[19] 中华人民共和国国家知识产权局

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

[12] 发明专利申请公开说明书

[21] 申请号 200480013412.3

[43] 公开日 2006 年 11 月 8 日

[11] 公开号 CN 1860474A

[22] 申请日 2004.4.16
[21] 申请号 200480013412.3
[30] 优先权
[32] 2003.4.16 [33] US [31] 10/417,709
[86] 国际申请 PCT/US2004/011772 2004.4.16
[87] 国际公布 WO2004/095178 英 2004.11.4
[85] 进入国家阶段日期 2005.11.16
[71] 申请人 雅虎公司
地址 美国加利福尼亚州
[72] 发明人 J·昌德
[74] 专利代理机构 北京市中咨律师事务所
代理人 杨晓光 李 峥
权利要求书 9 页 说明书 19 页 附图 3 页

[54] 发明名称

关联性分析的方法和产品

[57] 摘要

一种基于计算机的组织文本项的方法，其包括：接收大量单独的文本项的组；将各个唯一的整数项码赋给来自所述大量单独组的各所述单独文本项；识别来自所述大量组中的单独组中的文本项对；基于赋给各所述文本项对的所述单独文本项的所述唯一项码来对所述文本项对进行排序，从而，各所述对具有各自唯一的相对于其它对的排序位置；在计算机可读介质中提供对排序信息结构，其存储各所述已识别的对与其各自的所述唯一排序位置之间的关联。

1. 一种基于计算机的组织文本项的方法，包括以下步骤：

接收大量单独的文本项的组；

将各个唯一的整数项码赋给来自所述大量单独组的各所述单独文本项；

识别来自所述大量组中的单独组的文本项对；

基于赋给各所述文本项对的所述单独文本项的所述唯一项码来对所述文本项对排序，从而，各所述对具有各自唯一的相对于其它对的排序位置；以及

在计算机可读介质中提供对排序信息结构，其存储各所述已识别的对与其各自的所述唯一排序位置之间的关联。

2. 根据权利要求 1 的方法，其中，所述对所述文本项对进行排序的步骤进一步包括：

识别赋给各所述对的各所述单独文本项的各较低值项码值和较高值项码值；

识别所述文本项对的各个组，所述文本项对的组具有赋给各所述单独文本项的相同的所述较低值项码或赋给各所述单独文本项的相同的所述较高值项码中指定的一个；

通过以下步骤，确定所述各已识别的对的各自排序位置，

基于赋给各所述对的组的各单独文本项的各所述较低值项码或所述较高值项码中所述指定的一个，以指定的数字次序对所述各已识别的对的组进行排序；以及

基于赋给各所述对的组中的所述对的各文本项的所述较低值项码或所述较高值项码中的一个，以指定的数字次序对各所述对的组中的各对进行排序。

3. 根据权利要求 2 的方法，进一步包括以下步骤：

以所述各对的各自排序位置的数字次序将各唯一的整数对码值赋给所

述各对。

4. 根据权利要求 2 的方法，进一步包括以下步骤:

以根据所述各对的各自排序位置的数字次序将各唯一的整数对码值赋给所述各对;

其中，所述对排序信息结构存储各所述已识别的对与它们各自的所述已赋值的唯一对码值之间的关联。

5. 根据权利要求 1 的方法，进一步包括以下步骤:

在所述计算机可读介质中提供项码信息结构，其将所述文本项与所述已赋值的唯一整数项码相关联。

6. 根据权利要求 2 的方法，进一步包括以下步骤:

在所述计算机可读介质中提供项码信息结构，其将所述文本项与所述已赋值的唯一整数项码相关联; 并且

以根据所述各对的各自排序位置的数字次序将各唯一的整数对码值赋给所述各对;

其中，所述对排序信息结构存储各已识别的对与它们各自的已赋值的唯一对码值之间的关联。

7. 根据权利要求 1 的方法，其中，对所述文本项进行排序的步骤进一步包括:

识别赋给各所述对的各所述单独文本项的各较低值项码值和较高值项码值;

识别所述文本项对的各个组，所述文本项对的组具有赋给各所述单独文本项的相同的所述较低值项码或赋给各所述单独文本项的相同的所述较高值项码中指定的一个;

通过以下步骤，确定各已识别的对的各自排序位置，

基于赋给各所述对的组的各单独文本项的各所述较低值项码或所述较高值项码中所述指定的一个，以指定的数字次序对所述各已识别的对的组进行排序; 以及

基于赋给各所述对的组中所述对的各文本项的所述较低值项码或所述

较高值项码中的另一个，以指定的数字次序对各所述对的组中的各对进行排序。

8. 根据权利要求 6 的方法，

其中，所述指定的一个为各所述较低值项码；以及

其中，所述另一个为所述较高值项码。

9. 根据权利要求 6 的方法，

其中，所述指定的一个为各所述较高值项码；以及

其中，所述另一个为所述较低值项码。

10. 根据权利要求 1 的方法，其中，所述对文本项对进行排序的步骤进一步包括:

识别所述文本项对的各个组，所述文本项对的组具有赋给各所述单独文本项的相同的所述较低值项码；并且

通过以下步骤，确定各已识别的对的各自排序位置，

基于赋给各所述对的组的各单独文本项的各所述较低值项码，以指定的数字次序对所述各已识别的对的组进行排序；以及

基于赋给各所述对的组中的所述对的各文本项的所述较高值项码，以指定的数字次序对各所述对的组中的各对进行排序;

以根据所述各对的各自排序位置的数字次序将各唯一整数对码值赋给所述各对;

其中，所述对排序信息结构存储各已识别的对与它们各自的所述已赋值的唯一对码值之间的各关联；并且进一步包括:

在所述计算机可读介质中提供项码信息结构，其将所述文本项与所述已赋值的唯一整数项码相关联。

11. 根据权利要求 1 的方法，其中，所述对文本项对进行排序的步骤进一步包括:

识别赋给各所述对的各所述单独文本项的各较低值项码值和较高值项码值;

识别所述文本项对的各个组，所述文本项对的组具有赋给各所述文本

项的相同的所述较低值项码或赋给各所述文本项的相同的较高值项码中指定的一个；以及

通过以下步骤，确定各已识别的对的各自排序位置，

将各唯一整数对码值赋给所述对，从而使得

如果所述指定的一个为相同的较低值项码，则

每一对具有已赋值的对码值，该对码值大于具有较小较低值项码的对的组的对码值，并且小于具有较大较低值项码的对的组的对码值；并且

每一对具有已赋值的对码值，该对码值大于具有较小较高值项码的对的组中的对的对码值，并且小于具有较大较高值项码的对的组中的对的对码值；并且

如果所述指定的一个为相同的较高值项码，则

每一对具有已赋值的对码值，该对码值小于具有较大较高值项码的对的组的对码值，并且大于具有较小较高值项码的对的组的对码值；并且

每一对具有已赋值的对码值，该对码值小于具有较大较低值项码的对的组中的对的对码值，并且大于具有较小较低值项码的对的组中的对的对码值。

12. 根据权利要求 1 的方法，其中，所述对文本项对进行排序的步骤进一步包括:

识别赋给各所述对的各所述单独文本项的各较低值项码值和较高值项码值;

识别赋给各所述对的各所述单独文本项的各较低值项码值和较高值项码值;

识别所述文本项对的各个组，所述文本项对的组具有赋给各所述文本项的相同的较低值项码;

将唯一的整数对码值赋给所述对，从而使得，

每一对具有已赋值的对码值，该对码值大于具有较小较低值项码的对的组的对码值，并且小于具有较大较低值项码的对的组的对码值；并且

每一对具有已赋值的对码值，该对码值大于具有较小较高值项码的对

的组中的对的对码值，并且小于具有较大较高值项码的对的组中的对的对码值。

13. 根据权利要求 1 的方法，其中，所述对文本项对进行排序的步骤进一步包括:

