(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 107710266 A
(43)申请公布日 2018.02.16

(21)申请号 201680038125.0

(22)申请日 2016.08.08

(30)优先权数据
62/201,738 2015.08.06 US

(85)PCT国际申请进入国家阶段日
2017.12.28

(86)PCT国际申请的申请数据
PCT/US2016/046082 2016.08.08

(87)PCT国际申请的公布数据
WO2017/024316 EN 2017.02.09

(71)申请人 赫尔实验室有限公司
地址 美国加利福尼亚州

(72)发明人 许劼钧 T-C·卢

(74)专利代理机构 北京三友知识产权代理有限公司 11127
代理人 吕俊刚 杨薇

(51)Int.Cl.
G06Q 30/02(2012.01)
G06Q 50/00(2012.01)

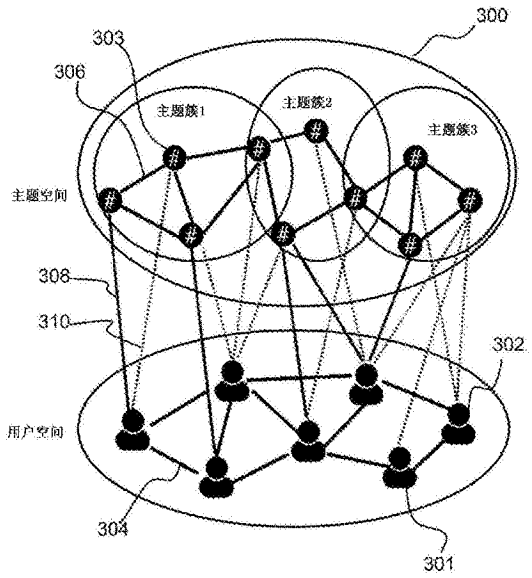
权利要求书2页 说明书11页 附图8页

(54)发明名称

用于借助社交媒体识别用户兴趣的系统和方法

(57)摘要

描述了一种用于借助在线社交媒体发现用户兴趣的系统，并且更具体地，涉及一种借助于双向图模型这样做的方式。在操作期间，该系统基于社交媒体平台上的用户交互和同现标签生成置信度矩阵F。置信度矩阵F指示社交媒体平台中的用户对特定主题感兴趣的可能性。基于这种可能性，对于对特定主题感兴趣的可能性超过预定阈值的那些用户，发起关于特定主题的动作。例如，该系统生成并且向对特定主题感兴趣的可能性超过预定阈值的那些用户呈现关于特定主题的针对用户的在线广告。



CN 107710266 A

1.一种用于借助社交媒体识别用户兴趣的系统,所述系统包括:

一个或更多个处理器和存储器,所述存储器为上面编码有可执行指令的非暂时性计算机可读介质,使得在执行所述指令时,所述一个或更多个处理器执行以下操作:

基于社交媒体平台上的用户交互和同现标签生成置信度矩阵F,所述置信度矩阵F指示所述社交媒体平台中的用户对特定主题感兴趣的可能性;以及

针对对特定主题感兴趣的可能性超过预定阈值的那些用户发起与所述特定主题有关的动作。

2.根据权利要求1所述的系统,所述系统还包括以下操作:

基于社交媒体平台上的用户交互的集合构造用户交互网络W;

基于所述社交媒体平台上的同现标记的集合构造标签同现网络R_h;

基于所述标签同现网络R_h构造主题相关网络R;

从所述用户交互网络W生成用户图拉普拉斯L_g;

从所述主题相关网络R生成主题图拉普拉斯L_c;以及

基于初始已知用户主题关联生成初始标签分配矩阵Y。

3.根据权利要求2所述的系统,其中,在生成主题相关网络R时,通过对R_h应用Louvain社区检测来生成所述主题相关网络。

4.根据权利要求3所述的系统,其中,置信度矩阵F的行表示用户,并且列表示主题,使得所述置信度矩阵F的各条目指示用户对特定主题感兴趣的可能性。

5.根据权利要求4所述的系统,其中,发起动作还包括以下操作:生成并且向对特定主题感兴趣的可能性超过预定阈值的那些用户呈现关于所述特定主题的针对用户的在线广告。

6.根据权利要求1所述的系统,其中,置信度矩阵F的行表示用户,并且列表示主题,使得所述置信度矩阵F的各条目指示用户对特定主题感兴趣的可能性。

7.根据权利要求1所述的系统,其中,发起动作还包括以下操作:生成并且向对特定主题感兴趣的可能性超过预定阈值的那些用户呈现关于所述特定主题的针对用户的在线广告。

8.一种用于借助社交媒体识别用户兴趣的方法,所述方法包括以下动作:

基于社交媒体平台上的用户交互和同现标签利用一个或更多个处理器生成置信度矩阵F,所述置信度矩阵F指示所述社交媒体平台中的用户对特定主题感兴趣的可能性;以及

利用所述一个或更多个处理器针对对特定主题感兴趣的可能性超过预定阈值的那些用户发起与所述特定主题有关的动作。

9.根据权利要求8所述的方法,所述方法还包括以下操作:

基于社交媒体平台上的用户交互的集合构造用户交互网络W;

基于所述社交媒体平台上的同现标签的集合构造标签同现网络R_h;

基于所述标签同现网络R_h构造主题相关网络R;

从所述用户交互网络W生成用户图拉普拉斯L_g;

从所述主题相关网络R生成主题图拉普拉斯L_c;以及

基于初始已知用户主题关联生成初始标签分配矩阵Y。

10.根据权利要求9所述的方法,其中,在生成主题相关网络R时,通过对R_h应用Louvain

社区检测来生成所述主题相关网络。

11.根据权利要求10所述的方法，其中，置信度矩阵F的行表示用户，并且列表示主题，使得所述置信度矩阵F的各条目指示用户对特定主题感兴趣的可能性。

12.根据权利要求11所述的方法，其中，发起动作还包括以下动作：生成并且向对特定主题感兴趣的可能性超过预定阈值的那些用户呈现关于所述特定主题的针对用户的在线广告。

13.根据权利要求8所述的方法，其中，置信度矩阵F的行表示用户，并且列表示主题，使得所述置信度矩阵F的各条目指示用户对特定主题感兴趣的可能性。

14.根据权利要求8所述的方法，其中，发起动作还包括以下动作：生成并且向对特定主题感兴趣的可能性超过预定阈值的那些用户呈现关于所述特定主题的针对用户的在线广告。

15.一种用于借助社交媒体识别用户兴趣的计算机程序产品，所述计算机程序产品包括：

非暂时性计算机可读介质，所述非暂时性计算机可读介质上面编码有可执行指令，使得在由一个或更多个处理器执行所述指令时，所述一个或更多个处理器执行以下操作：

基于社交媒体平台上的用户交互和同现标签生成置信度矩阵F，所述置信度矩阵F指示所述社交媒体平台中的用户对特定主题感兴趣的可能性；以及

针对对特定主题感兴趣的可能性超过预定阈值的那些用户发起与所述特定主题有关的动作。

16.根据权利要求15所述的计算机程序产品，所述计算机程序产品还包括以下操作：

基于社交媒体平台上的用户交互的集合构造用户交互网络W；

基于所述社交媒体平台上的同现标签的集合构造标签同现网络R_h；

基于所述标签同现网络R_h构造主题相关网络R；

从所述用户交互网络W生成用户图拉普拉斯L_g；

从所述主题相关网络R生成主题图拉普拉斯L_c；以及

基于初始已知用户主题关联生成初始标签分配矩阵Y。

17.根据权利要求16所述的计算机程序产品，其中，在生成主题相关网络R时，通过对R_h应用Louvain社区检测来生成所述主题相关网络。

18.根据权利要求17所述的计算机程序产品，其中，置信度矩阵F的行表示用户，并且列表示主题，使得所述置信度矩阵F的各条目指示用户对特定主题感兴趣的可能性。

19.根据权利要求18所述的计算机程序产品，其中，发起动作还包括以下操作：生成并且向对特定主题感兴趣的可能性超过预定阈值的那些用户呈现关于所述特定主题的针对用户的在线广告。

20.根据权利要求15所述的计算机程序产品，其中，置信度矩阵F的行表示用户，并且列表示主题，使得所述置信度矩阵F的各条目指示用户对特定主题感兴趣的可能性。

21.根据权利要求15所述的计算机程序产品，其中，发起动作还包括以下操作：生成并且向对特定主题感兴趣的可能性超过预定阈值的那些用户呈现关于所述特定主题的针对用户的在线广告。

