[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G06F 7/00 (2006.01)
G06F 17/30 (2006.01)

[12] 发明专利申请公布说明书

[21] 申请号 200680014980.4

[43] 公开日 2008 年 4 月 30 日

[11] 公开号 CN 101171568A

[22] 申请日 2006.4.21
[21] 申请号 200680014980.4
[30] 优先权
[32] 2005. 5. 2 [33] US [31] 11/119,667
[86] 国际申请 PCT/US2006/014975 2006.4.21
[87] 国际公布 WO2006/118814 英 2006.11.9
[85] 进入国家阶段日期 2007.11.1
[71] 申请人 微软公司
地址 美国华盛顿州
[72] 发明人 S·智恩 N·伊莫里卡
[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 3 页 说明书 18 页 附图 8 页

[54] 发明名称

用于查找语义相关的搜索引擎询问的方法

[57] 摘要

一种基于输入询问和存储询问的瞬态相关确定该输入询问和该存储询问之间语义相关的方法。 各实施例包括对输入询问和存储询问之间相关系数的形式计算。 可选实施例包括在牺牲少许或者不牺牲相关系数精确性的情况下使用简化数据模型和高效数据检查来计算该相关系数的方法。

1．一种用于确定两个搜索引擎询问之间语义关系的方法，包括如下步骤：

（a）归一化关于所述两个搜索引擎询问的询问数据；以及

（b）基于在所述步骤（a）中归一化的所述询问数据来识别所述两个询问之间的语义相关。

2．如权利要求 1 所述的用于确定两个搜索引擎询问之间语义关系的方法，其特征在于，关于所述两个搜索引擎询问的所述数据是与所述两个搜索引擎询问的所述频率有关的瞬态数据。

3．如权利要求 1 所述的用于确定两个搜索引擎询问之间语义关系的方法，其特征在于，归一化询问频率数据的所述步骤（a）包括测量在一离散时间单位内与所有其他搜索引擎询问的组合的瞬态询问频率数据相比的有关每个搜索引擎询问的瞬态询问频率数据的步骤。

4．如权利要求 1 所述的用于确定两个搜索引擎询问之间语义关系的方法，其特征在于，所述两个搜索引擎询问的第一询问是输入询问，而所述两个搜索引擎询问的第二询问是存储询问，本方法还包括在确定所述输入询问和所述存储询问之间存在高于阈值水平的语义关系的情况下向用户或广告商建议所述存储询问的步骤。

5．如权利要求 1 所述的用于确定两个搜索引擎询问之间语义关系的方法，其特征在于，识别所述两个询问之间语义相关的所述步骤（b）包括用一位串表示所述两个搜索引擎询问频率的所述数据，并且检查表示所述两个搜索引擎询问频率的所述位串内各对应位彼此匹配的程度的步骤。

6．如权利要求 1 所述的用于确定两个搜索引擎询问之间语义关系的方法，其特征在于，识别所述两个询问之间语义相关的所述步骤（b）包括计算表示所述第一询问的所述归一化频率的第一向量与表示所述第二询问的所述归一化频率的第二向量的点积的步骤。

7．一种用于确定两个搜索引擎询问之间语义关系的方法，包括如下步骤：

（a）用位串表示所述两个搜索引擎询问的每一个询问；以及

（b）基于表示所述两个搜索引擎询问的所述位串内各对应位彼此匹配的程度识别所述两个搜索引擎询问之间的语义相关。

8．如权利要求 7 所述的用于确定两个搜索引擎询问之间语义关系的方法，其

特征在于，关于所述两个搜索引擎询问的所述数据是与所述两个搜索引擎询问的所述频率有关的瞬态数据。

9．如权利要求 7 所述的用于确定两个搜索引擎询问之间语义关系的方法，其特征在于，用位串表示所述两个搜索引擎询问的每一个询问的所述步骤（a）包括用 128 位的位串表示所述两个搜索引擎询问的每一个询问的步骤。

10．如权利要求 7 所述的用于确定两个搜索引擎询问之间语义关系的方法，其特征在于，所述两个搜索引擎询问的询问 q 是输入询问，而所述两个搜索引擎询问的询问 p 是存储询问。

11．如权利要求 10 所述的用于确定两个搜索引擎询问之间语义关系的方法，其特征在于，还包括在确定所述输入询问和所述存储询问之间存在高于阈值水平的语义关系的情况下向用户建议所述存储询问的步骤。

12．如权利要求 10 所述的用于确定两个搜索引擎询问之间语义关系的方法，其特征在于，还包括计算所述两个搜索引擎询问之间的相关系数的步骤，所述相关系数指示所述两个搜索引擎询问之间的所述语义关系的程度。

13．如权利要求 7 所述的用于确定两个搜索引擎询问之间语义关系的方法，其特征在于，用位串表示所述两个搜索引擎询问的每一个询问的所述步骤（a）包括生成多个超平面，并且在用于所述两个搜索引擎询问的第一个询问的所述位串中生成一位，所述位由与所述超平面之一正交的第一向量与表示所述第一搜索引擎询问的第二向量的点积来确定的步骤。

14．如权利要求 7 所述的用于确定两个搜索引擎询问之间语义关系的方法，其特征在于，还包括如下步骤：

（c）扫描所述位串的每一位串中所述各位的一部分，以识别在所述第一和第二位串的所述部分内各对应位的某些部分是否彼此匹配；以及

（d）如果在所述步骤（c）中确定在所述第一和第二位串的所述部分内各对应位的某些部分彼此匹配，则对应地扫描所述位串内的剩余各位。

15．一种具有计算机可执行指令的计算机可读介质，所述计算机可执行指令用于执行的步骤包括：

（a）归一化关于多个搜索引擎询问中的每一询问的询问频率数据；

（b）将在所述步骤（a）中归一化的所述询问数据表示为相应的位串，所述多个搜索引擎询问中的每一个有一个位串；以及

（c）从所述多个搜索引擎询问中选择一个或多个搜索引擎询问，所选择搜索

引擎询问具有高于所述多个搜索引擎询问的基准搜索引擎询问的第一阈值水平的相关度。

16. 如权利要求 15 所述的具有计算机可执行指令的计算机可读介质，其特征在于，所述选择步骤（c）包括如下步骤：

（i）扫描与所述多个搜索引擎询问相关联的所述位串的每一位串中所述各位的一部分，以识别具有在所述位串的所述部分内各对应位的某些部分彼此匹配的一个或多个搜索引擎询问；以及

（ii）扫描所述的一个或多个搜索引擎询问以识别具有高于所述基准搜索引擎询问的所述第一阈值水平的相关度的所述一个或多个搜索引擎询问。

17. 如权利要求 16 所述的具有计算机可执行指令的计算机可读介质，其特征在于，扫描与所述多个搜索引擎询问相关联的所述位串的每一位串中所述各位的一部分的所述步骤（i）包括扫描所述位串的每一位串中的 20 位的步骤。

18. 如权利要求 15 所述的具有计算机可执行指令的计算机可读介质，其特征在于，将在所述步骤（a）中归一化的所述询问数据表示为相应的位串的所述步骤（b）包括生成多个超平面，并且在用于所述多个搜索引擎询问的第一个询问的所述位串中生成一位，所述位由与所述超平面之一正交的第一向量与表示所述第一搜索引擎询问的第二向量的点积所确定的步骤。

19. 如权利要求 15 所述的具有计算机可执行指令的计算机可读介质，其特征在于，还包括向客户建议所述一个或多个搜索询问的步骤。

20. 如权利要求 15 所述的具有计算机可执行指令的计算机可读介质，其特征在于，还包括计算所述基准搜索引擎询问与所述一个或多个搜索引擎询问之间的相关系数的步骤，所述相关系数指示所述两个搜索引擎询问之间的所述语义关系的程度。

用于查找语义相关的搜索引擎询问的方法

发明人

Steve Chien

Nicole lmmorlica

发明背景

发明领域

本发明涉及用于查找语义相关的搜索引擎询问的方法。

相关领域的描述

在线搜索引擎为以结构化且有区分的方案访问因特网上可用的海量信息提供了一个强有力的工具。诸如 MSN®、Google®和 Yahoo!®之类的流行搜索引擎每天为上千万的信息询问服务。典型的搜索引擎由一组协作的程序操作，这些程序包括聚集来自万维网上的各网页的信息以创建用于搜索引擎索引的条目的网络蜘蛛(spider)（也被称为“网络爬虫(crawler)”或“bot”）；从已被阅读的文件中创建该索引的索引程序；以及接收搜索询问，将其与索引上的各项条目相比较并返回适于该搜索询问的结果的搜索程序。

当前在搜索引擎技术领域中的一个重要研究方向是如何改善给定搜索询问结果的效率和质量。所谓的基于概念的搜索涉及对各种搜索准则进行统计学分析以识别并建议与输入搜索询问高度语义相关的可选搜索询问。识别可选的、高度相关的搜索询问有助于集中并改善给定搜索的搜索结果。此外，公司和广告商会在输入特定询问的情况下呈现广告。这将非常有利于这些公司和广告商把它们的广告与特定的询问以及其他语义相关的询问相关联。

在利用基于概念的现有技术的搜索系统示例中，取决于各询问内返回结果相同程度而将各询问相关在一起。于是，如果第一和第二询问返回几乎相同的搜索结果，则可以认为这两个询问彼此高度相关。基于概念的搜索的一个示例在 H. Daume 和 E. Brill 为 Human Language Technology Conference / North American Chapter of

the Association for Computational Linguistics (HTUNAACL), Boston, MA (2004)发布的题为“Web Search Intent Induction via Automatic Query Reformulation”的论文中有所阐述。

基于概念的搜索的另一个示例检查 click-through 数据作为相关搜索询问的指示符。这一模型观察来自不同搜索询问结果的 clicked-on 的链接。如果两个不同的询问导致用户点击相同的 URL，则可认为这两个询问高度相关。click-through 的基于概念的搜索的一个示例在 D. Beeferman 和 A. Berger 为 Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, MA (2000)发布的题为“Agglomerative Clustering of a Search Engine Query Log”的论文中有所阐述。

另一种有前途的基于语义的搜索技术涉及分析输入询问本身以揭示特定时序上的模式、趋势及周期性。例如，Vlachos, M.、Meek, C、Vagena, Z.和 Gunopulos, D.为 International Conference on Management of Data (SIGMOD), Paris, France (2004)发布的题为“Identifying Similarities, Periodicities and Bursts for Online Search Queries”的论文（Vlachos 等人），该论文全文结合在此作为参考。Vlachos 等人注意到不同的事件具有不同的瞬态搜索频率。例如，询问“电影院”的频率在每周末有一峰值，而询问“复活节”的频率在每年春天形成一单峰并在随后突然下降。瞬态相关后面的理论是如果两个搜索询问呈现足够类似的瞬态模式，则它们很有可能是语义相关的。Vlachos 等人使用存储在与搜索引擎（在他们的研究中是 MSN®）相关联的一个或多个服务器上的询问日志来为每个实际询问建立时间序列，在其中该时间序列的各元素是在给定的一天内询问被搜索的次数。

