不改变任何所述其他用户到 k 个集群的分配。

5. 一种用于将用户分配到集群的系统，所述系统包括：

用于为多个用户之中的每一个用户获得各自的兴趣集的装置，每个兴趣集是元素集，每个元素代表所述各个用户已经通过与数据处理系统进行交互而表达了兴趣的各个项目；

用于以下的装置：对所述多个用户之中的每一个用户，对于在 1 和 k 之间的每个整数 i，向用户的兴趣集的每个元素应用第 i 个散列函数以获得与各个元素相对应的各个函数值，k 个散列函数彼此不同，并且根据从所述 k 个散列函数获得的函数值来确定所述各自的兴趣集的 k 个散列值，其中所述各自的兴趣集的第 i 个散列值是通过向用户的兴趣集的元素应用第 i 个散列函数而获得的函数值中的最小值，其中 k 是大于或等于 1 的整数；以及

用于将所述多个用户之中的每一个用户分配到为所述各个用户建立的 k 个集群之中的每一个集群的装置，第 i 个集群由所述各个用户的各自的兴趣集的第 i 个散列值所代表，其中所述将多个用户之中的每一个用户分配到 k 个集群的完成不考虑任何其他用户到 k 个集群的分配。

6. 如权利要求 5 所述的系统，还包括：

用于将表达用户兴趣的行为记录在日志中的装置；以及

用于使用所述日志来为所述多个用户生成所述兴趣集的装置。

7. 如权利要求 5 所述的系统，还包括 ：

用于执行协同过滤计算机程序应用，以基于所述多个用户的第一用户到 k 个集群之中的一个或多个集群的分配而向所述第一用户提供信息的装置。

8. 如权利要求 5 所述的系统，还包括 ：

用于为所述多个用户的第一用户获得已改变的兴趣集的装置 ；

用于使用所述已改变的兴趣集，为所述第一用户确定 k 个散列值的装置 ；以及

用于将所述第一用户仅分配到由通过使用所述已改变的兴趣集所确定的 k 个散列值所代表的各个 k 个集群之中的每一个集群，而不改变任何所述其他用户到 k 个集群的分配的装置。

9. 一种用于将用户分配到集群的方法，所述方法包括 ：

为所述用户获得兴趣集，所述兴趣集是元素集，每个元素代表所述用户已经通过与数据处理系统进行交互而表达了兴趣的各个项目 ；

对于在 1 和 k 之间的每个整数 i，向所述兴趣集的每个元素应用第 i 个散列函数以获得与各个元素相对应的各个函数值，k 个散列函数彼此不同，并且根据从所述 k 个散列函数获得的函数值来确定所述兴趣集的 k 个散列值，其中第 i 个散列值是通过向所述用户的兴趣集的元素应用第 i 个散列函数而获得的函数值中的最小值，其中 k 是大于或等于 1 的整数 ；以及

将所述用户分配到 k 个集群之中的每个集群，第 i 个集群由所述第 i 个散列值所代表。

10. 如权利要求 9 所述的方法，其中 ：

所述兴趣集有 m 个元素 ；

所述第 i 个散列值是单向散列函数的 m 个应用的最小值，每一个应用将第 i 个种子值和所述兴趣集的所述 m 个元素中的对应元素进行散列。

11. 如权利要求 9 所述的方法，还包括 ：

使用所述 k 个用户集群来为所述用户执行协同过滤。

12. 如权利要求 9 所述的方法，还包括 ：

将表达用户兴趣的行为记录在日志中 ；以及

使用所述日志来为所述用户生成所述兴趣集。

13. 一种用于将用户分配到集群的系统，所述系统包括 ：

用于为所述用户获得兴趣集的装置，所述兴趣集是元素集，每个元素代表所述用户已经通过与数据处理系统进行交互而表达了兴趣的各个项目 ；

用于以下的装置 ：对于在 1 和 k 之间的每个整数 i，向所述兴趣集的每个元素应用第 i 个散列函数以获得与各个元素相对应的各个函数值，k 个散列函数彼此不同，并且根据从所述 k 个散列函数获得的函数值来确定所述兴趣集的 k 个散列值，其中第 i 个散列值是通过向所述用户的兴趣集的元素应用第 i 个散列函数而获得的函数值中的最小值，其中 k 是大于或等于 1 的整数 ；以及

用于将所述用户分配到 k 个集群之中的每个集群的装置，第 i 个集群由所述第 i 个散列值所代表。

14. 如权利要求 13 所述的系统，其中 ：

所述兴趣集有 m 个元素 ；

所述第 i 个散列值是单向散列函数的 m 个应用的最小值，每一个应用将第 i 个种子值和所述兴趣集的所述 m 个元素中的对应元素进行散列。

15. 如权利要求 13 所述的系统，还包括：

用于使用所述 k 个用户集群来为所述用户执行协同过滤的装置。

16. 如权利要求 13 所述的系统，还包括：

用于将表达用户兴趣的行为记录在日志中的装置；以及

用于使用所述日志来为所述用户生成所述兴趣集的装置。

用于将用户分配到集群的方法和系统

[0001] 本申请是国际申请日为 2006 年 8 月 15 日、国际申请号为 PCT/US2006/031868 的 PCT 国际申请的、进入中国国家阶段的国家申请号为 200680038100.7、题为“基于集的相似性的可扩展用户聚类”的专利申请的分案申请。

技术领域

[0002] 本发明涉及数字数据处理，并且尤其涉及将计算机应用或系统的用户分组为集群 (cluster)。

背景技术

[0003] 将用户分组为集群的操作是出于多种目的。为了实现用户的个性化，例如一种众所周知的技术，即协同过滤 (collaborative filtering)，涉及将用户进行聚类 (clustering) 并把在用户集群中的其它用户已经表达过兴趣的项目推荐给用户。一般可以认为用户以多种方式表达对项目的兴趣，例如，通过点击项目、购买项目、或将项目添加到购物车。推荐可采用多种方式，例如以部分搜索结果的形式呈现给用户，以用户可能想要阅读的新闻故事的形式进行展现，对用户可能想要购买的项目进行确定等等。

[0004] 一种实现用户聚类的方法是先定义两个用户之间的距离度量 (distance measure)，然后使用众所周知的诸如 k- 均值或分层合并聚类 (HAC) 的聚类算法将用户进行聚类。然而，这些技术有缺点。例如，HAC 的运行时间为 $O(n^2)$，对于数以亿计的 n 值是难以实现的；而 k- 均值算法需要代表数据点的均值，当数据点是集的时候，这是不可行的。

发明内容

[0005] 在特定实施方式中，本发明可提供可扩展的用户聚类，其中每个用户都以代表取自全体项目之中的项目的元素集的形式来表示。

[0006] 例如，当给定用户可以通过与计算机系统进行交互而选择的全体项目时，每个用户可以通过不同的行为（例如点击项目，购买项目，将项目添加到购物列表、查看项目等）来表达它们对项目的各个子集的兴趣。本发明的特定实施方式以此种方式将用户进行聚类（即将用户分配到集群），也就是在相同集群之中的用户可能在它们各自的项目子集之间具有高度的重叠。

