(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 101223525 B
(45)授权公告日 2012.04.25

(21)申请号 200680026035.6
(22)申请日 2006.06.05
(30)优先权数据
60/688,242 2005.06.06 US
(85)PCT申请进入国家阶段日
2008.01.16
(86)PCT申请的申请数据
PCT/US2006/021662 2006.06.05
(87)PCT申请的公布数据
WO2006/133050 EN 2006.12.14
(73)专利权人 加利福尼亚大学董事会
地址 美国加利福尼亚
(72)发明人 卡斯安·弗兰克斯
考内利亚·A.·迈尔斯
拉夫·M.·波多维斯基
(74)专利代理机构 中国国际贸易促进委员会专利商标事务所 11038
代理人 李向英
(51)Int.Cl.
G06F 17/30(2006.01)

(56)对比文件
US 5794178 A,1998.08.11,
US 6175828 B1,2001.01.16,
US 6138113 A,2000.10.24,
MARK EMBREE.Growth and decay of random Fibonacci sequences.《THE ROYAL SOCIETY》.1999,第2471－2484页.

审查员 李楠

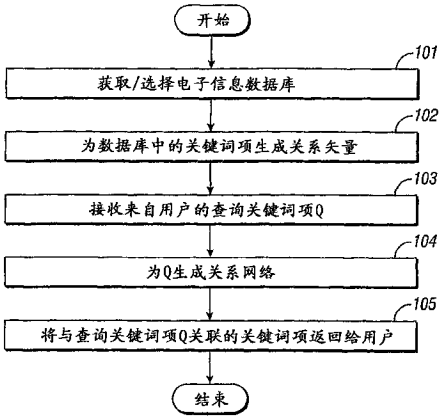
权利要求书 3 页 说明书 14 页 附图 16 页

(54)发明名称

关系网络

(57)摘要

说明了一种计算机实现的用于生成关系网络的系统和过程。该系统提供了待相关联的数据项的集合,并生成可变长度数据矢量以代表每一个数据项内的关键词项之间的关系。该系统可以用来生成文档、图像或任何其他类型的文件的关系网络。然后,可以查询此关系网络,以发现该数据项集合内的关键词项之间的关系。



CN 101223525 B

1. 一种计算机实现的用于生成关系网络的方法，包括：

(a) 提供待相关联的数据项的集合，其中，所述数据项包括多个关键词项；

(b) 选择要处理的第一数据项；

(c) 向所述第一数据项应用框架，其中，所述框架包括所述数据项内的关键词项的第一集合；

(d) 为所述框架内的关键词项计算数据矢量；

(e) 移动所述框架以包括所述数据项内的关键词项的第二集合；

(f) 通过重复步骤 (d)-(e) 直到已经计算出所述数据项中的所有关键词项的数据矢量来创建关系网络；

(g) 将所述关系网络存储在存储器中；

(h) 接收来自用户的查询关键词项；

(i) 基于接收到的查询关键词项，生成所存储的关系网络的查询；以及

(j) 显示查询结果，其中所述查询结果提供所述查询关键词项和所述数据项的集合中的其它关键词项之间的关系的可视表示，

其中，生成所存储的关系网络的查询包括：检索所述查询关键词项的查询对象矢量；配置一过滤器以与所述查询一起使用；将所述查询对象矢量展开为展开的查询对象矢量；基于所述查询对象矢量生成展开的关联的对象矢量；以及查找所述展开的关联的对象矢量与所述展开的查询对象矢量之间的关联关键词项，并且

其中，将所述查询对象矢量展开为展开的查询对象矢量包括：识别与所述查询关键词项最强相关的一组关键词项；将与所述查询关键词项最强相关的这组关键词项添加到所述展开的查询对象矢量的开头处；检索与所述查询关键词项最强相关的这组关键词项中的每个关键词项的矢量；以及对于与所述查询关键词项最强相关的这组关键词项的每个矢量，插入该矢量的多个最强关键词项。

2. 根据权利要求 1 所述的方法，包括向所述第一数据项中的独特关键词项的所述数据矢量添加权重值。

3. 根据权利要求 2 所述的方法，其中，所述权重值与所述第一数据项中的所述关键词项的频率相关。

4. 根据权利要求 1 所述的方法，其中，所述数据项包括文档，而所述关键词项包括单词。

5. 根据权利要求 4 所述的方法，其中，所述框架包括所述数据项中的至少三个句子。

6. 根据权利要求 1 所述的方法，进一步包括从所述数据项中删除特定的关键词项。

7. 根据权利要求 1 所述的方法，其中，基于所述查询对象矢量生成展开的关联的对象矢量包括：

识别与所述查询关键词项强相关的一组关键词项；

对于强相关的这组关键词项中的每个关键词项，检索该关键词项的矢量并基于检索的矢量生成第一维关联对象矢量；

识别与第一维关联对象矢量中的关键词项强相关的第二组关键词项；

对于强相关的第二组关键词项中的每个关键词项，检索该关键词项的矢量并基于检索的矢量生成第二维关联对象矢量。

8. 根据权利要求 1 所述的方法，其中，查找所述展开的关联的对象矢量与所述展开的查询对象矢量之间的关联关键词项包括：将展开的查询对象矢量传递到一比较函数，该比较函数被配置为确定所述展开的关联的对象矢量与所述展开的查询对象矢量中的关键词项之间的相似性。

9. 一种用于生成数据项之间的关系的系统，包括：

包括待相关联的数据项的集合的存储器，其中，每一个数据项都包括多个关键词项；

框架生成器，被配置为生成框架，该框架选择所述数据项中的多个关键词项以进行关联；

矢量生成器，被配置为生成数据矢量以代表所述框架内的数据项之间的关联；以及

网络生成引擎，被配置为：

接收来自用户的查询关键词项；

基于接收到的查询关键词项，生成所存储的关系网络的查询；以及

显示查询结果，其中所述查询结果提供所述查询关键词项和所述数据项的集合中的其它关键词项之间的关系的可视表示，

其中，所述网络生成引擎为了生成所存储的关系网络的查询还被配置为：检索所述查询关键词项的查询对象矢量；配置一过滤器以与所述查询一起使用；将所述查询对象矢量展开为展开的查询对象矢量；基于所述查询对象矢量生成展开的关联的对象矢量；以及查找所述展开的关联的对象矢量与所述展开的查询对象矢量之间的关联关键词项，并且

其中，所述网络生成引擎为了将所述查询对象矢量展开为展开的查询对象矢量还被配置为：识别与所述查询关键词项最强相关的一组关键词项；将与所述查询关键词项最强相关的这组关键词项添加到所述展开的查询对象矢量的开头处；检索与所述查询关键词项最强相关的这组关键词项中的每个关键词项的矢量；以及对于与所述查询关键词项最强相关的这组关键词项的每个矢量，插入该矢量的多个最强关键词项。

10. 根据权利要求 9 所述的系统，其中，所述矢量生成器被配置为基于框架内的关键词项的频率除以该关键词项在整个文档集合中的频率，来修改数据矢量。

11. 根据权利要求 9 所述的系统，被配置为从文档中删除停止列出的单词。

12. 根据权利要求 9 所述的系统，其中，包括多单词短语的关键词项被当作单个关键词项。

13. 根据权利要求 9 所述的系统，进一步包括用于接受来自用户的查询关键词项并确定与所述查询关键词项关联的关系矢量的输入模块。

14. 根据权利要求 13 所述的系统，进一步包括提取器模块，所述提取器模块被配置为提取所述查询关键词项的所有关系矢量。

15. 根据权利要求 14 所述的系统，进一步包括列表生成器，所述列表生成器被配置为根据独特性分数生成所述关系矢量的列表。

16. 根据权利要求 9 所述的系统，其中，所述矢量生成器被配置为计算关键词项之间的距离。

17. 根据权利要求 9 所述的系统，其中，所述网络生成引擎为了基于所述查询对象矢量生成展开的关联的对象矢量还被配置为：

识别与所述查询关键词项强相关的一组关键词项；

对于强相关的这组关键词项中的每个关键词项，检索该关键词项的矢量并基于检索的矢量生成第一维关联对象矢量；

识别与第一维关联对象矢量中的关键词项强相关的第二组关键词项；

对于强相关的第二组关键词项中的每个关键词项，检索该关键词项的矢量并基于检索的矢量生成第二维关联对象矢量。

18. 根据权利要求 9 所述的系统，其中，所述网络生成引擎为了查找所述展开的关联的对象矢量与所述展开的查询对象矢量之间的关联关键词项还被配置为：将展开的查询对象矢量传递到一比较函数，其中，该比较函数被配置为确定所述展开的关联的对象矢量与所述展开的查询对象矢量中的关键词项之间的相似性。

关系网络

[0001] 对相关申请的交叉引用

[0002] 本专利申请以 2005 年 6 月 6 日提出的美国临时专利申请 No.60/688,242 为基础要求优先权，这里引用了该申请的全部内容作为参考。

[0003] 关于联邦政府赞助的研究和开发的声明

[0004] 本发明是根据合同 No.DE-AC02-05CH11231 在美国能源部支持的工作过程中作出的。政府对本发明具有某些权限。

技术领域

[0005] 本发明涉及基于矢量的信息存储和检索系统。具体来说，本发明涉及用于存储、生成和检索上下文矢量以构建和可视化信息的关系网络的系统。

背景技术

[0006] 基于短语或关键字的搜索是用于对电子数据进行搜索的常用方法。通过关键字进行的搜索会在整个信息数据库中进行搜索，以查找搜索查询中的单词的实例。然而，通过关键字进行的搜索不会给出基于相关性的结果；除了搜索查询中的单词的实例之外，搜索查询结果常常还包括彼此没有相关性或关系的项目。例如，打算搜索技术公司 Apple 的产品的用户可能输入搜索查询“Apple”。然而，搜索结果将可能包括涉及水果苹果、带有音乐标记 Apple 的歌曲等等。因此，基于短语进行的搜索的搜索查询结果常常与用户的搜索意图毫无共同之处。

[0007] 将一个对象与另一个对象相关联的搜索方法常常被用来代替通过关键字进行的搜索，以便提供与搜索者的意图相关的搜索查询结果。这样的基于关系的搜索方法有很大的不同，范围从精确的到一般的各种各样的方法。涉及文本对象的方法在精确性和方法、质量和数量方面可以有很大的不同。例如，标题为“System and Method ofContext Vector Generation and Retrieval”的美国专利 No.5,619,709 的发明人 Caid 等人依靠上下文矢量生成和陈旧的神经网络方法而不是更加先进的自动关联方法。美国专利 No.6,816,857 的发明人 Weissman 等人使用距离计算方法来确定关系，以便在网站上放置基于意思的广告，或在当前使用的搜索引擎中评定文档相关性。

[0008] 然而，这些基于关系的搜索没有模拟人在分析相关信息以将对象彼此相关联时所使用的过程。从感兴趣的对象开始，研究人员通常在某些上下文本内进行研究，并形成在读取和分析文献的过程中收集的信息之间的关系。在此灵活的过程中，所感兴趣的上下文可以随着发现的信息或研究人员的思维过程而变化，变得精炼，或移动并呈现新的方向。在研究人员完成研究过程之后，给他留下了涉及特定主题或所感兴趣的上下文的有价值的信息集合。例如，如果研究人员的感兴趣的对象是音乐的时期，上下文是巴洛克风格，那么，研究人员可以将乐曲彼此相关联，将乐曲与作曲家相关联，将乐曲与地理位置或时期相关联。基于共同的关系的搜索不模拟此过程，因为它们两者都不灵活，也没有交互；它们既不允许用户在进行搜索过程中定义和控制上下文和单个关系，也不允许由用户交互地确定和可视化

关系的质量和数量。

发明内容

[0009] 这里某些实施例提供了与关系网络一起使用的用于分析、设计和实现从信息数据库创建的矢量的系统和计算机实现的方法。某些实施例还提供了与关系网络一起使用的基于关系的网络生成引擎。

[0010] 在一个实施例中，提供了用于确定电子数据库中的对象之间的关系的系统。首先，获取诸如原始文本文档或数据之类的对象。然后，通过过滤掉无关的数据并计算对象之间的距离，来处理对象。距离度量可以是，例如，指数式衰减计算。然后，使用距离分数来创建对象之间的关系的关系分数值。生成并存储使用关系分数值的矢量。在某些实施例中，可变长度矢量可以存储代表相对于操作对象的指定的框架内的对象之间的距离的数据。由于一个矢量内的每一个对象都可以具有其自己的矢量，因此，可以使用矢量来构建关系网络。此外，关系网络中的对象之间的连接的组织也可以供用户进行搜索、可视化或其他解释。在某些实施例中，可以突出显示独特对象，而在其他实施例中，则可以突出显示共同的对象。

[0011] 在另一个实施例中，提供了用于查找关系的网络生成引擎。当对两个或更多矢量之间发现的交叉属性进行操作时，网络生成引擎能够识别文本、单词或对象之间的明显的、独特的和隐藏关系。

[0012] 在一个实施例中，网络生成引擎可以在包含矢量集合的关系数据库上实现。使用输入查询对象作为指南以为查询对象从关系数据库中提取所有直接和关联的关系。引擎可以对这些关系进行评分并进行排序，并测量任何交叉对象的相似性分数，然后，使用相似性分数，构建另一个关系网络，该网络显示查询对象与其他对象的关系，以及它们的关系的强度。如有必要，可以可视化查询对象的所产生的关系网络，供进一步解释。为确保当正在构建关系网络时所提交的对象停留在特定上下文内，可以使用过滤器形式的主题上下文来控制在所产生的网络内提取的关系的类型。