使用傅立叶分析，Vlachos 等人通过傅立叶系数表现出了询问频率随时间变化的瞬态周期性，并在随后应用时间序列匹配技术来识别带有极类似瞬态模式的其他询问。他们利用的匹配技术基于傅里叶系数之间的欧几里德距离（Euclidean distance）来测量瞬态相似性。在此框架下，他们描述了一种使用有关每个询问的若干最佳傅立叶系数查找给定询问的最类似询问的方法。

有关搜索引擎询问的瞬态模式随时间变化。例如，搜索量在日间大于夜间，以及搜索量在平日大于周末。试图识别语义询问匹配的各模型可能会虚假地认为在两个搜索之间具有高度相关，这是因为没有将这一自然的随时间变化考虑在内。例如，图 1 是在一个月的时间段内每天测量的两个不同搜索询问的出现次数的采样图。第一曲线 n_a 是第一询问出现的次数而第二曲线 n_b 是第二询问出现的次数。如

图所示，两根曲线都示出了周末的出现次数相对于平日有所降低。在用于查找语义相关的搜索询问的典型瞬态模型下，这种类似的周末降低活动会导致这两个询问之间虚假的高度语义相关，但事实上这一相关其实是由于搜索询问随时间的自然变化所致。

发明概述

本发明的各实施例提供一种基于输入询问和存储询问的瞬态相关来确定它们之间语义相关的方法。各实施例包括输入询问和存储询问之间相关系数的形式计算。可选实施例包括在牺牲少许或者不牺牲相关系数精确性的情况下使用简化数据模型和有效数据检查来计算相关系数的方法。

一询问 q 随 d 时间单位变化的频率函数是 d 维向量 $Y_q = (Y_{q,1},...,Y_{q,d})$。在每个离散时间单位处的询问频率 $Y_{q,i}$ 是询问 q 在第 i 个时间单位期间的归一化频率。使用归一化频率（即，在给定时间单位期间询问 q 的频率与接收到的所有其他频率相比较）来归一化地去除询问流量随时间自然变化的影响。

形式进程存储有关所有时间单位及有关接收到的所有询问的归一化询问频率函数 $Y_{q,i}$。使用有关每个询问的归一化频率函数的平均值和标准差，就可以确定任何两个询问之间的相关系数。高于阈值的两个询问间的相关系数被认为有意义，并且指示了询问之间的瞬态和语义相关。

使用用于计算相关系数的该形式方法，带有 n 个询问和 d 个时间单位的询问流需要为每个询问存储 d 个频率值，即总共 dn 个的归一化频率函数值，并且还需要在每次期望确定所有存储的询问与输入询问的相关系数时线性传递所有这些数据。一个可选实施例利用来自嵌入和最近领域算法中的技术，允许用更少的数据表示每个询问的归一化频率函数。使用该简化数据模型的相关系数的计算还能够在更短的处理时间内并在能够实现相关系数实时计算的情况下提供对输入询问和其他询问之间相关系数的精确估计。

该简化数据模型通过为 d 维$\mathbb{R}^d$的欧几里德实空间内所有存储询问的归一化频率函数定义的单个询问频率向量$\tilde{Y}_q$而使用简化数据来表示询问频率函数。每个询问频率向量都可用作将每个询问频率向量$\tilde{Y}_q$表示为δ位串$v(\tilde{Y}_q)$的映射$v(\cdot): \mathbb{R}^d \rightarrow \{0,1\}^\delta$。用于位串或嵌入$v(\tilde{Y}_q)$内$\delta$位的每一位的值（1 或 0）可以使用利用嵌入理论的随机超平面来确定。对于某一输入询问 q 而言，位串$v(\tilde{Y}_q)$第 j 位的值由询问频率向量$\tilde{Y}_q$位于第 j 个随机选定超平面的哪一侧所支配，其中该超平面可由与该随机

生成的超平面正交的向量 r_j 表示。生成的超平面个数 δ 大到足以提供超平面的近似均匀分布以生成精确表示每个询问频率函数的嵌入 $v(\tilde{Y}_q)$。

在简化数据模型中，在输入询问 q 和任何存储询问 p 之间的相关系数 λ 可由嵌入 $v(\tilde{Y}_q)$和 $v(\tilde{Y}_p)$之间从第 1 位到第 δ 位对应相同位数的分数所近似。对应位的匹配数越大，询问 q 和 p 之间相关系数就越高，并且各询问也更为瞬态相关。

一旦生成并存储了所有询问频率函数的表示，就可以确定一输入询问与所有这些存储的询问之间的相关。虽然可能对所有存储数据进行线性传递，但是本发明的各实施例利用一种用缩短的处理时间找出相关询问的方案。这可以通过将检查的询问数限制为那些可能具有高于预定相关系数阈值的询问。

在这一实施例中，每一个存储询问 p 基于其嵌入 $v(\tilde{Y}_p)$的前 k 位被组织成缓冲区位置，在此被称为存储段（bucket）。其后，一旦接收到输入询问 q，该处理系统就检查在含有 q 的位串 $v(\tilde{Y}_q)$的存储段内的那些嵌入 $v(\tilde{Y}_q)$，以及那些与含有 q 的位串的存储段有预定匹配位数的存储段。

附图简述

图 1 是在用于查找语义相关搜索询问的常规系统中经过一个月长的时期所测得的两个不同询问的频率的曲线图。

图 2 是适于实现本发明实施例的计算机硬件的框图。

图 3 是能够由图 2 所示的计算机硬件操作和/或在其中操作的包括有代码和数据结构的搜索引擎的框图。

图 4 是根据本发明各实施例在一系统中经过多个离散时间单位测得的归一化询问涉及的曲线图。

图 5 是根据本发明各实施例用于在一系统中生成并存储归一化询问频率的一进程的流程图。

图 6 是用于从在图 4 内生成并存储的归一化询问频率中计算相关系数的一进程的流程图。

图 7 是归一化且成比例的向量 $\tilde{Y}_q$ 和 $\tilde{Y}_p$ 连同随机向量 r_i 的高斯分布采样的图示。

图 8 是基于各向量与随机向量 r_i 的关系从图 6 的归一化且成比例的向量至嵌入 $v(\tilde{Y}_q)$和 $v(\tilde{Y}_p)$的映射。

图 9 是根据本发明实施例用于确定两询问的相关系数的数据简化模型的流程图。

图 10 是根据本发明实施例用于识别相关询问的形式进程。

图 11 是根据本发明实施例示出用于识别相关询问的缩短时间模型的流程图。

详细描述

如下将参考图 2 至图 11 描述本发明的各实施例，各实施例一般地涉及用于查找语义相关的搜索引擎询问的方法。根据本发明各实施例的方法至少部分基于如果询问瞬态相关则这些询问就语义相关的观点。也就是说，如果不同搜索询问的流行度趋向于一起随时间上升和下降，这些询问就有可能彼此语义相关。因此本发明的各实施例识别与给定搜索询问瞬态相关的询问。

在各实施例中，本发明分析了在一具体时刻处一具体询问的密度，而非该询问的频率。密度是一询问频率在该具体时刻与所有其他频率的比较。该方法归一化地去除了由搜索引擎量自然变化引起的量改变所引入的差错。此外，比较频率函数的相关，而不是协方差，从而更好地捕捉虚假肯定。此外，为了让算法实用，使用近似算法来查找近似的相关，而非实际的相关。该近似算法显著降低了所需的存储空间，以及要求的处理时间，以便在少量牺牲或不牺牲各相关精确度的情况下查找语义相关的询问。

在此描述的方法可以在各种处理系统上实现。图 2 示出了可在其上实现本发明的合适的计算系统环境 100 的示例。计算系统环境 100 只是合适的计算环境的一个示例，并不旨在对本发明的使用范围或功能提出任何限制。也不应该把计算环境 100 解释为对示例性操作环境 100 中示出的任一组件或其组合有任何依赖性或要求。

本发明可用众多其它通用或专用计算系统环境或配置来操作。适合在本发明中使用的公知的计算系统、环境和/或配置的示例包括，但不限于，个人计算机、服务器计算机、多处理器系统、基于微处理器的系统、机顶盒、可编程消费者电子产品、网络 PC、小型机、大型机、膝上型和掌上计算机、手持设备、包含上述系统或设备中的任一个的分布式计算机环境等。

本发明可在诸如程序模块等由计算机执行的计算机可执行指令的通用语境中描述。一般而言，程序模块包括例程、程序、对象、组件、数据结构等，它们执行特定任务或实现特定抽象数据类型。本发明也可以在分布式计算环境中实现，其中任务由通过通信网络连接的远程处理设备执行。在分布式计算环境中，程序模块可以位于包括存储器存储设备在内的本地和远程计算机存储介质中。

参考图 2，用于实现本发明的一个示例性系统包括计算机 110 形式的通用计算设备。计算机 110 可以包括但不限于，处理单元 120、系统存储器 130 和将包括系统存储器在内的各种系统组件耦合至处理单元 120 的系统总线 121。系统总线 121 可以是若干类型的总线结构中的任一种，包括存储器总线或存储器控制器、外围总线和使用各种总线体系结构中的任一种的局部总线。作为示例，而非限制，这样的体系结构包括工业标准体系结构（ISA）总线、微通道体系结构（MCA）总线、扩展的 ISA（EISA）总线、视频电子技术标准协会（VESA）局部总线和外围部件互连（PCI）总线（也被称为 Mezzanine 总线）。

计算机 110 通常包括各种计算机可读介质。计算机可读介质可以是能够被计算机 110 访问的任何可用介质，且包括易失性和非易失性介质、可移动和不可移动介质。作为示例，而非限制，计算机可读介质可以包括计算机存储介质和通信介质。计算机存储介质包括以任何方法或技术实现的用于存储诸如计算机可读指令、数据结构、程序模块或其它数据等信息的易失性和非易失性、可移动和不可移动介质。计算机存储介质包括，但不限于，RAM、ROM、EEPROM、闪存或其它存储器技术、CD-ROM、数字多功能盘（DVD）或其它光盘存储、磁带盒、磁带、磁盘存储或其它磁性存储设备、或能用于存储所需信息且可以由计算机 110 访问的任何其它介质。通信介质通常具体化为诸如载波或其它传输机制等已调制数据信号中的计算机可读指令、数据结构、程序模块或其它数据，且包含任何信息传递介质。术语“已调制数据信号”指的是这样一种信号，其一个或多个特征以在信号中编码信息的方式被设定或更改。作为示例，而非限制，通信介质包括诸如有线网络或直接线连接的有线介质，以及诸如声学、RF、红外线和其它无线介质的无线介质。上述中任一个的组合也应包括在计算机可读介质的范围之内。

