(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 107729569 B
(45)授权公告日 2020.01.17

(21)申请号 201711143539.2

(22)申请日 2017.11.17

(65)同一申请的已公布的文献号
申请公布号 CN 107729569 A

(43)申请公布日 2018.02.23

(73)专利权人 杭州师范大学
地址 311121 浙江省杭州市余杭区余杭塘路2318号杭州师范大学

(72)发明人 张子柯 许帅帅 尤志强 周鸽 刘闯

(74)专利代理机构 杭州天正专利事务所有限公司 33201
代理人 王兵 黄美娟

(51)Int.Cl.
G06F 16/35(2019.01)
G06Q 10/04(2012.01)
G06Q 50/00(2012.01)

(56)对比文件
CN 105893611 A,2016.08.24,
US 2012317200 A1,2012.12.13,
CN 101075942 A,2007.11.21,
CN 106447505 A,2017.02.22,
莫靖杰等.基于多源信息融合的社交网络挖掘.《入选论文》.2017,(第9期),第73-76页.

审查员 程潇杰

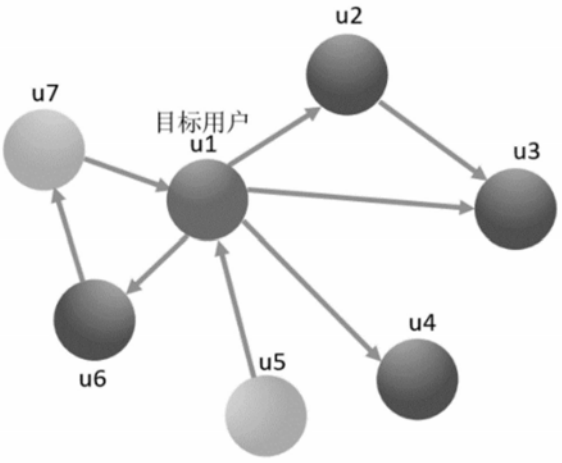
权利要求书2页 说明书6页 附图2页

(54)发明名称

一种融合网络结构和文本信息的社交关系预测方法

(57)摘要

一种融合网络结构和文本信息的社交关系预测方法,包括如下步骤:步骤1,构造原始兴趣向量;步骤2,构造修正兴趣向量;步骤3,重构用户兴趣向量;步骤4,预测社会关系。本发明综合考虑了社交网络中的文本信息和结构信息,解决了类似于微博和推特等社交网络平台上的链路预测问题和推荐问题;给出了一个可以应用在不同社交平台上的链路预测问题的统一解决框架;由于应用了word2vec,IKanalyzer等开源包,采用了兴趣向量,修正兴趣向量,桥接点等机制,所以达到了很高的预测准确度;丰富了对于链路预测方法的认识和理解。



CN 107729569 B

1.一种融合网络结构和文本信息的社交关系预测方法,包括如下步骤:

步骤1,构造原始兴趣向量;

使用微博和推特数据集中的用户关注关系和用户的文本信息内容,首先使用开源分词工具Ik Analyzer对数据集中的所有文本信息,即所有用户发表的微博内容或者推文内容,进行关键词的提取;这样可以得到用来刻画每一个用户的一系列关键词;然后将分词工具得到的所有的词语使用word2vec开源工具进行聚类,设置聚类个数为N,即将这些词划分为N个类别,这样就得到了N个话题类别,那么对于每一个用户来说,现在可以得到一个维度是N的兴趣向量,该向量的具体计算方法如下:针对一个用户i,构建一个长度为N维且每一个维度取值都为0的初始兴趣向量,然后依次扫描属于用户i的所有的关键词,若某个关键词属于第j个话题类别,那么用户i的特征向量中的第j个维度的值加1;直到扫描完所有属于该用户的关键词,就可以得到该用户i的兴趣向量;在该兴趣向量中,得分越高的维度说明该用户对于该维度的话题有更多的关注度和兴趣,使用T_i来表示第i个用户的兴趣向量,其分量具体的计算公式如下:

$$t_{ij}=\frac{Freq_{ij}}{\sum_j^n Freq_{ij}}$$

其中t_{ij}表示第i个用户在第j个话题上的得分,$Freq_{ij}$表示第i个用户的所有关键词出现在第j个话题上的数量,$\sum_j^n Freq_{ij}$表示第i个用户在所有的关键词出现在所有的话题上的数量,$t_{ij}=\frac{Freq_{ij}}{\sum_j^n Freq_{ij}}$为归一化项;