附图说明

[0013] 图 1 是用于生成关系网络的系统的一个实施例的流程图。

[0014] 图 2 是用于基于包含文本文档的电子信息数据库生成与关系网络一起使用的矢量的系统的一个实施例的流程图。

[0015] 图 3A 显示了来自包含文本文档的信息数据库的示例文档。

[0016] 图 3B 显示了图 3A 的文档在经过分析之后的情形。

[0017] 图 4 显示了与图 3A 和 3B 的样本数据一起使用的框架的一个实施例。

[0018] 图 5 显示了在框架中的正在被分析的当前关键词项是核心关键词项“red”的状态下图 4 的关键词项“red”的示例关联存储器模块。

[0019] 图 6A 显示了在系统完成了其对包含图 3A 的文档的信息数据库的分析之后关键词项“red”的关联存储器模块。

[0020] 图 6B 显示了图 6A 的关联存储器模块的示例查询对象矢量。

[0021] 图 7 显示了网络生成引擎的示例流程图。

[0022] 图 8A 显示了应用于查询对象矢量的示例排除过滤器矢量。

[0023] 图 8B 显示了使用图 8A 的经过滤的查询对象矢量生成展开的查询对象矢量的示例方法。

[0024] 图 8C 显示了使用图 8A 的经过滤的关联对象矢量生成展开的关联的对象矢量的示例方法。

[0025] 图 8D 显示了与展开的查询对象矢量一起使用展开的关联的对象矢量来查找关联的对象矢量和展开的查询对象矢量之间的关联的关键词项以便产生查询的搜索结果的一个示例方法。

[0026] 图 9 显示了响应关键词项“red”的查询创建的关系网络的图形可视化。

[0027] 图 10 显示了根据一个实施例的关系网络系统。

具体实施方式

[0028] 本发明的一个实施例是在一个集合中的不同项目之间创建并辨别关系的计算机方法和系统。在一个实施例中，在数据集中的数据项之间创建多对多关系。作为一个示例，数据项可以是基因，而数据集可以是 GENBANK 基因数据库。正如下面比较详细地描述的，系统的实施例对数据集中的数据项进行分析，此后创建反映数据集中的数据项之间的可变长度数据矢量，如查询对象矢量。然后，数据矢量可以存储起来，并被用作分析数据项之间的关系的数据挖掘工具的一部分。例如，可以搜索 Genbank 中的涉及胃癌的所有基因。

[0029] 在本发明的一个实施例中，通过首先分析两个数据项之间的直接相关性，然后寻找数据项之间的进一步的隐藏的关联，创建标记数据项之间的关联的数据矢量。在一个实施例中，通过反复分析数据集中的每一关键词项与其他关键词项的距离，确定这些隐藏关系。如此，例如，在数据集中发现两个单词彼此关联的次数越多，它们之间的关系就越近。在某些实施例中，通过跨每一个数据项地移动“框架”来分析关键词项。例如，如果数据项是文档，则框架可以一次一行地穿过文档，但是覆盖三行。随着框架沿着文档的每一行移动，对框架内的关键词项之间的距离进行分析。在此分析过程中，创建存储了框架中的每一关键词项之间的关系的数据矢量。在一个实施例中，整个数据集内的每一关键词项通过一个矢量来表示。该矢量提供了该关键词项以及其相关的关键词项之间的距离和关系。

[0030] 本发明的另一个实施例是使用存储的数据矢量来提供搜索查询的有用结果的系统和方法。当人或机器作为搜索的一部分输入一关键词项时，定位该关键词项的数据矢量，并根据数据矢量识别与搜索关键词项最相关的关键词项。然后，系统检索最相关关键词项的数据矢量，以便展开搜索。然后，可以识别与最相关关键词项相关的关键词项，并可以继续执行该过程，以构建原始搜索关键词项，以及所有其相关的关键词项之间的关系网络。一旦执行查询，并对包含最相关关键词项的矢量进行计分，构建关系网络。然后，可以准备提交的关键词项的所产生的网络，使其可视化，以便进行进一步的解释。在一个实施例中，关键词项显示在计算机屏幕上，带有链接的 Web，显示了每一个搜索关键词项与其结果如何相关。为确保当正在构建关系网络时所提交的关键词项停留在特定上下文内，可以使用过滤器形式的主题上下文来控制在所产生的网络内提取的关系的类型。

[0031] 这里所说明的系统和方法允许用户交互地参与到信息挖掘、隐藏关联和连接提取、关系网络构建和对象的比较中，同时交互地应用主题上下文控制，来精炼所提取的关系的类型。系统和方法给用户提供了有关信息数据库内的对象如何彼此相关、它们在什么上

下文中相关，以及它们的关系的强度的信息。

[0032] 通过组合用户的交互式角色，类似于研究人员在实验过程中所参与的事情，并将它应用到自动化文本挖掘方法的迭代过程，这里所讨论的某些实施例给用户提供了当在被搜索的信息中的感兴趣的对象之间建立连接时选择方向并定义关系。通过交互地定义并提取对象、主题及其他上下文之间的关系，使得在文本中进行关系探索和发现的精确性达到较高的水平。

[0033] 例如，如果用户在诸如因特网之类的电子信息数据库中搜索 Baroque 乐曲，则用户可以向关系网络系统提交关键词项“Baroque”。用户也可以通过使用诸如“compositions”之类的过滤器关键词项，指示在 Baroque 音乐的方向进行搜索，以便避免产生涉及巴洛克艺术的结果。然后，系统将不仅提供有关与关键词项“Baroque”强烈地关联的乐曲的信息，而且还提供诸如作曲家姓名“Bach”和“Handel”之类的与涉及“Baroque”的关键词项强烈地关联的乐曲，诸如“viola dagamba”或“harpsichord”之类的涉及与 Baroque 音乐关联的乐器的乐曲，或相关的艺术的时期“Classical”等等的乐曲。

[0034] 在一个实施例中，这里所说明的关系网络系统可以用于消除关键词项的歧义，这提供了区别完全相同但根据上下文具有不同含义的两个字符串（如兼作标识符或符号或实际单词的缩写词）的能力。例如，单词“cleave”具有彼此相反的两个含义。

[0035] 图 1 显示了使用电子信息数据库生成关系网络的过程 100。在某些实施例中，电子信息数据库可以包括，但不仅限于，字符集合或其他形式的文本、图像、音频、视频或可以以电子方式对其进行分析的任何其他数据。如此，信息数据库内的对象或关键词项可以是文档、字符、单词、图像、歌曲或视频（“关键词项”）。

[0036] 在所说明的实施例中，系统首先在状态 101 下选择待处理的电子信息数据库。在一个示例中，数据库是音乐乐曲的数据库。然后，系统在状态 102 下为数据库内的关键词项创建矢量。创建矢量，以捕获数据库内的乐曲之间的不同强度的关系。一旦创建了矢量，系统在状态 103 下接收来自用户的查询“Q”。例如，当用户希望查找类似于查询 Q 中列出的乐曲的乐曲时，进行查询。在某些实施例中，系统可以在接收查询之前创建矢量，以便缩小响应查询进行的数据处理的开销。在其他实施例中，可以在接收到查询之后创建矢量。虽然在某些实施例中，使用了矢量来存储关键词项之间的关系，但是，在其他实施例中，也可以使用其他数据结构。在使用矢量的某些实施例中，矢量空间表示方案使用了可变长度查询对象矢量。可变长度矢量可以具有基于关键词项之间的关系确定的多个组件值或元素。此外，可以基于每一个矢量内的关联的关键词项的数量，确定可变长度矢量的大小。

[0037] 在某些实施例中，关联的关键词项是彼此之间具有直接的或者间接的关系的关键词项。在某些实施例中，一个关键词项是“第一”关键词项，而第二个关键词项是“核心关键词项”。在某些实施例中，直接关系是在矢量中的同一个框架内发现一个核心关键词项作为关联的关键词项。在某些实施例中，间接关系是核心关键词项和关联的关键词项在它们的相应的矢量中的各自共享共同的关键词项。也可以生成关键词项之间的其他关系，与这里所讨论的某些实施例一起使用。

[0038] 回到图 1，在状态 103 下响应来自用户的关键词项 Q 的查询，系统基于关键词项 Q 的可变长度矢量，在状态 104 下生成 Q 的关系网络。在某些实施例中，关系网络包括其彼此的连接，以及这些连接的强度都基于定义的上下文和主题内的共享的独特属性的关系矢量

的网络。下面比较专门讨论了上下文和主题。一旦在状态 104 下已经生成了关系网络，系统就可以在状态 105 下返回与 Q 关联的关键词项。例如，返回的关键词项可以指向同一个作曲家 Q 所作的曲子，与 Q 相关的曲子，或基于 Q 的推荐作品。

[0039] I. 为关系网络生成矢量

[0040] 图 2 是从存储在数据库内的数据生成可变长度矢量的过程 102 的一个实施例的流程图。过程 102 在状态 201 下收集数据库中的每一个文档。对于收集的每一个文档，在状态 202 下对文档进行分析，以便删除不相关的或低值数据，如停止字符（诸如 a、of、as、the、on 等等之类的常用字。）。在状态 202 下对每一个文档进行分析之后，信息数据库只包含有价值的关键词项。

[0041] 然后，对于每一个经过分析的文档，系统在状态 203 下在文档中插入框架。框架可以被视为覆盖了文档中的一行或多行文本的覆盖层。例如，框架可以覆盖文档中的三行或句子。一旦已经在状态 203 下插入了框架，过程 102 进入状态 204，选择框架中处理的第一行中的第一个关键词项。图 4 显示了与图 3A 和 3B 中所显示的样本数据一起使用的框架的一个实施例。在状态 204 下选择了框架的活动句子中的第一个关键词项之后，在状态下 205 在第一个关键词项（“核心关键词项”）和框架内的其他关键词项（“关联的关键词项”）之间生成关系数据的集合。系统记录了核心关键词项的关系数据，包括诸如每一个核心关键词项与第一个关键词项的计算出的距离分数之类的数据。在某些实施例中，关系数据可以存储在关联存储器模块中，如图 5 所示。一旦为第一个关键词项生成了关系数据，过程 102 进入判断状态 206，就是否正在对框架的活动句子中的最后一个关键词项进行分析作出判断。如果当前关键词项不是最后一个关键词项，那么，过程 102 进入状态 207，捕获框架内的下一关键词项。然后，过程 102 返回到状态 205，在状态 205 下计算新捕获的关键词项和其他核心关键词项之间的关系数据。如果正在处理的关键词项是框架的活动句子中的最后一个关键词项，那么，过程 102 进入状态 208，在处于分析中的文档中将框架向前移动一个句子或一行。如果关键词项不是框架的活动句子中的最后一个关键词项，则过程 102 返回状态 205。

[0042] 一旦过程 102 将框架向前移动了另一行或句子，则在判断状态 209 下就框架是否位于文档的末尾处作出判断。如果判断过程 102 不在文档的末尾处，那么，过程 102 返回到状态 204，则选择移动的框架的活动句子内的第一个关键词项。如果判断框架在文档的末尾处，那么，过程 102 进入判断状态 210，就该过程是否处于数据库中的最后一个文档作出判断。如果过程 102 不在数据库中的最后一个文档中，那么，过程 102 进入状态 211，选择数据库内的下一个文档。然后，过程 102 返回到状态 203，将框架插入到新收集的文档中。

[0043] 如果在判断状态 210 下判断过程 102 在最后一个文档中，那么，过程进入状态 212，它例如从关联存储器模块中检索记录的关系数据，以查找数据库中的第一个关键词项。然后，过程进入状态 213，使用来自状态 212 的关系数据，创建可变长度查询对象矢量。在某些实施例中，可以存储在查询对象矢量中的来自状态 212 的关系数据值，当存储在查询对象矢量中时，可以得到增强。增强关系数据值的示例包括增大独特关联的数据值和缩小共同关联的数据值。图 6B 显示了图 6A 的关联存储器模块的示例查询对象矢量。接下来，过程进入判断状态 214，进行检查，以判断分析的关键词项是否是数据库中的最后一个关键词项。如果不是分析最后一个关键词项，则过程进入状态 215，选择数据库内的下一关键词项。

然后，过程 102 返回到状态 213，创建下一关键词项的查询对象矢量。如果在判断状态 214 判断过程 102 处于最后一个关键词项，那么，在结束状态 216 过程结束。

[0044] 图 3A 显示了来自包含文本文档的信息数据库的示例文档 300。图 3B 显示了图 3A 的文档在经过分析 310 之后的存储数据。从图 3A 和 3B 之间的差别可以看出，在此实施例中，系统删除了诸如“they”301 “from”302 “until”303 “they′ re”304 之类的停止字符，还根据它所在的文档 311 的标识以及其关键词项 312 来组织每一个句子。

[0045] 如图 4 所示，上下文或框架 400 的一个实施例包括周围的关联的关键词项，最终与框架中的正在被分析的当前核心关键词项“red”412 关联。在一个实施例中，通过使用文档内的距离阈值，构建框架 400 和它包围的空间。例如，在图 4 中，距离阈值是包含正在被分析的核心关键词项 410 的句子之前的一个句子和之后的一个句子。如果某一个关键词项在距离阈值内，则它被视为关联的关键词项，它变成上下文框架 400 的一部分。另一方面，如果某一个关键词项超出距离阈值之外，则它将不会变成上下文框架 400 的一部分，并且不接收到核心关键词项的距离分数（也被称为“分数关联”）。通过使用文档中的单词的数量以及句子、段落、字符或其他对象的数量，可以计算出距离阈值，框架上下文 400 的大小将增大，并随着文档被读取以及收集新的统计数据而波动。在一个实施例中，被分析的数字内容是原始文本文档，框架 400 被设置为每个框架三个、四个或五个句子。图 4 中的示例具有三个句子上下文框架 400。

