

[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G06F 17/30 (2006.01)



[12] 发明专利说明书

专利号 ZL 200510113312.4

[45] 授权公告日 2009 年 4 月 22 日

[11] 授权公告号 CN 100481075C

[22] 申请日 2005.9.20

[21] 申请号 200510113312.4

[30] 优先权

[32] 2004.10.20 [33] US [31] 10/969,567

[73] 专利权人 微软公司

地址 美国华盛顿州

[72] 发明人 B·E·安德森 小 F·H·贝尔

J·C·芬格 J·J·马尔斯曼

K·卡那瓦特 M·B·米德兰

P·M·玛罗莱普斯 清水刚

T·D·怀特 Y·张

[56] 参考文献

CN1299488A 2001.6.13

US2003/0046389A1 2003.3.6

Agglomerative clustering of a search engine query log. BEEFERMAN D, BERGER A. PROCEEDINGS OF THE ACM SIGKDD. INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, ACM, US. 2000

Query Clustering Using User Logs. WEN J. R ET AL. ACM TRANSACTIONS ON INFORMATION SYSTEMS, ACM, NEW YORK, NY, US, Vol. 20 No. 1. 2002

审查员 何明伦

[74] 专利代理机构 上海专利商标事务所有限公司

代理人 顾嘉运

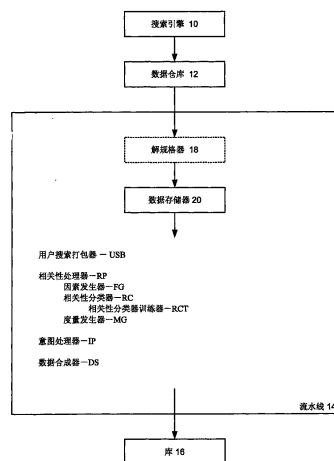
权利要求书 4 页 说明书 11 页 附图 2 页

[54] 发明名称

分析来自搜索系统的操作数据和其它数据的方法和系统

[57] 摘要

一种系统分析来自搜索引擎的数据。用户搜索打包器分析用户搜索，将相似的用户搜索组合成用户搜索包。因子发生器考虑用户搜索和相关信息以生成各因子，其中每个因子与来自搜索结果集的特定结果相关。相关性分类器接收各因子并基于各个因子操作以产生对每个结果的判断。度量发生器基于各个因子和各个判断生成度量，且数据合成器将已提取数据格式化到数据库中。



1. 一种用于分析来自搜索引擎的数据的系统，所述搜索引擎基于从请求用户接收的查询串生成一组搜索结果，所述查询串和所述搜索结果共同包括用户搜索，所述搜索结果包括至少一个结果，每个结果指向被认为是与所述查询串相关的特定内容项，从而一系列相关用户搜索包括一对话，所述搜索引擎存储每个用户搜索和相关信息，所述系统包括：

用户搜索打包器，分析用户搜索以找出这些用户搜索中的相似搜索，并将这些相似用户搜索组合成用户搜索包，

意图处理器，基于用户搜索包生成意图，每个意图是被认为是彼此相关的一个或多个对话的组合，所述意图处理器将具有在确定阈值之上的共性的强度的对话对打包到一个意图中，这种共性的强度表示了通过具有共同的目的的两个对话彼此相关的可能性，

因素发生器，考虑用户搜索和相关信息来产生因素，每个因素与来自一组搜索结果的特定结果相关，每个因素与一个或多个事件相关，每个事件是与查询用户所执行的动作相关的一段信息，

相关性分类器，接收由所述因素发生器对每个结果生成的因素，并基于所述因素以操作生成对所述结果的判断，所述判断表示对在用户决定访问来自各搜索结果的结果时，用户如何判断该结果的确定，

度量发生器，基于由因素发生器生成的因素和由相关性分类器生成的判断来产生度量，每个度量是有关结果、用户搜索或对话的度量，以及

数据合成器，提取由所述用户搜索打包器、意图处理器、因素发生器、相关性分类器和度量发生器生成的数据，将所述已提取数据格式化到一个或多个数据库，并将所述数据库存储到库中，从而可查看并聚集数据以提供反馈或生成报告。

2. 如权利要求 1 所述的系统，其特征在于，所述搜索引擎将每个查询串以及相应搜索结果和相关信息以规格化形式存储在数据仓库中，所述系统还包括一解规格器，它从所述数据仓库检索所述规格化数据、规格化所述数据、并将所述已规格化的数据存储在数据存储器中。

3. 如权利要求 1 所述的系统，其特征在于，为了查询串的相似性和搜索结果的相似性的至少之一，所述用户搜索打包器分析所述用户搜索。

4. 如权利要求 1 所述的系统, 其特征在于, 每个事件包括当用户执行选择和关闭特定结果的至少之一动作时的时间, 且其中所述因素发生器计算表示用户查看结果的时间长度的“停顿时间”因素, 所述停顿时间因素基于用户选择和关闭所述结果之间的时间差值, 每个时间都由相应的时间标记事件来表示。

5. 如权利要求 1 所述的系统, 其特征在于, 所述相关性分类器生成一判断, 包括“接受”判断、“浏览”判断、以及“拒绝”判断以及表示该判断正确性有多大可能的置信度的相应值。

6. 如权利要求 1 所述的系统, 其特征在于, 还包括相关性分类器训练器, 它接收来自所述因素发生器的显式判断因素并基于这些因素生成所述相关性分类器, 每个显式判断因素表示来自用户的有关相应结果的显式反馈, 所述相关性分类器训练器从所述显式判断因素学习什么因素暗示哪些判断并基于这个生成相关性分类器。