系统存储器 130 包括易失性或非易失性存储器形式的计算机存储介质，诸如只读存储器（ROM）131 和随机存取存储器（RAM）132。基本输入/输出系统 133（BIOS）包含有助于诸如启动时在计算机 110 中的元件之间传递信息的基本例程，它通常存储在 ROM 131 中。RAM 132 通常包含处理单元 120 可以立即访问和/或目前正在操作的数据和/或程序模块。作为示例而非限制，图 2 示出了操作系统 134、应用程序 135、其它程序模块 136 和程序数据 137。

计算机 110 也可以包括其它可移动/不可移动、易失性/非易失性计算机存储介质。仅作为示例，图 2 示出了从不可移动、非易失性磁介质中读取或向其写入的硬盘驱动器 141，从可移动、非易失性磁盘 152 中读取或向其写入的磁盘驱动器 151，

以及从诸如 CD ROM 或其它光学介质等可移动、非易失性光盘 156 中读取或向其写入的光盘驱动器 155。可以在示例性操作环境下使用的其它可移动/不可移动、易失性/非易失性计算机存储介质包括，但不限于，盒式磁带、闪存卡、数字多功能盘、数字录像带、固态 RAM、固态 ROM 等。硬盘驱动器 141 通常由诸如接口 140 等不可移动存储器接口连接至系统总线 121，磁盘驱动器 151 和光盘驱动器 155 通常由诸如接口 150 等可移动存储器接口连接至系统总线 121。

以上描述和在图 2 中示出的驱动器及其相关联的计算机存储介质为计算机 110 提供了对计算机可读指令、数据结构、程序模块和其它数据的存储。例如，在图 2 中，硬盘驱动器 141 被示为存储操作系统 144、应用程序 145、其它程序模块 146 和程序数据 147。注意，这些组件可以与操作系统 134、应用程序 135、其它程序模块 136 和程序数据 137 相同或不同。操作系统 144、应用程序 145、其它程序模块 146 和程序数据 147 在这里被标注了不同的标号是为了说明至少它们是不同的副本。用户可以通过输入设备（诸如键盘 162）和定点设备 161（诸如鼠标、跟踪球或触摸垫）向计算机 110 输入命令和信息。其它输入设备（未示出）可以包括麦克风、操纵杆、游戏垫、圆盘式卫星天线、扫描仪等。这些和其它输入设备通常由耦合至系统总线的用户输入接口 160 连接至处理单元 120，但也可以由其它接口或总线结构，诸如并行端口、游戏端口或通用串行总线（USB）连接。监视器 191 或其它类型的显示设备也经由接口，诸如视频接口 190 连接至系统总线 121。除监视器以外，计算机也可以包括其它外围输出设备，诸如扬声器 197 和打印机 196，它们可以通过输出外围接口 195 连接。

计算机 110 可使用至一个或多个远程计算机，诸如远程计算机 180 的逻辑连接在网络化环境下操作。远程计算机 180 可以是个人计算机、手持式设备、服务器、路由器、网络 PC、对等设备或其它常见的网络节点，且通常包括上文相对于计算机 110 描述的许多或所有元件。图 1 中所示逻辑连接包括局域网（LAN）171 和广域网（WAN）173，但也可以包括其它网络。这样的连网环境在办公室、企业范围计算机网络、内联网和因特网中是常见的。

当在 LAN 联网环境中使用时，计算机 110 通过网络接口或适配器 170 连接至 LAN 171。当在 WAN 联网环境中使用时，计算机 110 通常包括调制解调器 172 或用于通过诸如因特网等 WAN 173 建立通信的其它装置。调制解调器 172 可以是内置或外置的，它可以通过用户输入接口 160 或其它合适的机制连接至系统总线 121。在网络化环境中，相对于计算机 110 描述的程序模块或其部分可以存储在远程存储

器存储设备中。作为示例，而非限制，图 2 示出了远程应用程序 185 驻留在存储器设备 181 上。可以理解，所示的网络连接是示例性的，且可以使用在计算机之间建立通信链路的其它手段。

图 3 是其上可以实现本发明的包括软件模块和数据结构的搜索处理环境 300。该搜索处理环境 300 可以用上述计算系统环境 100 操作和/或作为其的一部分。搜索处理环境 300 可以是基于三种主要元件的基于网络爬虫的系统。第一个是网络蜘蛛，也被称为网络爬虫 302。该网络蜘蛛访问网页 390a、390b，读取这些网页，并在随后追寻至该站点内的其他页面的链接。网络蜘蛛定期返回该站点以查找变化。由任何网络爬虫执行的基础算法重复地采用一个种子 URL 列表作为输入： 从该 URL 列表中移去一 URL，确定其主机名的 IP 地址，下载相应的文档并且提取其内含有的任何链接。对于每个提取的链接，将其翻译成绝对 URL（如果需要的话），并且倘若以前尚未遇见，则将其添加至 URL 列表以供下载。如果期望的话，也可以按其他方式处理该下载文档（例如，索引其内容）。

网络蜘蛛找出的每件事物都进入搜索引擎的第二部分，即索引 306。有时被称为目录的索引 306 是含有网络蜘蛛找出的每张网页副本的储存库。如果网页改变，随后就使用新信息更新该工作簿。该索引存储在数据存储 310 内。

搜索处理环境 300 的第三部分是搜索引擎 312。搜索引擎 312 是一程序，它筛选记录在该索引内的数以百万计的页面，以找出其认为最相关的匹配搜索并排列搜索结果。搜索一索引包括用户建立一询问并通过搜索引擎将其提交。该询问可以很简单，最少一个单词，或者可以是一系列的词或短语。建立更为复杂的询问可通过使用允许用户精细化和扩展搜索项的布尔操作符而得以实现。

实践中，计算设备 325 的用户经由客户机侧的网络浏览器 316 和主机侧的网络服务器 314 访问搜索处理环境 300。一旦在客户机和主机之间建立通信链路，计算设备 325 的用户就可如上所述执行询问搜索。

根据本发明的各实施例，搜索处理环境 300 还可以包括相关引擎 350。相关引擎 350 是能够识别与计算设备 325 的用户输入的询问语义相关的询问，并且按将在下文中解释的方式将这些语义相关的询问呈现给用户的软件模块。相关引擎与搜索引擎 312 通信以接收输入询问。该相关引擎还与保存在数据存储 310 内的询问数据库通信。询问数据库 352 存储关于过去询问的所有原始频率数据，以及将在下文中解释的可由本发明使用的归一化询问数据和其他参数。相关引擎还可以与网络服务器通信以便将识别出的语义相关搜索询问呈现给设备 325 的用户，或者帮助其重新

定义搜索结果。

图 4 是例如由设备 325 的用户键入的可能输入搜索询问 q 的归一化询问频率函数 Y_q。归一化询问频率是经总数为 d 个的离散时间单位而测得的。询问 q 经 d 个时间单位的频率函数是 d 维向量 $Y_q = (Y_{q,1},...,Y_{q,d})$。询问 q 可以是使用在线搜索引擎（例如，MSN®搜索以及上述其他搜索引擎）搜索的任何询问。在此使用的输入询问及其表示通常被称为字母“q”而被存储的询问及其表示则通常被称为字母“p”。根据本发明的各实施例，某一时间单位内的频率 $Y_{q,i}$ 可以被看成随机变量，并且两个询问之间的相似性可以由将在下文中解释的其频率函数的相关系数给出。

在每个离散时间单位处的询问频率 $Y_{q,i}$ 是询问 q 在第 i 个时间单位期间的归一化频率。也就是说，归一化频率 $Y_{q,i}$ 并不简单地是询问 q 在第 i 个时间单位期间被搜索的次数。相反地，归一化频率 $Y_{q,i}$ 如下给出：

$$Y_{q,i} = \frac{n_{q,i}}{N_i},$$

其中 $n_{q,i}$ 是询问 q 在第 i 个时间单位期间出现的次数而 N_i 是在第 i 个时间单位期间的询问总数。正如在本发明的背景部分所指出的那样，询问流量随时间固有地变化。询问频率的上述定义是该询问的密度，即在给定的时间段内与所有其他的询问相比，询问 q 每隔多久被键入一次。这样就归一化地去除了询问流量随数据固有变化的影响，并且避免了某些询问对显示出虚假的高度相关，这些高度相关仅仅是因为每时间单位的询问总数比某些时期的要大，即白天与晚间相比以及平日与周末相比。为所有询问使用归一化频率就能使得本发明的各实施例更为精确地计算两个询问之间的真实相关性。

如下将参考图 5 解释一种用于计算输入询问 q 和存储询问 p 之间相关系数的形式进程。为了确定输入询问和其他询问之间的相关性，经多个离散时间单位测得的归一化询问数据的表示被存储在询问数据库 352 内。在一个实施例中，归一化询问数据例如可以在一年的时间里每一天获取一次。尽管如此，独立的离散时间单位在各个不同实施例中可以是秒、分、时、天、周、月或年。而在其他的可选实施例中，独立的离散时间单位还可以短于一秒，长于一年，或者可以是其间的任何离散时间段。考虑的这些离散时间单位的总数在各个不同实施例中可以总计为一小时、一天、一周、一个月、一年或十年。而在其他的可选实施例中，考虑的这些离散时

间单位的总数可以总计为短于一小时，长于十年，或者可以是其间的任何时间段。

在一个实施例中，如图 5 的流程图所示，保存在询问数据库 352 内的归一化询问数据的表示可以实时生成；也就是说在询问数据库内的各项在这些离散时间单位的总数时间内（例如，一年内）可以每离散时间单位更新一次（例如，每天一次）。在这一实施例中，在步骤 200 中检索来自数据库 352 的有关给定时间单位 i 的所有搜索询问的原始数据日志。该日志通常为了某一时间段搜索引擎 312 接收的所有询问而被保存。在步骤 202 中，相关引擎 350 可以从有关该离散时间单位 i 的搜索询问日志中确定该时间单位期间每个询问的密度 $Y_{q,i}$。如上所述，在一时间单位期间的每个询问的密度 $Y_{q,i}$ 如下给出：

$$Y_{q,i} = \frac{n_{q,i}}{N_i},$$

其中 $n_{q,i}$ 是询问 q 在第 i 个时间单位期间出现的次数而 N_i 是在第 i 个时间单位期间的询问总数。

在本发明的一个实施例中，在步骤 204 中，可以将一离散时间单位期间的所有询问的密度存储在数据库 352 内。如下将解释，存储该数据的步骤在执行本发明的应用的可选实施例中可以被省略。