进行数学计算，基于赋给各所述对的各所述单独文本项的各所述唯一项码确定各对码的各唯一排序位置。

14. 根据权利要求 1 的方法，

其中，所述对文本项对进行排序的步骤进一步包括进行数学计算，基于赋给各所述对的各所述单独文本项的各所述唯一项码确定各对码的各唯一排序位置；并且

其中，所述对排序信息结构存储各所述已识别的对与其各自的所述已确定的唯一对码值之间的关联。

15. 根据权利要求 1 的方法，

其中，所述接收包括从计算机网络接收大量单独的所述文本项的组。

16. 一种基于计算机的组织文本项的方法，包括以下步骤:

接收大量单独的文本项的组;

将各个唯一的整数值项码赋给来自所述大量单独组的各所述单独文本项;

在计算机可读介质中提供项码信息结构，其将所述文本项与所述已赋值的唯一整数项码相关联;

识别来自所述大量组的所述单独组的文本项对;

通过以下步骤，确定各所述已识别的对的各自排序位置，

基于赋给各所述对的组的各所述单独文本项的各较低值项码或较高值项码中指定的一个，以指定的数字次序对各所述已识别的对的组进行排序;以及

基于赋给各所述对的组中的所述对的各所述文本项的所述较低值项码或所述较高值项码中指定的一个，以指定的数字次序对各所述对的组中的各所述对进行排序;

以根据所述各对的各自排序位置的数字次序将各唯一的整数对码值赋给所述各对；以及

在计算机可读介质中提供对排序信息结构，其存储各所述已识别的对与其各自的所述对码值之间的关联。

17. 一种基于计算机的组织文本项的方法，包括以下步骤:

接收大量单独的文本项的组;

将各个唯一的整数值项码赋给来自所述大量单独组的各所述单独文本项;

在计算机可读介质中提供项码信息结构，其将所述文本项与所述已赋值的唯一整数值项码相关联;

识别来自所述大量组中的单独组的文本项对;

通过进行数学计算对所述文本项对进行排序，此计算基于赋给各所述对的各所述单独文本项的各所述唯一项码来确定各对码的各自唯一排序位置；以及

在计算机可读介质中提供对排序信息结构，其存储各所述已识别的对与其各自的所述已确定的唯一对码值之间的关联。

18. 一种产品，其包括:

计算机可读介质，在其中编码有:

项计数信息结构，其存储大量项中的每一个在大量项的分组中的出现计数;

代码赋值信息结构，其将各所述项与各唯一整数项码相关联；以及

对计数信息结构，其存储各个对计数，此计数指示大量的项对中的每一个在所述大量项的分组的一个或多个中的出现次数;

其中，各所述对计数被分别存储在所述对计数信息结构中，由从所述项码信息结构中与各所述项对的各组成项相关联的各所述项码计算得到的各个对码来索引所存储的位置。

19. 一种产品，其包括:

计算机可读介质，在其中编码有:

项计数信息结构，其存储大量项中的每一个在大量的项的分组中的出现计数；

代码赋值信息结构，其存储各所述项与各唯一整数项码的关联；

用于执行对码计算处理的计算机程序代码，所述计算处理利用在所述代码赋值信息结构中关联的各所述唯一项码为所述大量分组的一个或多个中的各个项对计算各唯一整数值对码；

对计数信息结构，其存储各所述对码与大量项对中的每一个在所述大量项的分组的单独一个中的出现计数的关联。

20. 根据权利要求 **19** 的产品，

其中，所述项计数信息结构存储大量所述项中的每一个在所述大量项的分组的单独一个中的出现计数。

21. 根据权利要求 **19** 的产品，

其中，所述项计数信息结构包括项计数散列表；以及

其中，所述对计数信息结构包括对计数散列表。

22. 根据权利要求 **19** 的产品，

其中，所述项计数信息结构和所述代码赋值信息结构被结合在一种信息结构中。

23. 根据权利要求 **19** 的产品，

其中，所述对计数信息结构基于对码整数值以指定的数字次序存储各所述对码与各所述计数的关联。

24. 根据权利要求 **19** 的产品，

其中，所述对计数信息结构基于对码整数值以指定的顺序数字次序存储各所述对码与各所述计数的关联。

25. 根据权利要求 **19** 的产品，

其中，所述计算机可读介质进一步编码有：

关联性确定处理，用于通过将利用所述项计数信息结构与被选择的项对的项中的一个相关联的计数和利用所述对计数信息结构与所述被选择的对相关联的计数进行比较，确定来自所述大量的项的分组的所述被选择的

项对之间的关联性。

26. 一种确定文本项之间关联性的方法，其包括以下步骤：

在计算机可读介质中提供项计数信息结构，其存储大量的项中的每一个在大量的项的分组的单独一个中的出现计数；

在计算机可读介质中提供项码赋值信息结构，其存储各所述项与各唯一整数项码的关联；

在计算机可读介质中提供对计数信息结构，其存储各唯一整数值对码与大量的项对中的每一个在所述大量项的分组的一个或多个中的出现计数的关联；

指定包括两个文本项的项对；

利用所述项码赋值信息结构，为所述指定的对中的所述两个指定的文本项确定两个项码；

利用所述已确定的两个项码为所述指定的对中的所述指定的文本项对计算各自的唯一整数值对码；

通过利用所述两个已确定的项码中的至少一个搜索所述项计数信息结构，为所述指定对的所述两个指定项中的至少一个确定项计数；

通过利用所述计算得到的唯一整数值对码搜索所述对计数信息结构，确定对计数；以及

将所述已确定的至少一个项计数与所述已确定的对码计数进行比较。

27. 根据权利要求 **26** 的方法，

其中，所述对计数信息结构基于对码整数值以指定的顺序数字次序存储各所述唯一整数值对码与各所述计数的关联。

28. 根据权利要求 **26** 的方法，

其中，所述对计数信息结构基于对码整数值以指定的顺序数字次序存储各所述唯一整数值对码与各所述计数的关联；并且

其中，确定所述对计数涉及以指定的顺序数字次序扫描所述对计数信息结构的至少一部分，以寻找所述计算得到的唯一整数值对码的匹配。

29. 一种确定文本项之间关联性的方法，其包括以下步骤：

在计算机可读介质中提供项计数信息结构，其存储大量的项中的每一个在大量的项的分组的单独一个中的出现计数;

在计算机可读介质中提供项码赋值信息结构，其存储各所述项与各唯一整数项码的关联;

在计算机可读介质中提供对计数信息结构，其存储各唯一整数值对码与大量的项对中的每一个在所述大量项的分组的一个或多个中的出现计数的关联;

指定包括两个文本项的项对;

利用所述项码赋值信息结构，分别为每一个所述两个指定的文本项的对确定两个所述项码;

利用所述已确定的两个项码，分别为每一个所述指定的文本项对计算各自的唯一整数值对码;

通过利用每一个所述对的所述两个已指定的项中的一个搜索所述项计数信息结构，为该项确定项计数;

通过利用各所述计算得到的唯一整数值对码搜索所述对计数信息结构，分别确定各所述对计数;