7. 如权利要求 1 所述的系统, 其特征在于, 所述度量发生器生成有关结果的以下至少之一:

位置度量, 有关判断用户如何已对所述结果排序,

相关性位置度量, 有关结果如何在各搜索结果内放置, 以及

错误排序结果度量, 有关对结果距离应当所在的位置有多“远”的度量, 这是基于所述位置度量和相关性位置度量的。

8. 如权利要求 1 所述的系统, 其特征在于, 通过基于所查看的用户搜索包在各对话上定位共同结果和共同查询项, 所述意图处理器确定对话之间的关系值并确定所发现的共性的强度。

9. 如权利要求 1 所述的系统, 其特征在于, 所述数据合成器将所述已提取数据格式化到关系数据库中。

10. 一种用于分析来自搜索引擎的数据的方法, 所述搜索引擎基于从请求用户接收的查询串生成一组搜索结果, 所述查询串和所述搜索结果共同包括用户搜索, 搜索结果包括至少一个结果, 每个结果指向被认为是与所述查询串相关的特定内容项, 从而一系列相关用户搜索包括一对话, 所述搜索引擎存储每个用户搜索和相关信息, 所述方法包括:

分析用户搜索以找出这些用户搜索中的相似搜索, 并将这些相似用户搜索组合成用户搜索包,

基于用户搜索包生成意图, 每个意图是被认为是彼此相关的一个或多个对话

的组合, 其中, 将具有在确定阈值之上的共性的强度的对话对打包到一个意图中, 这种共性的强度表示了通过具有共同的目的的两个对话彼此相关的可能性,

考虑用户搜索和相关信息来产生因素, 每个因素与来自一组搜索结果的特定结果相关, 每个因素与一个或多个事件相关, 每个事件是与查询用户所执行的动作相关的一段信息,

接收对每个结果生成的所述因素, 并基于所述因素以操作生成对所述结果的判断, 所述判断表示对在用户决定访问来自搜索结果的结果时用户如何判断该结果的确定,

基于生成的因素和生成的判断来产生度量, 每个度量是有关结果、用户搜索或对话的度量, 以及

提取包括所述用户搜索包、意图、因素、判断和度量的数据, 将所述已提取数据格式化到一个或多个数据库, 并将所述数据库存储到库中, 从而可查看并聚集数据以提供反馈或生成报告。

11. 如权利要求 10 所述的方法, 其特征在于, 包括将每个查询串以及所述相应搜索结果和相关信息以规格化形式存储在数据仓库中, 并且还包括从所述数据仓库检索所述规格化数据、规格化所述数据、并将所述已规格化的数据存储在数据存储库中。

12. 如权利要求 10 所述的方法, 其特征在于, 为了查询串的相似性和搜索结果的相似性的至少之一, 分析所述用户搜索。

13. 如权利要求 10 所述的方法, 其特征在于, 每个事件包括当用户执行选择和关闭特定结果的至少之一动作时的时间、且其中所述方法计算表示用户查看结果的时间长度的“停顿时间”因素, 所述停顿时间因素基于用户选择和关闭所述结果之间的时间差值, 每个时间都由相应的时间标记事件来表示。

14. 如权利要求 10 所述的方法, 其特征在于, 包括生成一判断, 它包括“接受”判断、“浏览”判断、以及“拒绝”判断以及表示该判断正确有多大可能的置信度的相应值。

15. 如权利要求 10 所述的方法, 其特征在于, 还包括接收显式判断因素并基于这些因素生成相关性分类器, 所述相关性分类器接收对每个结果生成的因素并基于这些因素操作以生成对所述结果的判断, 每个显式判断因素表示来自用户的有关相应结果的显式反馈, 从而可基于这些显式判断因素来学习什么因素暗示哪些判断。

16. 如权利要求 10 所述的方法，其特征在于，包括生成有关结果的以下至少之一：

位置度量，有关判断用户如何已对所述结果排序，

相关性位置度量，有关结果如何在各搜索结果内放置，以及

错误排序结果度量，有关对结果距离应当所在的位置有多“远”的度量，这是基于所述位置度量和相关性位置度量的。

17. 如权利要求 10 所述的方法，其特征在于，包括通过基于所查看的用户搜索包在各对话上定位共同结果和共同查询项，确定对话之间的关系值，并确定所发现的共性的强度。

18. 如权利要求 10 所述的方法，其特征在于，包括将所述已提取数据格式化到关系数据库中。

分析来自搜索系统的操作数据和其它数据的方法和系统

技术领域

本发明涉及用于分析数据的系统和方法，该数据包括在诸如用于响应搜索请求的系统的系统操作期间编译的数据。本发明尤其涉及为了包括改进响应和维持系统性能的目的而分析相对大量的数据的系统和方法。

背景技术

结合诸如典型的搜索引擎，访问该典型搜索引擎的用户通过输入包含一个或多个搜索项的可能带有布尔操作符的搜索串等来请求搜索。作为响应，搜索引擎基于该搜索串来搜索一个或多个数据库，生成一个搜索结果集，并可能以用户可查看的信息页面或信息链接的形式将这些搜索结果返回给请求用户。特别是在信息链接形式的情形中，用户可访问一个或多个链接，以查看与特定搜索结果相关的内容，且如果与搜索结果的一个或多个链接相关联的内容对用户而言是可接受的，则该用户通常继续以任何看起来适当的方式来采用这种可接受的内容。