[0007] 一方面，符合本发明实施方式的计算机程序产品可使得数据处理装置为多个用户之中的每一个用户获得各自的兴趣集，每个兴趣集表示在其中各个用户已通过与数据处理系统进行交互而表达了兴趣的项目；对多个用户之中的每一个用户，确定各个兴趣集的 k 个散列值 (hash value)，其中第 i 个散列值是在对应的第 i 个散列函数之下的各个兴趣集之中的最小值，其中 i 是在 1 和 k 之间的整数，并且其中 k 是大于或等于 1 的整数；并且将多个用户之中的每一个用户分配到为各个用户所建立的各个 k 个集群中的每一个集群，第 i 个集群由第 i 个散列值所代表，其中将多个用户之中的每一个用户分配到 k 个集群的完成不考虑任何其它用户到 k 个集群的分配。

[0008] 有利的实施方式可包括一个或多个下述特征。本产品可使得数据处理装置将表达用户兴趣的行为记录在日志中；并且使用该日志为多个用户生成兴趣集。

[0009] 本产品可使得数据处理装置为多个用户之中的第一个用户获得已改变的兴趣集；使用已改变的兴趣集为第一用户确定 k 个散列值；并且将第一用户仅分配到由使用已改变的兴趣集所确定的 k 个散列值所代表的各个 k 个集群之中的每一个集群，而不改变任何其它多个用户到集群的分配。

[0010] 在另一个方面，符合本发明实施方式的计算机程序产品可使得数据处理装置为用户获得兴趣集，兴趣集代表在其中用户已经通过与数据处理系统进行交互而表达了兴趣的项目；确定兴趣集的 k 个散列值，其中第 i 个散列值是在对应的第 i 个散列函数之下的兴趣集之中的最小值，其中 i 是在 1 和 k 之间的整数，并且其中 k 是大于或等于 1 的整数；并且将用户分配到 k 个集群中的每一个集群，第 i 个集群由第 i 个散列值所代表。

[0011] 有益的实施方式可包括一个或多个下述特征。兴趣集有 m 个元素；第 i 个散列值是单向散列函数的 m 个应用的最小值，每一个 m 应用将第 i 个种子值和兴趣集之中的 m 个元素的各个元素进行散列。产品可使得数据处理装置来使用 k 个用户集群来为用户完成协同过滤。

[0012] 另一方面，符合本发明实施例的系统包括：由多个用户使用数据处理系统所选择的项目的日志；用于使用指纹函数和项目的日志来将多个用户的每一个用户分配到 k 个集群的装置，其中 k 是大于或等于 1 的整数；并且基于第一用户到一个或多个 k 个集群的分配，可运行协同过滤计算机程序应用来将信息提供给多个用户的第一用户。

[0013] 有益的实施方式可以包括一个或多个下述特征。信息包括推荐、预计、或排名之中的至少一种。

[0014] 另一方面，符合本发明实施例的计算机程序产品可使得数据处理装置来使用 k 个元素的已排序的集来确定数据处理系统的用户，其中 k 是大于 1 的整数，其中 k 个元素中的每一个元素对应于在兴趣集之中的元素，每个在兴趣集之中的元素代表在其中用户已经通过其使用数据处理系统进行的行为表达了兴趣的项目。

[0015] 有益的实施方式可以包括一个或多个下述特征。在为用户执行的协同过滤中，产品可使得数据处理装置使用已排序的 k 个元素来确定用户。协同过滤包括将项目推荐给用户或为用户将项目进行排名。产品可使得数据处理装置从用户处接收输入，响应于输入的内容，数据处理系统将元素从兴趣集之中去除以生成修改后的兴趣集；确定已修改顺序的 k 个元素集，其中 k 个元素中的每一个元素对应于在已修改兴趣集之中的元素；并且使用已修改顺序的 k 个元素集而不是初始顺序的 k 个元素集来确定用户。k 个元素的已排序的集将用户确定为属于每一个 k 用户集群。产品可使得数据处理装置将表达用户兴趣的行为记录在日志中；并且使用日志为用户生成兴趣集。数据处理系统包括网站；并且用户的兴趣集包括以下代表：用户已经在网页中点击的一个或多个项目、用户已经从在线零售商购买的项目、或用户已经添加到购物车中的项目。由用户表达对项目感兴趣的行为包括隐含地表达兴趣的行为。由用户表达对项目感兴趣的行为包括清楚地表达兴趣的行为。用户是通过用户登录来确定的个体。用户是通过 cookie 来确定的个体。用户是具有共同的观察到的属性的一个或多个个体，其中属性是由一个或多个个体之中的每一个个体公开给数据处理系统的属性。用户是个体与数据处理系统进行交互的会话。在兴趣集之中的每一个元素

是用户在与数据处理系统进行交互中选定的项目。

[0016] 又一方面，本发明的实施方式可包括对应于上述程序和系统的方法，和对应于上述系统的程序。

[0017] 可实施本发明以实现一个或多个以下优点。聚类计算是可扩展的。可对由数亿个体用户所使用的应用执行计算，其中个体用户可具有在其兴趣集之中代表的数十个、数百个、或更多项目。可在由全体项目的子集代表被聚类的实体的情形下，执行聚类。不需要对全体进行预先定义。聚类是基于集的相似性度量。新用户的聚类的产生不改变任何现有的聚类。一个用户的聚类的产生不考虑其它用户已经聚类或正在聚类的方式。然而，一些全局值，例如种子值或排列 (permutation) 可以在聚类中共享。实际上，通过改变其选择——例如对项目的选择进行删除或添加，在之后计算或重新计算集群时，用户可以改变其已分配的集群。不使用取自其它用户的数据就可对新用户或具有已修改兴趣集的用户的集群的全体用户进行计算。聚类计算不限于对那些是个体的用户进行聚类。例如，不论每个用户是个体、每个用户是个体的集合、每个用户是与系统的交互、或是它们的某种组合，都可有效地完成聚类。

[0018] 在附图和以下说明中阐明本发明的一个或多个实施例的细节。从说明书、附图和权利要求书，本发明的其它特征和优点将是显而易见的。

附图说明

[0019] 图 1 是示出符合本发明实施例的用于将用户进行聚类的第一方法的流程图。

[0020] 图 2 是示出符合本发明实施例的用于将用户进行聚类的第二方法的流程图。

[0021] 图 3 是示出符合本发明实施例的使用用户集群的推荐系统的运行的流程图。

[0022] 图 4 是示出符合本发明一个实施例的具有新推荐引擎的新闻服务的示意图。

[0023] 在不同附图中的相同标记数字和名称指示相同的元素。

具体实施方式

[0024] 图 1 示出以下用于将用户进行聚类的最小散列 (minhash) 方法的逻辑说明。虽然此方法可实施，但是将其呈现于此主要是出于解释的目的。以下将参照图 2 描述用于在具有大量用户的系统中将用户进行聚类的实际实施方式。

[0025] 如图 1 所示，最小散列方法的输入是 :全体项目 110，标记为 U ;k 个排列的集 112，标记为 p1，p2，...，pk ;以及用户的兴趣集 114，对用户 A 标记为 X_A。

[0026] 排列是 U 范围内的排列，将其均匀地从 U 范围内的所有排列的集之中挑选出来，以使得每个排列被挑选的概率与其它排列相同。排列是 U 对 U 的每一个一对一的映射（双向单射）。只有 U 是固定并可数的，才能实现这种排列。整数 k 是选择参数。通常 k 的值是在 5 到 10 的范围内。然而，其可以是 1 或更大的任何整数。本方法给用户分配 k 个集群，标记为 C1，...，Ck。在排列被选定并被用于将用户分配到集群之后，如果排列变化，所有的聚类必须重新计算。

