(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 111699482 A
(43)申请公布日 2020.09.22

(21)申请号 201880089097.4

(22)申请日 2018.12.11

(30)优先权数据
62/597,012 2017.12.11 US
16/215,322 2018.12.10 US

(85)PCT国际申请进入国家阶段日
2020.08.11

(86)PCT国际申请的申请数据
PCT/US2018/064828 2018.12.11

(87)PCT国际申请的公布数据
WO2019/118388 EN 2019.06.20

(71)申请人 脸谱公司
地址 美国加利福尼亚州

(72)发明人 马蒂斯·杜兹
亚历山大·萨布利罗勒斯
埃尔韦·耶古

(74)专利代理机构 北京安信方达知识产权代理有限公司 11262
代理人 周靖 杨明钊

(51)Int.Cl.
G06F 16/33(2006.01)
G06F 16/56(2006.01)
G06F 16/901(2006.01)
G06F 16/9536(2006.01)
G06N 3/08(2006.01)

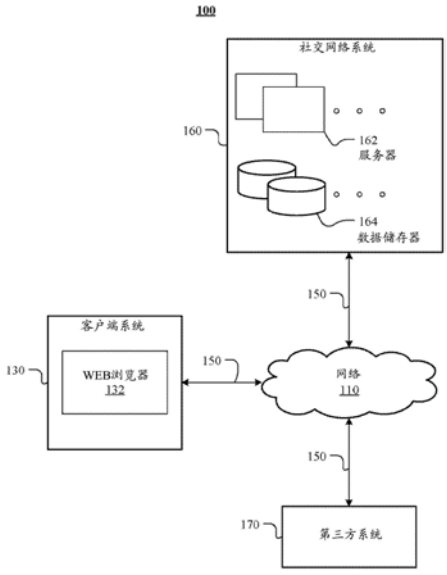
权利要求书6页 说明书41页 附图14页

(54)发明名称

在线社交网络上利用图和紧致回归码的快速索引

(57)摘要

在一个实施例中,一种方法包括接收包括查询内容对象和约束的查询,生成表示查询内容对象的特征向量,访问包括对应于由紧致码表示的候选内容对象的节点和连接这些节点的链接的稀疏图,选择入口节点,通过以下操作来迭代地选择相似内容对象:识别入口节点的链接节点,解压缩表示候选内容对象的紧致码以生成特征向量,基于表示查询内容对象的特征向量和表示候选内容对象的特征向量之间的比较来选择零个或更多个相似内容对象,如果满足完成条件,则返回所选择的相似内容对象,否则通过使用对应于最相似内容对象的链接节点作为入口节点来重复该迭代选择,以及发送用于呈现一个或更多个所选择的相似内容对象的指令。



CN 111699482 A

1.一种方法,包括由一个或更多个计算系统:

接收搜索查询,其中,所述搜索查询包括查询内容对象和一个或更多个查询约束;

生成表示所述查询内容对象的特征向量;

访问稀疏图,所述稀疏图包括对应于多个候选内容对象的多个节点和连接所述节点的多个链接,其中,连接到任何特定节点的链接的数量小于链接的阈值数量,并且其中,每个候选内容对象由紧致码表示;

从所述稀疏图的多个节点中选择入口节点;

从所述多个候选内容对象中选择关于所述查询内容对象的一个或更多个相似内容对象,其中,所述一个或更多个相似内容对象通过以下方式被迭代地选择:

识别所述入口节点的一个或更多个链接节点,其中,每个链接节点通过链接连接到所述入口节点;

对于对应于所述链接节点的每个候选内容对象,解压缩表示所述候选内容对象的紧致码,以生成表示所述候选内容对象的特征向量;

基于表示所述查询内容对象的特征向量和表示所述候选内容对象的特征向量之间的比较,从所述候选内容对象中选择零个或更多个相似内容对象;

确定是否达到选择所述一个或更多个相似内容对象的完成条件,其中,基于所述一个或更多个查询约束来确定所述完成条件;和

基于所述确定:

如果满足所述完成条件,则返回一个或更多个所选择的相似内容对象;否则

如果不满足所述完成条件,则通过使用来自所述一个或更多个链接节点的对应于最相似内容对象的链接节点作为所述入口节点来重复迭代选择;和

响应于接收到所述搜索查询,向与用户相关联的客户端系统发送用于呈现一个或更多个所选择的相似内容对象的指令。

2.根据权利要求1所述的方法,其中,所述一个或更多个查询约束包括以下项中的一个或更多个:

存储器利用率或CPU利用率约束;

选择精度约束;或者

搜索查询处理时间约束。

3.根据权利要求1所述的方法,其中,所述稀疏图的多个链接由一个或更多个机器学习算法生成。

4.根据权利要求1所述的方法,其中,基于一种或更多种编码来生成每个候选内容对象的紧致码。

5.根据权利要求4所述的方法,其中,所述一种或更多种编码包括乘积量化器或优化乘积量化器中的一个或更多个。

6.根据权利要求1所述的方法,其中,所述稀疏图中的每个节点与候选内容对象的聚类相关联,其中,候选对象的每个聚类与特征向量相关联。

7.根据权利要求6所述的方法,从所述稀疏图的多个节点中选择所述入口节点包括:

基于表示所述查询内容对象的特征向量和表示候选内容对象的所述聚类的特征向量之间的比较来选择所述入口节点。

8.根据权利要求1所述的方法，其中，解压缩表示每个候选内容对象的紧致码是基于以下项中的一个或更多个：

细化，所述细化基于表示所述候选内容对象的紧致码和与对应于所述候选内容对象的节点的一个或更多个链接节点相关联的一个或更多个紧致码；或者

表示所述候选内容对象的紧致码和包括多个回归权重向量的回归码本。

9.根据权利要求8所述的方法，其中，所述细化基于一个或更多个损失函数，并且其中，所述一个或更多个损失函数将表示所述候选内容对象的特征向量和与所述候选内容对象相关联的代码矩阵相关联，其中，所述代码矩阵包括表示所述候选内容对象的紧致码和与对应于所述候选内容对象的节点的链接节点相关联的紧致码。

10.根据权利要求8所述的方法，其中，基于所述多个候选内容对象和一个或更多个损失函数来学习所述回归码本。

11.根据权利要求10所述的方法，其中，所述回归码本包括每个候选内容对象的一组量化回归系数，并且其中，生成表示每个候选内容对象的特征向量还基于所述一组量化回归系数。

12.根据权利要求1所述的方法，其中，所述查询内容对象包括文本、音频剪辑、图像或视频剪辑中的一个或更多个。

13.根据权利要求1所述的方法，其中，表示所述查询内容对象的特征向量是基于一个或更多个机器学习算法生成的，并且其中，所述一个或更多个机器学习算法基于卷积神经网络。

14.根据权利要求1所述的方法，其中，所述入口节点的一个或更多个链接节点包括一个或更多个短程链接节点和至少一个长程链接节点。

15.根据权利要求14所述的方法，其中，所述一个或更多个短程链接节点中的每一个相对于所述入口节点具有大于或等于阈值水平的相似性水平。

16.根据权利要求14所述的方法，其中，所述至少一个长程链接节点相对于所述入口节点具有小于阈值水平的相似性水平。

17.根据权利要求1所述的方法，其中，基于所述一个或更多个查询约束来确定所述完成条件。

18.一个或更多个体现软件的计算机可读非暂时性存储介质，所述软件在被执行时能够操作来：

接收搜索查询，其中，所述搜索查询包括查询内容对象和一个或更多个查询约束；

生成表示所述查询内容对象的特征向量；

访问稀疏图，所述稀疏图包括对应于多个候选内容对象的多个节点和连接所述节点的多个链接，其中，连接到任何特定节点的链接的数量小于链接的阈值数量，并且其中，每个候选内容对象由紧致码表示；

从所述稀疏图的多个节点中选择入口节点；

从所述多个候选内容对象中选择关于所述查询内容对象的一个或更多个相似内容对象，其中，所述一个或更多个相似内容对象通过以下方式被迭代地选择：

识别所述入口节点的一个或更多个链接节点，其中，每个链接节点通过链接连接到所述入口节点；

对于对应于所述链接节点的每个候选内容对象，解压缩表示所述候选内容对象的紧致码，以生成表示所述候选内容对象的特征向量；

基于表示所述查询内容对象的特征向量和表示所述候选内容对象的特征向量之间的比较，从所述候选内容对象中选择零个或更多个相似内容对象；

确定是否满足选择所述一个或更多个相似内容对象的完成条件，其中，基于所述一个或更多个查询约束来确定所述完成条件；和

基于所述确定：

如果满足所述完成条件，则返回一个或更多个所选择的相似内容对象；否则

如果不满足所述完成条件，则通过使用来自所述一个或更多个链接节点的对应于最相似内容对象的链接节点作为所述入口节点来重复迭代选择；和

响应于接收到所述搜索查询，向与用户相关联的客户端系统发送用于呈现一个或更多个所选择的相似内容对象的指令。

19.一种系统，包括：一个或更多个处理器；以及耦合到所述处理器的非暂时性存储器，所述非暂时性存储器包括能够由所述处理器执行的指令，在执行所述指令时，所述处理器能够操作来：

接收搜索查询，其中，所述搜索查询包括查询内容对象和一个或更多个查询约束；

生成表示所述查询内容对象的特征向量；

访问稀疏图，所述稀疏图包括对应于多个候选内容对象的多个节点和连接所述节点的多个链接，其中，连接到任何特定节点的链接的数量小于链接的阈值数量，并且其中，每个候选内容对象由紧致码表示；

从所述稀疏图的多个节点中选择入口节点；

从所述多个候选内容对象中选择关于所述查询内容对象的一个或更多个相似内容对象，其中，所述一个或更多个相似内容对象通过以下方式被迭代地选择：

识别所述入口节点的一个或更多个链接节点，其中，每个链接节点通过链接连接到所述入口节点；

对于对应于所述链接节点的每个候选内容对象，解压缩表示所述候选内容对象的紧致码，以生成表示所述候选内容对象的特征向量；

基于表示所述查询内容对象的特征向量和表示所述候选内容对象的特征向量之间的比较，从所述候选内容对象中选择零个或更多个相似内容对象；

确定是否达到了选择所述一个或更多个相似内容对象的完成条件，其中，基于所述一个或更多个查询约束来确定所述完成条件；和

基于所述确定：

如果满足所述完成条件，则返回一个或更多个所选择的相似内容对象；否则

如果不满足所述完成条件，则通过使用来自所述一个或更多个链接节点的对应于最相似内容对象的链接节点作为所述入口节点来重复迭代选择；和

响应于接收到所述搜索查询，向与用户相关联的客户端系统发送用于呈现一个或更多个所选择的相似内容对象的指令。

20.一种方法，包括由一个或更多个计算系统：

接收搜索查询，其中，所述搜索查询包括查询内容对象和一个或更多个查询约束；

生成表示所述查询内容对象的特征向量；

访问稀疏图，所述稀疏图包括对应于多个候选内容对象的多个节点和连接所述节点的多个链接，其中，连接到任何特定节点的链接的数量小于链接的阈值数量，并且其中，每个候选内容对象由紧致码表示；

从所述稀疏图的多个节点中选择入口节点；

从所述多个候选内容对象中选择关于所述查询内容对象的一个或更多个相似内容对象，其中，所述一个或更多个相似内容对象通过以下方式被迭代地选择：

识别所述入口节点的一个或更多个链接节点，其中，每个链接节点通过链接连接到所述入口节点；

对于对应于所述链接节点的每个候选内容对象，解压缩表示所述候选内容对象的紧致码，以生成表示所述候选内容对象的特征向量；

基于表示所述查询内容对象的特征向量和表示所述候选内容对象的特征向量之间的比较，从所述候选内容对象中选择零个或更多个相似内容对象；

确定是否达到了选择所述一个或更多个相似内容对象的完成条件，其中，基于所述一个或更多个查询约束来确定所述完成条件；和

基于所述确定：

如果满足所述完成条件，则返回一个或更多个所选择的相似内容对象；否则

如果不满足所述完成条件，则通过使用来自所述一个或更多个链接节点的对应于最相似内容对象的链接节点作为所述入口节点来重复迭代选择；和

响应于接收到所述搜索查询，向与用户相关联的客户端系统发送用于呈现一个或更多个所选择的相似内容对象的指令。

21.根据权利要求20所述的方法，其中，所述一个或更多个查询约束包括以下项中的一个或更多个：

存储器利用率或CPU利用率约束；

选择精度约束；或者

搜索查询处理时间约束。

22.根据权利要求20或21所述的方法，其中，所述稀疏图的多个链接由一个或更多个机器学习算法生成。

23.根据权利要求20至22中任一项所述的方法，其中，基于一种或更多种编码来生成每个候选内容对象的紧致码；

可选地，其中所述一种或更多种编码包括乘积量化器或优化乘积量化器中的一个或更多个。

24.根据权利要求20至23中任一项所述的方法，其中，所述稀疏图中的每个节点与候选内容对象的聚类相关联，其中，候选对象的每个聚类与特征向量相关联。

25.根据权利要求24所述的方法，从所述稀疏图的多个节点中选择所述入口节点包括：

基于表示所述查询内容对象的特征向量和表示所述候选内容对象的聚类的特征向量之间的比较来选择所述入口节点。

26.根据权利要求20至23中任一项所述的方法，其中，解压缩表示每个候选内容对象的紧致码是基于以下项中的一个或更多个：

细化，所述细化基于表示所述候选内容对象的紧致码和与对应于所述候选内容对象的节点的一个或更多个链接节点相关联的一个或更多个紧致码；或者

表示所述候选内容对象的紧致码和包括多个回归权重向量的回归码本。

27.根据权利要求26所述的方法，其中，所述细化基于一个或更多个损失函数，并且其中，所述一个或更多个损失函数将表示所述候选内容对象的特征向量和与所述候选内容对象相关联的代码矩阵相关联，其中，所述代码矩阵包括表示所述候选内容对象的紧致码和与对应于所述候选内容对象的节点的链接节点相关联的紧致码。

28.根据权利要求26或27所述的方法，其中，基于所述多个候选内容对象和一个或更多个损失函数来学习所述回归码本；

可选地，其中所述回归码本包括每个候选内容对象的一组量化回归系数，并且其中，生成表示每个候选内容对象的特征向量还基于所述一组量化回归系数。

29.根据权利要求20至28中任一项所述的方法，其中，所述查询内容对象包括文本、音频剪辑、图像或视频剪辑中的一个或更多个。

30.根据权利要求20至29中任一项所述的方法，其中，表示所述查询内容对象的特征向量是基于一个或更多个机器学习算法生成的，并且其中，所述一个或更多个机器学习算法基于卷积神经网络。

31.根据权利要求20至30中任一项所述的方法，其中，所述入口节点的一个或更多个链接节点包括一个或更多个短程链接节点和至少一个长程链接节点。

32.根据权利要求31所述的方法，其中，所述一个或更多个短程链接节点中的每一个相对于所述入口节点具有大于或等于阈值水平的相似性水平；和/或

其中，所述至少一个长程链接节点相对于所述入口节点具有小于阈值水平的相似性水平。

33.根据权利要求20至32中任一项所述的方法，其中，基于所述一个或更多个查询约束来确定所述完成条件。

34.一个或更多个体现软件的计算机可读非暂时性存储介质，所述软件在被执行时能够操作来执行根据权利要求19至33中任一项的方法或者来：

接收搜索查询，其中，所述搜索查询包括查询内容对象和一个或更多个查询约束；

生成表示所述查询内容对象的特征向量；

访问稀疏图，所述稀疏图包括对应于多个候选内容对象的多个节点和连接所述节点的多个链接，其中，连接到任何特定节点的链接的数量小于链接的阈值数量，并且其中，每个候选内容对象由紧致码表示；

从所述稀疏图的多个节点中选择入口节点；

从所述多个候选内容对象中选择关于所述查询内容对象的一个或更多个相似内容对象，其中，所述一个或更多个相似内容对象通过以下方式被迭代地选择：

识别所述入口节点的一个或更多个链接节点，其中，每个链接节点通过链接连接到所述入口节点；

对于对应于所述链接节点的每个候选内容对象，解压缩表示所述候选内容对象的紧致码，以生成表示所述候选内容对象的特征向量；

基于表示所述查询内容对象的特征向量和表示所述候选内容对象的特征向量之间的

比较,从所述候选内容对象中选择零个或更多个相似内容对象;

确定是否满足选择所述一个或更多个相似内容对象的完成条件,其中,基于所述一个或更多个查询约束来确定所述完成条件;和

基于所述确定:

如果满足所述完成条件,则返回一个或更多个所选择的相似内容对象;否则

如果不满足所述完成条件,则通过使用来自所述一个或更多个链接节点的对应于最相似内容对象的链接节点作为所述入口节点来重复迭代选择;和

响应于接收到所述搜索查询,向与用户相关联的客户端系统发送用于呈现一个或更多个所选择的相似内容对象的指令。

35.一种系统,包括:一个或更多个处理器;以及耦合到所述处理器的非暂时性存储器,所述非暂时性存储器包括能够由所述处理器执行的指令,在执行所述指令时,所述处理器能够操作来执行根据权利要求19至33中任一项的方法或者来:

接收搜索查询,其中,所述搜索查询包括查询内容对象和一个或更多个查询约束;

生成表示所述查询内容对象的特征向量;

访问稀疏图,所述稀疏图包括对应于多个候选内容对象的多个节点和连接所述节点的多个链接,其中,连接到任何特定节点的链接的数量小于链接的阈值数量,并且其中,每个候选内容对象由紧致码表示;

从所述稀疏图的多个节点中选择入口节点;

从所述多个候选内容对象中选择关于所述查询内容对象的一个或更多个相似内容对象,其中,所述一个或更多个相似内容对象通过以下方式被迭代地选择:

识别所述入口节点的一个或更多个链接节点,其中,每个链接节点通过链接连接到所述入口节点;

对于对应于所述链接节点的每个候选内容对象,解压缩表示所述候选内容对象的紧致码,以生成表示所述候选内容对象的特征向量;

基于表示所述查询内容对象的特征向量和表示所述候选内容对象的特征向量之间的比较,从所述候选内容对象中选择零个或更多个相似内容对象;

确定是否达到了选择所述一个或更多个相似内容对象的完成条件,其中,基于所述一个或更多个查询约束来确定所述完成条件;和

基于所述确定:

如果满足所述完成条件,则返回一个或更多个所选择的相似内容对象;否则

如果不满足所述完成条件,则通过使用来自所述一个或更多个链接节点的对应于最相似内容对象的链接节点作为所述入口节点来重复迭代选择;和

响应于接收到所述搜索查询,向与用户相关联的客户端系统发送用于呈现一个或更多个所选择的相似内容对象的指令。

在线社交网络上利用图和紧致回归码的快速索引

技术领域

[0001] 本公开总体上涉及网络环境中的数据库和文件管理,并且尤其涉及在社交网络环境中执行对象搜索。

[0002] 背景

[0003] 可以包括社交网络网站的社交网络系统可以使它的用户(例如个人或组织)能够与它互动以及通过它彼此互动。社交网络系统可以利用来自用户的输入来创建并在社交网络系统中存储与用户相关联的用户简档(user profile)。用户简档可以包括人口统计信息、通信渠道信息以及关于用户的个人兴趣的信息。社交网络系统还可以用来自用户的输入来创建并存储用户与社交网络系统的其他用户的关系的记录,以及提供服务(例如,发帖墙(wall post)、照片分享、事件组织、消息传送、游戏或广告)以便于在用户之间或当中的社交互动。

[0004] 社交网络系统可以通过一个或更多个网络向用户的移动设备或其他计算设备发送与它的服务相关的内容或消息。用户还可以在用户的移动设备或其他计算设备上安装软件应用,用于访问用户的用户简档和在社交网络系统内的其他数据。社交网络系统可以生成一组个性化的内容对象以显示给用户,例如关连(connect)到该用户的其他用户的汇总动态(story)的动态消息(newsfeed)。

[0005] 社交图分析从由节点和边组成的网络理论方面来看待社交关系。节点表示网络中的单独参与者,以及边表示在参与者之间的关系。由此产生的基于图的结构常常是非常复杂的。可以有很多类型的节点和用于连接节点的很多类型的边。在它的最简单的形式中,社交图是在所有被研究的节点之间的所有相关边的布局图。

[0006] 特定实施例的概述

[0007] 在特定实施例中,社交网络系统可以响应于用户查询来执行内容对象的相似性搜索。基于图游走(graph walk)的相似性搜索方法最近在速度和精度之间取得了突出的折衷,抛开了存储器需求。在本文公开的实施例中,通过另外考虑在单个服务器上索引数十亿个内容对象(例如,图像)所需的存储器约束,重新考虑这些方法。在特定实施例中,社交网络系统可以利用一种方法,即链接和代码(L&C),该方法基于相似性搜索的紧致(compact)表示和图遍历(graph traversal)。L&C方法可以使用量化来对索引向量进行编码,并利用图结构来细化相似性估计。本质上,L&C方法可以充分利用这两个世界:搜索策略是基于嵌套图,从而利用相对较小的一组比较提供了高精度。同时,它可以提供显著的存储器压缩。因此,考虑到每个向量64-128字节,L&C方法在操作点上可以优于现有技术,20亿规模的公共基准(public benchmark)上的实验结果证明了这一点。尽管本公开描述了以特定方式用于特定相似性搜索的特定方法,但是本公开设想了以任何适当方式用于任何适当相似性搜索的任何适当方法。

[0008] 在特定实施例中,社交网络系统可以接收搜索查询。搜索查询可以包括查询内容对象和一个或更多个查询约束。社交网络系统然后可以生成表示查询内容对象的特征向量。在特定实施例中,社交网络系统可以访问稀疏图,该稀疏图包括对应于多个候选内容对

象的多个节点和连接这些节点的多个链接。连接到任何特定节点的链接的数量可以少于链接的阈值数量，并且每个候选内容对象可以由紧致码(compact code)表示。在特定实施例中，社交网络系统可以从稀疏图的多个节点中选择入口节点(entry node)。社交网络系统然后可以从多个候选内容对象中选择关于查询内容对象的一个或更多个相似内容对象。在特定实施例中，一个或更多个相似内容对象可以通过以下过程被迭代地选择。社交网络系统可以首先识别入口节点的一个或更多个链接节点。每个链接节点可以通过链接连接到入口节点。社交网络系统然后可以针对对应于链接节点的每个候选内容对象，解压缩表示候选内容对象的紧致码，以生成表示候选内容对象的特征向量。社交网络系统然后可以基于表示查询内容对象的特征向量和表示候选内容对象的特征向量之间的比较，从候选内容对象中选择零个或更多个相似内容对象。社交网络系统然后可以确定对于一个或更多个相似内容对象的选择是否满足完成条件。可以基于一个或更多个查询约束来确定完成条件。基于该确定，如果满足完成条件，则社交网络系统然后可以返回一个或更多个所选择的相似内容对象；如果不满足完成条件，则社交网络系统可以通过使用来自一个或更多个链接节点的对应于最相似内容对象的链接节点作为入口节点来重复该迭代选择。在特定实施例中，社交网络系统还可以响应于接收到搜索查询，向与用户相关联的客户端系统发送用于呈现一个或更多个所选择的相似内容对象的指令。

[0009] 实现基于图和代码的相似性搜索的目标存在某些技术挑战。一个技术挑战可以包括构建有助于高效搜索的稀疏图。由本文公开的实施例提出的解决上述挑战的解决方案是基于一个或更多个机器学习算法生成链接，因为连接到任何特定节点的链接的数量少于阈值数量，从而导致具有明显少于传统图的链接的稀疏图。另一个技术挑战可以包括重建表示候选内容对象的可靠特征向量。由本文公开的实施例提出的解决该挑战的解决方案是细化(refinement)和回归码本(regression codebook)，其可以利用图连通性(connectivity)并最小化重建误差，从而使得重建的特征向量更准确地表示候选内容对象。

[0010] 本文公开的某些实施例可以提供一个或更多个技术优势。实施例的技术优势可以包括在执行相似性搜索时由于基于候选内容对象的紧致表示和图遍历策略的利用的有限的存储器占用(memory footprint)，改善存储器和精度之间的折衷。实施例的另一个技术优势可以包括在相似性搜索期间对于图的聚类(cluster)中的入口节点提高所选择的候选内容对象的多样性，连接的节点包括短程链接节点(更类似于入口节点)和长程链接节点(不太类似于入口节点)。本文公开的某些实施例可以不提供、提供一些或全部上述技术优势。鉴于本公开的附图、说明书和权利要求，一个或更多个其他技术优势对于本领域技术人员来说可能是十分明显的。

[0011] 本文公开的实施例仅仅是示例，并且本公开的范围不限于它们。特定实施例可以包括本文公开的实施例的组件、元件、特征、功能、操作或步骤中的所有、一些或不包括它们。根据本发明的实施例在涉及方法、存储介质和系统的所附权利要求中被具体公开，其中在一个权利要求类别(例如方法)中提到的任何特征可以在另一个权利要求类别(例如存储介质、系统、计算机程序产品)中被要求保护。在所附权利要求中的从属性或往回引用仅为了形式原因而被选择。然而，也可以要求保护由对任何前面的权利要求的有意往回引用(特别是多项引用)而产生的任何主题，使得权利要求及其特征的任何组合被公开并且可被要

求保护，而不考虑在所附权利要求中选择的从属性。可以被要求保护的主题不仅包括如在所附权利要求中阐述的特征的组合，而且还包括在权利要求中的特征的任何其他组合，其中，在权利要求中提到的每个特征可以与在权利要求中的任何其他特征或其他特征的组合相结合。此外，本文描述或描绘的实施例和特征中的任一个可以在单独的权利要求中和/或以与本文描述或描绘的任何实施例或特征的任何组合或以与所附权利要求的任何特征的任何组合被要求保护。

[0012] 在一个实施例中，一种方法可以包括，由一个或更多个计算系统：

[0013] 接收搜索查询，其中该搜索查询包括查询内容对象和一个或更多个查询约束；

[0014] 生成表示查询内容对象的特征向量；

[0015] 访问稀疏图，该稀疏图包括对应于多个候选内容对象的多个节点和连接这些节点的多个链接，其中连接到任何特定节点的链接的数量小于链接的阈值数量，并且其中每个候选内容对象由紧致码表示；