仍然参见图 4 和图 5，在步骤 206 中计算并存储所有时间段 i 的询问频率的运行计数(running count)。该值用于为所有询问计算归一化询问频率 Y_q 的平均频率 $\mu(Y_q)$。该平均频率 $\mu(Y_q)$ 如下给出：

$$\mu(Y_q) = \frac{1}{d}\sum_{i-1}^{d} Y_{q,i},$$

其中 i 表示每个离散时间单位而 d 表示这一情况下时间单位 i 的总数。

有关所有时间段的询问频率的运行计数还可用于为所有询问计算归一化询问频率 Y_q 的标准差 $\sigma(Y_q)$。该标准差 $\sigma(Y_q)$ 如下给出：

$$\sigma(Y_q) = \sqrt{\frac{\sum_{i=1}^{d}(Y_{q,i} - \mu(Y_q))^2}{d}}。$$

在步骤 210 中，如果已收集到期望数目时间单位的所有数据，随后该数据收集进程结束。否则在步骤 212 中，相关引擎 350 等待下一个时间单位 i 经过，并且相关引擎返回步骤 200 以便计算并存储在下一个时间单位内获取的数据。图 5 描述了数据的实时收集。应该理解各实施例也可以代替地使用存储在数据库 352 上的询问频率的过去日志以生成并存储上述有关所有询问的归一化询问频率。

使用如图 5 所示在每个离散时间单位内为每个询问生成并存储的归一化询问频率，就可以计算任何两个询问 p 和 q 之间的相关系数λ，以指示询问 p 和 q 之间的瞬态相关。于是，在步骤 214 中（图 6）询问 q 被输入搜索引擎的情况下，就可以在步骤 216 中计算在该询问 q 和任何其他被存储的询问 p 之间的瞬态相关。询问 p 和 q 之间的相关系数如下给出：

$$\lambda_{p,q}=\sum_{i=1}^{d}\left(\frac{Y_{q,i}-\mu(Y_q)}{\sigma(Y_q)}\right)\left(\frac{Y_{p,i}-\mu(Y_p)}{\sigma(Y_p)}\right)$$

相关系数λ的范围在 1 至-1 之间。高度瞬态相关的询问 q 和 p 可以具有很高的λ值，完全相关的询问 q 和 p 所具有的相关系数为 1。逆相关的询问 p 和 q 可以具有负相关系数，而不相关的询问 p 和 q 具有的相关系数为 0。

如上所述，使用归一化频率函数来代替每个询问的单纯频率归一化地去除了询问流量随时间固有变动的影响，并且避免了某些询问对显示出虚假的高度相关，这些高度相关仅仅是因为每时间单位的询问总数比某些时期的要大。此外，以上述方式测量相关系数使用频率函数的相关，而这与另一种常规的相似度测量，即协方差相反。协方差无法归一化频率函数的变动，于是就使得大幅变动的询问不真实地呈现出与许多其他询问的瞬态相关。

如图 5 和图 6 所阐明的那样，任何两个询问之间的相关系数都可由相关引擎 350 使用上述等式连同为归一化询问 $Y_{q,i}$ 和 $Y_{p,i}$ 所存储的数据，以及与每个 Y_q 和 Y_p 有关的平均频率和标准差来计算。然而，对于带有 n 个询问和 d 个时间单位的询问流而言，这就在每次期望确定有关一输入询问的相关系数时要求存储 dn 个归一化频率函数并且还需要线性传递所有这些数据。

虽然本方法可用于本发明的各实施例，但是一可选方案，即简化的数据模型利用来自嵌入和最近领域算法中的技术，从而使得处理单元 120 能够用少得多的数据和小得多的存储空间为每个询问表示归一化频率函数。这一可选方法还允许相关

引擎 350 以更短的处理时间并能够实时计算相关系数的情况下从该简化数据中精确地估计输入询问和所有其他询问之间的相关系数。

如下将参考图 7 至图 9 描述根据本发明的数据简化方法。一般说来，该简化数据模型包括表示为例如 128 位的存储串的所有时间单位内的每个询问频率函数。任何两个询问频率函数之间的相关系数随后就可通过表示两个询问频率函数的位串内对应位彼此匹配的程度来给出。

如下将阐明该简化数据模型更为严格的数学解释，之后则是对在搜索处理环境 300 中软件执行的解释。嵌入理论在例如 P. lndyk 和 R. Motwani 为 the Thirtieth Annual ACM Symposium on the Theory of Computing, Dallas, Texas (1998)发表的题为“Approximate nearest neighbors: Towards removing the curse of dimensionality”的论文中另有描述，而最近邻域算法在例如 W. Johnson 和 J. Lindenstrauss 在 Contemporary Mathematics, 26:189-206 (1984)上发表的题为“Extensions of Lipschitz Maps Into a Hilbert Space”的论文中有所描述，这两篇论文都全文结合在此作为参考。

参考图 7 至图 9 描述的简化数据模型利用在第 i 个时间单位内给出的归一化输入询问频率能够被映射至成比例且归一化的向量 $\tilde{Y}_{q,i}$ 并可由该向量所描述，给出的该向量为：

$$\tilde{Y}_{q,i} = \frac{1}{\sqrt{n}} \frac{Y_{q,i} - \mu(Y_q)}{\sigma(Y_q)},$$

其中 n 等于在第 i 个时间单位内的询问总数。经过所有 d 个时间单位的所有这些向量 $\tilde{Y}_{q,i}$ 在 d 维欧几里德实空间 $\mathbb{R}^d$ 内为询问 q 定义了表示归一化频率函数的单个询问频率向量 $\tilde{Y}_q$。询问频率向量 $\tilde{Y}_q$ 是在原点具有第一点而在 d 维实空间内的 $(x_1, x_2, \ldots, x_d)$ 处具有第二点的向量，其中每个 x_i 是由经历 i 个时间单位的有关询问 q 的每一个成比例且归一化的向量 $\tilde{Y}_{q,i}$ 所确定。

每一个询问频率向量都可用于确定将每个询问频率向量 $\tilde{Y}_q$ 表示为 δ 位串 $v(\tilde{Y}_q)$ 的映射 $v(\cdot)$: $\mathbb{R}^d \rightarrow \{0, 1\}^\delta$。为 δ 选择的值取决于期望的近似精确程度，而不取决于原始维数 d。在各实施例中，δ 可以在 80 至 160 位之间，并且还可以是 128 位。应该理解在可选实施例中为 δ 选择的值也可以小于 80 或大于 160。

用于位串 $v(\tilde{Y}_q)$ 内 δ 位的每一位的值（1 或 0）可以按在先前结合以供参考的

P. lndyk 的论文和 W. Johnson 的论文中，以及在 M. Goemans 和 D. Williamson 在 Journal of the ACM (JACM), 42(6): 1 1 15-1145 (1995)上发表的题为“Improved Approximation Algorithms for Maximum Cut and Satisfiability Problems Using Semidefinite Programming”的论文中所阐明的类似方式使用随机超平面来确定。更具体地，映射 v(·)由δ个随机向量$\{r_1, r_2, \ldots, r_\delta\} \in \mathbb{R}^d$所定义，其中随机向量可被解释为从高斯分布中提取的随机选定超平面的法向量。对于某一输入询问 q 而言，位串 $v(\tilde{Y}_q)$第 j 位的值由询问频率向量$\tilde{Y}_q$位于第 j 个随机选定超平面的哪一侧所支配。这由点积$\tilde{Y}_q \cdot r_j$的符号给出。该点积的符号将由$\tilde{Y}_q$和r_j之间角度余弦的符号所给出。在r_j和$\tilde{Y}_q$之间的角度大于 90°（指示负点积）的情况下，则位串（在此还被称为“嵌入”）$v(\tilde{Y}_q)$的第 j 个值将会是 0。在r_j和$\tilde{Y}_q$之间的角度小于 90°（指示正点积）的情况下，则嵌入 $v(\tilde{Y}_q)$的第 j 个值将会是 1。生成的超平面个数δ大到足以提供超平面的平均分布以生成精确表示每个询问频率函数的位串 $v(\tilde{Y}_q)$。

这一概念将参考示出了表示询问 q 和 p 的一对询问频率向量$\tilde{Y}_q$和$\tilde{Y}_p$的图 7 和图 8 而得到更为详尽的解释。第一随机向量 r_1 被生成作为随机生成超平面的正交向量。点积$\tilde{Y}_q \cdot r_1$是负值，所以位串 $v(\tilde{Y}_q)$的第一位是 0。类似地，点积$\tilde{Y}_p \cdot r_1$是负值，所以位串 $v(\tilde{Y}_p)$的第一位是 0。第二随机向量 r_2 被生成作为第二随机生成超平面的正交向量。点积$\tilde{Y}_q \cdot r_2$ 是正值，所以位串 $v(\tilde{Y}_q)$的第二位是 1。类似地，点积$\tilde{Y}_p \cdot r_2$是正值，所以位串 $v(\tilde{Y}_p)$的第二位是 1。随后第三随机向量 r_3 被生成作为第三随机生成超平面的正交向量。点积$\tilde{Y}_q \cdot r_3$是正值，所以位串 $v(\tilde{Y}_q)$的第三位是 1。然而点积$\tilde{Y}_p \cdot r_3$是负值，所以位串 $v(\tilde{Y}_p)$的第三位是 0。这一生成随机向量 r_1 至 r_δ并获取其与$\tilde{Y}_q$和$\tilde{Y}_p$点积的进程持续，直到串 $v(\tilde{Y}_q)$和 $v(\tilde{Y}_p)$内从第一位到第δ位的每一位都已分配或 1 或 0 的位值。

在输入询问 q 和任何存储询问 p 之间的相关系数λ可由嵌入 $v(\tilde{Y}_q)$和 $v(\tilde{Y}_p)$之间从第 1 位到第δ位对应相同位数的数所近似。更具体地，询问 q 和 p 之间的相关系数λ由询问频率向量$\tilde{Y}_q$和$\tilde{Y}_p$的点积所给出。这是两向量之间角度的余弦，或由弧度测得的$\cos\theta_{qp}$。使用上述嵌入方法，期望在位串内彼此相合的对应位所占分数预计为$1-\theta_{qp}/\pi$。真实的相关系数与通过位一致所占分数$1-\theta_{qp}/\pi$给出的相关系数的近似并不彼此相等，但是在语义相关询问中已是彼此足够接近，从而能够给出相关系数的精确近似。例如，假设所关心的是具有 0.9 或更高相关系数λ的询问：

