(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 105723402 A
(43)申请公布日 2016.06.29

(21)申请号 201480058623.2
(22)申请日 2014.10.23
(30)优先权数据
61/895,539 2013.10.25 US
61/907,878 2013.11.22 US
62/020,833 2014.07.03 US
(85)PCT国际申请进入国家阶段日
2016.04.25
(86)PCT国际申请的申请数据
PCT/CA2014/051033 2014.10.23
(87)PCT国际申请的公布数据
WO2015/058309 EN 2015.04.30
(71)申请人 西斯摩斯公司
地址 加拿大安大略省多伦多市约克大街25号
(72)发明人 爱德华·东晋·金
布莱恩·佳利·耿
(74)专利代理机构 北京三聚阳光知识产权代理有限公司 11250
代理人 程纲
(51)Int.Cl.
G06Q 50/00(2012.01)
H04L 12/16(2006.01)

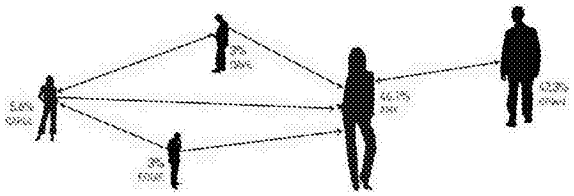
权利要求书2页 说明书19页 附图22页

(54)发明名称

用于确定社交数据网络中的影响者的系统和方法

(57)摘要

一种由服务器执行以确定对主题有影响力的至少一个用户账户的系统和方法，包括：获取所述主题；确定社交数据网络中与所述主题相关的多个用户账户；在连接图中将每一个所述用户账户表示为节点及确定每一个所述用户账户之间关系的存在；使用每一个作为节点的所述用户账户及每一个所述节点之间的作为边界的所述相应关系以计算主题网络图；将所述主题网络图中的所述用户账户排名以过滤所述主题网络图中的异常值节点；识别所述过滤的主题网络图中的所述用户账户中的至少两个不同社群，每一社群与所述用户账户的子集有关；识别与每一社群有关的属性；输出与所述相应属性有关的每一社群。



CN 105723402 A

1.一种由服务器执行以确定对主题有影响力的一个或多个用户的方法,包括:

获取主题;

确定社交数据网络内与所述主题相关的用户;

将每一所述用户建模为节点并确定每一所述用户之间的关系;

使用作为节点的所述用户及作为边界的所述关系以计算主题网络图;

将所述主题网络图中的所述用户排名;

识别并过滤所述主题网络图中的异常值节点;以及

根据其有关的排名输出所述主题网络图中剩余的用户。

2.如权利要求1所述的方法,其中,消耗和产生包括所述主题的内容中的至少一者的所述用户被视为与所述主题相关的用户。

3.如权利要求1所述的方法,其中,在所述主题网络图中,定义于至少两个用户之间的边界表示所述至少两个用户之间的朋友联系。

4.如权利要求1所述的方法,其中,在所述主题网络图中,定义于至少两个用户之间的边界表示所述至少两个用户之间的关注者与被关注者联系,且其中所述至少两个用户中的一者为关注者及所述至少两个用户中的另一者为被关注者。

5.如权利要求1所述的方法,其中,在所述主题网络图中,定义于至少两个用户之间的边界表示所述至少两个用户之间的回复联系,且其中所述至少两个用户中的一者回复所述至少两个用户中的另一者发布的帖。

6.如权利要求1所述的方法,其中,在所述主题网络图中,定义于至少两个用户之间的边界表示所述至少两个用户之间的转发联系,且其中所述至少两个用户中的一者转发所述至少两个用户中的另一者发布的帖。

7.如权利要求1所述的方法,其中,所述排名包括使用页面排名算法以衡量所述主题网络图中给定用户的重要性。

8.如权利要求1所述的方法,其中,所述排名包括使用以下至少一者:特征向量中心、权度、中间性及中心及权威性度量。

9.如权利要求1所述的方法,其中,识别及过滤所述主题网络图中的异常值节点包括:将集群算法、模块性算法及社群检测算法中的至少一者应用于所述主题网络图以输出多个社群;通过所述多个社群中的每一者中的若干用户将所述多个社群排序;选择n个具有最大数量用户的社群,其中所述n个社群中的所述用户的累积和至少达到所述主题网络图中的用户总数的百分比阈值;以及将未选的社群中的用户建立为所述异常值节点。

10.一种用于确定对主题有影响力的一个或多个用户的计算系统,包括:

通信装置;

存储器;以及

处理器,其配置为至少:

获取主题;

确定社交数据网络内与所述主题相关的用户;

将每一所述用户建模为节点并确定每一所述用户之间的关系;

使用作为节点的所述用户及作为边界的所述关系以计算主题网络图;

将所述主题网络图中的所述用户排名;

识别并过滤所述主题网络图中的异常值节点;以及

根据其有关的排名输出所述主题网络图中剩余的用户。

11.如权利要求10的所述计算系统,其中,消耗和产生包括所述主题的内容中的至少一者的所述用户被视为与所述主题相关的所述用户。

12.如权利要求10的所述计算系统,其中,在所述主题网络图中,定义于至少两个用户之间的边界表示所述至少两个用户之间的朋友联系。

13.如权利要求10的所述计算系统,其中,在所述主题网络图中,定义于至少两个用户之间的边界表示所述至少两个用户之间的关注者与被关注者联系,且其中所述至少两个用户中的一者为关注者及所述至少两个用户中的另一者为被关注者。

14.如权利要求10的所述计算系统,其中,在所述主题网络图中,定义于至少两个用户之间的边界表示所述至少两个用户之间的回复联系,且其中所述至少两个用户中的一者回复所述至少两个用户中的另一者发布的帖。

15.如权利要求10的所述计算系统,其中,在所述主题网络图中,定义于至少两个用户之间的边界表示所述至少两个用户之间的转发联系,且其中所述至少两个用户中的一者转发所述至少两个用户中的另一者发布的帖。

16.如权利要求10的所述计算系统,其中,所述排名包括使用页面排名算法以衡量所述主题网络图中给定用户的重要性。

17.如权利要求10的所述计算系统,其中,所述排名包括使用以下至少一者:特征向量中心、权度、中间性及中心及权威性度量。

18.如权利要求10的所述计算系统,其中,识别及过滤所述主题网络图中的异常值节点包括:将集群算法、模块性算法及社群检测算法中的至少一者应用于所述主题网络图以输出多个社群;通过所述多个社群中的每一者中的若干用户将所述多个社群排序;选择n个具有最大数量用户的社群,其中所述n个社群中的所述用户的累积和至少达到所述主题网络图中的用户总数的百分比阈值;以及将未选的社群中的用户建立为所述异常值节点。

用于确定社交数据网络中的影响者的系统和方法

相关申请的交叉引用

[0001] 本申请案请求2013年10月25日提交的标题为用于确定社交数据网络中的影响者的系统和方法”的美国临时专利申请案第61/895,539号、2013年11月22日提交的题为“用于识别社交数据网络中的影响者及其社群的系统和方法”的美国临时专利申请案第61/907,878号以及2014年7月3日提交的题为“以使用加权分析来动态确定社交数据网络中的影响者的系统和方法”的美国临时专利申请案第62/020,833号的优先权,所述申请案的全部内容以引用的方式并入本文中。

技术领域

[0002] 以下总体上涉及分析社交网络数据。

背景技术

[0001] 近年来,社交媒体已经成为个人和消费者在线(例如,在互联网上)交互的大众化方式。社交媒体还影响企业目的在于和其客户、粉丝、和潜在客户在线交互的方式。

[0002] 在特定主题上具有广大关注的一些部落客被识别并用于支持或赞助特定的产品。例如,大众化部落客的网站上的广告空间用于为相关产品及服务打广告。

[0003] 社交网络平台也用于影响成群的人。社交网络平台的实例包括那些通过商标名称所熟知的脸谱网、推特、领英(LinkedIn)、汤博乐及拼趣。社交网络平台中的大众化或专家个人可用于向其他人营销。当社交网络中的用户数量增长时,快速识别大众化或有影响力的个人变得越来越难。此外,难以准确识别特定主题中有影响力的个人。在社交网络中的专家或那些大众化用户在本文中可交换地称为“影响者”。

附图简要说明

[0004] 现在参考附图仅通过举例方式来描述实施例,在附图中:

[0005] 图1是展示在社交数据网络中相互联系的用户的图式。

[0006] 图2是与计算装置通信的服务器的示意图。

[0007] 图3是用于确定与主题有关的影响者的计算机可执行指令的实施例的流程图。

[0008] 图4是用于确定与主题有关的影响者的计算机可执行指令的另一实施例的流程图。

[0009] 图5是用于获取和储存社交网络数据的计算机可执行指令的实施例的流程图。

[0010] 图6是索引存储中的示例数据组件的框图。

[0011] 图7是简档存储的示例数据组件的框图。

[0012] 图8是示例用户列表及用户列入不同用户列表中的次数的统计的示意图。

[0013] 图9是用于确定其中给定用户被视作专家的主题的计算机可执行指令的实施例的流程图。

[0014] 图10是用于确定给定用户感兴趣的主题的计算机可执行指令的实施例的流程图。

[0015] 图11是用于搜索索引存储中被视作主题的专家的用户的计算机可执行指令的实

施例的流程图。

[0016] 图12是用于识别对主题感兴趣的用户的计算机可执行指令的实施例的流程图。

[0017] 图13是用于主题“McCafe”(麦咖啡)的示例主题网络图的图解。

[0018] 图14是图13中的主题网络图的图解,展示主集群和异常值集群的分解。

[0019] 图15是用于基于社群分解在主题网络中识别和过滤异常值的计算机可执行指令的实施例的流程图。

[0020] 图16是用于自每一主题网络识别和提供社群集群的计算机可执行指令的实施例的流程图。

[0021] 图17A至图17D展示用于与主题网络中显示影响者社群的GUI互动的示例性截屏。

[0022] 图18图示示例性社群网络图。

[0023] 图19A至图19C展示特定主题的示例性社群及特征。

[0024] 图20A至图20B展示第二选择主题的示例性社群及特征。

附图详细说明

[0025] 应当认识的是,为了说明的简化和清晰,在认为适当时,参考数字可在图中被重复以指示相应或相似的元件。此外,陈述了许多特定细节,以提供对本文中所描述的实施例的透彻理解。然而,本领域的普通技术人员将理解的是,没有这些特定细节也可以实践本文中所描述的实施例。在其他情形下,没有详细描述公知方法、程序和部件,以不使本文中所描述的实施例难理解。并且,本说明不被认为是限制本文中所描述的实施例的范围。

[0026] 社交网络平台包括(例如通过藉由与社交网络平台有关的网站通信的计算装置的网络)产生并发布内容给其他人看、听等的用户。社交网络平台的非限制性实例为Facebook、Twitter、LinkedIn、Pinterest、Tumblr、博客圈、网站、协作维基百科、在线新闻组、在线论坛、电子邮件以及即时消息业务。目前已知及未来可知的社交网络平台适用于本文中所描述的原理。社交网络平台可用于向平台的用户推广及发布广告。应认识到难以识别给定主题的相关用户。这包括识别给定主题上有影响力的用户。

[0027] 如本文所用,术语“影响者”是指首先产生并分享与主题有关的内容并被视为对社交数据网络中的其他用户有影响力的用户账户。如本文所用,术语“关注者”,是指关注第二用户账户(例如与第一用户账户的至少一个社交网络平台有关并通过计算装置存取的第二用户账户),以使得公开第二用户账户发布的内容供第一用户账户阅读、消耗等的第一用户账户(例如与一个或多个社交网络平台有关并通过计算装置存取的第一用户账户)。例如,当第一用户关注第二用户,第一用户(即关注者)将接收第二用户发布的内容。本文中对特定主题“感兴趣”的用户是指关注特定主题中的(例如与社交网络平台有关的)若干专家的用户账户。在某些情况下,关注者(例如通过分析或转发内容)参与其他用户发布的内容。

[0028] 公司需要识别关键影响者以(例如)将可潜在传播及支持品牌消息的个人作为目标。使所述个人参与可允许控制品牌的在线消息及可降低可能发生的潜在负面情绪。仔细管理该过程可(例如)在病毒式营销活动的情况下引起在线注意力份额的指数增长。

[0029] 过去用于确定影响者的大多数方式的关注点在于易计算的度量,例如关注者或朋友的数量、或发帖的数量。在合计的关注者或朋友计数可接近于整体社交网络时,其通过计算度量的方式提供小数据,所述度量表明相对于公司或品牌的用户或个人的影响力。此带来嘈杂影响者结果以及筛选大量潜在用户所浪费的时间。

[0030] 一些社交媒体分析公司宣称提供社交网络的影响者分数。然而，本文中认识到许多公司使用并非真实影响者度量的度量，而非关注者数量及提及的次数(例如Twitter的推文、帖子、消息等)的代数式。例如，一些已知方式使用所述数字的对数归一，其将约80％的权重分配至关注者计数以及提及的次数的余数。

[0031] 使用代数式的原因在于关注者和提及的计数或计算在社交网络的用户简档中是实时更新的。因此，计算迅速并且易于报告。这通常被称为权威度量或权威分数以将其与真实影响者分析区分开。然而，权威分数方式具有若干严重的缺点。

[0032] 本文中认识到所述权威分数为上下文非相关。其为与主题或查询无关的静态度量。例如，且不论主题，由于具有数百万关注者，如纽约时报(New York Times)或CNN的大众媒体可得到最高的排名。因此，其不是上下文相关的。

[0033] 本文中还认识到所述权威度量具有高关注者计数偏差。如果某一领域中存在拥有有限数量关注者的明确定义的专业人员，但他们不都是专家，由于其低关注者计数，他们绝不会出现在前20至100个结果中。实际上，所有的关注者均被当做具有相同的权重，这已被视为网络分析研究中错误的假定。

[0034] 本文中提出的系统和方法可动态计算关于查询主题的影响者，并且可对其关注者的影响力作出解释。

[0035] 本文中还认识到影响者关系的递归性是大规模执行影响者识别时的一种挑战。通过举例，假设存在个人A、B和C的情况下，其中A关注B和C；B关注C和A；以及C仅关注A。随后A的影响力取决于C，C的影响力又取决于A和B，等等。这样，这样，影响者关系具有递归性。

[0036] 更一般而言，提出的系统及方法提供一种确定社交数据网络中影响者的方式。

[0037] 作为实例，考虑图1中的特定主题的简化的关注者网络。展示与其他用户有关的每一用户(实际上为用户账户或与用户账户或用户数据地址有关的用户名)。所述用户之间的线，也被称为边界，代表用户之间的关系。例如，从用户账户“Dave”指向用户账户“Carol”的箭头表示Dave读取Carol发布的消息。换言之，Dave关注Carol。Amy和Brian之间的双向箭头表示，例如，Amy关注Dave，并且Dave关注Amy。除去图1中的每一用户账户，提供页面排名分数。页面排名算法为谷歌(Google)所用的已知算法用以衡量网络中的网站的重要性，并且还可应用于衡量社交数据网络中的用户的重要性。

[0038] 继续图1，Amy拥有大量关注者(即Dave、Carol和Eddie)，并且为所述网络中最具有影响力的用户(即页面排名分数为46.1％)。然而，仅拥有一个关注者(即Amy)的Brian比拥有两个关注者(即Eddie和Dave)的Carol更具影响力，主要是因为Brian具有很大一部分Amy的注意力份额。换言之，使用本文提出的系统及方法，虽然Carol比Brain具有更多的关注者，但是她不一定比Brian更具有影响力。因此，使用本文所提出的系统及方法，用户的关注者数量并非影响力的唯一决定因素。在实施例中，识别用户的关注者是谁也可作为影响力计算的因素。

[0039] 表1表示图1中的示例性网络，并且其说明页面排名可如何显著地与关注者数量区分开。

用户句柄	关注者计数	页面排名
Amy	4	46.1％
Brian	1	42.3％

Carol	2	5.6%
Dave	0	3.0%
Eddie	0	3.0%

表1:图1所示的Twitter关注者计数及样本网络的页面排名分数。

[0040] Amy拥有最大数量的关注者以及最高的页面排名分数,因而明显为最具影响力者。虽然Carol拥有两个关注者,但是与拥有1个关注者的Brian相比,她具有较低的页面排名度量。然而,Brian的一个关注者是最具影响力的Amy(拥有四个关注者),而Carol的两个关注者为低影响者(每人拥有0个关注者)。其直观表明,如果某人被少数专家认定为专家,则她/他也是专家。然而,与仅计数关注者数量相比,页面排名算法可更好的衡量影响力。如下所述,页面排名算法及其他类似排名算法可与本文中所提出的系统及方法一起使用。

[0041] 提出的系统及方法可用于确定社交数据网络中给定主题的关键影响者。

[0042] 在实施例中,提出的系统及方法可用于确定主题A中的影响者也是一个或多个其他主题(例如主题B、主题C等)中的影响者。

[0043] 转至图2,展示所提出的系统的示意图。服务器100通过网络102与计算装置101通信。服务器100获取并分析社交网络数据并且通过网络将结果提供至计算装置101。计算装置101可通过GUI接收用户输入以控制分析参数。

[0044] 可以认识到,社交网络数据包括社交网络平台的用户相关的数据,以及用户产生或组织,或产生并组织的内容。社交网络数据的非限制性实例包括用户账户ID或用户名、用户或用户账户的描述、用户发布的消息或其他数据、用户与其他用户之间的联系、位置信息等。联系的一个实例为“用户列表”(在本文中也称为“列表”),其包括列表名、列表描述以及给定用户关注的一个或多个其他用户。例如,用户列表由给定用户创建。

[0045] 继续图2,服务器100包括处理器103及存储装置104。在实施例中,服务器包括一个或多个处理器及大量存储器容量。在另一实施例中,存储装置104或存储装置为增加读写性能的固态驱动器。在另一实施例中,多个服务器用于实施本文中所述的方法。换言之,在实施例中,服务器100是指服务器系统。在另一实施例中,使用其他当前已知计算硬件或未来可知计算硬件,或两者。

[0046] 服务器100还包括通信装置105以通过网络102通信。网络102可为有线或无线网络,或有线及无线网络两者皆可。服务器100还包括GUI模块106,用于通过计算装置101显示和接收数据。服务器还包括:社交网络数据模块107;索引器模块108;用户账户关系模块109;专家识别模块110;兴趣识别模块111;用于识别对主题A(例如给定主题)感兴趣的用户的查询模块114;社群识别模块112及特征识别模块113。将描述,社群识别模块112被配置为基于专家识别模块识别的关系网络图来定义数据的社群或集群。

[0047] 服务器100还包括若干数据库,包括数据存储116;索引存储117;社交图数据库118;简档存储119;专家知识向量数据库120;兴趣向量数据库121;用于存储社群图信息的数据库128;以及用于存储每一社群的大众化特征以及存储在每一社群中将搜索的预定义特征的数据库129,社群识别模块112定义所述社群。