[0016] 从稀疏图的多个节点中选择入口节点；

[0017] 从多个候选内容对象中选择关于查询内容对象的一个或更多个相似内容对象，其中一个或更多个相似内容对象通过以下方式被迭代地选择：

[0018] 识别入口节点的一个或更多个链接节点，其中每个链接节点通过链接连接到入口节点；

[0019] 对于对应于链接节点的每个候选内容对象，解压缩表示候选内容对象的紧致码，以生成表示候选内容对象的特征向量；

[0020] 基于表示查询内容对象的特征向量和表示候选内容对象的特征向量之间的比较，从候选内容对象中选择零个或更多个相似内容对象；

[0021] 确定是否达到选择一个或更多个相似内容对象的完成条件，其中基于一个或更多个查询约束来确定完成条件；和

[0022] 基于该确定：

[0023] 如果满足完成条件，则返回一个或更多个所选择的相似内容对象；否则

[0024] 如果不满足完成条件，则通过使用来自一个或更多个链接节点的对应于最相似内容对象的链接节点作为入口节点来重复迭代选择；和

[0025] 响应于接收到搜索查询，向与用户相关联的客户端系统发送用于呈现一个或更多个所选择的相似内容对象的指令。

[0026] 一个或更多个查询约束可以包括以下项中的一个或更多个：

[0027] 存储器利用率或CPU利用率约束；

[0028] 选择精度约束；或者

[0029] 搜索查询处理时间约束。

[0030] 稀疏图的多个链接可以由一个或更多个机器学习算法生成。

[0031] 每个候选内容对象的紧致码可以基于一种或更多种编码来生成。

[0032] 一种或更多种编码可以包括乘积量化器(product quantizer)或优化乘积量化器中的一个或更多个。

[0033] 稀疏图中的每个节点可以与候选内容对象的聚类相关联，其中候选对象的每个聚类可以与特征向量相关联。

[0034] 从稀疏图的多个节点中选择入口节点可以包括：

[0035] 基于表示查询内容对象的特征向量和表示候选内容对象的聚类的特征向量之间的比较来选择入口节点。

[0036] 解压缩表示每个候选内容对象的紧致码可以基于以下项中的一个或更多个：

[0037] 细化，该细化基于表示候选内容对象的紧致码和与对应于候选内容对象的节点的一个或更多个链接节点相关联的一个或更多个紧致码；或者

[0038] 表示候选内容对象的紧致码和包括多个回归权重向量的回归码本。

[0039] 细化可以基于一个或更多个损失函数，并且一个或更多个损失函数可以将表示候选内容对象的特征向量和与候选内容对象相关联的代码矩阵相关联，其中代码矩阵可以包括表示候选内容对象的紧致码和与对应于候选内容对象的节点的链接节点相关联的紧致码。

[0040] 可以基于多个候选内容对象和一个或更多个损失函数来学习回归码本。

[0041] 回归码本可以包括每个候选内容对象的一组量化回归系数，并且生成表示每个候选内容对象的特征向量可以基于该组量化回归系数。

[0042] 查询内容对象可以包括文本、音频剪辑、图像或视频剪辑中的一个或更多个。

[0043] 表示查询内容对象的特征向量可以基于一个或更多个机器学习算法生成，并且一个或更多个机器学习算法可以基于卷积神经网络。

[0044] 入口节点的一个或更多个链接节点可以包括一个或更多个短程链接节点和至少一个长程链接节点。

[0045] 一个或更多个短程链接节点中的每一个相对于入口节点可以具有大于或等于阈值水平的相似性水平。

[0046] 至少一个长程链接节点相对于入口节点可以具有小于阈值水平的相似性水平。

[0047] 可以基于一个或更多个查询约束来确定完成条件。

[0048] 在一个实施例中，一个或更多个计算机可读非暂时性存储介质可以体现软件，该软件在被执行时可操作来：

[0049] 接收搜索查询，其中该搜索查询包括查询内容对象和一个或更多个查询约束；

[0050] 生成表示查询内容对象的特征向量；

[0051] 访问稀疏图，该稀疏图包括对应于多个候选内容对象的多个节点和连接这些节点的多个链接，其中连接到任何特定节点的链接的数量小于链接的阈值数量，并且其中每个候选内容对象由紧致码表示；

[0052] 从稀疏图的多个节点中选择入口节点；

[0053] 从多个候选内容对象中选择关于查询内容对象的一个或更多个相似内容对象，其中一个或更多个相似内容对象通过以下方式被迭代地选择：

[0054] 识别入口节点的一个或更多个链接节点，其中每个链接节点通过链接连接到入口节点；

[0055] 对于对应于链接节点的每个候选内容对象，解压缩表示候选内容对象的紧致码，以生成表示候选内容对象的特征向量；

[0056] 基于表示查询内容对象的特征向量和表示候选内容对象的特征向量之间的比较，从候选内容对象中选择零个或更多个相似内容对象；

[0057] 确定是否满足选择一个或更多个相似内容对象的完成条件，其中基于一个或更多个查询约束来确定完成条件；和

[0058] 基于该确定：

[0059] 如果满足完成条件，则返回一个或更多个所选择的相似内容对象；否则

[0060] 如果不满足完成条件，则通过使用来自一个或更多个链接节点的对应于最相似内容对象的链接节点作为入口节点来重复迭代选择；和

[0061] 响应于接收到搜索查询，向与用户相关联的客户端系统发送用于呈现一个或更多个所选择的相似内容对象的指令。

[0062] 在一个实施例中，一种系统可以包括：一个或更多个处理器；以及耦合到处理器的非暂时性存储器，该非暂时性存储器包括可由处理器执行的指令，当执行指令时，处理器可操作来：

[0063] 接收搜索查询，其中该搜索查询包括查询内容对象和一个或更多个查询约束；

[0064] 生成表示查询内容对象的特征向量；

[0065] 访问稀疏图，该稀疏图包括对应于多个候选内容对象的多个节点和连接这些节点的多个链接，其中连接到任何特定节点的链接的数量小于链接的阈值数量，并且其中每个候选内容对象由紧致码表示；

[0066] 从稀疏图的多个节点中选择入口节点；

[0067] 从多个候选内容对象中选择关于查询内容对象的一个或更多个相似内容对象，其中一个或更多个相似内容对象通过以下方式被迭代地选择：

[0068] 识别入口节点的一个或更多个链接节点，其中每个链接节点通过链接连接到入口节点；

[0069] 对于对应于链接节点的每个候选内容对象，解压缩表示候选内容对象的紧致码，以生成表示候选内容对象的特征向量；

[0070] 基于表示查询内容对象的特征向量和表示候选内容对象的特征向量之间的比较，从候选内容对象中选择零个或更多个相似内容对象；

[0071] 确定是否达到选择一个或更多个相似内容对象的完成条件，其中，基于一个或更多个查询约束来确定完成条件；和

[0072] 基于该确定：

[0073] 如果满足完成条件，则返回一个或更多个所选择的相似内容对象；否则

[0074] 如果不满足完成条件，则通过使用来自一个或更多个链接节点的对应于最相似内容对象的链接节点作为入口节点来重复迭代选择；和

[0075] 响应于接收到搜索查询，向与用户相关联的客户端系统发送用于呈现一个或更多个所选择的相似内容对象的指令。

[0076] 在根据本发明的实施例中，一个或更多个计算机可读非暂时性存储介质可以体现软件，该软件在被执行时可操作来执行根据本发明或任何上面提到的实施例的方法。

[0077] 在根据本发明的实施例中，一种系统可以包括：一个或更多个处理器；以及耦合到处理器并包括由处理器可执行的指令的至少一个存储器，处理器当执行指令时可操作来执行根据本发明或任何上面提到的实施例的方法。

[0078] 在根据本发明的实施例中，一种计算机程序产品优选地包括计算机可读非暂时性

存储介质，当计算机程序产品在数据处理系统上被执行时可以可操作来执行根据本发明或任何上面提到的实施例的方法。

[0079] 附图简述

[0080] 图1示出了与社交网络系统相关联的示例网络环境。

[0081] 图2示出了L&C方法的示例可视化。

[0082] 图3示出了关于选择性的示例精度变化。

[0083] 图4示出了精度和速度之间的示例折衷。

[0084] 图5示出了不同估计器的误差的示例分布。

[0085] 图6示出了取决于64字节的固定存储器预算是否被分配给代码(不同大小的QPQ代码)或链接的示例性能变化。

[0086] 图7A示出了操作点在搜索时间对精度方面的示例比较。

[0087] 图7B示出了操作点在搜索时间对精度方面的另一示例比较。

[0088] 图8示出了基于图和紧致码的相似性搜索的示例方法。

[0089] 图9示出了示例社交图。

[0090] 图10示出了用于存储社交网络系统的对象的示例划分。

[0091] 图11示出了嵌入空间的示例视图。

[0092] 图12示出了示例人工神经网络。

[0093] 图13示出了示例计算机系统。

[0094] 示例实施例的描述

[0095] 系统综述

[0096] 图1示出了与社交网络系统相关联的示例网络环境100。网络环境100包括通过网络110连接到彼此的客户端系统130、社交网络系统160和第三方系统170。尽管图1示出了客户端系统130、社交网络系统160、第三方系统170和网络110的特定布置，但是本公开设想了客户端系统130、社交网络系统160、第三方系统170和网络110的任何合适的布置。作为示例而不是作为限制，客户端系统130、社交网络系统160和第三方系统170中的两个或更多个可以直接连接到彼此，绕过网络110。作为另一示例，客户端系统130、社交网络系统160和第三方系统170中的两个或更多个可以全部或部分地在物理上或逻辑上彼此位于同一位置。此外，尽管图1示出了特定数量的客户端系统130、社交网络系统160、第三方系统170和网络110，但是本公开设想了任何合适的数量的客户端系统130、社交网络系统160、第三方系统170和网络110。作为示例而不是作为限制，网络环境100可以包括多个客户端系统130、社交网络系统160、第三方系统170和网络110。

[0097] 本公开设想了任何合适的网络110。作为示例而不是作为限制，网络110的一个或更多个部分可以包括自组织网络、内联网、外联网、虚拟专用网络(VPN)、局域网(LAN)、无线LAN(WLAN)、广域网(WAN)、无线WAN(WWAN)、城域网(MAN)、互联网的一部分、公共交换电话网(PSTN)的一部分、蜂窝电话网、或这些中的两个或更多个的组合。网络110可以包括一个或更多个网络110。

[0098] 链路150可以将客户端系统130、社交网络系统160和第三方系统170连接到通信网络110或连接到彼此。本公开设想了任何合适的链路150。在特定实施例中，一个或更多个链路150包括一个或更多个有线(例如数字用户线路(DSL)或基于电缆的数据服务接口规范

(DOCSIS))链路、无线(例如Wi-Fi或全球互通微波接入(WiMAX))链路、或光(例如同步光网络(SONET)或同步数字体系(SDH))链路。在特定实施例中,一个或更多个链路150各自包括自组织网络、内联网、外联网、VPN、LAN、WLAN、WAN、WWAN、MAN、互联网的一部分、PSTN的一部分、基于蜂窝技术的网络、基于卫星通信技术的网络、另一链路150、或两个或更多个这种链路150的组合。链路150不需要在整个网络环境100中是相同的。一个或更多个第一链路150可以在一个或更多个方面上不同于一个或更多个第二链路150。

[0099] 在特定实施例中,客户端系统130可以是包括硬件、软件或嵌入式逻辑组件、或两个或更多个这样的组件的组合,并且能够执行由客户端系统130实现或支持的适当功能的电子设备。作为示例而不是作为限制,客户端系统130可以包括计算机系统,例如台式计算机、笔记本或膝上型计算机、上网本、平板计算机、电子书阅读器、GPS设备、相机、个人数字助理(PDA)、手持电子设备、蜂窝电话、智能手机、其他合适的电子设备、或它们的任何合适的组合。本公开设想了任何合适的客户端系统130。客户端系统130可以使在客户端系统130处的网络用户能够访问网络110。客户端系统130可以使它的用户能够与在其他客户端系统130处的其他用户进行通信。

[0100] 在特定实施例中,客户端系统130可以包括web浏览器132,例如MICROSOFT INTERNET EXPLORER、GOOGLE CHROME或MOZILLA FIREFOX,并且可以具有一个或更多个附加件、插件或其他扩展件,例如TOOLBAR或YAHOO TOOLBAR。在客户端系统130处的用户可以输入统一资源定位符(URL)或将web浏览器132引导到特定的服务器(例如服务器162或与第三方系统170相关联的服务器)的其他地址,并且web浏览器132可以生成超文本传输协议(HTTP)请求并将HTTP请求传递到服务器。服务器可以接受HTTP请求,并响应于HTTP请求而向客户端系统130传递一个或更多个超文本标记语言(HTML)文件。客户端系统130可以基于来自服务器的HTML文件来再现web界面(例如网页)用于呈现给用户。本公开设想了任何合适的源文件。作为示例而不是作为限制,可以根据特定的需要从HTML文件、可扩展超文本标记语言(XHTML)文件或可扩展标记语言(XML)文件来再现web界面。这样的界面还可以执行脚本,例如且没有限制地,用JAVASCRIPT、JAVA、MICROSOFT SILVERLIGHT编写的脚本、标记语言和脚本(例如AJAX(异步JAVASCRIPT和XML))的组合等。在本文,在适当的情况下,对web界面的引用包括一个或更多个相应的源文件(浏览器可以使用这些源文件来再现web界面),反之亦然。

[0101] 在特定实施例中,社交网络系统160可以是可以托管在线社交网络的网络可寻址计算系统。社交网络系统160可以生成、存储、接收并发送社交网络数据(例如,用户简档数据、概念简档数据、社交图信息或与在线社交网络相关的其他合适的数据)。社交网络系统160可以由网络环境100的其他组件直接地或经由网络110来访问。作为示例而不是作为限制,客户端系统130可以使用web浏览器132或与社交网络系统160相关联的原生(native)应用(例如,移动社交网络应用、消息传送应用、另一合适的应用或其任何组合)直接地或经由网络110来访问社交网络系统160。在特定实施例中,社交网络系统160可以包括一个或更多个服务器162。每个服务器162可以是单一服务器(unitary server)或跨越多台计算机或多个数据中心的分布式服务器。服务器162可以具有各种类型,例如且没有限制,web服务器、新闻服务器、邮件服务器、消息服务器、广告服务器、文件服务器、应用服务器、交换服务器、数据库服务器、代理服务器、适合于执行本文描述的功能或过程的另一服务器、或者其任何

组合。在特定实施例中，每个服务器162可以包括硬件、软件或嵌入式逻辑组件、或用于执行由服务器162实现或支持的适当功能的两个或更多个这样的组件的组合。在特定实施例中，社交网络系统160可以包括一个或更多个数据储存器164。数据储存器164可以用于存储各种类型的信息。在特定实施例中，可以根据特定的数据结构来组织存储在数据储存器164中的信息。在特定实施例中，每个数据储存器164可以是关系数据库、纵列(columnar)数据库、相关性数据库或其他合适的数据库。尽管本公开描述或示出了特定类型的数据库，但是本公开设想了任何合适类型的数据库。特定实施例可以提供使客户端系统130、社交网络系统160或第三方系统170能够管理、检索、修改、添加或删除存储在数据储存器164中的信息的接口。

[0102] 在特定实施例中，社交网络系统160可以在一个或更多个数据储存器164中存储一个或更多个社交图。在特定实施例中，社交图可以包括多个节点——其可以包括多个用户节点(各自对应于特定用户)或多个概念节点(各自对应于特定概念)——以及连接节点的多条边。社交网络系统160可以向在线社交网络的用户提供与其他用户通信和互动的能力。在特定实施例中，用户可以经由社交网络系统160加入在线社交网络，然后向他们想要关连的社交网络系统160的多个其他用户添加关连(例如，关系)。在本文，术语“朋友”可以指社交网络系统160的任何其他用户，用户经由社交网络系统160与任何其他用户形成关连(connection)、关联(association)或关系。

[0103] 在特定实施例中，社交网络系统160可以向用户提供对由社交网络系统160支持的各种类型的项目或对象采取动作的能力。作为示例而不是作为限制，项目和对象可以包括社交网络系统160的用户可以属于的群组或社交网络、用户可能感兴趣的事件或日历条目、用户可以使用的基于计算机的应用、允许用户经由服务来购买或销售商品的交易、用户可以执行的与广告的互动、或其他合适的项目或对象。用户可以与能够在社交网络系统160中或者由第三方系统170的外部系统表示的任何事物进行互动，第三方系统170与社交网络系统160分离并且经由网络110耦合到社交网络系统160。

[0104] 在特定实施例中，社交网络系统160能够链接各种实体。作为示例而不是作为限制，社交网络系统160可以使用户能够彼此互动以及从第三方系统170或其他实体接收内容，或者允许用户通过应用编程接口(API)或其他通信渠道与这些实体互动。

[0105] 在特定实施例中，第三方系统170可以包括一种或更多种类型的服务器、一个或更多个数据储存器、一个或更多个接口(包括但不限于API)、一个或更多个web服务、一个或更多个内容源、一个或更多个网络或任何其他合适的部件(例如，服务器可以与这些部件通信)。第三方系统170可以由与操作社交网络系统160的实体不同的实体进行操作。然而，在特定实施例中，社交网络系统160和第三方系统170可以结合彼此来操作以向社交网络系统160或第三方系统170的用户提供社交网络服务。在这个意义上，社交网络系统160可以提供平台或骨干网，其他系统(例如第三方系统170)可以使用该平台或骨干网来在整个互联网上向用户提供社交网络服务和功能。

[0106] 在特定实施例中，第三方系统170可以包括第三方内容对象提供者。第三方内容对象提供者可以包括可以被传递到客户端系统130的内容对象的一个或更多个源。作为示例而不是作为限制，诸如，内容对象可以包括关于用户感兴趣的事情或活动的信息，例如电影放映时间、电影评论、餐馆评论、餐馆菜单、产品信息和评论或其他合适的信息。作为另一示

例而不是作为限制，内容对象可以包括激励内容对象（例如优惠券、折扣券、礼品券或其他合适的激励对象）。

[0107] 在特定实施例中，社交网络系统160还包括用户生成的内容对象，其可以增强用户与社交网络系统160的互动。用户生成的内容可以包括用户可以添加、上传、发送或“发布”到社交网络系统160的任何内容。作为示例而不是作为限制，用户将帖子从客户端系统130传送到社交网络系统160。帖子可以包括数据，例如状态更新或其他文本数据、位置信息、照片、视频、链接、音乐或其他类似数据或媒体。内容还可以由第三方通过“通信渠道”（例如动态消息或流）来添加到社交网络系统160。

[0108] 在特定实施例中，社交网络系统160可以包括各种服务器、子系统、程序、模块、日志和数据储存器。在特定实施例中，社交网络系统160可以包括下列项中的一个或更多个：web服务器、动作记录器、API请求服务器、相关性和排名引擎、内容对象分类器、通知控制器、动作日志、第三方内容对象暴露日志、推理模块、授权/隐私服务器、搜索模块、广告定位模块（advertisement-targeting module）、用户界面模块、用户简档储存器、关连储存器、第三方内容储存器或位置储存器。社交网络系统160还可以包括合适的组件，例如网络接口、安全机构、负载平衡器、故障转移服务器、管理和网络操作控制台、其他合适的组件、或其任何合适的组合。在特定实施例中，社交网络系统160可以包括用于存储用户简档的一个或更多个用户简档储存器。用户简档可以包括例如，传记信息、人口统计信息、行为信息、社交信息或其他类型的描述性信息（例如工作经历、教育历史、爱好或偏好、兴趣、喜好或位置）。兴趣信息可以包括与一个或更多个类别相关的兴趣。类别可以是一般的或特定的。作为示例而不是作为限制，如果用户“赞（like）”关于一种品牌的鞋的文章，则该类别可以是品牌，或者“鞋”或“衣服”的一般类别。关连储存器可以用于存储关于用户的关连信息。关连信息可以指示具有相似或共同的工作经历、组成员资格、爱好、教育历史或者以任何方式相关或共享共同属性的用户。关连信息还可以包括在不同用户和内容（内部和外部）之间的用户定义的关连。web服务器可以用于经由网络110将社交网络系统160链接到一个更或多个客户端系统130或一个或更多个第三方系统170。web服务器可以包括邮件服务器或用于在社交网络系统160和一个或更多个客户端系统130之间接收并按规定路线发送消息的其他消息传送功能。API请求服务器可以允许第三方系统170通过调用一个或更多个API来访问来自社交网络系统160的信息。动作记录器可以用于从web服务器接收关于用户在社交网络系统160上或之外的动作的通信。结合动作日志，可以维护用户暴露于第三方内容对象的第三方内容对象日志。通知控制器可以向客户端系统130提供关于内容对象的信息。信息可以作为通知被推送到客户端系统130，或者信息可以响应于从客户端系统130接收的请求而从客户端系统130中被拉取。授权服务器可以用于实施社交网络系统160的用户的一个或更多个隐私设置。用户的隐私设置确定与用户相关联的特定信息可以如何被分享。授权服务器可以例如通过设置适当的隐私设置来允许用户选择加入或选择退出使他们的动作由社交网络系统160记录或者与其他系统（例如，第三方系统170）分享。第三方内容对象储存器可以用于存储从第三方（例如第三方系统170）接收的内容对象。位置储存器可以用于存储从与用户相关联的客户端系统130接收的位置信息。广告定价模块可以组合社交信息、当前时间、位置信息或其他合适的信息以用通知的形式向用户提供相关广告。

[0109] 在线社交网络上利用图和紧致回归码的快速索引

[0110] 在特定实施例中，社交网络系统160可以响应于用户查询来执行内容对象的相似性搜索。基于图游走的相似性搜索方法最近在速度和精度之间取得了突出的折衷，抛开了存储器需求。在本文公开的实施例中，通过另外考虑在单个服务器上索引数十亿个内容对象(例如，图像)所需的存储器约束，重新考虑这些方法。在特定实施例中，社交网络系统160可以利用一种方法，即链接和代码(L&C)，该方法基于用于相似性搜索的紧致表示和图遍历。L&C方法可以使用量化来对索引向量进行编码，并利用图结构来细化相似性估计。本质上，L&C方法可以充分利用这两个世界：搜索策略是基于嵌套图，从而提供了相对较小的一组比较的高精度。同时，它可以提供显著的存储器压缩。因此，考虑到每个向量64-128字节，L&C方法在操作点上可以优于现有技术，20亿规模的公共基准上的实验结果证明了这一点。尽管本公开描述了以特定方式用于特定相似性搜索的特定方法，但是本公开设想了以任何适当方式用于任何适当相似性搜索的任何适当方法。

[0111] 在特定实施例中，社交网络系统160可以接收搜索查询。搜索查询可以包括查询内容对象和一个或更多个查询约束。社交网络系统160然后可以生成表示查询内容对象的特征向量。在特定实施例中，社交网络系统160可以访问稀疏图，该稀疏图包括对应于多个候选内容对象的多个节点和连接这些节点的多个链接。连接到任何特定节点的链接的数量可以少于链接的阈值数量，并且每个候选内容对象可以由紧致码表示。在特定实施例中，社交网络系统160可以从稀疏图的多个节点中选择入口节点。社交网络系统160然后可以从多个候选内容对象中选择关于查询内容对象的一个或更多个相似内容对象。在特定实施例中，一个或更多个相似内容对象可以通过以下过程被迭代地选择。社交网络系统160可以首先识别入口节点的一个或更多个链接节点。每个链接节点可以通过链接连接到入口节点。社交网络系统160然后可以针对对应于链接节点的每个候选内容对象，解压缩表示候选内容对象的紧致码，以生成表示候选内容对象的特征向量。社交网络系统160然后可以基于表示查询内容对象的特征向量和表示候选内容对象的特征向量之间的比较，从候选内容对象中选择零个或更多个相似内容对象。社交网络系统160然后可以确定对于一个或更多个相似内容对象的选择是否满足完成条件。可以基于一个或更多个查询约束来确定完成条件。基于该确定，如果满足完成条件，则社交网络系统160然后可以返回一个或更多个所选择的相似内容对象；如果不满足完成条件，则社交网络系统160可以通过使用来自一个或更多个链接节点的对应于最相似内容对象的链接节点作为入口节点来重复迭代选择。在特定实施例中，社交网络系统160还可以响应于接收到搜索查询，向与用户相关联的客户端系统130发送用于呈现一个或更多个所选择的相似内容对象的指令。

[0112] 介绍

[0113] 相似性搜索可能是大规模搜索中的关键问题，诸如，例如，响应于接收到搜索查询进行计算机视觉领域中的搜索。它可能是大规模图像搜索[32、38]、池化(pooling)[41]和半监督低样本分类(semi-supervised low-shot classification)[16]的核心组件。另一个示例可以包括具有大量类别的分类[22]。在特定实施例中，搜索查询中的查询内容对象可以包括文本、音频剪辑、图像或视频剪辑中的一个或更多个。然后可以在大量候选内容对象中执行相似性搜索。在过去的几年里，最近的大部分工作集中在紧致码上，其或者是二进制的[10、18]或者是基于各种量化方法的[26、11、3、44]。当使用诸如SIFT[31]的局部描述符时，采用向量的紧致表示可能是重要的，因为每幅图像可以被提取成千上万个这样的向

量。在这种情况下,描述符集通常比压缩图像本身需要更多的存储器。具有使用每个描述符8-32字节的压缩索引表示是由可扩展性和实际考虑驱动的需求。

[0114] 然而,视觉描述的最近进展主要考虑了描述方案[37、27、19],其中每个图像由唯一的向量表示。在特定实施例中,可以基于一个或更多个机器学习算法来生成表示查询内容对象的特征向量。一个或更多个机器学习算法可以基于卷积神经网络。更具体地,向量通常从卷积神经网络[5、42]的激活层提取。图像检索领域的现有技术端到端地对表示进行学习[20、39],使得余弦相似性或欧几里德距离反映语义相似性。最终的图像描述符由不超过几百个成分组成。

[0115] 在特定实施例中,搜索查询可以与一个或更多个查询约束相关联。一个或更多个查询约束可以包括下列项中的一个或更多个:存储器利用率或CPU利用率约束、选择精度约束、搜索查询处理时间约束、任何合适的约束或其任意组合。在这种情况下,为了更好的精度和/或效率,可能值得研究最近邻搜索交换存储器的方法。在许多情况下,128字节的图像表示是可以接受的,因为它与存储在数据库中的并且与其相关联的元数据相当(如果不小于该元数据)。虽然一些传统工作认为饱和性能超过16字节[8],但在Deep10M和Deep1B数据集(即,两个公共数据集)上用16字节实现的最佳结果在排序1(rank 1)处不超过50%的召回率(recall)[7、15]。虽然返回到原始向量可以提高召回率,但可能需要访问速度较慢的存储设备,这可能对整体效率不利。

[0116] 相比之下,一些方法(如在FLANN[35]中实现的方法)考虑了小得多的数据集,并以高精度和吞吐量为目标。在NMSLIB[9]中实现的现有技术的方法只关注速度和精度之间的折衷。这些方法在它们的评估中不包含任何存储器约束,并且只在最多包含几百万个向量的小数据集上对方法进行比较。值得注意的是,Malkov等人的成功方法[33、34]需要原始向量和链接向量的全图结构两者。这种存储器需求严重限制了这类方法的可扩展性,这类方法可能从未扩展到十亿个向量。