通过分别将所述已确定的至少一个项计数与各所述对的各个所述已确定的对码计数进行比较，生成各个单独的所述关联性；以及

比较所述单独的关联性。

30. 根据权利要求 **29** 的方法，

其中，所述对计数信息结构基于对码整数值以指定的顺序数字次序存储各所述唯一整数值对码与各所述计数的关联。

31. 根据权利要求 **29** 的方法，

其中，所述对计数信息结构基于对码整数值以指定的顺序数字次序存储各所述唯一整数值对码与各所述计数的关联；并且

其中，分别确定各所述对计数涉及以所述指定的顺序数字次序扫描所述对计数信息结构中的至少一部分，以寻找各个所述计算得到的唯一整数值对码的匹配。

关联性分析的方法和产品

技术领域

本发明主要涉及信息分析，更具体地，涉及关于各对文本项之间的关系的信息的组织。

背景技术

关联性是不同项之间的关联的量度。人们可能希望知道项之间的关联性，以识别或更好地理解在项之间的可能的相关性或关系，所述项可以是，例如，事件、兴趣、人或产品。关联性可用来预测优先选择。例如，关联性可用来预测一个对某种事物感兴趣的人可能同样也对另一种事物感兴趣。具体而言，例如，关联性可用来预测一个购买了某本特定书的人可能也对购买一本或多本其它特定的书感兴趣，或者，一个常玩某种特定在线视频游戏的人可能对一种或多种其它的视频游戏感兴趣。

图 1 为计算机用户界面屏幕的说明示图，示出了一种假设的关联性分析结果。所述关联性分析结果显示了在三种汽车中的关联性，该三种汽车是：本田雅阁（Honda Accord Sedan）、丰田佳美（Toyota Camry）和福特金牛（Ford Taurus）。此例中，在所述关联性分析中的基础车辆为本田雅阁。其它车辆为丰田佳美和福特金牛。屏幕左方示出了用户控制键，其用于选择将要对其进行关联性分析的车辆。此例中，所述关联性分析中的所述基础车辆为本田雅阁。其它车辆为所述丰田佳美和所述福特金牛。分析的时间范围为 2002 年 12 月。屏幕中心部分的上部示出了文氏图形式的图形表示，其表示雅阁与佳美的关联性以及雅阁与金牛的关联性。雅阁圆与佳美圆的重叠度图形化地表示了雅阁与佳美的关联性。同样地，所述雅阁圆与金牛圆的重叠度图形化地表示了雅阁与金牛的关联性。所述重叠表

示关联性程度。屏幕中心部分的下部为表格，示出了所述三种汽车之间的关联性。所述数据图表的上面一行示出了雅阁与佳美的关联性强度的数量（23.7%）以及其与金牛的关联性强度的数量（3.1%）。中间一行示出了佳美与雅阁的关联性强度（30.6%）以及其与金牛的关联性强度（4.2%）。下面一行示出了金牛与雅阁的关联性强度（18.3%）以及其与佳美的关联性强度（19.2%）。所述屏幕的右部示出了一个表格，其依序列出了所述基础车辆相对于与之具有最强关联性的十五种车辆的关联性强度。在此例中，所述右部的表格还列出了所述基础车辆相对于在左边选中的用于关联性分析的每一种其它车辆（即，金牛，第 63 位）的关联性，即使该其它汽车不在关联性排在前十五位的车辆中。

关联性分析可用于对给定的关键字寻找相似的关键字。例如，以下列表为关键字的假设示例列表，可以通过假设的关联性分析发现这些关键字相似于关键字“007”。

007 相似性列表

jamesbond

james bond 007

007.com

007 bond

bond

bond 007

james bond, 007

bond james bond

james bond 007:nightfire

james bond movies

007 nightfire

bond james

bond, james

die another day

james bond website

007 games

james bond characters

james bond nightfire

Nightfire

agent 007

die another day movie

许多以上关键字甚至不包括项“007”，尽管也可以发现它们是与“007”相似的关键字。

关联性分析实际使用的一个例子是回答一般类型的问题，如果用户使用一定关键字在因特网上搜索，那么该用户还可能希望在因特网上搜索其它什么呢？关联性分析可用来回答这个问题。例如，所述分析可引起对关于一个确定关键字的关联性在前 10、100 或 1000 的其它关键字的有序表的识别。关联性分析还可以用来回答所述一般类型的问题，如果有人买花，那么这个人可能还想要买什么其它东西呢？例如，这些类型的问题在交叉销售以及在市场研究中很有用。

通常，可以至少部分地基于项在一个或多个项的分组中一起发生的频繁程度来确定各项之间的关联性。有很多种方法来定义项的分组。在计算机网络环境中可能出现的分组的例子与 IP 地址、处理标识（TID）、URL 或‘cookie’有关。

IP 地址可用来识别特定用户的计算机。TID 可用来识别特定处理，诸如，商品或服务的购买。例如，用户可以使用具有给定 IP 地址的计算机与因特网可访问服务器站点形成连接，然后在因特网上购买许多项。所述给定 IP 地址可作为包括了由所述用户一起购买的项的项分组的组标识（组 ID）。并且，所述购买处理可以具有 TID，其作为包含了所述已购买项或服务的分组的组 ID。

关键字的分组可以与 URL 相关联。所述 URL 可作为所述组 ID，而所述关键字可作为所述分组中的项。例如，可以通过保持对基于关键字的因

特网搜索的记录，而逐渐建立这样的关键字分组，在所述搜索中关键字用于识别一组 URL，然后用户选择一个或多个已识别的 URL 来访问因特网上的网页。可以逐渐开发出分组的数据库。所选中的 URL 作为组 ID，并且用来识别所述 URL 的关键字是组内的项。

因特网 cookie 用来创建分组。cookie 是一种一般机制，其中，服务器侧连接（例如，CGI 脚本）可用来存储和检索在连接的客户端侧的信息。CGI（公共网关接口）用来作为接口连接外部应用程序与信息服务器，该信息服务器例如 HTTP 或 web 服务器。所述简单、持久的客户侧状态的添加显著地扩展了基于 web 的客户/服务器应用程序的能力。当服务器将 HTTP 对象返回给客户时，也发送所述客户将存储的一条状态信息。所述状态对象中包括了对其而言该状态合法的 URL 的范围的描述。由所述客户作出的落入所述范围的任何将来的 HTTP 请求，将包括从所述客户向所述服务器发送回所述状态对象的当前值。所述状态对象被称为 cookie。计算机的 cookie 识别符可以作为组 ID，而利用所述 cookie 存储的信息作为分组中的项。

因特网已经创造了巨大的机会来收集对项之间关联性的研究有用的数据。可以开发包括了诸如基于 IP 地址、TID、URL 或 cookie 的那些分组的分组的巨大数据库。随着新分组信息的加入，这些数据库可以发展。

一种确定关联性的已知方法涉及基于所述项的出现次数以及项的分组的出现次数的计算。例如，根据此种方法，项 t1 对于项 t2 的关联性可以采用关于以下内容的信息:

N(t1): 包括 t1 的组 ID 的个数，

N(t2): 包括 t2 的组 ID 的个数，

N(t1,t2): 同时包括 t1 和 t2 的组 ID 的个数。

项 t1 对于项 t2 的关联性可以计算为，

N(t1,t2)/N(t1)

相反，

项 t2 对于项 t1 的关联性可以计算为，

N(t1,t2)/N(t2)

虽然进行关联性分析的早期方法通常已然不错，但是，它们在使用中存在缺点。例如，随着项的分组的数据库变得非常大，进行关联性分析所涉及的计算也会变得非常复杂。例如，给定的数据库可包含关键字的分组。每一个不同的关键字被认为是不同的项。每一个不同的分组可包括一个、两个或几个关键字。可能有几百万个分组和几百万个关键字。然而，以上计算关联性分析的方法一次仅考虑两项的关联性。记录任意给定的两个关键字在几百万分组中的同时出现的次数是非常复杂的任务，随着关键字数和分组数的随时间增加和改变，这变得更加困难。