[0027] 兴趣集是代表取自全体 U 的项目的元素集。对于现在描述的使用来说，其中元素是项目本身，兴趣集是由用户 (X_A) 对全体 U 之中的项目所做的选择的集。可以如上所述来选择它们。在本说明书中出于便利，术语“项目”既可以指兴趣集之中的元素也可以指用

户的实际选择，其意义根据上下文是清楚的。

[0028] 使用此数据为用户确定了 k 个散列值（步骤 120），每个排列对应一个散列值。对排列 pi，散列值标记为 hi(X_A)。排列 pi 的散列值是在排列 pi 之下取自 X_A 的最小元素，即最小散列值。可以从元素的值或者从 U 的顺序来确定最小值。

[0029] 每个最小散列值用作集群的识别符，并且把用户分配到每个集群。用户将属于 k 个集群，第 i 个集群由第 i 个最小散列值所确定。所以，对给定的排列 pi，如果并仅如果在该排列之下的兴趣集的最小散列值相同，两个用户属于相同的集群。

[0030] 这种为每个数据元素关联一个散列值的最小散列技术是一种分类技术，被称为局部敏感 (locality sensitive) 散列技术，其拥有希望的属性：即两个数据元素有一定的概率具有相同的散列值，这个概率与两个数据元素之间相似性成正比。在本例中，如果两个用户 A、B（由其兴趣集 X_A 和 X_B 所代表）之间的相似性定义为用 (X_A 交 X_B) 的大小除以 (X_A 并 X_B) 的大小。那么，最小散列技术具有这样的属性：两个用户 A 和 B 的最小散列值相同的概率（在从中挑选实际使用的排列的排列集的范围内定义）与以上定义的相似性度量相等。由此，最小散列实现了概率聚类 (probabilistic clustering)，其中用户落在相同集群之中的概率与他们的相似性相等。

[0031] 因为确定了 k 个集群（步骤 122），如果两个用户在相同集群中的概率为 p(0 ≤ p ≤ 1)，那么即使在多次聚类中的一次中没有把他们聚类在一起，也会在多次聚类中的 p 次中把他们聚类在一起。这产生了平滑效应，使得每个用户均匀地属于 k 个不同集群并且在每次聚类中使每个用户与其他相似的用户进行聚类。应当挑选参数 k 以最优化在效率（较低的 k 产生更好效率）和质量（较高的 k 产生更好质量）之间的平衡。尽管不是严格必须如此，数字 k 通常是常数；并且象 10 这样的小值可以提供好的效果。

[0032] 最小散列聚类方法很有扩展性并且具有其它数个优点。例如，方法的运行时间与数据的大小（即，（用户，项目）对的总数）呈线性。

[0033] 并且，每个用户都独立聚类，即独立于所有其它用户。这在用户总是被添加、删除和更新的网络领域中尤其有意义。由此而带来的优点是可以轻松和增量地处理数个事例，这对于传统的聚类算法是困难的。如果将用户确定为垃圾信息，即出于影响使用聚类的系统的目的而表达虚假的兴趣，可以删除该用户而不影响任何其它用户，即，其它聚类不发生变化。并且，如果曾经对选择保密的用户决定公开她的选择，或者如果新用户被添加到系统，可将其添加到集群而不对其它用户重新聚类。最后，如果用户决定通过有效地编辑兴趣集来改变其简档 (profile)，那么可实时更新对该用户的聚类（与此相对的是通过批处理更新），还要考虑这点，即不影响任何其它用户的聚类。

[0034] 图 2 示出对在有大量用户（达到数以亿计、并且在每个用户的兴趣集之中可能有数百或更多的项目、在实际上或实践中不可计数的全体项目范围内）的系统之中的用户进行聚类的实际实施方式，本实施方式使用随后将描述的映射化简 (MapReduce) 编程模型和技术。

[0035] 本实施方式的输入是：标记为 D 的未以特定顺序存储的数据元素（诸如结果点击日志、购买日志等）的集合 210；标记为 s1，s2，...，sk 的 k 个种子值的已排序的集 212；和指纹函数 214。每个数据元素可认为是指示特定用户已经表达对特定项目感兴趣的对（用户，项目）。选择性地，可将后缀添加到项目的根形式以指示数据元素是指第一次、还是第二

次等用户表达兴趣的情形，以捕捉用户已这样做的频率。有益地是，项目的形式是文本串，使得项目可以通过任何网络应用（即通过任何使用网络浏览器来将用户界面呈现给用户的应用）容易地代表表达兴趣的任何用户行为。

[0036] 例如，在用户以回答在线问卷调查的形式将表示用户兴趣的信息提供给系统时，用户表达兴趣的行为可以被表达；或者用户表达兴趣的行为可以是隐含地，例如用户在新闻网站上选择新闻故事阅读时。

[0037] 将 k 个种子值 s1，s2，...，sk 看作是经挑选而随机出现的位串的数字，例如，这样二进制位表现为均匀的“0”或“1”。

[0038] 指纹函数将种子值和取自兴趣集的项目映射到一个大数，例如 64 位或 128 位的数字。

[0039] 在一个实施方式中，通过使用 unix 随机函数生成 k 个 32 位整数值来生成种子值。可能必须不止一次地调用随机函数来生成单个种子。在本实施方式中，指纹函数执行 MD5 单向散列算法，并且将与项目（其通常是字符串或二进制数）连接在一起的种子值进行散列以产生 128 位的值。

[0040] 参见图 1 所述，种子值和指纹函数逻辑上对应于 k 个排列 p1，...pk，并且在不需要全体项目是可数的情形下，提供项目的排序和排列。

[0041] 使用映射化简框架来处理集合 D，以下将对其进行描述。

[0042] 在映射阶段 220 中，对每个（用户，项目）对，以分布的形式将（键，值）对进行输出，其中键＝用户和值＝项目。

[0043] 在化简阶段 222，将所有这类有相同键（用户）的（键，值）对进行集合，并以分布的方式将其呈现给化简例程，该化简例程对每个不同的键（用户）值运行一次。

[0044] 化简例程（对特定用户）对用户的兴趣集之中的所有项目进行处理；对本说明书来说，将这些 m 个项目标记为 i1，i2，...，im。对每个种子值 si，化简例程计算 m 个值作为项目的指纹（每个项目一个）以及种子值，即指纹 (si，i1)。在 m 个项目范围内，计算这些指纹的最小值并且该最小值成为第 i 个最小散列值，对应于第 i 个种子 si。

[0045] 由此计算出 k 个最小散列值来代表用户。这些代表了用户属于的 k 个集群，并且认为用户被分配到这些集群。

[0046] 如图 3 所示，推荐计算机程序应用可以使用根据任何在此描述的方法所生成的用户集群。

[0047] 在一个实施方式中，系统将用户所做的选择记录在日志中（步骤 310）。可用任何方式存储日志，例如用非结构化文本的行或用结构化数据库之中的记录；并且可将其存储在任何计算机可读介质上，例如在文件服务器的磁盘驱动器上。系统可以是提供搜索结果、广告、购买选择、连接到网站之内或之外的网页的简单链接，或其他项目的网站。记入日志的选择可以是（但是不必须是）所有由系统的用户所做的选择。例如，应用可能仅对新闻网站而不是所有网站的选择或者仅对购买的项目而不是所有查看项目的选择感兴趣，另外，系统可以为不同的推荐应用维护不同选择类型的多个日志，该推荐应用可以计算其特有的各个用户聚类。例如，在使用种子和指纹函数的方法中，每个独立的聚类可以有其所有的独特的种子序列和指纹函数。