然而，可能是相反的情形，因为搜索结果中没有内容满足所请求的搜索，搜索结果对用户而言是不可接受的（至少从该用户的角度来看）。在这种情形中，用户可决定输入新的搜索串或修改先前输入的搜索串，并查看基于该新的或经修改的搜索串的来自搜索引擎的搜索结果。可以理解，这样的过程可以搜索对话的形式反复多次直到用户找到可接受的搜索结果或放弃为止。

一般而言，在诸如前述搜索引擎的高质量系统中，来自用户的如在搜索串中提出的每个查询应当准确地映射到表示回答该查询的内容的搜索结果。这种目标对提供好的搜索体验是重要的，并且实际上符合这种目标可显示出以下两者之间的差异：将以新的搜索对话返回该搜索引擎的快乐的满意的用户，和将相反访问另一搜索引擎的愤怒的不满意的用户。

为了实现该目标，已经认识到系统本身应当进行调整或“调谐”以改进响应并保持系统性能。如可以理解的那样，这种调整可主要（但不是只）基于系统操作期间编译的操作数据和其它数据来进行。然而，如果系统特别大，具有较高的通信量等，可用于分析的这种数据量会特别大，因而不易使用并以其它方式显现为难以

工作。因此，需要用于分析大量数据，特别是来自搜索系统的大量数据的系统和方法。

发明内容

前述需要至少部分地通过本发明得到了满足，在本发明中提供用于分析来自搜索引擎的数据的系统。该搜索引擎基于从请求用户处接收的查询串来生成一个搜索结果集，其中查询串和搜索结果一起组成用户搜索。搜索结果包括至少一个结果，其中每个结果指向相信与查询串相关的特定内容项。一系列相关用户搜索组成一个对话，并且搜索引擎存储每个用户搜索及相关信息。

在该系统中，用户搜索打包器（USB）分析用户搜索，以找出这些用户搜索中相似的用户搜索，并将这些相似的用户搜索组成用户搜索包，而意图处理器（IP）基于来自 USB 的用户搜索包产生意图。每个意图是被认为是彼此相关的一个或多个对话的组合。

因素发生器（FG）考虑用户搜索及相关信息以产生因素，其中每个因素关于来自一个搜索结果集的一个特定结果。每个因素涉及一个或多个事件，其中，每个事件是与查询用户进行的动作有关的一段信息。相关性分类器（RC）接收 FG 为每个结果生成的因素，并基于它们操作以产生对结果的判断，其中该判断表示对在用户决定访问来自各搜索结果的结果之后用户如何判断该结果的确定。度量发生器（MG）基于由 FG 生成的因素和由 RC 生成的判断来产生度量，其中每个度量是有关结果、用户搜索或对话的度量。最后，数据合成器（DS）提取由 USB、IP、FG、RC 和 MG 生成的数据，将已提取数据格式化到一个或多个数据库，并将各数据库存储到库中，借此，可查看并聚集数据以提供反馈或生成报告。

附图说明

结合附图阅读前面的发明内容和以下本发明各实施例的详细描述，能得到更好的理解。为了说明本发明，在附图中示出当前优选的各实施例；然而，应理解本发明并不限于所示的精确排列和装置。在附图中：

图 1 是示出其中可组合本发明各方面和/或各部分的通用计算机系统的框图，以及

图 2 是根据本发明各实施例示出用于分析来自诸如搜索引擎之类的较大系统的数据的数据分析流水线的框图。

具体实施方式

计算机环境

图 1 和以下讨论旨在提供一种本发明和/或其部分可在其中实现的适当计算环境的简要一般说明。尽管不是必需的,本发明将在由诸如客户机工作站或服务器的计算机执行的诸如程序模块的计算机可执行指令的一般环境中进行说明。通常,程序模块包括执行特定任务或实现具体抽象数据类型的例程、程序、对象、组件、数据结构等等。此外,本领域技术人员将理解,本发明和/或其部分可在其它计算机系统配置中实践,包括手持式装置、多处理器系统、基于微处理器的或可编程的电器消费品、网络 PC、小型计算机、大型计算机等等。本发明还可在任务由经通信网络链接的远程处理装置执行的分布式计算环境中实践。在分布式计算环境中,程序模块可被置于本地和远程存储器存储设备中。

如图 1 所示,示例性通用计算系统具有常规个人计算机 120 等,包括处理单元 121、系统存储器 122 以及把包括系统存储器的各种系统组件耦合到处理单元 121 的系统总线 123。系统总线 123 可能是若干总线结构类型中的任何一种,包括存储器总线或存储器控制器、外围总线、以及使用多种总线体系结构的任一种的局部总线。系统存储器包括只读存储器(ROM) 124 和随机存储器(RAM) 125。含有帮助如启动期间在个人计算机 120 中元件之间传递信息的基本例程的基本输入/输出系统(BIOS) 126 被存储在 ROM 124 中。

个人计算机 120 还包括读取和写入硬盘(未示出)的硬盘驱动器 127、读取或写入可移动磁盘 129 的磁盘驱动器 128、以及读取或写入诸如 CD ROM 或其它光学介质等可移动光盘 131 的光盘驱动器 130。硬盘驱动器 127、磁盘驱动器 128 和光盘驱动器 130 分别通过硬盘驱动器接口 132、磁盘驱动器接口 133 和光盘驱动器接口 134 连接至系统总线 123。诸驱动器及其相关联计算机可读介质为计算机 120 提供计算机可读指令、数据结构、程序模块和其它数据的非易失性储存。