因此，需要改进对用于关联性分析的项的组的组织。还需要改进项之间关联性的确定。本发明满足了这些需求。

发明内容

一方面，提供了基于计算机的组织文本项的方法。例如，所述方法在进行关联性分析时是有用的。提供了大量的文本项的单独组。将唯一的整数项码(item code)赋给单独文本项。识别来自单独组的文本项对(text item pair)。基于赋给所述对的组成文本项的所述唯一项码对所述文本项对进行排序。于是，各文本项对被赋予相对于其它文本项对的唯一的排序位置。在计算机可读介质中提供对排序信息结构，从而存储文本项对与其唯一的排序位置之间的关联。

在本发明的另一个方面，提供了一种产品，其包括计算机可读介质，此介质中编码有项计数信息结构、代码赋值信息结构以及对计数信息结构。所述项计数信息结构存储来自大量项组的大量项中的每一个的出现计数。所述代码赋值信息结构将各项与各个唯一的整数项码相关联。所述对计数信息结构存储各对的计数，此计数指示大量项对中的每一个在所述大量项组的一个或多个中的出现次数。对应于单独项对的对计数被存储在所述对计数信息结构中，此结构的位置由从在所述项码信息结构中与单独项对的组成项相关联的项码所计算得到的对码（pair code）来索引。

本发明的另一方面提供了一种产品，其包括计算机可读介质，此介质中编码有项计数信息结构、代码赋值信息结构、用于进行对码计算处理的计算机程序代码，以及对计数信息结构。所述项计数信息结构存储大量项组中的大量项中的每一个的出现计数。所述代码赋值信息结构存储项与唯一的整数项码之间的关联。所述计算机程序代码利用在所述代码赋值信息结构中相关的唯一项码为在所述大量分组的一个或多个中的项对计算唯一的整数值对码。所述对计数信息结构存储对码与在所述大量项组的一个或多个中的大量项对的每一个的出现次数的关联。

在本发明的另一方面，提供了改进的处理，以利用项码信息结构和对码信息结构进行关联性分析。

使用整数项码来表示文本项，连同使用整数对码来表示项对，使得可以改进对与文本项间关系的分析相关的信息的组织。更具体地，项码信息结构和对码信息结构在与文本项对之间的关系相关的大量信息的组织中尤其有用。从以下的详细描述和附图中，本发明的这些及其它的特点和优点将更加明显。

附图说明

图 1 是计算机用户接口屏幕的说明性示图，示出了假设的关联性分析结果;

图 2 是说明性的框图，示出了在其中应用了本发明原理的一种因特网环境;

图 3 是说明性的流程图，示出了这样的处理，其用于产生根据本发明实施例所用的计算机可读介质中的信息结构。

具体实施方式

提出以下描述是为了使得任何本领域技术人员都可以实施并使用所述发明，并且其配置于特定应用及其要求的语境下。对于本领域技术人员，对优选实施例的各种变型将十分明显，并且，无需脱离本发明的精神和范

围，在此定义的所述一般原理即可应用于其它实施例和应用。此外，在以下描述中，阐述了多个细节以用于解释的目的。然而，本领域普通技术人员应认识到无需使用这些具体细节即可实现本发明。在其它例子中，以框图形式所示的众所周知的结构和装置并非是以不必要的细节来模糊对本发明的描述。因此，本发明不限于所示实施例，而是依据与在此公开的原理和特点相符合的最广范围。

概述

在发明的当前实施例中，文本项被组织以用于关联性分析。关联性分析的目的在于基于这些分组确定文本项之间的关联性。所述发明的当前实施例通过提供一种对文本项对的新的组织来促进所述关联性分析，在此新的组织中，从项的分组中识别对，并且其中，基于单独对中的文本项的标识来系统地开发所述对的新的组织。这种文本项对的新的组织简化了随后的关联性分析。

根据本实施例来开发文本项对的新的组织方法涉及对大量项分组的单独项赋予唯一的整数值。这些唯一的整数值被称为项码。从所述大量分组中识别所述大量项组。基于所述文本项对的项码对其进行相互排序。具体地，相对于文本项的其它对来对文本项的单独对进行排序，从而使得每一个对具有相对于其它对的排序位置的唯一的排序位置。

提供于计算机可读介质中的新的对排序信息结构可存储在唯一的对排序位置与其它数据，诸如在给定数据库中所述对的出现次数的计数之间的关联。因此，在关联性分析时，项码可用于访问所述对排序信息结构。例如，在关联性分析时，给定项对的组成项的项码可以用来确定在所述对排序信息结构中所述给定对的唯一位置。这种唯一位置信息可用来定位信息，例如，通过所述对排序信息结构与所述给定对相关联的计数。

在本发明的一种实施例中，单独文本项被映射到项码，而文本项的单独项对被映射到称为对码的唯一整数值。基于所述单独对中的文本项的项码来确定所述单独对到单独对码的映射。单独对的所述排序位置取决于相

应的单独对码，而该对码又取决于所述单独对的组成项的项码。在关联性分析中，给定项对的组成项的项码可用来确定所述给定对的唯一对码。这种唯一位置信息可用来定位信息，例如通过所述对排序信息结构与所述给定对相关联的计数。

文本项和分组

如在此所使用的，文本项可包括一组一个或多个字符，诸如，字母、数字、符号或其组合。所述字符可以具有单词或短语的意义，但是其本身并不需要具有任何特殊的意义。通常，根据一些规则将所述文本项编成大量分组，这不构成本发明的部分。例如，可以基于所述文本项与相同 IP（因特网协议）地址、处理标识（TID）、URL 或 cookie 的关联将它们一起分组。例如，用户可以与特定站点建立因特网连接，并进行在线购买处理，在其中该用户购买名为“Encyclopaedia（百科全书）”的书，标为“Popular Songs（流行歌曲）”的 CD，以及参加名为“Luxury Voyage（豪华旅行）”的旅行的票。用于此在线购买的处理标识将与该三个文本项相关联。一个为文本串“Encyclopaedia”。另一个为文本串“Popular Songs”。再一个为文本串“Luxury Voyage”。

文本代码的赋值

以下为根据本发明的实施例对项分组内的项进行项码赋值的例子。以下说明性的分组将用于该例子。

G1={x，y，z}

G2={x，y}

G3={x，z}

尽管在实际实现中可能有大量分组，甚至是几百万个，然而，为解释清楚起见，在此例中仅使用三个分组。G1、G2 和 G3 是这三个说明性的分组的组标识。可以理解，G1、G2 和 G3 可能为不同的 IP 地址，TID、URL、cookie 或一些其它形式的分组标识。并且，甚至可以是混合类型的分组，

例如，G1 标示 IP 地址，G2 标示 TID，而 G3 标示 URL。项 x 为文本项，是 G1、G2 和 G3 的组成元素。项 y 为文本项，是 G1、G2 和 G3 的组成元素。项 z 为文本项，仅为 G1 的组成元素。

将整数值项码选择性地赋给所述 G1、G2 和 G3 的文本项。例如，可对项 x 赋值 1；对项 y 赋值 2；而对项 z 赋值 3。顺序地进行整数项码赋值。此外，可采用项阈值处理来选择进行项码赋值的项。例如，可利用这样的项阈值条件，其要求项至少在一些指定的最小阈值数量的分组内存在，从而具有项码赋值的资格。这样的项阈值处理为可选的优化，其目的在于保证仅将项码赋给在所述大量分组中具有指定的使用级别的项。在此例中，如果所述项阈值设置为二，则项 x 和 y 可接受项码，而项 z 则不能。而如果将所述项阈值设置为三，则仅有项 x 可接受项码。