[0048] 通过用户的注册或登录、通过 cookie、或其它，系统可将个体确定为用户。选择性

地，如果不希望在与系统交互的多个会话之中维护有关个体用户的信息，出于聚类的目的，系统可以将用户会话视为用户。也可使用 cookie 来维护会话。cookie 是由服务器发送到网络浏览器，并且之后浏览器每次访问该服务器时，由浏览器发送回去的信息包。选择性地，系统可允许个体来确定其是否要参加日志记录，即将自身包括进或将自身排除出他们的选择的日志记录。

[0049] 选择性地，系统可将个体与系统进行交互的某种属性或者属性的组合视为是用户。属性可以由系统进行观察，例如正使用的 IP（因特网协议）地址或正使用的语言，或其可以是由个体提供的信息，例如居住的城市或国家，或对由系统提供服务的订购。所以，例如，系统可以将来自 Cupertino 的个体视为是一个用户，并将来自 Redmond 的个体视为不同的用户。这种集合性聚类的优点是其允许系统在不需要登录或注册的情形下，提供一定程度的个性化。另外，系统可选择性地为均在相同集群之中的所有类型的用户（例如个体或群体）进行聚类，或者可以为不同类型的用户建立不同的集群。

[0050] 由系统的用户进行的选择可以是简单选择或者是复合选择。复合选择是选择的序列，例如从到第一网页，接着直接到第二网页浏览的序列。网页是由网络服务器提供给网络浏览器的资源，典型地是 HTML（超文本标识语言）文档。网络服务器是接受通常经由网络接收的 HTTP（超文本传送协议）请求，并将 HTTP 响应提供给请求者的计算机程序。HTTP 响应通常由 HTML 文档，但也可以是文本文件、图像、或其它类型的文档所组成。

[0051] 如本说明书其它地方所描述的那样，基于已计入日志的选择，把每个用户分配到 k 个集群（步骤 312）。当新用户出现在系统中时以及当选择添加到日志或从日志中去除时，可更新该用户的聚类。选择性地，在某些情形下，并不是把所有的用户都分配到 k 个集群。在这类情形下，可获得一个或多个（但是少于 k 个）集群识别符以便为特定用户找到推荐。例如，如果系统接收到为具有选择集的新用户提供推荐的请求，系统可使用该选择来选择性地计算第一集群的确定 (identity)，使用其来找到推荐，继续并相似地计算并使用第二集群，以此类推，直到找到系统定义的足够数目的推荐。

[0052] 然后，推荐应用可使用用户集群来对特定用户做出推荐（步骤 314）。可以把任何基于将每个用户分组为单个集群进行推荐的方法用于此处描述的多个集群。例如，可应用这种方法 k 次并且合并 k 个结果来为用户提供推荐项目的合并集。或者，可以使用在其中项目出现的不同的结果的数目来对项目进行排名。或者，可将一些取自每个基于集群的推荐结果的项目提供给用户，以给用户多种推荐。分配到用户的多个集群可反映当使用系统时用户具有的不同类型的兴趣，并且于是相比于如果仅使用单个集群，给用户提供这种多种类型的推荐使得推荐更可能包括用户当前兴趣的内容。

[0053] 推荐应用是协同过滤的一个实例，并且在本说明书中描述的用户聚类的方法也可应用到其它类型的协同过滤。在协同过滤中，找到与当前用户相类似的用户，并且从其偏好或行为，为当前用户做出排名、推荐或预计。通过将用户分组为多个集群，系统通过对用户的分组隐含地确定出用户的偏好并对项目进行分组。

[0054] 如图 4 图解所示，在本说明书中描述的将用户分配到集群的技术，可以在新闻推荐引擎 410 中实施，该新闻引擎基于由那些用户先前对文章所作出的选择，可以提供对要呈现给用户 402a、402b 的新闻文章的推荐。用户 402a、402b 通过他们各自的浏览器，经由诸如本地、广域、或虚拟个人网络、或因特网的数据通信网络 404，与一个或多个网络服务器

430 进行通信。新闻服务 420 以托管在一个或多个服务器 430 上的计算机程序的形式来实施，并将网页提供给用户 402a、402b 以对用户的请求进行响应。在由新闻服务 420 提供的页面之中，是用户可以从中选择一篇或多篇新闻文章用于由用户的浏览器显示的网页。响应于用户的选择，新闻服务 420 把选定的文章提供给用户（功能 424）。如果新闻推荐引擎 410 已经为特定用户提供了推荐，那么根据对该用户的推荐，新闻服务可以提供显示为该用户选择的文章的网页（功能 422）。

[0055] 新闻推荐引擎 410 以运行在一个或多个服务器 430 之上的计算机程序来实施。新闻推荐引擎 410 从新闻服务 420 的用户接收选择，并且把这些选择记入日志 440 中（功能 412）。如本发明书其它地方所描述的那样，通过使用在日志 440 中的信息，引擎把用户分配到集群（功能 414）。对任何已经分配到集群的特定用户，引擎基于用户已经分配到的集群来确定推荐（功能 416），并且将这些推荐提供给新闻服务 420。

[0056] 在确定向特定用户做何种推荐时，引擎考虑已经分配到与特定用户相同的一个或多个集群的其它用户所做的选择。引擎可以选择性地将用户已经选定的新闻文章从可能的推荐之中删除。引擎或服务可基于多种标准来对推荐进行排名，包括：由分配到该用户已经分配到的集群的其它用户选择该新闻文章的次数，新闻文章有多新，具有文章是关于讨论中新闻文章的主题的来源的数目等。用这种方法，新闻服务可以为客户提供新闻文章的个性化呈现和排名。

[0057] 在一个实施方式中，新闻推荐引擎 410 把用户确定为个体，并且要求用户登录并注册以得到个性化推荐。如本说明书其它地方所描述的那样，在其它实施方式中，可隐含地确定用户，或将其确定为集合性分组。

[0058] 可依此方法实施推荐引擎以支持其它类型服务的个性化，例如提供图像、博客、或购物信息的选择的服务。

[0059] 尽管引擎和服务的功能在图 4 中是以独立模块的形式示出，其实并不必须以这种方式实施；尤其是，可用一部分服务的实施的形式来实施引擎。

[0060] 以下部分描述了映射化简编程模型和用于处理和生成大数据集的模型的实施方式。该模型及其库的实施方式都称为映射化简。使用映射化简，程序员对处理键 / 值对来生成中间键 / 值对的映射函数，以及把所有与相同中间键相关联的中间值都进行合并的化简函数进行规定。按这种函数形式编写的程序可自动并行化并在普通计算机的大集群上执行。可实施运行时系统或框架来：划分输入数据、调度在机器集之中程序的执行、处理机器故障、并且管理必需的机器间通信。

[0061] 映射化简计算接受输入键 / 值对集，并产生输出键 / 值对集。用户把计算表达为两个函数：映射和化简。

[0062] 映射，由用户所编写，接受输入键 / 值对并产生中间键 / 值对集。映射化简库把所有与相同中间键 I 相关联的中间值分组在一起并把它们传递给化简函数。

[0063] 化简函数，也是由用户所编写，接收中间键 I 和此键的值的集。其将这些值合并在一起以形成的值的可能的更小集。通常，对每个化简调用，只产生 0 或 1 输出值。经由迭代器 (iterator)，把中间值提供给用户化简函数。用这种方法，可以处理太大不能放入存储器的值的列表。