[0117] 这两种观点(即压缩域搜索和基于图的探索)考虑了操作点谱中关于存储器要求的极端方面。虽然在可扩展性方面,存储器的紧致性可能具有明显的实际优势,但在精度和基本向量比较的数量之间的折衷方面,HNSW(分层可导航小世界[33、34])明显优于倒排多索引(IMI)[6],这要归功于快速收敛到给定查询的最近邻居的有效图游走。

[0118] 本文公开的实施例旨在通过提出一种解决方案来调和相似性搜索中的这两种趋势,该解决方案由于有限的存储器占用而可扩展到十亿个向量,并且给出了由图遍历策略提供的良好的精度/速度折衷。为此,L&C方法可以通过以下项来表示每个索引向量:i)基于优化的乘积量化(OPQ)[17]的紧致表示,以及ii)通过来自邻居的新的量化回归来细化它。该细化可以利用图的连通性,并且向量只需几个字节。L&C方法可以通过交替优化来学习回归码本,以最小化重建误差。结果,由于基于候选内容对象的紧致表示和图遍历策略的利用的有限存储器占用,本文公开的实施例在执行相似性搜索时可以具有改善存储器和精度之间的折衷的技术优势。

[0119] 图2示出了L&C方法的示例可视化。在特定实施例中,可以采用保持所有数据库点之间连通性的图遍历策略。L&C方法可以基于压缩描述符以节约存储器:每个数据库向量(由图2中的圆圈表示)可以用量化(由图2中的正方形表示)来近似。例如,向量201、203、205、207和209分别由量化202、204、206、208和210近似。在特定实施例中,可以通过从其编

码的邻居中回归每个数据库向量来进一步改进该估计，这可以提供极好的表示基础。回归系数$\beta=[\beta_0,\ldots,\beta_k]$可以从被学习的码本中选择，以最小化向量重建误差。

[0120] 本文公开的实施例的贡献可以包括评估不同假设的初步分析，以及采用图结构和紧致码的索引方法。具体来说，

[0121] • 本公开的实施例表明，在典型设置中，使用粗略质心(coarse centroid)可以提供比其最近邻居更好的描述符近似，这表明向量的一级近似应该是质心，而不是数据集的另一个点[8]。本文公开的实施例还表明向量可以通过其邻居的小集合的线性组合来被更好地近似，其中固定的混合权重通过封闭形式的方程获得。如果可以根据每个向量存储权重，则可以进一步改进该估计器。

[0122] • 本文公开的实施例表明，与基于倒排列表的竞争方法相比，HNSW可以提供更好的选择性。这可以有利于HNSW用于大型表示，而不是非常短的代码(8-16字节)的情况。

[0123] • 本文公开的实施例引入了利用紧致码和来自邻居的量化回归的基于图的相似性搜索方法。它在十亿级高精度范围的基准上实现了目前技术水平的性能。

[0124] 相关工作

[0125] 考虑一组N个元素$\mathcal{X}=\{x_1,\ldots,x_N\}\subset\Omega$和距离$d:\Omega\times\Omega\to\mathbb{R}$(或相似性)，本文公开的实施例解决了寻找查询$y\in\Omega$的最近邻居$\mathcal{N}_{\mathcal{X}}(y)\subset\mathcal{X}$的问题，即最小化距离d(y，x)(或分别最大化相似性)的x的元素{x}。本文公开的实施例通常考虑$\Omega=\mathbb{R}^d$和$d=\ell_2$这种情况，这在计算机视觉应用中可能是高度感兴趣的。

[0126] 某种程度上让人想起90年代的压缩研究领域，人们见证了从无损压缩到有损压缩的快速转变，最近在这一领域的研究工作集中在近似近邻或最近邻搜索[24、23、14、26、18、6]，在这种搜索中，精度的保证与高效率增益是相互矛盾的。

[0127] 近似方法通常通过将距离评估限制到元素的子集来提高效率，元素的子集是基于由空间划分引起的局部性标准来选择的。例如，位置敏感哈希(LSH)方案[12、1]利用了从Johnson-Lindenstrauss引理得到的散列属性。如果真阳性(true positive)不属于所选子集，则会出现误差。

[0128] 近似的另一个来源来自压缩表示，这是由Weber等人[43]首创的以提高搜索效率[43]。随后，Charikar关于草图(sketch)的开创性工作[10]推广了紧致二进制代码，使其成为可扩展性的推动者[25、32]。在这些工作中，以及随后采用向量量化[26]的工作中，距离的近似会引起误差，这导致在排序后的结果列表中交换元素。典型地，通过主成分分析(PCA)降维，然后进行某种形式的量化(例如标量量化[40]，二进制量化[21]和乘积量化或其变体[26、17])来缩减向量。最近的相似性搜索方法通常将这两种近似且互补的策略结合在一起，正如Jégou等人[26]最初提出的那样。量化是分级的，即第一级量化器产生向量的近似版本，并且附加的代码细化该近似[28、3]。

[0129] [26]中的IVFADC方法和IMI[6]是采用两个量化级别的代表性搜索算法。具有相同第一级量化码的所有代码都存储在一个连续的阵列中，称为倒排列表，它被顺序地扫描。Ann Arbor[8]对向量关于一组固定的最近向量进行编码。本文公开的实施例表明，这种选择可能是有害的，并且学习锚向量的集合对于达到良好的精度可能是必要的。

[0130] 基于图的方法。与基于空间划分的方法不同，启发性的NN下降算法[13]构建了一

个knn图来求解所有邻居问题:目标是为每个x∈x关于d找到x中的k个最近邻居。搜索过程通过局部更新进行,并且不是穷举的,即算法在不考虑所有对(x,x′)∈x^2的情况下收敛。NN下降算法也被考虑用于近似最近邻搜索。

[0131] Yuri Malkov等人[33、34]介绍了NN下降算法的最完善的版本,即HNSW。该解决方案选择一系列嵌套的数据库向量子集或“层”。第一层仅包含单个点,并且基础层(base layer)是整个数据集。层的大小遵循几何级数,但它们以其他方式被随机抽样。对于这些层中的每一层,HNSW都构造一个邻居图。搜索从第一层开始。对该层执行贪婪搜索,直到它到达该层中查询的最近邻居。该向量用作下一层的入口点,作为种子点(seed point)再次执行搜索。在由所有点组成的基础层,过程不同:从种子开始的广度优先搜索(bread first search)产生最终的邻居。

[0132] 重要的是,与每个子集相关联的图不是该子集的精确knn图:应当包括长程边。这类似于模拟退火(simulated annealing)或优化中的其他多样化技术:一小部分被评估的元素应当远离。在HNSW中,这种多样化是以自然的方式提供的,这要归功于结构的上层所加强的长程链接,这种链接建立在更少的点上。然而,这还不够,这导致Malkov等人设计了“收缩(shrinking)”算子,该算子以不一定保持最近的邻居的方式来减少向量的邻居列表。

[0133] 初步分析

[0134] 这一节介绍了指导本文公开的实施例的设计的几项研究。所有这些评估都是在$\mathcal{X} = \mathrm{Deep1}M \subset \mathbb{R}^{96}$(即Deep1B数据集[7]的第一个一百万幅图像)上执行的。

[0135] 在特定实施例中,首先执行基于图的HNSW遍历和IMI中采用的基于聚类的哈希方案之间的比较。目标是衡量一种方法在用最少数量的比较识别包含邻居的子集时的有效性。第二种分析考虑了向量的不同估计器,以便在某些假设条件下最好地逼近它,包括oracle提供额外信息(例如邻居)的情况。最后,对用于以压缩形式编码描述符的不同方法进行了比较评估,假设使用近似表示的穷举搜索是可能的。这导致识别目标操作点的吸引人的选择。

[0136] i选择性:HNSW相比于IMI

[0137] 在特定实施例中,考虑了两种用于识别元素子集的流行方法,即HNSW的多尺度图遍历[34]和在IMI中采用的空间划分[6],其依赖于乘积量化器[26]。这两种方法都包括(i)识别阶段,其中查询向量与相对小的向量集(质心或HNSW中的高层)进行比较;以及(ii)比较阶段,其中执行大多数实际距离评估。为了更直接的比较,计算向量之间的精确距离。测量在精度和距离计算数量之间的折衷。这与选择性线性相关:该度量[36]测量必须查找的元素的比例(fraction)。

[0138] 为该设置选择标准设置:对于IMI,使用2^{10}个质心的2个码本,产生大约1M(100万个)倒排列表。对于HNSW,在基础层使用64个邻居,并且在其他层使用32个邻居。在细化阶段,两种方法都从最有希望的候选开始执行代码比较,并存储k个最佳搜索结果。在两种方法中,搜索停止之后的比较次数是搜索时间参数T。图3示出了关于选择性的示例精度变化。在图3中,精度被示为对两种方法(即,IMI 301和HNSW 302)执行的距离计算的数量的函数。该图显示,对于期望的精度水平,HNSW 302比IMI 301的选择性高5至8倍。这种更好的选择性可能不会直接转化为相同的加速,因为HNSW 302需要来自主存储器的许多随机探测,而不是连续的倒排列表。然而,这表明,当随机访问的代价不再支配搜索时间时,对于较大的

向量表示,HNSW 302可能变得总是比IMI 301更好。图4示出了精度和速度之间的示例折衷。在图4中,示出了两种方法(即,中等规模数据集(Deep100M)上的IMI 401和HNSW 402)的折衷。两种方法都使用描述符的96字节编码。图4证实了具有标量量化器的HNSW 402比在两个级别都采用非常精细的量化器的IMI 401更快且更精确。然而,这需要每个向量224字节,当包括数据结构的所有开销时,这相当于50GB的RAM。

[0139] ii质心、邻居或回归?

[0140] 此后,研究了获得向量x∈x的粗略近似值的几种方法。在特定实施例中,这些估计中的一些可以取决于附加信息,例如当假设给定数据库向量x的所有其他顶点都可用时。为了本文公开的研究,它们被报告为顶线结果(topline result)。

[0141] 质心。在特定实施例中,可以通过k均值来学习包括16k个元素的粗略码本C。它可以直接对x来学习,或者可以使用100万个向量的不同训练集来学习。x可以用它的最近邻居$q(x)\in C$来近似。

[0142] 最近邻居。可以假设x的最近邻居$n_1(x)$是已知的,并且它可以用作近似值。该选择显示了搜索算法通过在x中选择单个向量所能达到的上限。

[0143] 加权平均。在这里,可以假设按距离递减排序的x的k=8个最近邻居是可访问的,以矩阵形式存储为$N(x)=[n_1,\cdots,n_k]$。。x可以被估计为加权平均值:

[0144] $$\bar{x} = \beta^{*T}\mathbf{N}(x), \qquad (1)$$

[0145] 其中β^*是x中所有元素共享的固定权重向量常数。β^*的封闭形式计算将在后面详述。

[0146] 回归。同样,可以使用N(x)来估计x,但是可以另外假设最小化x的重建误差的最佳回归系数β(x)是完全已知的。换句话说,计算下面的等式:

[0147] $$\hat{x} = \beta(x)^{T}\mathbf{N}(x), \qquad (2)$$

[0148] 其中β(x)作为超定系统的最小二乘最小值$||x-\beta(x)^TN(x)||^2$被获得。

[0149] 图5示出了不同估计器的误差的示例分布。对于x的不同的估计器,通过基于平方损失的累积概率质量函数的平方欧几里德距离来测量误差。q(x)501表示离对作为不同集合的$q(x)^*$502或x学习的码本的最近质心;$n_1(x)$503表示x中的最近邻居;$\bar{x}$504表示8个邻居的加权平均值;$\hat{x}$505表示来自其8个最近邻居的最佳估计器。根据图5,可以得出几个观察结果。首先,在16k个向量的码本中选择质心q(x)501可能比在100万个向量的64×较大集合x当中选择最近邻居$n_1(x)$503更有效。因此,使用HNSW高层图的向量作为参考点来计算剩余向量可能不是一个令人感兴趣的策略。第二,如果连通性是免费授予的或者是由像在HNSW中那样的设计所要求的,则由$\bar{x}$504实现的性能可以表明,如果存在N(x)的合理的近似,来自其邻居的x的估计可以在没有额外存储的情况下得到改善。第三,在相同的假定下,并且另外假设所有x的参数β(x)都是可用的,可以用等式(2)获得更好的估计器。该观察结果可能是本文公开的实施例中引入的重新排序策略的关键。

[0150] iii编码方法:一级近似

[0151] 在特定实施例中,在给定存储器预算的情况下,评估哪个向量压缩本身是最准确的,以执行初始搜索。这种类型的许多研究集中在非常紧致的代码上,比如每个向量8字节

或16字节。本文公开的实施例对更高精度的操作点感兴趣。此外，对于由几个因素(短列表、探针数量等)参数化的系统，通常会报告结果，这使得难以将编码引起的假阴性(false negative)与搜索过程引起的假阴性分开。

[0152] 为了避免这个比较问题，对数据库向量进行压缩和解压缩，并执行穷举搜索。表1中的所有实验都是在Deep1M(Deep1B数据集的1M个第一图像)上进行的。报告查询的百分比，最近邻居在top1(分别为top10)的结果之中。表1中Ann Arbor[8]的结果取决于亲子关系链接(parenthood link)(4字节)。码本是在所提供的不同训练集上进行训练的。特别地，考虑了乘积量化(PQ[26])和优化乘积量化器(OPQ[17])。采用PQ16x8或OPQ14x2形式的符号，其中这些值分别表示码本的数量和每个子量化器的位数。

[0153] 表1.HNSW:用不同的编码方法在96D的1M向量中进行穷举搜索。

[0154]

编解码器	大小	精度	
	(字节)	召回率@1	召回率@10
无	384	1.000	1.000
标量量化器	96	0.978	1.000
PQ16x8	16	0.335	0.818
PQ8x16	16	0.394	0.881
PQ2x8+OPQ14x8	16	0.375	0.867
PQ1x16+OPQ14x8	16	0.422	0.899
PQ2x16+OPQ12x8	16	0.421	0.904
PQ2x12+OPQ13x8	16	0.382	0.870
AnnArbor [8]	(*) 16	0.421	
OPQ32x8	32	0.604	0.982
PQ1x16+OPQ30x8	32	0.731	0.997
PQ2x16+OPQ28x8	32	0.713	0.996
PQ2x14+OPQ28x8	32	0.693	0.995
PCA8	32	0.017	0.074

[0155] 此外，利用剩余向量的量化器的组合被认为实现了更高精度的性能。在这种情况下，考虑2级残差编解码器，其中第一级是具有65536个质心的向量量化器(在本文使用的符号中被表示为PQ1x16)或乘积量化器(PQ2x12或PQ2x14)。剩余的存储器预算用于存储用于细化编解码器的OPQ码，该编解码器对剩余向量进行编码。请注意，基于IVFADC的方法和变体(如IMI)利用2级编解码器。只有数据结构不同。

[0156] 结果显示2级编解码器比1级编解码器更精确。它们解码的计算成本也更高。对于32字节的操作点，观察到仅仅通过PCA来缩减向量或者用标量量化器来对它们进行编码在精度方面是次优的。使用OPQ给出了更高的精度。由于基于表查找的搜索，它在典型设置中也比标量量化器快。为了与Ann Arbor方法进行比较，还报告了一些关于每个向量16字节的结果。同样的结论成立:具有65536个质心的简单的2级编解码器(例如，PQ1x16+OPQ14 x 8)获得与Ann Arbor相同的编解码器性能。

[0157] 链接和代码方法

[0158] 这一节公开了L&C方法。它可以在考虑非常短的代码(8-32字节)的方法和不考虑存储器约束的方法(如FLANN和HNSW)之间提供目前技术水平的折衷。在给出了索引结构和搜索过程的综述之后，显示了如何在没有额外存储器的情况下从其近似邻居改进索引向量

的重建。然后介绍了具有量化回归系数的一种新的细化过程和用于学习回归码本的优化过程的细节。最后，在固定每个向量的存储器占用时，进行分析以讨论连通性和编码之间的折衷。

[0159] i索引和搜索综述

[0160] 向量近似。在特定实施例中，可以基于一种或更多种编码来生成每个候选内容对象的紧致码。一种或更多种编码可以包括乘积量化器或优化乘积量化器中的一个或更多个。换句话说，可以首先用独立于结构的编码方法对所有索引向量进行压缩。它是一个量化器，其可以形式上映射任何向量$x \in \mathbb{R}^d \mapsto q(x) \in C$，其中C是$\mathbb{R}^d$的有限子集，这意味着q(x)被存储为代码。根据初步分析一节的结果，所有实验都采用两级编码。对于第一级，选择大小为2x12或2x14位的乘积量化器(PQ2x12和PQ2x14)，其计算起来可能更便宜。对于第二级，使用具有任意长度代码的OPQ。

[0161] 基于图的结构。在特定实施例中，可以采用HSNW索引结构(即，稀疏图)，除了可以对其进行修改以使其与本文公开的编码向量一起工作。更准确地说，所有向量可以以编码格式存储，但是加法和查询操作可以使用不对称距离计算来被执行[26]：要插入的查询或向量没有被量化，只有已经被索引的元素被量化。在特定实施例中，稀疏图的多个链接可以由一个或更多个机器学习算法生成。基于一个或更多个机器学习算法生成链接可能是解决构建有助于高效搜索的稀疏图的技术挑战的有效解决方案，因为连接到由机器学习算法生成的任何特定节点的链接的数量少于阈值数量，从而导致稀疏图的链接明显少于传统图。在特定实施例中，图的度数可以在较高的级别固定为k＝32，并且两个图级别之间的大小比可以固定为30，即，图级别1中的元素比图级别0中的元素少30倍。

[0162] 在特定实施例中，稀疏图中的每个节点可以与候选内容对象的聚类相关联。该聚类可以包括一个以上的单个候选内容对象。每个聚类可以与特征向量相关联。在特定实施例中，从稀疏图的多个节点中选择入口节点可以包括基于表示查询内容对象的特征向量和表示候选内容对象的聚类的特征向量之间的比较来选择入口节点。在特定实施例中，入口节点的一个或更多个链接节点可以包括一个或更多个短程链接节点和至少一个长程链接节点。一个或更多个短程链接节点中的每一个可以相对于入口节点具有大于或等于阈值水平的相似性水平。在特定实施例中，至少一个长程链接节点相对于入口节点可以具有小于阈值水平的相似性水平。因此，本文公开的实施例可以具有在相似性搜索期间提高所选择的候选内容对象的多样性的技术优势，因为对于图的聚类中的入口节点，连接节点包括短程链接节点(更类似于入口节点)和长程链接节点(不太类似于入口节点)。

[0163] 细化策略。在特定实施例中，可以采用两阶段搜索策略[28]。在第一阶段期间，选择潜在邻居候选的短列表可以仅依赖于由q(•)引起的一级近似。索引向量可以从它们的紧致码中动态(on-the-fly)重构。第二阶段可能需要对每个向量进行更多的计算，并且可能仅应用于该短列表来对候选进行重新排序。在特定实施例中，可以使用该细化过程的两种变型。第一种变型可以包括0字节细化，其不需要每个向量任何额外的存储。这可以通过从用q(x)编码的其连接的邻居中重新估计候选元素来被执行。另一种变型可以包括通过使用为每个向量存储的一组量化的回归系数来细化向量近似。这些系数可以在构建时为每个索引向量离线学习并选择。在特定实施例中，解压缩表示每个候选内容对象的紧致码可以

基于以下项中的一个或更多个:基于表示候选内容对象的紧致码和与对应于候选内容对象的节点的一个或更多个链接节点相关联的一个或更多个紧致码的细化(即,上述第一变型);或者表示候选内容对象的紧致码和包括多个回归权重向量的回归码本(即,上述第二变型)。可以利用图的连通性并最小化重建误差的细化和回归码本可以是解决重建表示候选内容对象的可靠特征向量的技术挑战的有效解决方案,因为它们使得重建的特征向量更准确地表示候选内容对象。

[0164] ii0字节细化

[0165] 在特定实施例中,细化可以基于一个或更多个损失函数。一个或更多个损失函数可以将表示候选内容对象的特征向量和与候选内容对象相关联的代码矩阵相关联。代码矩阵可以包括表示候选内容对象的紧致码和与对应于候选内容对象的节点的链接节点相关联的紧致码。更具体地说,每个索引向量x可以在图中连接到一组k个其他向量$g_1(x),\ldots,g_k(x)$,按照到x的距离增加来排序。该组可以包括x的一些最近邻居,但不是必须的。根据它们的代码,x可以被重构为q(x),并且每个g_i都可以被重构为q(gi(x))。在特定实施例中,矩阵G(x)可以被定义为叠加重构向量的$G(x)=[q(x),q(g_1(x)),\ldots,q(g_k(x))]$。目标可以包括使用该矩阵来设计比q(x)更好的x的估计器,即,最小化期望的平方重建损失。为此,经验损失在x上可最小化为

[0166] $L(\beta)=\sum_{x\in X}||x-\beta^{T}G(x)||^{2}$。 (3)

注意,考虑到k+1参数的小集合,使用x的子集在实践中可能没有任何区别。在特定实施例中,垂直连接的向量和矩阵可以定义为

[0167] $$X=\begin{bmatrix}x_1\\ \vdots\\ x_N\end{bmatrix} \text{ 和 } Y=\begin{bmatrix}\mathbf{G}(x_1)\\ \vdots\\ \mathbf{G}(x_N)\end{bmatrix} \quad (4)$$

[0168] 以及$L(\beta)=||X-\beta^{T}Y||^{2}$。这可以是具有封闭形式解$\beta^{*}=Y^{*}X$的常规最小二乘问题,其中$Y^{*}$是Y的Moore-Penrose伪逆。在特定实施例中,最小化器(minimizer)β^{*}可以用标准回归器来计算。该回归权重可以由所有索引元素共享,因此可以不涉及任何每向量代码。索引向量可以从与x及其连接向量相关联的紧致码中细化为

[0169] $$\bar{x}=\beta^{*T}\mathbf{G}(x)。 \quad (5)$$

[0170] 在预期和设计中,$\bar{x}$可能是比q(x)更好的x近似,即,它可以减少量化误差。有趣的是,在最终的近似值中,看一下与向量q(x)相对应的β^{*}中的权重系数。如果量化器非常粗糙,它可以小到0.5:在这种情况下,量化误差可能很大,并且可以通过利用邻居来显著减小。相反,如果量化误差有限,权重通常为0.9。

[0171] iii回归码本

[0172] 在特定实施例中,所公开的0字节细化步骤可以被免费授予,假设连接每个索引元素和附近点的图是可访问的。如前所述,如果最优回归系数已知,向量x可以更好地从它的邻居近似得到。这可能需要在每个向量的基础上存储它们,这可能会使每个向量的存储器占用增加4×k字节的浮点值。为了限制额外的存储器开销,可以使用一种方法来学习回归权重向量的码本$B=\{\beta_1,\ldots,\beta_B\}$,其公开如下。在特定实施例中,可以基于多个候选内容对

象和一个或更多个损失函数来学习回归码本。回归码本可以包括每个候选内容对象的一组量化回归系数。在特定实施例中，生成表示每个候选内容对象的特征向量可以进一步基于该组量化回归系数。特别地，目标可以包括最小化经验损失(empirical loss)

[0173] $$L'(\beta)=\sum_{x\in X}\min_{\beta\in B}\|x-\beta^{T}\mathbf{G}(x)\|^{2}$$ 。 (6)

[0174] 直接对回归权重向量执行k均值可以优化回归向量β(x)的ℓ_2重构，但不能优化等式(6)中的损失。在特定实施例中，仅k均值可以用于初始化回归码本。然后，可以在下面两个步骤中交替使用类EM算法(EM-like algorithm)。

[0175] 1.分配。可以给每个向量x分配最小化其重建误差的码本：

[0176] $$\beta(x)=\arg\min_{\beta\in B}\|x-\beta^{T}\mathbf{G}(x)\|^{2}$$ (7)

[0177] 2.更新。对于由β_i方便识别的每个聚类，最佳回归权重可以被确定为

[0178] $$\beta_i^{*}=\arg\min_{\beta}\sum_{x\in X:\,\beta(x)=\beta_i}\|x-\beta^{T}\mathbf{G}(x)\|^{2}$$ (8)

[0179] 并相应地更新$\beta_i\leftarrow\beta_i^{*}$。对于给定的聚类，等式(8)可以与等式(3)的等式相同，除了仅可以在分配给β_i的向量的子集上计算解之外。它可以是如前面所讨论的封闭形式。

[0180] 在实践中，由于B相对较小(B＝256)，可能只需要x的子集来学习代表性码本B。该细化阶段可能需要每个索引向量1个字节来存储从码本B中选择的权重向量。

[0181] 乘积码本。如后面的实验所示，这种回归码本带来的性能改进可能值得每个向量增加额外的存储器。然而，性能可能随着码本大小B的增加而迅速饱和。这是可以预期的，因为估计器$\beta(x)^{T}G(x)$仅跨越$\mathbb{R}^{d}$的(k+1)维子空间k＜＜d。因此，位于零空间ker(G)中的x的投影不能被恢复。

[0182] 在特定实施例中，这个问题可以通过采用由乘积量化[26]启发的策略来规避。在特定实施例中，每个向量可以被平均分割为$x=[x^{1};\ldots,x^{M}]$，其中每个$x^{j}\in\mathbb{R}^{d/M}$，且乘积回归码本$B^{1}\times\ldots\times B^{M}$：(即每个子空间一个码本)可以被学习。在这种情况下，将上标符号扩展到β和G，向量可以被估计为

[0183] $$\hat{x}=[\beta^{1}(x^{1})\mathbf{G}^{1}(x),\ldots,\beta^{M}(x^{M})\mathbf{G}^{M}(x)]$$ (9)

[0184] 其中$\forall j,\beta^{j}(x^{j})\in B^{j}$。这一组可能的估计器可以跨越维度达M×k的子空间。这种细化方法可能需要每个向量M个字节。

[0185] iv L&C:存储器/精度折衷

[0186] 如前所述，HNSW方法可能是目前最快且最准确的索引方法，但是它的可扩展性可能受到其高存储器使用率的限制。由于这个原因，它从未被证明能在十亿规模上工作。在这一小节中，在对每个向量施加固定的存储器预算时，对L&C方法进行分析。三个因素可能导致边际存储器占用(marginal memory foorprint)：

[0187] • 用于初始向量近似的代码，例如OPQ32(32字节)；

[0188] • 每个向量的图链接数k(每个链接4字节)；

[0189] • [可选地]来自具有乘积回归码本的邻居的细化方法使用的M字节。

[0190] L&C符号。为了明确地识别参数设置，采用了L6&OPQ40形式的符号。L6表示在图中使用了每个向量6个链接，并且OPQ40表示首先用OPQ对向量进行编码，每个向量分配40字节。如果可选地使用回归码本，则在表格和附图中用符号M＝4来表示。0编码的情况用M＝0表示。