[0048] 社交网络数据模块107用于接收社交网络数据流。在实施例中,每天有数百万新消息实时传送至社交网络数据模块107。社交网络数据模块107接收的社交网络数据存储在数据存储116中。

[0049] 索引器模块108对数据存储116中的数据执行索引器进程并在索引存储117中存储已编索引的数据。在实施例中，可更容易地搜索索引存储117中已编索引的数据，并且索引存储中的标识符可用于检索实际数据(例如全消息)。

[0050] 社交图同样获取自社交网络平台服务器(未展示)并存储于社交图数据库118中。当作为输入给予用户以供查询时，社交图可用于返回关注所查询用户的所有用户。

[0051] 简档存储119存储与用户简档有关的元数据。与简档有关的元数据的实例包括给定用户的关注者合计数量、给定用户的自揭个人信息、给定用户的位置信息等。可查询简档存储119中的数据。

[0052] 在实施例中，用户账户关系模块109可使用社交图118及简档存储119以确定有哪些用户关注特定用户。

[0053] 专家识别模块110被配置为识别其中列出了用户账户的所有用户列表组，所述所有用户列表组被称为专家知识向量。用户的专家知识向量存储在专家知识向量数据库120中。兴趣识别模块111被配置为识别给定用户感兴趣的主题，所述主题被称为兴趣向量。用户的兴趣向量存储在兴趣向量数据库121中。

[0054] 再次参见图2，服务器100进一步包括社群识别模块112，所述社群识别模块被配置为识别通过专家识别模块110识别的主题网络以及有关的影响者中的社群(例如查询如主题A的主题中的信息集群)。根据图3将描述的，主题网络说明(例如由专家识别模块110及/或社交图118定义的)有影响力的用户及其关系的图。自社群识别模块112的输出包括定义为主题网络的社群的集群的可视识别(例如颜色编码)，所述集群包含共同特征及/或被同一社群的其他实体(例如影响者)影响至与另一社群中的所述集群相比更高的程度。服务器100进一步包括特征识别模块113。

[0055] 特征识别模块113被配置为接收自社群识别模块112识别的社群并提供对社群成员中大众化特征(例如会话主题)的识别。特征识别模块113的结果能可视地链接至如社群识别模块112中提供的社群的相应可视化。将描述，一方面，社群识别模块112的结果(例如多个社群)及/或特征识别模块113(例如每一社群的多个大众化特征)在显示于显示屏125上作为对计算装置101的输出。在另一个方面，GUI模块106被配置为接收自计算装置101输入以选择由社群识别模块112识别的特定社群。GUI模块106随后被配置为与特征识别模块113通信，以提供与所选社群(例如所选社群中的所有有影响力的用户)有关的特定特征(例如定义大众化会话)的结果的输出。特征识别模块112的结果(例如可视地定义所选社群的用户中的大众化会话的字云)可于特定所选社群及/或特定所选社群中用户的列表旁显示于显示屏125上。

[0056] 继续图2，计算装置101包括通过网络102与服务器100通信的通信装置122、处理器123、存储装置124、显示屏125及因特网浏览器126。在实施例中，计算装置101通过因特网浏览器显示服务器100提供的GUI。在另一实施例中，在计算装置101上可使用分析应用127的情况下，计算装置通过分析应用127显示GUI。可以认识到，显示装置125可为计算装置的一部分(例如，与移动装置、平板电脑、笔记本电脑等的情况一样)或可与计算装置分离(例如，与台式电脑等的情况一样)。

[0057] 虽然未展示，但是各种用户输入装置(例如触控屏幕、滚珠、光学鼠标、按钮、键盘、麦克风等)可用于促进用户与计算装置101之间的互动。

[0058] 可以认识到本文中举例说明的执行指令的任一模块或元件可包括或可以利用计算机可读媒体，例如存储媒体、计算机存储媒体或(可移动及/或固定的)数据存储装置，所述数据存储装置例如(如)磁盘、光盘或磁带。计算机可读媒体可包括以任何方法或技术实施以存储信息的易失性及非易失性、可移动及固定媒体，例如计算机可读指令、数据结构、程序模块或其他数据。计算机可读媒体的实例包括RAM、ROM、EEPROM、闪存或其他存储器技术、CD-ROM、数字通用光盘(digital versatiledisks，DVD)或其他光学存储装置、盒式录音带、磁带、磁盘存储器或其他磁性存储装置，或可用于存储所需信息并可由应用、模块或两者存取的任何其他媒体。任何所述计算机可读媒体可为服务器100或计算装置101的一部分或者可存取至或可连接至服务器100或计算装置101。可使用由所述计算机可读媒体存储或具有的计算机可读/可执行指令实施本文所述的任何应用或模块。

[0059] 转至图3，展示计算机可执行指令的实施例以确定给定主题的一个或多个影响者。图3中展示的过程假定服务器100可存取社交网络数据，并且社交网络数据包括表示为组U的多个用户。在框301处，服务器100获取表示为T的主题。例如，用户可通过显示在计算装置101处的GUI输入主题，并且计算装置101将主题发送至服务器100。主题也可不由用户输入而获取。在框302处，服务器使用主题以自与主题有关的社交网络数据确定用户。所述确定可以各种方式实施，并将在下文中详细描述。与主题有关的用户组表示为U_T，其中U_T是U的子集。

[0060] 继续图3，服务器将用户组UT中的每一用户建模为节点并确定用户U_T之间的关系(框303)。服务器计算分别对应于用户U_T的节点和边界以及用户U_T之间的关系的网络(框304)。换言之，服务器创建分别对应于用户U_T的节点和边界以及器关系的网络图。网络图被称为“主题网络”。可以认识到，在此适用图论原则。定义两个实体或用户U_T之间的边界或连通性的关系可包括(例如)：特定社交网络平台中的两个实体之间的朋友联系及/或关注者与被关注者联系。在另一个方面，关系可包括定义两个实体之间的社交媒体连通性的其他类型的关系，例如朋友的朋友的联系。在另一方面，关系可包括跨越不同社交网络平台(例如Instagram及Facebook)的朋友或关注者联系的连通性。在另一个方面，由边定义义的用户U_T之间的关系可包括(例如)：通过由一个用户转发原本由另一用户用户发布的消息来联系的用户(例如在Twitter上重新发推文)，及/或通过经由社交网络平台回复由一个用户发布并由另一用户推荐的消息来联系的用户。再次参见图3，两个实体之间存在边界表明一个或多个社交网络平台中至少一种类型的关系或连通性(例如两个用户之间的朋友或关注者连通性)。

[0061] 服务器随后排名主题网络中的用户(框305)。例如，服务器使用页面排名以衡量主题网络中的用户的重要性并基于衡量来排名用户。可使用的排名算法的其他非限制性实例包括：特征向量中心、权度、中间性、中心及权威性度量。

[0062] 服务器识别并过滤出主题网络中的异常值节点(框306)。异常值节点为在主题网络中被视为与大量用户或用户集群不同的异常值用户。主题网络中的异常值用户或节点组表示为U_0，其中U_0是U_T的子集。以下描述关于识别及过滤异常值节点的进一步细节。

[0063] 在框307，服务器根据排名输出用户U_T，移除用户U_0。

[0064] 在替代的实施例中，在框305之前执行框306。

[0065] 在框308，服务器识别移除用户U_0的用户U_T的社群(例如C_1，C_2，…，C_n)。社群识别可

取决于与另一社群的节点相比一个社群中的节点之间的连通性程度。即，通过相对于所定义的社群以外的实体内在具有更高程度的连通性的实体或节点(例如相对于同一社群的其他节点)来定义社群。将定义，可预定义用于将一个社群与另一个社群分开的连通性程度的值或阈值(例如，如社群图数据库128及/或自计算装置101的用户定义所提供)。因此，解析度定义社群中的节点的互联性的密度。因此，每一识别的社群图为框304中针对每一社群定义的节点和边界(主题网络)的网络图的子集。一方面，社群图进一步用社群图显示社群(例如节点)中的用户的视觉表示及(例如图1中提供至显示屏125的)社群中的用户的文字列表。在另一个方面，根据(例如图1中提供至显示屏125的)社群及/或主题T的所有社群中的影响力程度排名社群中的用户的列表的显示。根据框308，用户U_T随后分为其社群图分类，例如U_{C1}，U_{C2}，…U_{Cn}。

[0066] 在框309处，对于每一给定社群(例如C_1)，服务器基于其社交网络数据确定与给定社群中用户(例如U_{C1})相关的预定义特征(例如以下一个或多个：共同词组及短语、会话主题、共同位置、共同图片、共同元数据)的大众化特征值。所选特征(例如主题或位置)可为用户定义的(例如通过自计算装置101的输入)及/或自动生成的(例如基于同一主题网络中的其他社群的特征、或基于针对同一主题T的以前所用特征)。在框310处，服务器输出所识别的社群(例如C_1，C_2，…，C_n)及与每一给定社群有关的大众化特征。可将所识别的社群输出(例如通过服务器显示于显示屏125上)为与针对每一社群的预定义特征的特征值视觉关联的社群图。

[0067] 转至图4，展示计算机可执行指令的另一实施例以确定给定主题的一个或多个影响者。框401至404对应框301至304。框404之后，服务器100使用第一排名过程排名主题网络中的用户(框405)。第一排名过程可或不可与框305中的排名过程一样。完成排名以识别哪些用户在针对给定主题的给定主题网络中最具有影响力。

[0068] 在框406，服务器识别并过滤出主题网络中的异常值节点(用户U_0)，其中U_0是U_T的子集。在框407，服务器在移除用户U_0的情况下使用第二排名过程调整用户U_T的排名，所述第二排名过程是基于某一时间段内来自用户的发帖的数量。例如，服务器确定如果与第二用户在相同时间段之内的发帖数量相比，第一用户在过去两个月内具有更大的帖子数量，则可提高第一用户的(自框405)原始排名，而保持第二用户的排名不变或降低。

[0069] 应认识到基于所有用户U的网络图可以是相当大的。例如，在U组中可有亿万用户。分析与U相关的整个数据集可能花费太多计算以及时间。因此，使用上述过程以找到与主题T相关的较小用户组U_T减少了需要分析的数据量。此举也减少了处理时间。在实施例中，在分析Twitter的整个社交网络平台时已产生影响者的近实时结果。使用较小组的用户U_T及与用户U_T有关的数据，计算新的主题网络。与包括所有用户U的社交网络图相比，主题网络更小(即更少节点及更少边界)。基于主题网络排名用户比基于所有用户U的社交网络图排名用户快得多。

[0070] 此外，识别及过滤主题网络中的异常值节点有助于进一步提高结果的质量。

[0071] 在框409，服务器被配置为通过类似前述于框308中的方式在移除用户U_0的情况下(例如使用图2的社群识别模块112)识别用户U_T之间的社群(例如C_1，C_2，…，C_n)。在框410，服务器被配置为针对每一给定社群(例如C_1)基于其社交网络数据通过类似前述于框309中的方式确定与给定社群(例如C_1)中的用户(例如U_{C1})有关的预定义特征(例如共同关键词及短

语、会话主题、共同位置、共同图片、共同元数据)的大众化特征值。在框411，服务器被配置为通过类似于框310的方式输出所识别的社群以及与每一给定社群(例如C_1-C_n)有关的大众化特征的特征值(例如，如图2中所示，通过与服务器100及/或计算装置101有关的显示屏)。

[0072] 以下描述图3及图4中所述的方法的进一步细节。

[0073] 获取社交网络数据：

[0074] 关于获取社交网络数据，虽然未展示于图3或图4中，可以认识到，服务器100获取社交网络数据。可通过各种方式获取社交网络数据。以下为获取社交网络数据的非限制性实施例。

[0075] 转至图5，展示计算机可执行指令的实施例以获取社交网络数据。数据可实时接收为数据流，包括消息及元数据。例如，使用压缩行格式将所述数据存储在数据存储116中(框501)。在非限制性实施例中，使用MySQL数据库。例如，通过社交网络数据模块107实施框500及501。

[0076] 在实施例中，复制由社交网络模块107接收的社交网络数据，并且社交网络数据的副本跨越多个服务器存储。此举便于在分析社交网络数据时的平行处理。换言之，一个服务器分析社交网络数据的一个方面而同时另一服务器分析社交网络数据的另一方面是可能的。

[0077] 服务器100使用索引器进程为消息编索引(框502)。例如，索引器进程与包括当其在数据存储116中具体化时扫描消息的存储过程分离。在实施例中，索引器进程自身在单独的服务器上运行。此举便于平行处理。例如，索引器进程为具体化针对每天或某些其他给定的时间段的已编索引的数据的表的多线程的过程。已编索引的数据输出并存储于索引存储117(框504)。

[0078] 简单参见图6，其展示示例性索引存储117，表中的每一行为唯一的用户账户标识符及所述那天或给定时间段生成的所有消息标识符的相应列表。在实施例中，每天数百万行的数据可读取及写入索引存储117中，并且所述过程可在新数据具体化或加至数据存储116时发生。在实施例中，压缩行格式用于索引存储117中。在另一实施例中，通过运行放松的事务处理语义避免死锁，因为这样再读取和写入表时增加了跨越多个线程的吞吐量。借助上述背景，当通过每一任务均锁定一个其他任务尝试锁定的资源而使两个或更多任务永久地相互阻挡时发生死锁。

[0079] 转回至图5，服务器100进一步获取关于哪些用户账户关注其他用户账户的信息(框503)。所述过程包括识别简档相关元数据并将其存储在简档存储中(框505)。

[0080] 在图7中，简档存储119的实例展示针对每一用户账户，存在有关的简档相关元数据。例如，简档相关元数据包括用户关注者合计数量、自揭个人信息、位置信息及用户列表。

[0081] 在获取并存储数据后，可分析数据以，例如，识别专家和兴趣。

[0082] 确定有关主题的用户：

[0083] 关于确定有关主题的用户，如框302和402中每一者所述，可以认识到所述操作可以各种方式执行。以下为可用于确定有关主题的用户的非限制性实施例。

[0084] 在实施例中，确定有关主题的用户的操作(例如框302及框402)是基于Sysomos搜索引擎，并描述于2009年7月10日申请的名为“信息发现及文本分析的方法和系统(Method and System for Information Discovery and Text Analysis)”的美国专利申请案第

2009/0319518号中，该申请案特此以全文引用的形式并入。根据美国专利申请案第2009/0319518号中所述的过程，主题用于识别在一定的时间间隔内的大众化文档。本文中认识到所述过程还可用于识别有关主题的用户。特别是，当主题(例如关键词)提供至美国专利申请案2009/0319518号中的系统时，系统返回与主题相关并大众化文档(例如帖子、推文、消息、文章等)。使用本文所述并提出的系统及方法，可执行指令包括服务器100确定大众化文档的一个或多个作者。这样，一个或多个作者被识别为与给定主题相关的最佳用户。可提供上限n以识别前n个与给定主题相关的最佳用户，其中n为整数。虽然可用其他数字，但在实施例中，n为5000。可根据已知或未来所知的排名算法或使用用于社交媒体分析的已知或未来可知权威评分算法确定前n个用户。对于前n个用户中的每一者，服务器确定关注前n个用户中的每一者的用户。所述不被认为是前n个用户中的一部分或不关注前n个用户的用户不是主题网络的用户U_T的一部分。在实施例中，用户U_T的组包括前n个用户及其关注者。

[0085] 在执行确定有关主题的用户的操作的另一实施例中(例如框302及框402)，计算机可执行指令包括：确定与给定主题有关的文档(例如帖子、文章、推文、消息等)；确定文档的一个或多个作者；并将一个或多个作者建立为与给定主题有关的用户U_T。

[0086] 在执行确定有关主题的用户的操作的另一实施例中(例如框302及框402)，操作包括识别用户的专家知识向量。所述实施例使用图8至11进行解释。

[0087] 举例来说并转至图8，用户可具有他或她可能关注的其他用户的列表。例如，用户A具有用户A关注的用户B、用户C及用户D的列表。用户(例如用户B、用户C及用户D)在名为列表A的列表中分组，并且列表具有有关的列表描述n(例如描述A)。换言之，用户A相信用户B、用户C及用户D关于主题A是专家或知识渊博的。

[0088] 另一用户(用户E)可具有相同或相似的列表名和描述(例如与列表A、描述A相同或相似)，但可具有与用户A的列出的用户不用的用户。例如，用户E关注用户B、用户C及用户G。换言之，用户E相信用户B、用户C及用户G关于主题A是专家或知识渊博的。

[0089] 另一用户(用户F)可具有相同或相似的列表名和描述(例如与列表A、描述A相同或相似)，但可具有与用户A的列出的用户不用的用户。例如，用户F关注用户B、用户H及用户I，因为用户F相信所述用户关于主题A是专家或知识渊博的。

[0090] 基于上述示例性情景，可以认识到不同用户可具有相同或相似名字或相似描述的列表，但每一主题中的用户可不同。换言之，不同用户可认为其他不同用户关于给定主题是专家。

[0091] 继续图8的实例，基于针对给定主题用户被列出到另一用户的列表上的次数，服务器100可确定用户是否被其他用户视为专家。例如，用户B被列出在关于主题A的三个不同列表上；用户C被列出在两个不同列表上；并且用户D、用户G、用户H及用户I的每一者仅被列出在一个列表上。因此，在本实例中，用户B被视作主题A的最重要的专家，其次为用户C。

[0092] 转至图9，提供计算机可执行指令的实施例以确定给定用户被视作专家的主题。在框901处，服务器100获取给定用户被列出的列表组。在框902处，服务器100使用列表组以确定与给定用户有关的主题。在框903处，服务器输出给定用户被视作专家的主题。所述主题形成给定用户的专家知识向量。例如，如果用户Alice被列出在Bob的钓鱼列表、Celine的艺术列表及David的摄影列表，则Alice的专家知识向量包括：钓鱼、艺术及摄影。

[0093] 在实施例中，由于用户动态更新用户列表并且新列表经常被创建，所以通过不断

抓取所述列表来获取用户列表。在实施例中，使用Apache Lucene索引处理用户列表。使用Lucene算法处理给定用户的专家知识向量以填入与给定用户有关的主题的索引。所述索引支持(例如)全Lucene查询语法，包括短语查询与Boolean逻辑。借助上述背景，Apache Lucene为适合全文索引及搜索的信息检索软件库。Lucene也因其在实施Internet搜索引擎及本地单站搜索中的用途而广知。可以认识到，可使用其他当前所知或未来所知搜索及索引算法。