用于借助社交媒体识别用户兴趣的系统和方法

[0001] 政府权利

[0002] 本发明在由IARPA发布的美国政府合同号D12PC00285下由政府支持做出。政府在本发明中具有特定权利。

[0003] 相关申请的交叉引用

[0004] 这是2015年8月6日提交的第62/201,738号美国临时申请的非临时专利申请,在此以引证方式将该申请的全文并入。

技术领域

[0005] 本发明涉及一种用于发现用户兴趣的系统,并且更具体地,涉及一种用于使用双向图模型借助在线社交媒体发现用户兴趣的系统。

背景技术

[0006] 对从在线社交媒体发现用户兴趣和主题越来越受关注(参见所并入参考文献列表,第3和4号参考文献)。一种常见方法是使用从用户的所有帖子的文本生成的向量表示来表示用户的兴趣。然后,可以由两个用户的特征向量的相似性得分测量两个用户之间的相似性。这还被称为词袋方法。然而,这种方法非常易受嘈杂文本影响。这在社交媒体环境中更为严重,这是因为用户自由地发布可能不反映他们感兴趣的真正主题的关于他们生活的任意帖子。用于隐藏用户主题发现的另一种深入研究方法是基于LDA(潜在狄利克雷分配,Latent Dirichlet Allocation)的方法。已经使用基于LDA的方法的一些研究可以在第1、4和8号参考文献中看到。因为LDA依赖词袋假定,所以它具有类似缺点。另外,对LDA的计算要求通常较高,并且它对方法的可扩展性形成显著瓶颈。

[0007] 识别兴趣的另一种方法是分析如在社交和主题空间中构造的网络拓扑。在第2号参考文献中,作者调查互易Twitter追随者网络中的用户社区,并且将用户兴趣总结为几类。在第5号参考文献中,作者提出通过对用户感兴趣的主题建模来链接由用户张贴的推文中的实体指代(mentions)的基于图的框架。前述方法的一个共性是两种方法在它们的分析中仅聚焦于一种类型的网络拓扑(例如,用户为中心的网络或主题为中心的网络),这不允许以统一方式查阅多个网络中的双关系方面。

[0008] 由此,持续需要一种可以用于通过以统一方式将两个(用户和主题)网络的拓扑用于用户兴趣建模,借助在线社交媒体高效且有效地发现用户兴趣的系统。

发明内容

[0009] 本公开提供了一种用于借助在线社交媒体识别用户兴趣的系统。该系统包括一个或更多个处理器和上面编码有指令的关联存储器(例如,硬盘驱动器等)。在执行指令时,一个或更多个处理器执行多个操作。例如,在操作期间,该系统基于社交媒体平台(例如,Twitter、Tumblr或任意其它社交媒体平台)上的用户交互和同现标签生成置信度矩阵F。置信度矩阵F指示社交媒体平台中的用户对特定主题感兴趣的可能性。基于这种可能性,可以

针对对特定主题感兴趣的可能性超过预定阈值的那些用户发起关于特定主题的动作。例如，该系统可以生成并且向对特定主题感兴趣的可能性超过预定阈值(例如，大于50％或如操作员认为适当的任意其它预定阈值)的那些用户呈现关于特定主题的针对用户的在线广告。

[0010] 在另一个方面中，该系统执行以下操作：基于社交媒体平台上的用户交互的集合构造用户交互网络W；基于社交媒体平台上的同现标签的集合构造标签同现网络R_h；基于标签同现网络R_h构造主题相关网络R；从用户交互网络W生成用户图拉普拉斯L_g；从主题相关网络R生成主题图拉普拉斯L_c；以及基于初始已知用户主题关联(association)生成初始标签(label)分配矩阵Y。

[0011] 进一步地，在生成主题相关网络R时，通过对R_h应用Louvain社区检测来生成主题相关网络。

[0012] 在又一个方面中，置信度矩阵F的行表示用户，并且列表示主题，使得置信度矩阵F的各条目指示用户对特定主题感兴趣的可能性。

[0013] 最后，本发明还包括计算机程序产品和计算机实现方法。计算机程序产品包括存储在非暂时性计算机可读介质上的计算机可读指令，这些指令可由具有一个或更多个处理器的计算机执行，使得在执行指令时，一个或更多个处理器执行这里所列出的操作。另选地，计算机实现方法包括使得计算机执行这种指令并执行所得到的操作的动作。

附图说明

[0014] 本发明的目的、特征以及优点将从本发明的各种方面的以下具体描述连同以下附图而变得清晰，附图中：

[0015] 图1是描绘了根据本发明的各种实施方式的系统的组件的框图；

[0016] 图2是具体实现本发明的一方面的计算机程序产品的例示；

[0017] 图3是根据本发明的各种实施方式的用户兴趣建模的双关系图的例示；

[0018] 图4A是示例标签网络的例示。

[0019] 图4B是与在图4A中描绘的标签网络关联的示例主题网络的例示；

[0020] 图4C是示例标签网络的例示。

[0021] 图4D是与在图4C中描绘的标签网络关联的示例主题网络的例示；以及

[0022] 图5是例示了根据本发明的各种实施方式的用于识别用户兴趣的过程的流程图。

具体实施方式

[0023] 本发明涉及一种用于发现用户兴趣的系统，并且更具体地，涉及一种用于使用双向图模型借助在线社交媒体发现用户兴趣的系统。以下描述被提出为使得本领域普通技术人员能够进行并使用本发明并将本发明并入特定应用的环境中。各种修改以及不同应用中的各种使用对于本领域技术人员来说更明显，并且这里所定义的一般原理可以应用于广泛方面。由此，本发明不旨在限于所提出的方面，而是符合与这里所公开的原理和新型特征一致的最宽范围。

[0024] 在以下具体描述中，为了提供本发明的更彻底理解，阐述了大量具体细节。然而，将对本领域技术人员显而易见的是，本发明可以在不必限于这些具体细节的情况下被实

践。在其它情况下，为了避免使本发明模糊，以框图形式而不是详细示出公知结构和装置。

[0025] 读者的注意力在于与本说明书同时提交且与本说明书一起对公众审查公开的所有文献，并且以引证的方式将所有这种文献的内容并入于此。本说明书(包括任意所附权利要求、摘要以及附图)中所公开的所有特征可以用服务相同、等效或类似目的的另选特征来替换，除非另外明确阐述。由此，除非另外明确阐述，所公开的各特征仅为一般系列等效或类似特征的一个示例。

[0026] 此外，权利要求中未明确阐述用于执行指定功能的“装置”或用于执行特定功能的“步骤”的任何元素不被解释为如35 U.S.C第112章节第6段中指定的“装置”或“步骤”条款。具体地，这里权利要求中“的步骤”或“的动作”的使用不旨在涉及35 U.S.C 112第6段中的规定。

[0027] 在详细描述本发明之前，首先提供所列举参考文献的列表。接着，提供本发明的各种主要方面的描述。随后，引言给读者提供本发明的一般理解。最后，提供本发明的各种实施方式的具体细节，以给出特定方面的理解。

[0028] (1)所并入参考文献的列表

[0029] 贯穿本申请列举以下参考文献。为了清晰和方便起见，参考文献在这里被列出为读者的中心资源。以下参考文献通过引用被并入于此，就像被完全阐述一样。参考文献如下通过参考对应参考文献号列举在本申请中：

[0030] 1.Harvey,M.,Crestani,F.,&Carman,M.J.(2013).Building User Profiles from Topic Models for Personalised.Conference on Information and Knowiedge Managemeni(CIKM),San Francisco.

[0031] 2.Java,A.,Song,X.,Finin,T.,&Tseng,B.(2007).Why we twitter:Understanding microblogging usage and communities.In Proc,9th WebKDD and 1st SNA-KDD Workshop on Web Mining and Social Neiwork Analysis.

[0032] 3.Michelson,M.,&Macskassy,S.A.(2010).Discovering Users′Topics of Interest on Twitter:A First Look.Proceedings of the fourth workshop on Analytics for noisy unstructured txet data(AND).Toronto.

[0033] 4.Ovsjanikov,M.,&Chen,Y,(2010).Topic modeling for personalized recommendation of volatile items.European conference on Machine learning and knowledge discovery in daiabases:Part II.

[0034] 5.Shen,W.,Wang,J.,Luo,P.,&Wang,M.(2013).Linking Named Entities in Tweets with Knowledge Base via User Interest Modeling.ACM SIGKDD iniernational conference on Knowledge discovery and data mining.Chicago.

[0035] 6.Wang,H.,Huang,H.,&Ding,C.(2009).Image annotation using multi-label correlated Green′s function.IEEE 12th International Conference on Computer Vision.Kyoto.

[0036] 7.Weng,L.,&Menczer,F.(2014).Topicality and Social Impact:Diverse Messages but Focused Messengers.CoRR abs/1402.5443.

[0037] 8.Xu,J.,Compton,R.,Lu,T.-C.,&Allen,D.(2014).Rolling through Tumblr:Characterizing Behavioral Patterns of the Microblogging Platform.ACM Web

Science.Bloomington.

[0038] 9.Xu,J.,Jagadeesh,V.,&Manjunath,B.(2014).Multi-label Learning with Fused Multimodal Bi-relational Graph、IEEE Transaction on Multimedia.

[0039] 10.Xu,Z.,Lu,R.,Xiang,L.,&Yang,Q.(2011).Discovering User Interest on Twitter with a Modified Author-Topic Model.IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology.

[0040] 11.Jiejun Xa,Tsai-Ching Lu.Toward Precise User-Topic Alignment in Online Social MediaIn IEEE International Conference on Big Data(IEEE BigData),Santa Clara,California,2015.

[0041] 12.D.Zhou,O.Bousquet,T.N.Lal,J.Weston,and B.Schlkopf.Learning with local and global consistency.In NIPS.MIT Press,2004.

[0042] 13.X.Zhu.Semi-supervised Iearning literature survey.In University of Wisconsin Madison,Computer Sciences TR-1530,2008.