[0191] 编码与链接。在特定实施例中，可以首先考虑链接数量和分配给压缩编解码器的字节数量之间的折衷。图6示出了取决于64字节的固定存储器预算是否被分配给代码(不同大小的QPQ代码)或链接的示例性能变化。性能变化是基于一个实验，在这个实验中，它从全HNSW表示开始，然后或者是链接的数量或者是为向量存储的维数逐渐减少。在这个实验中，考虑了达到相同的64字节预算的所有设置，并报告了参数T的几个选择的结果。例如，T等于16、64、256和1024，分别由601、602、603和604表示。回想一下，T是限制距离评估数量的参数，它控制比较的总数。在特定实施例中，参数T可以与相似内容对象的迭代选择的完成条件相关联。可以基于一个或更多个查询约束来确定完成条件。基于图6，可以观察到存在由存储器约束施加的明确的折衷。由于链接太少，搜索可能无效，因为算法无法到达所有点。另一方面，当分配给压缩的存储器预算不足时，精度也可能会受到向量的过于强大近似值的影响。有趣的是，增加T会将最佳的折衷朝向为代码分配更多的字节转移。这可能意味着在图中可以到达邻居，但可能需要更多的跳数。

[0192] 编码向量对回归系数。在特定实施例中，链接的数量被固定为6，然后评估在固定总存储器约束下的细化策略。在这种情况下，在分配给压缩编解码器的字节数和分配给细化过程的字节数之间可能存在折衷。在特定实施例中，可以在64字节并且每个索引向量使用k＝6个链接的约束下考虑折衷。从表2中得出的第一个观察结果可以是，本节中提出的两种细化方法都显著降低了总平方损失(square loss)。0编码可能会出现这种情况，因为这可能正是该方法优化的内容。然而，这种更好的重构性能可能不会转化为这种设置中更好的召回率。这种观察结果的原因可能是0编码方法在用精确的邻居进行回归时可以给出明显的增益，但是由图结构提供的那些可能有更多的长程链接。相比之下，第二种细化策略非常有效。用码本对回归系数进行编码显著改善了重构损失和召回率：基于图的细化编码可能比第一级编码更有效，第一级编码可能不知道向量的局部分布。

[0193] 表2.在用于重建的代码和邻居的代码之间分配字节的不同折衷。

[0194]

编解码器	向量量化误差 ($x10^3$)		R@1 穷尽		R@1 T=1024	R@1 T=16384
深度:	100M	1B	100M	1B	100M	1B
L6&OPQ40	24.3	24.3	0.608	0.601	0.427	0.434
L6&OPQ40 M=0	22.7	22.5	0.611	0.600	0.429	0.435
L6&OPQ36 M=4	21.9	21.5	0.608	0.607	0.428	0.434
L6&OPQ32 M=8	20.0	19.8	0.625	0.612	0.438	0.438

[0195] 实验

[0196] 实验通常评估搜索时间与精度的折衷，还考虑向量表示的大小。精度是以查询的

实际最近邻居在排序1或在某个其他排序之前(召回率@排序)返回的情况的分数(fraction)来衡量的。在2.5GHz服务器上,用一个线程的情况下,每个查询的搜索时间以毫秒为单位给出。用多线程批处理搜索是微不足道的,但是计时的可重复性较差。

[0197] i基线和实施

[0198] IMI被选为基线方法,因为大规模索引的最新工作建立在它的基础上[30、4、7、15]并且十亿规模搜索的顶级结果(top results)是由依靠它的方法报告的。Faiss[29]的竞争性实施(在CPU模式下)被用作IMI基线。自动超参数调谐用于优化IMI的操作点。这些参数是所访问代码的数量(T)、多探针数量和用于比较多义代码(polysemous code)的汉明阈值[15]。

[0199] HNSW的实施遵循最初的NM-SLIB版本[9]。最明显的区别是(i)向量是分批添加的,因为整个向量集不适合装在RAM中,以及(ii)HNSW结构是逐层构建的,这可以提高图的质量。利用L&C索引10亿个向量需要大约26个小时:每秒钟可以将超过10000个向量添加到索引中。细化最多执行10个向量。

[0200] 对于编码,系统地执行用优化的乘积量化(OPQ)估计的旋转,以便于二级乘积量化器中的编码。

[0201] ii大规模评估

[0202] 在计算机视觉领域广泛采用的两个大型数据集上对该方法进行评估。BIGANN[28]是一个256维的1B SIFT向量数据集,而Deep1B是一个由CNN提取的图像描述符数据集。这两个数据集都带有一组10000个查询向量,为这些向量提供了真值(ground-truth)最近邻,以及被用于学习量化器的码本的一组不相关的训练向量。使用200万个向量来训练IMI码本,并且使用25万个向量和10次迭代来训练L&C的回归码本。

[0203] 图7A-图7B示出了关于搜索时间对精度的操作点的示例比较。搜索时间是在一个核心上对每个查询进行测量的。图7A示出了关于搜索时间对精度的操作点的示例比较。该比较基于Deep1B,用于每个向量使用96字节的编码,包括IMI(2×14,PQ96)701、L13&C2×14+PQ40 702、L13&C2×14+PQ40 M=1 703和L13&C2×14+PQ40 M=8 704。图7B示出了关于搜索时间对精度的操作点的另一个示例比较。该比较基于BIGANN,包括IMI(2×14,PQ64)705、L7&C2×14+PQ32 706、L7&C2×14+PQ32 M=1707以及L7&C2×14+PQ32 M=8 708。如图7A-图7B所示,对于大多数操作点,L&C方法快得多。例如,在Deep1B上获得50%的召回率@1要快2.5倍。改进可能是由于细化步骤,即从邻居的回归也是显著的。每个向量要多消耗几个字节(最多8个)。对于计算量大的操作点,IMI更适合召回率@1,因为用于链接的4k=52字节可以用来更准确地表示顶点。

[0204] iii与现有技术的比较

[0205] L&C方法进一步与文献中报告的其他结果进行比较。表3示出了这种比较。L&C方法使用的存储器可能比其他主要关注优化存储器和精度之间的折衷的方法要多得多。然而,与HNSW不同,L&C方法可以在一台服务器上轻松扩展到10亿个向量。当时间预算很小时,L&C方法可能是有竞争力的,并且人们对更高的精度感兴趣。竞争方法要么慢得多,要么明显不够准确。在Deep1B上,只有多义代码获得了与L&C方法相似的效率,并且占用的存储器更少。然而,它仅获得召回率@1=45.6%,相比而言,L&C为66.7%。考虑到召回率@1,L&C方法在精度/速度折衷方面远远优于BIGANN的现有技术。可以注意到,用其他方法增加编码大小会增

加精度，但也会不可避免地增加搜索时间。考虑到，在一般应用中，存储器占用可以保持等同于或小于与图像相关联的其他元数据，对于这些元数据，L&C方法可以在大多数应用中提供吸引人且实用的解决方案。

[0206] 表3.在20亿大小的数据集上与现有技术进行比较。

	R@1	R@10	R@100	时间（ms）	字节
BIGANN					
Multi-LOPQ [30]	0.430	**0.761**	0.782	8	16
OMulti-D-OADC-L [6]	0.421	0.755	0.782	7	16
FBPQ [4]	0.179	0.523	0.757	1.9	16
[0207]	0.186	0.556	**0.894**	9.7	16
多义[15]	0.330		0.856	2.77	16
L7&C32 M=8	**0.461**	0.608	0.613	2.10	72
Deep1B					
GNO-IMI [7]	0.450	0.8		20	16
多义[15]	0.456			3.66	20
L13&C40 M=8	**0.668**	**0.826**	**0.830**	3.50	108

[0208] 结论

[0209] 在本文公开的实施例中，已经公开了一种用于在十亿大小的数据集中精确近似最近邻搜索的方法，即L&C。L&C方法可以针对高精度方案，这对于大量应用是重要的。L&C方法可以在成功的压缩域和基于图的方法之间架起一座桥梁。基于图的候选生成可以提供比基于倒排列表的传统结构更高的选择性。压缩域搜索可以让搜索在普通服务器上扩展到数十亿个向量。作为一个关键的新颖性，它显示了图结构可以被用来改善对于中等甚至零存储器预算的距离估计。其结果是，在高精度方案中，两个公共十亿规模基准的现有技术结果被报告。L&C方法在Faiss库中是开源的。

[0210] 参考文献

[0211] 以下参考文献列表对应于上述引用：

[0212] [1]Alexander I Rudnicky,Eric H Thayer,Paul C Constantinides,Chris Tchou,R Shern,Kevin A Lenzo,Wei Xu和Alice Oh.Creating natural dialogs in the carnegie mellon communicator system.In Eurospeech,1999.

[0213] [2]Antoine Raux,Brian Langner,Dan Bohus,Alan W Black和Maxine Eskenazi.Let's go public!taking a spoken dialog system to the real world.In Interspeech,2005.

[0214] [3]Steve Young,Milica Gašić,Blaise Thomson和Jason D Williams.Pomdp-based statistical spoken dialog systems:A review.Proceedings of the IEEE,101(5):1160-1179,2013.

[0215] [4]Bing Liu,Gokhan Tur,Dilek Hakkani-Tur,Pararth Shah和Larry Heck.End-to-end optimization of task-oriented dialogue model with deep reinforcement learning.arXiv preprint arXiv:1711.10712,2017.

[0216] [5]Bing Liu,Gokhan Tur,Dilek Hakkani-Tur,Pararth Shah和Larry

Heck.Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems.arXiv preprint arXiv:1804.06512,2018.

[0217] [6]Pararth Shah,Dilek Hakkani-Tur,Bing Liu和Gokhan Tur.Bootstrapping a neural con¬versational agent with dialogue self-play,crowdsourcing and on-line reinforcement learning.In Proceedings of the 2018Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies,Volume 3(Industry Papers),volume 3,第41-51页,2018.

[0218] [7]Yang Liu,Kun Han,Zhao Tan和Yun Lei.Using context information for dialog act classification in dnn framework.In Proceedings of the 2017Conference on Empirical Methods in Natural Language Processing,第2170-2178页,2017.

[0219] [8]Jian Pei,Jiawei Han,Behzad Mortazavi-Asl,Helen Pinto,Qiming Chen,Umeshwar Dayal和Mei-Chun Hsu.Prefixspan:Mining sequential patterns efficiently by prefix-projected pattern growth.In icccn,page 0215.IEEE,2001.

[0220] [9]Milica Gasic和Steve Young.Gaussian processes for pomdp-based dialogue manager opti¬mization.IEEE/ACM Transactions on Audio,Speech,and Language Processing,2014.

[0221] [10]Pararth Shah,Dilek Hakkani-Tür和Larry Heck.Interactive reinforcement learning for task-oriented dialogue management.In NIPS 2016Deep Learning for Action and Interaction Workshop,2016.

[0222] [11]Pei-Hao Su,Milica Gasic,Nikola Mrksic,Lina Rojas-Barahona,Stefan Ultes,David Vandyke,Tsung-Hsien Wen和Steve Young.On-line active reward learning for policy optimisation in spoken dialogue systems.In ACL,2016.

[0223] [12]Pei-Hao Su,Pawel Budzianowski,Stefan Ultes,Milica Gasic和Steve Young.Sample-efficient actor-critic reinforcement learning with supervised data for dialogue management.In SIGDIAL,2017.

[0224] [13]Tsung-Hsien Wen,David Vandyke,NikolaMrkšié,MilicaGašié,Lina M.Rojas-Barahona,Pei-Hao Su,Stefan Ultes和Steve Young.A network-based end-to-end trainable task-oriented dialogue system.In EACL,2017.

[0225] [14]Antoine Bordes和Jason Weston.Learning end-to-end goal-oriented dialog.In International Conference on Learning Representations,2017.

[0226] [15]Bing Liu和Ian Lane.An end-to-end trainable neural network model with belief tracking for task-oriented dialog.In Interspeech,2017.

[0227] [16]Tiancheng Zhao和Maxine Eskenazi.Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning.In SIGDIAL,2016.

[0228] [17]Xuijun Li,Yun-Nung Chen,Lihong Li和Jianfeng Gao.End-to-end task-completion neural dialogue systems.arXiv preprint arXiv:1703.01008,2017.

[0229] [18]Baolin Peng,Xiujun Li,Lihong Li,Jianfeng Gao,Asli Celikyilmaz,Sungjin Lee和Kam-Fai Wong.Composite task-completion dialogue policy learning via hierarchical deep reinforcement learning.In Proceedings of EMNLP,2017.

[0230] [19]Jan Ruben Zilke,Eneldo Loza Mencía和Frederik Janssen.Deepred-rule extraction from deep neural networks.In International Conference on Discovery Science,第457-473页.Springer,2016.

[0231] [20]Hiroshi Tsukimoto.Extracting rules from trained neural networks.IEEE Transactions on Neural Networks,11(2):377-389,2000.

[0232] [21]Rudy Setiono和Wee Kheng Leow.Fernn:An algorithm for fast extraction of rules from neural networks.Applied Intelligence,12(1-2):15-25,2000.

[0233] [22]Ismail A Taha和Joydeep Ghosh.Symbolic interpretation of artificial neural networks.IEEE Transactions on knowledge and data engineering,11(3):448-463,1999.

[0234] [23]Marco Tulio Ribeiro,Sameer Singh和Carlos Guestrin.Why should i trust you?:Explaining the predictions of any classifier.In Proceedings of the22nd ACM SIGKDD international conference on knowledge discovery and data mining,第1135-1144页,ACM,2016.

[0235] [24]Dumitru Erhan,Yoshua Bengio,Aaron Courville和Pascal Vincent.Visualizing higher-layer features of a deep network.University of Montreal,1341(3):1,2009.

[0236] [25]Quoc V Le.Building high-level features using large scale unsupervised learning.In Acoustics,Speech and Signal Processing(ICASSP),2013IEEE International Conference on,第8595-8598页.IEEE,2013.

[0237] [26]Karen Simonyan,Andrea Vedaldi和Andrew Zisserman.Deep inside convolutional networks:Visualising image classification models and saliency maps.arXiv preprint arXiv:1312.6034,2013.

[0238] [27]Rahul Iyer,Yuezhang Li,Huao Li,Michael Lewis,Ramitha Sundar和Katia Sycara.Transparency and explanation in deep reinforcement learning neural networks.2018.

[0239] [28]Rakesh Agrawal,Ramakrishnan Srikant等人.Fast algorithms for mining association rules.In Proc.20th int.conf.very large data bases,VLDB,volume 1215,第487-499页,1994.

[0240] [29]Jiawei Han,Jian Pei和Yiwen Yin.Mining frequent patterns without candidate generation.In ACM sigmod record,volume 29,第1-12页.ACM,2000.

[0241] 图8示出了基于图和紧致码的相似性搜索的示例方法800。该方法可以开始于步骤810,其中社交网络系统160可以接收搜索查询,其中搜索查询包括查询内容对象和一个或更多个查询约束。在步骤820,社交网络系统160可以生成表示查询内容对象的特征向量。在步骤830,社交网络系统160可以访问稀疏图,该稀疏图包括对应于多个候选内容对象的多

个节点和连接这些节点的多个链接，其中连接到任何特定节点的链接的数量小于链接的阈值数量，并且其中每个候选内容对象由紧致码表示。在步骤840，社交网络系统160可以从稀疏图的多个节点中选择入口节点。在步骤850，社交网络系统160可以从多个候选内容对象中选择关于查询内容对象的一个或更多个相似内容对象，其中通过以下子步骤迭代地选择一个或更多个相似内容对象。在子步骤850a，社交网络系统160可以识别入口节点的一个或更多个链接节点，其中每个链接节点通过链接连接到入口节点。在子步骤850b，社交网络系统160可以为对应于链接节点的每个候选内容对象解压缩表示候选内容对象的紧致码，以生成表示候选内容对象的特征向量。在子步骤850c，社交网络系统160可以基于表示查询内容对象的特征向量和表示候选内容对象的特征向量之间的比较，从候选内容对象中选择零个或更多个相似内容对象。在子步骤850d，社交网络系统160可以确定是否达到选择一个或更多个相似内容对象的完成条件，其中完成条件是基于一个或更多个查询约束来确定的。基于该确定，如果满足完成条件，则在子步骤850e，社交网络系统160可以返回一个或更多个所选择的相似内容对象；否则，如果不满足完成条件，则在子步骤850f，社交网络系统160可以通过使用来自一个或更多个链接节点的对应于最相似内容对象的链接节点作为入口节点来重复迭代选择。在步骤860，社交网络系统160可以响应于接收到搜索查询，向与用户相关联的客户端系统发送用于呈现一个或更多个所选择的相似内容对象的指令。在适当的情况下，特定实施例可以重复图8的方法的一个或更多个步骤。尽管本公开将图8的方法的特定步骤描述并示出为以特定顺序发生，但是本公开设想了以任何适当顺序发生的图8的方法的任何适当步骤。此外，尽管本公开描述并示出了基于图和紧致码的、包括图8的方法的特定步骤的相似性搜索的示例方法，但是本公开设想了基于图和紧致码的、包括任何合适的步骤的相似性搜索的任何合适的方法，在适当的情况下，可以包括图8的方法的所有、一些步骤或不包括这些步骤。此外，尽管本公开描述并示出了执行图8的方法的特定步骤的特定组件、设备或系统，但是本公开设想了执行图8的方法的任何适当步骤的任何适当组件、设备或系统的任何适当组合。

[0242] 社交图

[0243] 图9示出了示例社交图900。在特定实施例中，社交网络系统160可以在一个或更多个数据储存器中存储一个或更多个社交图900。在特定实施例中，社交图900可以包括多个节点——该多个节点可以包括多个用户节点902或多个概念节点904——以及连接这些节点的多条边906。每个节点可以与唯一的实体(即，用户或概念)相关联，每个实体可以具有唯一的标识符(ID)，例如唯一的编号或用户名。出于教导的目的，图9中以二维视觉图表示(two-dimensional visual map representation)示出了示例社交图900。在特定实施例中，社交网络系统160、客户端系统130或第三方系统170可以访问社交图900和相关社交图信息以用于合适的应用。社交图900的节点和边可以作为数据对象被存储在例如数据储存器(例如社交图数据库)中。这种数据储存器可以包括社交图900的节点或边的一个或更多个可搜索或可查询的索引。

[0244] 在特定实施例中，用户节点902可以对应于社交网络系统160的用户。作为示例而不是作为限制，用户可以是与社交网络系统160或通过社交网络系统160互动或通信的个人(人类用户)、实体(例如，企业、公司或第三方应用)或(例如，个人或实体的)团体。在特定实施例中，当用户向社交网络系统160注册账户时，社交网络系统160可以创建对应于用户的

用户节点902，并将用户节点902存储在一个或更多个数据储存器中。本文描述的用户和用户节点902在适当的情况下可以指注册的用户和与注册的用户相关联的用户节点902。另外或作为备选方案，在适当的情况下，本文描述的用户和用户节点902可以指没有向社交网络系统160注册的用户。在特定实施例中，用户节点902可以与由用户提供的信息或由各种系统(包括社交网络系统160)收集的信息相关联。作为示例而不是作为限制，用户可以提供他或她的姓名、简档图片、联系信息、出生日期、性别、婚姻状况、家庭状况、职业、教育背景、偏好、兴趣或其他人口统计信息。在特定实施例中，用户节点902可以与对应于与用户相关联的信息的一个或更多个数据对象相关联。在特定实施例中，用户节点902可以对应于一个或更多个web界面。

[0245] 在特定实施例中，概念节点904可以对应于概念。作为示例而不是作为限制，概念可以对应于地点(例如，电影院、餐馆、地标或城市)；网站(例如，与社交网络系统160相关联的网站或与web应用服务器相关联的第三方网站)；实体(例如，个人、企业、团体、运动队或名人)；资源(例如，音频文件、视频文件、数字照片、文本文件、结构化文档或应用)，其可以位于社交网络系统160内或外部服务器(例如web应用服务器)上；不动产或知识产权(例如，雕塑、绘画、电影、游戏、歌曲、想法、照片或书面作品)；游戏；活动；想法或理论；在增强/虚拟现实环境中的对象；另一个合适的概念；或者两个或更多个这样的概念。概念节点904可以与由用户提供的概念的信息或由各种系统(包括社交网络系统160)收集的信息相关联。作为示例而不是作为限制，概念的信息可以包括名称或标题；一个或更多个图像(例如，书籍的封面的图像)；位置(例如，地址或地理位置)；网站(其可以与URL相关联)；联系信息(例如，电话号码或电子邮件地址)；其他合适的概念信息；或者这样的信息的任何合适的组合。在特定实施例中，概念节点904可以与一个或更多个数据对象相关联，一个或更多个数据对象对应于与概念节点904相关联的信息。在特定实施例中，概念节点904可以对应于一个或更多个web界面。

[0246] 在特定实施例中，社交图900中的节点可以表示web界面(其可以被称为“简档界面”)或者由web界面表示。简档界面可以由社交网络系统160托管或是社交网络系统160可访问的。简档界面也可以在与第三方系统170相关联的第三方网站上被托管。作为示例而不是作为限制，对应于特定外部web界面的简档界面可以是特定外部web界面，并且简档界面可以对应于特定概念节点904。简档界面可以由其他用户的全部或选定子集可查看。作为示例而不是作为限制，用户节点902可以具有相应的用户简档界面，其中相应的用户可以添加内容、作出声明或以其他方式表达他或她自己。作为另一示例而不是作为限制，概念节点904可以具有相应的概念简档界面，其中一个或更多个用户可以添加内容、作出声明或表达他们自己，特别是关于对应于概念节点904的概念。

[0247] 在特定实施例中，概念节点904可以表示由第三方系统170托管的第三方web界面或资源。第三方web界面或资源除了其它元素以外还可以包括表示动作或活动的内容、可选择的图标或其他图标或其他可交互对象(其可以例如用JavaScript、AJAX或PHP代码来实现)。作为示例而不是作为限制，第三方web界面可以包括可选择的图标，例如“赞”、“签到(check-in)”、“吃”、“推荐”，或其他合适的动作或活动。查看第三方web界面的用户可以通过选择图标之一(例如，“签到”)来执行动作，使客户端系统130向社交网络系统160发送指示用户的动作的消息。响应于该消息，社交网络系统160可以在对应于用户的用户节点902

和对应于第三方web界面或资源的概念节点904之间创建边(例如,签到类型边),并将边906存储在一个或更多个数据储存器中。

[0248] 在特定实施例中,社交图900中的一对节点可以通过一条或更多条边906连接到彼此。连接一对节点的边906可以表示在该对节点之间的关系。在特定实施例中,边906可以包括或表示对应于在一对节点之间的关系的一个或更多个数据对象或属性。作为示例而不是作为限制,第一用户可以指示第二用户是第一用户的“朋友”。响应于该指示,社交网络系统160可以向第二用户发送“朋友请求”。如果第二用户确认“朋友请求”,社交网络系统160可以在社交图900中创建将第一用户的用户节点902连接到第二用户的用户节点902的边906,并将边906作为社交图信息存储在一个或更多个数据储存器164中。在图9的示例中,社交图900包括指示在用户“A”和用户“B”的用户节点902之间的朋友关系的边906,以及指示在用户“C”和用户“B”的用户节点902之间的朋友关系的边。尽管本公开描述或示出了连接特定用户节点902的具有特定属性的特定边906,但是本公开设想了连接用户节点902的具有任何适当属性的任何适当边906。作为示例而不是作为限制,边906可以表示友谊、家庭关系、商业或雇佣关系、粉丝关系(包括例如,赞等)、关注者(follower)关系、访问者关系(包括例如,访问、查看、签到、分享等)、订购者关系、上级/下级关系、互惠关系、非互惠关系、另一种合适类型的关系、或两种或更多种这样的关系。此外,尽管本公开一般将节点描述为被连接,但是本公开也将用户或概念描述为被连接。在本文,对被连接的用户或概念的引用在适当的情况下可以指在社交图900中由一条或更多条边906连接的对应于那些用户或概念的节点。分别由两个节点表示的两个对象之间的分离度(degree of separation)是连接社交图900中的两个节点的最短路径中的边数。作为示例而非限制,在社交图900中,用户“C”的用户节点902经由多条路径连接到用户“A”的用户节点902,例如,直接通过用户“B”的用户节点902的第一路径,通过公司“Acme”的概念节点904和用户“D”的用户节点902的第二路径,以及通过代表学校“斯坦福”、用户“G”、公司“Acme”和用户“D”的用户节点902和概念节点904的第三路径。用户“C”和用户“A”具有分离度2,因为连接它们相应节点的最短路径(即,第一路径)包括两条边906。

[0249] 在特定实施例中,在用户节点902和概念节点904之间的边906可以表示由与用户节点902相关联的用户朝着与概念节点904相关联的概念执行的特定动作或活动。作为示例而不是作为限制,如图9所示,用户可以“赞”、“出席”、“播放”、“收听”、“烹饪”、“工作于”或“观看”概念,其中每个可以对应于边类型或子类型。对应于概念节点904的概念简档界面可以包括例如可选择的“签到”图标(例如,可点击的“签到”图标)或可选择的“添加到收藏夹”图标。类似地,在用户点击这些图标之后,社交网络系统160可以响应于对应于相应动作的用户动作来创建“收藏夹”边或“签到”边。作为另一示例而不是作为限制,用户(用户“C”)可以使用特定的应用(声田(SPOTIFY),其为在线音乐应用)来收听特定的歌曲(“想象(Imagine)”)。在这种情况下,社交网络系统160可以在对应于用户的用户节点902和对应于歌曲和应用的概念节点904之间创建“收听”边906和“使用”边(如图9所示),以指示用户收听了歌曲并使用了应用。此外,社交网络系统160可以在对应于歌曲和应用的概念节点904之间创建“播放”边906(如图9所示),以指示特定歌曲由特定应用播放。在这种情况下,“播放”边906对应于由外部应用(声田)对外部音频文件(歌曲“想象”)执行的动作。尽管本公开描述了连接用户节点902和概念节点904的具有特定属性的特定边906,但是本公开设想了

连接用户节点902和概念节点904的具有任何适当属性的任何适当边906。此外，尽管本公开描述了表示单个关系的在用户节点902和概念节点904之间的边，但是本公开设想了表示一个或更多个关系的在用户节点902和概念节点904之间的边。作为示例而不是作为限制，边906可以表示用户喜欢并使用了特定概念。可选地，另一条边906可以表示用户节点902和概念节点904之间的每种类型的关系(多个单一关系)(如图9中用户“E”的用户节点902和“声田”的概念节点904之间所示)。