尽管在此所述示例性环境采用了硬盘、可移动磁盘 129 和可移动光盘 131,但本领域技术人员应理解也可在示例性操作环境中使用其它类型计算机可访问的能够存储数据的计算机可读介质。这样的其它类型的介质包括诸如磁带盒、闪存卡、数字视频盘、Bernoulli 卡式磁带、随机存取存储器(RAM)、只读存储器(ROM)等等。

包括操作系统 135、一个或多个应用程序 136、其它程序模块 137 和程序数据 138 的众多程序模块，可存储在硬盘、磁盘 129、光盘 131、ROM 124 或 RAM 125 中。用户可通过诸如键盘 140 和定位装置 142 的输入装置把指令和信息输入个人计算机 120。其它输入装置（未示出）可包括诸如话筒、操纵杆、游戏垫、卫星接收器、扫描仪等等。这些和其它输入装置常常通过与系统总线相耦合的串行端口接口 146 连接到处理单元 121，但是它也可通过其它接口相连，如并行端口、游戏端口或通用串行总线（USB）。监视器 147 或其它类型的显示装置也通过诸如视频适配器 148 的接口和系统总线 123 相连。除了显示器 147，个人计算机通常包括其它外围输出装置（未示出），如扬声器和打印机。图 1 的示例性系统还包括主机适配器 155、小型计算机系统接口（SCSI）总线 156、以及与 SCSI 总线 156 相连的外部存储装置 162。

个人计算机 120 可以在使用与一台或多台远程计算机，诸如远程计算机 149 的逻辑连接的网络化环境中运行。远程计算机 149 可以是另一台个人计算机、服务器、路由器、网络 PC、对等装置或其它公共网络节点，而且通常包括上述与个人计算机 120 相关的许多或全部元件，尽管在图 1 中仅显示了存储器存储装置 150。图 1 中所描绘的逻辑连接包括局域网（LAN）151 和广域网（WAN）152。这样的网络化环境在办公室、企业范围计算机网络、企业内联网和因特网上是常见的。

当用于 LAN 网络环境中时，个人计算机 120 通过网络接口或适配器 153 与 LAN 151 连接。当用于 WAN 网络环境中时，个人计算机 120 通常包括调制解调器 154 或其它用于在诸如因特网的广域网 152 上建立通信的装置。可以是内置式或外置式的调制解调器 154，与系统总线 123 通过串行端口接口 146 相连接。在网络化环境中，所述与个人计算机 120 相关的程序模块或其一部分，可以存储在远程存储器存储装置中。可以理解，所示网络连接是示例性的，也可以使用其它在计算机之间建立通信连接的装置。

数据分析流水线

作为准备并作为术语，可以理解，结合诸如可结合本发明采用的搜索引擎，每个访问该搜索引擎的用户通过输入包括具有一个或多个搜索项的可能具有布尔算子的搜索串的查询来请求搜索。作为响应，搜索引擎基于该搜索串生成一个搜索结果集（假设这些搜索结果实际上可用的），并将这些搜索结果返回给请求用户。所返回的搜索结果可包括被认为是与搜索请求相关的特定内容项或结果，尽管更可

能是相反每个特定结果是通过搜索结果中的相应链接来访问的。

特别地，如果所返回的搜索结果是不可接受的，则用户可用新的搜索串或先前输入的搜索串的修改来输入另一查询串，从而基于该另一查询来产生另一个搜索结果集。然后一系列相关查询组成整个搜索对话，并且该对话有望在用户在搜索结果中找到可接受的结果时结束。

在用户输入查询串并查看搜索结果的过程中，搜索引擎或相关实体可以并常常标识并存储与该对话相关的数据。具体地，除标识该对话之外，搜索引擎或相关实体（此后称为“搜索分析器”）可标识并存储该对话的每个查询串、以及对每个查询串返回的搜索结果等。此外，搜索分析器可对搜索结果的每个返回结果的每个链接标识并存储用户是否借此访问了该结果，以及因此用户花了多长时间来查看该结果等等。因而应当理解，搜索分析器或具有这种信息的另一实体可导出关于用户对每个来自对话的返回搜索结果集有多满意或“愉悦”的定性的（即使是非定量的）度量。

注意，执行前述功能的搜索分析器是众所周知的，或应当对相关人群是显而易见的，因此无需作任何详细阐述。因此，可结合本发明采用任何适当的搜索分析器。

作为形成对话的查询串的组合的一个示例，考虑以下查询串：“Cars”（汽车）、“Fords”（福特）、“Ford Edsel”（福特埃德塞尔）、“Cheap Vacations”（便宜的旅行）、和“London trip prices”（伦敦旅行价格）。应当理解，每个查询串是用户在搜索时输入的实际文本。此外，每个查询串产生具有对结果的链接的一个搜索结果集，并可通过搜索分析器使相关数据与之相关联，这些相关数据诸如是否选择了每个链接、查看该结果所花的停顿时间、对于该结果采取的滚动和其它动作、以及其它类似用户动作。

基于所有前述信息，应当理解前三个查询串（即“Cars”、“Fords”、“Ford Edsel”）是第一个对话部分，而后两个查询串（即“Cheap Vacations”、“London trip prices”）是第二个对话部分，因为用户显得具有两个不同的意图。这样就应当理解，对话是一组由单个用户在时间上连续作出的具有相同意图或目的的查询串及其相应搜索结果（即用户搜索）。