[0094] 在实施例中，图9的计算机可执行指令可由模块110执行。

[0095] 转至图10，通过计算机可执行指令的实施例以确定给定用户感兴趣的主题。在框1001处，服务器100获取给定用户关注的从属用户。

[0096] 在框1002处，针对每一从属用户执行大量指令。特别是，在框1003处，服务器获取列出从属用户的列表组(例如从属用户的专家知识向量)。在框1004处，服务器使用列表组以确定与从属用户有关的主题。框1004的输出为与从属用户有关的主题(框1005)。在实施例中，框1002可简单请求图9中提出的算法，应用于每一从属用户。

[0097] 在实施例中，在框1006处，服务器组合来自所有从属用户的主题。所组合的主题形成给定用户感兴趣的主题的输出1007(例如给定用户的兴趣向量)。

[0098] 在另一实施例中，框1006及1007的替代方式为确定哪些主题是从属用户之间共同或最共同的(框1008)。例如，给定用户Alice关注从属用户Bob、Celine及David。Bob被视作钓鱼与摄影(例如Bob的专家知识向量)的专家。Celine被视作钓鱼、摄影与艺术的专家(例如Celeine的专家知识向量)。David被视作钓鱼与音乐的专家(例如David的专家知识向量)。因此，由于钓鱼的主题是所有从属用户之间最共有的，因此识别Alice对钓鱼的主题有兴趣。或者，由于摄影的主题是从属用户之间较为共有的(例如钓鱼之后第二最共有主题)，则摄影的主题也被识别为Alice感兴趣的主题。由于艺术和音乐不是从属用户之间所共有的，因此，所述主题不被视为Alice感兴趣的主题。

[0099] 在实施例中，模块111执行图10中所示的计算机可执行指令。

[0100] 在实施例中，将来自专家知识向量的数据和来自兴趣向量的数据提供至Lucene算法用于编索引。

[0101] 转至图11，提供示例性计算机可执行指令用于搜索在索引存储117中被视为主题中的专家的用户。在框1101处，服务器获取用于查询的主题。在框1102处，服务器100识别将主题A(例如正在查询的主题)列入其专家知识向量的用户。在框1103处，服务器确定在所识别的用户中在与主题A有关的列表中出现最多次的用户。在框1104处，在列表中出现最多次的前n个用户为主题A的专家。换言之，服务器创建用户组U_T以包括前n个用户及其关注者。

[0102] 在包括图8至11中所述的原则的用于确定用户的另一实施例的中，关注者的最大范围可用于识别前n个用户最大范围计算确定有多少与用户组(例如专家、影响者)有关的唯一关注者。例如，如果第一专家和第二专家一共具有总数为200的唯一关注者，并且第二专家和第三专家一共具有总数为300的唯一关注者，则与第一专家和第二专家相比，第二专家和第三专家具有较大的关注者“范围”。转至图12，示例性计算机可执行指令用于识别对主题A感兴趣的用户，其可由模块114执行。在框1201处，例如，服务器100通过GUI中的用户输入以获取主题A。在框1202处，服务器(例如通过分析每一用户的兴趣向量)搜索对主题A感兴趣的用户。在框1203处，输出自框1202识别用户。

[0103] 为确定对主题A感兴趣的用户的最大范围，服务器确定n个用户的哪种组合提供最高数量的用户的唯一关注者(框1204)。将确定的前n个用户与其关注者一同输出(框1205)。换言之，主题网络中的用户U_T包括前n个用户及其关注者。

[0104] 可以认识到，用于识别有关主题的用户的其他已知及未来所知方式可用于其他实施例中。

[0105] 识别并过滤主题网络中的异常值用户：

[0106] 关于识别及过滤主题网络中的异常值节点(例如用户)，如框306及406中的每一者所示，可以认识到，可使用不同计算。以下为实施框306及406的非限制性实施例。

[0107] 应认识到，可通过移除有问题的异常值来改善来自主题网络的数据。例如，使用参考麦当劳咖啡品牌的主题“McCafe”的查询也会碰巧带回一些来自菲律宾的喜欢同样名字的卡拉OK吧/咖啡店的用户。由于他们碰巧为一个紧密的社群，所以其影响者分数通常足够高以在关键的前十名排行列表中排名。

[0108] 转至图13，提供了展示未过滤结果的主题网络1301的实施例的图解。节点表示关于主题McCafe的用户组U_T。一些节点1302或用户为来自菲律宾的喜欢同样名字McCafe的卡拉OK吧/咖啡店的用户。

[0109] 在测试案中有时会出现所述现象，并不限于主题McCafe的测试案。本文中认识到寻找McCafe用户并非同时寻找麦当劳咖啡和菲律宾的卡拉OK吧，因此所述子网络1302被视为杂讯。

[0110] 为完成杂讯的减少，在实施例中，服务器使用称为模块性的网络社群检测算法来识别并过滤主题查询中的所述类型的异常值集群。模块性算法描述于引用的Newman，M.E.J.(2006)名为“网络中的模块性与社群结构(Modularity and community structure in networks)”的文章中(美国科学国内进展研究(PROCEEDINGS-NATIONAL ACADEMY OF SCIENCES USA)103(23):8577-8696，文章在此以全文引用的方式并入)。

[0111] 可以认识到，其他类型的集群及社群检测算法可用于确定主题网络中的异常值。过滤有助于移除非计划的或由寻找与主题相关的影响者的用户搜寻之后的结果。

[0112] 如图14所示，相对于主题网络1301的主集群1402识别异常值集群1401。用户$U_0$1401的异常值集群自主题网络移除，且主集群1402中剩余的用户用于形成输出的影响者的排名列表。

[0113] 在实施例中，服务器100以下指令以过滤出异常值：

[0114] 1.对主题网络执行模块性算法。

[0115] 2.模块性功能将主题网络分解为模块化社群或子网络，并且将每一节点标注为X集群/社群中的一者。在实施例中，$X<N/2$，由于社群具有不止一个成员，因此N为组U_T中的用户。

[0116] 3.通过社群中的用户数量排序社群，并接纳最多人数的社群。

[0117] 4.当节点人数的累积和超过总数的80％，将剩余较小的社群自主题网络移除。

[0118] 关于图15描述用于识别及过滤主题网络的计算机可执行指令的一般实施例。可以认识到，所述指令可用于执行框306及406。

[0119] 在框1501处，服务器100将社群寻找算法应用于主题网络以将网络分解为社群。用于寻找社群的算法的非限制性实例包括最小切法、Hierarchical集群、Girvan-Newman算

法、以上参考的模块性算法及基于Clique的方法。

[0120] 在框1502处，服务器将每一节点(即用户)标注为X社群中的一者，其中X<N/2且N为主题网络中的节点的数量。

[0121] 在框1503处，服务器识别每一社群中的节点的数量。

[0122] 如果所述社群尚未添加至过滤的主题网络，服务器则将具有最大数量的节点的社群添加至过滤的主题网络(框1504)。可以认识到，最初过滤的主题网络包括0个社群，并且添加至过滤的主题网络的第一个社群为最大社群。来自未过滤主题网络的相同社群不能再次被添加至过滤的主题网络。

[0123] 在框1505处，服务器确定过滤的主题网络的节点数量是否超过或大于原始或未过滤主题网络的节点数量的Y%。在实施例中，Y%为80%。Y的其他百分比值也可行。如果没有超过，则过程循环返回至框1504。当框1505的条件为真实的，过程前进至框1506。

[0124] 通常，当过滤的主题网络的节点数量达到或超过未过滤主题网络的节点数量的多数百分比，则识别主集群并且还识别为异常值节点(例如U_0)的剩余节点。

[0125] 在框1506处，输出过滤的主题网络，其不包括异常值用户U_0。

[0126] 实例:McCafe案例研究

[0127] McCafe是由麦当劳创建的咖啡屋类型的食物和饮料品牌。其具有各种各样的菜单项目，例如咖啡、拿铁、意式浓咖啡及冰沙。针对“McCafe”使用本文中所描述的系统和方法的影响者结果展示于表2中。社交网络数据来自于Twitter。

按影响力排序的Twitter用户	权威分数	页面排名	按权威排序的Twitter用户	权威分数	页面排名
McCafe©	8	2.255%	McDonald's Corp.	10	1.682%
McDonald's Corp.	10	1.682%	McDonald's	10	0.959%
McDonald's Philly	6	1.478%	Divine Lee	10	0.558%
Marti	7	1.236%	Victor Basa	10	0.558%
McDonald's SoCal	7	1.174%	Tyler Fox-Banks	10	0.279%
The Mommy-Files	8	1.164%	McDonald's Venezuela	10	0.234%
McDonalds Eastern NE	6	1.091%	hashtags	10	0.203%
McDonaldsDMV	6	1.017%	GUYEL	10	0.136%
Rick Wion	7	1.012%	The Product Poet	10	0.107%
McDonald's	9	0.960%	Mia Farrow	10	0.074%
Canada					
McDonald's	10	0.959%	Maxene Magalona	10	0.065%
McDonalds NYTriState	8	0.916%	XIAN LIM	10	0.065%
Utah McDonald's	6	0.913%	Xeni Jardin	10	0.000%
Me Encanta	6	0.910%	Manado Kota	10	0.000%

表2.针对主题查询“McCafe”的按影响力分数及权威分数排序的最高排名Twitter句柄

[0128] 对所述结构有若干观察。

[0129] 影响力分数准确地将句柄McCafe列为针对查询的最具影响力者，而权威分数为8。其未出现在权威分数的第一页。

[0130] 许多本地/区域性的McDonald的句柄基于影响力排名很高但具有低于10的权威分数。

[0131] 具有为7的低权威分数的Rick Wion是基于影响力的第九高的排名用户。Rick Wion是McDonald的社交媒体互动副总裁，其明显为Twitter上的McCafe的影响者。

[0132] 权威分数列表中存在很多不适当的名字，其可能提过McCafe并具有许多关注者，但其明显不是影响者。

[0133] 以上观察证明当使用本文中所述的系统和方法时可得到更好质量的影响者结果。

[0134] 实例:Fanexpo案例研究

[0135] Fanexpo是加拿大多伦多市举办的漫画、科幻和奇幻娱乐的年度大会。针对主题查询“Fanexpo”的最高排名的影响者展示于表3左侧，基于权威分数的对照结果展示于右侧。使用本文中所述的系统及方法确定影响者。

按影响力排序的Twitter用户	权威分数	页面排名		按权威排序的Twitter用户	权威分数	页面排名
Fan Expo Canada	8	1.241%		Dark Horse Comics	10	0.749%
C.B. Cebulski	9	0.966%		Torontoist	10	0.778%
Silver Snail	7	0.822%		Michael Rooker	10	0.580%
SpaceChannel	8	0.790%		Amanda Tapping	10	0.563%
Torontoist	10	0.778%		National Post	10	0.432%
Dark Horse Comics	10	0.749%		CTV Toronto	10	0.322%
Mark Brooks	8	0.671%		CBC Top Stories	10	0.310%
Michael Shanks	9	0.661%		Nathan Fillion	10	0.358%
Katie Cook	8	0.659%		Brent Spiner	10	0.350%
Kelly Sue DeConnick	8	0.637%		Jessica Nigri	10	0.338%
Ramon Perez	7	0.632%		Meg Turney	10	0.132%
Shaun Hatton	7	0.627%		The Walking Dead	10	0.215%
Fearless Fred	9	0.614%		Eduardo Benvenuti	10	0.119%
Alice Quinn	7	0.583%		Randy Pitchford	10	0.118%

表3.针对主题查询“Fanexpo”的按影响力分数及权威分数排序的最高排名Twitter句柄

[0136] 当分析所述结果时可见若干有趣的观察。

[0137] 本文中所述的影响者方式准确列出句柄Fan Expo Canada为针对查询的最具影响力者，而权威方式给出8分的分数。

[0138] 第二排名的影响者，C.B.Cebulski，为在所述领域被视为非常有影响力的漫威漫

画的著名编剧。

[0139] 注意最高权威排名中，上述两位影响者(即Fan Expo Canada以及C.B.Cebulski)并没有出现在关键的第一页上。

[0140] 下面四位影响者，Silver Snail、SpaceChannel、Torontoist及Dark Horse Comics为多伦多的漫画存储库、科幻电视频道、多伦多娱乐博客以及漫画出版商。

[0141] 最高权威排名的大众新闻媒体National Post，CTV Toronto，CBC Top Stories为不适合所述主题的用户账户。

[0142] 其次的一系列影响者(例如Twitter账户名)不是漫威或DC漫画的编剧即使科幻或奇幻电影或电视剧的演员。注意其中许多人具有低于10的权威分数。

[0143] 此外，上述观察证明当使用本文中所述的系统和方法时可得到更好质量的影响者结果。

[0144] 实例:Nike Livestrong案例研究

[0145] Livestrong是由现在蒙受耻辱的自行车骑手兰斯·阿姆斯特朗建立的组织以资助癌症研究。在阿姆斯特朗因兴奋剂丑闻被起诉之后，Nike最近断绝了与Livestrong的关系。使用来自Twitter的社交网络数据查询“Nike Livestrong”的影响者结果展示于表4的右侧。使用权威方式得到的结果展示于右侧。

按影响力排序的Twitter用户	权威分数	页面排名	按权威排序的Twitter用户	权威分数	页面排名
Darren Rovell	10	0.63%	Darren Rovell	10	0.63%
The Associated Press	10	0.45%	The Associated Press	10	0.45%
Juliet Macur	8	0.40%	Nice Kicks	10	0.37%
Deadspin	10	0.37%	Deadspin	10	0.37%
Nice Kicks	10	0.37%	NBC Nightly News	10	0.32%
Joseph Weisenthal	9	0.34%	Jim Roberts	10	0.34%
Jim Roberts	10	0.34%	Bloomberg News	10	0.34%
Bloomberg News	10	0.34%	Sports Illustrated	10	0.32%
NBC Nightly News	10	0.32%	Business Insider	10	0.29%
Sports Illustrated	10	0.32%	CBSSports.com	10	0.28%
NYT Sports	9	0.29%	Complex	10	0.26%
Business Insider	10	0.29%	Cyclingnews.com	10	0.25%
CBSSports.com	10	0.28%	Fast Company	10	0.20%

表4.针对主题查询“Nike Livestrong”的按影响力分数及权威分数排序的最高排名Twitter句柄

[0146] 表4中有若干有趣之处。

[0147] 许多具有权威分数10的最具影响力者为体育新闻句柄或大量撰写阿姆斯特朗兴

奋剂丑闻的体育记者。

[0148] 特别是，Juliet Macur为基于影响力排名为第三，而她的权威分数为8。她是写了名为《连绵不绝的谎言：兰斯·阿姆斯特朗的坠落(Cycle of Lies:the Fall of Lance Armstrong)》的书的纽约时报体育记者。

[0149] Joseph Weisenthal是发了有关Nike Livestrong合伙关系的兴奋剂丑闻的推文的体育行业知情人。

[0150] 虽然难以用权威分数10来将所有的Twitter用户账户区分开，但是影响力排名为影响者的相关排名提供了更多的特异性。

[0151] 如特定与社群识别、大众化特征识别及其每一社群中的值相关的图3及图4中描述的方法步骤的更多细节以及结果的显示描述于下方。

[0152] 识别社群

[0153] 转至图16，展示计算机可执行指令的实施例以自社交网络数据识别社群。

[0154] 社交网络平台的特征在于用户正在关注(或定义为朋友)另一用户。如早前所述，其他类型的关系或互联性可存在于主题网络内多个节点和边界说明的用户之间。在主题网络内，影响者可影响不同集群的用户至各种程度。即，基于根据图16中描述的识别社群的过程，服务器被配置为识别单一主题网络内的多个集群，称为社群。由于影响力并非均匀地跨越社交网络平台，根据图16定义的社群识别过程为有利的，因为其识别跨越主题网络的每一影响者(例如将一个社群与另一个社群有关)的影响力的程度或深度。

[0155] 将定义于图16中，服务器被配置为提供不同社群组(例如C1，…，Cn)，以及每一社群中的一个或多个最具影响力者。在又一优选方面，服务器被配置为提供跨越所有社群的最具影响力者的合计的列表以提供所有影响者的相关顺序。

[0156] 在步骤1601，服务器被配置为如前所述自社交网络数据获取主题网络图信息(例如图3至图4)。主题网络可视地说明节点之间的关系，用户组(U_T)中的每一用户表示为主题网络图中的节点并通过边界连接以表示主题网络图中的两个用户之间的关系(例如朋友或关注者与被关注者，其他社交媒体互联性)。在框1602处，服务器获取用于定义社群之间的界限的内部及/或外部互联性(例如解析度)的预定义程度或衡量值。

[0157] 在框1603处，服务器被配置为根据互联性(例如解析度)的预定义程度计算节点(例如影响者)和边界的每一者的评分。即，在一个实例中，每一用户句柄分配有模块性等级标识符(Mod ID)及页面排名分数(定义影响力的程度)。一方面，解析度参数被配置为控制识别的社群的密度和数量。在优选方面，服务器使用提供2至10个社群的为2的默认解析度值。在另一方面，用户定义解析度值(例如通过图2中的计算装置101)以根据社群信息可视化的需要而产生更高或更低的社群粒度。

[0158] 在框1604处，服务器被配置为定义并输出不同社群集群(例如C_1，C_2，…，C_n)从而将用户U_T分区为U_{C1}至U_{Cn}以使得将网络中的节点定义的每一用户映射至相应社群。一方面，模块性分析用于定义社群以使得每一社群在社群中的节点的集群之间具有密集联系(高连通性)但与不同社群中的节点稀疏联系(低连通性)。一方面，可使用模块性算法及/或密度算法(其衡量内部连通性)实施社群检测过程步骤1603-1606。此外，在一个方面使用Gephi、开源图分析包及/或javascript库实施结果的可视化。

[0159] 在框1605处，服务器被配置为定义及输出跨越所有社群的最具影响力者及/或每

一社群中的最具影响力者并提供所有影响者的相对顺序。一方面，当选定特定社群时，将最具影响力者与其社群可视地并列显示。在另一个方面，在框1605处，服务器被配置为提供跨越所有社群的所有最具影响力者的合计的列表以提供所有影响者的相对顺序。

[0160] 在框1606处，服务器被配置为可视地描述与区分每一社群集群(例如通过色彩编码或其他可视识别以将社群彼此区分开)。在另一个方面，在框1606处，服务器被配置为提供每一社群中的最具影响力者组，其可视地链接至相应社群。在另一个方面，在框1606处的服务器，服务器被配置为改变社群图的每一节点的大小以对应相应影响者的分数(例如影响力分数)。作为自框1606的输出，自节点的边界展示其社群中及跨越其他社群的每一用户之间的联系。

[0161] 因此，如将在图19A至19C及20A至20B中所示，社群及影响者的可视化(例如每一社群中的排名的最具影响力者及/或跨越所有社群的最具影响力者的列表)允许终端用户(例如图2中的计算装置101的用户)视觉化其相关社群中的每一影响者的等级及相对重要性。