[0250] 在特定实施例中，社交网络系统160可以在社交图900中的用户节点902和概念节点904之间创建边906。作为示例而不是作为限制，(例如，通过使用由用户的客户端系统130托管的web浏览器或专用应用)查看概念简档界面的用户可以通过点击或选择“赞”图标来指示他或她喜欢由概念节点904表示的概念，这可以使用户的客户端系统130向社交网络系统160发送指示用户喜欢与概念简档界面相关联的概念的消息。响应于该消息，社交网络系统160可以在与用户相关联的用户节点902和概念节点904之间创建边906，如由在用户和概念节点904之间的“赞”边906所示的。在特定实施例中，社交网络系统160可以将边906存储在一个或更多个数据储存器中。在特定实施例中，边906可以由社交网络系统160响应于特定用户动作而自动形成。作为示例而不是作为限制，如果第一用户上传图片、观看电影或收听歌曲，则可以在对应于第一用户的用户节点902和对应于那些概念的概念节点904之间形成边906。尽管本公开描述了以特定方式形成特定边906，但是本公开设想了以任何合适的方式形成任何合适的边906。

[0251] 在线社交网络上的搜索查询

[0252] 在特定实施例中，社交网络系统160可以从在线社交网络的用户的客户端系统接收由用户输入的查询。用户可以通过例如选择查询输入或将文本输入到查询字段(query field)来将查询提交给社交网络系统160。在线社交网络的用户可以通过向搜索引擎提供描述主题的短语(常常被称为“搜索查询”)来搜索与特定主题(例如，用户、概念、外部内容或资源)相关的信息。该查询可以是非结构化文本查询，并且可以包括一个或更多个文本串(其可以包括一个或更多个n-gram)。通常，用户可以将任何字符串输入到查询字段中以在社交网络系统160上搜索与文本查询匹配的内容。社交网络系统160然后可以搜索数据储存器164(或者特别是社交图数据库)以识别与查询匹配的内容。搜索引擎可以使用各种搜索算法基于查询短语来进行搜索，并生成识别最可能与搜索查询相关的资源或内容(例如，用户简档界面、内容简档界面或外部资源)的搜索结果。为了进行搜索，用户可以向搜索引擎输入或发送搜索查询。作为响应，搜索引擎可以识别可能与搜索查询相关的一个或更多个资源，每个资源可以单独地被称为“搜索结果”，或者共同被称为对应于搜索查询的“搜索结果”。所识别的内容可以包括例如社交图元素(即，用户节点902、概念节点904、边906)、简档界面、外部web界面或其任何组合。社交网络系统160然后可以生成具有对应于所识别的内容的搜索结果的搜索结果界面，并将搜索结果界面发送给用户。搜索结果常常可以以在搜索结果界面上的链接的列表的形式被呈现给用户，每个链接与包含一些所识别的资源或内容的不同界面相关联。在特定实施例中，在搜索结果中的每个链接可以是以统一资源定位符(URL)的形式，该URL指定相应界面所位于的地方以及用于取回它的机制。社交网络系统160然后可以将搜索结果界面发送到在用户客户端系统130上的web浏览器132。然后，用户可以点击URL链接或以其他方式从搜索结果界面选择内容，以视情况从社交网络系统160或

从外部系统(例如,第三方系统170)访问内容。资源可以根据它们与搜索查询的关联的相对程度而被排名并呈现给用户。搜索结果也可以根据它们与用户的关联的相对程度而被排名并呈现给用户。换句话说,可以基于例如社交图信息、用户信息、用户的搜索或浏览历史或者与用户相关的其他合适的信息,针对查询用户来个性化搜索结果。在特定实施例中,资源的排名可以由搜索引擎所实现的排名算法来确定。作为示例而不是作为限制,与搜索查询或与用户更相关的资源可以比与搜索查询或与用户不太相关的资源排名更高。在特定实施例中,搜索引擎可以将它的搜索限制到在线社交网络上的资源和内容。然而,在特定实施例中,搜索引擎还可以在其他源(例如第三方系统170、互联网或万维网或者其他合适的源)上搜索资源或内容。尽管本公开描述了以特定方式查询社交网络系统160,但本公开设想了以任何合适的方式查询社交网络系统160。

[0253] 输入提示(Typeahead)过程和查询

[0254] 在特定实施例中,一个或更多个客户端侧和/或后端(服务器侧)过程可以实现并利用“typeahead”特征,该特征可以自动尝试将社交图元素(例如,用户节点902、概念节点904或边906)与由用户当前在结合所请求的界面(例如,用户简档界面、概念简档界面、搜索结果界面、与在线社交网络相关联的原生应用的用户界面/视图状态、或者在线社交网络的另一个合适的界面)再现的输入表格中输入的信息进行匹配,所请求的界面可以由社交网络系统160托管或者在社交网络系统160中可访问。在特定实施例中,当用户输入文本以作出声明时,typeahead特征可以尝试将在声明中输入的文本字符串与对应于用户、概念或边及它们在社交图900中的对应元素的字符串(例如,名称、描述)进行匹配。在特定实施例中,当匹配被找到时,typeahead特征可以用对现有社交图元素的社交图元素的引用(例如,节点名称/类型、节点ID、边名称/类型、边ID或另一个合适的引用或标识符)来自动填充表格。在特定实施例中,当用户将字符输入到表格框中时,typeahead过程可以读取所输入的文本字符串。当进行每个键击时,前端typeahead过程可以将所输入的字符串作为请求(或调用)发送到在社交网络系统160内执行的后端typeahead过程。在特定实施例中,typeahead过程可以使用一个或更多个匹配算法来尝试识别匹配的社交图元素。在特定实施例中,当一个或更多个匹配被找到时,typeahead过程可以向用户的客户端系统130发送响应,该响应可以包括例如匹配的社交图元素的名称(名称字符串)或描述以及潜在地与匹配的社交图元素相关联的其他元数据。作为示例而不是作为限制,如果用户将字符“pok”输入到查询字段中,则typeahead过程可以显示下拉菜单,该下拉菜单显示匹配的现有简档界面和相应的用户节点902或概念节点904的名称(例如被命名为或专用于“poker”或“pokemon”的简档界面),然后用户可以点击或以其他方式选择这些名称,从而确认声明对应于选定节点的匹配用户或概念名称的期望。

[0255] 可以在2010年4月19日提交的第12/763162号美国专利申请和2012年7月23日提交的第13/556072号美国专利申请中找到关于typeahead过程的更多信息,这些专利申请通过引用被并入。

[0256] 在特定实施例中,本文描述的typeahead过程可以应用于由用户输入的搜索查询。作为示例而不是作为限制,当用户将文本字符输入到查询字段中时,typeahead过程可以尝试识别匹配当用户输入字符时被输入到查询字段中的字符串的一个或更多个用户节点902、概念节点904或边906。当typeahead过程从文本查询接收包括字符串或n-gram的请求

或调用时，typeahead过程可以执行或被促使执行搜索以识别具有相应的名称、类型、类别或与所输入的文本匹配的其他标识符的现有的社交图元素(即，用户节点902、概念节点904、边906)。typeahead过程可以使用一个或更多个匹配算法来尝试识别匹配的节点或边。当一个或更多个匹配被找到时，typeahead过程可以向用户的客户端系统130发送响应，该响应可以包括例如匹配节点的名称(名称字符串)以及潜在地与匹配节点相关联的其他元数据。typeahead过程然后可以显示下拉菜单，该下拉菜单显示匹配的现有简档界面和相应的用户节点902或概念节点904的名称，并且显示可以连接到匹配的用户节点902或概念节点904的匹配的边906的名称，然后用户可以点击或以其他方式选择该下拉菜单，从而确认搜索对应于选定节点的所匹配的用户或概念名称或者搜索通过匹配的边连接到所匹配的用户或概念的用户或概念的愿望。可选地，typeahead过程可以简单地用排名最高的匹配的名称或其他标识符来自动填充表格，而不是显示下拉菜单。然后，用户可以简单地通过在键盘上键入“enter”或通过点击自动填充的声明来确认自动填充的声明。当用户确认匹配的节点和边时，typeahead过程可以发送请求，该请求通知社交网络系统160用户确认了包含匹配的社交图元素的查询。响应于所发送的请求，社交网络系统160可以自动地(或者可选地基于在请求中的指令)调用或者以其他方式搜索社交图数据库，以找到匹配的社交图元素，或者视情况找到连接到匹配的社交图元素的社交图元素。尽管本公开描述了以特定方式应用typeahead过程来搜索查询，但本公开设想了以任何合适的方式应用typeahead过程来搜索查询。

[0257] 结合搜索查询和搜索结果，特定实施例可以利用在2006年8月11日提交的第11/503093号美国专利申请、2010年12月22日提交的第12/977027号美国专利申请和2010年12月23日提交的第12/978265号美国专利申请中公开的一个或更多个系统、部件、元件、功能、方法、操作或步骤，这些专利申请通过引用被并入。

[0258] 结构化搜索查询

[0259] 在特定实施例中，响应于从第一用户(即，查询用户)接收到的文本查询，社交网络系统160可以解析文本查询并识别对应于特定社交图元素的文本查询的部分。然而在一些情况下，查询可以包括模糊的一个或更多个项，其中模糊项是也许可能对应于多个社交图元素的项。为了解析模糊项，社交网络系统160可以访问社交图900，并然后解析文本查询以从文本查询识别对应于模糊n-gram的社交图元素。社交网络系统160然后可以生成一组结构化查询，其中每个结构化查询对应于可能匹配的社交图元素之一。这些结构化查询可以基于由语法模型生成的字符串，使得它们用对相关社交图元素的引用以自然语言语法来被再现。作为示例而不是作为限制，响应于文本查询“向我展示我女朋友的朋友”，社交网络系统160可以生成结构化查询“斯蒂芬妮(Stephanie)的朋友”，其中在结构化查询中的“朋友”和“斯蒂芬妮”是对应于特定社交图元素的引用。对“斯蒂芬妮”的引用将对应于特定的用户节点902(其中社交网络系统160已经解析n-gram“我的女朋友”以对应于用户“斯蒂芬妮”的用户节点902)，而对“朋友”的引用将对应于将该用户节点902连接到其他用户节点902的朋友类型边906(即，连接到“斯蒂芬妮的”一级朋友的边906)。当执行该结构化查询时，社交网络系统160可以识别由朋友类型边906连接到对应于“斯蒂芬妮”的用户节点902的一个或更多个用户节点902。作为另一示例而不是作为限制，响应于文本查询“在脸书(facebook)工作的朋友”，社交网络系统160可以生成结构化查询“在脸书工作的我的朋友”，其中在结构

化查询中的“我的朋友”、“在...工作”和“脸书”是对应于如前所述的特定社交图元素的引用(即,朋友类型边906、工作类型边906和对应于公司“脸书”的概念节点904)。通过响应于用户的文本查询而提供所建议的结构化查询,社交网络系统160可以为在线社交网络的用户提供强大方式以基于他们的社交图属性及他们与各种社交图元素的关系来搜索在社交图900中表示的元素。结构化查询可以允许查询用户搜索通过特定的边类型连接到社交图900中特定用户或概念的内容。结构化查询可以被发送给第一用户并显示在下拉菜单中(经由例如客户端侧typeahead过程),其中第一用户然后可以选择适当的查询以搜索期望的内容。使用本文描述的结构化查询的一些优点包括基于有限的信息来找到在线社交网络的用户,基于来自在线社交网络的内容与各种社交图元素的关系来聚集该内容的虚拟索引,或者找到与你和/或你的朋友相关的内容。尽管本公开描述了以特定方式生成特定的结构化查询,但本公开设想了以任何合适的方式生成任何合适的结构化查询。

[0260] 可以在2012年7月23日提交的第13/556072号美国专利申请、2012年12月31日提交的第13/731866号美国专利申请和2012年12月31日提交的第13/732101号美国专利申请中找到关于元素检测和解析查询的更多信息,这些申请中的每一个通过引用被并入。可以在2012年7月23日提交的第13/556072号美国专利申请、2012年11月12日提交的第13/674695号美国专利申请和2012年12月31日提交的第13/731866号美国专利申请中找到关于结构化搜索查询和语法模型的更多信息,这些申请中的每一个通过引用被并入。

[0261] 生成关键字和关键字查询

[0262] 在特定实施例中,当用户将文本串输入到查询字段内时,社交网络系统160可以向查询用户提供定制的关键字补全(keyword completion)建议。关键字补全建议可以以非结构化格式提供给用户。为了生成关键字补全建议,社交网络系统160可以访问在社交网络系统160内的多个源以生成关键字补全建议,对来自多个源的关键字补全建议评分,并然后将关键字补全建议返回给用户。作为示例而不是作为限制,如果用户键入查询“朋友斯坦(stan)”,则社交网络系统160可以建议例如“朋友斯坦福(stanford)”、“朋友斯坦福大学(stanford university)”、“朋友斯坦利(stanley)”、“朋友斯坦利库珀(stanley cooper)”、“朋友斯坦利库布里克(stanley kubrick)”、“朋友斯坦利杯(stanley cup)”和“朋友斯坦龙斯基(stanlonski)”。在该示例中,社交网络系统160正在建议作为模糊n-gram“斯坦”的修改的关键字,其中这些建议可以从各种关键字生成器生成。社交网络系统160可能已经选择了关键字补全建议,因为用户以某种方式连接到这些建议。作为示例而不是作为限制,查询用户可以在社交图900内例如由赞类型或出席类型的边906连接到对应于斯坦福大学的概念节点904。查询用户也可能有名叫斯坦利·库珀的朋友。尽管本公开描述了以特定方式生成关键字补全建议,但本公开设想了以任何合适的方式生成关键字补全建议。

[0263] 可以在2014年4月3日提交的第14/244748号美国专利申请、2014年8月27日提交的第14/470607号美国专利申请和2014年12月5日提交的第14/561418号美国专利申请中找到关于关键字查询的更多信息,这些申请中的每一个通过引用被并入。

[0264] 基于对象类型的索引

[0265] 图10示出了用于存储社交网络系统160的对象的示例划分。多个数据储存器164(其也可以称为“垂直(vertical)”)可以存储社交网络系统160的对象。存储在数据储存器中的数据(例如,社交图200的数据)的数量可能非常大。作为示例而不是作为限制,由加利

福尼亚州门洛帕克(Menlo Park)的脸书公司使用的社交图可以具有10^8量级的节点数量和10^{10}量级的边数量。一般,数据的大集合(例如大数据库)可以被分成多个分区。因为数据库的每个分区的索引小于总数据库的索引,划分可以提高在访问数据库时的性能。因为分区可以分布在大量服务器上,分区还可以提高在访问数据库时的性能和可靠性。通常,可以通过单独地存储数据库的行(或列)来划分数据库。在特定实施例中,可以基于对象类型来划分数据库。数据对象可以存储在多个分区中,每个分区保持单个对象类型的数据对象。在特定实施例中,社交网络系统160可以通过将搜索查询提交到存储与搜索查询的预期结果相同的对象类型的对象的特定分区,而响应于搜索查询来检索搜索结果。尽管本公开描述了以特定方式存储对象,但本公开设想了以任何合适的方式存储对象。

[0266] 在特定实施例中,每个对象可以对应于社交图200的特定节点。连接特定节点和另一节点的边206可以指示在对应于这些节点的对象之间的关系。除了存储对象之外,特定的数据储存器还可以存储与对象相关的社交图信息。可选地,关于特定对象的社交图信息可以存储在与对象不同的数据储存器中。社交网络系统160可以基于新近接收的对象以及与所接收的对象相关联的关系来更新数据储存器的搜索索引。

[0267] 在特定实施例中,每个数据储存器164可以被配置成在相应的数据存储设备1040中存储多个对象类型中的特定对象类型的对象。对象类型可以是例如用户、照片、帖子、评论、消息、事件列表、web界面、应用、位置、用户简档界面、概念简档界面、用户群组、音频文件、视频、出价/优惠券或另一合适类型的对象。尽管本公开描述了特定类型的对象,但本公开设想了任何合适类型的对象。作为示例而不是作为限制,图10中示出的用户垂直P1可以存储用户对象。存储在用户垂直P1中的每个用户对象可以包括在线社交网络的用户的标识符(例如,字符串)、用户名和简档图片。社交网络系统160还可以在用户垂直P1中存储与用户对象相关联的信息,例如语言、位置、教育、联系信息、兴趣、关系状态、朋友/联系人的列表、家庭成员的列表、隐私设置等。作为示例而不是作为限制,图10所示的帖子垂直P2可以存储帖子对象。存储在帖子垂直P2中的每个帖子对象可以包括标识符(发布到社交网络系统160的帖子的文本串)。社交网络系统160还可以在帖子垂直P2中存储与帖子对象相关联的信息,例如时间戳、作者、隐私设置、点赞帖子的用户、点赞的计数、评论、评论的计数、位置等。作为示例而不是作为限制,照片垂直P3可以存储照片对象(或其他媒体类型(例如视频或音频)的对象)。存储在照片垂直P3中的每个照片对象可以包括标识符和照片。社交网络系统160还可以在照片垂直P3中存储与照片对象相关联的信息,例如时间戳、作者、隐私设置、在照片中标记的用户、点赞照片的用户、评论等。在特定实施例中,每个数据储存器还可以被配置为在数据存储设备1040中存储与每个所存储的对象相关联的信息。

[0268] 在特定实施例中,存储在每个垂直164中的对象可以由一个或更多个搜索索引来被编索引。搜索索引可以由包括一个或更多个计算设备(例如,服务器)的相应索引服务器1030来托管。索引服务器1030可以基于由用户或社交网络系统160(或第三方系统)的其他进程提交给社交网络系统160的数据(例如,照片和与照片相关联的信息)来更新搜索索引。索引服务器1030还可以周期性地(例如,每24小时)更新搜索索引。索引服务器1030可以接收包括搜索项的查询,并且从对应于该搜索项的一个或更多个搜索索引中访问并检索搜索结果。在一些实施例中,对应于特定对象类型的垂直可以包括多个物理或逻辑分区,每个分区包括相应的搜索索引。

[0269] 在特定实施例中，社交网络系统160可以从PHP(超文本预处理器)进程1010接收搜索查询。PHP进程1010可以包括由社交网络系统160的一个或更多个服务器162托管的一个或更多个计算进程。搜索查询可以是由用户或社交网络系统160(或第三方系统170)的另一进程提交给PHP进程的文本串或搜索查询。在特定实施例中，聚合器1020可以被配置成从PHP进程1010接收搜索查询，并将搜索查询分发到每个垂直。聚合器可以包括由社交网络系统160的一个或更多个计算设备(例如服务器)托管的一个或更多个计算进程(或程序)。如图10所示，特定实施例可以维持多个垂直164。如早些时候描述的，每个垂直164可以被配置成存储由搜索索引编索引的单一类型的对象。在特定实施例中，聚合器1020可以接收搜索请求。例如，聚合器1020可以从PHP进程1010接收搜索请求。在特定实施例中，搜索请求可以包括文本串。搜索请求可以是由用户经由PHP进程提交的结构化或实质上非结构化的文本串。搜索请求也可以是从社交网络系统的另一进程接收的结构化或者实质上非结构化的文本串。在特定实施例中，聚合器1020可以基于接收到的搜索请求来确定一个或更多个搜索查询。在特定实施例中，每个搜索查询对于它的预期结果可以具有单个对象类型(即，单个结果类型)。在特定实施例中，对于搜索查询的每一个，聚合器1020可以从至少一个垂直164中访问并检索搜索查询结果，其中至少一个垂直164被配置为存储搜索查询的对象类型(即，搜索查询的结果类型)的对象。在特定实施例中，聚合器1020可以聚集相应搜索查询的搜索查询结果。例如，聚合器1020可以向特定的垂直提交搜索查询并访问垂直的索引服务器1030，使索引服务器1030返回搜索查询的结果。

[0270] 可以在2012年7月27日提交的第13/560212号美国专利申请、2012年7月27日提交的第13/560901号美国专利申请、2012年12月21日提交的第13/723861号美国专利申请和2013年4月25日提交的第13/870113号美国专利申请中找到关于索引和搜索查询的更多信息，这些专利申请的中每一个通过引用被并入。

[0271] 向量空间和嵌入

[0272] 图11示出了向量空间1100的示例视图。在特定实施例中，对象或n-gram可以在d维向量空间中表示，其中d表示维度的任何合适数量。尽管向量空间1100被图示为三维空间，但是这仅仅是为了说明的目的，因为向量空间1100可以是任何合适的维度。在特定实施例中，n-gram可以在向量空间1100中表示为称为项嵌入(term embedding)的向量。每个向量可以包括对应于在向量空间1100中的特定点(即，向量的终点)的坐标。作为示例而非限制，如图11所示，向量1110、1120和1130可以被表示为向量空间1100中的点。n-gram可以被映射到相应的向量表示。作为示例而非限制，通过应用由字典定义的函数$\vec{\pi}$，n-gramt_1和t_2可以分别映射到向量空间1100中的$\overrightarrow{v_1}$和$\overrightarrow{v_2}$，使得$\overrightarrow{v_1}=\vec{\pi}(t_1)$和$\overrightarrow{v_2}=\vec{\pi}(t_2)$。作为另一个示例而非限制，可以利用被训练为将文本映射到向量表示的字典，或者这种字典本身可以通过训练来生成。作为另一个示例而非限制，可以使用诸如Word2vec的模型来将n-gram映射到向量空间1100中的向量表示。在特定实施例中，可以通过使用机器学习模型(例如，神经网络)将n-gram映射到向量空间1100中的向量表示。可能已经使用训练数据序列(例如，各自包括n-gram的对象的语料库(corpus))对机器学习模型进行了训练。

[0273] 在特定实施例中，对象可以在向量空间1100中被表示为称为特征向量或对象嵌入的向量。作为示例而非限制，可以通过应用函数$\vec{\pi}$将对象e_1和e_2分别映射到向量空间1100中

的向量$\overrightarrow{v_1}$和$\overrightarrow{v_2}$，使得$\overrightarrow{v_1}=\vec{\pi}(e_1)$和$\overrightarrow{v_2}=\vec{\pi}(e_2)$。在特定实施例中，可以基于对象的一个或更多个特性、属性或特征、对象与其他对象的关系或与对象相关联的任何其他合适的信息，将对象映射到向量。作为示例而非限制，函数$\vec{\pi}$可以通过特征提取将对象映射到向量，特征提取可以从测量数据的初始集合开始，并建立导出值(例如，特征)。作为示例而非限制，可以通过使用算法来检测或隔离对象的各种期望部分或形状而将包括视频或图像的对象映射到向量。用于计算向量的特征可以基于从边缘检测、拐角检测、斑点检测(blob detection)、脊检测、尺度不变特征变换、边缘方向、变化的强度、自相关、运动检测、光流、阈值、斑点提取、模板匹配、霍夫变换(例如，线、圆、椭圆、任意形状)获得的信息，或者任何其他合适的信息。作为另一示例而非限制，可以基于诸如频谱斜率、音调系数(tonality coefficient)、音频频谱质心、音频频谱包络、Mel频率倒谱或任何其他合适的信息的特征而将包括音频数据的对象映射到向量。在特定实施例中，当对象具有太大而不能被有效处理的数据或者包括冗余数据时，函数$\vec{\pi}$可以使用变换后的简化特征集(例如，特征选择)将对象映射到向量。在特定实施例中，函数$\vec{\pi}$可以基于与对象e相关联的一个或更多个n-gram将对象e映射到向量$\vec{\pi}(e)$。尽管本公开描述了以特定方式在向量空间中表示n-gram或对象，但是本公开设想了以任何合适的方式在向量空间中表示n-gram或对象。

[0274] 在特定实施例中，社交网络系统160可以计算向量空间1100中向量的相似性度量。相似性度量可以是余弦相似性、Minkowski距离、Mahalanobis距离、Jaccard相似性系数或任何合适的相似性度量。作为示例而非限制，$\overrightarrow{v_1}$和$\overrightarrow{v_2}$的相似性度量可以是余弦相似性$\frac{\overrightarrow{v_1}\cdot\overrightarrow{v_2}}{\|\overrightarrow{v_1}\|\|\overrightarrow{v_2}\|}$。作为另一个示例而非限制，$\overrightarrow{v_1}$和$\overrightarrow{v_2}$的相似性度量可以是欧几里德距离$\|\overrightarrow{v_1}-\overrightarrow{v_2}\|$。两个向量的相似性度量可以表示分别对应于两个向量的两个对象或n-gram彼此之间的相似程度，如通过向量空间1100中两个向量之间的距离所测量的。作为示例而非限制，基于各个向量之间的距离，向量1110和向量1120可以对应于彼此比对应于向量1110和向量1130的对象更相似的对象。尽管本公开描述了以特定方式计算向量之间的相似性度量，但是本公开设想了以任何合适的方式计算向量之间的相似性度量。

[0275] 可以在2015年11月23日提交的第14/949436号美国专利申请、2016年10月5日提交的第15/286315号美国专利申请和2016年11月30日提交的第15/365789号美国专利申请中找到关于向量空间、嵌入、特征向量和相似性度量的更多信息，这些专利申请中的每一个通过引用被并入。

[0276] 人工神经网络

[0277] 图12示出了示例人工神经网络(“ANN”)1200。在特定实施例中，ANN可以指包括一个或更多个节点的计算模型。示例ANN 1200可以包括输入层1210，隐藏层1220、1230、1240和输出层1250。ANN 1200的每一层可以包括一个或更多个节点，例如节点1205或节点1215。在特定实施例中，ANN的每个节点可以连接到ANN的另一个节点。作为示例而非限制，输入层1210的每个节点可以连接到隐藏层1220的一个或更多个节点。在特定实施例中，一个或更

多个节点可以是偏置节点(例如,层中不连接到前一层中的任何节点并且不从前一层中的任何节点接收输入的节点)。在特定实施例中,每层中的每个节点可以连接到前一层或后续层的一个或更多个节点。尽管图12描绘了具有特定层数、特定节点数和节点间特定连接的特定ANN,但是本公开设想了具有任何适当层数、任何适当节点数和节点间任何适当连接的任何适当ANN。作为示例而非限制,尽管图12描绘了输入层1210的每个节点和隐藏层1220的每个节点之间的连接,但是输入层1210的一个或更多个节点可以不连接到隐藏层1220的一个或更多个节点。

[0278] 在特定实施例中,ANN可以是前馈ANN(例如,没有循环或回路的ANN,其中节点之间的通信从输入层开始在一个方向上流动,并前进到后续的层)。作为示例而非限制,隐藏层1220的每个节点的输入可以包括输入层1210的一个或更多个节点的输出。作为另一示例而非限制,输出层1250的每个节点的输入可以包括隐藏层1240的一个或更多个节点的输出。在特定实施例中,ANN可以是深度神经网络(例如,包括至少两个隐藏层的神经网络)。在特定实施例中,ANN可以是深度残差网络。深度残差网络可以是前馈ANN,其包括被组织成残差块的隐藏层。第一残差块之后的每个残差块的输入可以是前一残差块的输出和前一残差块的输入的函数。作为示例而非限制,到残差块N的输入可以是F(x)+x,其中F(x)可以是残差块N-1的输出,x可以是到残差块N-1的输入。尽管本公开描述了特定的ANN,但是本公开设想了任何合适的ANN。