现在参看图2，可看到本发明假定诸如搜索引擎10的系统用于执行来自冷酷的请求者的请求搜索，并用适当的搜索结果来响应。该搜索引擎10当然可以是任何适当的搜索引擎，而不背离本发明的精神和范围。该搜索引擎10的操作通常是

众所周知的或应当对相关人群是显而易见的，因此除非另外提出，否则无需作任何详细阐述。更一般地，系统可以是与搜索引擎 10 不同的系统，这也不背离本发明的精神和范围。

如在图 2 中所见，假设与系统/搜索引擎 10 相连，维护有数据仓库 12。可以理解，数据仓库 12 结合搜索引擎 10 的操作存储数据，可能包括但不限于由搜索分析器生成的前述数据。可由该数据仓库 12 存储的这种数据当然可以是任何其它适当的数据而不背离本发明的精神和范围，且类似地数据仓库 12 及其组织可以是任何适当的数据仓库和组织而不背离本发明的精神和范围。

可以理解，数据仓库 12 内的数据可无限增长或可周期性地过期并删除。因而，在本发明中，流水线 14 被用来分析该数据，并且如以下详述，可分析数据仓库 12 内数据的任何部分，包括全部数据或数据的滑动窗口。可以看到，由流水线 14 执行的这种数据分析可包括若干组件，每个组件从仓库中取得数据的一部分并在其上操作。在本发明一实施例中，流水线 14 的输出被置于库 16 中，同时输出数据组织成诸如关系数据库、多维表格等的形式。因而，用适当的控制，web 浏览器上的用户可查看这种经组织的数据并可能还对其进行分析。

可以理解，由流水线 14 执行的处理提供有关如何使用搜索引擎 10 的统计数据。此外，并在本发明一实施例中，这种执行过程用来标识用户如何使用搜索引擎 10 的模式，特别是可提供真正有价值信息的如此之多的模式。

注意仓库 12 中的数据可高度规格化，特别是因为可以理解规格化这种数据使新数据能以有效方式添加。然而，这种规格化数据在分析之前必须进行解规格化，且相应地数据流水线 14 可包括解规格器 18 作为其输入。这种解规格器 18 可以是任何适当的解规格器并可以任何适当方式操作。这种解规格器 18 的操作是众所周知的或应对相关人群是显而易见的，因此无需在此作任何详细阐述。

在本发明的流水线 14 中采用的解规格器 18 解规格仓库 12 中的数据，并还可创建解规格化的批数据用于后来的处理。批处理可根据任何适当标准来执行。例如，批数据可以是上个批数据到达之后的所有数据，其中如果解规格器 18 每天运行一次则批数据可以是一天的数据。类似地，解规格器 18 可每周运行一次，但被设置成生成该周内每一小时的批数据。

可见，解规格器 18 的批数据输出被存储在数据存储器 20 中，用于由流水线 14 的其它组件作进一步处理。这种数据存储器 20 和其中批数据的组织可以是任何适当的存储器和组织，而不背离本发明的精神和范围。如图 2 所示，采用数据存储

器 20 中批数据的组件可包括用户搜索打包器 (USB)、相关性处理器 (RP)、意图处理器 (IP)、数据合成器 (DS) 等等。

数据流水线 14 的用户搜索打包器 (USB) 分析用户搜索, 每个用户搜索是用户在搜索引擎 10 上调用查询的一个实例, 并包括有关查询串和返回搜索结果等的信息。两个调用相同查询串的用户产生两个用户搜索, 而调用同一查询串两次的一个用户也可导致两个用户搜索。在本发明一实施例中, USB 尝试找出“看起来相像”并因此相似的用户搜索, 并将这种相似的用户搜索组合成用户搜索包。USB 可用任何适当的方式操作而不背离本发明的精神和范围。操作 USB 的方法是众所周知的或对相关人群是显而易见的, 因此无需在此作任何详细描述。

例如, USB 可采用分析查询串的“相似”(Look Alike) 算法。例如, 如果两个用户都调用“Wilkes-Barre, PA”的搜索, 则那些查询串看起来相像。更有趣的是, 如果一个用户搜索“duck”而另一个搜索“ducks”(一个复数一个单数), 则那两个查询串也因为算法采用词干化、案例规格化、以及其它类似查询压缩技术而“相似”。

然而, USB 并不限于分析查询串的相似性。相反, 在本发明一实施例中, USB 还可采用返回搜索结果分析算法。例如, USB 可基于所返回的搜索结果分析算法来推断, 即使相应的查询串并非都类似, 如果每个用户搜索产生一组类似的搜索结果, 则两个用户搜索相似。此外或作为替代, USB 可根据内容满意分析算法来推断, 如果每个用户搜索产生用户确定为满意的特定片段的内容或结果, 则两个用户搜索相似。类似地, USB 可基于内容不满分析算法来推断, 如果每个用户搜索产生用户确定为不满意的特定内容或结果, 则两个用户搜索不相似。

USB 的输出可存储在作为打包的用户搜索的数据库的库 16 中。或者, 这种输出可存储回数据存储器 20 中, 以对流水线 14 的其它组件可用。

本发明的相关性处理器 (RP) 可再划分成因素发生器 (FG)、相关性分类器 (RC) 和度量发生器 (MG)。但是注意, RP 的这种再分元件还可独立呈现在本发明的流水线 14 中, 而不组合成 RP。