步骤2,构造修正兴趣向量;

使用用户的关注者的兴趣向量修正用户本身的原始兴趣向量;具体方法如下:针对一个特定的目标用户u1,该目标用户u1所有关注的用户是u2和u3,并且只关心用户兴趣向量中取值最大的top-K个维度,那么可以得到目标用户u1的兴趣向量$T_{u1}=(t_{11},t_{12},\cdots,t_{1K})$,用户u2的兴趣向量为$T_{u2}=(t_{21},t_{22},\cdots,t_{2K})$,以及用户u3的兴趣向量为$T_{u3}=(t_{31},t_{32},\cdots,t_{3K})$,那么目标用户u1的兴趣向量的修正的方法为将$t_{11},t_{12},\cdots,t_{1K},t_{21},t_{22},\cdots,t_{2K},t_{31},t_{32},\cdots,t_{3K}$中相同的维度上的值相加,不同的维度上的值全部保留而得到的结果;通过这样的方法可以得到用户的修正兴趣向量;

步骤3,重构用户兴趣向量;

在得到了目标用户的修正兴趣向量之后,需要对目标用户和潜在的目标用户的关注用户的兴趣向量进行重构,目标用户u1的修正兴趣向量为$T_{u1}=(t_{11},t_{12},\cdots,t_{1n})$,其中的n的取值小于等于原始兴趣向量的维度N并且大于等于在修正兴趣向量模块中取top-K个维度中选取的K值;目标用户u1的潜在关注用户u5的原始兴趣向量为$T_{u5}=(t_{51},t_{52},\cdots,t_{5N})$;首先如在修正兴趣向量模块中所述,抽取该用户的top-K,这里K取值为4,即值最大的前4个维度组成新的兴趣向量,$T_{u5}=(t_{51},t_{52},\cdots,t_{5K})$;然后考虑u1的修正兴趣向量和u5的Top-4兴趣向量的维度的并集,即,$(t_{11},t_{12},\cdots,t_{1n})\cup(t_{51},t_{52},\cdots,t_{5K})$;并按照并集的结果重新分别构造u1和u5的兴趣向量,若某一个用户没有某一个维度上的特征,则使用0补齐,这样就得到了目标用户和目标用户的潜在关注用户的重构的兴趣向量;

步骤4,预测社会关系;

对于给定的目标用户ui和ui的潜在关注用户uj,定义关注uj并且同时是ui的关注者为ui到uj的桥接点;将微博数据集和推特数据集随机的划分为两个部分,分别用作训练集和测试集;训练集中包括已知连边的90%;这样,对于测试集中的任一条边E_{ij},通过构建该边E_{ij}所连接的两个用户的修正兴趣向量和识别这两个用户之间的桥节点,即综合考虑通过修正兴趣向量对文本信息的利用和桥节点对网络结构的利用,得到如下的用于计算用户ui关注用户uj的概率计算公式,也就是边E_{ij}存在的概率:

$$P_{ij}=\sum_{k\in s_{if}}\left(I_{kj}\times\left(\vec{A}\cdot\vec{B}\right)\right)+\frac{I_{\vec{A}}\cdot I_{\vec{B}}}{K}$$

其中,S_{if}表示用户ui关注的所有对象;任何一个属于S_{if}的用户k,如果该用户k也关注了用户j,那么$I_{kj}=1$;否则$I_{kj}=0$;$I_{\vec{A}}$是值为0或者1的二元向量,该向量中每个维度上的值由向量A决定,如果向量A在该维度上的权值为正,那么$I_{\vec{A}}$在这个维度上的值为1;否则为0;所以$I_{\vec{A}}\cdot I_{\vec{B}}$表示用户ui和用户uj的兴趣点重叠的个数。