[0279] 在特定实施例中,激活函数可以对应于ANN的每个节点。节点的激活函数可以为给定的输入定义节点的输出。在特定实施例中,节点的输入可以包括一组输入。作为示例而非限制,激活函数可以是恒等函数(identity function)、二进制阶跃函数(binary step function)、逻辑函数或任何其他合适的函数。作为另一个示例而非限制,节点k的激活函数可以是S型函数(sigmoid function) $F_k(s_k)=\frac{1}{1+e^{-s_k}}$、双曲线正切函数 $F_k(s_k)=\frac{e^{s_k}-e^{-s_k}}{e^{s_k}+e^{-s_k}}$、整流器(rectifier)$F_k(s_k)=\max(0,s_k)$或任何其他合适的函数$F_k(s_k)$,其中$s_k$可以是节点k的有效输入。在特定实施例中,对应于节点的激活函数的输入可以被加权。每个节点可以使用基于加权输入的相应激活函数来生成输出。在特定实施例中,节点之间的每个连接可以与权重相关联。作为示例而非限制,节点1205和节点1215之间的连接1225可以具有0.4的加权系数,这可以指示0.4乘以节点1205的输出被用作节点1215的输入。作为另一个示例而非限制,节点k的输出y_k可以是$y_k=F_k(s_k)$,其中F_k可以是对应于节点k的激活函数,$S_k=\sum_j(w_{jk}x_j)$可以是对节点k的有效输入,x_j可以是连接到节点k的节点j的输出,并且w_{jk}可以是节点j和节点k之间的加权系数。在特定实施例中,对输入层的节点的输入可以基于表示对象的向量。尽管本公开描述了节点的特定输入和输出,但是本公开设想了节点的任何合适的输入和输出。此外,尽管本公开可以描述节点之间的特定连接和权重,但是本公开设想了节点之间的任何合适的连接和权重。

[0280] 在特定实施例中,可以使用训练数据来训练ANN。作为示例而非限制,训练数据可以包括到ANN 1200的输入和预期输出。作为另一个示例而非限制,训练数据可以包括向量,每个向量表示训练对象和每个训练对象的预期标签。在特定实施例中,训练ANN可以包括通

过优化目标函数来修改与ANN的节点之间的连接相关联的权重。作为示例而非限制，可以使用训练方法(例如，共轭梯度法、梯度下降法、随机梯度下降法)来反推作为表示训练对象的每个向量之间的距离测量的平方和误差(例如，使用最小化平方和误差的成本函数)。在特定实施例中，可以使用丢弃法技术(dropout technique)来训练ANN。作为示例而非限制，在训练时可以暂时省略一个或更多个节点(例如，不接收输入并且不生成输出)。对于每个训练对象，ANN的一个或更多个节点可能有被忽略的某种可能性。针对特定训练对象省略的节点可以不同于针对其他训练对象省略的节点(例如，可以逐个对象地暂时省略的节点)。尽管本公开描述了以特定方式训练ANN，但是本公开设想了以任何合适的方式训练ANN。

[0281] 社交图亲和度(affinity)和系数

[0282] 在特定实施例中，社交网络系统160可以确定各种社交图实体对于彼此的社交图亲和度(在本文可以称为“亲和度”)。亲和度可以表示在与在线社交网络关联的特定对象(诸如用户、概念、内容、动作、广告)、与在线社交网络关联的其他对象、或其任何合适的组合之间的关系强度或感兴趣程度。还可以针对与第三方系统170或其他合适的系统相关联的对象来确定亲和度。可以为每个用户、主题或内容类型建立对社交图实体的总体亲和度。基于对与社交图实体相关联的动作或关系的持续监控，总体亲和度可以改变。尽管本公开描述了以特定方式确定特定亲和度，但是本公开设想了以任何合适的方式确定任何合适的亲和度。

[0283] 在特定实施例中，社交网络系统160可以使用亲和度系数(在本文可以称为“系数”)来度量或量化社交图亲和度。系数可以表示或量化与在线社交网络相关联的特定对象之间的关系强度。系数还可以表示基于用户对特定动作的兴趣来估量用户将执行该动作的预测概率的概率或函数。以这种方式，可以基于用户的先前动作来预测用户的未来动作，其中可以至少部分地基于用户的动作的历史来计算系数。系数可以用于预测可以位于在线社交网络内部或外部的任何数量的动作。作为示例而不是作为限制，这些动作可以包括各种类型的通信，诸如发送消息、发布内容或对内容进行评论；各种类型的观察动作，诸如访问或查看简档界面、媒体或其他合适的内容；关于两个或更多个社交图实体的各种类型的一致性信息，诸如在同一组中、在同一照片中被标记、在同一位置处签到、或参加同一活动(event)；或其他合适的动作。尽管本公开描述了以特定方式估量亲和度，但是本公开设想了以任何合适的方式估量亲和度。

[0284] 在特定实施例中，社交网络系统160可以使用各种因素来计算系数。这些因素可以包括例如，用户动作、对象之间的关系类型、位置信息、其他合适的因素或其任意组合。在特定实施例中，当计算系数时，不同的因素可以被不同地加权。每个因素的权重可以是静态的，或者可以根据例如用户、关系类型、动作类型、用户的位置等来改变权重。可以根据因素的权重来组合这些因素的等级(rating)，以确定用户的总系数。作为示例而不是作为限制，特定用户动作可以被分配等级和权重，而与特定用户动作相关联的关系被分配等级和相关权重(例如，因此权重总计为100％)。为了计算用户对特定对象的系数，分配给用户动作的等级可以包括例如，总系数的60％，而用户和该对象之间的关系可以包括总系数的40％。在特定实施例中，当确定用于计算系数的各种因素的权重时，社交网络系统160可以考虑各种变量，例如自信息被访问以来的时间、衰减因子、访问的频率、与信息的关系或与信息被访问了的对象的关系、与连接到对象的社交图实体的关系、用户动作的短期或长期平均值、用

户反馈、其他合适的变量、或它们的任意组合。作为示例而不是作为限制，系数可以包括衰减因子，该衰减因子导致由特定动作提供的信号强度随时间衰减，使得在计算系数时更近的(more recent)动作更相关。等级和权重可以基于系数所基于的动作的持续跟踪而被持续更新。可以采用任何类型的过程或算法来对每个因素的等级和分配给这些因素的权重进行分配、组合、平均等。在特定实施例中，社交网络系统160可以使用根据历史动作和过去的用户响应训练的机器学习算法，或者通过将用户暴露于各种选项并测量响应而从用户获得的数据，来确定系数。尽管本公开描述了以特定方式计算系数，但是本公开设想了以任何合适的方式计算系数。

[0285] 在特定实施例中，社交网络系统160可以基于用户的动作来计算系数。社交网络系统160可以监控在线社交网络上、第三方系统170上、其他合适的系统上、或它们的任意组合上的这样的动作。可以跟踪或监控任何合适类型的用户动作。典型的用户动作包括查看简档界面、创建或发布内容、与内容进行互动、在图像中标记或被标记在图像中、加入群组、列出并确认事件的出席情况、在位置处签到、点赞特定界面、创建界面以及执行其他便于社交动作的任务。在特定实施例中，社交网络系统160可以基于用户对特定类型的内容的动作来计算系数。内容可以与在线社交网络、第三方系统170或另一合适的系统相关联。内容可以包括用户、简档界面、帖子、新闻动态(news stories)、标题、即时消息、聊天室会话、电子邮件、广告、图片、视频、音乐、其他合适的对象或其任意组合。社交网络系统160可以分析用户的动作，以确定这些动作中的一个或更多个是否指示对主题、内容、其他用户等的亲和度。作为示例而不是作为限制，如果用户频繁发布与“咖啡”或其变体相关的内容，则社交网络系统160可以确定用户相对于概念“咖啡”具有高的系数。特定动作或动作类型可以被分配比其他动作更高的权重和/或等级，这可能影响计算的总系数。作为示例而不是作为限制，如果第一用户向第二用户发送电子邮件，则该动作的权重或等级可能高于第一用户简单地查看第二用户的用户简档界面的情况。

[0286] 在特定实施例中，社交网络系统160可以基于特定对象之间的关系的类型来计算系数。参考社交图900，当计算系数时，社交网络系统160可以分析连接特定用户节点902和概念节点904的边906的数量和/或类型。作为示例而不是作为限制，通过配偶型边(表示两个用户结婚了)连接的用户节点902可以被分配比通过朋友型边连接的用户节点902更高的系数。换句话说，根据分配给特定用户的动作和关系的权重，可以确定对关于用户配偶的内容的总体亲和度高于对关于用户朋友的内容的总体亲和度。在特定实施例中，用户与另一对象的关系可以影响关于计算该对象的系数的用户动作的权重和/或等级。作为示例而不是作为限制，如果用户在第一张照片中被标记，但仅仅赞第二张照片，则社交网络系统160可以确定用户关于第一张照片具有比第二张照片更高的系数，因为具有与内容的标记型(tagged-in-type)关系可以被分配比具有与内容的点赞型(like-type)关系更高的权重和/或等级。在特定实施例中，社交网络系统160可以基于一个或更多个第二用户与特定对象所具有的关系来计算第一用户的系数。换句话说，其他用户与对象的关连和系数可能影响第一用户关于该对象的系数。作为示例而不是作为限制，如果第一用户关连到一个或更多个第二用户或者对于一个或更多个第二用户具有高的系数，并且那些第二用户关连到特定对象或者对于特定对象具有高的系数，则社交网络系统160可以确定第一用户对于该特定对象同样应当具有相对高的系数。在特定实施例中，系数可以基于特定对象之间的分离

度。较低的系数可以表示第一用户将分享对在社交图900中间接地关连到第一用户的用户的内容对象有兴趣的可能性降低。作为示例而不是作为限制，在社交图900中更靠近(即，分离度更少)的社交图实体可以比在社交图900中更远离的实体具有更高的系数。

[0287] 在特定实施例中，社交网络系统160可以基于位置信息来计算系数。在地理上彼此更靠近的对象可以被认为比更远的对象彼此更相关或更感兴趣。在特定实施例中，用户对特定对象的系数可以基于对象的位置与和用户相关联的当前位置(或用户的客户端系统130的位置)的接近度。第一用户可能对更靠近第一用户的其他用户或概念更感兴趣。作为示例而不是作为限制，如果用户距机场一英里且距加油站两英里，则社交网络系统160可以基于机场与用户的接近度来确定用户对机场具有比加油站更高的系数。

[0288] 在特定实施例中，社交网络系统160可以基于系数信息来执行关于用户的特定动作。系数可以用于基于用户对特定动作的兴趣来预测用户是否将执行该动作。当生成或向用户呈现任何类型的对象(例如广告、搜索结果、新闻动态、媒体、消息、通知或其他合适的对象)时，可以使用系数。该系数也可以被用来适当地对这些对象进行排列(rank)和排序(order)。以这种方式，社交网络系统160可以提供与用户的兴趣和当前环境相关的信息，增加了他们将找到这样的感兴趣的信息的可能性。在特定实施例中，社交网络系统160可以基于系数信息来生成内容。可以基于特定于用户的系数来提供或选择内容对象。作为示例而不是作为限制，系数可以用于为用户生成媒体，其中可以向用户呈现用户关于媒体对象具有高的总系数的媒体。作为另一个示例而不是作为限制，系数可以用于为用户生成广告，其中可以向用户呈现用户对于被登广告的对象具有高的总系数的广告。在特定实施例中，社交网络系统160可以基于系数信息生成搜索结果。可以基于与查询用户的搜索结果相关联的系数来对特定用户的搜索结果进行评分或排序。作为示例而不是作为限制，对应于具有较高系数的对象的搜索结果在搜索结果界面上的排名可以高于对应于具有较低系数的对象的结果。

[0289] 在特定实施例中，社交网络系统160可以响应于来自特定系统或过程的系数的请求来计算系数。为了预测用户在给定情况下可能采取动作(或者可能是动作的对象)的可能性，任何过程都可以请求关于用户的计算出的系数。请求还可以包括用于计算系数的各种因素使用的一组权重。该请求可以来自在线社交网络上运行的进程(process)、来自第三方系统170(例如，经由API或其他通信渠道)、或者来自另一个合适的系统。响应于该请求，社交网络系统160可以计算系数(或者如果先前已经计算并存储了系数信息，则访问系数信息)。在特定实施例中，社交网络系统160可以度量关于特定过程的亲和度。不同的过程(在线社交网络的内部和外部)可以请求关于特定对象或对象集的系数。社交网络系统160可以提供与请求了亲和度的度量的特定过程相关的亲和度的度量。以这种方式，每个过程接收针对不同上下文调整的亲和度度量，其中该过程将使用亲和度度量。

[0290] 结合社交图亲和度和亲和度系数，特定实施例可以利用在2006年8月11日提交的第11/503093号美国专利申请、2010年12月22日提交的第12/977027号美国专利申请、2010年12月23日提交的第12/978265号美国专利申请和2012年10月01日提交的第13/632869号美国专利申请(其中的每一个都通过引用被并入)中公开的一个或更多个系统、部件、元件、功能、方法、操作或步骤。

[0291] 系统和方法

[0292] 图13示出了示例计算机系统1300。在特定实施例中，一个或更多个计算机系统1300执行本文描述或示出的一个或更多个方法的一个或更多个步骤。在特定实施例中，一个或更多个计算机系统1300提供本文描述或示出的功能。在特定实施例中，在一个或更多个计算机系统1300上运行的软件执行本文描述或示出的一个或更多个方法的一个或更多个步骤，或者提供本文描述或示出的功能。特定实施例包括一个或更多个计算机系统1300的一个或更多个部分。在本文，在适当的情况下，对计算机系统的引用可以包括计算设备，反之亦然。此外，在适当的情况下，对计算机系统的引用可以包括一个或更多个计算机系统。

[0293] 本公开设想了任何合适数量的计算机系统1300。本公开设想了计算机系统1300采取任何合适的物理形式。作为示例而不是作为限制，计算机系统1300可以是嵌入式计算机系统、片上系统(SOC)、单板计算机系统(SBC)(例如，模块上计算机(COM)或模块上系统(SOM))、台式计算机系统、膝上型或笔记本计算机系统、交互式信息亭、大型机、计算机系统网状网、移动电话、个人数字助理(PDA)、服务器、平板计算机系统、或者这些系统的两个或更多个的组合。在适当的情况下，计算机系统1300可以包括一个或更多个计算机系统1300；是整体式的或分布式的；跨越多个位置；跨越多台机器；跨越多个数据中心；或者驻留在云中，云可以包括在一个或更多个网络中的一个或更多个云组件。在适当的情况下，一个或更多个计算机系统1300可以在没有实质性空间或时间限制的情况下执行本文描述或示出的一个或更多个方法的一个或更多个步骤。作为示例而不是作为限制，一个或更多个计算机系统1300可以实时地或以批处理模式来执行本文描述或示出的一个或更多个方法的一个或更多个步骤。在适当的情况下，一个或更多个计算机系统1300可以在不同的时间或在不同的位置处执行本文描述或示出的一个或更多个方法的一个或更多个步骤。

[0294] 在特定实施例中，计算机系统1300包括处理器1302、存储器1304、存储装置1306、输入/输出(I/O)接口1308、通信接口1310和总线1312。尽管本公开描述并示出了具有在特定布置中的特定数量的特定组件的特定计算机系统，但是本公开设想了具有在任何合适布置中的任何合适数量的任何合适组件的任何合适的计算机系统。

[0295] 在特定实施例中，处理器1302包括用于执行指令(例如构成计算机程序的那些指令)的硬件。作为示例而不是作为限制，为了执行指令，处理器1302可以从内部寄存器、内部高速缓存、存储器1304或存储装置1306中检索(或取回)指令；将他们解码并执行它们；以及然后将一个或更多个结果写到内部寄存器、内部高速缓存、存储器1304或存储装置1306。在特定实施例中，处理器1302可以包括用于数据、指令或地址的一个或更多个内部高速缓存。在适当的情况下，本公开设想了处理器1302包括任何合适数量的任何合适的内部高速缓存。作为示例而不是作为限制，处理器1302可以包括一个或更多个指令高速缓存、一个或更多个数据高速缓存、以及一个或更多个转译后备缓冲器(TLB)。在指令高速缓存中的指令可以是在存储器1304或存储装置1306中的指令的副本，并且指令高速缓存可以加速处理器1302对那些指令的检索。在数据高速缓存中的数据可以是在存储器1304或存储装置1306中的数据的副本，用于使在处理器1302处执行的指令进行操作；在处理器1302处执行的先前指令的结果，用于由在处理器1302处执行的后续指令访问或者用于写到存储器1304或存储装置1306；或其他合适的数据。数据高速缓存可以加速由处理器1302进行的读或写操作。TLB可以加速关于处理器1302的虚拟地址转译。在特定实施例中，处理器1302可以包括用于

数据、指令或地址的一个或更多个内部寄存器。在适当的情况下，本公开设想了处理器1302包括任何合适数量的任何合适的内部寄存器。在适当的情况下，处理器1302可以包括一个或更多个算术逻辑单元(ALU)；是多核处理器；或者包括一个或更多个处理器1302。尽管本公开描述并示出了特定的处理器，但是本公开设想了任何合适的处理器。

[0296] 在特定实施例中，存储器1304包括用于存储用于使处理器1302执行的指令或用于使处理器1302操作的数据的主存储器。作为示例而不是作为限制，计算机系统1300可以将指令从存储装置1306或另一个源(例如，另一个计算机系统1300)加载到存储器1304。处理器1302然后可以将指令从存储器1304加载到内部寄存器或内部高速缓存。为了执行指令，处理器1302可以从内部寄存器或内部高速缓存中检索指令并将它们解码。在指令的执行期间或之后，处理器1302可以将一个或更多个结果(其可以是中间结果或最终结果)写到内部寄存器或内部高速缓存。处理器1302然后可以将这些结果中的一个或更多个写到存储器1304。在特定实施例中，处理器1302仅执行在一个或更多个内部寄存器或内部高速缓存中或在存储器1304(而不是存储装置1306或其他地方)中的指令，并且仅对在一个或更多个内部寄存器或内部高速缓存中或在存储器1304(而不是存储装置1306或其他地方)中的数据进行操作。一个或更多个存储器总线(其可以各自包括地址总线和数据总线)可以将处理器1302耦合到存储器1304。如下所述，总线1312可以包括一个或更多个存储器总线。在特定实施例中，一个或更多个存储器管理单元(MMU)驻留在处理器1302和存储器1304之间，并且便于由处理器1302请求的对存储器1304的访问。在特定实施例中，存储器1304包括随机存取存储器(RAM)。在适当的情况下，该RAM可以是易失性存储器。在适当的情况下，该RAM可以是动态RAM(DRAM)或静态RAM(SRAM)。此外，在适当的情况下，该RAM可以是单端口RAM或多端口RAM。本公开设想了任何合适的RAM。在适当的情况下，存储器1304可以包括一个或更多个存储器1304。尽管本公开描述并示出了特定的存储器，但是本公开设想了任何合适的存储器。

[0297] 在特定实施例中，存储装置1306包括用于数据或指令的大容量存储装置。作为示例而不是作为限制，存储装置1306可以包括硬盘驱动器(HDD)、软盘驱动器、闪存、光盘、磁光盘、磁带或通用串行总线(USB)驱动器、或这些中的两个或更多个的组合。在适当的情况下，存储装置1306可以包括可移动或不可移动(或固定)介质。在适当的情况下，存储装置1306可以在计算机系统1300的内部或外部。在特定实施例中，存储装置1306是非易失性固态存储器。在特定实施例中，存储装置1306包括只读存储器(ROM)。在适当的情况下，该ROM可以是掩模编程ROM、可编程ROM(PROM)、可擦除PROM(EPROM)、电可擦除PROM(EEPROM)、电可变ROM(EAROM)、或闪存、或这些中的两个或更多个的组合。本公开设想了采用任何合适的物理形式的大容量存储装置1306。在适当的情况下，存储装置1306可以包括便于在处理器1302和存储装置1306之间的通信的一个或更多个存储装置控制单元。在适当的情况下，存储装置1306可以包括一个或更多个存储装置1306。尽管本公开描述并示出了特定的存储装置，但是本公开设想了任何合适的存储装置。

[0298] 在特定实施例中，I/O接口1308包括为在计算机系统1300和一个或更多个I/O设备之间的通信提供一个或更多个接口的硬件、软件或两者。在适当的情况下，计算机系统1300可以包括这些I/O设备中的一个或更多个。这些I/O设备中的一个或更多个可以实现在人和计算机系统1300之间的通信。作为示例而不是作为限制，I/O设备可以包括键盘、小键盘、麦克风、监视器、鼠标、打印机、扫描仪、扬声器、静态摄像机、触笔、平板计算机、触摸屏、跟踪

球、视频摄像机、另一个合适的I/O设备、或这些中的两个或更多个的组合。I/O设备可以包括一个或更多个传感器。本公开设想了任何合适的I/O设备以及用于它们的任何合适的I/O接口1308。在适当的情况下，I/O接口1308可以包括使处理器1302能够驱动这些I/O设备中的一个或更多个的一个或更多个设备或软件驱动器。在适当的情况下，I/O接口1308可以包括一个或更多个I/O接口1308。尽管本公开描述并示出了特定的I/O接口，但是本公开设想了任何合适的I/O接口。

[0299] 在特定实施例中，通信接口1310包括提供用于在计算机系统1300和一个或更多个其他计算机系统1300或一个或更多个网络之间的通信(例如，基于包的通信)的一个或更多个接口的硬件、软件或两者。作为示例而不是作为限制，通信接口1310可以包括用于与以太网或其他基于有线的网络进行通信的网络接口控制器(NIC)或网络适配器，或用于与无线网络(例如WI-FI网络)进行通信的无线NIC(WNIC)或无线适配器。本公开设想了任何合适的网络和用于它的任何合适的通信接口1310。作为示例而不是作为限制，计算机系统1300可以与自组织网络、个域网(PAN)、局域网(LAN)、广域网(WAN)、城域网(MAN)或互联网的一个或更多个部分、或这些中的两个或更多个的组合进行通信。这些网络中的一个或更多个的一个或更多个部分可以是有线的或无线的。作为示例，计算机系统1300可以与无线PAN(WPAN)(例如，蓝牙WPAN)、WI-FI网络、WI-MAX网络、蜂窝电话网络(例如，全球移动通信系统(GSM)网络)、或其他合适的无线网络、或这些中的两个或更多个的组合进行通信。在适当的情况下，计算机系统1300可以包括用于这些网络中的任一个的任何合适的通信接口1310。在适当的情况下，通信接口1310可以包括一个或更多个通信接口1310。尽管本公开描述并示出了特定的通信接口，但是本公开设想了任何合适的通信接口。

[0300] 在特定实施例中，总线1312包括将计算机系统1300的组件耦合到彼此的硬件、软件或两者。作为示例而不是作为限制，总线1312可以包括加速图形端口(AGP)或其他图形总线、扩展工业标准体系结构(EISA)总线、前端总线(FSB)、HYPERTRANSPORT(HT)互连、工业标准体系结构(ISA)总线、INFINIBAND互连、低引脚数(LPC)总线、存储器总线，微通道体系结构(MCA)总线、外围部件互连(PCI)总线、PCI-Express(扩展)(PCIe)总线、串行高级技术附件(SATA)总线、视频电子标准协会本地(VLB)总线、或任何其他合适的总线、或这些中的两个或更多个的组合。在适当的情况下，总线1312可以包括一个或更多个总线1312。尽管本公开描述并示出了特定总线，但是本公开设想了任何合适的总线或互连。

[0301] 在本文，在适当的情况下，一个或更多个计算机可读非暂时性存储介质可以包括一个或更多个基于半导体的或其他的集成电路(IC)(例如，现场可编程门阵列(FPGA)或专用IC(ASIC))、硬盘驱动器(HDD)、混合硬盘驱动器(HHD)、光盘、光盘驱动器(ODD)、磁光盘、磁光盘驱动器、软盘、软盘驱动器(FDD)、磁带、固态驱动器(SSD)、RAM驱动器、安全数字(SECURE DIGITAL)卡或驱动器、任何其他合适的计算机可读非暂时性存储介质、或这些中的两个或更多个的任何合适组合。在适当的情况下，计算机可读非暂时性存储介质可以是易失性的、非易失性的或者易失性和非易失性的组合。

[0302] 其他

[0303] 本文中，除非另有明确指示或通过上下文另有指示，否则“或”是包括一切的而非排他性的。因此在本文，除非另有明确指示或通过上下文另有指示，否则“A或B”意指“A、B或两者”。此外，除非另有明确指示或通过上下文另有指示，否则“和”既是联合的又是各自的。

因此在本文，除非另有明确指示或通过上下文另有指示，否则“A和B”意指“A和B，联合地或各自地”。

[0304] 本公开的范围包括本领域中的普通技术人员将理解的对本文描述或示出的示例实施例的所有改变、替换、变化、变更和修改。本公开的范围不限于本文描述或示出的示例实施例。此外，尽管本公开将本文的相应实施例描述并示出为包括特定的组件、元件、特征、功能、操作或步骤，但是这些实施例中的任何一个可以包括本领域中的普通技术人员将理解的在本文任何地方描述或示出的任何组件、元件、特征、功能、操作或步骤的任何组合或置换。此外，在所附权利要求中对适合于、被布置成、能够、被配置成、实现来、可操作来、或操作来执行特定功能的装置或系统或装置或系统的组件的引用包括该装置、系统、组件，无论它或那个特定功能是否被激活、开启或解锁，只要该装置、系统或组件是这样被调整、被布置、使能够、被配置、被实现、可操作的、或操作的。此外，尽管本公开将特定实施例描述或示为提供特定优点，但是特定实施例可以提供这些优点中的一些、全部或不提供这些优点。