[0064] 考虑对在文档的大集合之中出现的每个词的计数问题。用户可以编写类似于下列

伪代码的代码：

[0065]

```
map(String key, String value):
  // key:文档名
  // value:文档内容
for each word w in value:
  EmitIntermediate (w,"1");
reduce (String key, Iterator values):
  // key:词
  // values:计数列表
  int result=0;
  for each v in values:
  result+=ParseInt(v);
  Emit(AsString(result));
```

[0066] 映射函数的发出是每个词加上相关联出现的计数（在本简单实例中正好为“1”）。化简函数把所有对特定词发出的计数进行加总。

[0067] 在一个实施方式中，为了完成计算，用户编写代码以用输入和输出文件名和可选调整参数来填充进规定对象。接着用户调用映射化简函数，将其传递给规定对象。用户的代码与映射化简库链接在一起。

[0068] 尽管前述伪代码按照字符串输入和输出的形式编写，但是从概念上来说，由用户提供的映射和化简函数有关联的类型：

[0069] map(k1，v1) → list(k2，v2)

[0070] reduce(k2，list(v2)) → list(v2)

[0071] 也就是，输入键和值取自与输出键和值不同的域。另外，中间键和值取自与输出键和值相同的域。

[0072] 映射化简模型的很多不同的实施方式是可能的。

[0073] 下面部分描述的实施方式定位于具有用交换以太网连接在一起的普通个人计算机的大集群的计算环境。在该环境中，机器典型地是每个机器具有 2-4GB（十亿字节）的存储器，集群具有数百或数千台机器，存储器由直接连到个人机器的廉价 IDE（集成驱动器电子标准）磁盘所提供，使用分布式文件系统来管理存储在这些磁盘（其使用副本在不可靠的硬件之上提供可用性和可靠性）之上的数据，并且用户将作业递交给调度系统。每个作业由任务集所组成，并且由调度系统的调度器将其映射到集群内的可用机器的集。

[0074] 通过自动地把输入数据划分进 M 个分隔的集，将映射调用分布在多个机器之中。不同的机器可以对输入分隔进行并行处理。通过使用划分函数（例如，hash(key)mod R）把中间键空间划分为 R 块，将化简调用分布化。由用户来规定划分 (R) 的数目和划分函数。

[0075] 当用户程序调用映射化简函数时，发生下列动作序列：

[0076] 1. 用户程序的映射化简库首先把输入文件分隔为通常每块是16兆字节到64兆字节 (MB)(用户可控制)的M个块。接着其启动在机器的集群上的程序的多个复本。

[0077] 2. 程序的多个复本中的一个是主控程序 (master)。其它是由主控程序分配工作的工作程序 (worker)。要分配M个映射任务和R个化简任务。主控程序挑选空闲的工作程序并把映射任务或化简任务分配给每个工作程序。

[0078] 3. 已分配映射任务的工作程序读取相应的输入分隔中的内容。其从输入数据中分析出键/值对,并把每个对传递给用户定义的映射函数。由映射函数产生的中间键/值对在存储器中进行缓冲。

[0079] 4. 周期性地,将已缓冲的对写入本地磁盘,通过划分函数将其划分进R个区。这些已缓冲的对在本地磁盘上的位置信息被返回给主控程序,其负责将这些位置信息转发给化简工作程序。

[0080] 5. 当主控程序将这些位置信息通知给化简工作程序时,化简工作程序使用远端过程调用来从映射工作程序的本地磁盘读取已缓冲的数据。当化简工作程序读取了所有中间数据,其通过中间键来对中间数据进行排序以使得所有发生相同键的都被分组在一起。这种排序是有用处的,因为通常很多不同的键映射到相同的化简任务。如果中间数据的量太大以至于不能放入存储器,那么使用外部排序。

[0081] 6. 化简工作程序遍历已排序的中间数据,并且每遇到一个唯一的中间值,其将键和相应的中间值的集传递给用户的化简函数。把化简函数的输出附加到该化简划分的最后的输出文件之中。

[0082] 7. 当完成了所有的映射任务和化简任务,主控程序唤醒用户程序。在该点上,在用户程序中调用的映射化简返回到用户代码。

[0083] 成功完成之后,执行的输出在R个输出文件中可用(每个化简任务一个,具有的文件名与由用户规定的一样)。用户不需要将这些R个输出文件组合为一个文件;其可以把这些文件作为输入传递到其它映射化简调用,或者从其它可对划分为多个文件的输入进行处理的分布式应用中使用这些文件。

[0084] 主控程序具有数种数据结构。对每个映射任务和化简任务,其存储状态(空闲、进行中、或完成)和工作程序机器(用于非空闲任务)的标志。

[0085] 主控程序是管道,通过它把中间文件区的位置信息从映射任务传递到化简任务。所以,对每个已完成的映射任务,主控器将由映射任务产生的R个中间文件区的位置信息和大小进行存储。当映射任务完成时,接收对这些位置信息和大小信息的更新。将信息增量地推给具有进行中的化简任务的工作程序。

[0086] 因为本实施映射化简库是被设计用来使用数百或数千台机器处理极大量的数据信息,所以库适度地容忍机器的错误。

[0087] 主控程序周期地查验 (ping) 每个工作程序。如果在特定时间内没有从工作程序接收到响应,主控程序可以将该工作程序标记为故障。任何由该工作程序完成的映射任务都复位回到其初始空闲状态,并因此变得可在其它工作程序上进行调度。类似地,任何在故障工作程序上进行的映射任务或化简任务也复位到空闲,并变得可重新调度。

[0088] 要重新执行在故障工作程序上的已完成映射任务,因为其输出存储在故障机器的本地磁盘上并因此不可访问。不需要重新执行已完成的化简任务,因为其输出存储在全局

文件系统中。

[0089] 当映射任务首先由工作程序 A 执行，然后接着由工作程序 B 执行（因为 A 出现故障）时，把重新执行通知给所有执行化简任务的工作程序。任何还未从工作程序 A 读取数据的化简任务将从工作程序 B 读取数据。

[0090] 因为只有单个主控程序，所以其发生故障是糟糕的；所以如果主控程序失效，则中止映射化简计算。用户或用户程序可以对这种情形进行检查并且如果其希望可以重试映射化简操作。

[0091] 当用户（提供映射和化简的操作者）是其输入值的确定的函数时，本分布式实施方式产生的输出与可能由整个程序的无故障顺序执行所产生的输出相同。每个进行中的任务将其输出写入专用临时文件。当映射任务完成时，工作程序将消息发送给主控程序并将 R 个临时文件的名字包括在消息中。如果对已完成的映射任务，主控程序接收到完成消息，其忽略此消息。不然，其将 R 个文件的名字记入主控程序数据结构中。当化简任务完成时，化简工作程序将其临时输出文件重命名为最后输出文件。如果在多个机器上运行相同的化简任务，那么对相同的最后输出文件将执行多个重命名调用。由底层文件系统提供的原子重命名操作保证了最后的文件系统状态恰包含由化简任务的一个执行产生的数据。

[0092] 利用输入数据存储在组成集群的机器的本地磁盘上的事实，实施方式节约了网络带宽。文件系统将每个文件分割成 64MB 的块，并将每个块的复本存储在不同的机器上。映射化简主控程序考虑到输入文件的位置信息，并试图对包含对应输入信息数据副本的机器之上的映射任务进行调度。如果此操作失败，其试图对该任务输入数据副本附近的映射任务进行调度（例如，在与包含数据的机器相同的网络交换器上的工作程序机器）。