[0043] 14.R.Compton,D.Jurgens,and D.Allen.Geotagging one hundred million twitter accounts with total variation mini-mization.In IEEE International Conference on Big Data,volume abs/1404.7152,2014.

[0044] 15.R.Ottoni,D.B.L.Casas,J,P.Pesce,W,M,Jr.,C,Wilson,A、Misloye,and V.Almeida.Of pins and tweets:Investigating how users behave across image-and text-based social net-works.In Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM),2014.

[0045] 16.L.Weng and F.Menczer.Topicality and impact in social media:Diverse messages,focused messengers.PLoS ONE,10(2):e0118410,02 2015.

[0046] 17.Y.Yamaguchi,T.Amagasa,and H.Kitagawa.Tag-based user topic discovery using twitter lists.In Internutional Conference on Advances in Social Networks Analysis and Mining(ASONAM),Kaohsiung,Taiwan,25-27July 2011.

[0047] 18.V.Blondel,J.Guillaume,R.Lambiotte,and E.Mech、Fast unfolding of communities in large networks.J.Stat、Mech,page P10008,2008.

[0048] (2)主要方面

[0049] 本发明的各种实施方式包括三个“主要”方面。第一主要方面为一种用于借助在线社交媒体发现用户兴趣的系统,并且更具体地为借助于双向图模型这样做的方式。该系统通常为运行软件的计算机系统的形式或“硬编码”指令集的形式。该系统可以并入提供不同功能的各种装置中。第二主要方面为通常以软件形式使用数据处理系统(计算机)操作的方法。第三主要方面为一种计算机程序产品。该计算机程序产品通常代表非暂时性计算机可读介质(诸如光学存储装置(例如,光盘(CD)或数字视频盘(DVD))或磁存储装置(诸如软盘或磁带))上所存储的计算机可读指令。另外,计算机可读媒体的非限制性示例包括硬盘、只读存储器(ROM)以及闪存。下面将更详细地描述这些方面。

[0050] 图1中提供了描绘了本发明的系统(即,计算机系统100)的示例的框图。计算机系统100被配置为执行与程序或算法关联的计算、处理、操作和/或功能。在一个方面中,这里所讨论的特定处理和步骤被实现为位于计算机可读存储器单元内且由计算机系统100的一

个或更多个处理器执行的一系列指令(例如,软件程序)。当被执行时,指令使得计算机系统100执行特定动作并展示特定行为,诸如这里所述的。

[0051] 计算机系统100可以包括被配置为传输信息的地址/数据总线102。另外,一个或更多个数据处理单元(诸如处理器104(或多个处理器))与地址/数据总线102耦接。处理器104被配置为处理信息和指令。在一方面中,处理器104为微处理器。另选地,处理器104可以为不同类型的处理器(诸如并行处理器、专用集成电路(ASIC)、可编程逻辑阵列(PLA)、复杂可编程逻辑器件(CPLD)或现场可编程门阵列(FPGA))。

[0052] 计算机系统100被配置为使用一个或更多个数据存储单元。计算机系统100可以包括与地址/数据总线102耦接的易失性存储器单元106(例如,随机存取存储器(“RAM”)、静态RAM、动态RAM等),其中,易失性存储器单元106被配置为存储用于处理器104的信息和指令。计算机系统100还可以包括与地址/数据总线102耦接的非易失性存储器单元108(例如,只读存储器(“ROM”)、可编程ROM(“PROM”)、可擦除可编程ROM(“EPROM”)、电可擦除可编程ROM(“EEPROM”)、闪存等),其中,非易失性存储器单元108被配置为存储用于处理器104的静态信息和指令。另选地,计算机系统100诸如在“云”计算中可以执行从在线数据存储单元检索的指令。在一方面中,计算机系统100还可以包括与地址/数据总线102耦接的一个或更多个接口(诸如接口110)。一个或更多个接口被配置为使得计算机系统100能够与其它电子装置和计算机系统对接。由一个或更多个接口实现的通信接口可以包括有线(例如,串行电缆、调制解调器、网络适配器等)和/或无线(例如,无线调制解调器、无线网络适配器等)通信技术。

[0053] 在一方面中,计算机系统100可以包括与地址/数据总线102耦接的输入装置112,其中,输入装置112被配置为向处理器100传输信息和命令选择。根据一方面,输入装置112是可以包括字母数字键和/或功能键的字母数字输入装置(诸如键盘)。另选地,输入装置112可以为除了字母数字输入装置之外的输入装置。在一方面中,计算机系统100可以包括与地址/数据总线102耦接的光标控制装置114,其中,光标控制装置114被配置为向处理器100传输用户输入信息和/或命令选择。在一方面中,光标控制装置114使用诸如鼠标、追踪球、追踪垫、光学追踪装置或触摸屏的装置来实现。尽管存在前述内容,但在一方面中,光标控制装置114经由来自输入装置112的输入(诸如,响应于与输入装置112关联的特定键和键序列命令的使用)被定向和/或启动。在另选方面中,光标控制装置114被配置为由语音命令来定向或引导。

[0054] 在一方面中,计算机系统100还可以包括与地址/数据总线102耦接的一个或更多个可选计算机可用数据存储装置(诸如存储装置116)。存储装置116被配置为存储信息和/或计算机可执行指令。在一方面中,存储装置116为诸如磁或光盘驱动器(例如,硬盘驱动器(“HDD”))、软盘、光盘只读存储器(“CD-ROM”)、数字视频盘(“DVD”)等的存储装置。依据一方面,显示装置118与地址/数据总线102耦接,其中,显示装置118被配置为显示视频和/或图形。在一方面中,显示装置118可以包括阴极射线管(“CRT”)、液晶显示器(“LCD”)、场发射显示器(“FED”)、等离子体显示器或适于显示用户可识别的视频和/或图形图像以及字母数字字符的任意其它显示装置。

[0055] 这里所提出的计算机系统100为根据一方面的示例计算环境。然而,计算机系统100的非限制性示例不严格限于计算机系统。例如,一方面规定:计算机系统100表示可以根

据这里所述的各种方面使用的一种类型的数据处理分析。而且，还可以实现其它计算系统。实际上，本技术的精神和范围不限于任何单个数据处理环境。由此，在一方面中，本技术的各种方面中的一个或更多个操作使用由计算机执行的计算机可执行指令(诸如程序模块)来控制或实现。在一个实现中，这种程序模块包括被配置为执行特定任务或实现特定抽象数据类型的例程、程序、对象、组件和/或数据结构。另外，一方面规定：本技术的一个或更多个方面通过使用一个或更多个分布式计算环境(诸如，任务由借助通信网络链接的远程处理装置来执行的环境或诸如各种程序模块位于包括存储-储存装置的本地和远程计算机存储媒体中的环境)来实现。

[0056] 图2中描绘了具体实现本发明的计算机程序产品(即，存储装置)的例示图。计算机程序产品被描绘为软盘200或光盘202(诸如CD或DVD)。然而，如之前所提及的，计算机程序产品通常表示任意可兼容非暂时性计算机可读介质上存储的计算机可读指令。如这里关于本发明使用的术语“指令”通常指示要在计算机上执行的一组操作，并且可以表示整个程序或独立可分离软件模块的区段。“指令”的非限制性示例包括计算机程序代码(源代码或目标代码)和“硬编码”电子器件(即，被编码到计算机芯片中的计算机操作)。“指令”被存储在任意非暂时性计算机可读介质上(诸如，存储在计算机的存储器中或软盘、CD-ROM以及闪盘驱动器上)。在任一种情况下，指令被编码在非暂时性计算机可读介质上。

[0057] (3)引言

[0058] 本公开描述了一种基于双关系图从在线社交媒体(例如，Tumblr等)发现用户兴趣的技术。具体地，图模型包含两个子结构：用户的网络和主题(由标签表示)的网络。前者用于捕捉社交空间中的用户交互(例如，转载等)，并且后者用于捕捉主题空间中的标签同现。随后，用户兴趣发现问题被明确地公式化为关于所提出的双关系图的多标签学习问题。给定用户与标签的一些初始关联，该系统可以估计跨两个子网络的剩余用户节点与标签节点的关联。

[0059] 在一些实施方式中，系统和方法的目的是发现特定社交媒体用户的感兴趣话题。这允许基于用户的兴趣更好地簇并搜索。作为示例，聚焦于Tumblr平台，目标是基于用户张贴或转载的内容和用户如何与其它人交互生成针对各用户的一组“主题标签”。双关系图表示同时允许用户相似性和主题相关性的有效利用。这与独立地考虑两个因素的先前工作形成对比。

[0060] 如本领域技术人员可以理解的，该系统和方法例如可以用于科学技术分析(例如，基于用户的兴趣预测他们之间的将来合作)、用于从个性化或营销服务的兴趣模型建立用户配置文件以及其它数据收集用途。