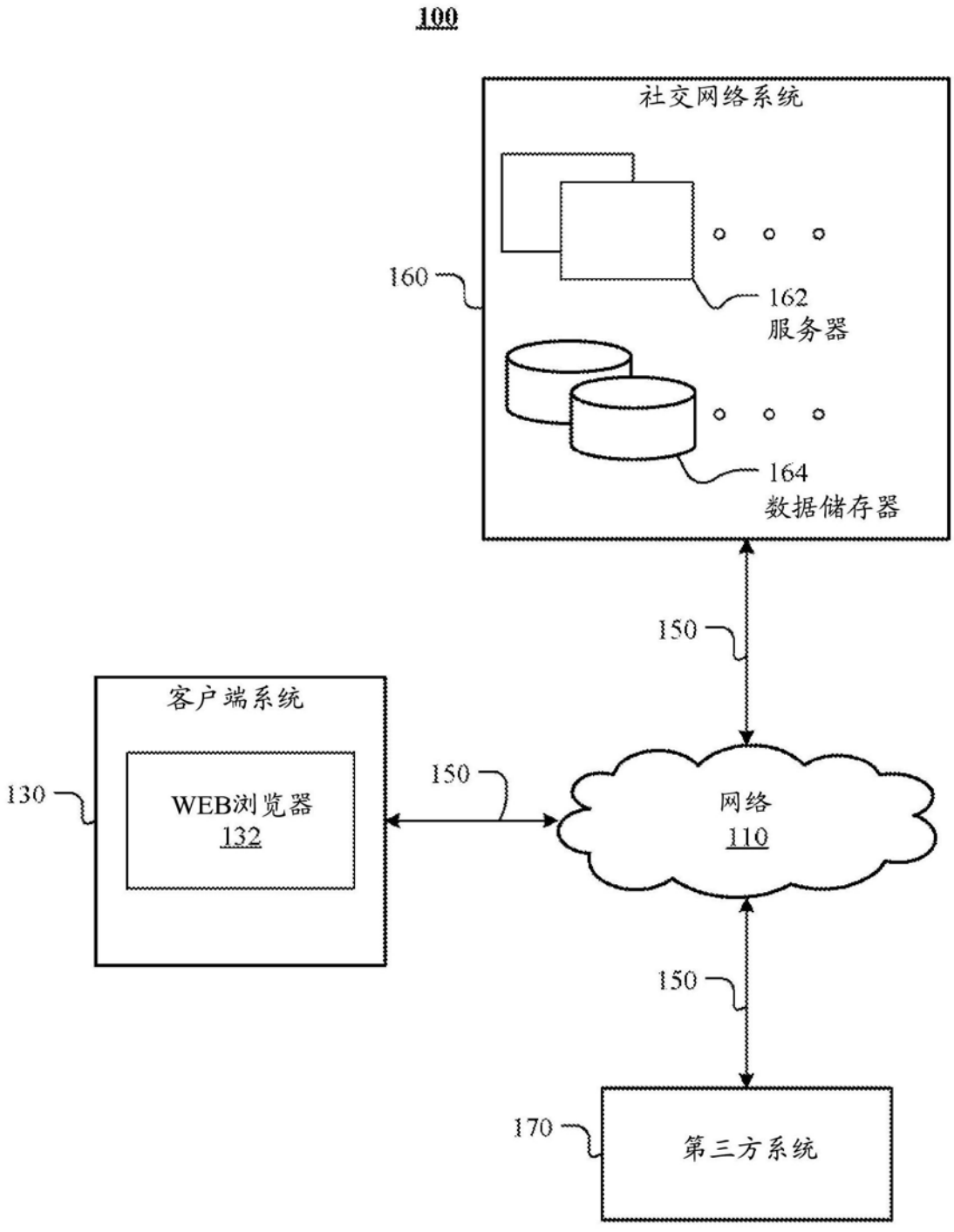

100
社交网络系统
160
162
服务器
164
数据储存器
150
客户端系统
130
WEB浏览器
132
150
网络
110
150
170
第三方系统

图1

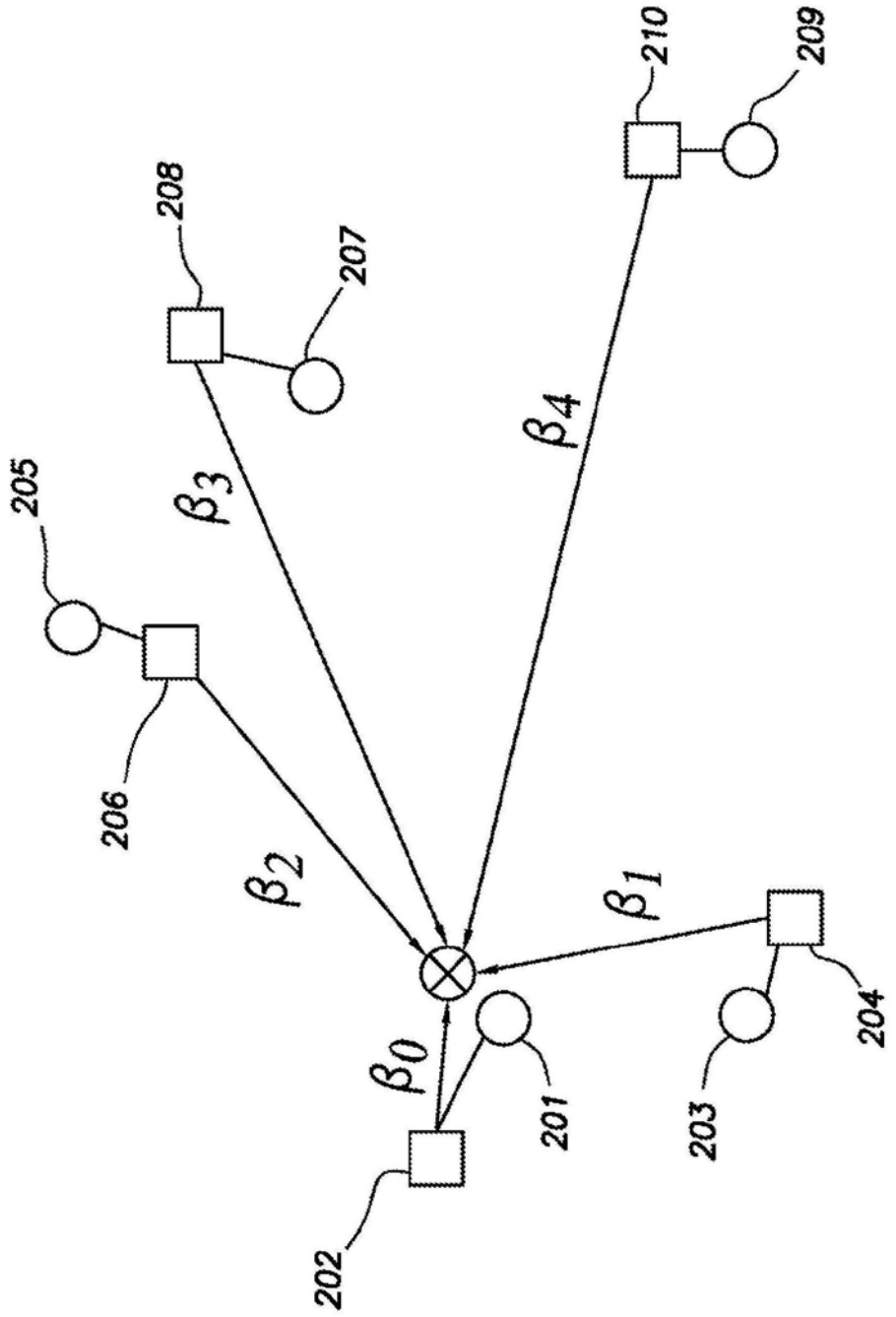


图2

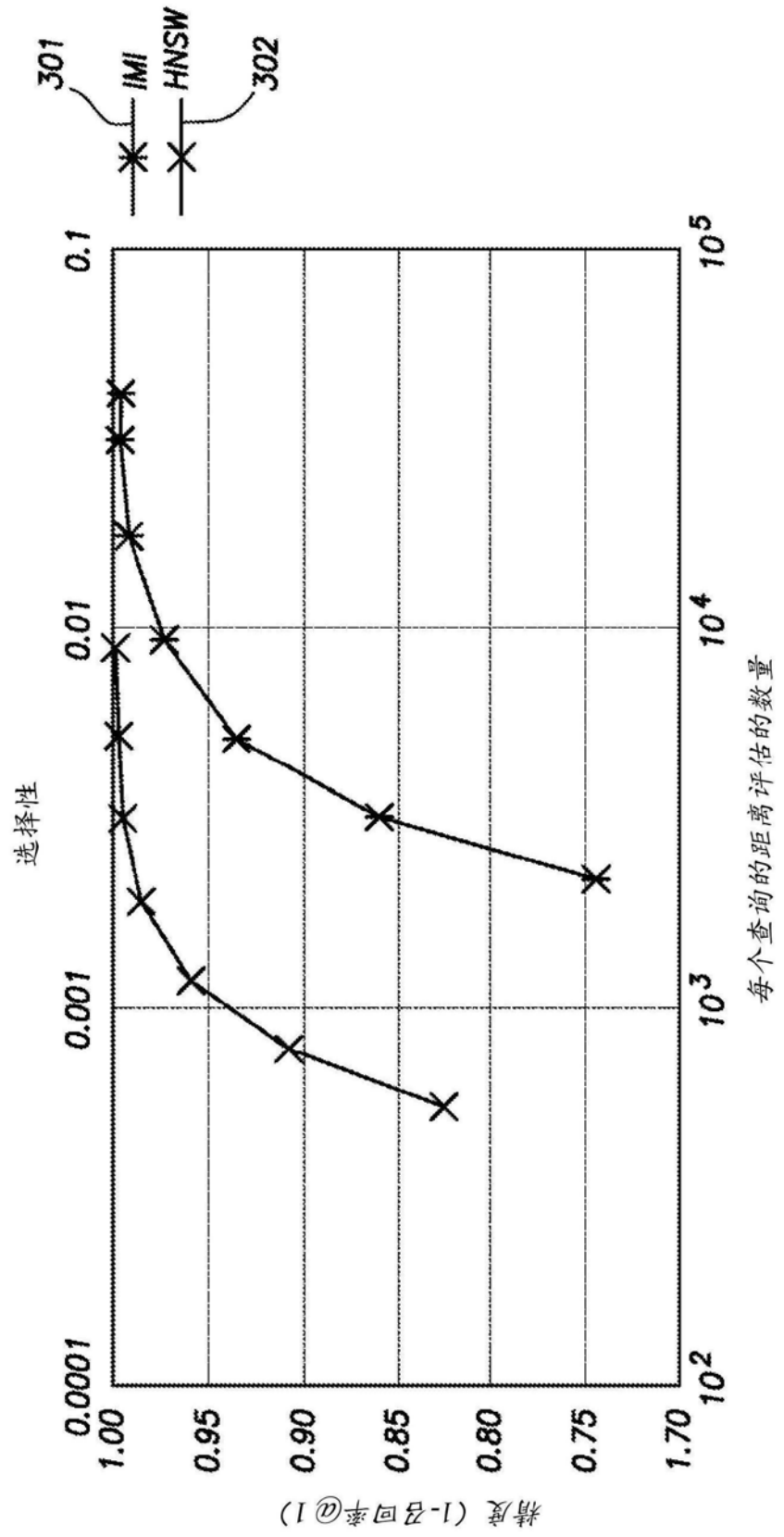

301
IMI
HNSW
302
0.1
0.01
0.001
0.0001
选择性
1.00
0.95
0.90
0.85
0.80
0.75
1.70
精度 (1-召回率@1)
10^2
10^3
10^4
10^5
每个查询的距离评估的数量

图3

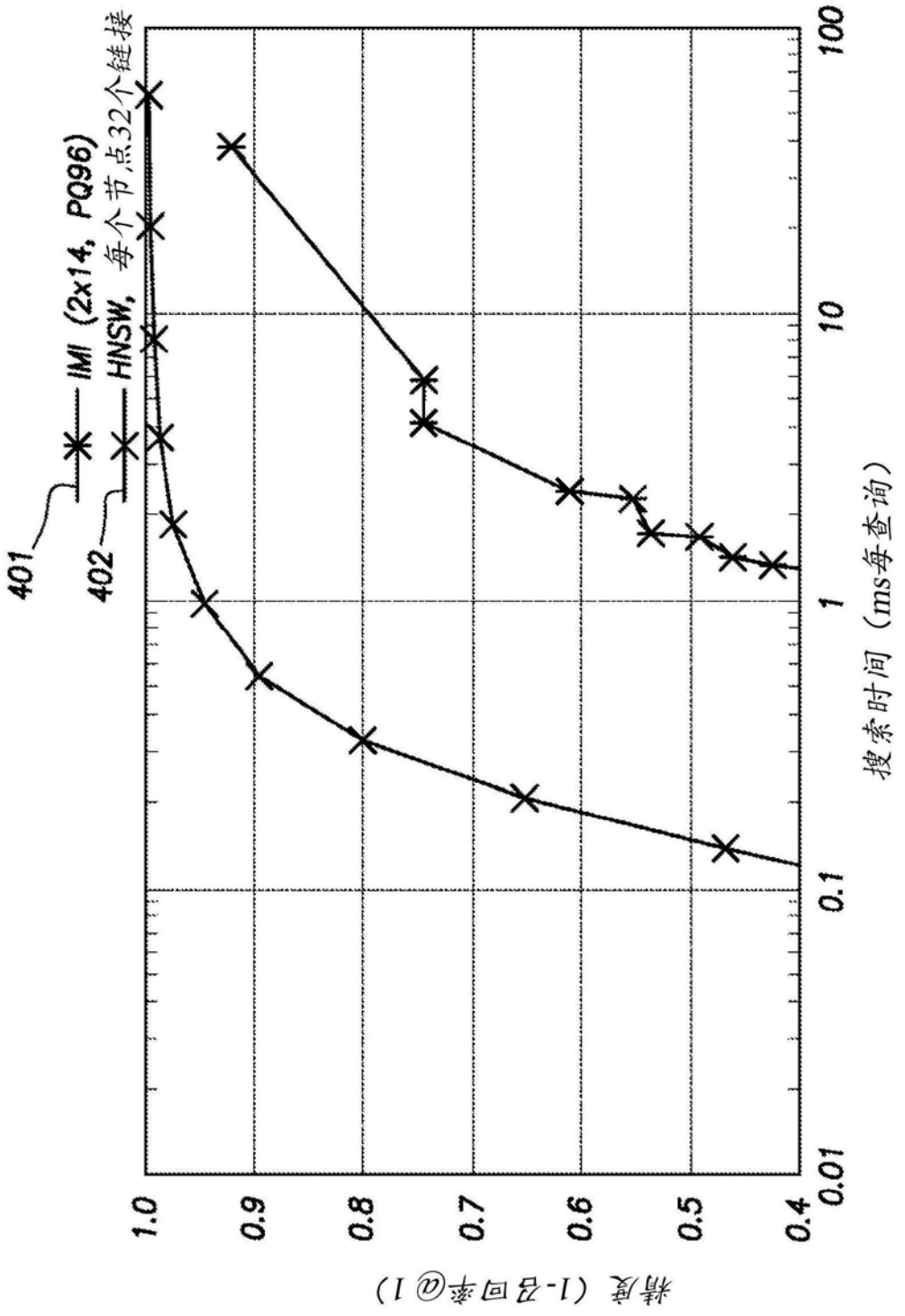

IMI (2x14, PQ96)
HNSW, 每个节点32个链接
401
402
精度（1-召回率@1）
1.0
0.9
0.8
0.7
0.6
0.5
0.4
搜索时间（ms每查询）
0.01
0.1
1
10
100

图4

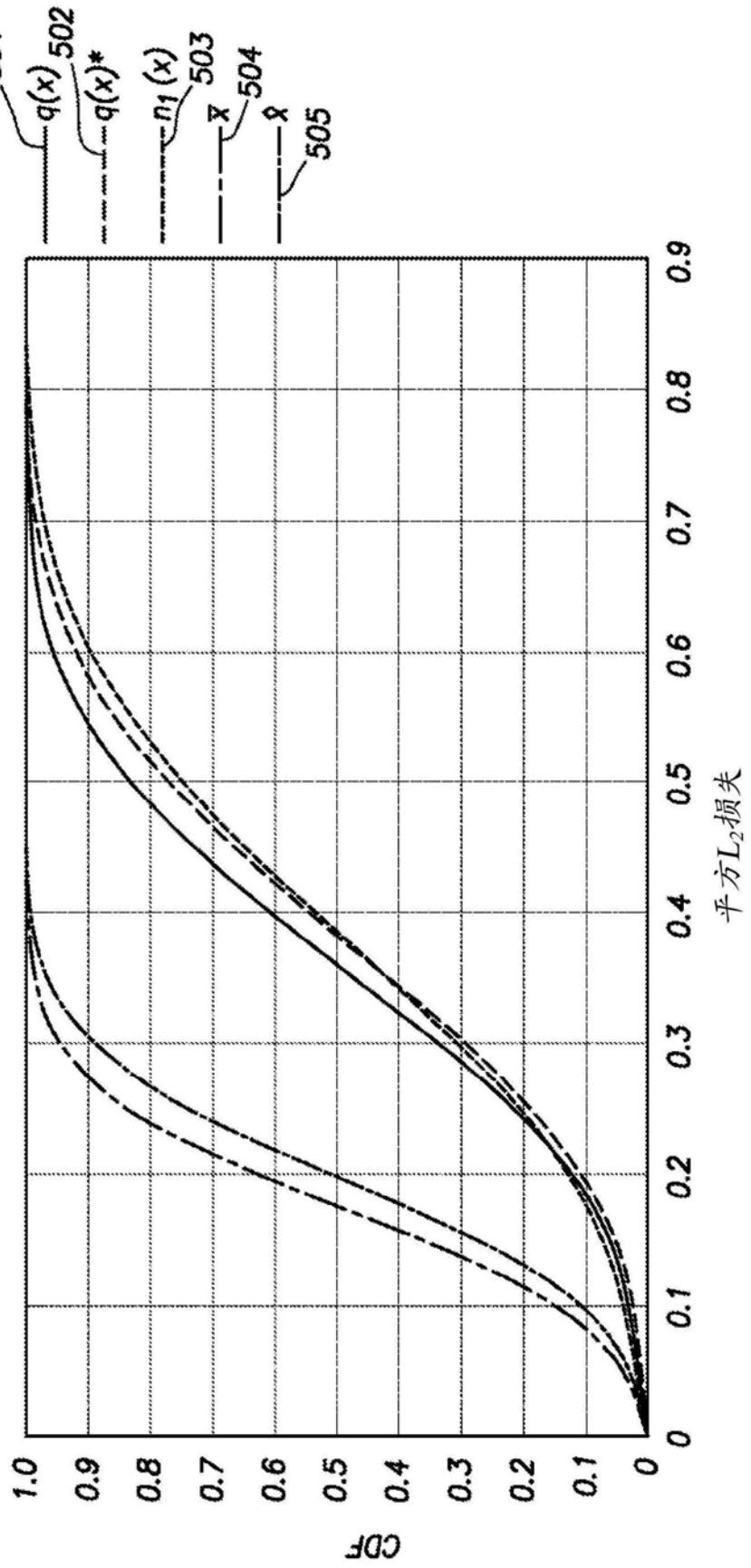

501
q(x)
502
q(x)*
n1(x)
503
504
505
CDF
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
平方L2损失

图5

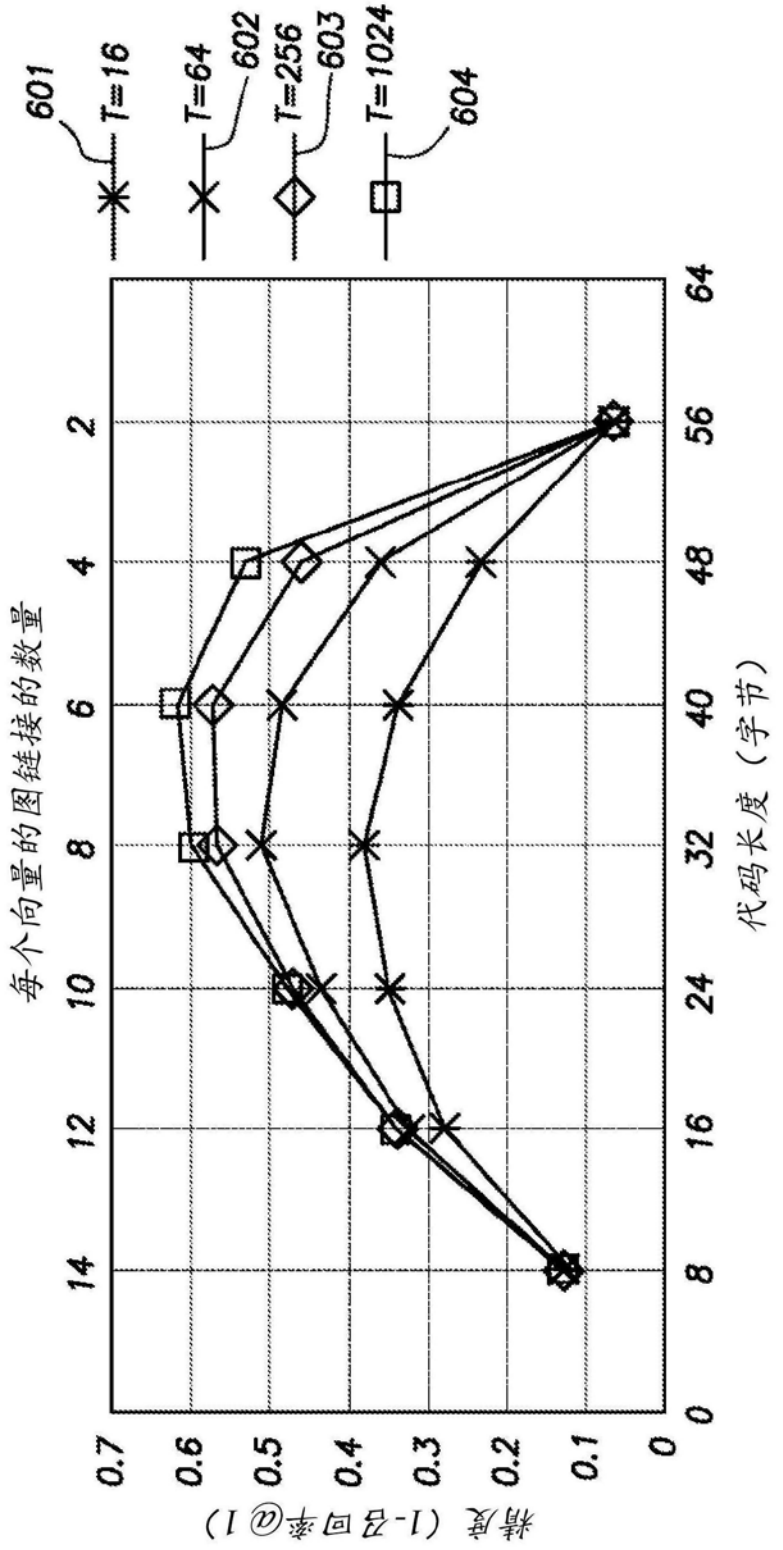

601
T=16
T=64
602
T=256
603
T=1024
604
每个向量的图链接的数量
代码长度（字节）
精度（1-召回率@1）
0
2
4
6
8
10
12
14
0
8
16
24
32
40
48
56
64
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

图6

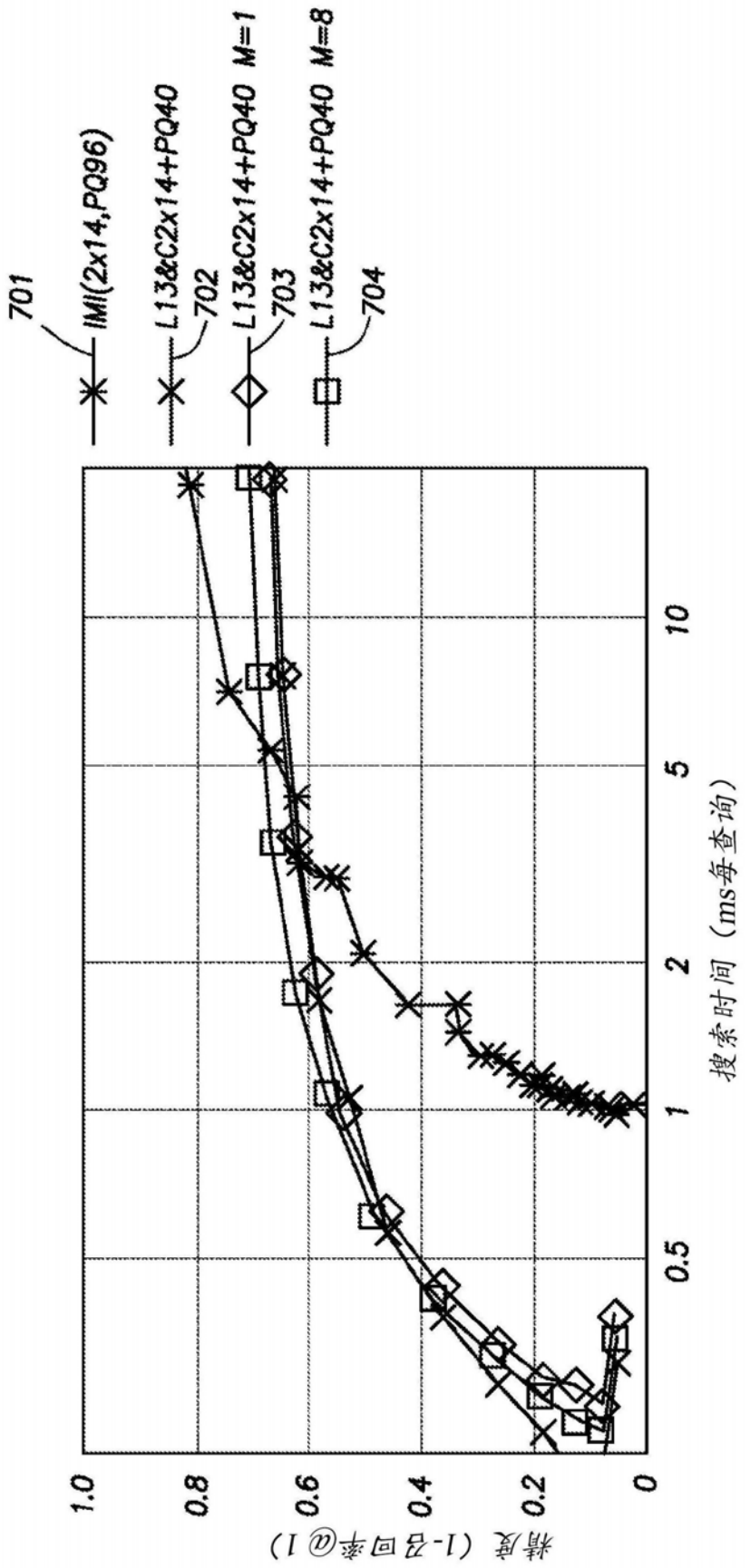

701
IMI(2x14,PQ96)
702
L13&C2x14+PQ40
703
L13&C2x14+PQ40 M=1
704
L13&C2x14+PQ40 M=8
精度（1-召回率@1）
1.0
0.8
0.6
0.4
0.2
0
0.5
1
2
5
10
搜索时间（ms每查询）