识别给定社群中的大众化特征

[0162] 如根据图3及图4所述，在另一方面，服务器被配置为针对框1603提供的每一给定社群(例如C_1)基于社交网络数据确定与给定社群(例如C_1)中的用户(例如U_{C1})有关的预定义特征(例如共同关键词及短语、会话主题、共同位置、共同图像、共同元数据)的大众化特征值。因此，可通过检查每一社群C_1中的用户U_{C1}的预定义特征组(例如会话主题)定义趋势或共性。一方面，与每一社群有关的特征值的最高列表(例如每一社群的所有用户之间的最高会话主题)描绘于在框1605，并输出至计算装置101(展示于图2中)用于显示。

显示社群及大众化特征

[0163] 参见图17A至图17D，展示由服务器的GUI模块106提供并输出至计算装置的显示屏125(图2)的截屏以用于可视化来自主题网络的社群集群及可视化每一社群中的大众化特征。如图17A至图17D中所示，服务器提供交互接口以选择主题网络/特定社群中的社群及/或节点以可视地展现关于、每一节点(例如用户、社群信息及影响力程度)的细节。因此，图17A至图17D展示影响者社群及其特征(例如WordCloud可视化技术中的每一社群的会话)的交互可视化。还如图17A至图17D中所示，每一社群(例如由边界和节点组成)可视地与另一社群区分开(例如通过色彩编码)并且根据整个主题网络中的影响力程度调节每一节点尺寸。此外，通过选择特定社群(例如使用鼠标或指针从主题网络中可视地选择社群)，随后描述社群值(例如标亮在所述主题网络图中的社群，展现社群中最具影响力者并展现针对所选社群的最高会话主题的大众化特征值)。在图17A至图17D中，大众化特征值在显示屏(例如图2中的计算装置101的屏幕)上的可视化展示为描述所选社群中的最高会话主题的字云以及特定社群的所有用户中的每一主题使用的频率的指示。

[0164] 参见图17A，展示(例如图2中的计算装置101的)画面1701，说明在主题搜索(例如搜索术语“阿迪达斯”(adidas))期间，存在发生在社交网络的数个社群(集群、节段)中的多个会话。

[0165] 参见图18，展示画面，说明在另一主题搜索期间，主题网络具有每一者可视地相互区别的多个社群集群以及节点，所述节点调整尺寸以反映优选在整个主题网络中的影响力程度。

[0166] 参见图17B，展示画面1702，其描述节点经彩色编码以可视地与其相应社群有关，

并且每一节点的尺寸与相对于整个主题网络(经彩色编码)的其社群中的影响者分数成比例。图17B进一步展示通过选择节点(例如在节点上悬停鼠标指针),弹出Twitter句柄(例如adidasrunning),并且针对所述句柄的信息显示于画面1702上(例如信息下方的右边列表)。

[0167] 参见图17C,展示画面1703,以及选择子图,所述子图可视地标亮所选社群中的最具影响力者,并在画面1703视觉表示(例如所述社群中的会话的字云)。如图17中所示,展示了对社群行为的洞察以及积极/消极情绪。

[0168] 参见图17D,展示画面1704,其中,(例如通过图2中的计算装置101由用户输入选择)选择社群(例如社群1),并且社群中的最具影响力者可视地描述于标亮的主题网络旁,以显示所选社群。图17D展示(使用页面排名)针对社群检测(例如模块性)及影响力的高级网络分析的示例性用法。图17A至17D中的方式的有益之处在于其允许大规模处理社交网络数据(例如全Twitter、流水型(Firehose)),而非取样社交网络数据,所述取样可能遗漏影响者的小型但潜在重要的社群。

[0169] 定义社群中的大众化特征(例如会话主题)

[0170] 参见图19A至19C、20A及20B,展示两个不同主题网络中的各种影响者社群的示例性截屏(例如分别的阿迪达斯及多芬)。如图所示,当每一社群的用户句柄的身份可对社群的人口统计资料给出一些见解,需要展示更具体的社群描述。因此,一方面(例如图3及图4的示例性实施),识别自主题搜索查询返回的维特的样品,并且产生相关术语的频率计数以产生每一社群的会话中的大众化术语的字云。使用所述可视化,可因此易于可视地识别每一社群的行为特征,并且使用所述信息制作更具针对性的消息至每一社群的影响者。

[0171] 图19A至19C、20A及20B展示用于确定及视觉化主题网络中的社群集群以及针对每一社群的有关的大众化特征值的示例性实施(例如图3或图4的示例性实施)。根据一个实施,图19A至19C、20A及20B使用从Sysomos(社交媒体分析和监测公司)系统获取的潜在的Twitter数据,在一个示例性实施中,所述数据由跨越整个特定时间段的布尔型(Boolean)关键字搜索术语的用户定义列表形成。

[0172] 实例:Adidas Running案例研究——图19A至19C

[0173] 图19A至19C中的较暗阴影组分别对应“阿迪达斯跑步”主题中的三个最大社群。图19A中的标亮社群(蓝色)对应影响者的最大集。

[0174] 由图19A可见,字云及用户句柄说明所述社群中的会话似乎是关于阿迪达斯运动鞋和鞋子。

[0175] 在图19B中,第二大社群(橘色)具有关于阿迪达斯训练用运动教练(Micoach)智能手表的会话。在所述社群中还有许多配套审视句柄,例如Engadget、CNET、Mashable、FastCompany及Gizmodo。

[0176] 图19C中,主要的阿迪达斯跑步句柄是所述较小社群的一部分(绿色),其具有重要的句柄,例如YohanBlake、RunBlogRun、LondonMarathon、B_A_A(Boston Athletic Association)、RunningNetwork等。

[0177] 在审视图19A至19C中社群及其特征的可视化画面之后,可见阿迪达斯跑步可较好联系至重要跑步社群(绿色),但非较好联系至运动鞋迷(蓝色)及配套审视(绿色)社群的较大影响者社群。因此,可确定对于有效影响者市场营销,阿迪达斯跑步应联系其他社群中的

主要影响者，并且其消息能经剪裁以发送至其他社群，以与其他社群具有更好的重叠和联系。

[0178] 实例：多芬(Dove)案例研究

[0179] 图20A及20B展示较暗底纹中多芬(肥皂)产品主题的两个最大社群。图20A具有较低影响者的最大社群(蓝色)。由图20A及20B的用户句柄及字云可见地得知，用户句柄及字云反映有影响力的用户/有影响力的主题看起来是对储蓄、购物、取胜、奖品、克罗格(Kroger(超市))感兴趣的“妈咪博客”。

[0180] 同样，多芬的“无人能挡的女孩”活动也在所述社群中具有影响力。

[0181] 图20B描述具有官方多芬公司句柄(多芬加拿大、多芬英国、联合利华等)及一些半影响力的美妆博客的小型社群。

[0182] 因此，回顾图20A及20B之后，可见地得知，当多芬(作为主题查询)与有影响力的美妆博客良好地联系，由于与美妆博客相比，妈咪博客为更大的社群，其与所述妈咪博客之间可存在更强的联系。再次，可在不隔离其他人的情况下，以不同的方式裁剪消息至所述社群中的影响者。

[0183] 因此，如参考附图(例如图2、3、4及16-20b)所论述的，提出(基于获取的社交网络数据)针对给定查询主题识别其社交社群中的影响者的系统和方法。同样可见影响者不具有统一的特征，并且实际上在给定主题网络中甚至存在影响者的社群。本文所提出的系统及方法用于输出网络图中的可视的可视化于计算装置(例如计算装置101)上以显示实体或个人的相对影响者及其相应社群。额外大众化特征值(例如基于预定义特征，如会话的主题)可视地表述于计算装置的显示屏为每一社群展示最高或相关主题。可将主题描述为每一社群的会话的字云以可视地展现个人的社群的行为特征。

[0184] 所述方法及系统的大体实例提供如下。

[0185] 在一个实施例中提供一种由服务器执行以确定对主题有影响力的一个或多个用户的方法。所述方法包括：获取主题；确定社交数据网络内与所述主题相关的用户；将每一所述用户建模为节点并确定每一所述用户之间的关系；使用作为节点的所述用户及作为边界的所述关系以计算主题网络图；将所述主题网络图中的所述用户排名；识别并过滤所述主题网络图中的异常值节点；以及根据其有关的排名输出所述主题网络图中剩余的用户。

[0186] 在一个示例方面，消耗和产生包括所述主题的内容中的至少一者的所述用户被视为与所述主题相关的所述用户。

[0187] 在另一示例方面，在所述主题网络图中，定义于至少两个用户之间的边界表示所述至少两个用户之间的朋友联系。

[0188] 在另一示例方面，在所述主题网络图中，定义于至少两个用户之间的边界表示所述至少两个用户之间的关注者与被关注者联系，且其中所述至少两个用户中的一者为关注者及所述至少两个用户中的另一者为被关注者。

[0189] 在另一示例方面，在所述主题网络图中，定义于至少两个用户之间的边界表示所述至少两个用户之间的回复联系，且其中所述至少两个用户中的一者回复所述至少两个用户中的另一者发布的帖。

[0190] 在另一示例方面，在所述主题网络图中，定义于至少两个用户之间的边界表示所述至少两个用户之间的转发联系，且其中所述至少两个用户中的一者转发所述至少两个用

户中的另一者发布的帖。

[0191] 在另一示例方面，所述排名包括使用页面排名算法以衡量所述主题网络图中给定用户的重要性。

[0192] 在另一示例方面，所述排名包括使用以下至少一者：特征向量中心、权度、中间性及中心及权威性度量。

[0193] 在另一示例方面，识别及过滤所述主题网络图中的异常值节点包括：将集群算法、模块性算法及社群检测算法中的至少一者应用于所述主题网络图以输出多个社群；通过所述多个社群中的每一者中的若干用户将所述多个社群排序；选择n个具有最大数量用户的社群，其中所述n个社群中的所述用户的累积和至少达到所述主题网络图中的用户总数的百分比阈值；以及将未选的社群中的用户建立为所述异常值节点。

[0194] 在另一实施例中提供一种用于确定对主题有影响力的一个或多个用户的计算系统。所述计算系统包括：通信装置；存储器；以及处理器。所述处理器其配置为至少：获取主题；确定社交数据网络内与所述主题相关的用户；将每一所述用户建模为节点并确定每一所述用户之间的关系；使用作为节点的所述用户及作为边界的所述关系以计算主题网络图；将所述主题网络图中的所述用户排名；识别并过滤所述主题网络图中的异常值节点；以及

[0195] 将认识到，如本文中所描述的系统和方法的实施例的不同特征可以用不同的方式相互组合。换言之，尽管没有具体陈述，但根据其他实施例，不同的模块、操作和部件可以一起使用。

[0196] 本文中描述的所述流程图中的步骤或操作仅是示例。在不脱离本发明或这些发明的精神的情况下，这些步骤或操作可以有许多变化。例如，这些步骤可以按不同的顺序进行，或者可以被添加、删除或修改。

[0197] 本文中描述的GUIs及截屏仅是示例。在不脱离本发明或这些发明的精神的情况下，这些图形化的和交互式的部件可以有许多变化。例如，可将所述部件置于不同位置、或被添加、删除或者修改。

尽管已经参照某些特定实施例对以上内容进行了描述，但在不脱离所附权利要求书的范围的情况下，其各种修改对于本领域的技术人员而言将是明显的。

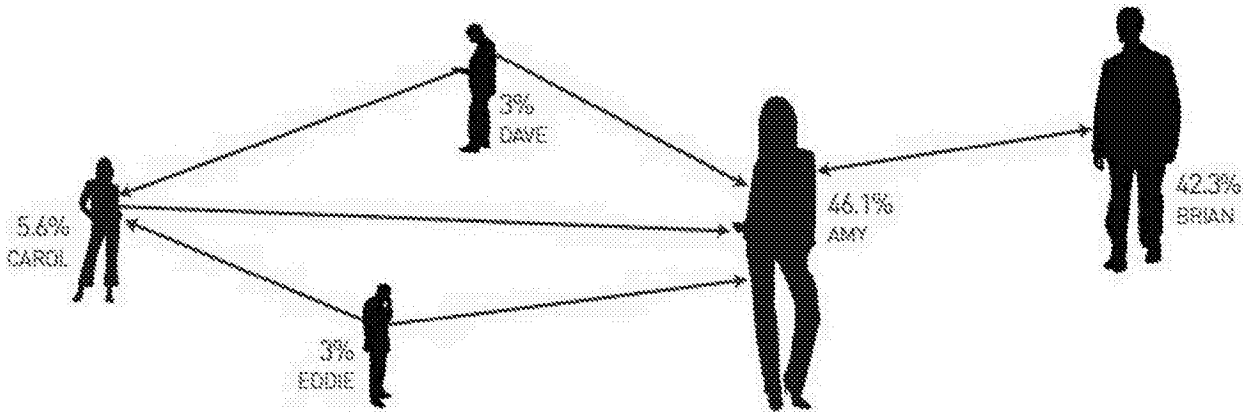


图1

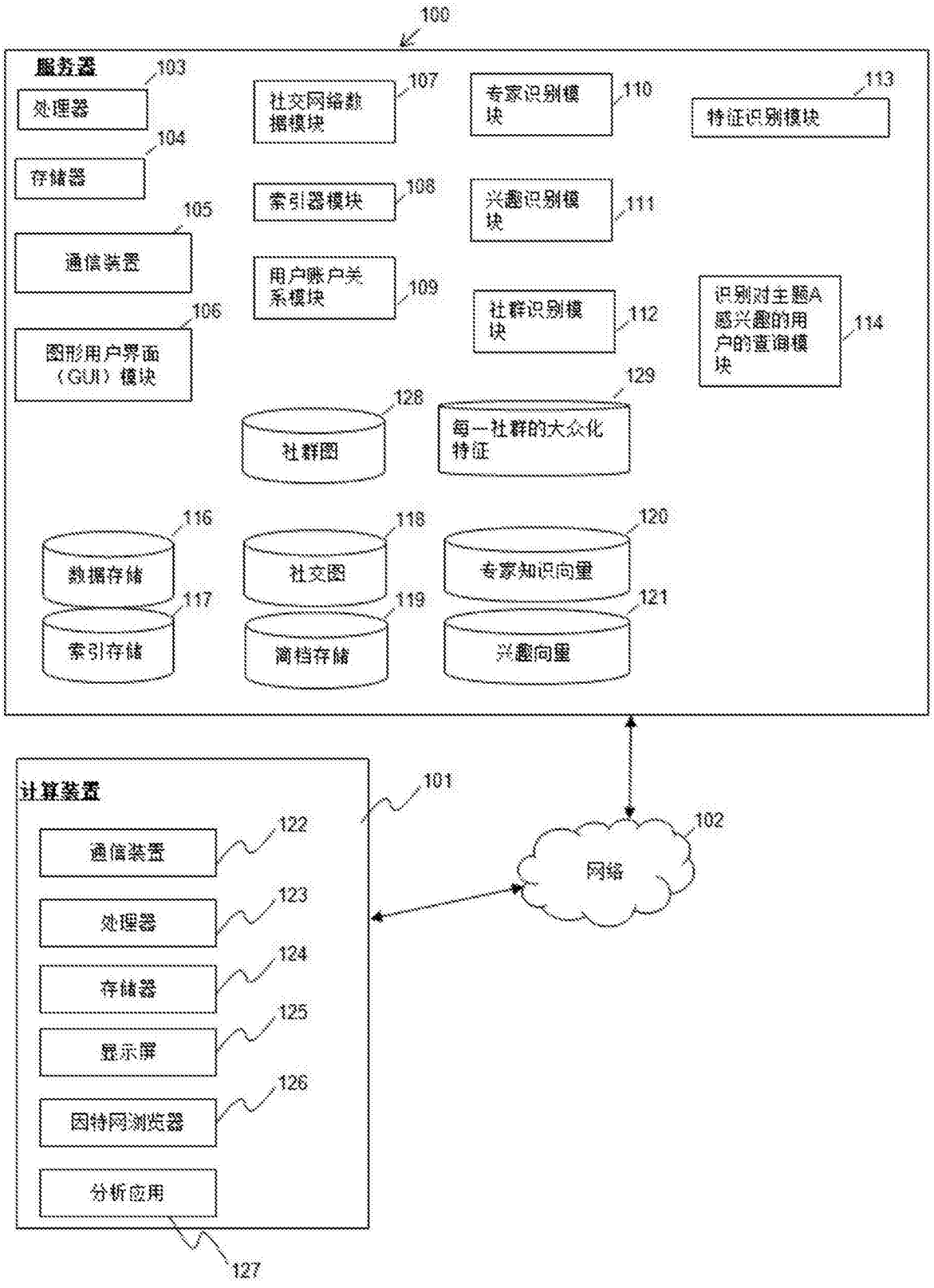

100
服务器
103
处理器
104
存储器
105
通信装置
106
图形用户界面（GUI）模块
社交网络数据模块
107
索引器模块
108
用户账户关系模块
109
专家识别模块
110
兴趣识别模块
111
社群识别模块
112
特征识别模块
113
识别对主题A感兴趣的用户的查询模块
114
社群图
128
每一社群的大众化特征
129
数据存储
116
索引存储
117
社交图
118
简档存储
119
专家知识向量
120
兴趣向量
121
计算装置
101
通信装置
122
处理器
123
存储器
124
显示屏
125
因特网浏览器
126
分析应用
127
网络
102