[0046] 系统可以在文档中或包括文件信息数据库的其他分析的数据中移动框架 400。随着框架经过文档集合地逐行移动，关键词项可以自动地彼此关联起来，并包括表示操作文档 311 的标识符。随着关键词项流入和流入出穿过文档的框架，关联的关键词项可以通过距离分数定义它们与核心关键词项关联的强度。例如，在图 4 中，在系统计算出核心关键词项“red”的距离分数之后，框架的焦点将进入下一关键词项“pink”，直到焦点到达框架的中间行中的最后一个关键词项“raspberry”。在系统计算出与关键词项“raspberry”关联的关键词项的距离分数之后，框架将前进一行，核心关键词项焦点将从下一行上的第一个关键词项“Hummingbirds”开始。此外，以关键词项“bloom”开始的句子将流出框架，以关键词项“one”开始的句子将流入框架。

[0047] 通过给每一个关联的关键词项提供距离分数，文档中的每一个核心关键词项 410 变成从统计学上来讲重要的对象，包含关系分数关联的关键词项的家族作为其关联存储器模块的元素。然后，可以在过程完成对整个信息数据库的分析之后，使用两个关键词项之间的距离分数来创建关系分数。例如，在一个实施例中，两个关键词项在反复地出现在信息数据库中的框架内时这两个关键词项之间的距离分数可以相加起来，以创建关系分数。

[0048] 当对数千或数百万文档生成关系分数时，单个文档中的框架 400 使用变得特别有利。在这里的某些实施例中，通过两个或更多关键词项之间的强的和独特连接，随着时间的过去，定义单词之间的有效的关系。对某一个关键词项的关系分数可以与人可以通过重复进行学习的方式进行比较。如果一个人重复地一起听到两个关键词项，则他将容易记住这两个关键词项，并将这两个关键词项关联在一起，而如果一个人不经常一起听到两个关键词项，则他可能不会记住这两个关键词项或将这两个关键词项关联在一起。在这里所讨论的某些实施例中，系统给经常在一起出现的两个或更多关键词项提供较高的关系分数。在某些其他实施例中，共享非常独特的属性集合的两个或更多关键词项得到较高的分数。

[0049] 如上文所讨论的，系统可以将核心关键词项 410 以及其关联的关键词项之间的关系存储在为核心关键词项创建的叫做“关联存储器模块”的文件中。在一个实施例中，关联存储器模块是存储了与统计的和基于距离的对象关联相关的信息，以及文档统计的数据库架构。如此，关联存储器模块可以捕获待搜索的数据中的意义敏感性，这需要已知每一对关键词项的靠近程度，评定了距离分数并存储起来。如此，关联存储器模块可以有利地存储诸如单词、段落、搜索查询、对象、文档、文档标识符、图像的某些部分、关键词项的某些部分、文本的某些部分、序列的某些部分或已经分成多个部分、关键词项和文档的任何一段之类的信息，以及类似地表示的许多其他信息项目类型，如数值、金融和科学数据。在一个实施例中，关联存储器模块和矢量中的每个关联的关键词项也是其自己的关联存储器模块和矢量的核心关键词项，从而实现了高维度的多对多分数关联的关系网络。在某些实施例中，这又可以在例如，关键词项的各部分之间，关键词项之间，以及关键词项和它们所在的文档之间进行较强的比较。

[0050] 在某些实施例中，可以对关联存储器模块和矢量的长度进行限制，以便更快地创建关系网络或适当的存储空间约束，因为矢量或模块的长度可能影响数据库的大小以及系统的性能、功能。在其他实施例中，关联存储器模块或矢量可以包含与可以支持的数量相同的元素。在某些实施例中，系统可以呈现一定数量的具有高分的或具有高于某一阈值的分数的关键词项，以便较好地表示被查询的信息数据库，并有助于用户进行查看。

[0051] 图 5 显示了在框架 400 中的正在被分析的当前关键词项是核心关键词项“red”410 的状态下，图 4 的关键词项“red”500 的示例关联存储器模块。所显示的关联存储器模块 500 具有三个部分 :与关键词项相关的统计 510，与包含关键词项的文档相关的统计 520，以及与关联的关键词项相关的统计 530。在所显示的实施例中，第一部分，与关键词项相关的统计 510，可以包含诸如分析的文本中的该关键词项的出现次数 511，包含该关键词项的句子的数量 512，与核心关键词项关联的其他关键词项的数量 513，以及其他关键词项与核心关键词项之间的关联的数量 514。由于所显示的关联存储器模块 500 只包含对数据库中的分析过的第一文档中的关键词项“red”410 进行分析后得到的数据（图 3A），因此，图 5 中的数据反映的是不完整的分析。如此，由于关键词项“red”410 到目前为止只出现了一次，并只在一个句子 412 中，因此，关键词项“red”410 的出现次数 510 和句子数量 511 两者都等于 1。类似地，由于到目前为止分析的所有十八关键词项也是框架 400 中的当前的所有关键词项，因此，它们也都与关键词项“red”410 关联 513。此外，由于这些关联的关键词项中一个也没有出现两次，因此，它们都是关键词项“red”410 的单个关联 514。

[0052] 文档统计部分 520 有利地识别包含该关键词项的文档 521，包含该关键词项的文档中的句子的数量 522，以及文档相对于该关键词项的分数 523。在所显示的示例中，只列出了一个文档 524，因为它是分析的包含关键词项“red”的唯一文档。文档 524 是通过其标题进行标识的，虽然也可以使用任何其他已知的标识系统来记录文档标识，如统一资源定位器（“URL”）地址。此外，在文档中只找到了一个包含关键词项“red”的句子 525。因此，给该文档分配了分数 526，该分数是 1。在所显示的实施例中，与文档关联的分数 526 是该关键词项在文档内出现的次数，虽然在其他实施例中可以使用其他评分方法。

[0053] 关联的关键词项部分 530 包括，但不仅限于，诸如关联的关键词项 531、每一个关联的关键词项相对于核心关键词项的出现次数 532，以及关联的关键词项 / 核心关键词项

对的对应的距离分数 533 之类的数据。在其他实施例中，关联的关键词项部分 530 也可以包括有关到目前为止处理的包含关联的关键词项相对于核心关键词项的句子的数量，以及关联的关键词项与核心关键词项的距离的数据。

[0054] 在移动框架内应用测量关键词项之间的关联的距离分数 533。例如，图 4 显示了围绕核心关键词项“red”的三个句子框架 400。随着框架 400 以及其核心关键词项焦点 410 穿过文档，进行计算，以向框架 400 内的每一关键词项分配相对于核心关键词项 410 的距离分数。

[0055] 可以通过任意数量的已知的方法来计算距离分数 533。此外，为了给予与核心关键词项更靠近的关联的关键词项较大的值，分配给关联的关键词项的距离分数值 533 随着它们与核心关键词项的距离的增大而可以衰减。这可以相反使用斐波纳契序列来进行应用。换句话说，在相反使用斐波纳契序列的一个实施例中，从核心关键词项到关联的关键词项的距离分数是：

[0056] $S_{ij} = \phi^{\Delta x}$，

[0057] 其中：

[0058] S_{ij} = 核心关键词项 i 和关联的关键词项 j 之间的距离分数，

[0059] $\varphi = 0.618$ 是黄金比率组件“phi”，以及

[0060] $\Delta x = |x_i - X_j|$ 是核心关键词项 i 和关联的关键词项 j 之间的相对位置。

[0061] φ 是黄金比率 $\varphi = 1.618034$ 的十进制组件。

[0062] 返回到图 5，对于与关键词项“red”关联的关键词项“cardinal”（它们是相邻的关键词项（$\Delta x = 1$））使用此公式的距离分数 536 是 $0.618 = 0.618^1$。类似地，与关键词项“red”关联的关键词项“bloom”的距离分数 537 是 $0.008 = 0.618^{10}$，因为“bloom”相距“red”十个关键词项（$\Delta x = 10$）。在某些实施例中，随着系统遇到与第一次出现分开的关联的关键词项和核心关键词项之间的第二次出现，系统可以向第一次出现中添加第二次出现的距离分数，以便保持关联的距离分数的总和。例如，在图 5 中，如果系统在包含“red”的框架内再次遇到关键词项“cardinal”534，第二次出现的距离分数是 0.008，那么，系统可以将“red”关联存储器模块 500 中的“cardinal”的距离分数 536 更新为 $0.626 = 0.618+0.008$。在其他实施例中，也可以在系统处理信息数据库的过程中使用其他方法来更新距离分数值。

[0063] 可以有利地使用基于斐波纳契数的计算，因为在许多自然现象（包括生物学和材料科学中）中发现基于连续的斐波纳契数的比率（黄金比率）的序列。如此，斐波纳契数与语法和人生成的模式有关系，并对信息的解释有影响。

[0064] 在另一个实施例中，可以使用增强的指数加权移动平均数 (EEMA)，EWMA（指数加权移动平均数）时间序列计算方式的一种变种，来计算框架内的关键词项之间的距离分数。使用 EEMA 的示例公式可以被定义为：

[0065] EEMA = \//((K*(C-P)+P)

[0066] 其中：

[0067] C = 核心关键词项的位置

[0068] P = 前一周期的简单移动平均数 (SMA)

[0069] N = EEMA 的周期数

[0070] $K = e^{(-C/5.0)}$ 平滑常数

[0071] 在再一个实施例中，可以应用标准指数式衰减算法。下面是两个可以用来计算距离分数的指数式衰减的公式：

[0072] 如果核心关键词项 i 在关联的关键词项 j 之前，那么

[0073] $Sij = 1/e(j-i)$

[0074] 如果核心关键词项 i 在关联的关键词项 j 之后，那么

[0075] $Sij = 1/e(i-j)$

[0076] 其中，Sij =对象 i 和 j 之间的关系分数，

[0077] 图 6A 显示了在系统完成了其对包含图 3A 的文档的信息数据库的分析之后关键词项“red”的关联存储器模块 600。在示例关联存储器模块 600 中，系统判断分析的信息数据库在总共十二个句子 612 中包含关键词项“红色”的十二次出现 611。此外，有 319 个与“red”关联的关键词项，在那些关键词项和“red”之间有 450 个关联。文档“Gardening Journal”625 包含四个句子 626，“red”总计出现了四次，文档“Top News Stories”628 只包含一个句子 630，出现了一次。另外，关联的关键词项“cardinal”634 与 red 具有六个关联，其单个距离分数共计等于 4.124 的总计距离分数 636，关联的关键词项“paste”只具有一个与“red”关联的出现，总计距离分数为 0.008。

[0078] 在系统处理信息数据库中的每一个文档之后，每一个关联存储器模块都可以用来创建查询对象矢量。图 6B 显示了从图 6A 的关联存储器模块 600 创建的示例查询对象矢量 650。在所显示的实施例中，使用来自关联存储器模块 650 的距离分数 633，通过强调共同关联，来计算查询对象矢量 650 的关系分数 653，如下面将进一步详细讨论的。然后，系统根据查询对象矢量 650 中的关联的关键词项的关系分数 653，对它们进行排序。例如，在图 6B 中，关联的关键词项“Cardinal”654 被排在第一位，因为它具有最高的关系分数，关键词项“Paste”655 被排在第 319 位（等于与“red”关联的关键词项的总数），因为它具有最低的关系分数。如此，每一个关联存储器模块都用于创建查询对象矢量 213。

[0079] 如此，图 6B 说明了这里所描述的系统和方法的一个优点。在基于关键字的搜索中，如果查找红色毛衣的用户在她的查询中使用了关键词项“red”，那么，她将只接收到毛衣被专门与关键词项“red”列在一起的结果。另一方面，如果用户向这里所描述的系统的实施例提交了搜索，则用户将不会只接收到“red”毛衣的结果，而是可以接收到具有红色的其他深浅程度的毛衣，如 cardinal、maroon 和 raspberry。

[0080] 在某些实施例中，系统可以有利地使用来自关联存储器模块的数据，以便为查询对象矢量创建不同关系分数值。例如，在一个实施例中，可以以强调独特关联为目标来修改距离分数，以便帮助查找隐藏关系。隐藏关系可以用来通过呈现用户不知道的可能的重要新关系的列表，帮助假设公式化。在一个实施例中，可以使用下面的独特性函数来计算强调独特性的关系分数：

[0081] $U_{ij} = S_{ij} \cdot B_{ij}$

[0082] 其中：

[0083] S_{ij} =关键词项 i 和 j 之间的基于距离的关系分数

[0084] B_{ij} =关键词项 i 与关键词项 j 的关联的偏离，

[0085] 其中：

[0086] $B_{ij} = A_i/A_j$

[0087] A_i =关键词项 i 的关联的总数

[0088] A_j =关键词项 j 的关联的总数

[0089] 在另一个实施例中，可以以强调共同关联为目标来修改距离分数，以便基于直接的关联来生成清晰的定义。直接的关联可以用来生成非常类似的对象的列表。在一个实施例中，可以使用下面的共同性性函数来计算强调共同关联的关键词项的关系分数：

[0090] $B_{ij} = A_j/A_i$

[0091] 其中：

[0092] A_i =关键词项 i 的关联的总数

[0093] A_j =关键词项 j 的关联的总数

[0094] 如此，等到图 2 的过程完成时，每一个经过分析的文档中的每一关键词项都将具有其自己的查询对象矢量；即，每一关键词项将是查询对象矢量的核心关键词项，以及其他关键词项的查询对象矢量的关联的关键词项。在某些实施例中，每一个查询对象矢量可以强调独特的或者共同的关系。此外，在某些实施例中，每一个文档也将具有其自己的关联存储器模块和查询对象矢量。然后，可以使用这些矢量来构建关系网络。