一种融合网络结构和文本信息的社交关系预测方法

技术领域

[0001] 本发明涉及社交网络上的关系预测，尤其适用于解决在类似于微博这样不仅仅有网络结构信息而且还包含丰富的文本信息的网络上的链路预测问题。

背景技术

[0002] 链路预测由于其在复杂网络、社会网络和生物网络等领域的广泛应用，已经吸引了各个领域研究者的关注。链路预测的目标是根据网络中已观测到的部分信息来估计我们尚未观察到的边的存在可能性。迄今为止，链路预测算法已经成功地应用于从生物学到电子商务的许多领域。例如，使用有效的链路预测方法可以给出蛋白质-蛋白质相互作用网络中最有可能存在的连边，这样就不用对每一个可能的连边进行实验，大大降低了实验成本。链路预测方法也可以用于推荐，最近的研究结果表明它们比标准的协同过滤算法的表现更好。

[0003] 在微博这样的社交网络平台上，如果可以准确的预测用户之间的关注关系，这将有助于帮助新用户构建其社交圈，并且也将增强用户的参与感。对于这样的平台来说这是至关重要的。在网络科学领域，一系列基于节点属性和网络拓扑结构的链路预测方法已经被提出。其中基于局部相似性指标的方法包括common neighbors，Jaccard coefficient和Adamic/Adar。例如common neighbors计算用户间共有的邻居数量，因为基于经验可以发现，拥有更多共同邻居的用户之间更容易存在连边。考虑全局的网络拓扑结构信息的链路预测算法包括Katz，Hitting Time，Commute Time，local random walk等等。然而这些已经存在的方法大多是基于只有网络拓扑结构信息可用而没有文本信息可用的网络。通过对微博和推特数据的分析发现，有关注关系的用户之间存在共同的兴趣，信息传播中的关键节点有助于提高信息的扩散。另外根据以往的研究表明，人们通常会有在社交平台上表达情感和展现意愿的倾向，这将有利于我们收集用于描述用户兴趣的有用信息。基于以上的讨论本分发明提出了一种新的算法，称为Maximum Preference on Interest Similarity (MPIS)，该算法充分利用了文本内容和网络结构信息来解决社交网络上的链路预测问题。

[0004] 针对类似于推特和微博这样的社交网络，由于其不同于传统的只有网络结构信息的网络，经典的链路预测方法不能够有效的利用其所包含的丰富的文本信息，这将会导致大量有用信息的丢失，降低链路预测的效果。

发明内容

[0005] 本发明要解决现有技术只考虑网络拓扑结构而忽略文本信息，以及计算量大，计算效率低下的缺点，提供一种基于网络结构和文本信息的社交关系预测方法。

[0006] 本发明利用微博和推特社交网络平台上拥有的数量丰富的用户文本信息，结合网络拓扑结构信息发明了一种链路预测的方法。在技术上实现了对用户关注关系的预测问题，丰富了对于链路预测问题的认识和理解。

[0007] 一种融合网络结构和文本信息的社交关系预测方法，包括如下步骤：

[0008] 步骤1,构造原始兴趣向量;

[0009] 本发明提出的方法主要使用微博和推特数据集中的用户关注关系和用户的文本信息内容,首先使用开源分词工具Ik Analyzer对数据集中的所有文本信息,即所有用户发表的微博内容或者推文内容,进行关键词的提取。这样可以得到用来刻画每一个用户的一系列关键词。然后将分词工具得到的所有的词语使用word2vec开源工具进行聚类,设置聚类个数为N,即将这些词划分为N个类别,这样就得到了N个话题类别.那么对于每一个用户来说,现在可以得到一个维度是N的兴趣向量,该向量的具体计算方法如下:针对一个用户i,构建一个长度为N维且每一个维度取值都为0的初始兴趣向量,然后依次扫描属于用户i的所有的关键词,若某个关键词属于第j个话题类别,那么用户i的特征向量中的第j个维度的值加1。直到扫描完所有属于该用户的关键词,就可以得到该用户i的兴趣向量。在该兴趣向量中,得分越高的维度说明该用户对于该维度的话题有更多的关注度和兴趣.使用T_i来表示第i个用户的兴趣向量,其分量具体的计算公式如下:

[0010] $$t_{ij}=\frac{Freq_{ij}}{\sum_j^n Freq_{ij}}$$

[0011] 其中t_{ij}表示第i个用户在第j个话题上的得分,$Freq_{ij}$表示第i个用户的所有关键词出现在第j个话题上的数量,$\sum_j^n Freq_{ij}$表示第i个用户在所有的关键词出现在所有的话题上的数量,该项为归一化项.