[0061] (4)各种实施方式的具体细节

[0062] 如以上所注释的，本公开提供了一种用于用户兴趣发现的基于唯一双关系图的模型。这具有广泛应用(包括准确用户归档和个性化推荐)。主题或兴趣在该环境中被当作“标签”，并且用户兴趣发现的问题被公式化为关于图的多标签分类问题。多标签分类的一般处理已经在图像注释领域中被广泛研究(参见第6号和第9号参考文献)。根据本发明的各种实施方式的基于图的多标签分类技术表示将标签信息(即，兴趣、主题)从用户的小子集扩散到图中的剩余部分的直推式半监督学习处理。借助双关系图的仔细构造，在扩散处理中联合地利用用户相似性和标签相关性。当前分析针对Tumblr数据进行。平台的选择由第8号参

考文献启示(因为它示出了Tumblr极大地受用户兴趣驱动)。

[0063] (4.1)公式化

[0064] 图3中示出了双关系图的示例构造。如图所示,存在至少两个网络(主题空间300和社交或用户空间302)。用户空间实线304指示用户节点301之间的亲和力(affinity)关系(即,用户相似性),并且主题空间实线306指示主题节点303之间的亲和力关系(即,主题相关性)。跨两个网络的交叉网络实线308表示初始标签(即,主题/兴趣)关联,并且交叉网络虚线310表示要估计的标签分配。由此,两个子图中的每个内的实线304和306指示社交同质性关系和主题相关性,而跨两个子图的实心黑线308表示初始已知用户主题分配。

[0065] 在分类方面,大多数现有的基于图的半监督学习框架尝试使将以下两个特性考虑在内的成本函数最小化:数据(即,用户)图的平滑度和初始分配的偏差。这里,将第三个特性引入到正则化框架中,即,标签(主题)图中的平滑度。下面进一步详细地提供用于构造图的过程。

[0066] 假设存在N个用户$U=\{u_1,u_2,\ldots,u_N\}$和K个感兴趣主题$=\{t_1,t_2,\ldots,t_K\}$的集合。假定U中的一些用户针对他们感兴趣的主题被(部分)标注,则目标是利用标签子集$l\subseteq I$预测针对集合中的剩余未标注用户u_i的感兴趣主题。

[0067] 根据本发明的各种实施方式的基于图的多标签学习技术表示基于固有图结构将标签信息从小节点子集扩散到剩余节点的直推式半监督学习过程。注意,术语“感兴趣主题”和“标签”可以可互换地使用。传统基于图的学习中的基本步骤是构造顶点表示数据实例且边权重表示它们之间的亲和力的图。基于图的多标签学习的关键是一致性的先验假定:附近数据实例或位于同一结构上的数据实例很可能共享同一标签。通常,它在正则化框架中被公式化如下:

[0068] $F^*=\mathrm{argmin}_F\{\Omega_{smooth}(F)+\Omega_{prior}(F)\}$,其中,F是包含图节点的标签分配的待学习矩阵。

[0069] 第一项与通过对相邻标签强加平滑度约束来反映一致性假定的损失函数对应。第二项是用于拟合约束的正则化项,该正则化项意味着应尽可能少地改变初始分配的标签(参见第12和13号参考文献)。

[0070] 在本系统的环境中,数据实例与用户对应,并且它们的亲和力可以通过以社会互动被特征化或基于任意其它相似性措施(诸如用户人口统计资料和地理位置)来计算。注意,上述正则化框架的第一项根据社交同质化假定。除了用户图之外,传统基于图的学习框架还通过引入新图来强调主题之间的相关性来增强。两个图一起组成如图3例示的双关系图模型。

[0071] 给定用于小数据集合的标签关联(即,用户节点与主题节点之间的初始分配),目标是估计剩余部分中的两种类型节点之间的隐藏链接(link)。这种模型允许有效利用对两个子图以及它们之间的相互影响的平滑度约束。

[0072] (4.2)图构造

[0073] 该工作中的用户图的构造基于社交媒体平台中的一级交互。比如,可以聚焦于Twitter中的@mention4动作。Twitter用户经常通过在所指代的用户姓名之前加上“@”来“@mention”彼此。虽然存在其它类型的交互(诸如点赞(like)和转推(retweet)),但@mention已经被示出为指示社会纽带(参见第14号参考文献)。类似地,系统聚焦于Tumblr上的转载

(reblog)动作(该转载动作是再发布在Tumblr中的另一个用户帖子的内容的正式机制),因为该转载动作被示出为指示用户之间的共同爱好和兴趣(参见第8号参考文献)。为了获得强社会纽带,在各种实施方式中,系统聚焦于互换的@mention和reblog(注意,虽然使用@mention和reblog,但它们被提供为非限制示例,并且系统不限于这种线索)。换言之,如果在某一时间点,u_i@mentions(reblogs)u_j且u_j@mentions(reblogs)u_i,则在用户i与j之间仅引入双向边。边的权重基于两个用户之间的往复频率(即,@mentions(reblogs))的最小数来确定。

[0074] 主题图的构造基于主题之间的同现。然而,微博平台中通常不明确定义主题,在第15号参考文献中具有少数稀有例外。另选地,该系统可以被设计为将用户定义标签认为是研究社交媒体中的主题的通道。现有文献中已经研究了该策略(参见第16号和第17号参考文献)。

[0075] 为了例示性目的,图4A和图4C示出了用Twitter和Tumblr数据构造的标签同现网络的快照,而图4B和图4D分别描绘了对应主题网络。作为示例,节点的尺寸和/或颜色与其度数(degree)成比例;边的宽度与同现频率成比例。网络中的节点的"度数"是到其它节点的连接数。例如,节点可以被例示为使得它们的颜色例如从绿色到紫色到白色逐渐变化。在该非限制示例中,节点越绿,度数越高(即,连接到许多其它节点,或者为中心节点);另一方面,白色/紫色指示对应节点更少地连接到其它节点(即,外围节点)。作为另一个示例,节点越大,度数越高,而更小节点指示它们更少地被连接。

[0076] 如可以看到的,各个网络中的标签与单个相干主题有关。比如,Twitter网中的标签与是流行漫画出版商的"Marvel"有关。图中的节点包括漫画标题(及其名称变体)、漫画人物以及漫画书改编电影的角色成员。相同观察可以从源于Tumblr平台的样本标签网络看到,在该平台中,与"足球"有关的节点经常一起同现。

[0077] 因为社交媒体站点上的标签由数百万内容生成器自主创建,所以关于如何将它们分组到多个主题没有预定共识。可以开发多个副本标签来表示同一事件、题目或对象。比如,#loki、#thor、#odin、#asgard全部与Marvel电影中的虚构角色有关;#worldcup2014、#brazilwc2014、#wc2014、#fifawc14全部与发生在2014年6月的重大足球事件有关。为了减少副本和噪音,原始标签可以被聚集并抽象化为语义相关标签的更一般等级簇(被称为主题)。这些簇通过找到基于标签的同现网络中的社区来检测。例如,Louvain社区检测方法(参见第18号参考文献)可以由于其计算效率而用于识别主题簇。Louvain方法的基本思想是通过优化在所有节点上本地地优化模块性来重复地找到小社区,然后将这些小社区中的每个分组到单个节点。图4B和图4D示出了所得到的主题图的示例。可以观察强主题位置。

[0078] (4.3)关于双关系图的多标签学习

[0079] 如以上所提及的,传统基于图的学习框架使具有两项的成本函数最小化。将新主题图引入框架导致如下关于F的已更新正则化框架:

[0080] 使W为表示用N个数据点(用户)构造的数据图的N×N亲和力矩阵,并且R为表示针对K个主题构造的标签图的K×K亲和力矩阵。W和R中的基于频率的权重被标准化到相同动态范围。使$F=(F_1,\ldots,F_N)$ $T=(C_1,\ldots,C_K)$为表示每个用户主题对之间的最终关联的N×K矩阵。$(C_1,\ldots,C_K)$为与K个标签对应的F的列。类似地,使$Y=(Y_1,\ldots,Y_N)^T$为表示初始标签分配的N×K矩阵。各Y_{ij}具有1或0作为可能值:如果用户i被标注有主题j,则为1,如果用户i未

被标注，则为0。总成本函数被表达为：

[0081] $$\Omega(F)=\underbrace{\frac{1}{2}\eta\sum_{i,j}^{N}W_{ij}\left\|\frac{F_i}{\sqrt{D_{ii}}}-\frac{F_j}{\sqrt{D_{jj}}}\right\|^2}_{\text{关于用户图的平滑度}}+\underbrace{\frac{1}{2}\mu\sum_{i,j}^{N}W'_{ij}\left\|\frac{C_i}{\sqrt{D'_{ii}}}-\frac{C_j}{\sqrt{D'_{jj}}}\right\|^2}_{\text{关于主题图的平滑度}}+\underbrace{\sum_{i}^{N}\|F_i-Y_i\|^2}_{\text{先验约束}},\quad(1)$$

[0082] 其中，D和D'都是(i,i)条目等于W和R的第i行的总和的对角矩阵，即，$D_{ii}=\sum_{j=1}^{N}W_{ij}$和$D'_{ii}=\sum_{j=1}^{N}R_{ij}$。F的解可以通过使上述成本函数最小化来找到。

[0083] 上述方程(1)的第一项是对用户图的平滑度约束。使其最小化意味着相邻顶点应共享相似标签。比如，如果两个用户基于他们的频繁转载活动(例如，@mention、reblog)靠近彼此，则他们将可能具有共同兴趣(由此具有类似标签)。第二项是对主题或标签图的平滑度约束。使其最小化意味着相邻顶点应包括类似用户。比如，如果两个主题彼此高度相关，那么同一组用户可能对它们感兴趣。第三项指示初始已知的用户主题对应该尽可能少地改变。