[0095] II. 构建关系网络

[0096] 图 7 显示了与上文所讨论的关系网络的实施例一起使用的网络生成引擎的过程 700。具体来说，说明了一个实施例，使用从如上文所描述的包含文本文档的电子信息数据库生成的查询对象矢量，生成关系网络。响应由用户输入的搜索查询关键词项，可以基于搜索查询关键词项，根据从查询对象矢量提取的关系，生成关系网络。在某些实施例中，关系网络将由关键词项的展开的矢量的网络、它们彼此的连接以及这些连接的强度构成，其中，连接基于所定义的框架内共享的属性。虽然所显示的示例流程图使用文本文档以及关键词项讨论了一个实施例，但是，在其他实施例中，查询关键词项可以是音频数据、视频数据、图像数据，或任何其他种类的电子数据。

[0097] 首先，用户在状态 701 下向系统提交至少一个查询关键词项。在某些实施例中，可以向系统提交多个关键词项，多个关键词项可以当做一个查询关键词项或多个查询关键词项。在某些实施例中，如果在信息数据库中不存在 Q，那么，系统不会返回任何数据。响应接收到的查询，系统在状态 702 下检索查询关键词项的矢量，即，查询对象矢量（“QOV”）。然后，过程 700 进入状态 703，其中，用户或系统配置一个与查询一起使用的过滤器，以便聚焦查询结果。此过滤器可以在状态 703 下通过例如，从为搜索关键词项 Q 检索到的矢量中过滤掉关键词项来进行设置。下面将参考图 8A 更加详细地对此进行讨论。接下来，系统在状态 704 下将矢量展开为展开的 QOV。下面将参考图 8B 更加详细地对此过程进行讨论。然后，过程 700 进入状态 705，其中，系统使用 QOV 生成展开的关联的对象矢量（“AOV”）。下面将参考图 8C 更加详细地对此进行讨论。然后，系统进入状态 706，以查找展开的 AOV 和展开的 QOV 之间的关联的关键词项。然后，在状态 707 下提供查询 Q 的搜索结果。下面将参考图 8D 讨论提供搜索结果的过程。最后，过程 700 基于查询结果，呈现关系网络的可视表示。

[0098] 在一个实施例中，系统使用诸如相关主题和类别的本体论的形式之类的过滤器，来控制在搜索过程中推导出的关系的类型，确保当正在构建关系网络时关键词项停留在某

一定义的上下文内。在某些实施例中，可以使用过滤器，因为为过滤器选择的关键词项也存在于正在被搜索的信息数据库中，如此，过滤器关键词项具有它们自己的矢量。过滤器可以与查询一起提供，以便聚焦查询结果。过滤器可以单词、符号或对象的列表，通过它们，对查询的结果进行控制。例如，可以使用过滤器短语“基因和与药物的推断的关系”来对与遗传数据相关的信息数据库进行的基因组搜索。

[0099] 在某些实施例中，过滤器可以是完整的矢量，其中，其元素表示整个框架数据集或文档数据库中的上下文，以控制关系提取过程。根据所使用的过滤器的类型，处理被发现与矢量过滤器交叉的任何搜索结果。

[0100] 可以有许多不同种类的过滤器与这里所说明的系统和方法一起使用。一种过滤器，即，排除过滤器，可以主动地删除与过滤器不匹配的关键词项和矢量。排除过滤器可以用来确保从查询对象矢量和过程的任何方面的关联的对象矢量中删除特定主题的元素。图 8A 显示了包含关键词项 Z_1 到 Z_n 的示例排除过滤器矢量 810。向为查询 Q801 检索到的查询对象矢量 820 应用过滤器矢量，以便聚焦查询的结果。如图 8A 所示，系统有利地删除了关键词项的出现在过滤器矢量中的实例。已经从最终的查询对象矢量 825 中过滤掉了关键词项 Z_1、Z_2 和 Z_3，因为这些关键词项出现在排除过滤器 810 中。

[0101] 另一方面，选择型过滤器可以主动地选择与过滤器匹配的关键词项和矢量。选择型过滤器可以用来确保只有特定主题的元素可以用于特定过程。在一个实施例中，过程包括选择顶部查询关键词项矢量元素和关联的关键词项矢量元素，以便生成展开的查询关键词项矢量和关联的关键词项矢量。过滤器元素还将用于展开的查询关键词项矢量的最终的关键词项的选择，用在展开的关联比较和关联分数计算上。

[0102] 另一种过滤器，加权过滤器，可以调节某些关键词项和矢量的关系分数，以便对关键词项或矢量进行重新排序。加权过滤器可以用来修改特定的关键词项的组，从而影响它们对算法过程和计算结果的影响。

[0103] 可以有利地在系统正在响应查询展开检索到的查询对象矢量的任何一点应用过滤器。使用过滤器会导致系统能够将关系基于可以包括主题的特定的关键词项的集合。如果没有主题过滤，系统可以检索各种各样的推断的关系，如果知道要查找什么样的关系，这是没有益处的。例如，向信息数据库提交搜索查询关键词项“red”而没有过滤器的用户可能接收到范围非常宽的结果。另一方面，如果用户使用选择型过滤器（它们将排除过滤器中未找到的所有关键词项），如过滤器短语或矢量“flowers”作为“red”的上下文，则在查询结果中将最有可能发现涉及红色的植物群的特定的关键词项。在某些实施例中，过滤器可以是预先定义的，并且是可互换的，以便允许用户修整搜索查询。通过利用这种上下文控制创建关键词项关系的网络，可以使得预先未识别的连接被放在显著位置，因为系统的用户可能希望查找在指定的上下文中存在与此查询关键词项的什么关系。

[0104] 图 8B 是显示了使用图 8A 的经过滤的 QOV 825 生成展开的 QOV 850 的一个示范性方法的数据流程图。首先，系统识别与查询关键词项 Q 801 相关的最强的三十个关键词项 A1 到 A30 826。将这三十个最强关键词项添加到展开的 QOV 850 的开始 826 中。接下来，系统检索这三十个关键词项 A1 到 A30 830 中的每一个关键词项的矢量，并插入这三十个矢量 831（即，A_1 的 $A_{1,1}$ 到 $A_{1,3}$，A_2 的 $A_{2,1}$ 到 $A_{2,3}$，...A_{10} 的 $A_{10,1}$ 到 $A_{10,3}$）中的每一个矢量中的前三个最强的关键词项，以完成展开的 QOV 850。虽然所显示的系统的实施例选择了三十个关

键词项进行处理，但是，在其他实施例中，也可以使用任何其他关键词项数来进行处理。

[0105] 图 8C 是显示了使用图 8A 的经过滤的 QOV 825 生成展开的 AOV 875 的一个方法的数据流程图。首先，系统识别与 Q 801 相关的三十个最强的关键词项 A_1 到 A_{30} 826，检索它们的矢量 827，并开始每一个关键词项 A_1 到 A_{30} 的展开的 AOV 875。然后，系统从与 A_1 到 A_{30} 中的每一个（即，A_1 的 $A_{1,1}$ 到 $A_{1,3}$，A_2 的 $A_{2,1}$ 到 $A_{2,3}$，...A_{30} 的 $A_{30,1}$ 到 $A_{30,3}$）830 相关的第一维矢量识别三个最强的关键词项，将这些关联的关键词项添加到对应的展开的 AOV 875A_1 到 A_{30} 中，并检索它们的矢量 831。类似地，系统从与每一个 $A_{1,1}$ 到 $A_{30,3}$（即，$A_{1,1}$ 的 $A_{1,1,1}$ 到 $A_{1,1,3}$，$A_{1,2}$ 的 $A_{1,2,1}$ 到 $A_{1,2,3}$，...$A_{30,3}$ 的 $A_{30,3,1}$ 到 $A_{30,3,3}$）840 相关的第二维矢量检索三个最强的关键词项，并检索它们的矢量 841。再次，系统从与每一个 $A_{1,1,1}$ 到 $A_{30,3,3}$（即，$A_{1,1}$ 的 $A_{1,1,1,1}$ 到 $A_{1,1,1,3}$，$A_{1,2}$ 的 $A_{1,1,2,1}$ 到 $A_{1,1,2,3}$，...$A_{30,3,3}$ 的 $A_{30,3,3,1}$ 到 $A_{30,3,3,3}$）850 相关的第三维矢量检索三个最强的关键词项。然后，记那个来自第三维矢量 850 的前三个关联的关键词项插入在已经在展开的 AOV 875 中的第一维关键词项 830 之后，以完成展开的 AOV 875。虽然图 8C 显示了生成 A_1 的展开的 AOV 875，但是，在所显示的实施例中，为每一个 A_1 到 A_{30} 826 总共产生了 30 个展开的 AOV。

[0106] 图 8D 是显示了与展开的 QOV 850 一起使用展开的 AOV875 的一个示范性方法的数据流程图，该方法用于查找 AOV 875 和展开的 QOV 850 之间的关联的关键词项，以便产生查询 Q 801 的搜索结果。展开的矢量 850 和 875 被传递到确定展开的矢量 850 和 875 中的交叉关键词项之间的相似性的函数。在一个实施例中，如图 8D 所示，系统本该取每一个展开的 AOV 875 和 QOV 850 的交叉，以便定位查询关键词项 Q 801 的关联的关键词项 880。在其他实施例中，也可以使用其他函数来定位关联的关键词项。

[0107] 在某些实施例中，在定位 Q 的关联的关键词项之后，可以计算查询关键词项 Q 和每一个关联的关键词项之间的相似性分数。然后，可以按照各自的相似性分数值，对关联的关键词项进行排序，以便将具有最高相似性分数的关联的关键词项排在第一位。在某些实施例中，相似性分数函数可以是相关系数距离测量，其值可以作为表示关联的关键词项和初始查询关键词项之间的最终的相似性测量的分数，分配到所产生的匹配关键词项，即，有多少结果匹配初始查询关键词项。

[0108] 在一个实施例中，可以通过求交叉关键词项的关系分数的总和，并将它乘以只包括交叉关键词项的矢量的长度，来计算两个矢量之间的相似性分数。在另一个实施例中，两个矢量之间的相似性分数可以是使用下列公式的相关系数距离测量函数：

[0109] $n\left(\sum_{i=1}^{n}(V \cap W)_k\right)$ 或

[0110] $\|X\| \sum_{i=1}^{n} X_k$

[0111] 其中

[0112] $X = (V \cap W)_k$

[0113] V = 查询矢量，而

[0114] W = 与查询矢量相比的任何矢量。

[0115] 在另一个实施例中，可以使用非居中的皮尔逊相关系数距离测量来计算不同大小的矢量之间的相似性分数，其中：

[0116] $$r_U = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i}{\sigma_x^{(0)}}\right)\left(\frac{y_i}{\sigma_y^{(0)}}\right)$$

[0117] 其中

[0118] $$\sigma_x^{(0)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

[0119] 并且其中距离由下列公式进行定义

[0120] $d_U \equiv 1-r_U$

[0121] 在某些实施例中，在定位查询结果关键词项 880 之后，返回查询的每一个元素的矢量，还提取并比较这些矢量，评定相似性分数。此步骤有利地允许通过交叉结果的矢量的内容来使结果形成网络。可以使用通过交叉形成的网络来确定初始查询结果如何相关，它们在什么上下文中相关，它们的连接是直接的还是间接的，以及它们的关系的强度。

[0122] 查询结果数据和使用该数据构建的关系网络，可以如此有益地使用相似性分数，显示查询关键词项 801 与其他关键词项的关系、矢量彼此之间的关系，以及它们的关系的强度。在某些实施例中，如有必要，可以可视化查询结果关键词项 880 和 / 或查询相关的矢量的所产生的关系网络，供进一步解释。例如，图 9 显示了响应关键词项“red”的查询创建的关系网络的图形可视化 900（没有按比例绘制）。与关键词项“red”具有较高的关系分数的关键词项与“red”更靠近，如“cardinal”654。具有较低的关系分数的关键词项在较远处，如“paste”655。用户可以有利地使用类似于图 9 的可视化，以便快速地理解信息数据库中的关键词项之间的关系。

[0123] III. 示例系统组件

[0124] 图 10 显示了根据一个实施例的关系网络系统 1000。关系网络系统 1000 包括 Web 服务器 1010，该 Web 服务器 1010 生成主机网站的页面，并向最终用户的计算设备 1020 提供这些页面。虽然被描述为台式计算机 1020，但是，计算设备 1020 也可以包括各种其他种类的设备，如蜂窝电话和个人数字助理 (PDA)。Web 服务器 1010 可以作为单个物理服务器来实现，也可以作为物理服务器的集合来实现。某些实施例也可以以另一种多用户交互式系统来实现，如交互式电视系统、在线服务网络或基于电话的系统，其中，用户通过电话小键盘进行输入和 / 或语音，来选择要获取的项目。

[0125] Web 服务器 1010 给用户提供对数据库或数据库集合 1020 内表示的电子信息进行访问的途径。在该 Web 服务器上运行的，或与该 Web 服务器相关联的信息获取处理器 1015，给用户提供输入他们希望查找的信息的搜索查询的功能。在一个实施例中，在数据库 1020 中表示的信息可以包括文档、字符、单词、图像、歌曲或视频或可以以电子方式存储的任何其他数据。数据库中可以存储成千上万或数百万字节的数据。

[0126] 在一个实施例中，可以使用信息获取处理器 1015 来检索信息数据库 1020 中的文档或其他对象。可以通过，例如，利用信息获取处理器 1015 来对项目进行搜索，或通过从浏览树列表中选择对象，定位每一个对象。

[0127] 如图 10 所示，关系网络系统 1000 包括关系处理器 1030，该关系处理器 1030 除了其他任务之外，还负责为信息数据库 1020 中的数据创建关系矢量。然后，这些关系矢量存储在关系数据库 1040 中。在某些实施例中，关系处理器 1030 定期运行，并笼统地分析或