图2

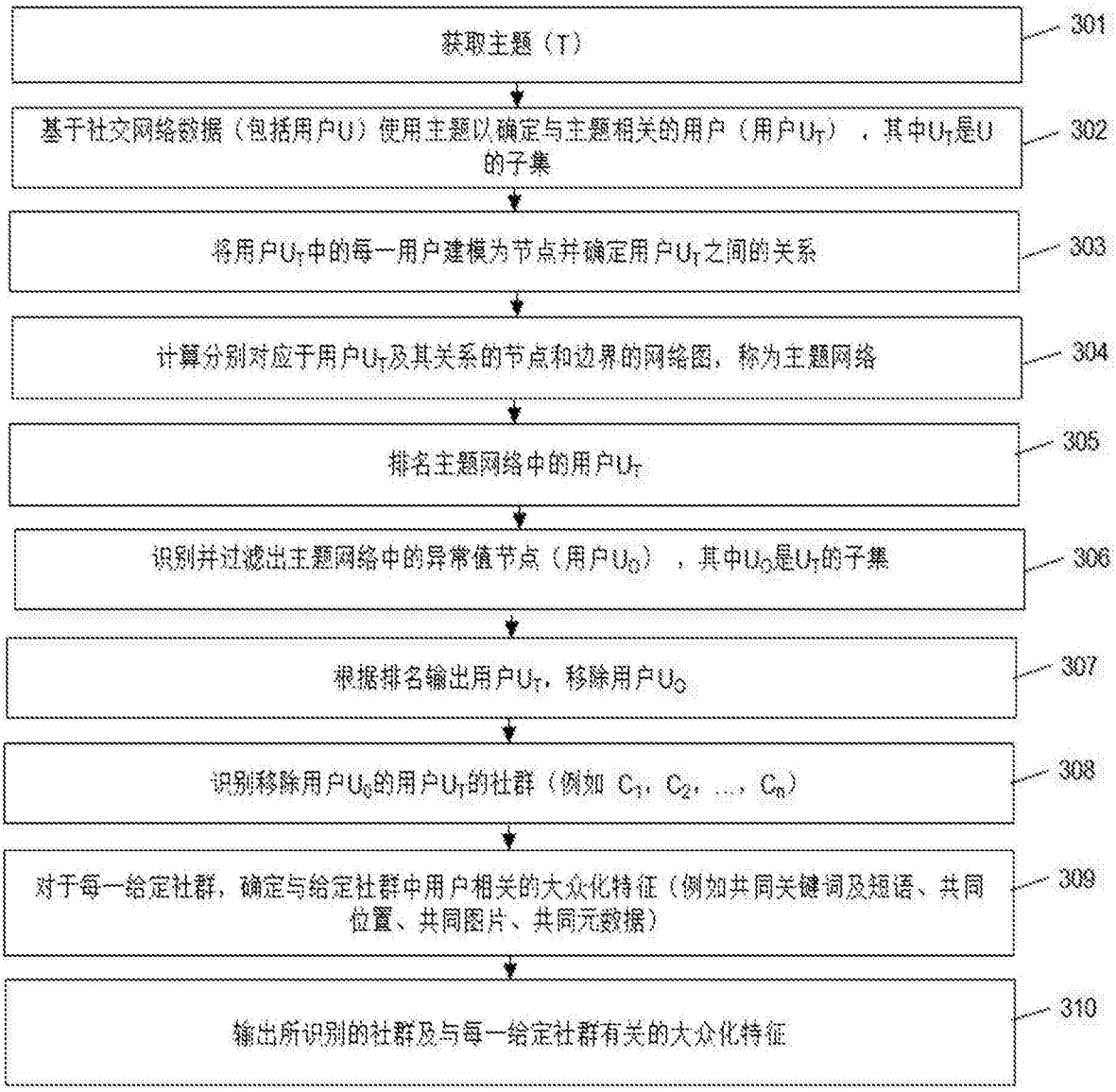

获取主题（T）
301
基于社交网络数据（包括用户U）使用主题以确定与主题相关的用户（用户U_T），其中U_T是U的子集
302
将用户U_T中的每一用户建模为节点并确定用户U_T之间的关系
303
计算分别对应于用户U_T及其关系的节点和边界的网络图，称为主题网络
304
排名主题网络中的用户U_T
305
识别并过滤出主题网络中的异常值节点（用户U_O），其中U_O是U_T的子集
306
根据排名输出用户U_T，移除用户U_O
307
识别移除用户U_O的用户U_T的社群（例如 C_1，C_2，…，C_n）
308
对于每一给定社群，确定与给定社群中用户相关的大众化特征（例如共同关键词及短语、共同位置、共同图片、共同元数据）
309
输出所识别的社群及与每一给定社群有关的大众化特征
310

图3


获取主题（T）
401
基于社交网络数据（包括用户U）使用主题以确定与主题有关的用户（用户U_T），其中U_T是U的子集
402
将用户U_T中的每一用户建模为节点并确定用户U_T之间的关系
403
计算分别对应于用户U_T及其关系的节点和边界的网络图，称为主题网络
404
使用第一排名过程排名主题网络中的用户U_T
405
识别并过滤出主题网络中的异常值节点（用户U_O），其中U_O是U_T的子集
406
在移除用户U_O的情况下使用第二排名过程调整用户U_T的排名，所述第二排名过程是基于某一时间段内来自用户的发帖的数量
407
根据调整的排名输出用户U_T，移除用户U_O
408
识别移除用户U_O的用户U_T的社群（例如C_1，C_2，…，C_n）
409
对于每一给定社群，确定与给定社群中用户相关的大众化特征（例如共同关键词及短语、共同位置、共同图片、共同元数据）
410
输出所识别的社群及与每一给定社群有关的大众化特征
411

图4

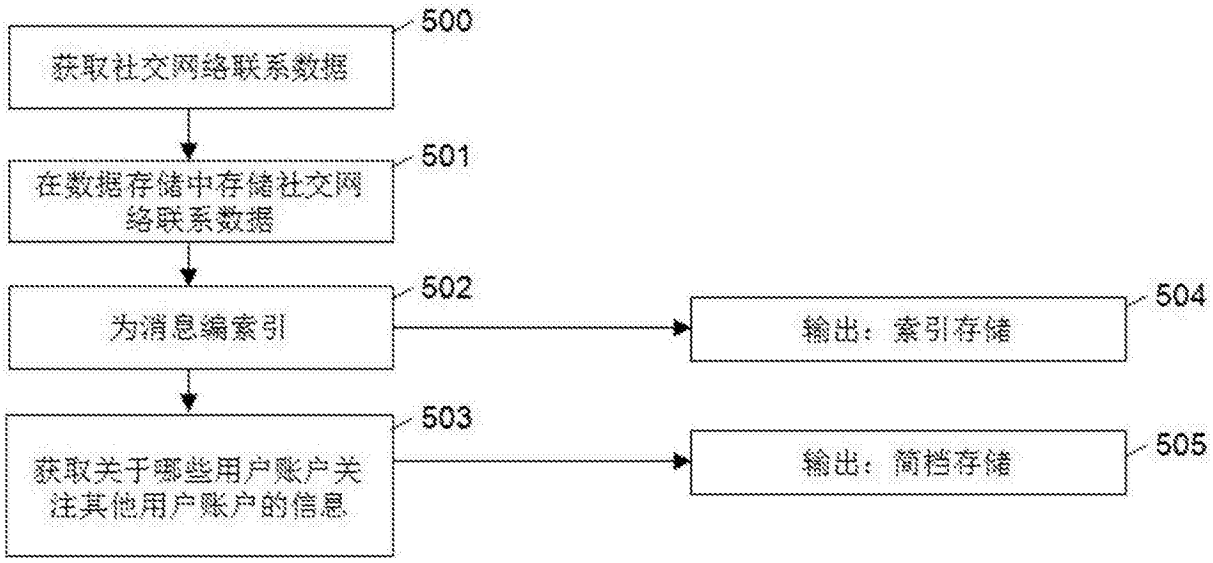

500
获取社交网络联系数据
501
在数据存储中存储社交网络联系数据
502
为消息编索引
503
获取关于哪些用户账户关注其他用户账户的信息
504
输出：索引存储
505
输出：简档存储

图5

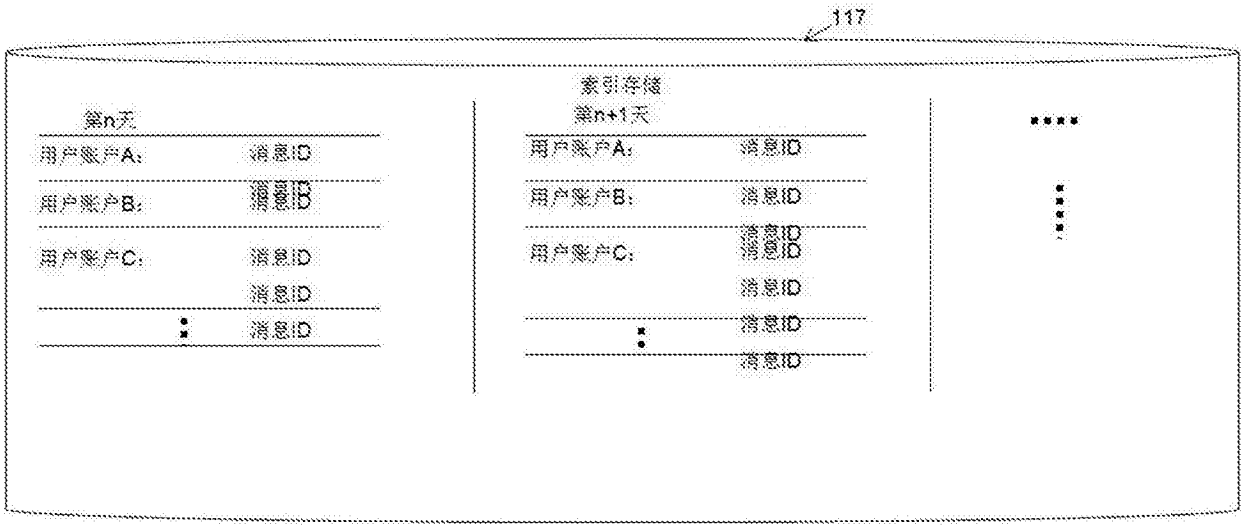

117
索引存储
第n天
第n+1天
用户账户A：
用户账户B：
用户账户C：
消息ID

图6

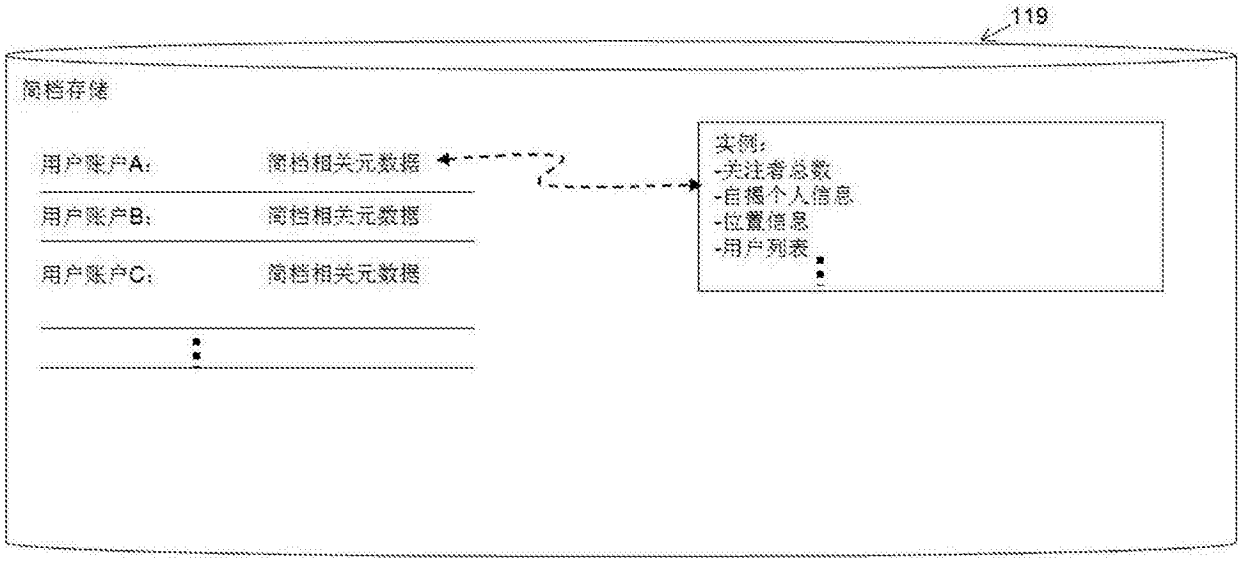

119
简档存储
用户账户A:
简档相关元数据
用户账户B:
简档相关元数据
用户账户C:
简档相关元数据
实例:
-关注者总数
-自我个人信息
-位置信息
-用户列表

图7

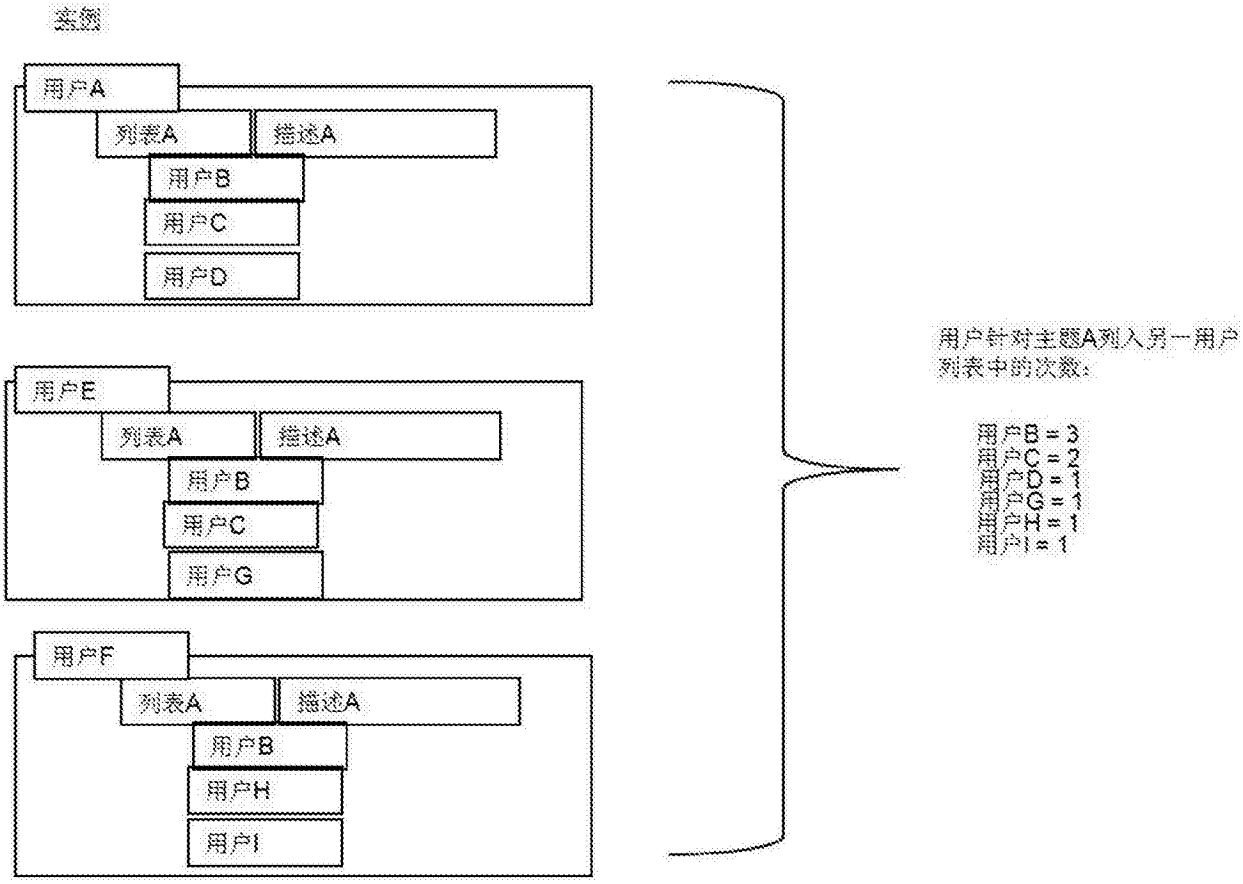

实例
用户A
列表A
描述A
用户B
用户C
用户D
用户E
列表A
描述A
用户B
用户C
用户G
用户F
列表A
描述A
用户B
用户H
用户I
用户针对主题A列入另一用户列表中的次数:
用户B = 3
用户C = 2
用户D = 1
用户G = 1
用户H = 1
用户I = 1

图8

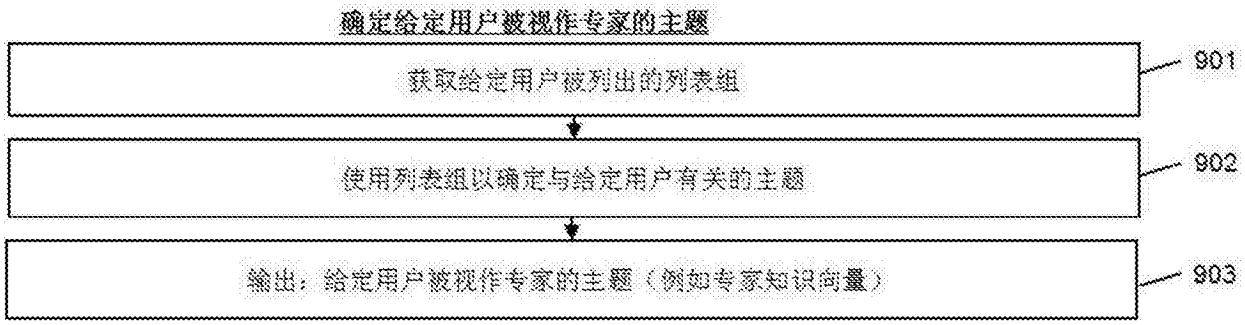

确定给定用户被视作专家的主题
获取给定用户被列出的列表组
901
使用列表组以确定与给定用户有关的主题
902
输出：给定用户被视作专家的主题（例如专家知识向量）
903

图9

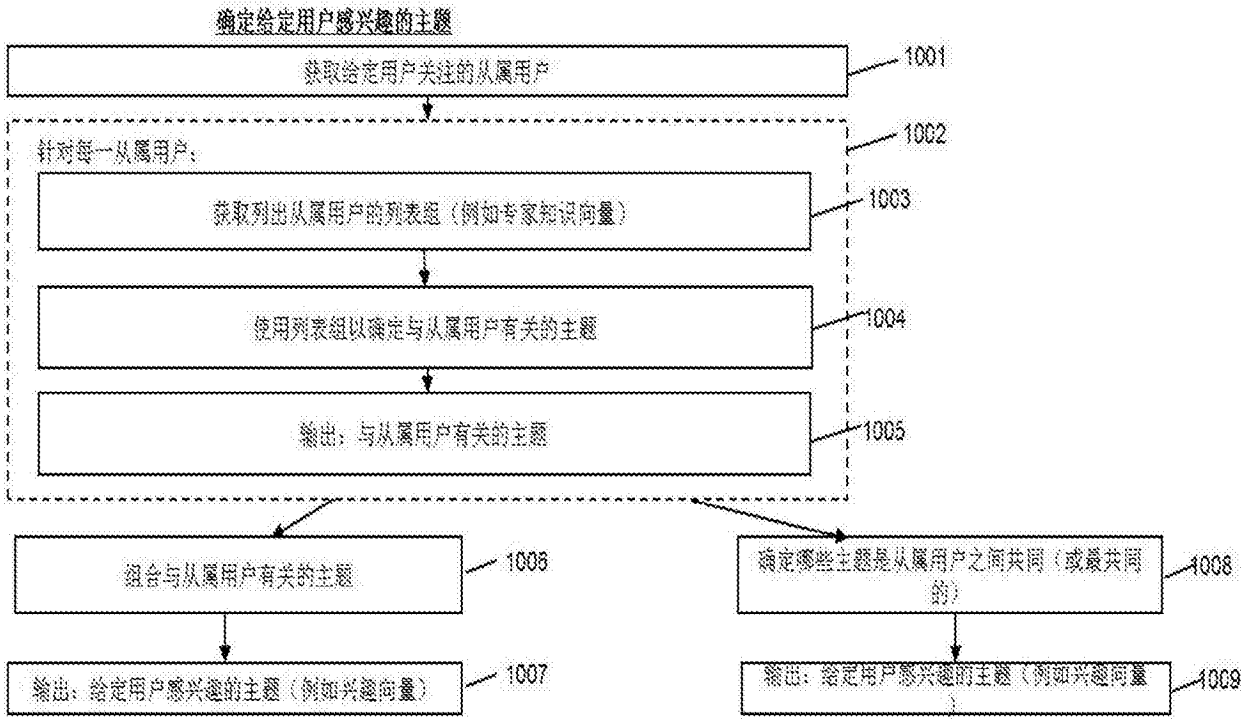

确定给定用户感兴趣的主题
获取给定用户关注的从属用户
1001
针对每一从属用户：
1002
获取列出从属用户的列表组（例如专家知识向量）
1003
使用列表组以确定与从属用户有关的主题
1004
输出：与从属用户有关的主题
1005
组合与从属用户有关的主题
1006
确定哪些主题是从属用户之间共同（或最共同的）
1008
输出：给定用户感兴趣的主题（例如兴趣向量）
1007
输出：给定用户感兴趣的主题（例如兴趣向量）
1009