生成项码信息结构，以将文本项与唯一整数项码相关联。假设所述项阈值设置为一，则在此例中项的可能的项码信息结构可如下表所示。

项码信息结构

文本项	项码
项 x	1
项 y	2
项 z	3

所述项码信息结构提供了从选中的文本项到赋值的项码的映射。以上表格仅为用来将项映射到项码的一种类型的结构的一个例子。所述项码信息结构可存储在计算机可读存储介质中。

识别项对

以下为根据本发明的一种实施例在项的分组内识别项对的例子。在此例中将使用以上示出的说明性的分组。在一种实施例中，仅对通过项阈值处理的项来识别对。如果所述项阈值设置为一，则对 G1、G2 和 G3 可识别出识别的对：（x，y），（x，z），（y，z）。如果所述项阈值设置为二，则所述已识别的对将为（x，y）。

根据本发明的实施例，依照其组成项的所述项码来表示所述项对。如果所述项阈值设置为一，并且所述已赋值的项码为 x=1、y=2、z=3，则将所述项对表示为（1，2），（1，3），（2，3）。如果所述项阈值设置为二，且所述已赋值的项码为 x=1、y=2，则将所述项对将被表示为（1，2）。

于是，项码信息结构将文本项与项码相关联。这些项码用于表示项对。如下解释，项对的所述唯一项码被用来排序所述项对，从而使得各对具有相对于其它对的唯一的排序位置。在一种实施例中，单独对的组成项的唯一项码被用来计算单独的唯一对码，其指示所述单独对的单独的唯一排序位置。

排序文本项对

以下图表说明了基于组成的文本项的对码进行对排序的例子，从而使得每个对具有相对于其它对的排序位置的唯一排序位置。这些图表中的每一个示出了基于所述对中的组成项的项码所进行的项对的不同的可能排序。为了解释的简洁和清楚，这些例子各自仅包括六个项对。

这些图表说明了项对的可选排序。每一个图表说明了基于所述对的较高或较低值项码中的一个进行的项对的分组。具体地，图表的每一行保持不同的项码组，其基于所述较高或较低值项码中的一个被分组。每一个图表还说明了已分组的项对的指定排序。各图表还说明了在项对组内的对的指定排序。

例如，参照图表 1，在单独图表位置的左侧的整数值指示此位置处的项对的排序位置。例如，对（1，2）在排序位置“1”，而对（2，4）在排序位置“5”。顶行具有一个对的组，其具有等于整数 1 的较低值项码。第二行具有一个对的组，其具有等于整数 2 的较低值项码。第三行具有一个对的组，其具有等于整数 3 的较低值项码。

图表 1 说明了项对的排序，在其中，基于较低值项码对项进行分组。例如，在所述项码对（1，2）中，1 为较低值项码，而 2 为较高值项码。

图表 1 说明了组的排序，在其中，具有较小的较低值项码的组排在具

有较大的较低值项码的组之前（从上到下读行）。于是，组对{（1，2），（1，3），（1，4）}排在组对{（2，3），（2，4）}之前（上）。类似地，组对{（2，3），（2，4）}排在组对{（3，4）}之前（上）。

图表 1 说明了组内的对的排序，在其中，具有较小的较高值项码的对排在具有较大的较高值项码的对之前（从右到左读列）。于是，组对{（1，2），（1，3），（1，4）}被排序为，对（1，2）排在第一，接着（1，3）排在第二，再接着（1，4）排在第三。

图表 1

较低值组/较小的较低值优先的组间排序/较小的较高值优先的组内排序

1X	1（1，2）	2（1，3）	3（1，4）
2X	X	4（2，3）	5（2，4）
3X	X	X	6（3，4）
4X	X	X	X

图表 2 说明了项对的排序，在其中，基于较低值项码对项进行分组。所述组被排序，使得具有较小的较低值项码的组排在具有较大的较低值项码的组之前（上）。组内的对被排序，使得具有较大的较高值项码的对排在具有较小的较高值项码的对之前。于是，组对{（1，4），（1，3），（1，2）}的排序为，对（1，4）排在第一，接着（1，3）排在第二，再接着（1，2）排在第三。

图表 2

较低值组/较小的较低值优先的组间排序/较大的较高值优先的组内排序

1X	1（1，4）	2（1，3）	3（1，2）
2X	X	4（2，4）	5（2，3）
3X	X	X	6（3，4）
4X	X	X	X

图表 3 说明了项对的排序，在其中，基于较高值项码对项进行分组。

所述组被排序，使得具有较大的较高值项码的组排在具有较小的较高值项码的组之前（上）。于是，例如，组{（1，4），（2，4），（3，4）}排在组{（1，3），（2，3）}之前。组内的对被排序，使得具有较小的较低值项码的对排在具有较大的较低值项码的对之前。于是，例如，组对{（1，4），（2，4），（3，4）}的排序为，对（1，4）排在第一，接着（2，4）排在第二，再接着（3，4）排在第三。

图表3

较高值组/较大的较高值优先的组间排序/较小的较低值优先的组内排序

1X	1（1，4）	2（2，4）	3（3，4）
2X	X	4（1，3）	5（2，3）
3X	X	X	6（1，2）
4X	X	X	X

图表4说明了项对的排序，在其中，基于较高值项码对项进行分组。所述组被排序，使得具有较大的较高值项码的组排在具有较小的较高值项码的组之前（上）。于是，例如，组{（3，4），（2，4），（1，4）}排在组{（2，3），（1，3）}之前。组内的对被排序，使得具有较大的较低值项码的对排在具有较小的较低值项码的对之前（从左向右读）。于是，例如，组对{（3，4），（2，4），（1，4）}的排序为，对（3，4）排在第一，接着（2，4）排在第二，再接着（1，4）排在第三。

图表4

较高值组/较大的较高值优先的组间排序/较大的较低值优先的组内排序

1X	1（3，4）	2（2，4）	3（1，4）
2X	X	4（2，3）	5（1，3）
3X	X	X	6（1，2）
4X	X	X	X

这些图表仅说明了四种可能的系统的方法，以根据本发明的原理，基于项码来对项对进行排序。将项码赋给项使得项对可以由其组成项的项码

来表示。以上图表说明了与单独项关联的项码的对可以用来对所述对进行确定地排序，从而使得每个对具有相对于其它对的唯一排序位置。重要的是，各单独对具有唯一的排序位置，基于赋给组成所述对的项的唯一项码的对来确定此位置。

排序位置的计算

通过数学计算来计算排序位置。以下处理包括用来计算在图表 1 中说明的项对排序位置的数学计算。基于赋给单独对的组成项的项码，确定该单独对的单独排序位置。

以下处理可以利用编码在计算机可读介质中的计算机程序代码来实现。所述处理根据本发明的实施例对给定的任意项对（t1，t2）计算唯一的整数对码值。假设已对给定项对的每一项赋予项码。再假设所述对中的项和赋予那些项的项码之间的关联已存储在项码信息结构中。还假设 MAX 为赋值给任何项的最大项码。

初始步骤是为所述给定项对产生已排序的项码对，其中较低值项码排在第一位，而较高值项码排在第二位。于是，对于项对（t1，t2），从项码信息结构检索 t1 的项码和 t2 的项码。假设对于所述给定的项码对（t1，t2），code1 为已赋予文本项 t1 的项码，而 code2 为已赋予文本项 t2 的项码。

根据所述对排序处理，

If (code2 = MAX), then paircode (t1, t2) = code1*MAX - SUM(x) where x=1 to code1,

Else pair (t1, t2) = (code2 - code1) +(code1-1)*MAX- SUM(x)，where x=1 to code1-1。