[0084] η和μ是控制正则化项的折衷的两个常数。如果μ被设置为零，则它意味着忽视主题之间的相关性，并且公式化被减少至关于单个(社交)图的传统多标签学习。

[0085] 上述成本函数的第一项可以被重写为：

[0086] $$\begin{aligned}&\frac{1}{2}\eta\sum_{i,j}^{N}W_{ij}\left\|\frac{F_i}{\sqrt{D_{ii}}}-\frac{F_j}{\sqrt{D_{jj}}}\right\|^2\\&=\frac{1}{2}\eta\left(\sum_{i=1}^{N}F_i^TF_i+\sum_{j=1}^{N}F_j^TF_j-2\sum_{i,j=1}^{N}\frac{W_{ij}F_i^TF_j}{\sqrt{D_iD_j}}\right)\\&=\eta\left(\sum_{i=1}^{N}F_i^TF_i-\sum_{i,j=1}^{N}\frac{W_{ij}F_i^TF_j}{\sqrt{D_iD_j}}\right)\\&=\eta tr(F^T(I-D^{-1/2}WD^{-1/2})F).\end{aligned}\quad(2)$$

。

[0087] 类似地，成本函数的第二项和第三项可以用多个代数步骤以矩阵形式重写。由此，上述初始成本函数可以以更简洁形式写为：

[0088] $$\Omega(F)=\eta tr(F^TL_gF)+\mu tr(FL_cF^T)+tr((F-Y)^T(F-Y)),\quad(3)$$

[0089] 其中，$L_g=I-D^{-1/2}WD^{-1/2}$，并且$L_c=I-D'^{-1/2}RD'^{-1/2}$。它们分别是用户图和主题图的标准化拉普拉斯。

[0090] 通过应用以下矩阵特性：

[0091] $$\frac{\partial tr(X^TAX)}{\partial X}=(A+A^T)X,\quad\frac{\partial tr(XAX^T)}{\partial X}=X(A+A^T),\quad(4)$$

[0092] 方程可以关于F微分如下：

[0093] $$\frac{\partial\Omega(F)}{\partial F}=\eta LF+\mu FL_c+(F-Y),\quad(5)。$$

[0094] 这是因为L_g和L_c是对称矩阵。F的解可以通过要求$\frac{\partial \Omega(F)}{\partial F}$为零来获得。凭借一些简单代数步骤，变得明显的是$(\eta L_g+I)F+\mu F L_C=Y$，该方程基本上为具有AX+XB=C形式的矩阵方程。该方程的解可以从现有数字库(诸如，线性代数包(LAPACK)和矩阵实验室(Matlab))容易地获得。LAPACK是由田纳西(Tennessee)大学、加州大学伯克利分校(California, Berkeley)、科罗拉多大学丹佛分校(Colorado Denver)以及NAG公司提供的软件包。注意，F_{ij}基本上是对主题t_j感兴趣的用户u_i的置信度值。

[0095] 一旦找到F或F_{ij}，则可以使用简单阈值向用户分配标签(即，感兴趣主题)。基本上，具有更高值的用户可以被分配给具有更高置信度的对应主题。用于推断用户感兴趣主题的全过程在以下算法中概述。

[0096] 输入：设置包含用户交互的集合的$\mathbf{E}=\{(e_i^a,e_i^b,w_i)|i=1,2,\ldots,N_{|E|}\}$(例如，$e_i^a$转载$e_i^b$ w_i次)。设置包含同现标签的集合的$\mathbf{H}=\{(h_j^1,h_j^2,\ldots h_j^{n(j)})|j=1,2,\ldots,N_{|H|}\}$(例如，$h_j^1,\ldots h_j^{n(j)}$与第j个社交媒体帖子关联)。输出：置信度矩阵F，其中，F_{ij}指示用户u_i对主题t_j感兴趣的可能性。如图5所示，算法根据以下步骤来继续进行：

[0097] 1.从E构造(或生成)用户交互网络W 500。

[0098] 2.从H构造标签同现网络R_h 502。

[0099] 3.通过对R_h应用Louvain社区检测来构造主题相关网络R 504。

[0100] 4.从W计算用户图拉普拉斯L_g 506。

[0101] 5.从R计算主题图拉普拉斯L_c 508。

[0102] 6.基于初始已知用户主题关联计算Y 510。

[0103] 7.通过使方程(3)中的成本函数最小化来计算F 512，即，对以下矩阵方程求解：$\eta L_g F+\mu F L_c+(F-Y)=0$。

[0104] 8.通过对F中的条目分类并排序来返回最置信用户-主题对。

[0105] 系统然后可以用于通过使用源于如在上述算法中描述的在线社交网络的信息估计F矩阵来特征化社交媒体用户的感兴趣主题。F矩阵的行表示用户，并且列表示主题。矩阵的各条目指示用户对特定主题感兴趣的可能性。

[0106] 因为研究结果允许在线用户的更好簇和搜索，并且本发明对在线体验增强的个性化、推荐以及许多其它方面具有直接影响，所以本发明很重要。该系统已经被应用于特征化两个社交媒体平台(Twitter和Tumblr)上的在线用户感兴趣主题。在这两种情况下，与现有方法相比，获得显著改进。例如，如这里所描述的过程由如在第11号参考文献中描述的实验研究支持。

[0107] 如以上所注释的，存在系统可以通过针对对特定主题感兴趣的可能性超过预定阈值(例如，大于50％可能性)的用户自动发起关于特定主题的动作来实现。例如，基于F中的用户主题对和已排序条目，系统然后可以诸如通过自动地生成并且向对特定主题感兴趣的可能性超过预定阈值的那些用户呈现针对用户的关于特定主题的在线广告514，基于用户的兴趣向特定个人销售服务或商品。作为非限制示例，如果特定用户具有对与Marvel人物关联的主题感兴趣的高可能性(例如，大于50％)，那么可以借助互联网向用户呈现与卡通人物关联的即将上映电影的横幅广告。作为另一个非限制示例，如果特定用户具有对与足

球游戏(诸如世界杯)关联的主题感兴趣的高可能性,那么可以向用户呈现用于各种足球游戏的旅行包的横幅广告(例如,用于到国际足球事件的主办城市的航班和酒店住宿的横幅广告)。作为又一个非限制示例,如果用户具有对与汽车性能关联的主题感兴趣的高可能性,那么可以向用户呈现关于新车辆的邮件或横幅广告。

[0108] 最后,虽然已经鉴于多个实施方式描述了本发明,但本领域普通技术人员将容易地认识到,本发明可以具有其它环境中的其它应用。应注意,许多实施方式和实现是可以的。进一步地,以下权利要求并不旨在将本发明的范围限于以上所描述的具体实施方式。另外,“用于……的装置”的任意叙述旨在唤起元素和权利要求的装置加功能阅读,而未具体使用叙述“用于……装置”的任何元素不旨在被阅读为装置加功能元素,即使权利要求以其它方式包括词语“装置”。进一步地,虽然已经以特定顺序列举特定方法步骤,但方法步骤可以以任意期望顺序发生,并且落在本发明的范围内。