图10

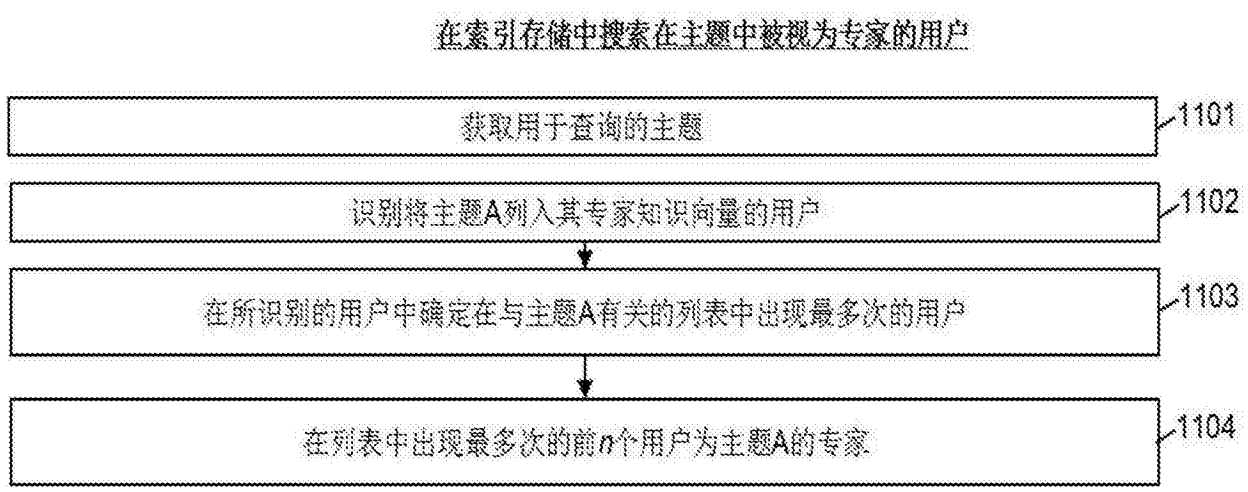

在索引存储中搜索在主题中被视为专家的用户
获取用于查询的主题
1101
识别将主题A列入其专家知识向量的用户
1102
在所识别的用户中确定在与主题A有关的列表中出现最多次的用户
1103
在列表中出现最多次的前n个用户为主题A的专家
1104

图11

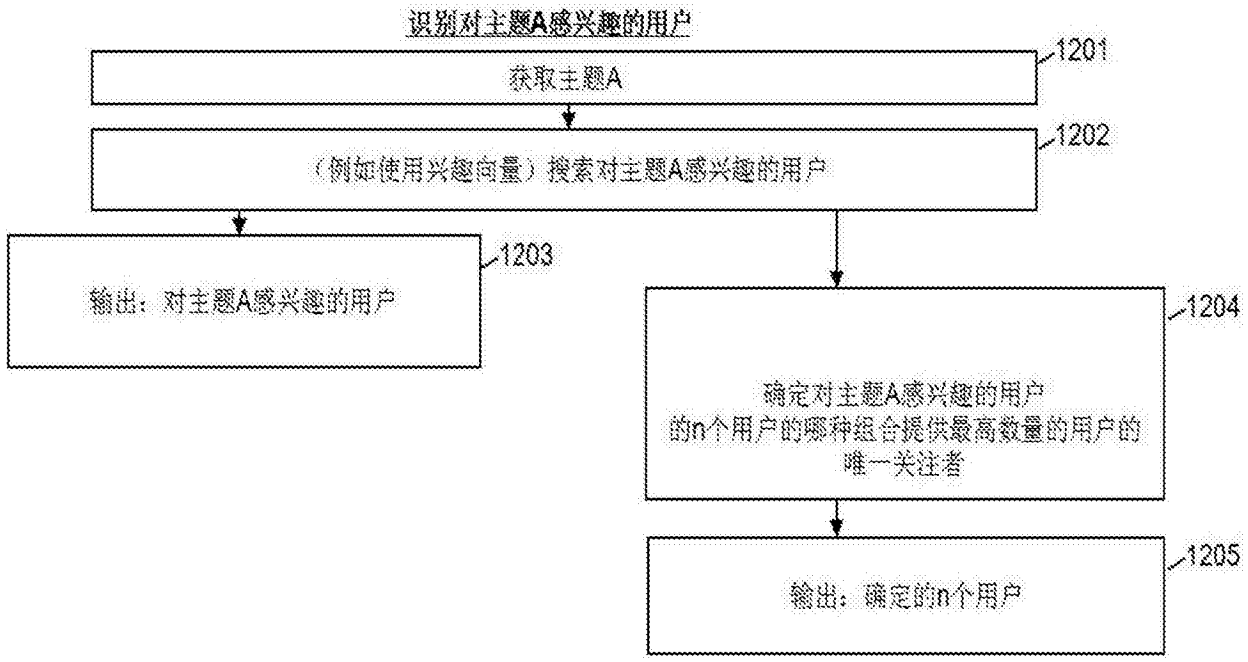

识别对主题A感兴趣的用户
获取主题A
1201
（例如使用兴趣向量）搜索对主题A感兴趣的用户
1202
输出：对主题A感兴趣的用户
1203
确定对主题A感兴趣的用户
的n个用户的哪种组合提供最高数量的用户的
唯一关注者
1204
输出：确定的n个用户
1205

图12

实例：主题“McCafe”的主题网络

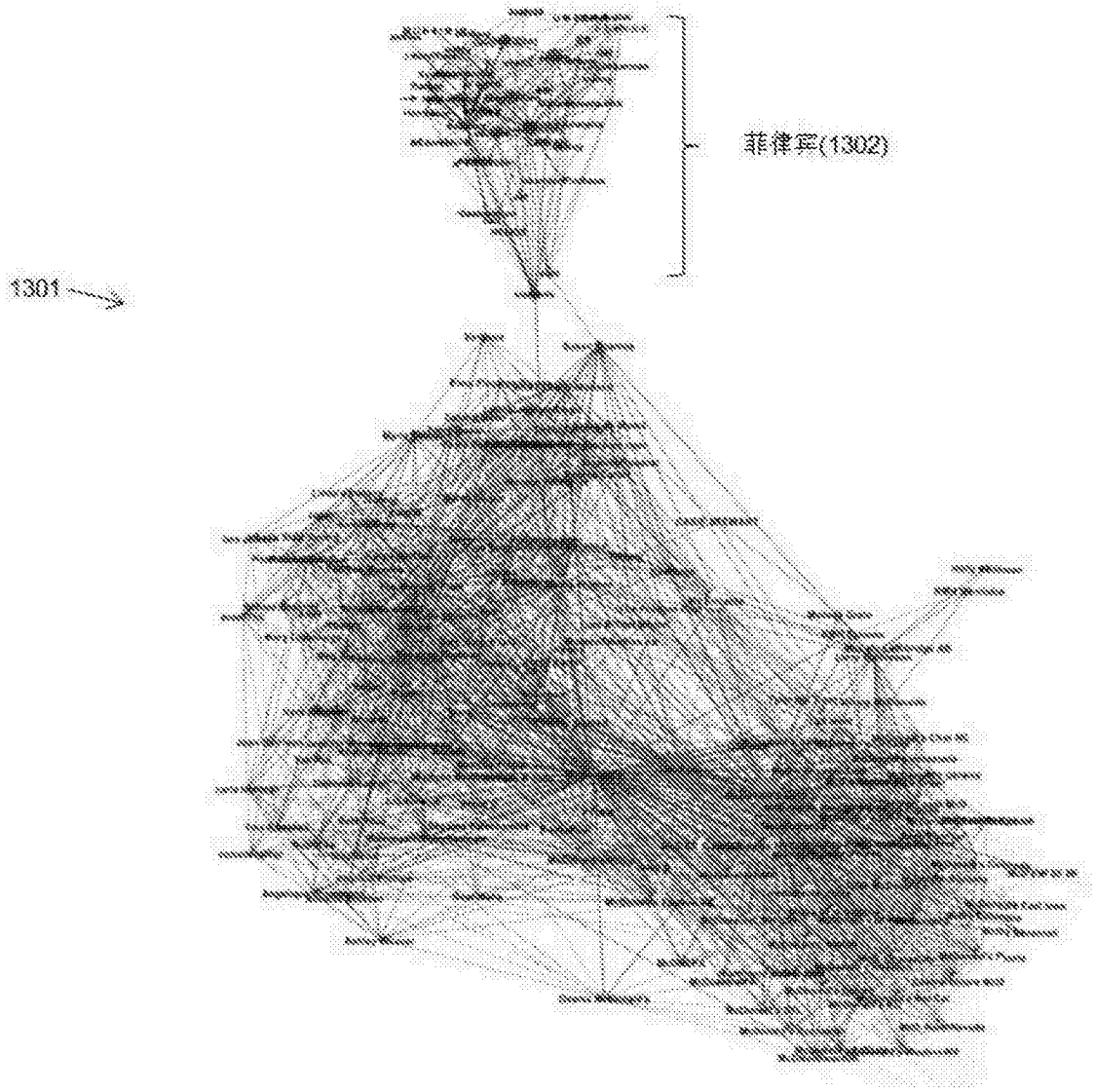


图13

实例：包括集群的主题“McCafe ”的主题网络

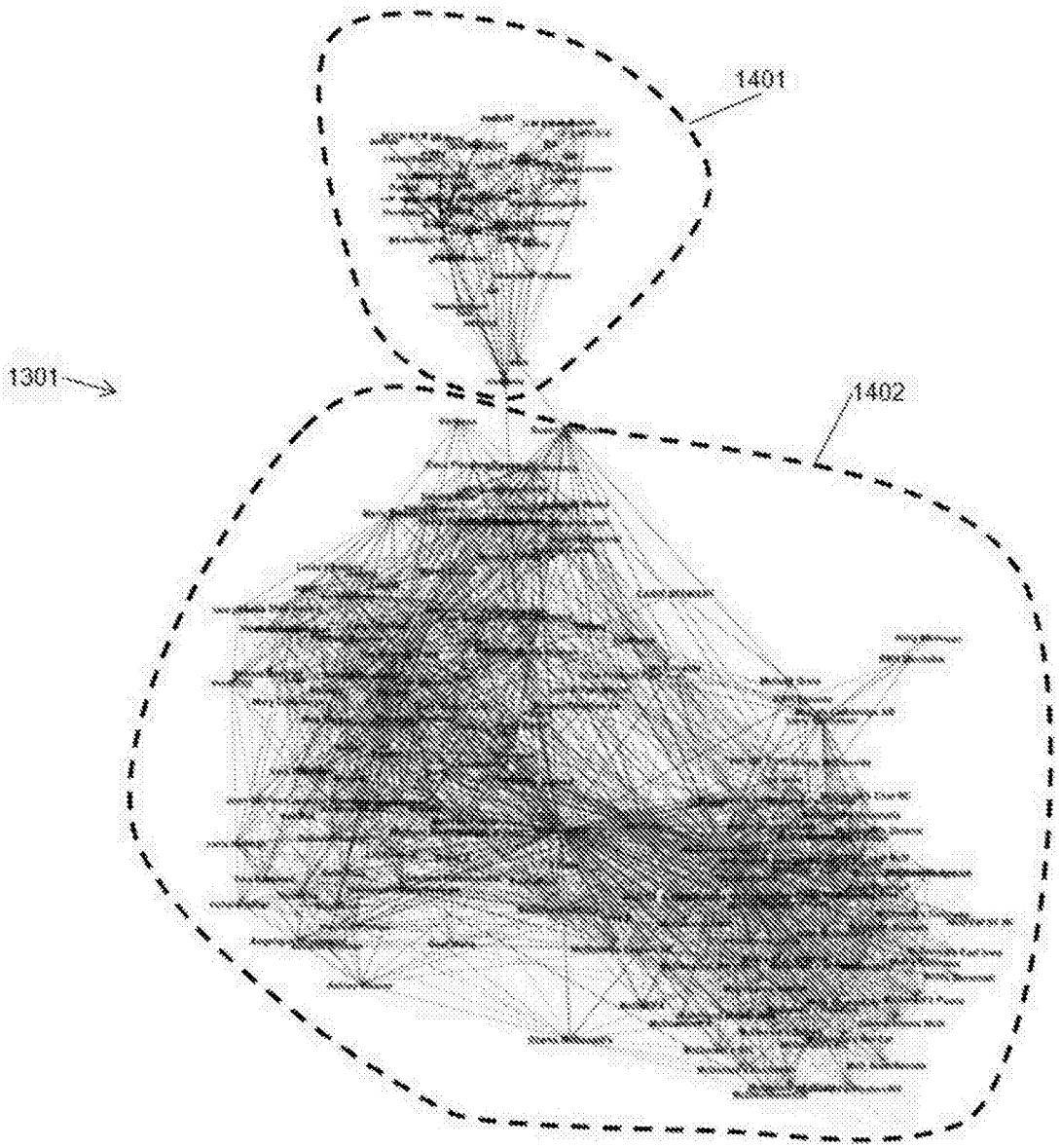


图14

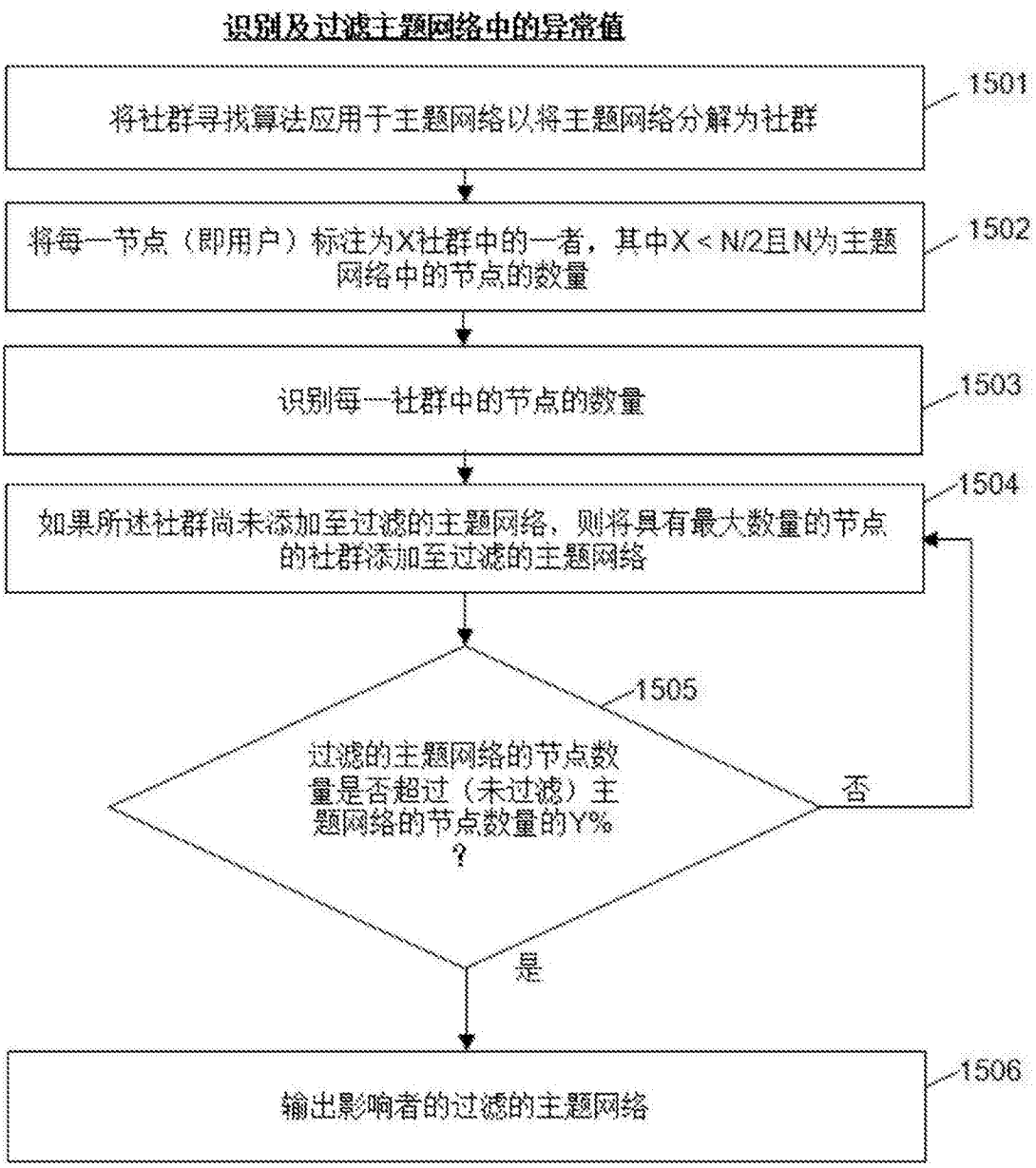

识别及过滤主题网络中的异常值
将社群寻找算法应用于主题网络以将主题网络分解为社群
1501
将每一节点（即用户）标注为X社群中的一者，其中X < N/2且N为主题网络中的节点的数量
1502
识别每一社群中的节点的数量
1503
如果所述社群尚未添加至过滤的主题网络，则将具有最大数量的节点的社群添加至过滤的主题网络
1504
过滤的主题网络的节点数量是否超过（未过滤）主题网络的节点数量的Y%？
1505
否
是
输出影响者的过滤的主题网络
1506