对于图表 1 中对码集合的例子，MAX = 4。

以下是为图表 1 中的所述项码对的代表性抽样计算唯一对码和相应的唯一排序位置的例子。

对于所述项码对（1，2），code2 = 2。因此，对于（1，2），code2

≠MAX。从而，paircode (1,2)= (2-1)+ (1-1)x 4 - (0)=1。

对于所述项码对（2，3），code2 = 3。因此，对于（2，3），code2 ≠MAX。从而，paircode (2,3)= (3-2)+ (2-1)x 4 - (1)=4。

对于所述项码对（2，4），code2 = 4。因此，对于（2，4），code2=MAX。从而，paircode (2,4)= (2 x 4)- (1+2)= 5。

此计算处理可用于建造有关项对的信息的数据库。此相同的计算处理可用来访问信息的数据库，以检索关于所述项对的已存储信息。在数据库建造期间，通过所述计算处理计算得到的对码可用来确定在存储介质中的位置，在该处存储与单独项对相关联的信息。此后，假设单独项对的信息实际上已存储在由其对码确定的存储位置中，相同的计算处理可用来为给定的项对计算对码，从而定位并从所述存储介质中检索关于所述项对的信息。当然，应该认识到，信息的数据库可以不断地建造和修改。因此，延续的建造和检索可以同时进行。

此外，此计算处理可用来生成可以被高效搜索的项对信息数据库。如上所解释的，唯一的对码可以表示项对相关信息的唯一排序位置。依照本发明的一个方面，基于所述计算处理确定对排序位置，并且将项对相关信息以排序位置的顺序存储在计算机可读介质中。于是，可以使用线性扫描型处理更方便地定位已存储的项对相关信息。在信息检索期间，利用所述计算处理计算将被访问以检索项对相关信息的位置。如果已将项对信息按照由所述计算处理确定的唯一位置顺序进行存储，则所述计算处理可以用来计算在所述存储介质中将被访问的位置的线性序列。

例如，参考图表 1 的所述假设例，所述顶行的项码与对码之间的相关性如下。

项码对 （1，2）→ 对码 =1

项码对 （1，3）→ 对码 =2

项码对 （1，4）→ 对码 =3

对码与存储地址位置之间相关性的假设例如下。

对码 =1 → 存储位置 1000

对码 =2 → 存储位置 1001

对码 =3 → 存储位置 1002

因此，可以认识到，可以通过对所述存储介质的线性扫描搜索来访问与项码=1 的项相关联的所有对相关信息，所述扫描搜索从位置 1000 开始，然后是 1001，最后在 1002 结束。

以上各示例图表仅包括六个项码对以及六个相应的对码。在实际实现中，可能有几百万个项和几百万个对。由于所述用于排序项对的处理可延展（scaleable），相同的基本对排序处理可用于为大量对确定项对排序位置和项对码。实际上，所述处理在进行对排序和实现高效信息存储策略中的优势随着项以及对的数量增加，而变得更为明显。

尽管图表 1-4 提供了四种可能的根据本发明的原理的对排序的例子，但是，本领域技术人员应该认识到，在本发明的范围内可以有其它排序方法。此外，尽管描述了一种具体的计算处理以用来为图表 1 的所述示例排序计算对码，可以相信，本领域技术人员可以容易地知道，类似的基本计算原理也可应用于示例图表 2-4 所示类型的对排序的计算处理。

组织关联性分析信息

图 2 为说明性框图，其示出了一种因特网计算语境，在其中可以应用本发明的原理。服务器系统 20 通过因特网 30 与大量与因特网连接的计算机设备 22、24、26 和 28 进行通信。例如，所述服务器 20 可以是因特网门户，如 yahoo 站点：www.yahoo.com，并具有 yahoo 的所有特性。用户从计算机（22-28）进入此站点，以获取各种信息和服务，例如，搜索、邮件财务等。在此处理中收集的数据可用于进行关联性分析。例如，所述计算机设备 22-28 可以是任何用来登录到所述计算机并从服务器 20 在因特网上访问信息的用户设备。

所述服务器 20 通过在所述因特网 30 上与计算机 22-28 的交互来收集文本项的分组。应该认识到，尽管在图 2 中仅示出了四个代表性的计算机 22-28，所述服务器 20 每天与几百万的计算机进行通信。例如，计算机 22-28

可以产生各种对于信息和服务的请求。这些请求可以涉及如前解释的文本项的分组。所述服务器可以处理这些分组，从而产生对于关联性分析有用的信息的数据库。

从用户计算机 22-28 接收的文本项的分组组成了输入数据，该输入数据被处理以生成对关联性分析有用的信息的数据库。识别分组中的项。将项码赋给所述项。例如，单独的项码可以赋给以下的各文本项。

Honda

Honda Motor

Honda Motor Company

如上解释的，使用阈值处理去除那些使用很少的项，无需将项码赋给这些项。在计算机可读介质中生成项码信息结构，所述介质存储了项与其被赋予的项码之间的关联。

此外，识别所述输入数据中的分组中的项对。为已选中的项对计算对码，相应的项码对被赋值给该项对。以上参照图表 1 描述的计算处理可以用于从项码计算对码。

图 3 为说明性流程图，其示出了处理 38，所述处理利用项码和对码生成用于关联性分析的计算机可读介质中的信息结构。在步骤 40 中，提供了项计数散列结构，在其中项名称被映射到码和计数。在此阶段应用阈值，从而仅将阈值以上的项放入此散列结构。在本实施例中，所述项计数散列结构包括具有项码和项计数的散列表。在步骤 42 中，提供了对计数散列结构，在其中对码被映射到项对计数。在本实施例中，所述对计数散列结构包括具有对码和对计数的散列表。

在步骤 44 中，扫描所述输入数据。对于每个已识别的项的分组，确定所述组中的项是否存在项码。对于已经存在项码的项，增加所述项码计数散列结构中的相应项码计数。在本实施例中，忽略不存在项码的项，因为它们低于阈值，对分析影响很小。

在步骤 46 中，扫描所述输入数据。对每个已识别的项的分组，为所述分组中的每一项识别项码，并为所述分组中已识别的各项对计算对码。诸

如以上参照图表 1 所描述的计算处理可被用于计算对码。

对所述输入数据中的每个分组，如下所述更新所述对计数散列结构。在步骤 48 中，确定是否已将所述对码输入所述对计数散列结构。如果给定项对的对码已经存在于所述对计数散列结构中，则在步骤 50 中，将对应于所述先存在的对码的计数增加一。如果并非如此，则在步骤 52 中，确定在所述对计数散列结构中的条目的数目是否小于所允许的条目的最大数目 MAXIMUM。如果其小于 MAXIMUM，则在步骤 54 中，将所述新的对码加入所述结构，并且相关计数=1。如果所述条目的数目等于（或超出）MAXIMUM，则在步骤 56 中，按照对码将所述对计数散列结构中的所有条目排列到对计数中间信息结构。清空所述对计数散列结构，并将所述新的对码与相关计数=1 的条目一起加入所述新近清空的结构。

重复以上处理，直到处理完所述输入数据中的所有分组。当完成了对所述输入数据的扫描，将所述项计数散列结构中的所有项相关的信息写入项计数信息结构。类似地，当完成了对所述输入数据的扫描，将所述对计数散列结构中的所有对相关的信息写入对计数中间信息结构。然后，将所有对计数中间信息结构合并在一起，并将具有阈值以上计数的对以对码排序写入对计数信息结构。