[0093] 为了动态负载平衡，M 和 R 应当比工作程序机器的数目要大很多。在本实施方式中，对 M 和 R 的大小有实际限制，因为如上所述，主控器必须做出 O(M+R) 个调度决定并在存储器中保持 O(MxR) 个状态。另外，R 经常受限于用户，因为每个化简任务的输出结束于独立的输出文件。实践中，应当挑选 M 以使得每个单个的任务具有大约 16MB 到 64MB 的输入数据，以使得上述的局部优化更有效，并且 R 应当是期望使用的工作程序机器数量的小的倍数。

[0094] 落后者（花费超出常理的时间来完成计算中的最后几个映射或化简任务之中的一个的机器）可对映射化简操作花费的总时间造成负面影响。为了减轻落后者的问题，当映射化简操作接近完成时，主控程序对剩余正在处理任务的备份执行进行调度。不论是主要还是备份执行完成时，将任务标记为完成。

[0095] 除了上述的基本功能以外，本实施方式提供了下述有益的扩展。

[0096] 在一些情形中，通过键的某种特定函数来划分数据是有用处的。为了支持该点，映射化简库的用户可以提供划分函数。

[0097] 本实施方式保证在给定的划分内，以增加的键顺序来处理中间键 / 值对。这使得对每个划分生成一个排序后的输出文件变得容易，这对于当输出文件格式需要支持用键的有效随机查询或者输出的用户发觉将数据排序是便利的时候是有用处的。

[0098] 在一些情形中，由每个映射任务产生的中间键具有很高的重复率，并且用户规定的化简函数是可变换和可结合的。一个这种实例是上述的词计数实例。每个映射任务可产生数百或数千种形式为＜ the，1＞的记录。所有这些计数将经由网络被发送给单独的化简

任务并通过化简函数相加在一起以产生一个数字。为了提供这类情形，本实施方式在将数据发送在网络上之前，允许用户对执行数据的部分合并的可选组合函数进行规定。

[0099] 组合函数在每个执行映射任务的机器上执行。可用相同的代码来实施组合和化简函数。化简函数和组合函数唯一的不同是映射化简库处理函数的输出的方式。把化简函数的输出写入最后的输出文件。把组合函数的输出写入将要发送给化简任务的中间文件。

[0100] 映射化简库对读入数种不同格式的输入数据提供支持。例如，“文本”模式输入将每行视为键 / 值对 :键是在文件中的偏移量，而值是该行的内容。另一种通常支持的格式将按键排序的键 / 值对序列进行存储。输入类型的每一种实施方式都知道如何将该类型的数据分隔为有意义的范围用于以独立的映射任务的形式来处理（例如，文本模式的范围分隔保证范围分隔只发生行边界）。通过提供简单读者界面的实施方式，用户可以增加对新输入类型的支持。另外，读者不限于提供从文件读取的数据。例如，用户可以从数据库或者从映射进存储器的数据结构中读取记录。

[0101] 以类似的方式，实施方式支持输出类型的集以用于产生不同格式的数据，并且对用户代码来说增加对新输出类型的支持也是容易的。

[0102] 有时，在用户或第三方代码中的错误 (bug) 使得映射或化简函数在特定记录上确定地发生崩溃。有时忽略一些记录是可以接受的，例如，当在大数据集上做统计分析时。实施方式提供执行的可选模式，其中映射化简库侦测引起确定性崩溃的记录并且跳过这些记录以继续运行。

[0103] 对该模式，每个工作程序进程安装有捕捉分段冲突和总线错误的信号处理程序。在调用用户映射或化简操作之前，映射化简库将自变量的序号存储进全局变量。如果用户代码生成一个信号，信号处理程序将包含序号的“临终”(last gasp)UDP(用户数据报协议)包发送给映射化简主控程序。当主控程序观察到在特定记录上有超过一个以上的故障时，当其发布相应映射或化简任务的下一次重新执行时，其指明应当跳过该记录。

[0104] 关于映射化简的更多信息可以在 J.Dean 和 S.Ghemawat 于 2004 年 12 月 6 日在“操作系统设计和实施的第六次讨论会”的大会论文集第 137-150 页的《映射化简 :大集群上的简单数据处理》(Simplified Data Processing on Large Clusters，Proceedings of the 6th Symposium on Operating Systems Design and Implementation) 中找到，将该文章以引用的方式包括在此。

[0105] 以下将简单描述另一种使用局部敏感聚类方案将用户聚类进多个集群的方法。在该方法中，每个用户有一个由表征用户的高维向量所表示的简档。挑选运行在这类向量上的 k 个散列函数的集。用户简档的第 i 个散列值代表用户被分配到的第 i 个集群。在 Charika 在 2002 年 5 月 19-21 日于加拿大魁北克蒙特利尔的“计算原理的第 34 届 ACM 讨论会”上的《来自化整算法的相似性估计计数》(Similarity Estimation Techniques from Rounding Algorithms，34th ACM Symposium on Theory of Computing) 中描述了对本方法有用处的局部敏感散列函数。

[0106] 在这种方法的一个实施方式中，用户由 < 项，权重 > 对的列表所代表。如前所述，k 是为用户计算的集群的数目和散列值的数目。为说明，将种子值的数目指定为 8k，尽管指定为常数 8 的数通常为参数。8k 个随机种子值表示为标记为 s_1，s_2，...s_8k 的字符串，并且对其进行挑选以随机出现，例如在二进制的表示中的位是均匀的“0”或“1”。对每个用

户，用以下方式计算第 i 个散列值：

[0107]

```
for b from 1 to 8:
do
   initialize sum=0;
   for all <term_j, weight_j> pairs in the user's list:
do
   if(fingerprint(term_j+s_((i-l)*8+b))has least significant bit=1)
      sum=sum+weight_j
   else
      sum=sum-weight_j
  done
  if(sum > 0)
     b-th bit of i-th hash value is set to 1
  else
     b-th bit of i-th hash value is set to 0.
  done.
```

[0108] 项 fingerprint(term_j+s_((i-1)*8+b) 代表与种子串 s_((i-1)*8+b)，即第((i-1)*8+b) 个种子串相连接的第 j 个项 (term_j) 的指纹函数（如上所述进行计算）。

[0109] 在本说明书中描述的本发明的实施例和所有功能操作，可以用包括本说明书公开的结构及其结构对等物的数字电子电路、或计算机软件、固件或硬件、或这些的组合来实施。可以以一个或多个计算机产品的形式来实施本发明的实施例，即编码在诸如机器可读存储设备、机器可读存储介质、存储器装置的计算机可读介质上，或机器可读传导信号上的一个或多个计算机程序指令的模块，其由数据处理装置来执行或者控制数据处理装置的运行。术语“数据处理装置”包括所有用于处理数据的装置、设备、和机器，包括举例来说，可编程处理器、计算机或多个处理器或计算机。装置可包括，除硬件以外，创建用于讨论中的计算机程序的执行环境的代码，例如，组成处理器固件、协议栈、数据库管理系统、操作系统或它们的组合的代码。传导信号是人工生成的信号，例如机器生成的电、光或电磁信号，生成该信号用来将传输的信息编码为适合于接收器装置。

[0110] 计算机程序（也称为程序、软件、软件应用、脚本、或代码）可以用包括汇编或解释性语言的任何编程语言编写，并且可以用任何方式进行布署，包括用单机程序或用模块、组件、子程序、或其它适于在计算环境中使用的单元。计算机程序并不必要对应于文件系统中的文件。程序可以存储在含有其它程序或数据（例如一个或多个存储在标识语言文档中的脚本）的文件的一部分之中、在专用于讨论中的程序的单独文件之中、或在多个协调文件之中（例如存储一个或多个模块、子程序、或代码部分的文件）。可以把计算机程序布署地