“挖掘”信息数据库，以便响应可以存储在信息数据库 1020 中的新的数据，创建和维持关系数据库 1040。

[0128] 响应由信息获取处理器 1015 接收到的查询，关系网络系统 1000 向网络生成器 1050 发送查询，该网络生成器 1050 除查询之外还从关系数据库 1030 接收关系矢量信息，以便基于查询生成关系网络。在某些关系网络系统实施例中，可以对创建的关系的数量施加固定的限制，以便满足可以在 Web 空间创建的大量的关系，如上文所讨论的。

[0129] 然后，向查询结果处理器 1060 发送所产生的关系网络，而查询结果处理器 1060 对结果进行处理，可选地创建关系网络的可视表示，并将此数据发送到信息获取处理器 1015。然后，可以通过因特网将结果数据返回到提交了查询的计算设备 1020。

[0130] IV. 示例：音乐数据库

[0131] 可以实现本发明的一个实施例，以发现与音乐的数据库相关的人生成的内容之间的关系。播放列表、博客，以及推荐列表是涉及音乐的人生成的内容的一些示例。系统可以基于音乐文件在诸如因特网之类的大的数据空间上的目录或知识库内的位置，确定音乐文件之间的关系。此关系数据可以包括诸如艺术家、专辑、歌曲的标题以及发行年份之类的信息，可以存储在关联存储器模块中，然后传输到查询对象矢量，如上文所描述的。然后，响应诸如针对艺术家或歌曲之类的查询，系统可以创建相关的艺术家或歌曲的关系网络，并呈现给查询，可选地可视化关系网络。

[0132] V. 结束语

[0133] 上文所描述的所有特征都可以以由通用计算机执行的软件模块实现，并自动化。软件模块可以存储在任何种类的计算机存储设备或介质中。这里所描述的各种实施例和特征的所有组合都在本发明的范围内。

[0134] 虽然是利用某些优选实施例来描述各种发明特征和服务的，但是，对那些精通本技术的普通人员显而易见的其他实施例，包括没有提供这里所阐述的所有优点和特征的实施例并没有解决这里所阐述的所有问题的实施例，也都在本发明的范围内。例如，虽然上文所描述的某些示例涉及向用户呈现搜索查询结果，但是，本发明也可以用于其他系统中，如拼写检查程序、金融关系网络、基因解释，或基于搜索查询结果向用户呈现广告。本发明的范围只能通过参考所附的权利要求来进行定义。

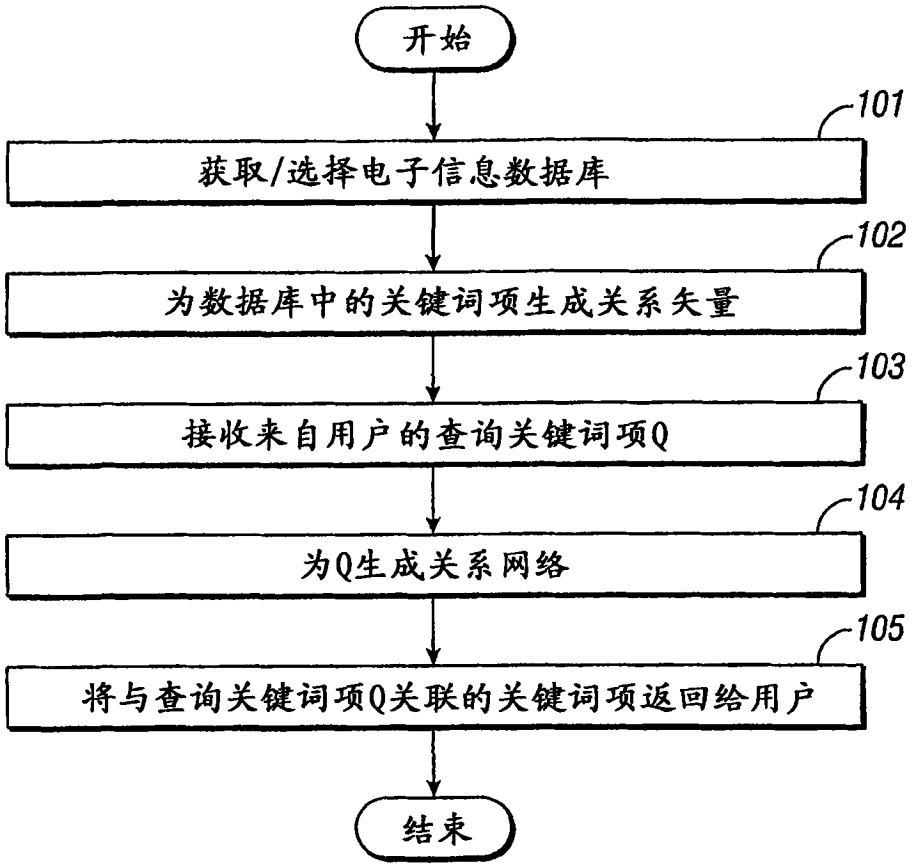

开始
101
获取/选择电子信息数据库
102
为数据库中的关键词项生成关系矢量
103
接收来自用户的查询关键词项Q
104
为Q生成关系网络
105
将与查询关键词项Q关联的关键词项返回给用户
结束

图1

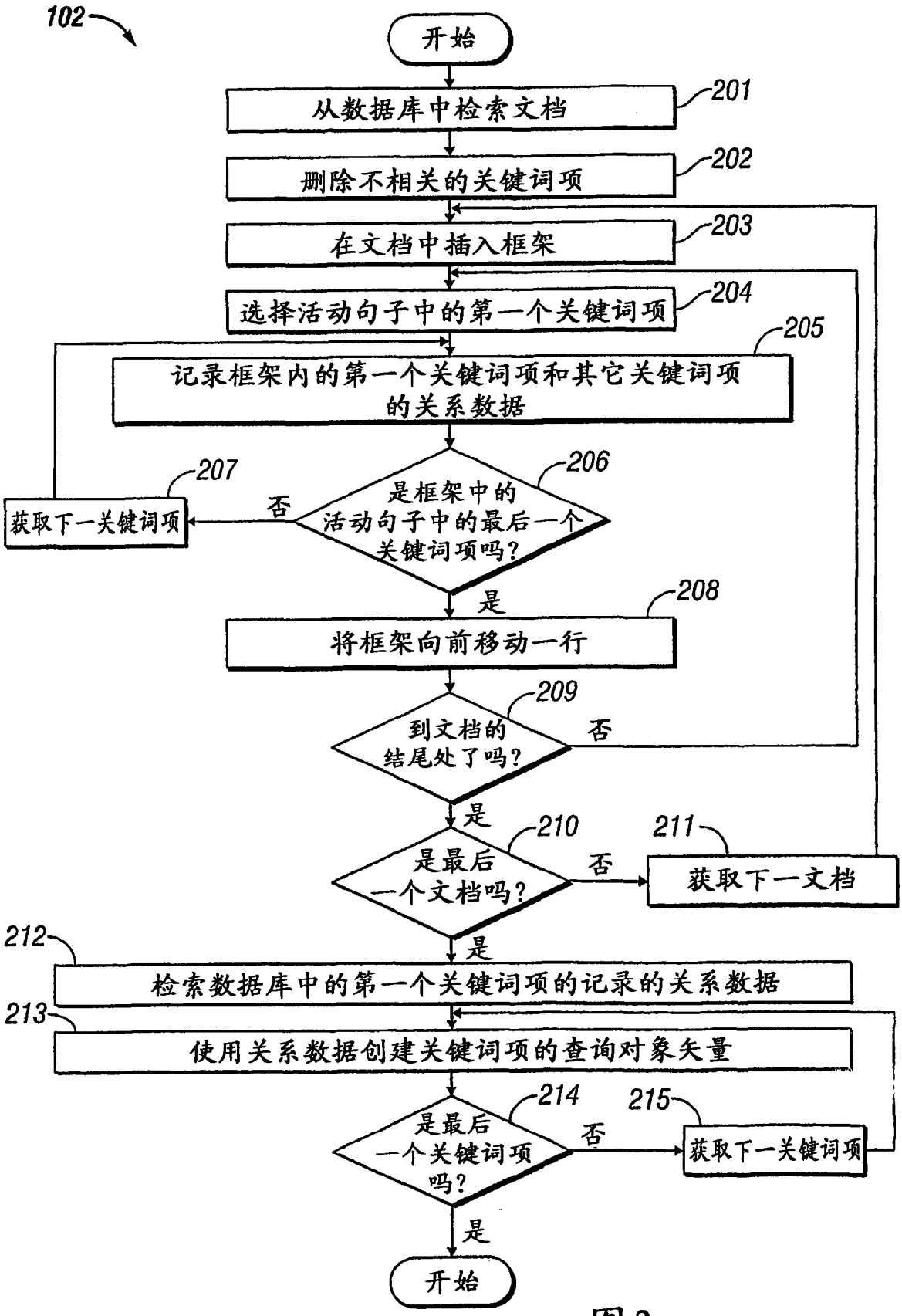

102
开始
从数据库中检索文档
201
删除不相关的关键词项
202
在文档中插入框架
203
选择活动句子中的第一个关键词项
204
记录框架内的第一个关键词项和其它关键词项的关系数据
205
是框架中的活动句子中的最后一个关键词项吗?
206
否
获取下一关键词项
207
是
将框架向前移动一行
208
到文档的结尾处了吗?
209
否
是
是最后一个文档吗?
210
否
获取下一文档
211
是
检索数据库中的第一个关键词项的记录的关系数据
212
使用关系数据创建关键词项的查询对象矢量
213
是最后一个关键词项吗?
214
否
获取下一关键词项
215
是
开始

图 2

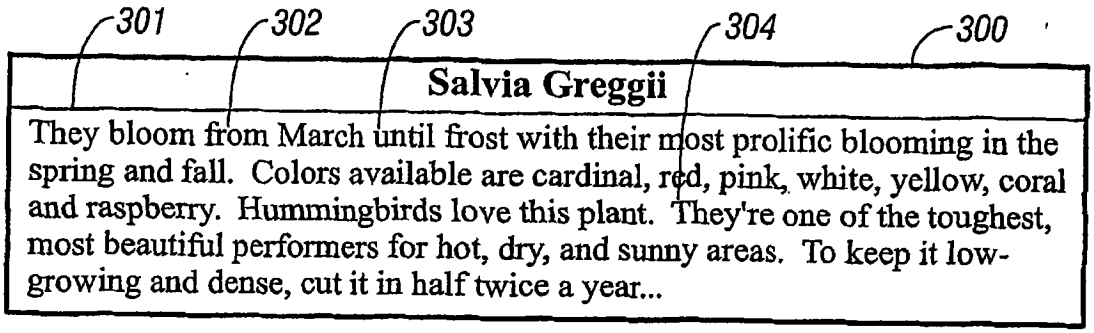

301
302
303
304
300
Salvia Greggii
They bloom from March until frost with their most prolific blooming in the spring and fall. Colors available are cardinal, red, pink, white, yellow, coral and raspberry. Hummingbirds love this plant. They're one of the toughest, most beautiful performers for hot, dry, and sunny areas. To keep it low-growing and dense, cut it in half twice a year...

图 3A

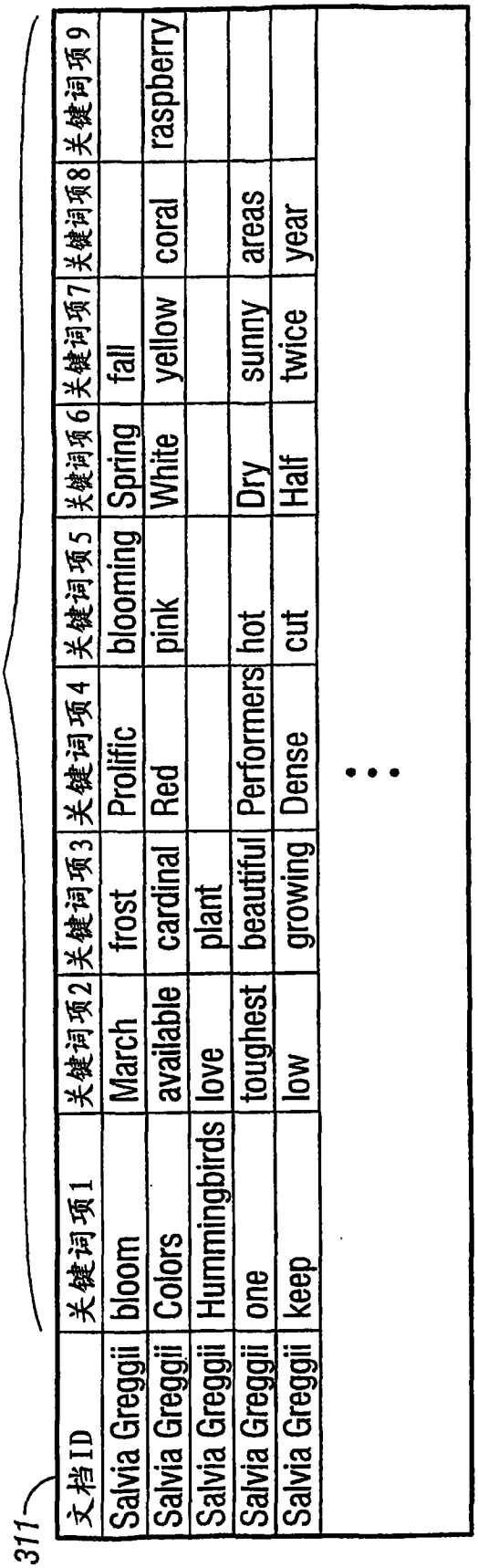

310
312
311
文档ID
关键词项1
关键词项2
关键词项3
关键词项4
关键词项5
关键词项6
关键词项7
关键词项8
关键词项9
Salvia Greggii
bloom
March
frost
Prolific
blooming
Spring
fall
Salvia Greggii
Colors
available
cardinal
Red
pink
White
yellow
coral
raspberry
Salvia Greggii
Hummingbirds
love
plant
Salvia Greggii
one
toughest
beautiful
Performers
hot
Dry
sunny
areas
Salvia Greggii
keep
low
growing
Dense
cut
Half
twice
year