$\theta_{qp} = \mathrm{Cos}^{-1}(0.9) = 0.45$ 弧度

$1-\theta_{qp}/\pi = 0.856$

于是在本发明的各实施例中，在两个询问之间位一致分数是0.856或更高的情况下，这些询问就可具有0.9或更高的相关系数，并且可以被认为是语义相关的询问。

对应位的匹配分数越大，询问q和p之间相关系数的近似度就越高，并且各询问也更为瞬态相关。可以看出在嵌入$v(\tilde{Y}_q)$和$v(\tilde{Y}_p)$中的点积符号只会在随机生成的超平面落入两个询问频率向量$\tilde{Y}_q$和$\tilde{Y}_p$之间的情况下才不一致。于是，两个询问频率向量$\tilde{Y}_q$和$\tilde{Y}_p$之间的角度越小，它们被随机超平面截开的可能性也越小，并且询问q和p的相关系数也就越大。

简化数据模型允许以高效的时间和空间方式精确确定询问之间相关系数的相当接近的近似。在把嵌入$v(\tilde{Y}_p)$存储为128位的一个实施例中，16字节的存储位置能够存储每个嵌入。于是，例如40,000,000个询问的存储仅需要占用约640兆字节的存储空间。

如下将参考图9描述执行上述简化数据数学运算的软件步骤。在图9的描述中，假设经过所有期望的离散时间单位1至d的所有搜索询问的原始数据日志是从前一时间段中获取并且已经存在于数据库352内。应该理解这些原始数据日志还可以另外方式实时生成（即，这些原始数据不是先前存储的，而是在第i个时间单位经过之后被实时记录的）。在任一实施例中，正如前述所指示的那样，原始频率数据的日志通常在一预定时段内被存储并保持在数据库352内。

在参考图9的实施例中，在步骤230中检索给定时间单位内所有搜索询问的原始数据日志。在步骤232中，相关引擎350可以从该离散时间单位i的搜索询问日志中确定该时间单位期间每个存储询问的密度$Y_{p,i}$。如前所述，在一时间单位期间的每个询问的密度$Y_{p,i}$如下给出：

$$Y_{p,i}=\frac{n_{p,i}}{N_i},$$

其中$n_{p,i}$是询问p在第i个时间单位期间出现的次数而N_i是在第i个时间单位期间的询问总数。

在步骤236中，可以计算并存储有关所有询问的归一化询问频率Y_p的平均频率$\mu(Y_q)$的运行计数。并且可以在步骤238中计算并存储有关所有询问的归一化询问频率Y_p的标准差$\sigma(Y_q)$的运行计数。

在本发明的一个实施例中，相关引擎 350 在步骤 240 中检查在 d 个时间段总数内是否还存在额外的离散时间段。如果存在额外的离散时间段，就在步骤 242 内增加该时间单位并且相关引擎返回步骤 230 直到已经为所有询问计算了 i = 1 至 d 时所有的 $Y_{p,i}$ 值。一旦已经为每个询问计算了离散时间单位 1 至 d 的所有归一化询问频率 $Y_{p,i}$，就可以在步骤 244 中计算用于每个询问的归一化且成比例询问频率向量 $\tilde{Y}_p$。无需存储用于每个离散时间段并用于每个询问的归一化询问频率 $Y_{p,i}$。然而，即使在应用了数据简化技术的情况下，也可以在可选实施例中存储用于每个离散时间段并用于每个询问的归一化询问频率 $Y_{p,i}$。

随后在步骤 246 中生成定义随机向量 r_j 的随机超平面。这通过随机生成与给定超平面正交的向量 r_j 的每一分量来实现。也就是说，向量 r_j 由从均值为 0 方差为 1 的标准高斯分布中随机选出的 d 个值组成。在步骤 248 中计算每个询问频率向量 $\tilde{Y}$ 与随机向量 r_j 的点积。随后在步骤 250 中，相关引擎根据询问频率向量 $\tilde{Y}$ 位于某一随机向量 r_j 的哪一侧，把 0 或 1 赋予每个串 $v(\tilde{Y}_p)$ 的第 j 位。随后在步骤 252 中存储串 $v(\tilde{Y}_p)$ 有关每个询问的更新值。相关引擎 350 随后在步骤 254 中检查以观察串 $v(\tilde{Y}_p)$ 中从 1 至 δ 的每一位是否已被分配一位值。如果是，随后该数据收集进程结束。如果不是，则在步骤 256 中递增 j 并且相关引擎返回步骤 246 进行该位生成进程。

正如本领域普通技术人员所能领会的那样，以上识别步骤中的一部分在可选实施例中可以按不同的次序执行。例如，代替确定有关所有串 $v(\tilde{Y}_p)$ 的第 j 位并在随后递增 j，可以确定单个串 $v(\tilde{Y}_p)$ 内的每一位，并在随后为所有剩余 $v(\tilde{Y}_p)$ 重复该进程。

使用根据参考图 5 至图 6 所述实施例或者参考图 7 至图 9 描述的数据简化实施例而存储的任何排列询问数据，就可以确定键入询问 q 和所有存储询问 p 之间的相关系数。如下将参考图 10 描述用于在接下来识别与输入询问具有一阈值相关系数的所有存储询问的形式模型。当在步骤 260 中键入询问 q 时，就可以在步骤 262 从存储器中检索与其相对应串 $v(\tilde{Y}_q)$（在由数据简化技术操作的实施例中）。随后在步骤 264 中做出对所有串 $v(\tilde{Y}_q)$ 中所有 δ 位的线性传递以确定输入询问和所有其他存储询问之间的相关系数 λ。

在步骤 266 中识别其与键入询问的相关系数在一预定阈值之上的所有询问。在本发明的各实施例中，相关系数的阈值可以是 0.80（可转化为嵌入中对应位的 0.795 的一致性）。在另一些实施例中，相关系数的阈值可以是 0.90（可转化为上

述嵌入中对应位的 0.856 的一致性）。可以认为与输入询问的相关系数等于这些值或在这些值之上的经识别询问与键入的询问具有高度的瞬态和语义相关。应该理解在可选实施例中该相关系数的阈值可以低于 0.8 或者高于 0.9。

一旦已经识别出与输入询问相关联的一个或多个搜索询问，就在步骤 268 中将这些结果呈现给用户。对这一个或多个相关搜索询问的呈现在本发明的可选实施例中可以被不同地处理。在一个实施例中，可以向用户提供查看已确定与键入搜索询问语义相关的一个或多个搜索询问的选项。在一可选实施例中，这些相关搜索询问可以不经提示地呈现给用户。在另一个可选实施例中，可以在输入询问的同时自动搜索与该输入询问的相关系数在阈值之上的存储询问，并连同有关该输入询问的搜索结果返回各结果。高度相关搜索询问的指示对广告业而言也颇有价值。

一旦一输入询问已被键入，对所有串 v($\tilde{Y}_q$)内所有 δ 位的线性传递可能会消耗过多的时间。因此，本发明的各实施例利用一种能够在不牺牲或牺牲少量精确度的情况下用缩短的时间找出相关询问的方案。查看图 11 描述的这一方案利用有关所有询问 n 的存储串 v($\tilde{Y}_p$)。根据这一方案，算法的运行时间通过将相关引擎 350 检查的询问数限制为那些其相关可能具有高于预定阈值 λ 的询问而得以缩短。更具体地，只检查那些与输入询问 q 在嵌入 v($\tilde{Y}_q$)和 v($\tilde{Y}_p$)的前 k 位高度相关的存储询问 p，因为这些询问 p 最有可能与询问 q 高度相关。

随后将参考图 11 更为详尽地解释这一进程。在步骤 280 中，分配 2^k 个存储段（即，存储器缓冲区）。随后在步骤 282 中将每个存储的询问 p 组织成由其嵌入 v($\tilde{Y}_p$)的前 k 位指示的存储段。也就是说，前 k 位相同的所有串都被组织成在相同的存储段中。在本发明的各实施例中，k 可以在 10 至 30 位之间，进一步地可以在 15 至 25 位之间，并且在本发明的一个实施例中可以是 20 位。

一旦在步骤 284 中接收一输入询问 q，处理单元在步骤 286 中只检查在含有 q 的嵌入 v($\tilde{Y}_q$)的存储段内的那些嵌入 v($\tilde{Y}_q$)（即，在 q 的存储段内的所有串），以及那些被定义为与含有 q 的嵌入 v($\tilde{Y}_q$)的存储段相当接近的存储段。如果一存储段 b' 的 k 位与存储段 b 内对应位的一致性分数为 ρ，则存储段 b'被定义为与含有询问 q 的存储段 b 相当接近。在本发明的各实施例中，ρ 可以小于或等于相关系数 λ。相关系数 λ 为 0.9 时，ρ 的范围可以位于 0.6 至 0.9 之间，或者可选地位于 0.7 至 0.8 之间，或者在本发明的实施例中可选地为 0.85。应该理解在本发明的可选实施例中，ρ 可以小于 0.6、大于 0.9 并且大于相关系数 λ。

一旦在步骤 286 中确定了存储段 b 和 b'，则在步骤 288 的本发明各实施例中

将上述存储段中每一个内的询问的所有δ位与位串 $v(\tilde{Y}_q)$进行比较，以便从仅对前 k 位的检查所得中去除虚假肯定，并且识别那些实际上等于或大于相关系数λ的询问。一旦识别成相关询问，随后就在步骤 290 中如上参考图 10 所述将这些询问呈现给用户。

通过仅搜索询问 q 的存储段和相当接近的存储段，需要执行位比较的询问数大幅减低。例如，当 k = 20 且 p = 0.85，对于任何给定的存储段而言，仅需要检查 2^{20} = 1,048,576 个存储段中的 1351 个。

本发明描述的实施例还进一步地测量给定时间单位内的进入询问数据。然而，由于不同的时区，例如来自美国东海岸的询问的发送时刻就与例如来自美国西海岸的询问发送时刻不同，即使这些询问是同时到达搜索引擎位置的。在另外各可选实施例中，可以在计算归一化询问频率时考虑发送该询问的时区位置，从而使得例如在美国东部时区 4pm 发送的询问能够被包括在与在美国西部时区 4pm 发送的询问所在相同的时间单位内，即使这些询问到达搜索引擎位置的时间不同。

在上述实施例中，为所有时间单位存储在第 i 个时间单位期间的每个询问的密度 $Y_{p,i}$。在本发明的另一实施例中，本发明的上述方法可以在不存储有关各时间单位的每个询问的密度 $Y_{p,i}$ 的情况下开展。更具体地，如前所述，随机向量 r 和询问频率向量 $\tilde{Y}_q$ 的点积符号可用于确定从中可以精确近似两询问的相关系数的嵌入 $v(\tilde{Y}_q)$。然而可以在不存储询问密度 $Y_{p,i}$ 的过去值的情况下计算 r 和 $\tilde{Y}_q$ 的点积符号。r 和 $\tilde{Y}_q$ 的点积给出如下：

$$\sum_{i=1}^{d}\frac{(Y_{q,i}-\mu(Y_q))}{\sigma(Y_q)}\times r_i$$

$\sigma(Y_q)$的值不影响点积的符号，所以在确定点积符号时无需该$\sigma(Y_q)$。于是，所需要的是如下的符号：

$$\sum_{i=1}^{d}\left(Y_{q,i}-\mu(Y_q)\right)\times r_i \text{，等于} \sum_{i=1}^{d}Y_{q,i}\times r_i-\mu(Y_q)\times\sum_{i=1}^{d}r_i$$