在一个计算机上或者在位于一个站点上或在通过通信网络互联的分布于多个站点的多个计算机上执行。

[0111] 本说明书中描述的过程和逻辑流程可以通过一个或多个可编程处理器（执行一个或多个计算机程序以通过用输入数据运行并生成输出来实现函数）。过程和逻辑流程可由专用逻辑电路，例如FPGA（在线可编程门阵列）或ASIC（特定用途集成电路）来执行，装置也可以用此方式来实施。

[0112] 适合于执行计算机程序的处理器包括，举例来说，通用或专用微处理器，和任何其它一个或多个任何种类的数字计算机。一般说来，处理器将从只读存储器或随机存取存储器或两者接收指令和数据。计算机的基本单元是用于执行指令的处理器和用于存储指令和数据的一个或多个存储器设备。一般说来，计算机还包括一个或多个用于存储数据的诸如磁盘、磁光盘、或光盘的海量存储器设备，或可操作地耦合到这些存储设备以从其接收数据或对其发送数据，或既接收又发送。然而，计算机不必要具有这类设备。另外，计算机可以嵌入到其它设备，这里仅列举几个，例如，移动电话，个人数字助理 (PDA)、移动音频播放器、全球定位系统 (GPS) 接收机。适合于承载计算机程序指令和数据的信息载体包括所有形式的非易失存储器，包括举例来说半导体存储器设备，例如 EPROM、EEPROM 和闪存设备；诸如内部硬盘或可移动盘的磁盘等；磁光盘；和 CD-ROM 和 DV-ROM 盘。处理器和存储器可以由专用逻辑电路所增加也可以被其集成进去。为了提供与用户的交互，可以在具有用于将信息显示给用户的诸如 CRT（阴极射线管）或 LCD（液晶显示）监视器的显示设备，以及用户通过其可以将输入提供给计算机的键盘和诸如鼠标或轨迹球的定点设备的计算机上实施本发明的实施例。也可用其它种类的设备来提供与用户的交互；例如，提供给用户的反馈可以是任何形式的感觉的反馈，例如视觉反馈、听觉反馈、或触觉反馈；并且来自用户的输入可以用任何形式接收，包括声音、话音、或触觉输入。

[0113] 可以在计算机系统中实施本发明的实施例，该计算机系统包括：诸如数据服务器的后端组件、或包括诸如应用程序服务器的中间件组件、或包括通过其用户可以与本发明的实施例交互的诸如具有图形用户界面或网络浏览器的客户端计算机的前端组件、或任何这类后端、中间件或前端组件的组合。系统的组件可以通过任何形式或介质的数字数据通信相互连接，例如通信网络。通信网络的实例包括局域网（“LAN”）和诸如因特网的广域网（“WAN”）。

[0114] 计算系统可包括客户端和服务器。客户端和服务器通常相互远离，并且典型地通过通信网络进行交互。借助运行在各自计算机上的计算机程序，客户端和服务器产生关系，并且具有相互具有客户端 - 服务器关系。

[0115] 已经描述了本发明的特定实施例。其它实施例在以下权利要求的范围之内。例如，权利要求中叙述的步骤可以用不同的顺序来执行并仍然实现需要的结果。

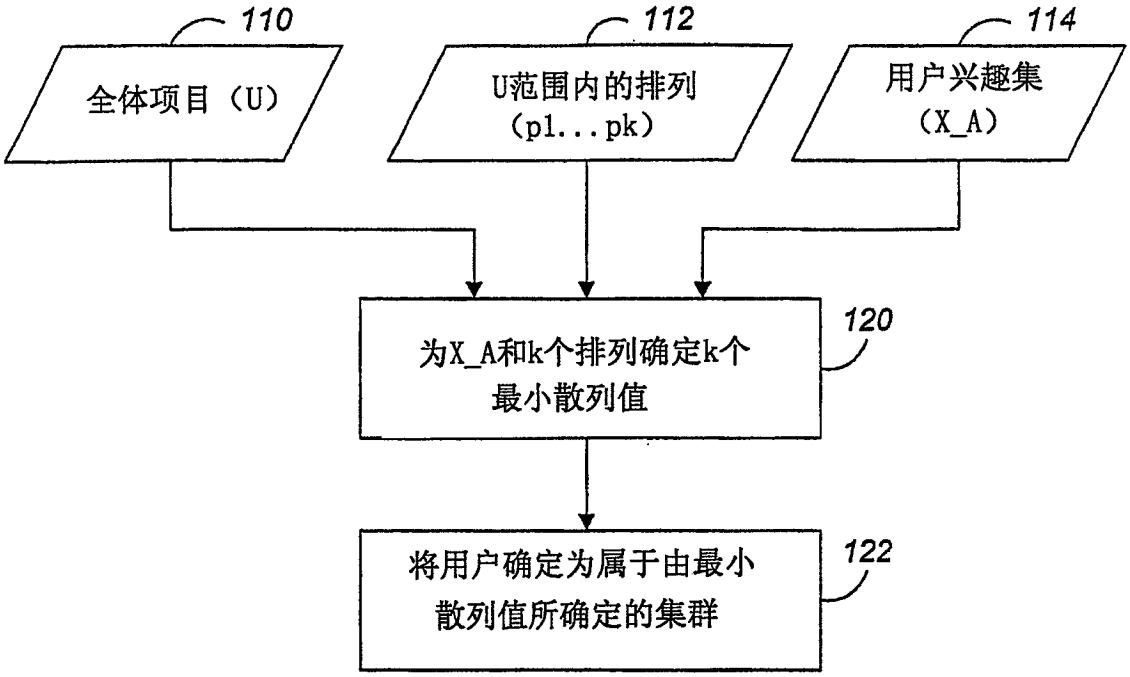

110
全体项目（U）
112
U范围内的排列
（p1...pk）
114
用户兴趣集
（X_A）
120
为X_A和k个排列确定k个
最小散列值
122
将用户确定为属于由最小
散列值所确定的集群