图15

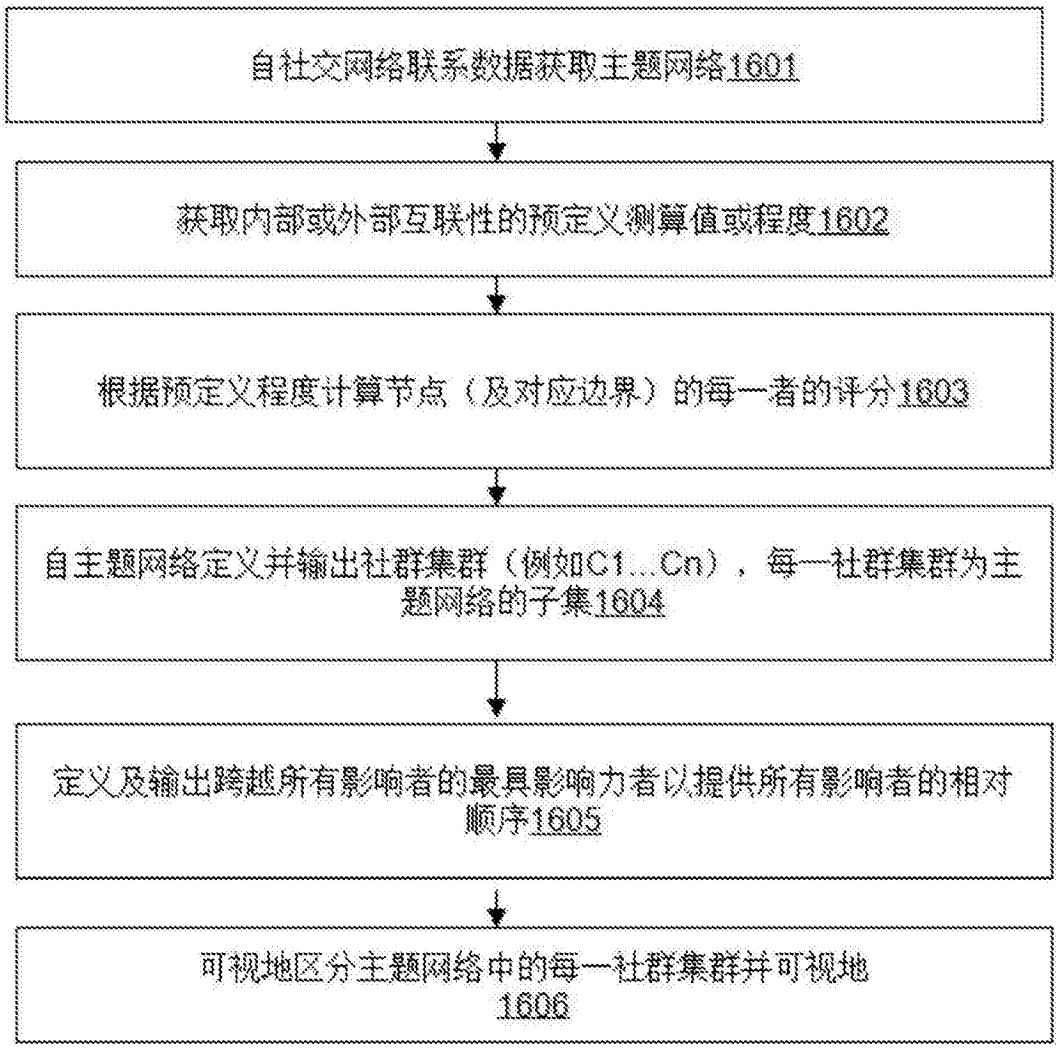

自社交网络联系数据获取主题网络1601
获取内部或外部互联性的预定义测算值或程度1602
根据预定义程度计算节点（及对应边界）的每一者的评分1603
自主题网络定义并输出社群集群（例如C1...Cn），每一社群集群为主题网络的子集1604
定义及输出跨越所有影响者的最具影响力者以提供所有影响者的相对顺序1605
可视地区分主题网络中的每一社群集群并可视地1606