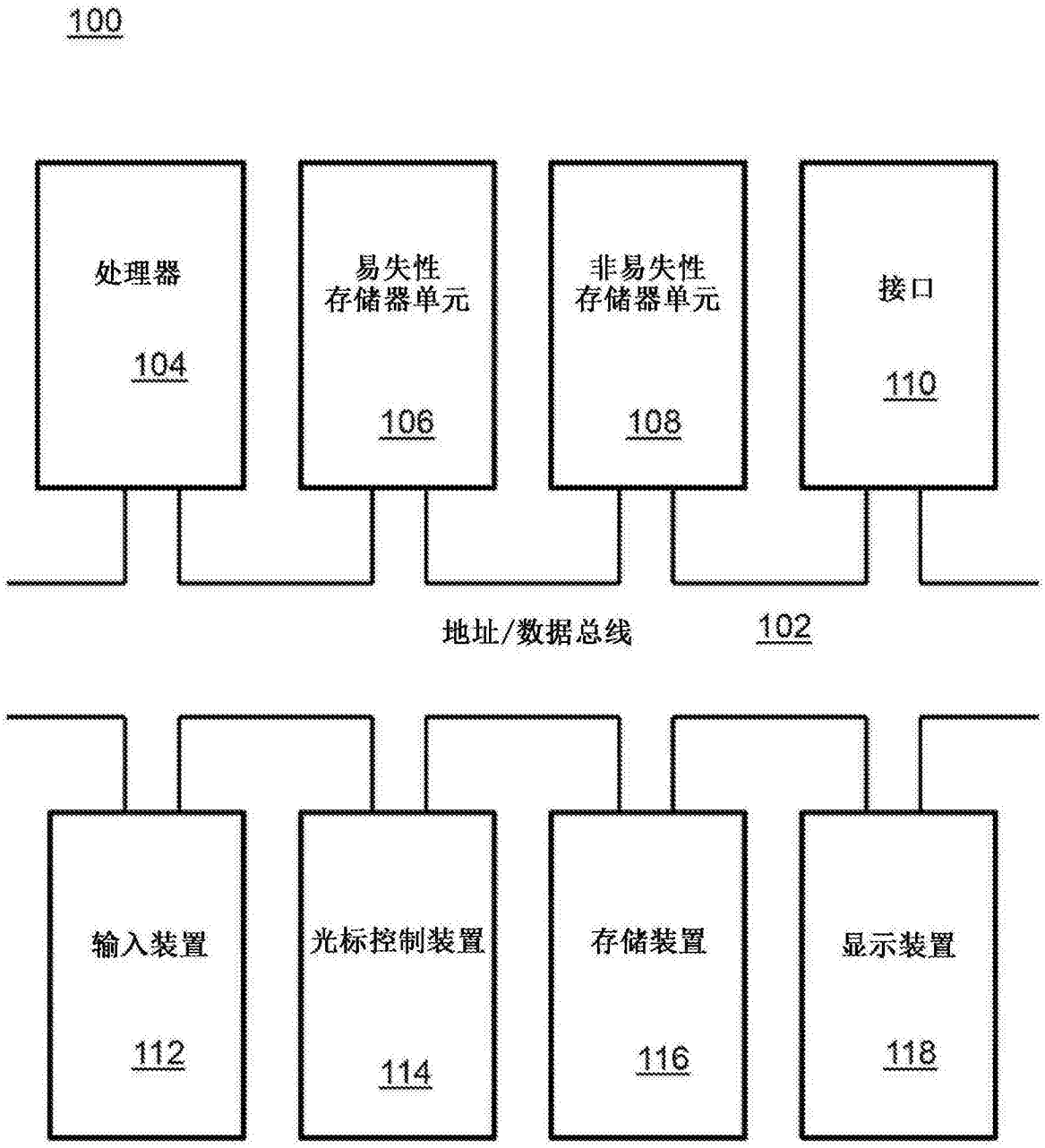

100
处理器
104
易失性
存储器单元
106
非易失性
存储器单元
108
接口
110
地址/数据总线
102
输入装置
112
光标控制装置
114
存储装置
116
显示装置
118