图 3B

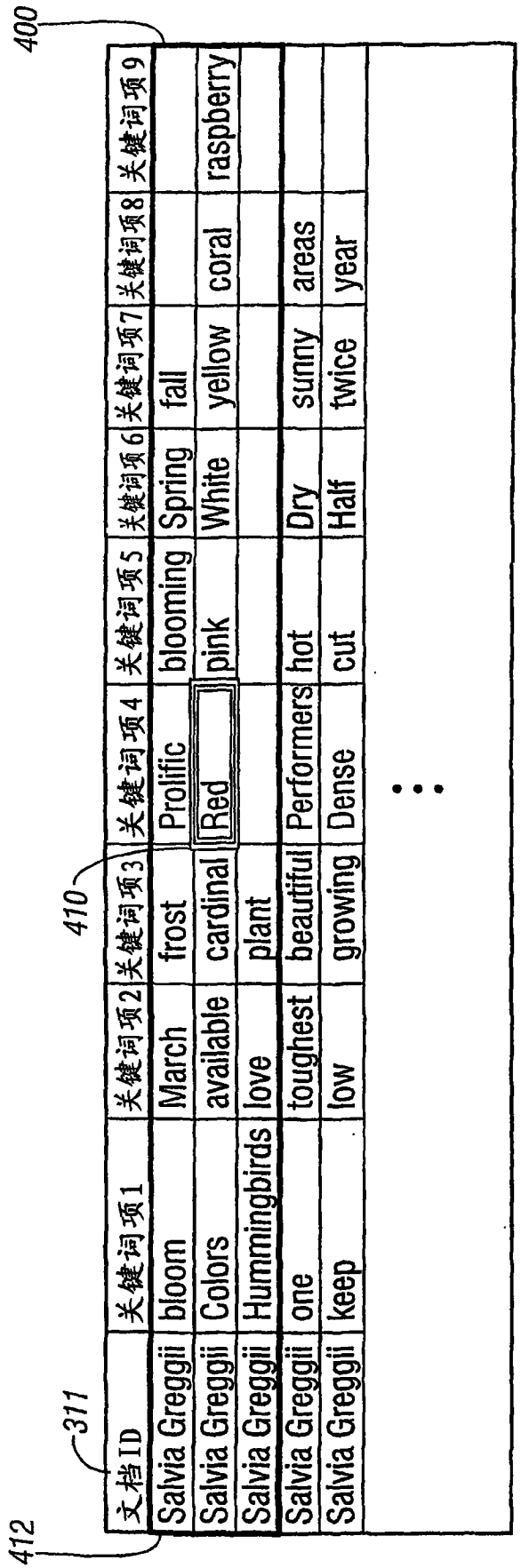

400
410
311
412
文档ID
关键词项1
关键词项2
关键词项3
关键词项4
关键词项5
关键词项6
关键词项7
关键词项8
关键词项9
Salvia Greggii
bloom
March
frost
Prolific
blooming
Spring
fall
Salvia Greggii
Colors
available
cardinal
Red
pink
White
yellow
coral
raspberry
Salvia Greggii
Hummingbirds
love
plant
Salvia Greggii
one
toughest
beautiful
Performers
hot
Dry
sunny
areas
Salvia Greggii
keep
low
growing
Dense
cut
Half
twice
year

图 4

500

510 关键词项统计信息：		
511 # 分析的文本中的出现次数	1	
512 # 句子数	1	
513 # 关联的关键词项数	18	
514 # 单个关联数	18	
520 文档统计信息：	522	523
521 文档ID	# 句子数	总分
524 Salvia Greggii	1	1
530 关联的关键词项统计信息：	525	526
	532	533
531 关联的关键词项	# 关联数	距离分数
534 Cardinal	1	0.618
Pink	1 (535)	0.618 (536)
White	1	0.381
⋮		
Plant	1	0.021
March	1	0.013
Bloom	1	0.008

537

图5

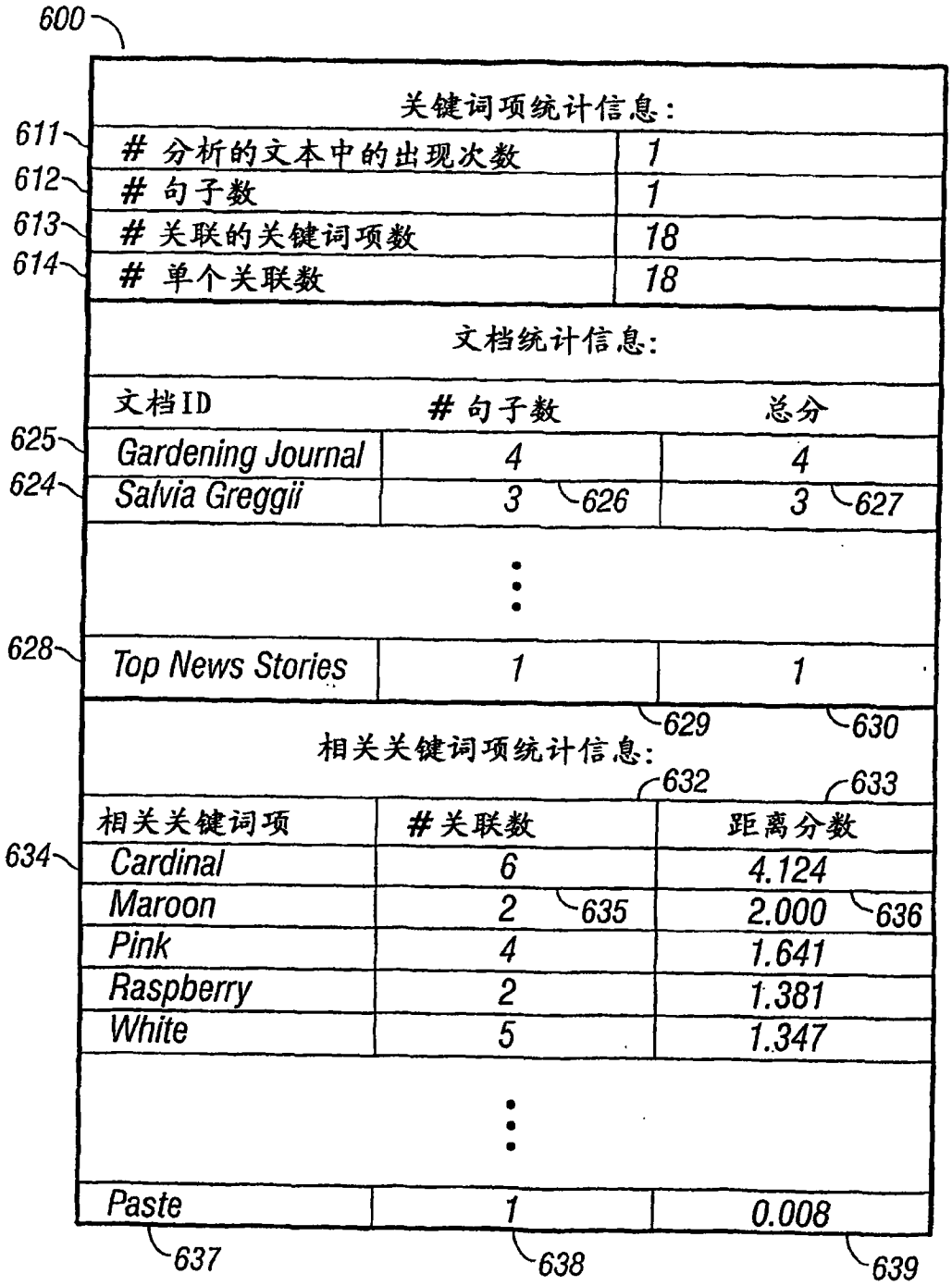
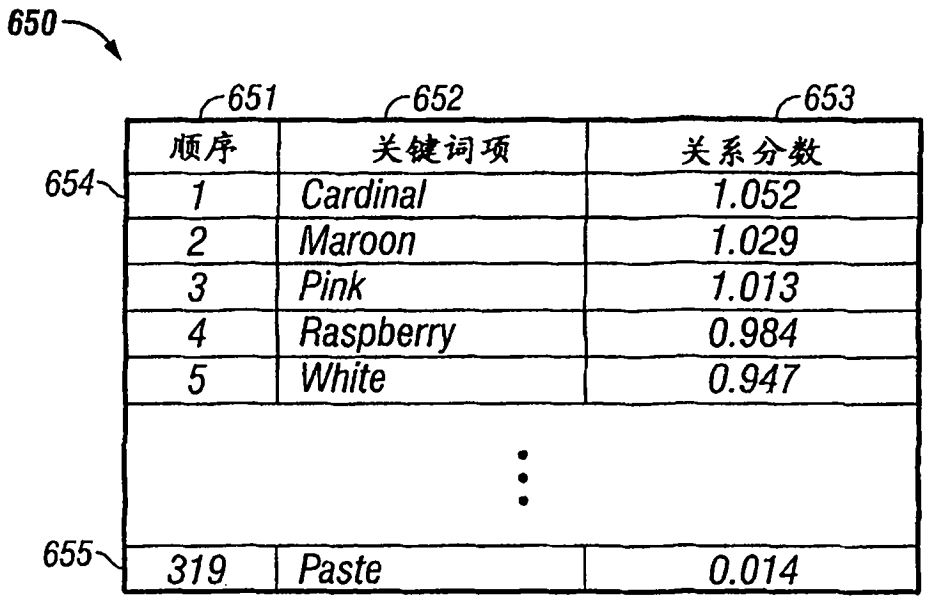

600
关键词项统计信息：
611
分析的文本中的出现次数
1
612
句子数
1
613
关联的关键词项数
18
614
单个关联数
18
文档统计信息：
文档ID
句子数
总分
625
Gardening Journal
4
4
624
Salvia Greggii
3
626
3
627
628
Top News Stories
1
1
629
630
相关关键词项统计信息：
632
633
相关关键词项
#关联数
距离分数
634
Cardinal
6
4.124
Maroon
2
635
2.000
636
Pink
4
1.641
Raspberry
2
1.381
White
5
1.347
Paste
1
0.008
637
638
639

图6A



顺序	关键词项	关系分数
1	Cardinal	1.052
2	Maroon	1.029
3	Pink	1.013
4	Raspberry	0.984
5	White	0.947
⋮		
319	Paste	0.014

图6B

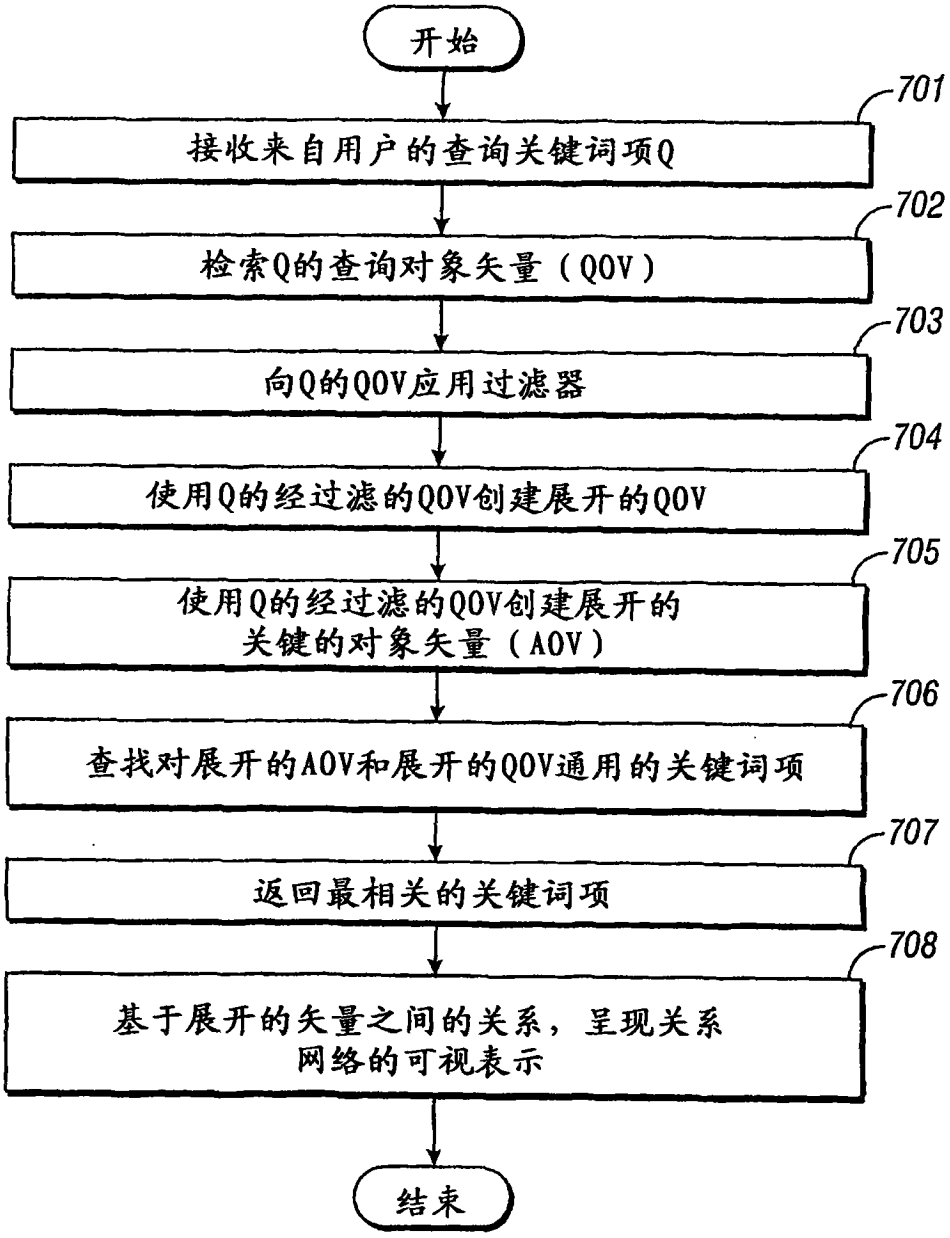

开始
701
接收来自用户的查询关键词项Q
702
检索Q的查询对象矢量（QOV）
703
向Q的QOV应用过滤器
704
使用Q的经过滤的QOV创建展开的QOV
705
使用Q的经过滤的QOV创建展开的
关键的对象矢量（AOV）
706
查找对展开的AOV和展开的QOV通用的关键词项
707
返回最相关的关键词项
708
基于展开的矢量之间的关系，呈现关系
网络的可视表示
结束