图16

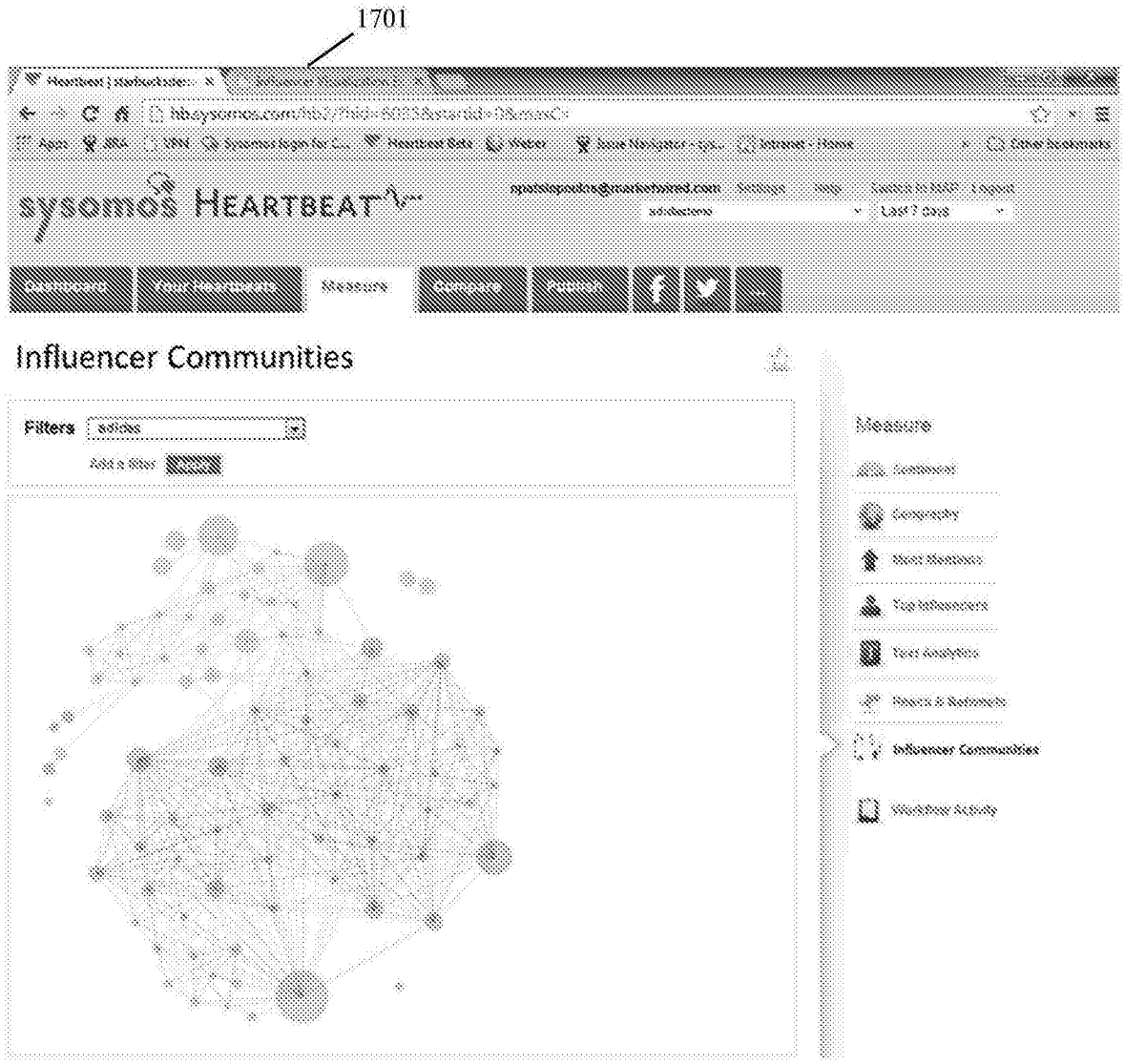

1701
sysomos HEARTBEAT
Influencer Communities
Filters
Measure

图17A

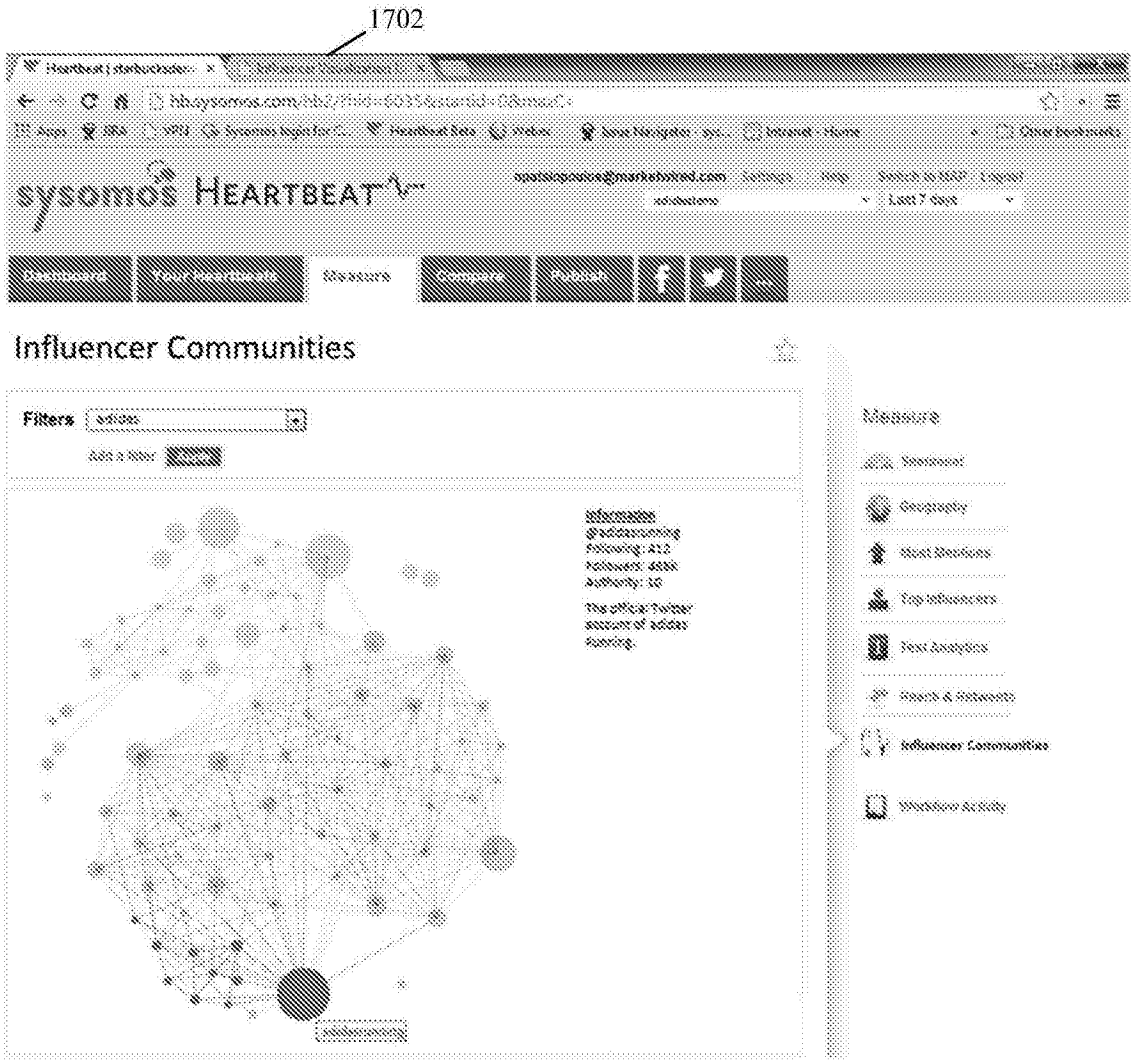

1702
sysomos HEARTBEAT
Influencer Communities
Filters
Measure

图17B

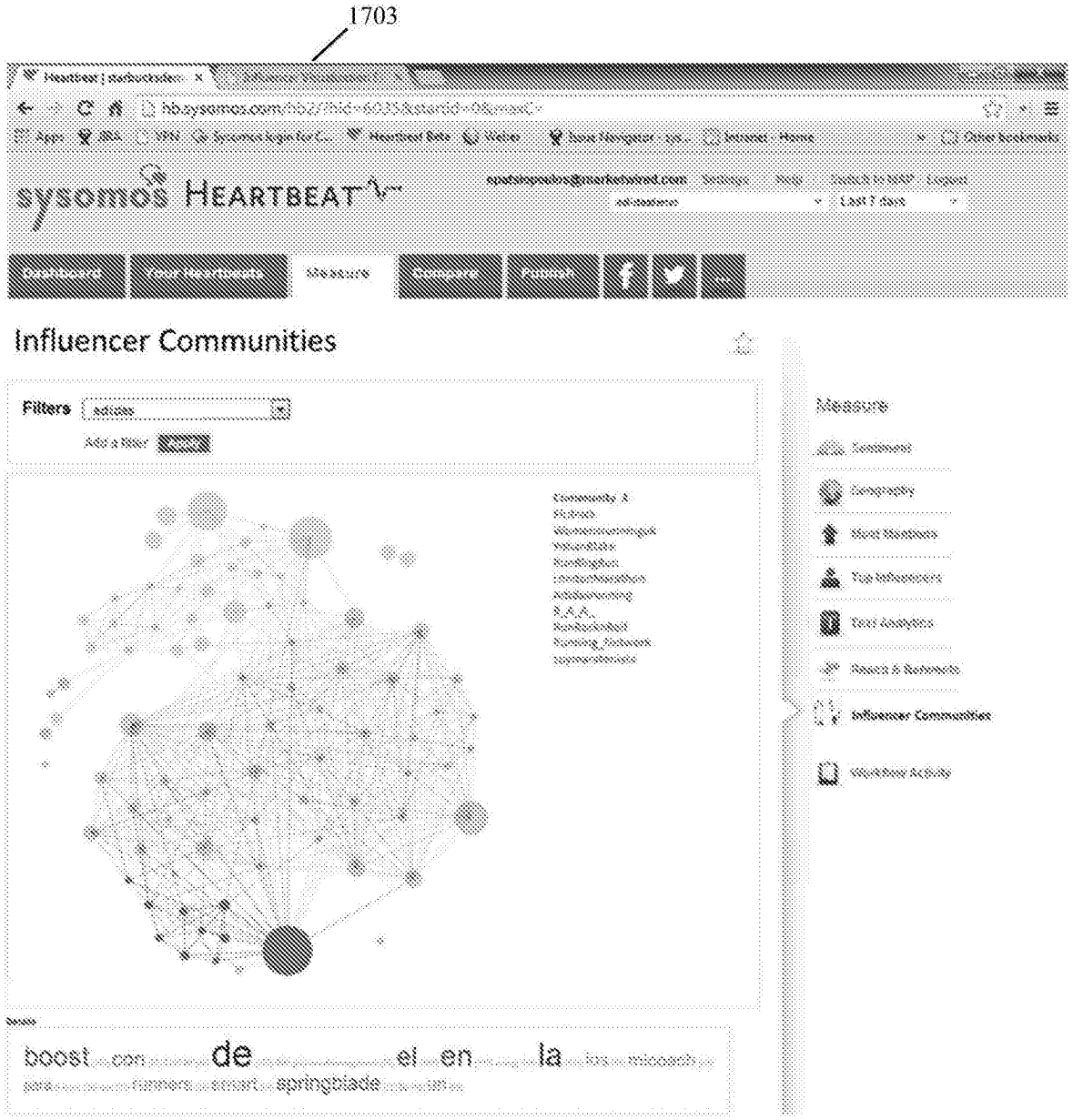

1703
sysomos HEARTBEAT
Influencer Communities
Filters
Measure
boost
de
el
en
la

图17C

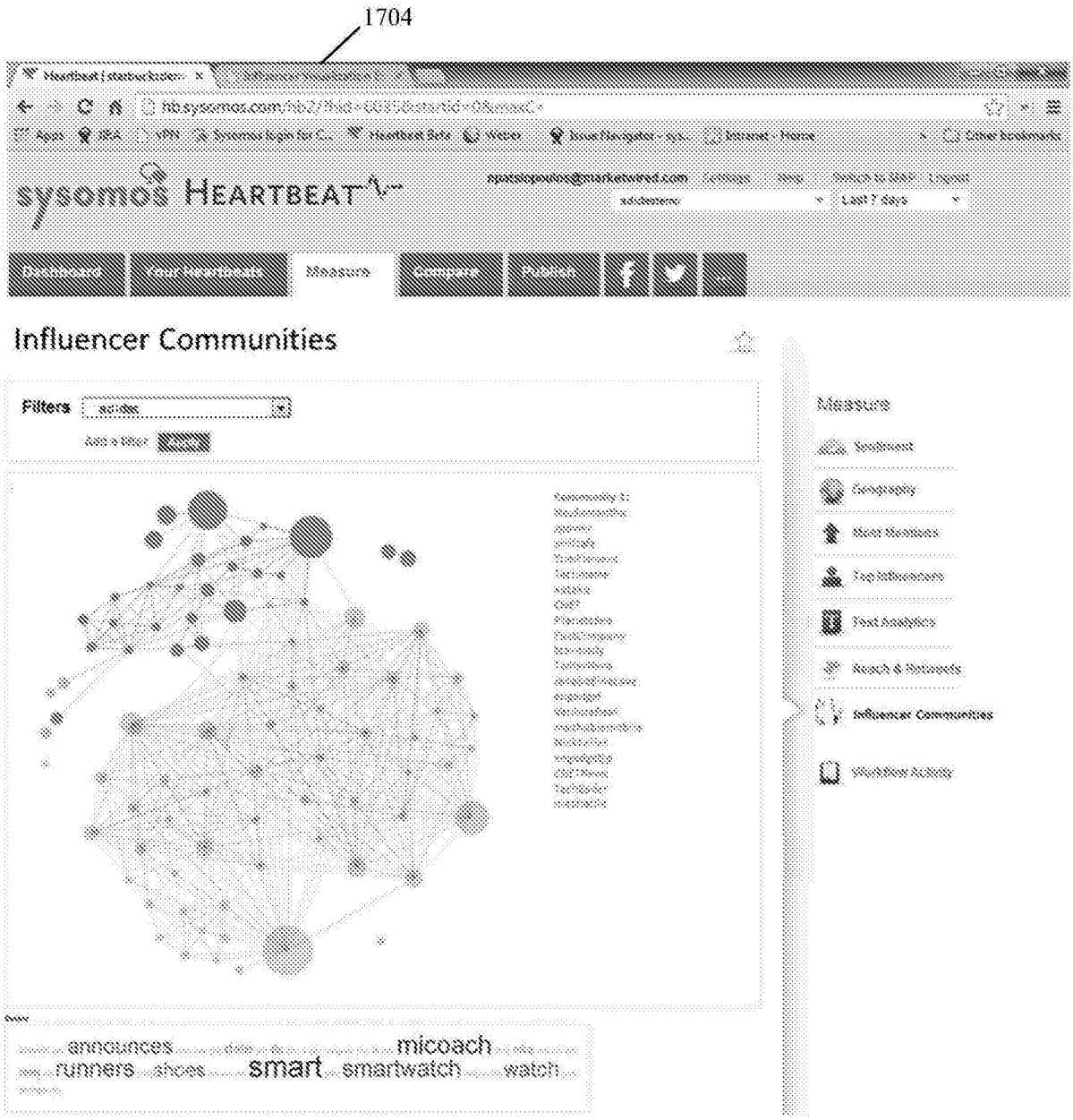
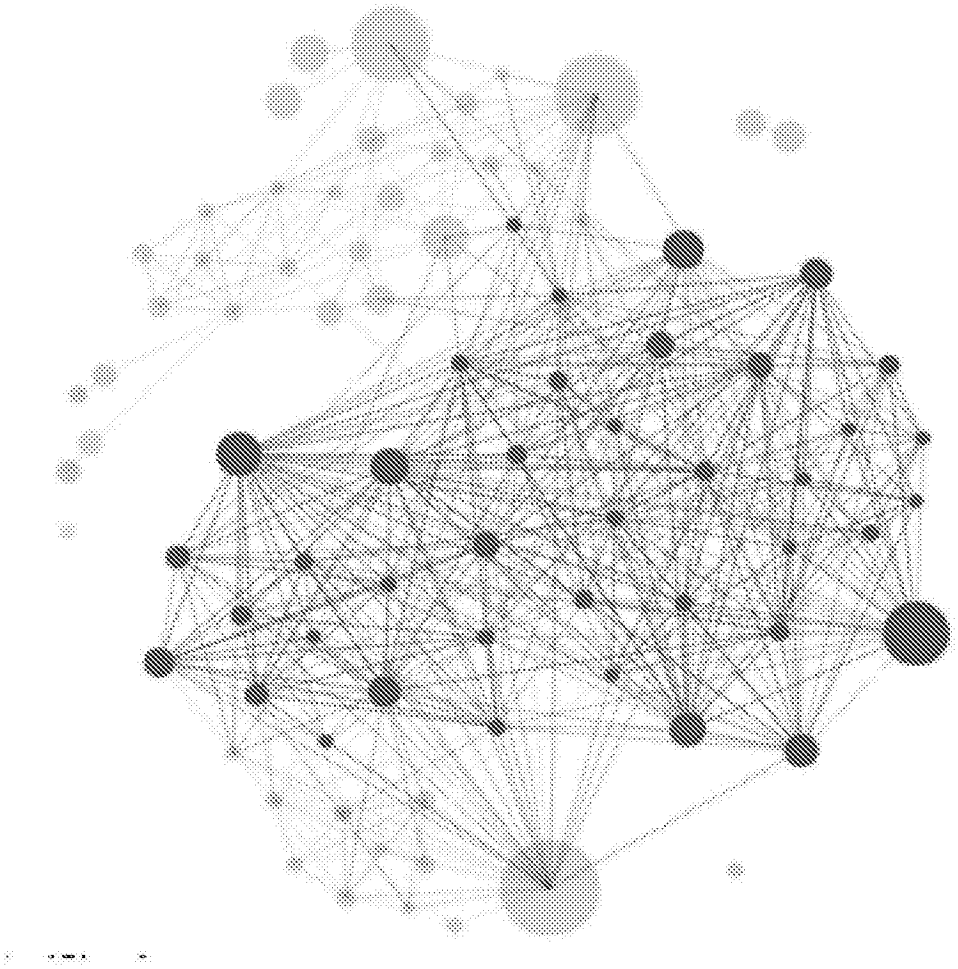
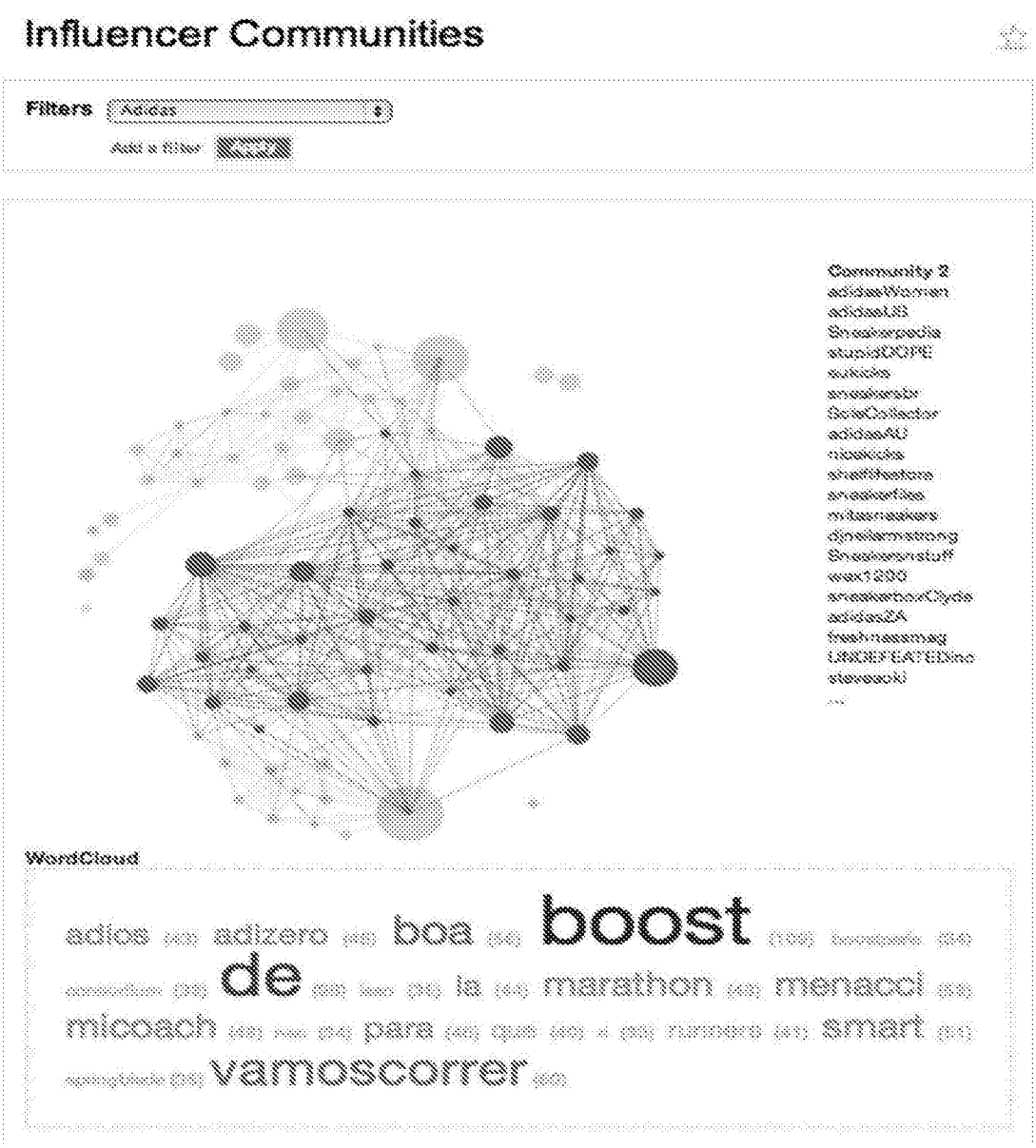

1704
sysomos HEARTBEAT
Dashboard
Your Heartbeats
Measure
Compare
Publish
Influencer Communities
Filters
Measure
announces
runners
shoes
smart
smartwatch
micoach
watch

图17D



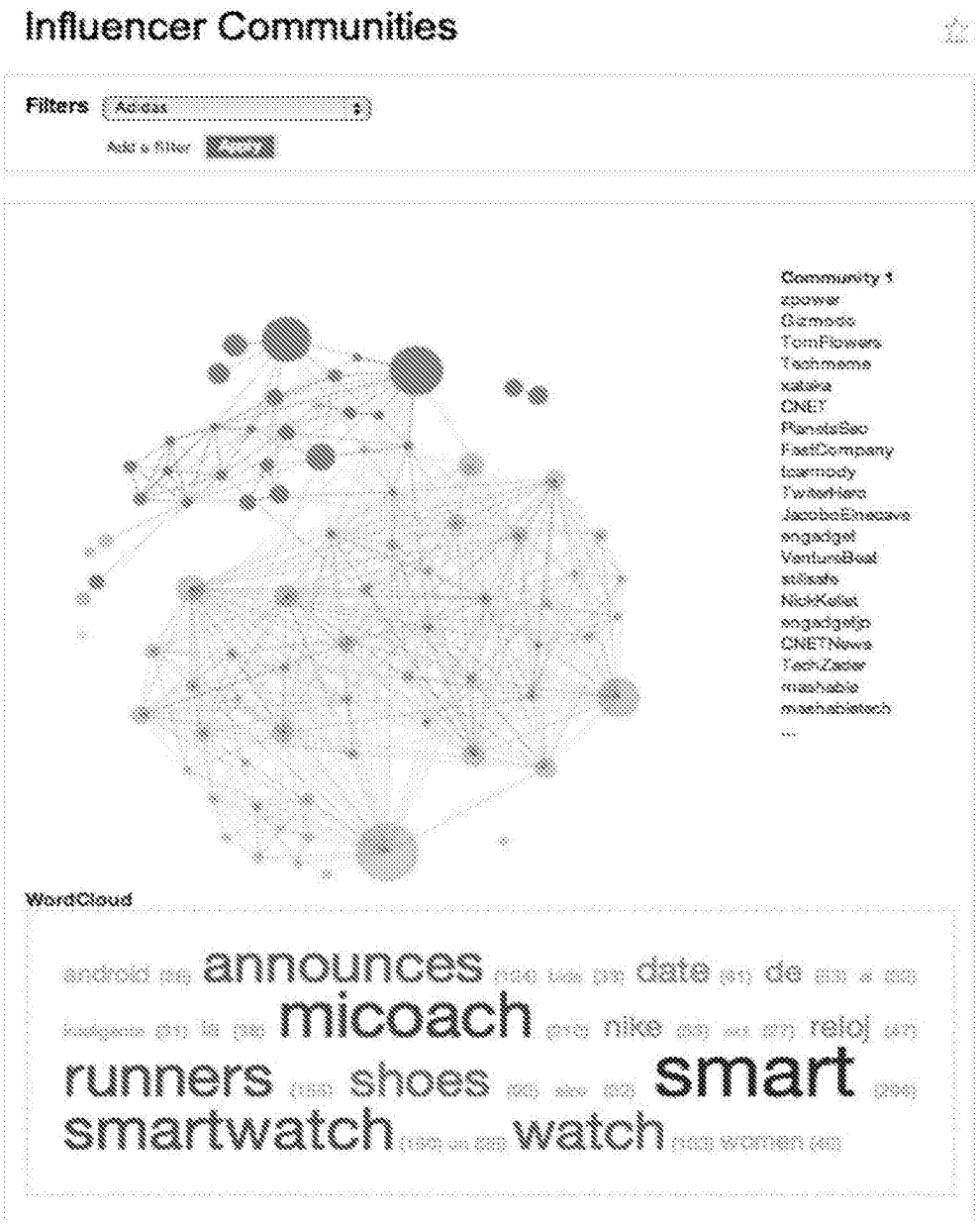
网络图社群

图18



Adidas Running 主题的影响者社群图

图19A

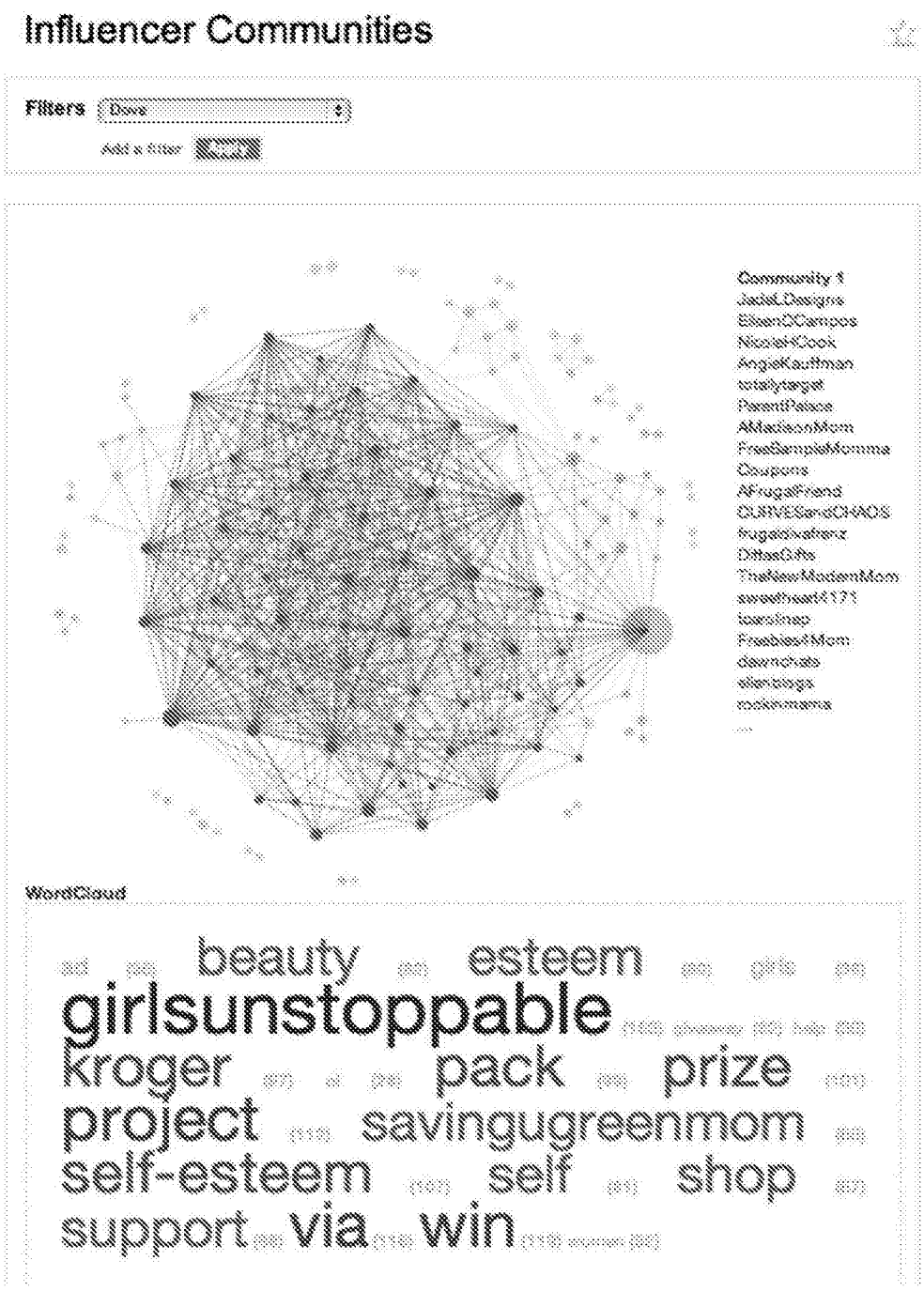


Adidas Running 主题的影响者社群 1

图19B

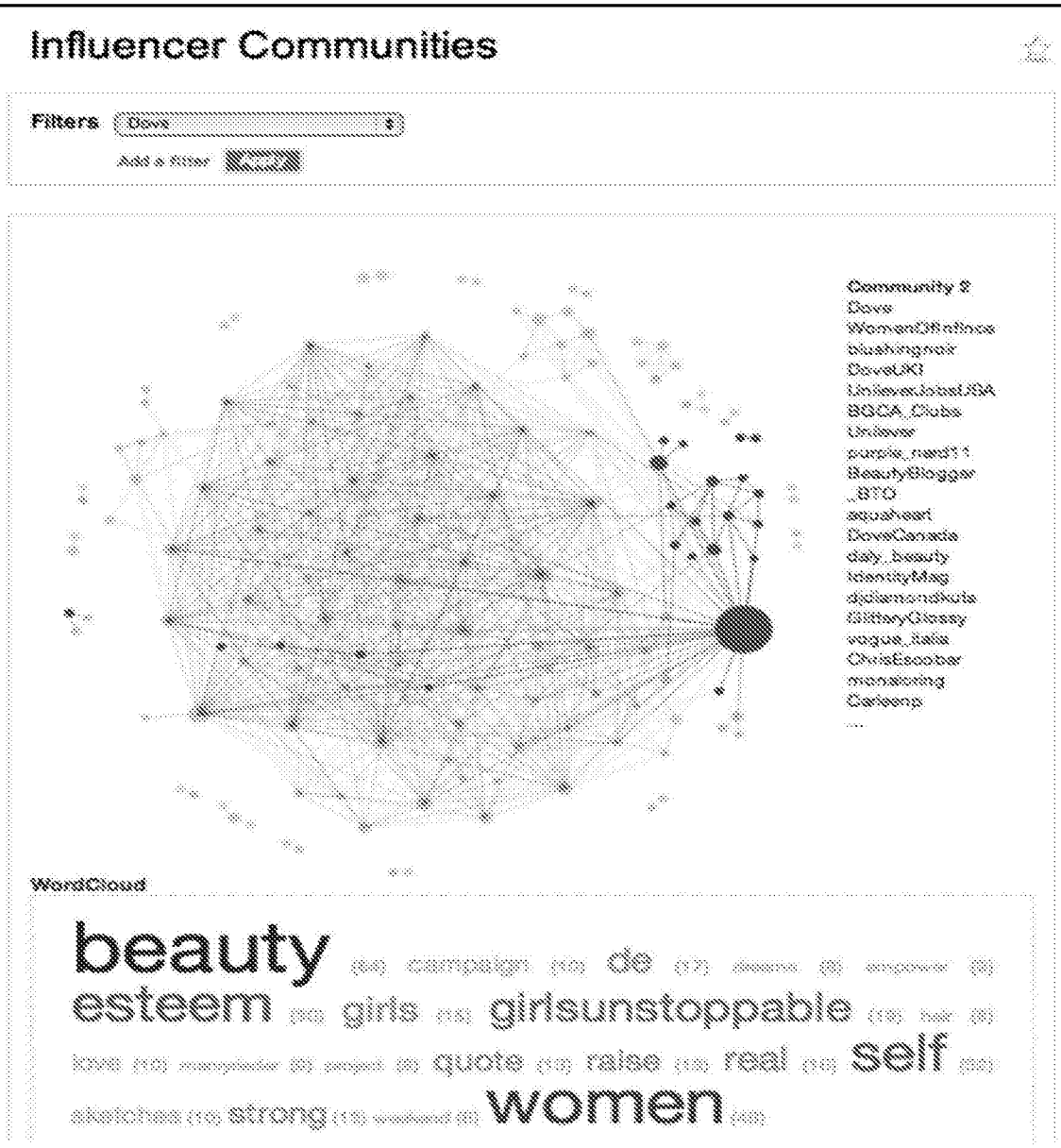
Adidas Running 主题的影响者社群 3

图19C



Dove 主题的影响者社群 1 图

图20A



Dove 主题的影响者社群 2 图

图20B