[0012] 步骤2,构造修正兴趣向量;

[0013] 通过对微博和推特的数据分析发现,尽管用户会主动给自己贴一些标签,但是如果仅仅使用用户自己给出的标签来刻画用户的兴趣会导致大量的信息丢失并且会存在大量的噪声。同样如果只是使用兴趣向量构造模块所构建的用户本身的原始兴趣向量来描绘用户兴趣,也会导致大量的信息缺失。另外由于用户经常可能会发布一些例如吃晚饭等等对于描述用户的真实兴趣向量没有贡献甚至会形成噪声的信息,所以只使用用户兴趣向量中权重比较大的top K个维度来描述用户会得到更为准确的结果。基于上述讨论并且通过更近一步的研究发现,用户的关注者的兴趣向量可以很好的用来修正用户本身的原始兴趣向量。该修正的方法如下:假设针对一个特定的目标用户u1,该目标用户u1所有关注的用户是u2和u3,并且假设只关心用户兴趣向量中取值最大的top-K个维度,那么可以得到目标用户u1的兴趣向量$T_{u1}=(t_{11},t_{12},\cdots,t_{1K})$,用户u2的兴趣向量为$T_{u2}=(t_{21},t_{22},\cdots,t_{2K})$,以及用户u3的兴趣向量为$T_{u3}=(t_{31},t_{32},\cdots,t_{3K})$,那么目标用户u1的兴趣向量的修正的方法为将$t_{11},t_{12},\cdots,t_{1K},t_{21},t_{22},\cdots,t_{2K},t_{31},t_{32},\cdots,t_{3K}$中相同的维度上的值相加,不同的维度上的值全部保留而得到的结果。通过这样的方法可以得到用户的修正兴趣向量,该向量由于融合了目标用户的关注者的兴趣特征而可以更加准确和全面的描述目标用户。

[0014] 步骤3,重构用户兴趣向量;

[0015] 在得到了目标用户的修正兴趣向量之后,需要对目标用户和潜在的目标用户的关注用户的兴趣向量进行重构,假设目标用户u1的修正兴趣向量为$T_{u1}=(t_{11},t_{12},\cdots,t_{1n})$,其中的n的取值小于等于原始兴趣向量的维度N并且大于等于在修正兴趣向量模块中取top-K个维度中选取的K值。假设目标用户u1的潜在关注用户u5的原始兴趣向量为$T_{u5}=(t_{51},t_{52},\cdots,t_{5N})$。首先如在修正兴趣向量模块中所述,抽取该用户的top-K,这里K取值为4,即值最大的前4个维度组成新的兴趣向量,假设为$T_{u5}=(t_{51},t_{52},\cdots,t_{5K})$。然后考虑u1的修正兴趣

向量和u5的Top-4兴趣向量的维度的并集，即，$(t_{11},t_{12},\cdots,t_{1n})\cup(t_{51},t_{52},\cdots,t_{5K})$。并按照并集的结果重新分别构造u1和u5的兴趣向量，若某一个用户没有某一个维度上的特征，则使用0补齐，这样就得到了目标用户和目标用户的潜在关注用户的重构的兴趣向量。

[0016] 步骤4，预测社会关系；

[0017] 考虑到网络结构对于社会关系预测的作用，本发明引入了桥节点的概念来利用网络的结构信息。对于给定的目标用户ui和ui的潜在关注用户uj，定义关注uj并且同时是ui的关注者为ui到uj的桥接点。通过实验研究发现，桥接点对于信息的传播有着非常重要的影响。如果在ui到uj之间桥节点的个数很多，信息越有可能从uj传到ui，即直观上来讲，桥节点可以放大信息的传播。基于以上的讨论，在这里提出一种社交网络上的链路预测的算法Maximum Preference on Interest Similarity(MPIS)，用于预测边E_{ij}存在的可能性，即预测用户ui是否会关注用户uj。该算法同时考虑了网络的结构信息和网络中包含的文本信息。为了测试算法的表现，将微博数据集和推特数据集随机的划分为两个部分，分别用作训练集和测试集。训练集中包括已知连边的90％。这样，对于测试集中的任一条边E_{ij}，我们通过构建该边所连接的两个用户的修正兴趣向量和识别这两个用户之间的桥节点，即综合考虑通过修正兴趣向量对文本信息的利用和桥节点对网络结构的利用，得到如下的用于计算用户ui关注用户uj的概率计算公式，也就是边E_{ij}存在的概率：