图1

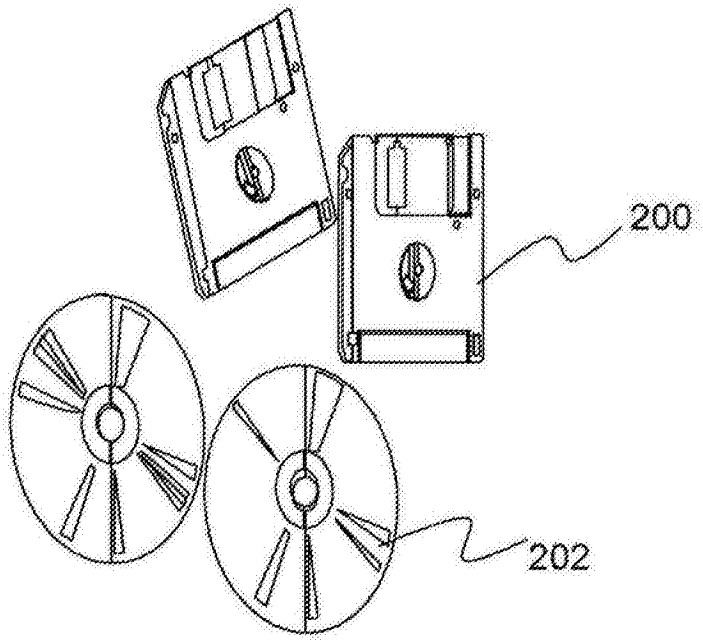


图2

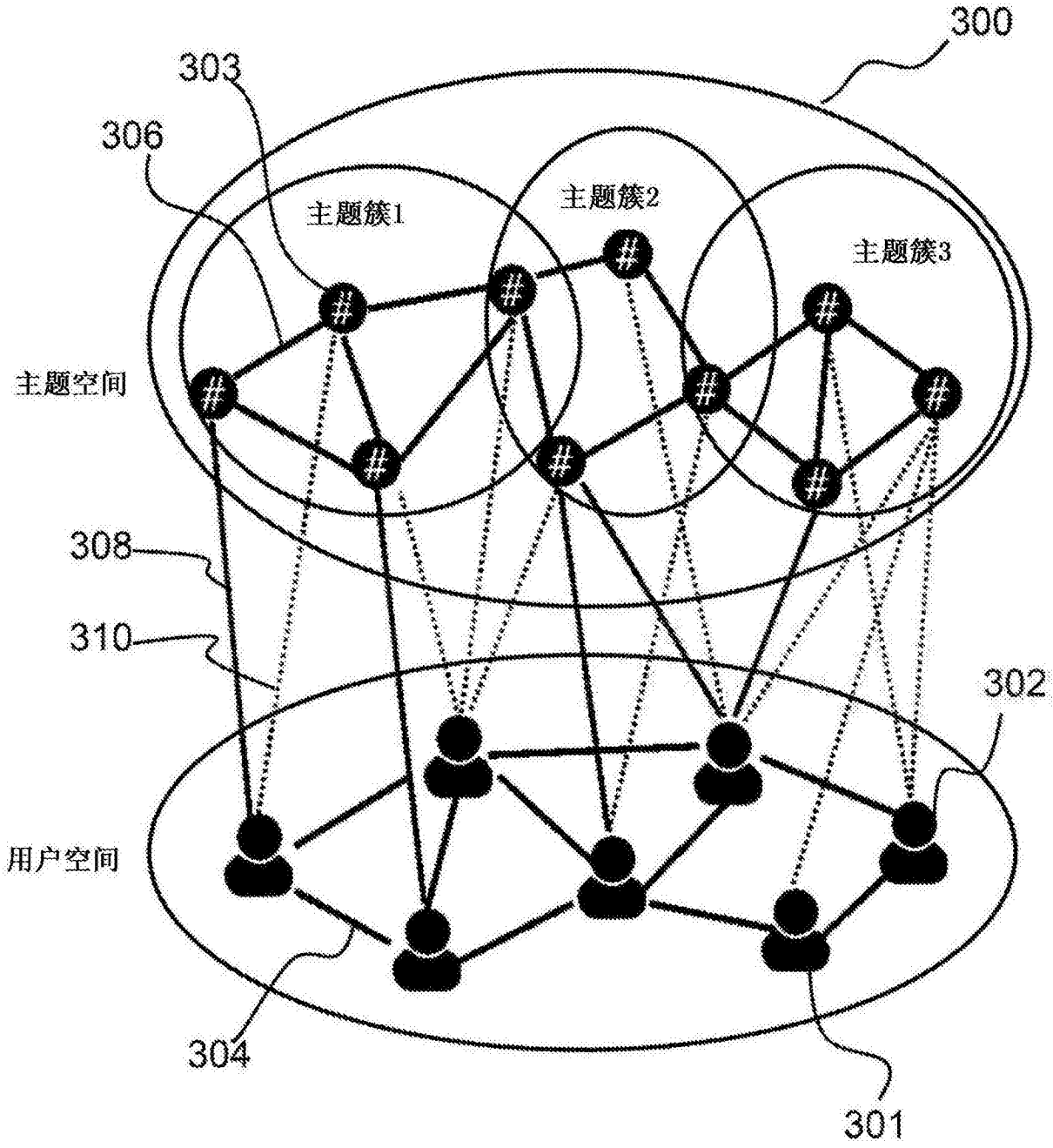
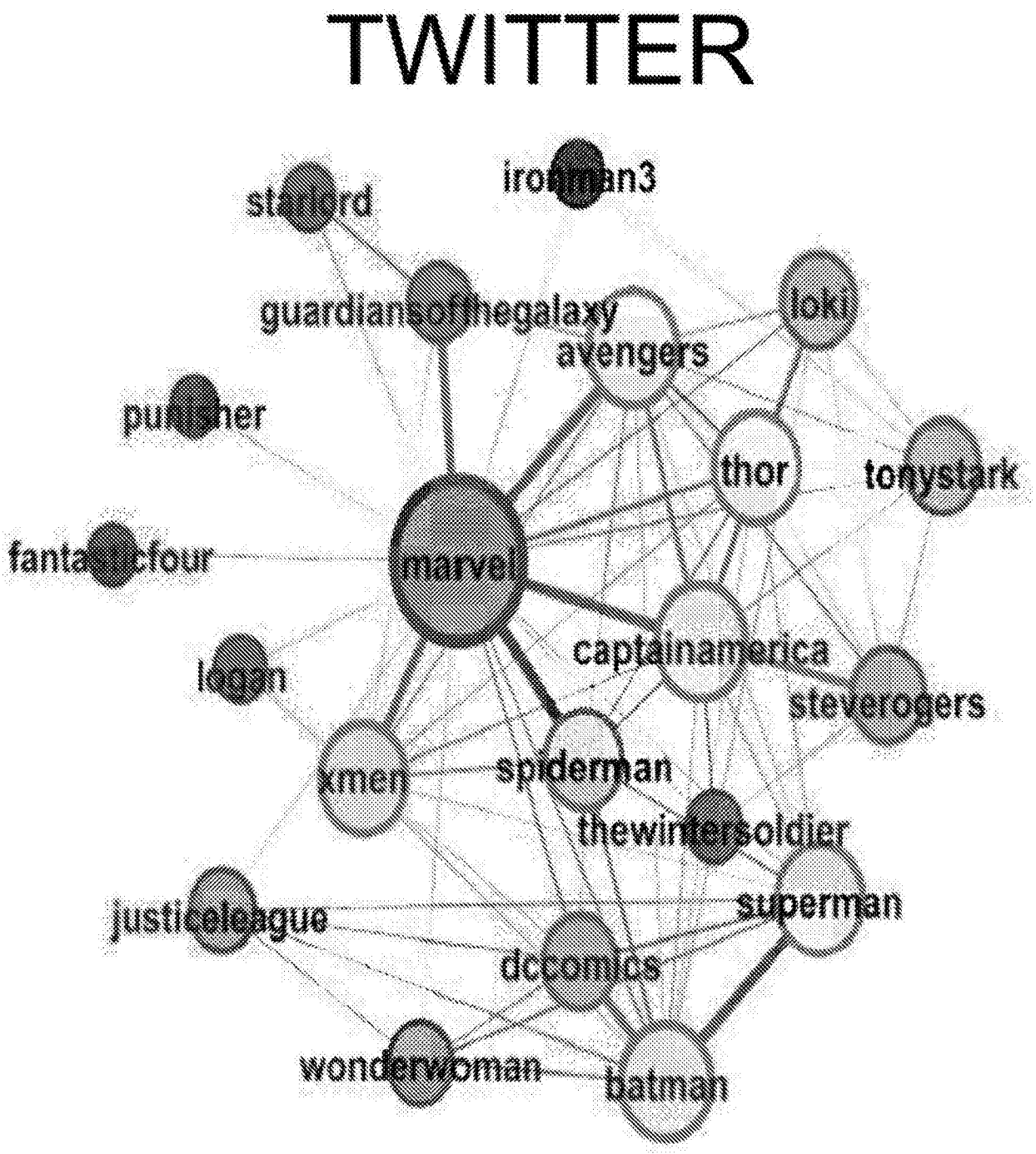
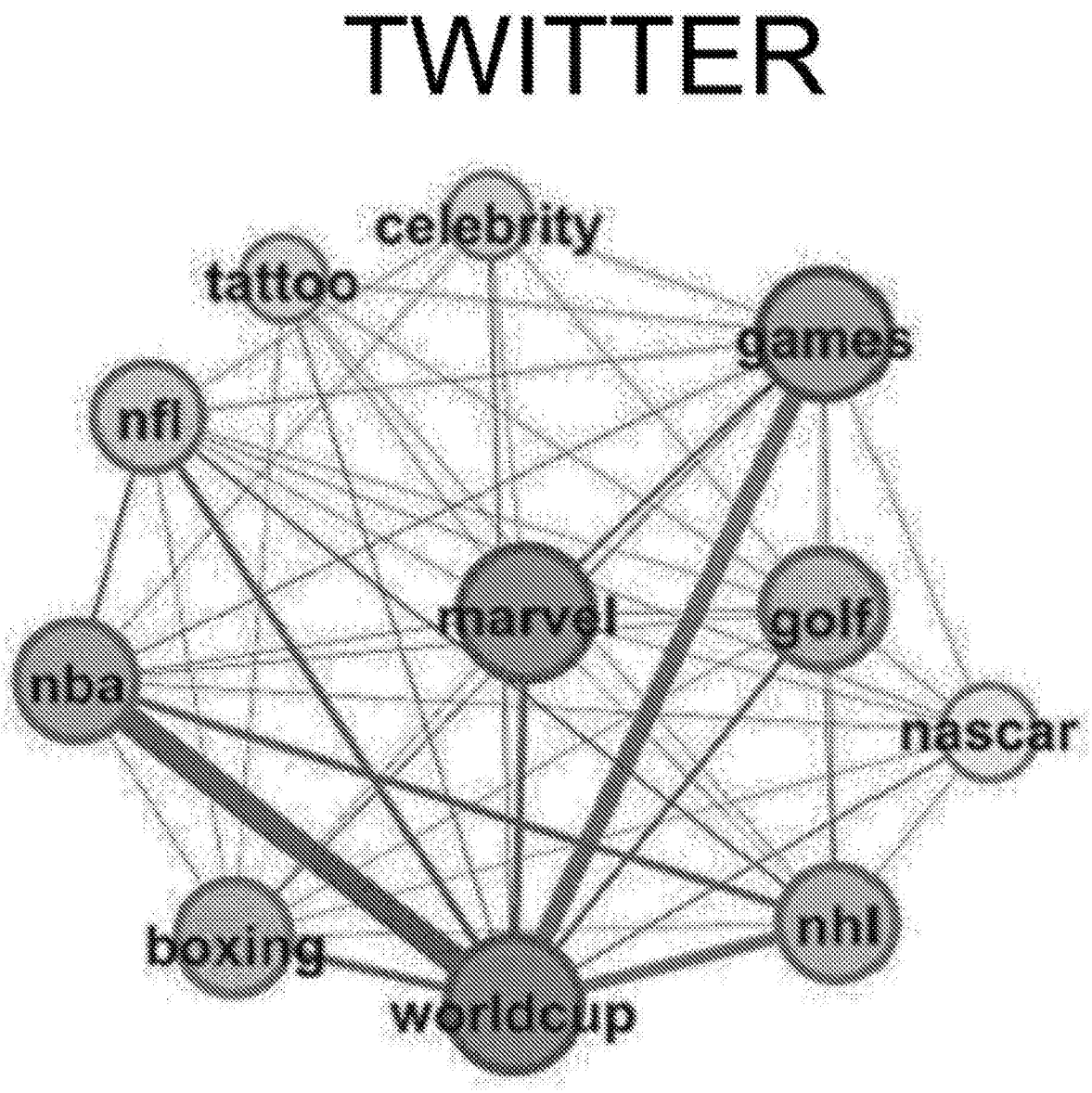


图3


TWITTER
ironman3
guardiansofthegalaxy
loki
avengers
thor
tonystark
marvel
captainamerica
steverogers
spiderman
thewintersoldier
superman
dccomics
batman
wonderwoman
justiceleague
xmen
logan

标签网络

图4A


TWITTER
celebrity
tattoo
games
nfl
marvel
golf
nba
nascar
boxing
nhl
worldcup

主题网络

图4b

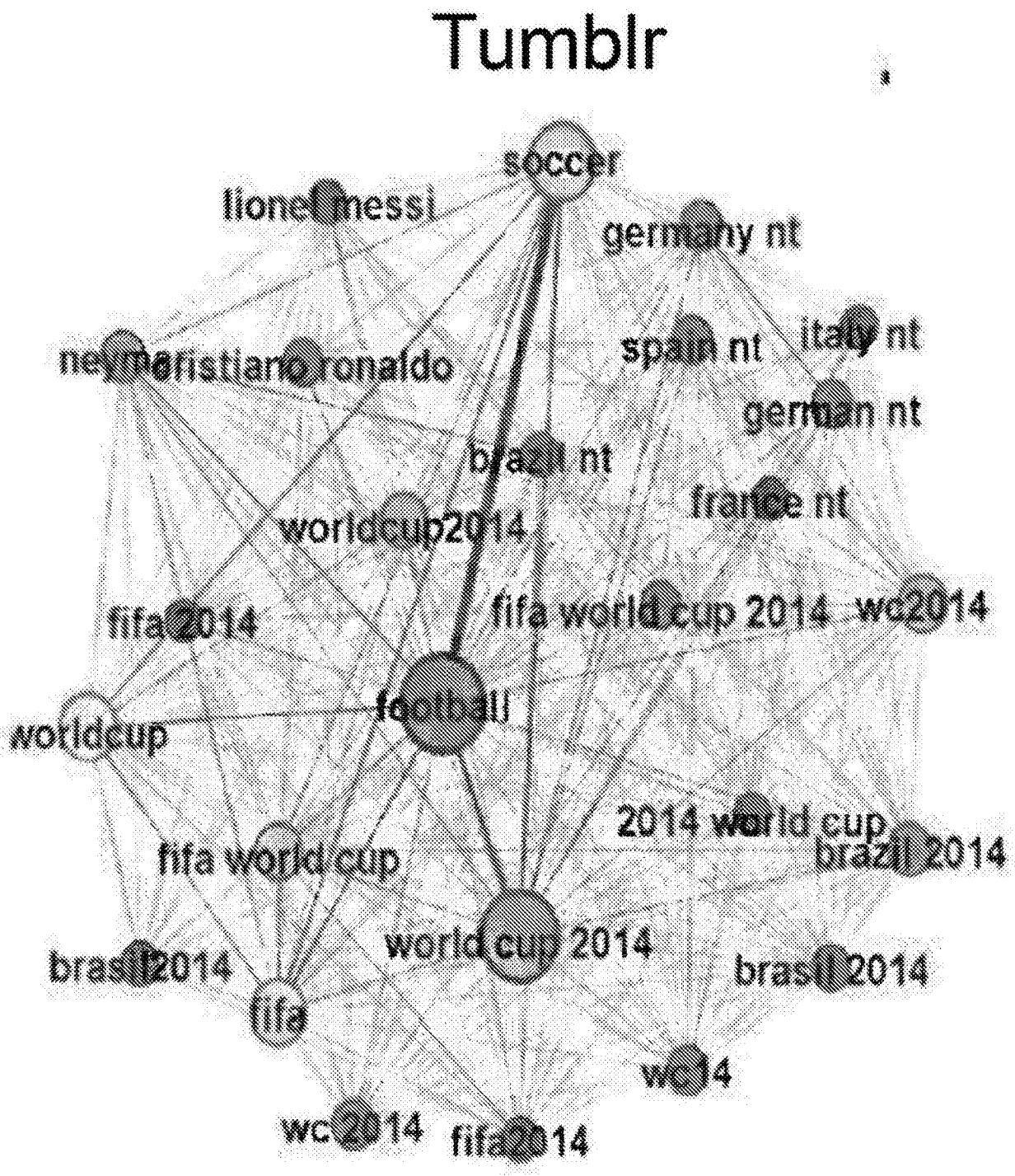

Tumblr
soccer
lionel messi
germany nt
spain nt
italy nt
cristiano ronaldo
german nt
brazil nt
france nt
worldcup2014
fifa 2014
fifa world cup 2014
wc2014
football
worldcup
2014 world cup
brazil 2014
fifa world cup
world cup 2014
brasil2014
brasil 2014
fifa
wc14
wc 2014
fifa2014
标签网络