因素发生器 (FG) 考虑有关用户搜索的信息, 并产生因素, 其中这些因素作为输入应用于相关性分类器 (RC)。每个因素与来自搜索结果集的特定结果相关。因素常常但并非总是通过考虑若干事件而产生, 其中每个事件通常是与查询用户执行的动作相关的一段信息。例如, 选择特定结果可以是一事件, 像结果可被显示、关闭、打印、添加到特定列表、重新打开等。事件还可包括显式的用户反馈 (如果

有的话)。

然后在本发明一实施例中, FG 基于事件计算值。例如, FG 可计算作为因素的“停顿时间”, 它表示用户查看结果的时间长度, 并可基于用户开始和结束查看该结果的时间差来计算, 每个时间都由相应的时间标记事件来表示。但是注意, 在产生停留时间因素时, FG 也可考虑其它事件, 包括用户在查看结果时是否被打断, 如由其它适当时间标记事件所表示。

由 FG 生成的另一因素可以是用户是否将结果添加到诸如喜爱列表之类的特定列表的因素。在这种情形中, 这种“喜爱”因素被设置为真值。可以理解, 基于事件的因素类型可以是任何适当因素而不背离本发明的精神和范围。这种因素是众所周知的, 或应当对相关人群是显而易见的, 因此无需在此作任何详细阐述。

注意, 因素也可来自不同于事件的东西。例如, 因素可以是制作该结果的用户标识符。因而, 因素可以是结果的属性。其它因素可包括查询串的所属地区(例如美式英语、加拿大英语、巴西葡萄牙语), 并可从用户搜索中的信息中导出。更一般地, 因素可由 FG 或另一元件从可用的任何信息片段中导出, 而不背离本发明的精神和范围。

由 FG 或可能其它地方生成的因素可存储在适当数据库内的库 16 中。或者, 这种因素可存储回数据存储器 20 以对流水线 14 的其它组件可用, 或可直接传送到这些组件中。

相关性处理器(RP)的相关性分类器(RC)接收由 FG 和可能其它地方生成的因素。RC 是机器生成的决策树, 它基于作为输入接收因素进行操作。在由 RP 起动之后, RC 从数据存储器 20 或其它地方读取结果的因素, 并生成对该结果的判断。一般而言, 这种判断是在决定访问来自搜索结果集的结果之后对用户如何判断该结果的确定。

判断可根据任何适当的判断系统来表达, 而不背离本发明的精神和范围。例如, 判断可以是数字或字母分级记分, 可以是“接受”(即用户对结果满意)、“浏览”(即用户对结果既不是满意也不是不满意)、以及“拒绝”(即用户对结果不满意)等。此外, 对特定结果的判断还可包括表示该判断有多正确的置信度的值, 由 RC 来确定。

RC 和可能其它地方生成的判断可存储在适当数据库内的库 16 中。或者, 这种因素可存储回数据存储器 20 以对流水线 14 的其它组件可用, 或可直接传送到这些组件中。

为了机器生成 RC 的决策树，可以是如下情形：RP 考虑具有显式判断因素的每个结果。显式判断是表示来自用户的有关结果的显式反馈的一类因素。通过取具有显式判断因素的每个结果并分析该结果的其它可用因素，RP 或其它地方的相关性分类器训练器（RCT）能够“学习”什么因素暗示哪些判断并基于这个构建 RC。学习因素暗示什么并从中构建 RC 可用任何适当方式完成，而不背离本发明的精神和范围。这样做是众所周知的，或对相关人群显而易见的，因此无需在此作任何详细阐述。

RP 的度量发生器（MG）产生度量，其中度量是有关结果、用户搜索、对话等的度量。一般而言，MG 基于由 FG 和可能其它地方生成的因素、RC 生成的判断、以及其它可用的相关信息来产生这种度量。生成这种度量可用任何适当方式完成，而不背离本发明的精神和范围。这样做是众所周知的，或对相关人群显而易见的，因此无需在此作任何详细阐述。

MG 可生成以下有关结果的度量：

- 接受、浏览、拒绝：每一个都从结果的判断中导出并可包括置信度。
- 错误排序结果：对结果距离应当所在的位置有多“远”的度量，这是基于结果在搜索结果内如何放置以及判断用户如何已对该结果排序。
- 位置：判断用户如何已对结果排序。
- 相关性位置：结果如何在各搜索结果内放置。

MG 可生成以下有关用户搜索的度量：

- 结果集排序记分：类似于对每个结果生成的错误排序结果度量，但与用户搜索的所有结果相关。该值尝试反映结果集距离应当所在位置有多远。
- 经归纳的接受、浏览、拒绝：分别归纳用户搜索中每个结果的所有接受、浏览、拒绝度量。

MG 可生成以下有关对话的度量：

- 内容量：对话是否显示缺乏用户在搜索什么的可用信息。
- 意图确定：对话是否显示出不能确定用户在搜索什么。

由 MG 和可能其它地方生成的度量可存储在适当数据库内的库 16 中。或者，这种因素可存储回数据存储器 20 以对流水线 14 的其它组件可用，或可直接传送到这些组件中。

流水线 14 的意图处理器（IP）生成意图，其中每个意图是被认为是因为用户在每个实例中查寻相同搜索结果而彼此相关的一个或多个对话的组合。即，意图是

一组共享共同目的的对话。

IP 通过考虑每个对话、每个对话的每个用户搜索、以及每个用户搜索的每个结果来生成意图。在本发明一实施例中，并记住具有共同结果的对话可能是相关的，只有具有判断“接受”(Judgement of Accept)的结果才进行研究。然而，也可采用具有其它判断的结果，而不背离本发明的精神和范围。