[0018] $$P_{ij}=\sum_{k\in s_{if}}\left(I_{kj}\times(\vec{A}\cdot\vec{B})\right)+\frac{I_{\vec{A}}\cdot I_{\vec{B}}}{K}$$

[0019] 其中，S_{if}表示用户ui关注的所有对象。任何一个属于S_{if}的用户k，如果该用户k也关注了用户j，那么$I_{kj}=1$；否则$I_{kj}=0$。$I_{\vec{A}}$是值为0或者1的二元向量，该向量中每个维度上的值由向量A决定，如果向量A在该维度上的权值为正，那么$I_{\vec{A}}$在这个维度上的值为1；否则为0。所以$I_{\vec{A}}\cdot I_{\vec{B}}$表示用户ui和用户uj的兴趣点重叠的个数。

[0020] 本发明的优点是：综合考虑了社交网络中的文本信息和结构信息，解决了类似于微博和推特等社交网络平台上的链路预测问题和推荐问题；给出了一个可以应用在不同社交平台上的链路预测问题的统一解决框架；由于应用了word2vec，IKanalyzer等开源包，采用了兴趣向量，修正兴趣向量，桥接点等机制，所以达到了很高的预测准确度；丰富了对于链路预测方法的认识和理解。

附图说明

[0021] 图1给出了用户之间的关注关系网络图，图中共有7个用户，图中箭头的方向表示关注的方向，如图所示，设定u1为目标用户，根据箭头的方向可知其关注的用户有u2，u3，u4和u6.从图中可以看出，目前尚没有已知的信息来表明是否用户u1关注了用户u5，即没有从目标用户u1指向用户u5的箭头。

[0022] 图2给出了对于目标用户的兴趣向量进行修正的计算过程。由于在图1中已知目标用户u1所关注的用户的是u2，u3，u4和u6。所以要使用u2，u3，u4和u6的兴趣向量来对u1进行修正。在图2(a)中给出了所有用户的原始兴趣向量，在该示例中假设用户原始兴趣向量的维度为10，并且对每一个用户的兴趣向量都已经做了归一化处理，例如假设用户u1的原始的兴趣向量为(0.02,0.12,0.091,0.21,0.002,0.006,0.05,0.3,0.14,0.061)；用户u2的原

始的兴趣向量为(0.15,0.019,0.23,0.22,0.001,0.03,0,0.022,0.13,0.198)。图2(b)中首先选取目标用户u1和目标用户的关注用户u2,u3,u4和u6的兴趣向量中权值最大的4个维度,构成新的用户的兴趣向量。例如此时用户u1的新的兴趣向量为(0.21,0.14,0.12,0.091),对应的特征的维度为(4,9,2,3);用户u2的新的兴趣向量为(0.23,0.22,0.198,0.15),对应的特征的维度为(3,4,10,1)。图2(c)中对图2(b)的结果进行归一化处理,即首先将每一个用户的新的兴趣向量求和,然后将各个维度除以求和所得的结果,得到归一化之后的兴趣向量,例如经过上述计算得到u1的兴趣向量(0.374,0.250,0.214,0.162);u2的兴趣向量(0.288,0.276,0.248,0.188)。图2(d)通过将图2(c)中各个向量的相应的维度相加得到了用户u1的最终的修正兴趣向量,因为此时u1保留的兴趣向量的维度为(4,9,2,3),u2为(3,4,10,1),u3为(4,6,1,9),u4为(1,4,3,2),u6为(8,3,1,9),所以最终的u1的修正的兴趣向量的维度为(4,9,2,3)∪(3,4,10,1)∪(4,6,1,9)∪(1,4,3,2)∪(8,3,1,9)=(1,2,3,4,6,8,9,10),并且各个维度相应的值通过计算可以得到为(1.007,0.35,0.919,1.348,0.25,0.309,0.569,0.248)。图2(e)中,为了计算用户u1是否会关注u5,需要对u1和u5的兴趣向量进行重构。在重构之前,针对用户u5的原始的兴趣向量,首先提取值最大的前4个维度得到(0.21,0.134,0.131,0.12),对应的维度是(1,6,9,10)。然后考虑u1的修正兴趣向量和u5的Top-4兴趣向量的维度的并集,即,(1,2,3,4,6,8,9,10)∪(1,6,9,10)=(1,2,3,4,6,8,9,10)。并重新分别构造u1和u5的兴趣向量,如图2(f)所示,重构之后的u1和u5的兴趣向量分别为(0.21,0,0,0,0.134,0,0.131,0.12)和(1.007,0.35,0.919,1.348,0.25,0.309,0.569,0.248)。