图7A

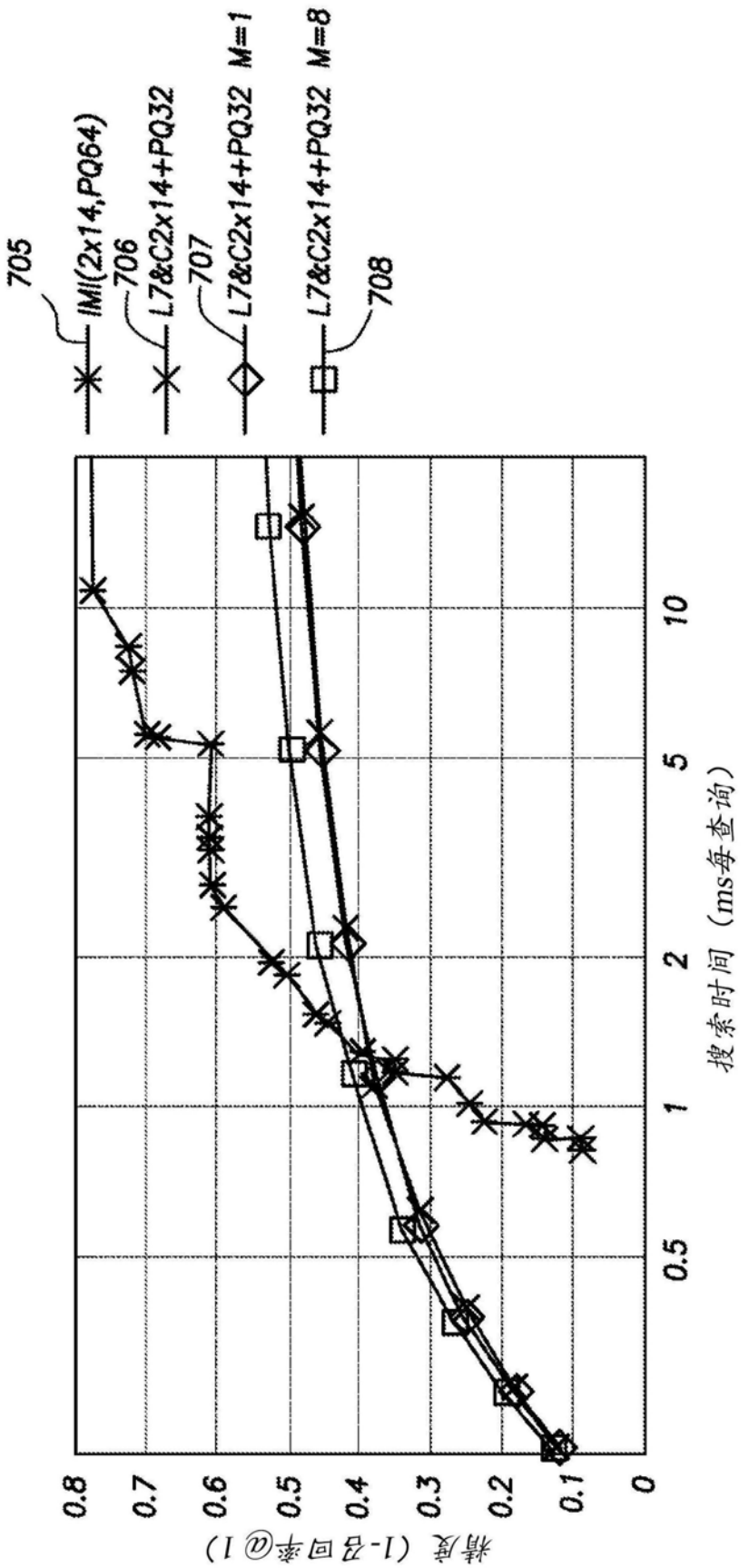

705
IMI(2x14,PQ64)
706
L7&C2x14+PQ32
707
L7&C2x14+PQ32 M=1
708
L7&C2x14+PQ32 M=8
精度（1-召回率@1）
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
搜索时间（ms每查询）
0.5
1
2
5
10

图7B

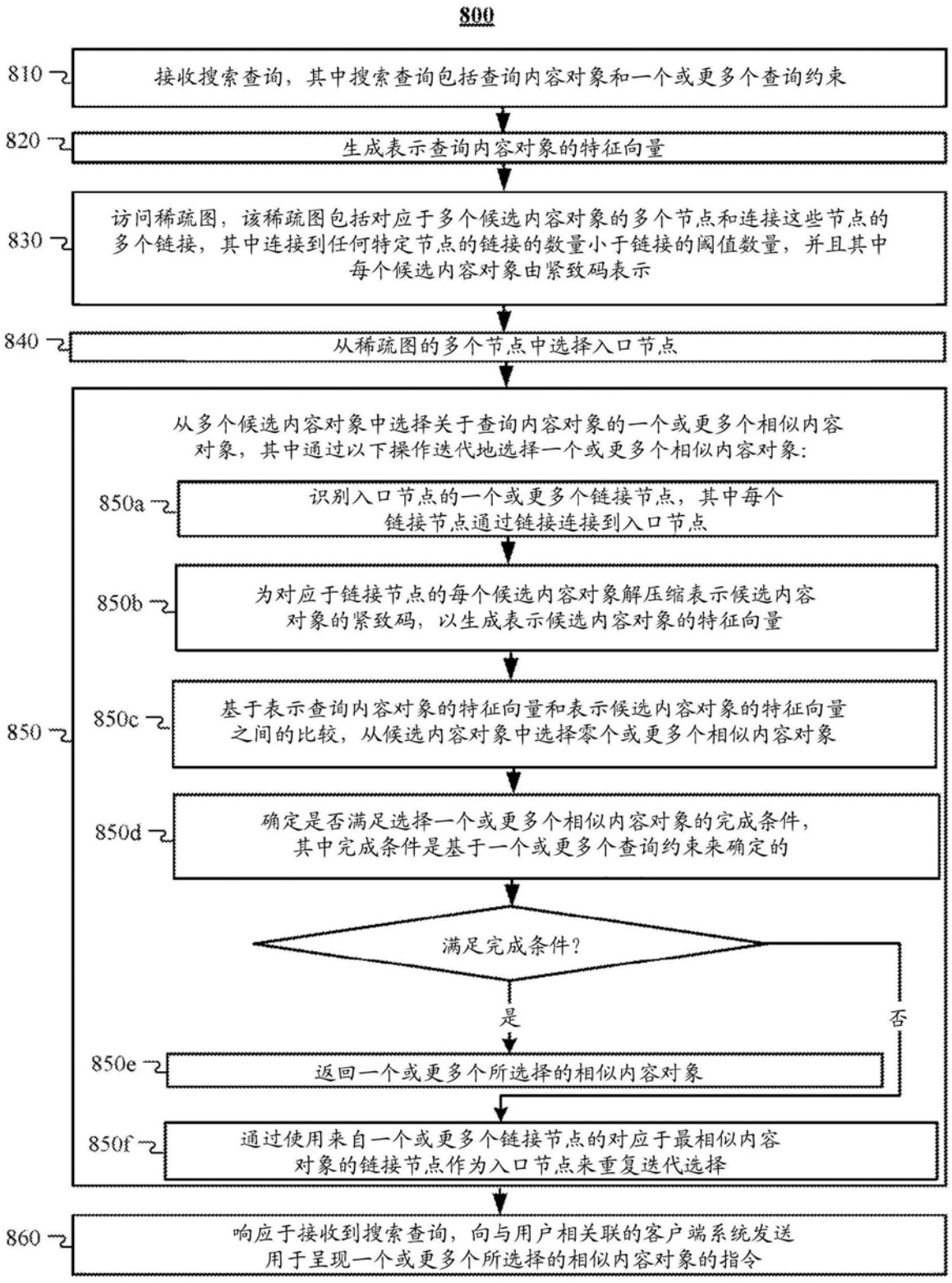

800
810
接收搜索查询，其中搜索查询包括查询内容对象和一个或更多个查询约束
820
生成表示查询内容对象的特征向量
830
访问稀疏图，该稀疏图包括对应于多个候选内容对象的多个节点和连接这些节点的多个链接，其中连接到任何特定节点的链接的数量小于链接的阈值数量，并且其中每个候选内容对象由紧致码表示
840
从稀疏图的多个节点中选择入口节点
850
从多个候选内容对象中选择关于查询内容对象的一个或更多个相似内容对象，其中通过以下操作迭代地选择一个或更多个相似内容对象:
850a
识别入口节点的一个或更多个链接节点，其中每个链接节点通过链接连接到入口节点
850b
为对应于链接节点的每个候选内容对象解压缩表示候选内容对象的紧致码，以生成表示候选内容对象的特征向量
850c
基于表示查询内容对象的特征向量和表示候选内容对象的特征向量之间的比较，从候选内容对象中选择零个或更多个相似内容对象
850d
确定是否满足选择一个或更多个相似内容对象的完成条件，其中完成条件是基于一个或更多个查询约束来确定的
满足完成条件?
是
否
850e
返回一个或更多个所选择的相似内容对象
850f
通过使用来自一个或更多个链接节点的对应于最相似内容对象的链接节点作为入口节点来重复迭代选择
860
响应于接收到搜索查询，向与用户相关联的客户端系统发送用于呈现一个或更多个所选择的相似内容对象的指令

图8

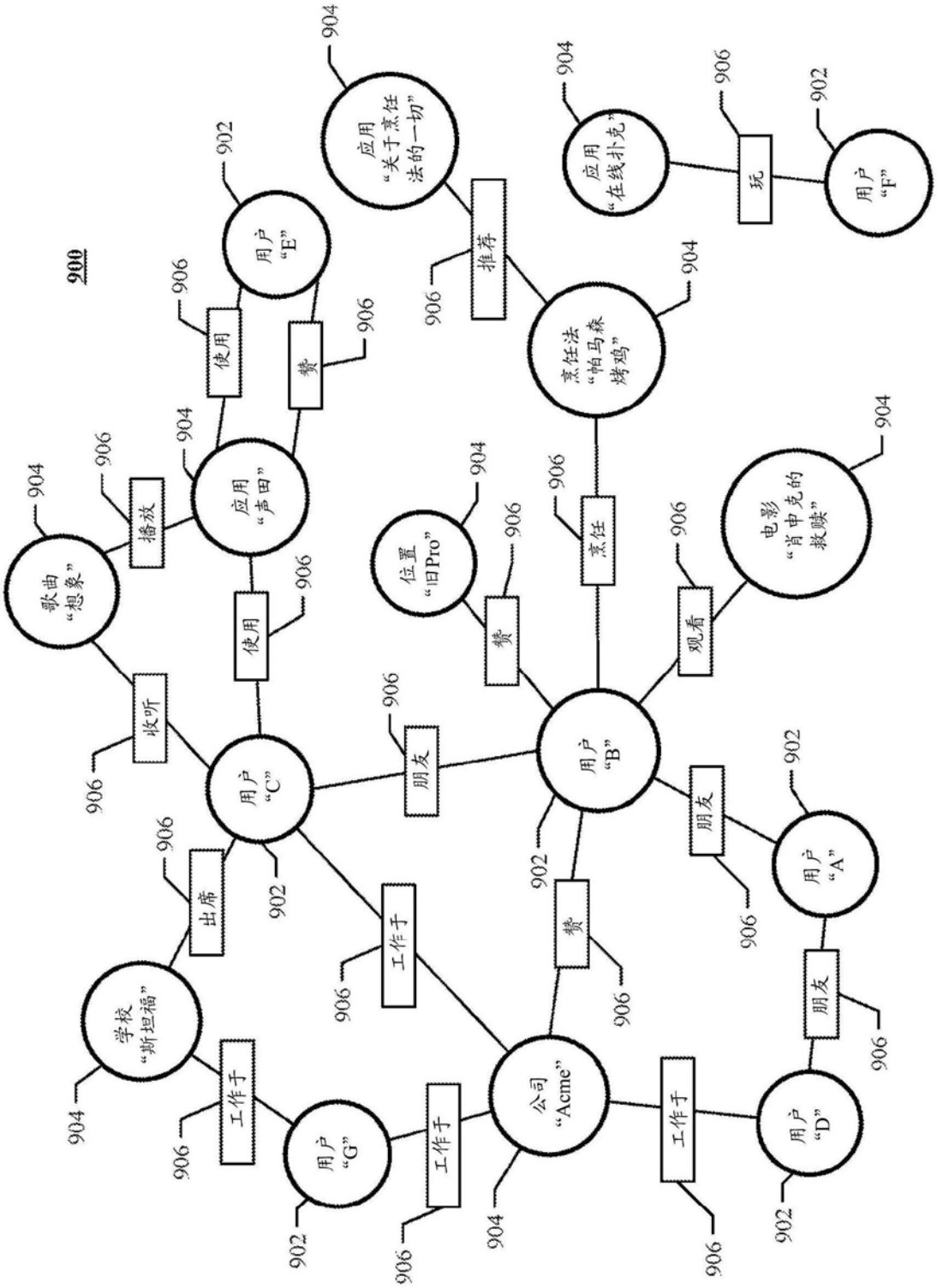

200
用户 "E"
902
应用 "关于烹饪法的一切"
904
应用 "在线扑克"
904
玩
906
用户 "F"
902
推荐
906
使用
906
赞
906
烹饪法 "帕马森烤鸡"
904
应用 "声田"
904
播放
906
歌曲 "想象"
904
位置 "旧Pro"
904
赞
906
烹饪
906
电影 "肯申克的救赎"
904
使用
906
观看
906
收听
906
用户 "C"
902
朋友
906
用户 "B"
902
出席
906
工作于
906
朋友
906
用户 "A"
902
学校 "斯坦福"
904
工作于
906
工作于
906
赞
906
朋友
906
用户 "G"
902
工作于
906
公司 "Acme"
904
工作于
906
用户 "D"
902

图9

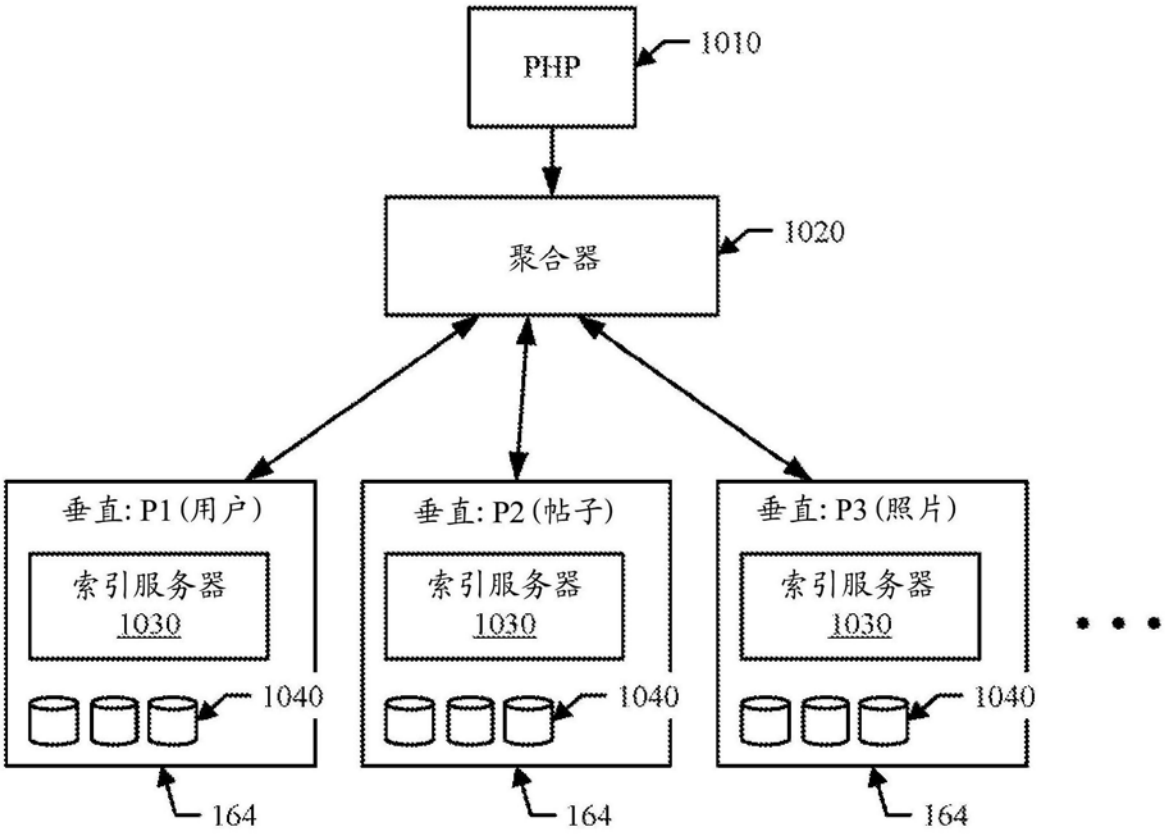

PHP
1010
聚合器
1020
垂直: P1 (用户)
索引服务器
1030
1040
164
垂直: P2 (帖子)
索引服务器
1030
1040
164
垂直: P3 (照片)
索引服务器
1030
1040
164

图10

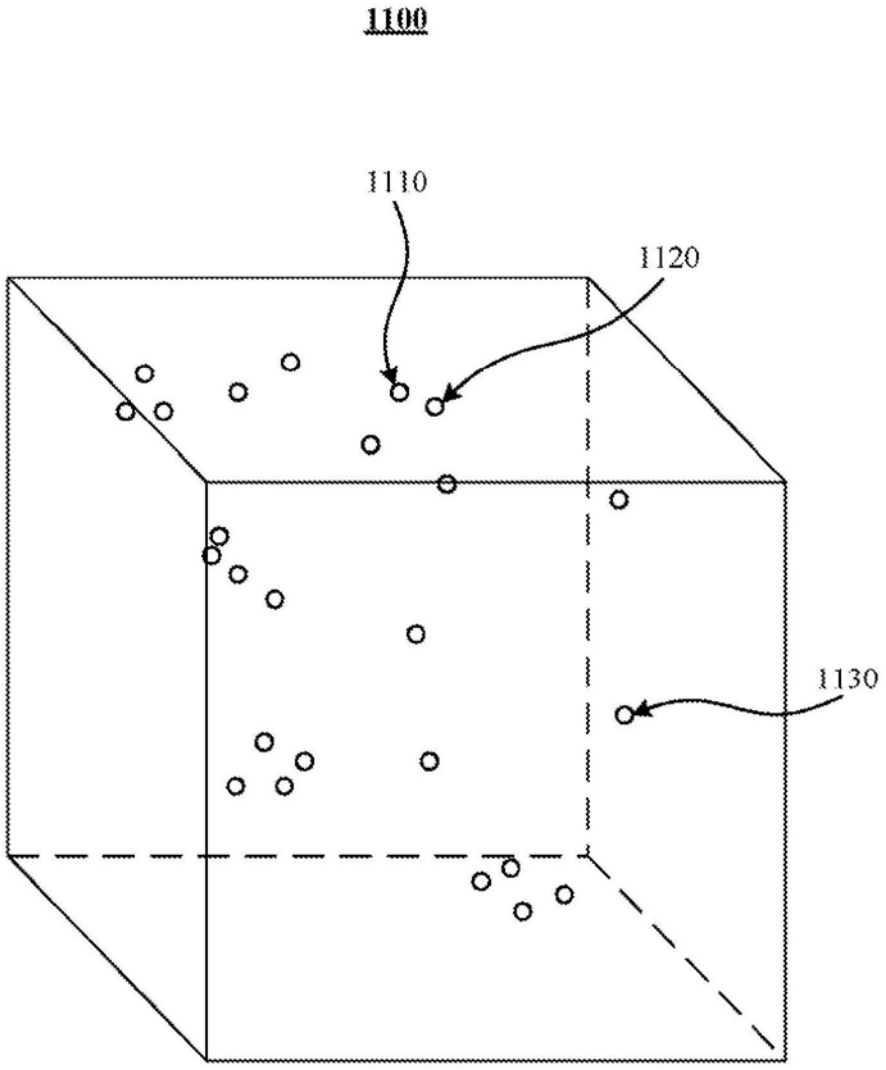


图11

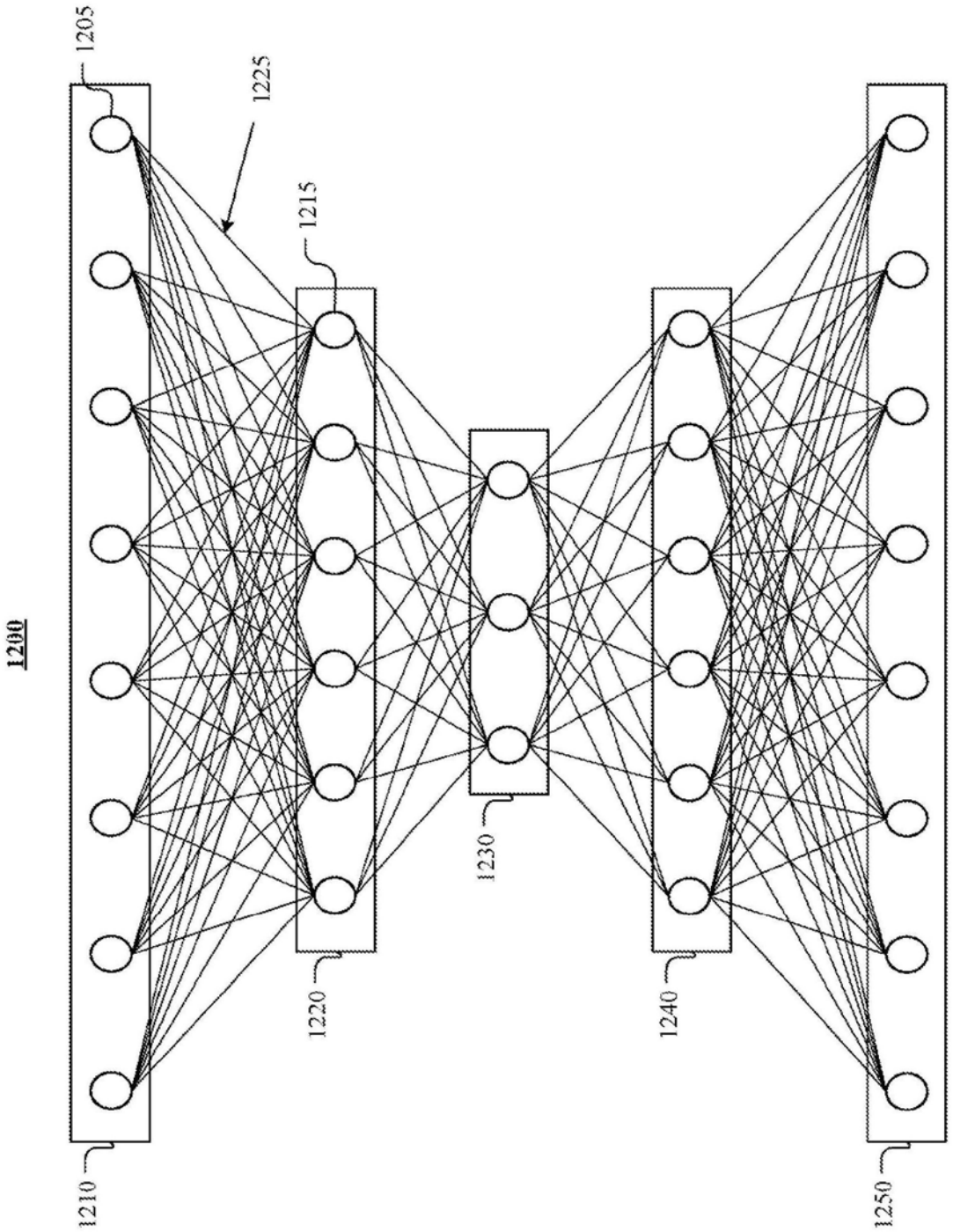


图12

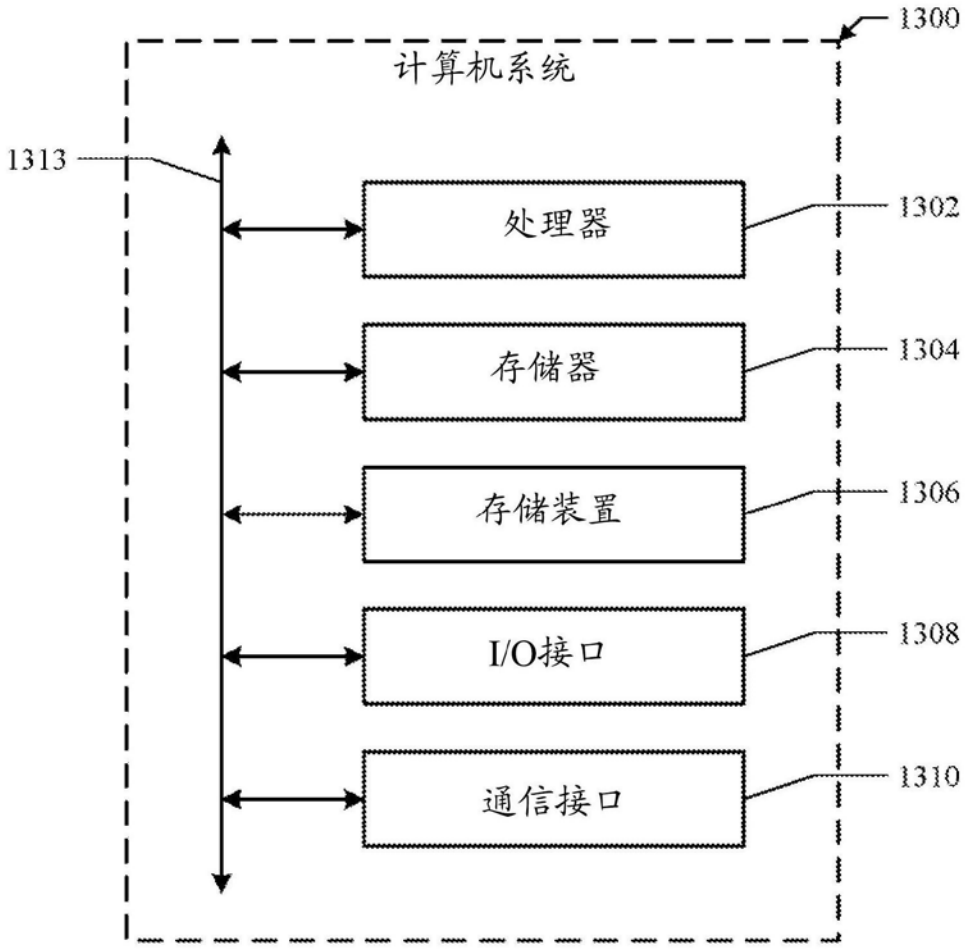

1300
计算机系统
1313
处理器
1302
存储器
1304
存储装置
1306
I/O接口
1308
通信接口
1310

图13