以下表格为根据图 3 的处理生成的项码结构、项计数信息结构和对计数信息结构的例子。这些结构中包含的信息被编码进由服务器 20 访问的计算机可读介质。应该认识到，这些结构仅为用于解释目的而生成的假设的例子。此外，为了解释的简洁清楚，在这些结构中仅列出了很少的项和对。然而，本领域技术人员可以认识到，相同的原理可应用于大量的项和对。

项码结构

文本项	项码
A	1
B	2
C	3
D	4

项计数信息结构

项码	项计数
1	10
2	20
3	15
4	30

对计数信息结构

对码	相关项码	对码计数
1	（1，2）	5
2	（1，3）	6
3	（1，4）	7
4	（2，3）	4
5	（2，4）	5
6	（3，4）	4

关联性分析举例

以上例子中的所述信息结构可被用于进行关联性分析。

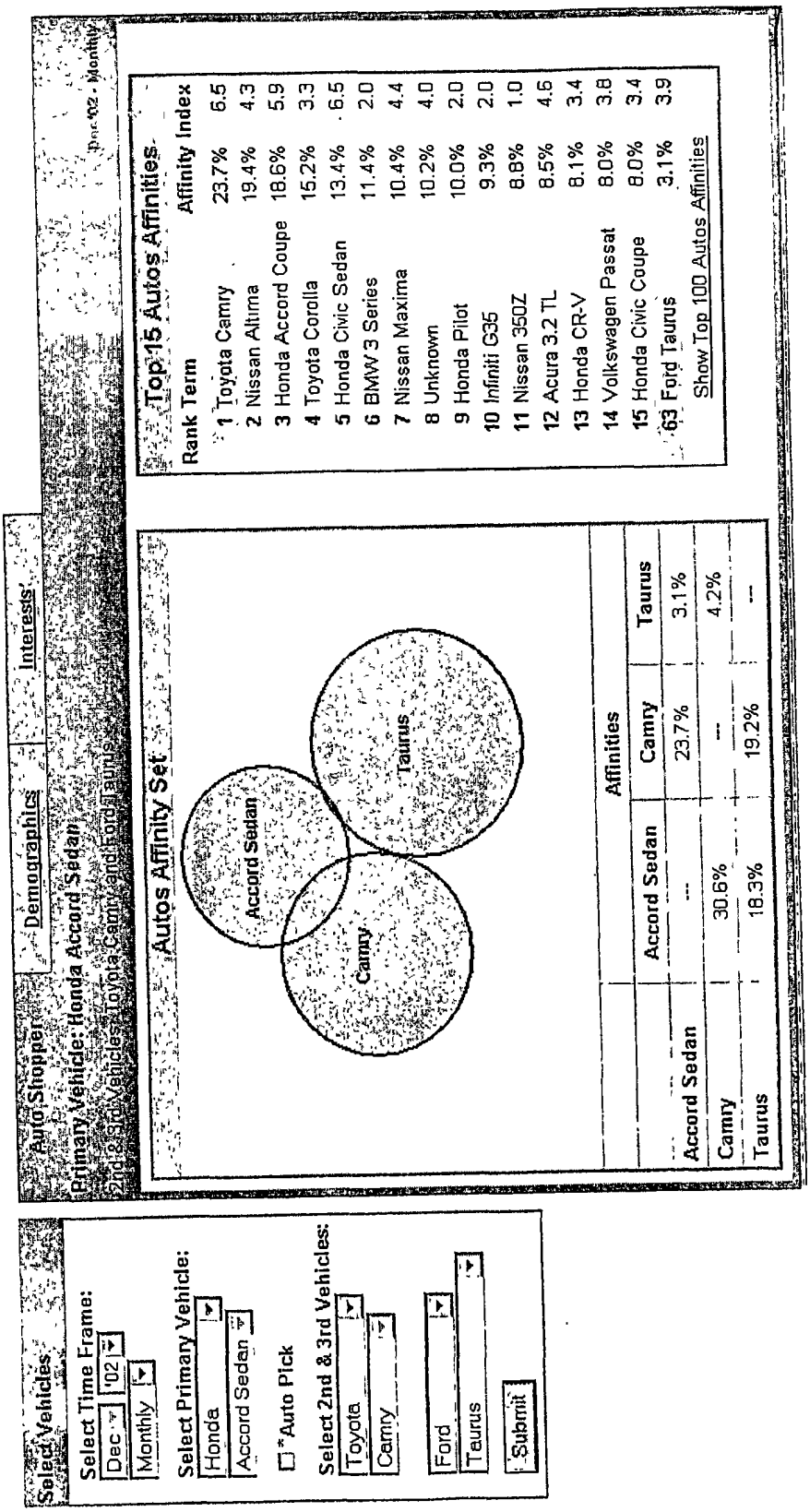
例如，利用这些结构，可以如下确定文本项 A 对于文本项 B 的关联性。为了此例的目的，A 对 B 的所述关联性可定义为 count(A,B)/countA。

从所述项码信息结构检索文本项 A 和 B 的项码。A 的项码为 1。B 的项码为 2。将所述项码作为在所述项计数信息结构中的索引，并检索 A 的项计数。A 的项计数为 10。利用 A 和 B 的项码对，即，项码对（1，2），通过进行计算处理来计算所述项对（A，B）的对码。在此例中，所述计算处理产生对码 1。将计算得到的对码 1 作为在所述对计数信息结构中的索引，并检索对码 1 的对计数。检索得到的对计数为 5，这意味着（A，B）的对计数为 5。因此，A 对于 B 的关联性为 5/10=0.50。

反之，例如，利用这些结构，可以通过相同的程序来确定文本项 B 对于文本项 A 的关联性。为了此例的目的，B 对 A 的关联性可定义为 count(A,B)/countB。以上示例结构可被用于确定 countB=20，且 count(A,B)=5。因此，B 对于 A 的关联性为，5/20=0.25。

本领域技术人员可认识到，可以通过分析得到关联性信息。例如，对诸如 A 对 B 和 B 对 A 的关联性进行比较可以决定什么文本项最有意义。例如，关联性信息可被用于获得按关联性排序的与给定项有高关联性的项的列表。例如，图 1 中右侧的表格中有与所述基础型号具有关联性的汽车型号的列表。在该例子中，对于所有对计算关联性，在其中，基础型号为所述项中的一个。然后，利用关联性对这些项排序。使用本发明的当前实施方式，可以在一次扫描中完成整个分析。

应该认识到，根据本发明的优选实施例的以上描述和图表仅仅是对本发明原理的说明。无需脱离本发明的精神和范围，本领域技术人员即可进行各种修改。


Select Vehicles
Select Time Frame:
Dec
'02
Monthly
Select Primary Vehicle:
Honda
Accord Sedan
*Auto Pick
Select 2nd & 3rd Vehicles:
Toyota
Camry
Ford
Taurus
Submit
Auto Shopper
Demographics
Interests
Primary Vehicle: Honda Accord Sedan
2nd & 3rd Vehicles: Toyota Camry and Ford Taurus
Autos Affinity Set
Accord Sedan
Camry
Taurus
Affinities
Accord Sedan
Camry
Taurus
Accord Sedan

23.7%
3.1%
Camry
30.6%

4.2%
Taurus
18.3%
19.2%

Top 15 Autos Affinities
Rank Term
Affinity Index
1 Toyota Camry 23.7% 6.5
2 Nissan Altima 19.4% 4.3
3 Honda Accord Coupe 18.6% 5.9
4 Toyota Corolla 15.2% 3.3
5 Honda Civic Sedan 13.4% 6.5
6 BMW 3 Series 11.4% 2.0
7 Nissan Maxima 10.4% 4.4
8 Unknown 10.2% 4.0
9 Honda Pilot 10.0% 2.0
10 Infiniti G35 9.3% 2.0
11 Nissan 350Z 8.8% 1.0
12 Acura 3.2 TL 8.5% 4.6
13 Honda CR-V 8.1% 3.4
14 Volkswagen Passat 8.0% 3.8
15 Honda Civic Coupe 8.0% 3.4
63 Ford Taurus 3.1% 3.9
Show Top 100 Autos Affinities