项$\sum_{i=1}^{d}Y_{q,i}\times r_i$和$\mu(Y_q)\times\sum_{i=1}^{d}r_i$的计算都无需参考询问密度 $Y_{p,i}$ 的过去值。第一项可以由接收到 $Y_{p,i}$ 新值计算。第二项可以仅通过存储 i 个时间单位经过时 $Y_{p,i}$ 的和与

r_i的和来计算。于是，r 和$\tilde{Y}_q$的点积符号就可以在不存储询问密度 $Y_{p,i}$的过去值的情况下算出。

在上述实施例中，不同时间单位 i 的存储询问密度 $Y_{p,i}$在计算归一化且成比例的频率向量$\tilde{Y}_p$时的权重相同。然而在可选实施例中，可以将指数衰减函数应用于更早时间单位的询问密度 $Y_{p,i}$，从而使得更近时间单位的询问密度在确定归一化且成比例的频率向量$\tilde{Y}_p$时的权重更重。例如，当使用当前时间单位的新询问密度 $Y_{p,i}$来更新频率向量$\tilde{Y}_p$时，在添加当前时间单位的询问密度之前可以将小于 1 的乘数（例如，0.5 至 0.99，并且作为另一示例的 0.9）应用于所有过去询问密度之和。于是，来自更早时间段的询问密度信息在频率向量$\tilde{Y}_p$中权重就更轻。

在上述各实施例中，语义关系由瞬态相关所确定，即由给定时间单位内两个询问的密度关系确定。应该理解也可以使用给定时间单位内密度之外的其他准则作为确定各询问之间相关系数和语义关系的基础。例如，在一个可选实施例中，由大量用户中单个用户生成的询问可以使用根据本发明上述的方法来进行相关。这可以是完全时间无关的。作为另一个实施例，由特定时间段内单个用户生成的询问可以使用上述方法进行相关。可以预计本发明的各方法在可选实施例中还可用于以地理区域或其他准则为基础进行询问相关。

在前已出于例证和描述的目的呈现了本发明的详细描述。并不旨在穷竭地或将本发明限制在公开的精确形式。根据上述教示可以进行许多修改和变化。选择描述各实施例是为了最好地解释本发明的原理及其实际应用，从而让本领域普通技术人员能够在适于特定预期使用的各种不同的实施例中使用各种不同的修改最好地利用本发明。本发明的范围应由所附权利要求所限定。

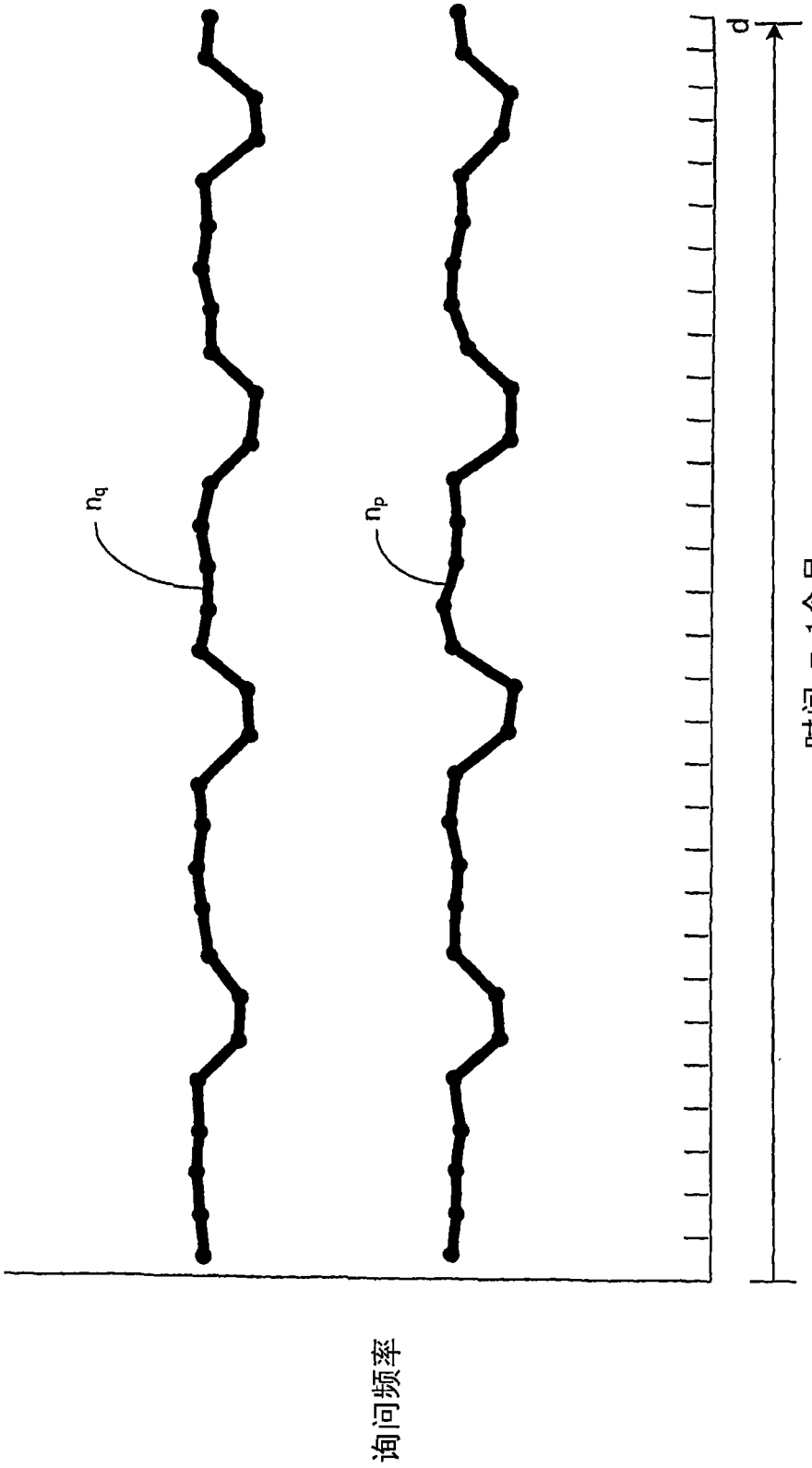

图 1
现有技术

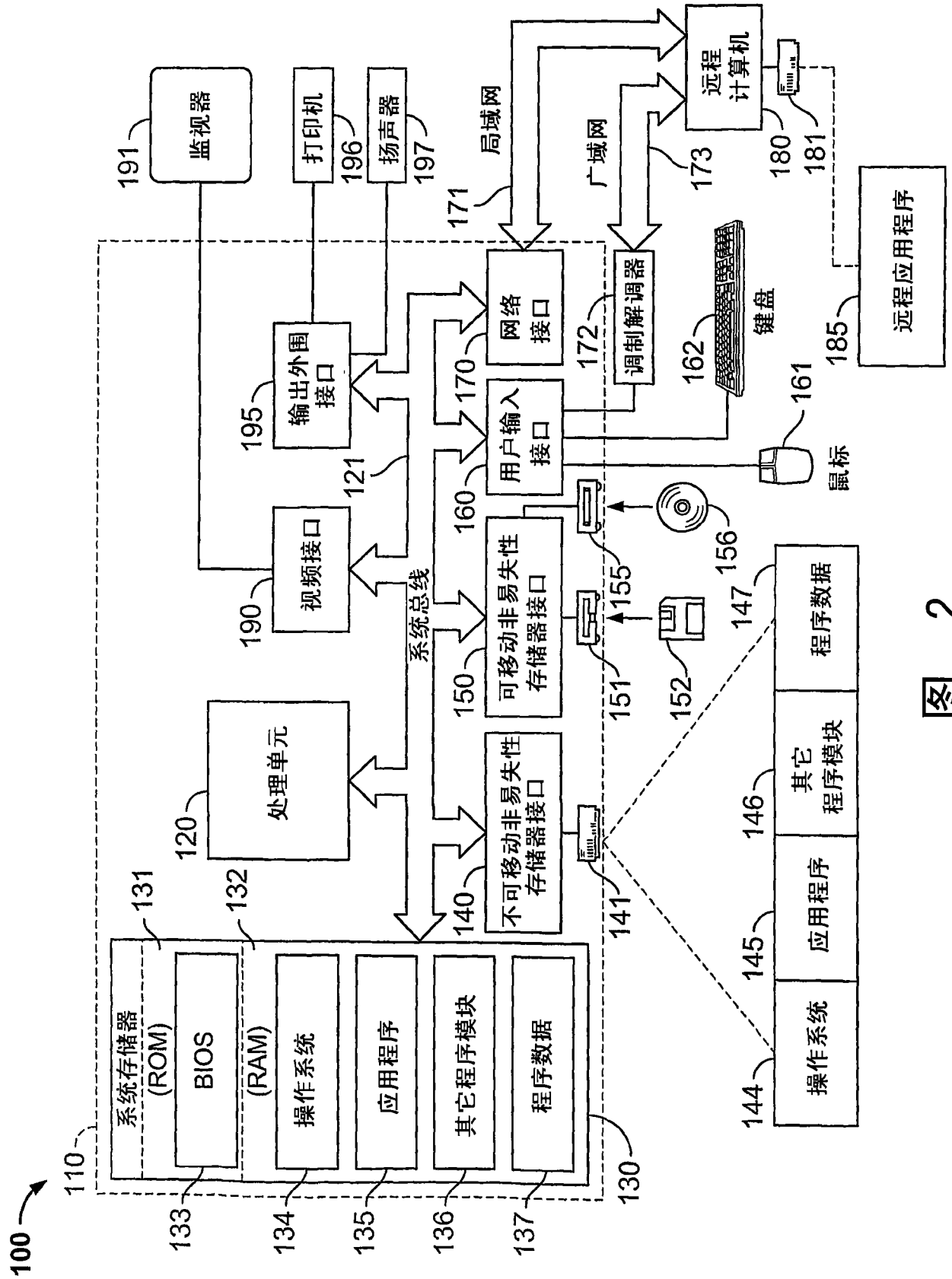
100
110
系统存储器
(ROM)
BIOS
(RAM)
操作系统
应用程序
其它程序模块
程序数据
131
132
133
134
135
136
137
130
处理单元
120
系统总线
121
视频接口
190
监视器
191
输出外围接口
195
打印机
196
扬声器
197
网络接口
170
用户输入接口
160
161
鼠标
162
键盘
调制解调器
172
局域网
171
广域网
173
远程计算机
180
181
远程应用程序
185
可移动非易失性存储器接口
150
不可移动非易失性存储器接口
140
151
152
155
156
141
操作系统
144
应用程序
145
其它程序模块
146
程序数据
147