一般而言，通过基于所查看的用户搜索包在对话上定位共同结果和共同查询项，并确定所发现的共性的强度，IP 确定对话之间的关系值。这种强度表示通过具有共同的目的，两个对话有多大可能彼此相关。具有在某一确定阈值之上强度的对话对可由 IP 打包到一个意图中。

由 IP 和可能其它地方生成的意图可存储在适当数据库内的库 16 中。或者，这种因素可存储回数据存储器 20 以对流水线 14 的其它组件可用，或可直接传送到这些组件中。

最后，用所有迄今为止生成的数据，流水线 14 包括从数据存储器 12、库 20、或其它地方中提取这些数据，并将这些数据格式化到要存储在库 20 中的一个或多个数据库的数据合成器(DS)。特别是如果不同的组件将数据写入不同的表格、数据库等，这种格式化是必需的。例如，USB 可将有关每个用户搜索的数据添加到用户搜索打包表格，且 MG 可能已经将数据添加到度量表格。因而，DS 在适当时将这些表格和其它结合到有意义的表格中，从中可对这些数据进行搜索、聚集、呈现等等。

注意 DS 可编程为只按需格式化数据的一部分。例如，如果一部分存储数据被认为是没有用的，DS 可略去这些数据。或者，DS 可将相同的数据复制到多个表格或将数据分到多个表格中。更一般地，DS 可编程为将来自流水线 14 的数据格式化成任何适当格式，而不背离本发明的精神和范围。

在本发明一实施例中，DS 将数据格式化到诸如 SQL 数据库等的关系数据库中。这样，数据可表示为“星”形，其中事实表(fact table)在中间而维表(dimension table)在四周。可以理解，通过这样做，事实表和维表可建成诸如 OLAP 立方体等的数据立方体，它使搜索引擎 10 的管理员能基于不同标准有效地聚集数据。例如，该管理员可请求该立方体找出存在于某日期范围中的每个对话根据意图来聚集对话的意图确定度量的平均值。类似地，报告可根据该立方体来运行，且其结果可被转发给管理员等。当然，该管理员还可根据该立方体布置大多数任何其它适当的查询，或接收根据该立方体运行的任何其它报告。

结论

本发明可实践关于分析来自搜索引擎 10 的大多数任何适当系统的数据。应理解，使用在此阐述的本发明，可为任何目的分析系统的数据表示，包括调整或“调节”系统以改进响应、维护系统性能、并以其它方式确保系统在令人满意地运行。

实现结合本发明执行的过程所需的编程是相对简单的，且应当对相关编程人员显而易见。因此，这些程序未附于此。任何具体的编程可用来实现本发明而不背其精神和范围。

在前面的描述中，可以看到本发明包括分析大量数据尤其是来自搜索引擎 10 等的大量数据的新颖和有用的装置。这种装置在以下情形特别有用：搜索引擎 10 特别大、具有较高通信量等，且可用来分析的这种数据量特别大，因此不易使用并在其它方面难以工作。