图 1

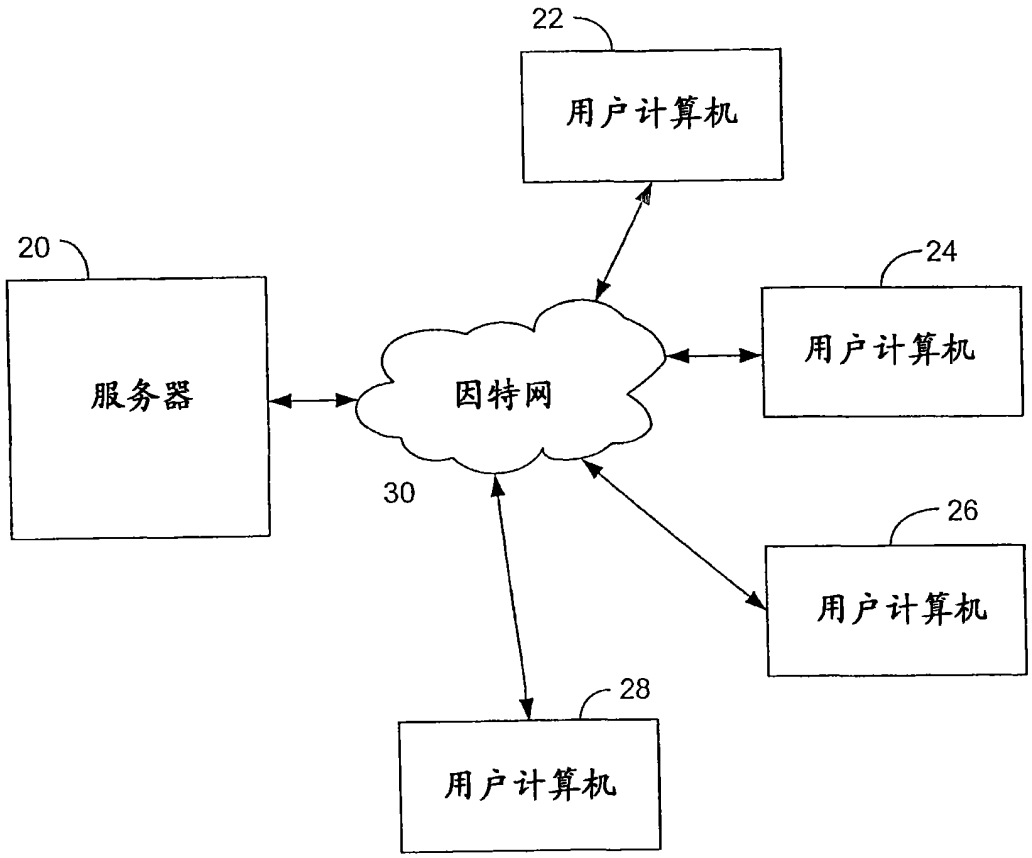


图 2

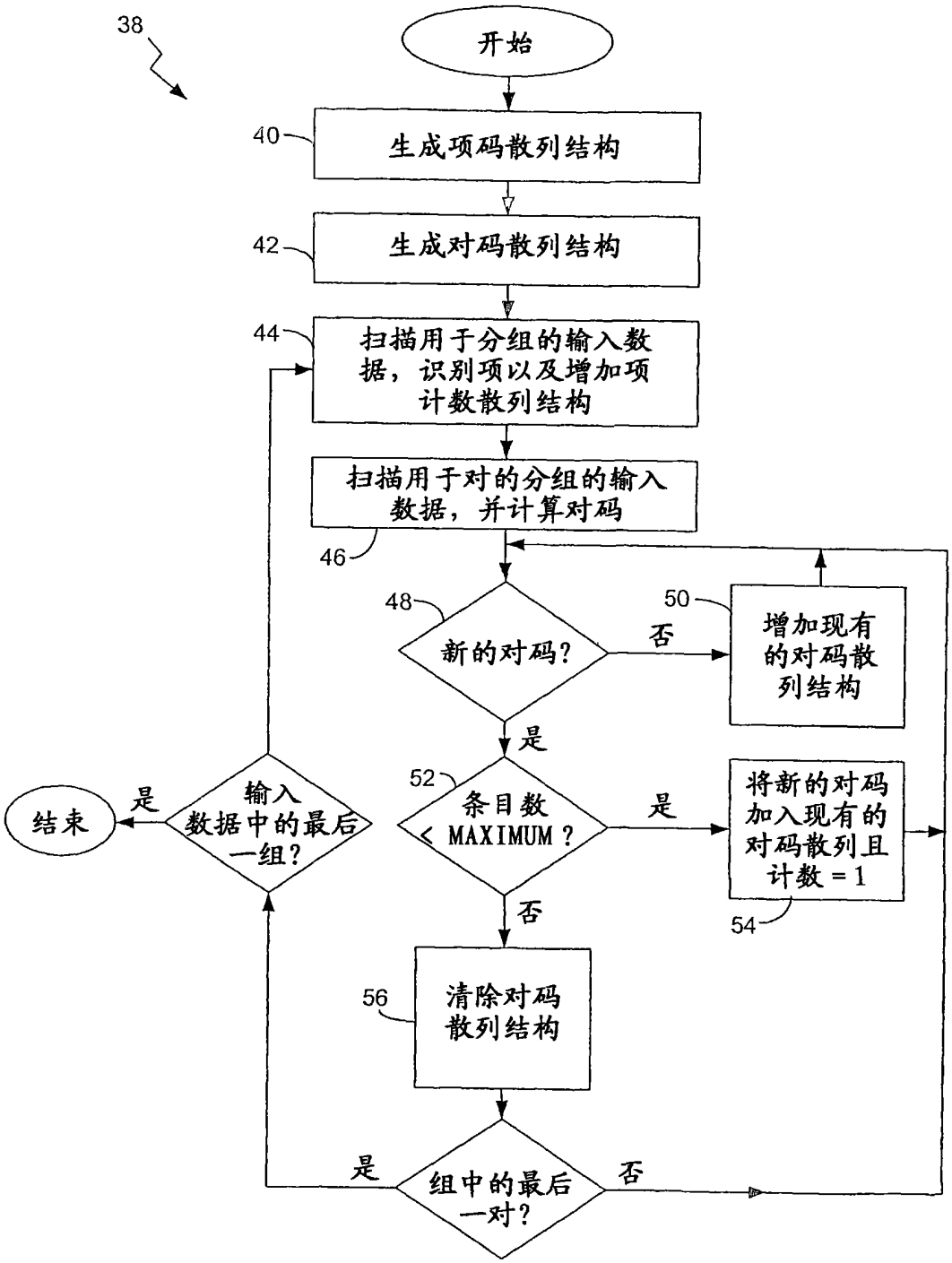

38
开始
40
生成项码散列结构
42
生成对码散列结构
44
扫描用于分组的输入数据，识别项以及增加项计数散列结构
扫描用于对的分组的输入数据，并计算对码
46
48
新的对码?
否
50
增加现有的对码散列结构
是
52
条目数 < MAXIMUM ?
是
将新的对码加入现有的对码散列且计数 = 1
54
否
输入数据中的最后一组?
是
结束
56
清除对码散列结构
组中的最后一对?
是
否

图 3