图 2

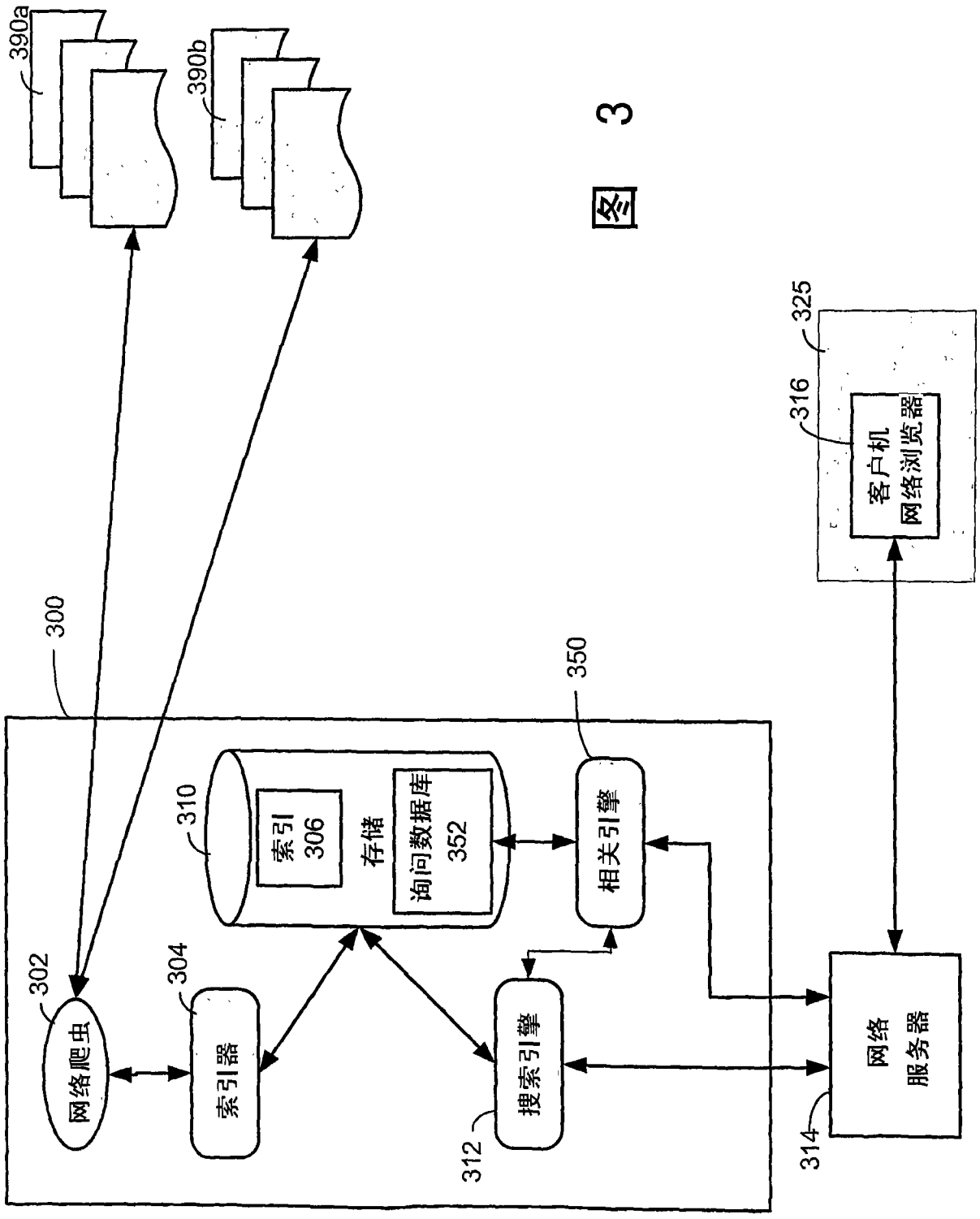
390a
390b
300
302
网络爬虫
304
索引器
310
索引
306
存储
询问数据库
352
350
相关引擎
312
搜索引擎
314
网络
服务器
316
客户机
网络浏览器
325

图 3

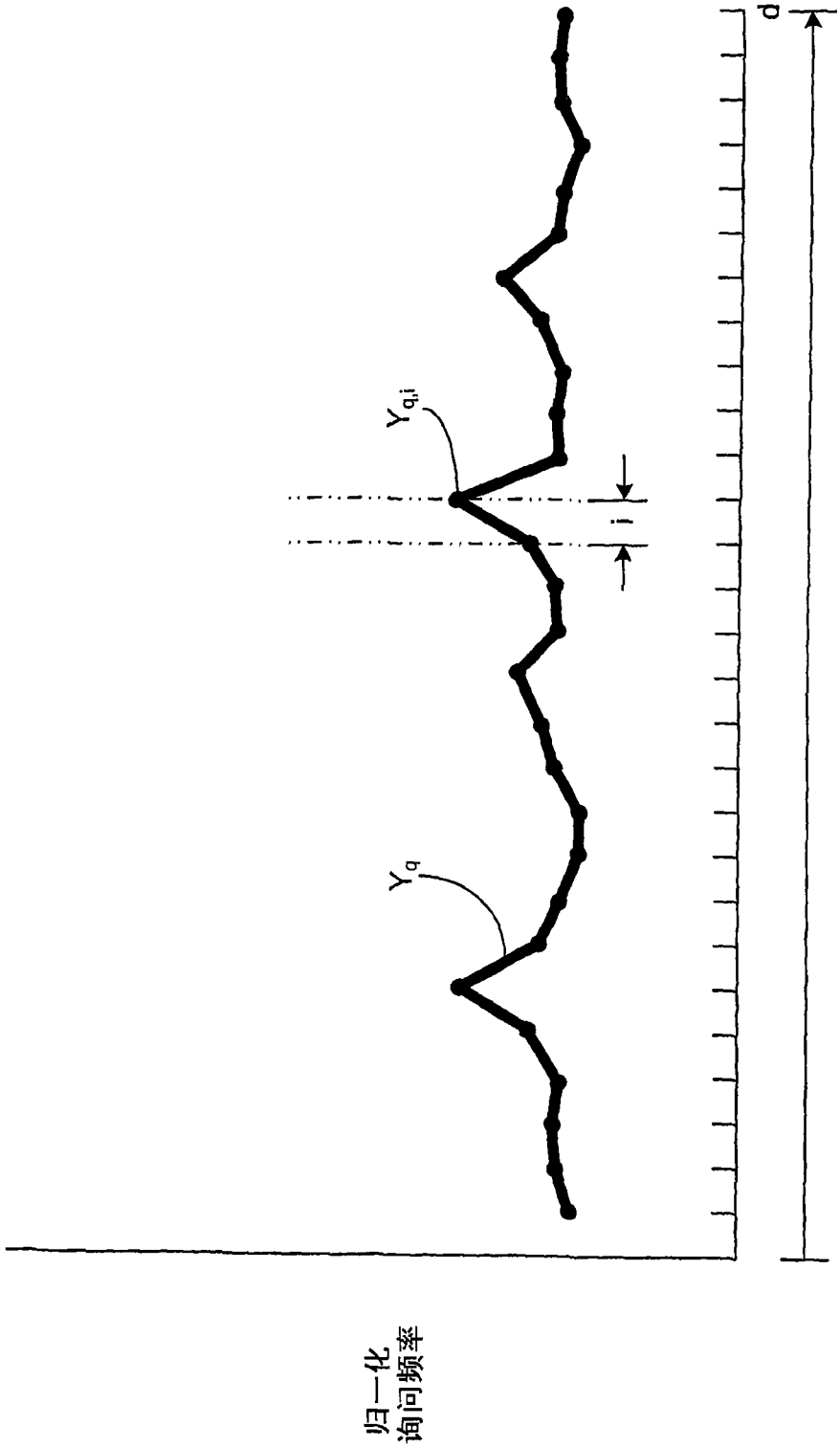

图 4

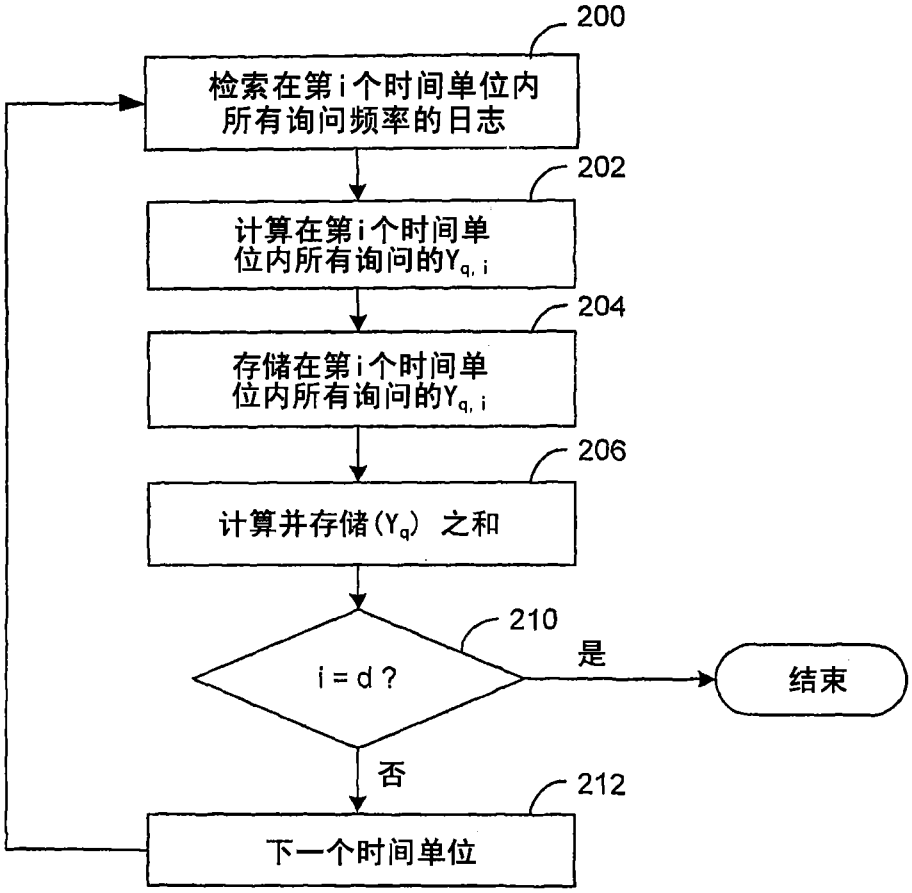
200
检索在第i个时间单位内所有询问频率的日志
202
计算在第i个时间单位内所有询问的$Y_{q,i}$
204
存储在第i个时间单位内所有询问的$Y_{q,i}$
206
计算并存储(Y_q) 之和
210
i = d ?
是
结束
否
212
下一个时间单位