图4C

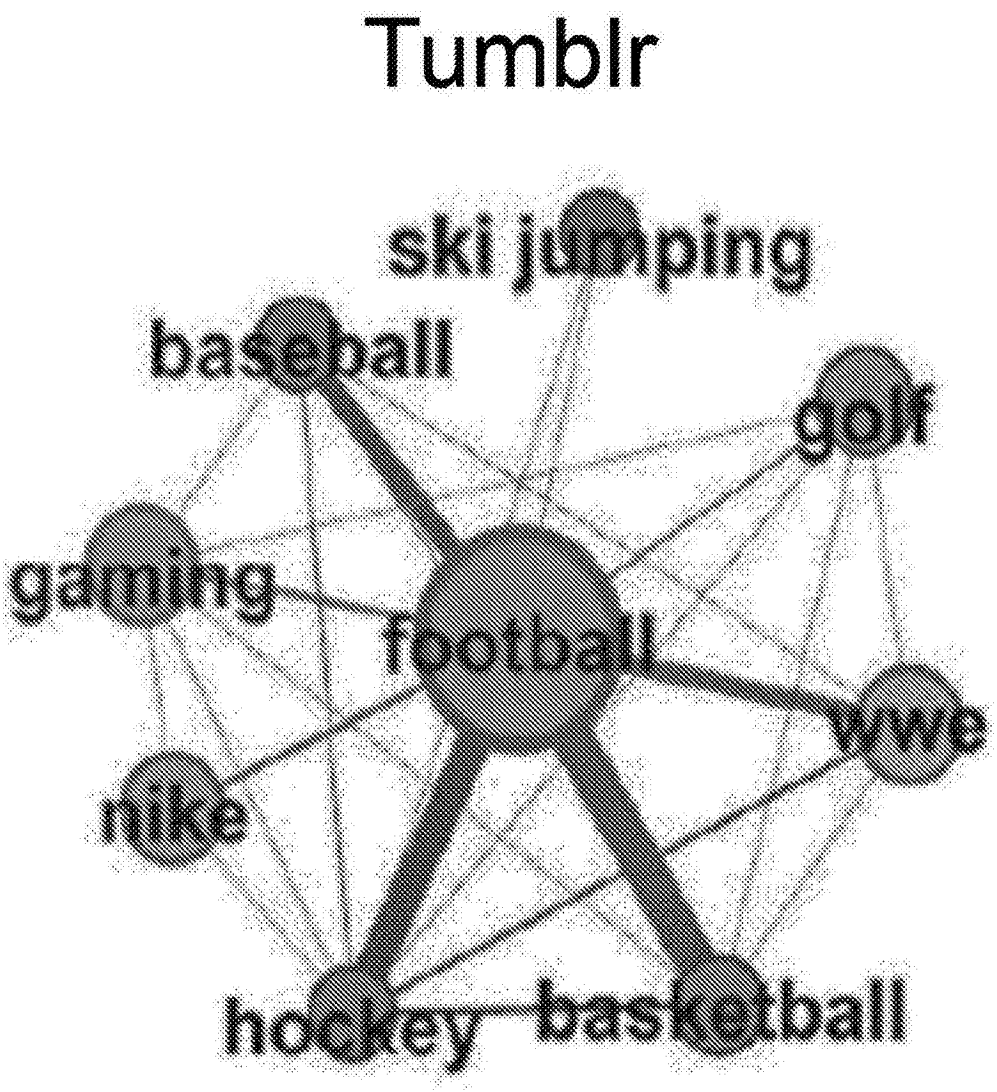

Tumblr
ski jumping
baseball
golf
gaming
football
wwe
nike
hockey
basketball
主题网络

图4D

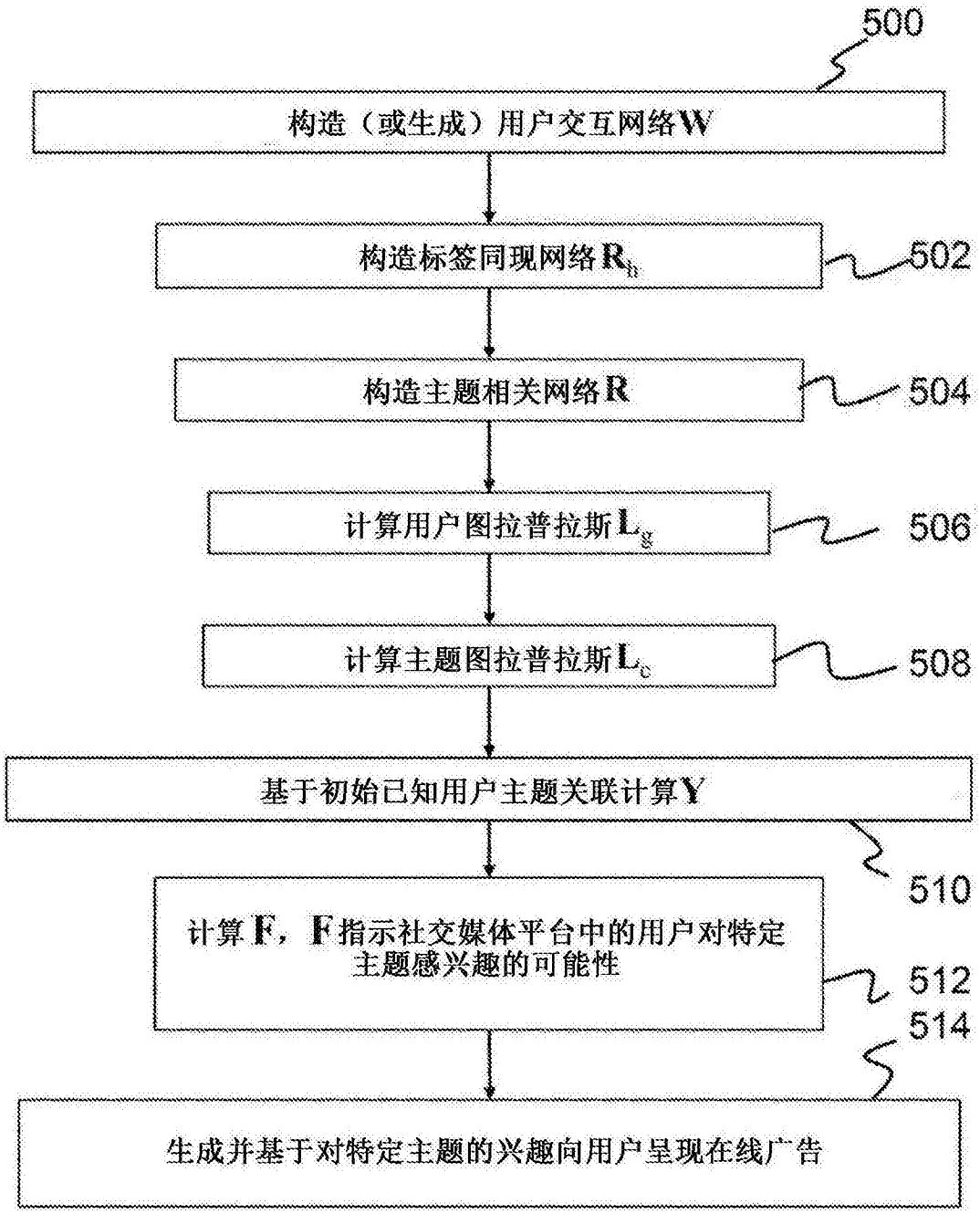

500
构造（或生成）用户交互网络W
构造标签同现网络R_{ti}
502
构造主题相关网络R
504
计算用户图拉普拉斯L_g
506
计算主题图拉普拉斯L_c
508
基于初始已知用户主题关联计算Y
510
计算F，F指示社交媒体平台中的用户对特定主题感兴趣的可能性
512
514
生成并基于对特定主题的兴趣向用户呈现在线广告

图5