图 1

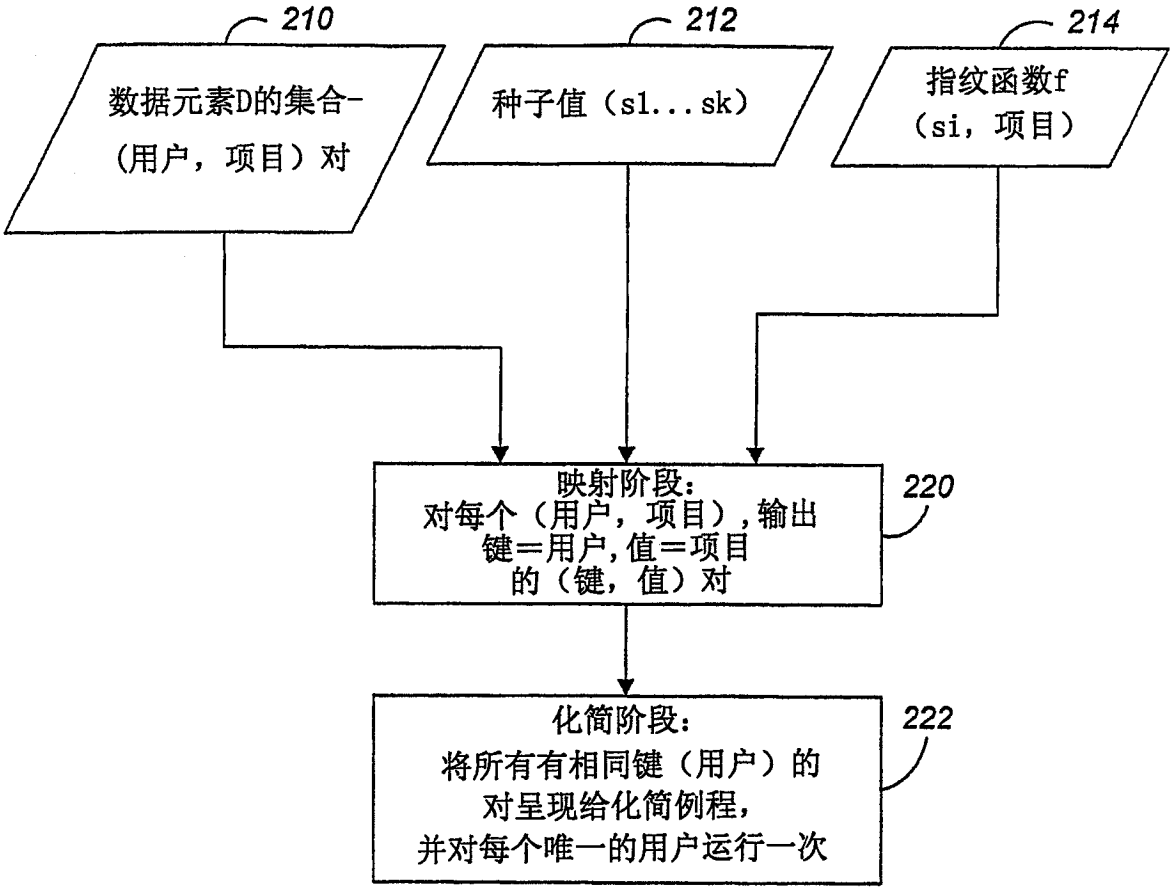

210
数据元素D的集合-
（用户，项目）对
212
种子值（s1...sk）
214
指纹函数f
（si，项目）
220
映射阶段：
对每个（用户，项目），输出
键＝用户，值＝项目
的（键，值）对
222
化简阶段：
将所有有相同键（用户）的
对呈现给化简例程，
并对每个唯一的用户运行一次

图 2

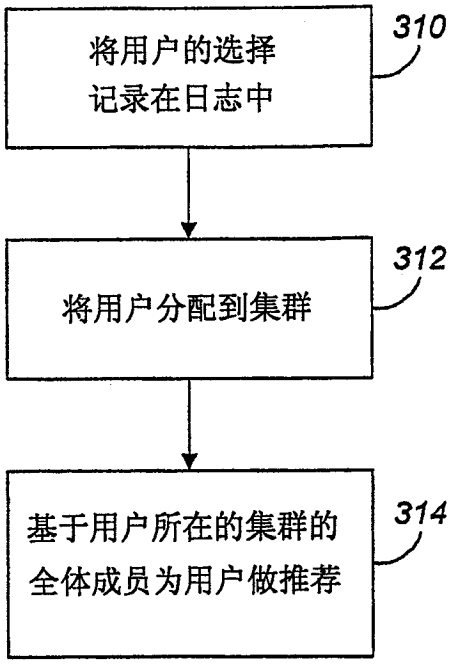


图 3

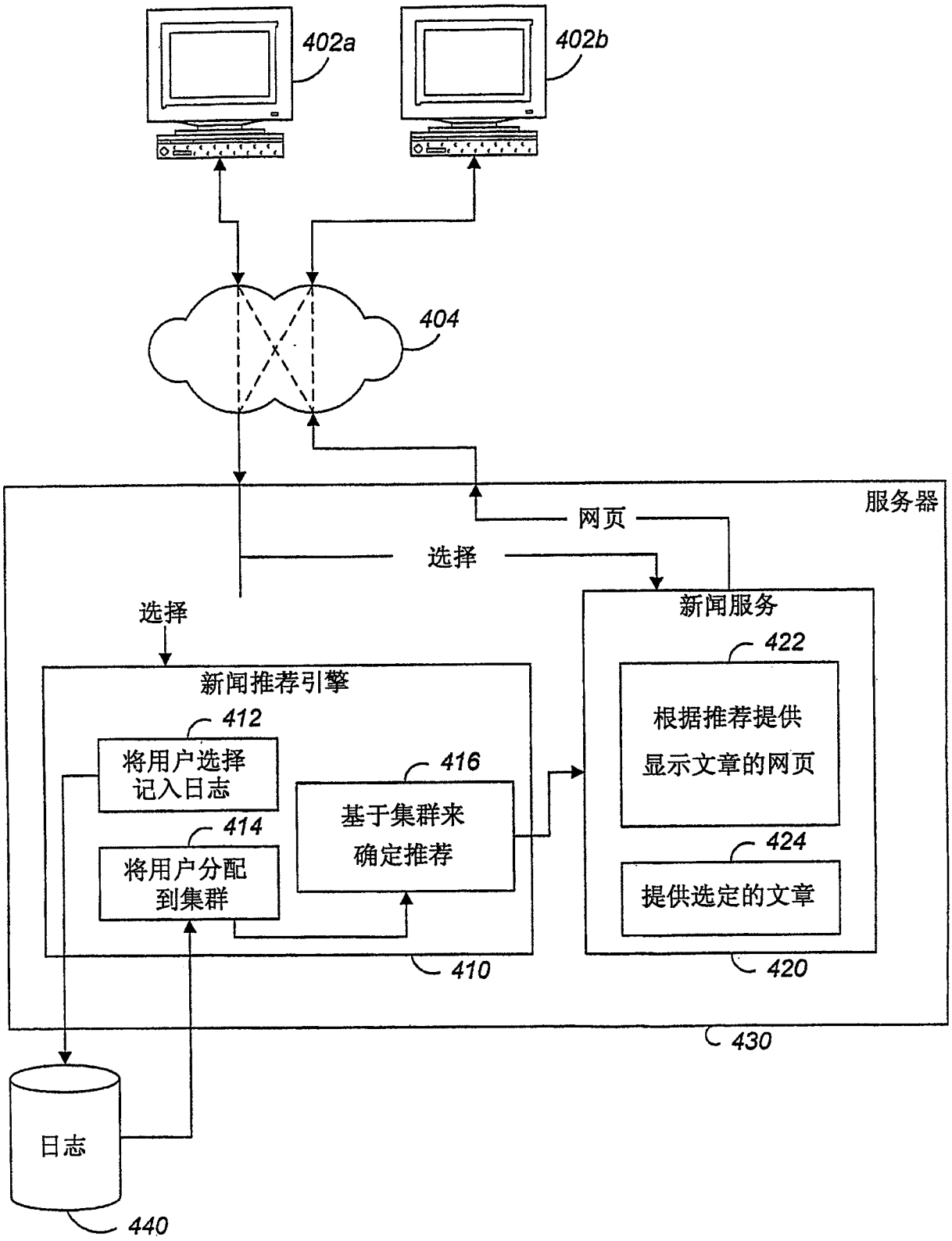

402a
402b
404
服务器
网页
选择
选择
新闻服务
422
新闻推荐引擎
412
将用户选择
记入日志
416
基于集群来
确定推荐
414
将用户分配
到集群
根据推荐提供
显示文章的网页
424
提供选定的文章
410
420
430
日志
440

图 4