图7

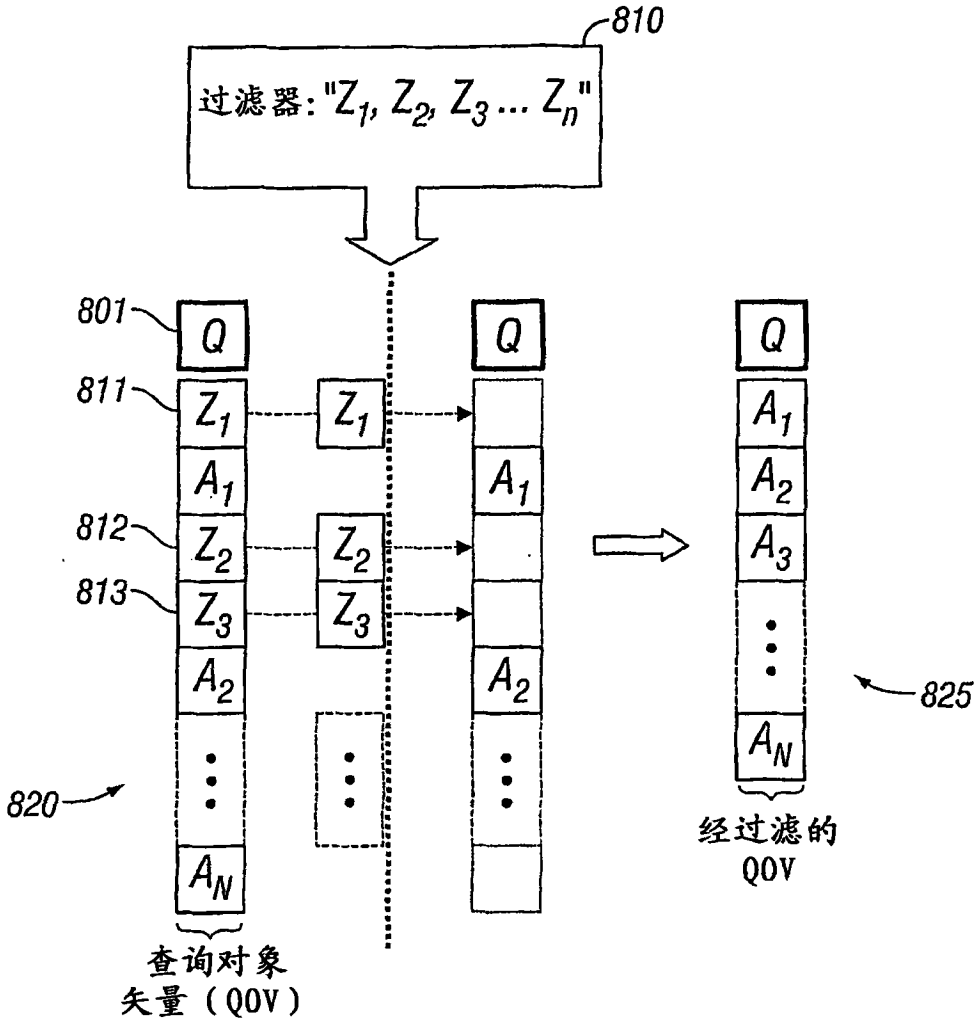

810
过滤器："Z_1, Z_2, Z_3 ... Z_n"
801
Q
811
Z_1
A_1
812
Z_2
813
Z_3
A_2
A_N
820
查询对象
矢量（QOV）
Q
Z_1
Z_2
Z_3
A_1
A_2
Q
A_1
A_2
A_3
A_N
825
经过滤的
QOV

图8A

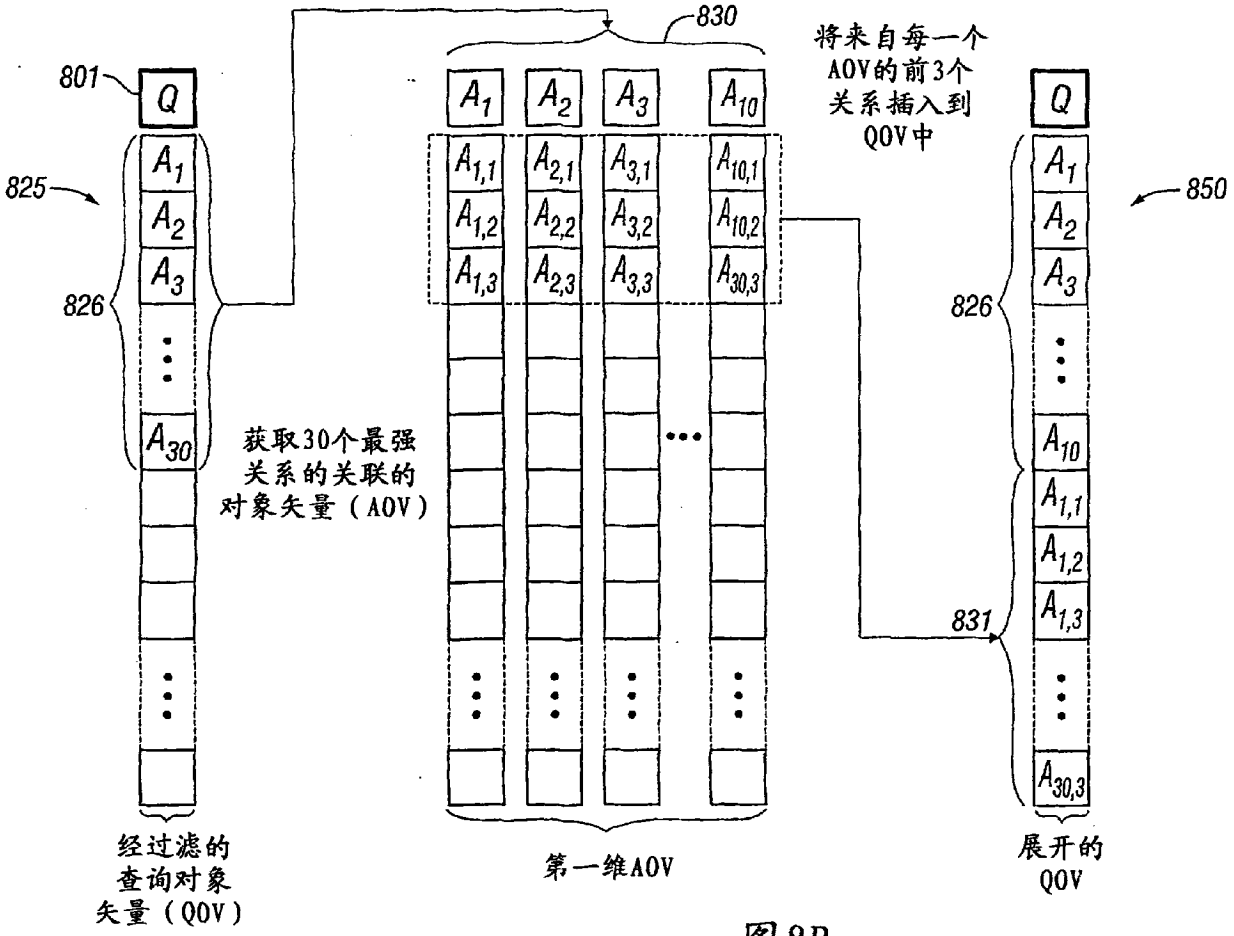

801
Q
825
826
A_1
A_2
A_3
A_{30}
830
A_1 A_2 A_3 A_{10}
$A_{1,1}$ $A_{2,1}$ $A_{3,1}$ $A_{10,1}$
$A_{1,2}$ $A_{2,2}$ $A_{3,2}$ $A_{10,2}$
$A_{1,3}$ $A_{2,3}$ $A_{3,3}$ $A_{30,3}$
获取30个最强关系的关联的对象矢量（AOV）
将来自每一个AOV的前3个关系插入到QOV中
850
831
A_{10}
$A_{1,1}$
$A_{1,2}$
$A_{1,3}$
$A_{30,3}$
经过滤的查询对象矢量（QOV）
第一维AOV
展开的QOV

图8B

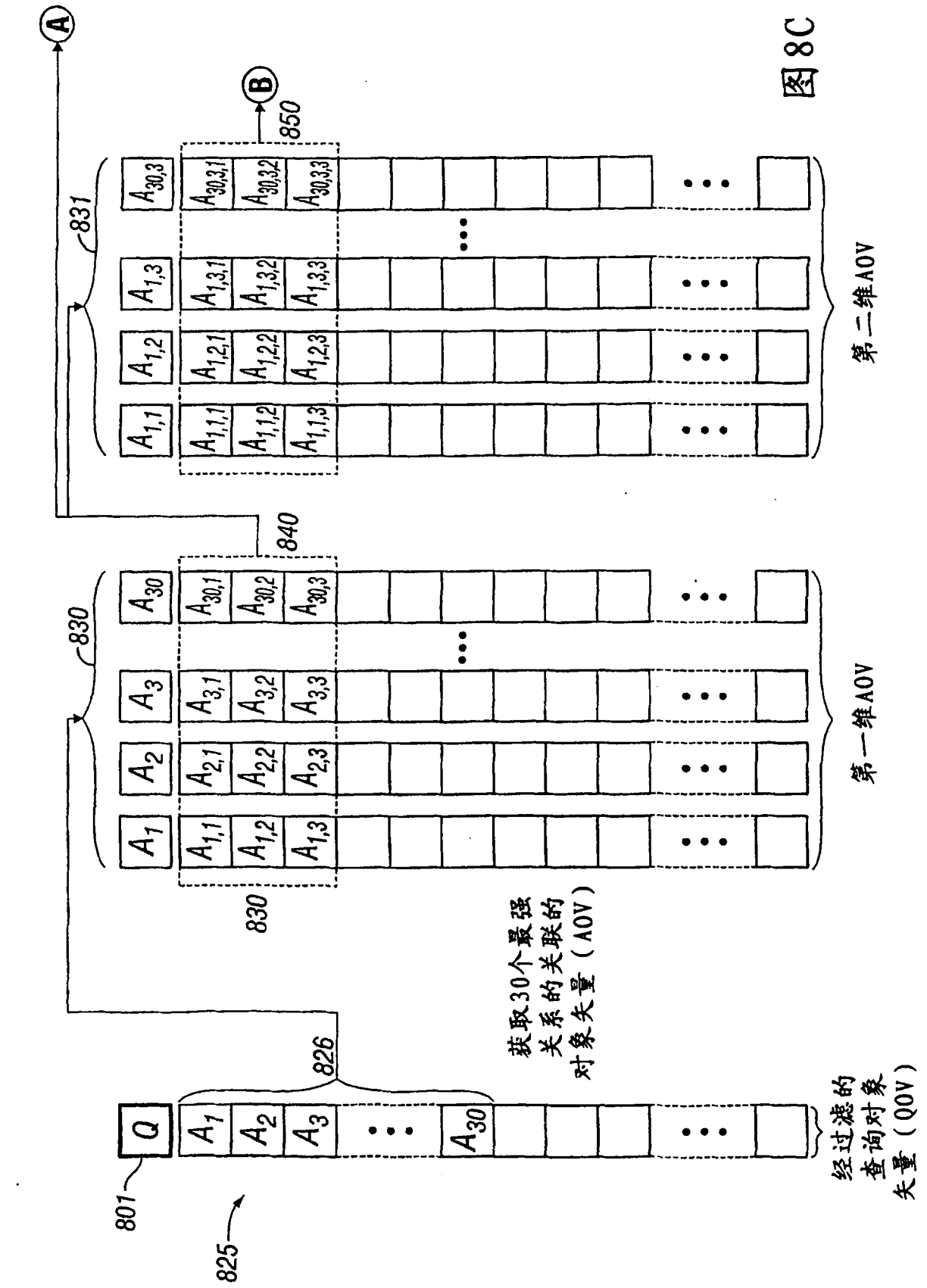

A
B
850
831
840
830
826
825
801
Q
A_1 A_2 A_3 ... A_{30}
$A_{1,1}$ $A_{1,2}$ $A_{1,3}$
$A_{2,1}$ $A_{2,2}$ $A_{2,3}$
$A_{3,1}$ $A_{3,2}$ $A_{3,3}$
$A_{30,1}$ $A_{30,2}$ $A_{30,3}$
$A_{1,1,1}$ $A_{1,1,2}$ $A_{1,1,3}$
$A_{1,2,1}$ $A_{1,2,2}$ $A_{1,2,3}$
$A_{1,3,1}$ $A_{1,3,2}$ $A_{1,3,3}$
$A_{30,3,1}$ $A_{30,3,2}$ $A_{30,3,3}$
第一维AOV
第二维AOV
获取30个最强关系的关联的对象矢量（AOV）
经过滤的查询对象矢量（QOV）