图 5

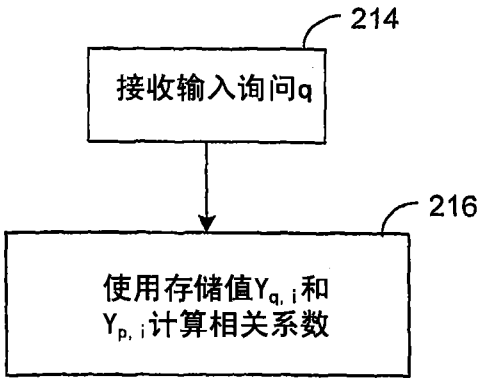
214
接收输入询问q
216
使用存储值$Y_{q,i}$和$Y_{p,i}$计算相关系数

图 6

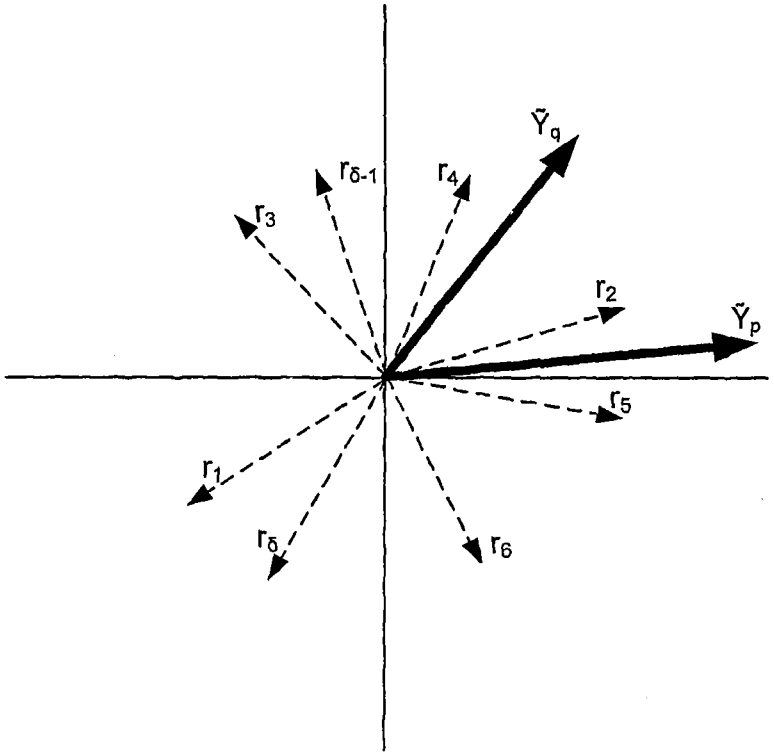

图 7

	1	2	3	4	5	6	...	δ-1	δ
$v(\tilde{Y}_q)$	0	1	1	1	1	0		1	0
$v(\tilde{Y}_p)$	0	1	0	1	1	1		0	0

图 8

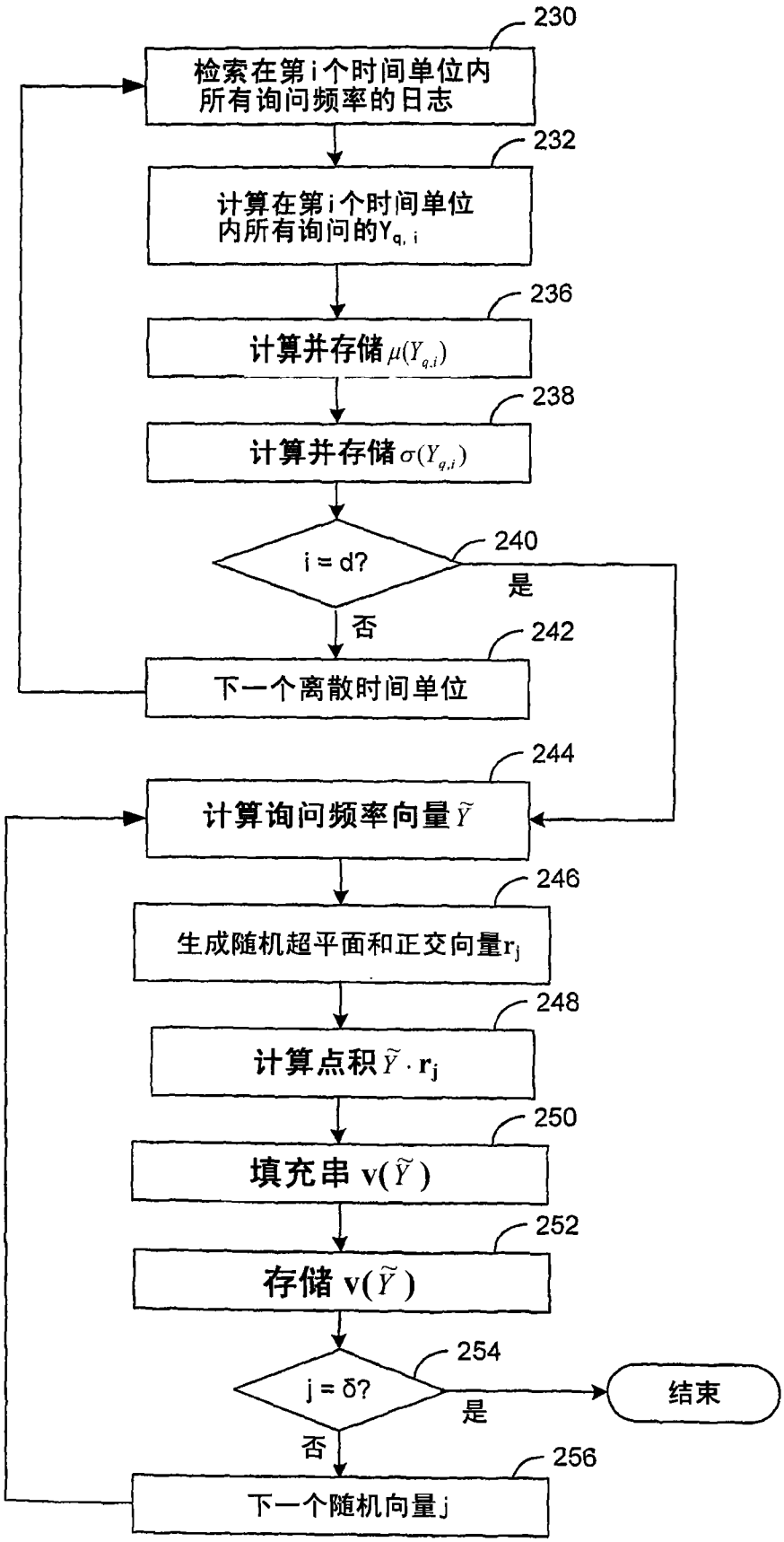
230
检索在第i个时间单位内所有询问频率的日志
232
计算在第i个时间单位内所有询问的$Y_{q,i}$
236
计算并存储$\mu(Y_{q,i})$
238
计算并存储$\sigma(Y_{q,i})$
240
i = d?
是
否
242
下一个离散时间单位
244
计算询问频率向量$\tilde{Y}$
246
生成随机超平面和正交向量r_j
248
计算点积$\tilde{Y} \cdot r_j$
250
填充串 $v(\tilde{Y})$
252
存储 $v(\tilde{Y})$
254
j = δ?
是
结束
否
256
下一个随机向量j

图 9

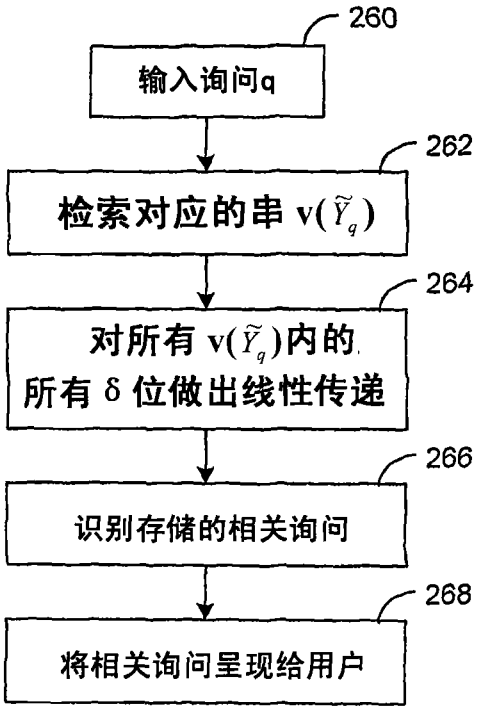
260
输入询问q
262
检索对应的串 $v(\tilde{Y}_q)$
264
对所有 $v(\tilde{Y}_q)$内的
所有 δ 位做出线性传递
266
识别存储的相关询问
268
将相关询问呈现给用户

图 10

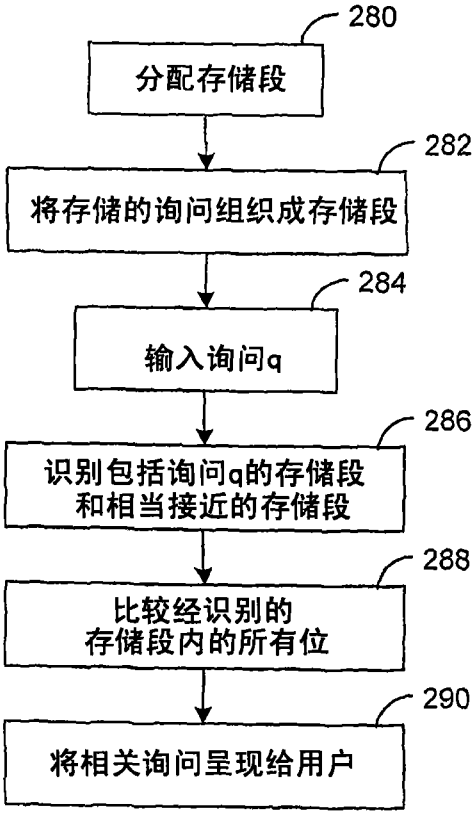
280
分配存储段
282
将存储的询问组织成存储段
284
输入询问q
286
识别包括询问q的存储段
和相当接近的存储段
288
比较经识别的
存储段内的所有位
290
将相关询问呈现给用户

图 11