应理解，可对上述实施例作改变而不背离其中发明概念。因此，应理解，本发明并不限于所揭示的特定实施例，而旨在涵盖由所附权利要求定义的本发明精神和范围内的更改。

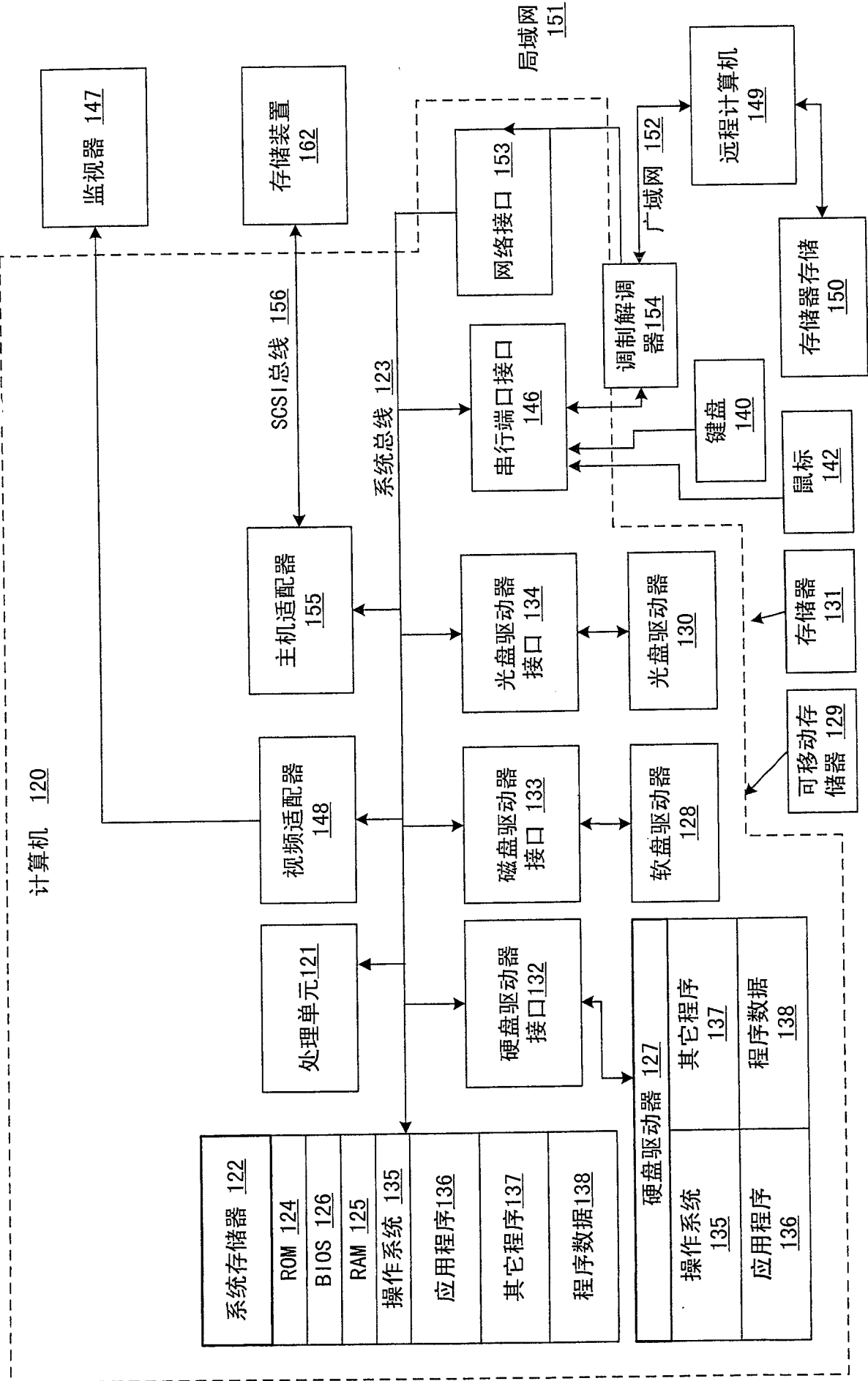


图 1

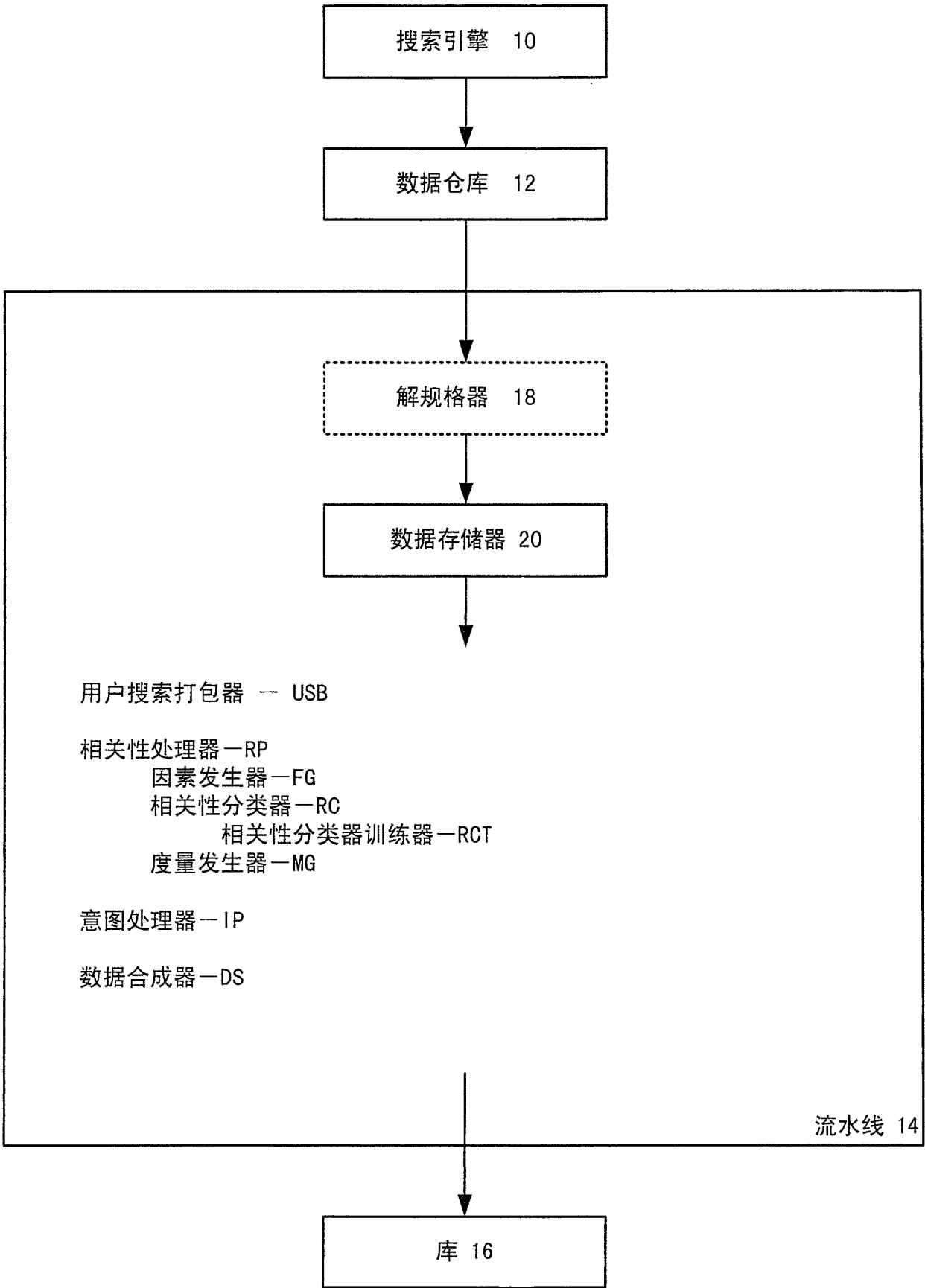


图 2