图8C

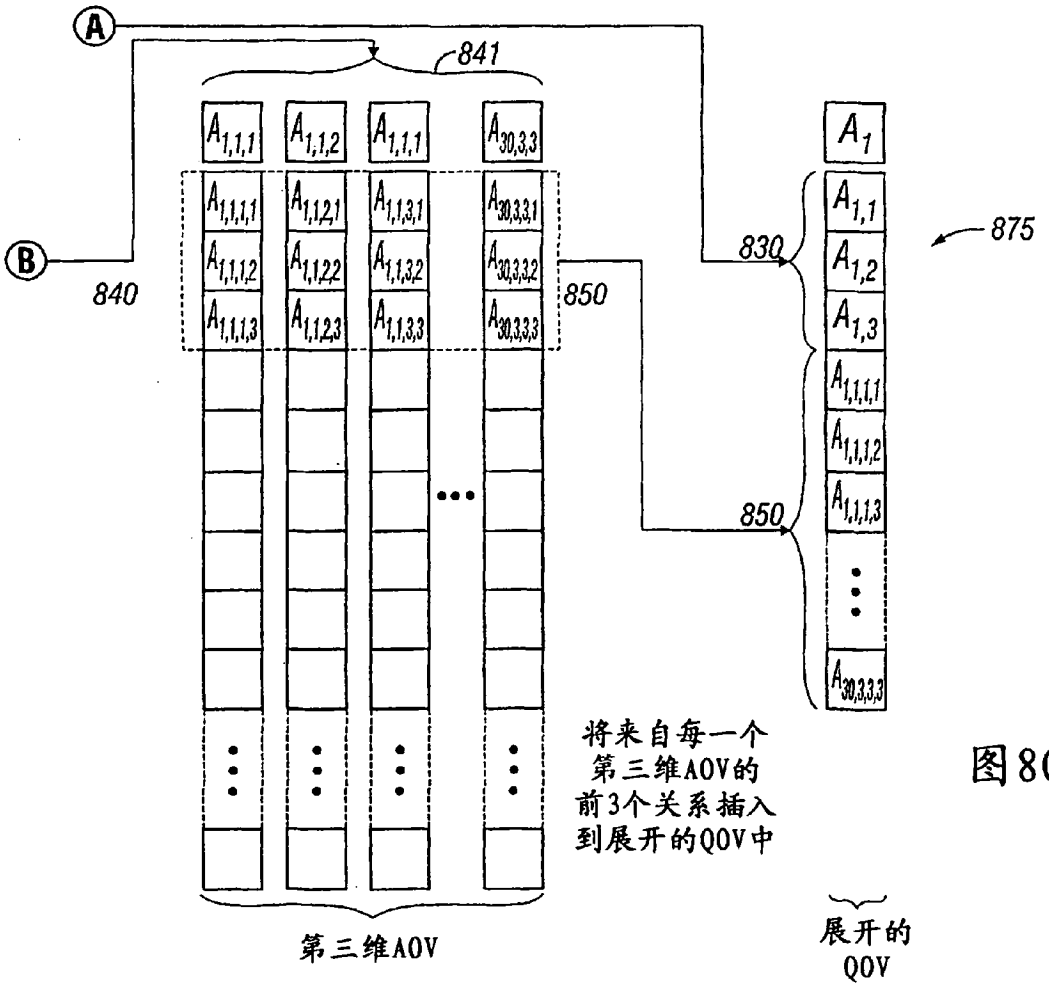

A
B
841
840
850
830
850
875
将来自每一个第三维AOV的前3个关系插入到展开的QOV中
第三维AOV
展开的QOV

图8C（续）

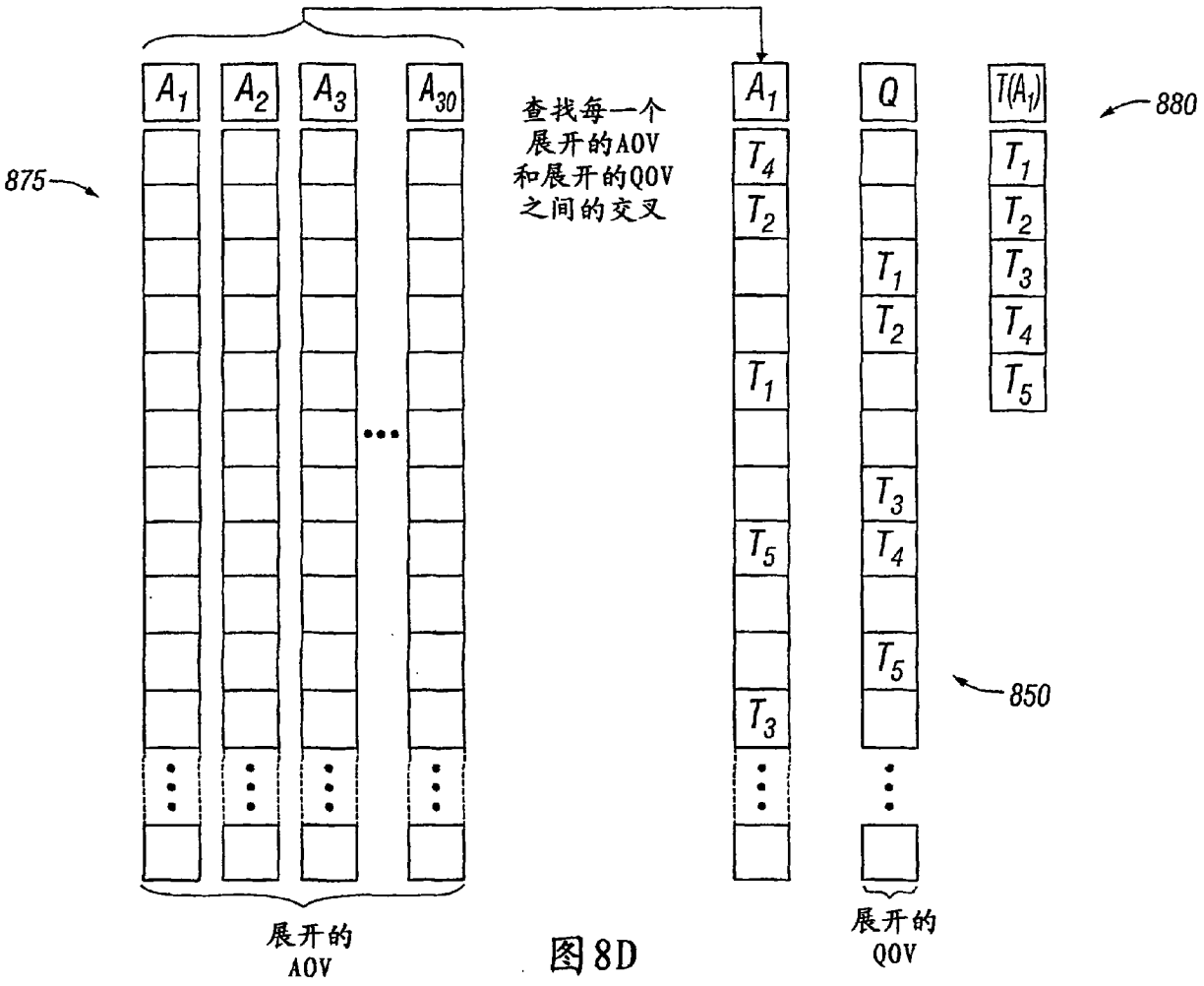

875
A_1
A_2
A_3
A_{30}
展开的
AOV
查找每一个
展开的AOV
和展开的QOV
之间的交叉
A_1
T_4
T_2
T_1
T_5
T_3
Q
T_1
T_2
T_3
T_4
T_5
850
展开的
QOV
$T(A_1)$
T_1
T_2
T_3
T_4
T_5
880

图8D

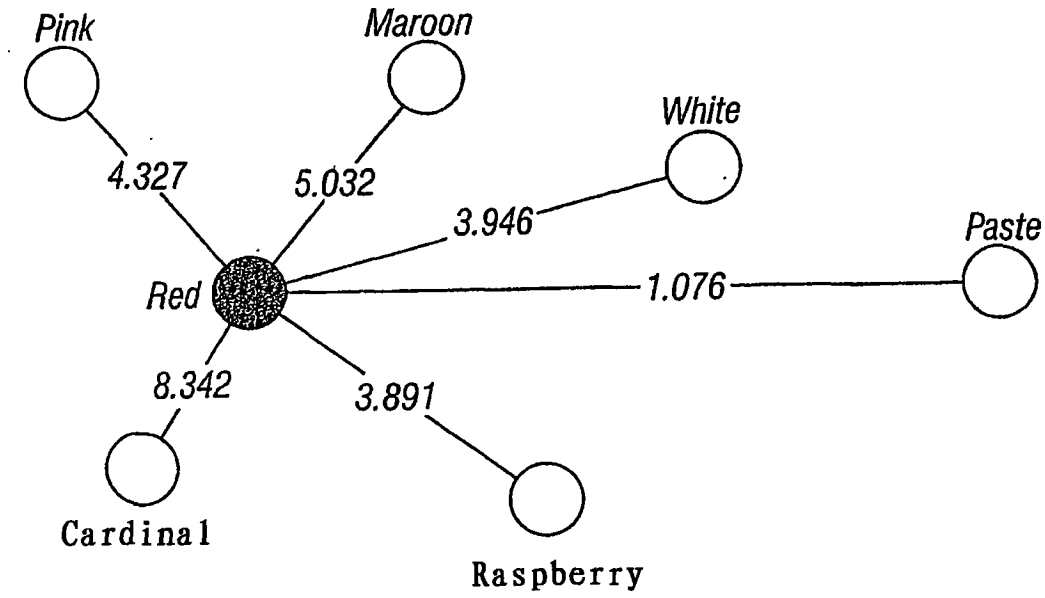

Pink
Maroon
White
4.327
5.032
3.946
Paste
Red
1.076
8.342
3.891
Cardinal
Raspberry

图 9

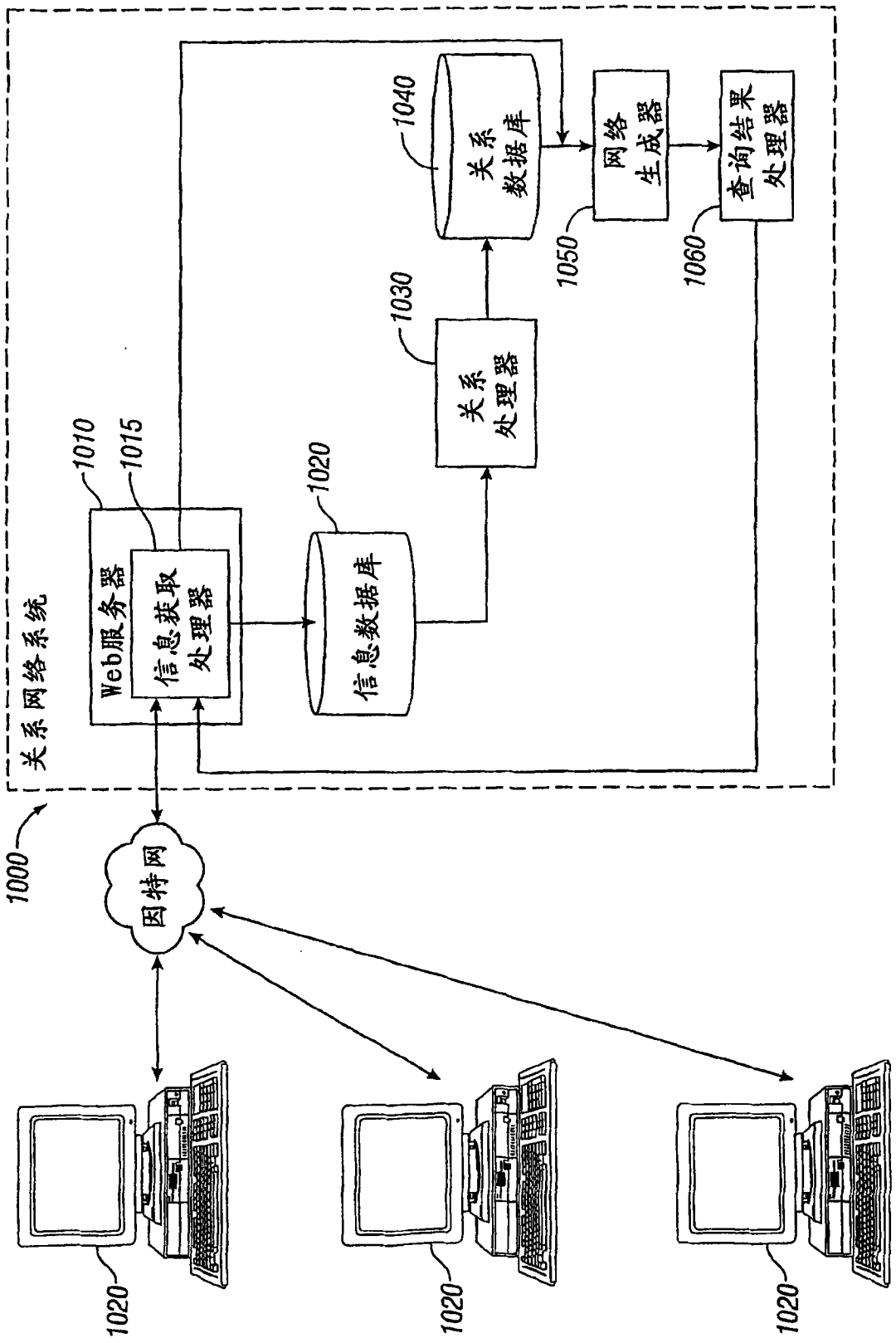

关系网络系统
1000
因特网
Web服务器
信息获取处理器
1010
1015
信息数据库
1020
关系处理器
1030
关系数据库
1040
网络生成器
1050
查询结果处理器
1060
1020
1020
1020

图10