[0023] 图3中给出了桥接点的示意图,当想要预测u1是否会关注u4,这时在u1所有关注的用户中,那些关注u4的用户被称为是桥接点。所以在该图中u2和u3用户被称为桥接点。

具体实施方式

[0024] 下面结合附图,进一步说明本发明的技术方案。

[0025] 一种融合网络结构和文本信息的社交关系预测方法,包括如下步骤:

[0026] 步骤1.构造原始兴趣向量;

[0027] 针对微博和推特数据集,分别使用开源分词工具Ik Analyzer对采样得到的数据集中的所有文本信息,即所有用户发表的微博内容或者推文内容,进行关键词的提取。这样可以得到用来刻画每一个用户的一系列关键词。然后将分词工具得到的所有的词语使用word2vec开源工具进行聚类,设置聚类个数为N,即将这些词划分为N个类别,这样就得到N个话题类别.那么对于每一个用户来说,现在可以得到一个维度是N的兴趣向量来描述该用户,该向量的具体计算方法如下:针对一个用户i,构建一个长度为N维且每一个维度取值都为0的初始兴趣向量,然后依次扫描属于用户i的所有的关键词,若某个关键词属于第j个话题类别,那么用户i的特征向量中的第j个维度的值加1。直到扫描完所有属于该用户的关键词,就可以得到该用户i的兴趣向量。在该兴趣向量中,得分越高的维度说明该用户对于该维度的话题有更多的关注度和兴趣.使用T_i来表示第i个用户的兴趣向量,其分量具体的计算公式如下:

[0028] $$t_{ij}=\frac{Freq_{ij}}{\sum_j^n Freq_{ij}}$$

[0029] 其中t_{ij}表示第i个用户在第j个话题上的得分,$Freq_{ij}$表示第i个用户的所有关键词出现在第j个话题上的数量,$\sum_j^n Freq_{ij}$表示第i个用户在所有的关键词出现在所有的话题上的数量,该项为归一化项.

[0030] 步骤2.构造修正兴趣向量;

[0031] 图1给出了用户之间的关注关系网络图,其中箭头的方向表示关注方向。在本示例中,针对目标用户u1进行兴趣向量的修正。从图中可以得到目标用户u1关注的用户有u2,u3,u4和u6.这四个用户用来对目标用户u1的兴趣向量做修正,并且从图中可得,u1用户尚没有关注用户u5。接下来将利用本发明给出的方法演示计算目标用户u1关注u5的可能性大小的过程,即计算边E_{15}存在的概率的大小。

[0032] 在图2中给出了具体的用户兴趣向量修正的过程:在图2(a)中给出了所有用户的原始兴趣向量,在该示例中假设用户原始兴趣向量的维度为10,并且对每一个用户的兴趣向量做归一化处理,例如用户u1的原始的兴趣向量为(0.02,0.12,0.091,0.21,0.002,0.006,0.05,0.3,0.14,0.061);用户u2的原始的兴趣向量为(0.15,0.019,0.23,0.22,0.001,0.03,0,0.022,0.13,0.198)。图2(b)中首先选取目标用户u1和目标用户的关注用户u2,u3,u4和u6的兴趣向量中权值最大的4个维度,构成新的用户的兴趣向量。例如此时用户u1的新的兴趣向量为(0.21,0.14,0.12,0.091),对应的特征的维度为(4,9,2,3);用户u2的新的兴趣向量为(0.23,0.22,0.198,0.15),对应的特征的维度为(3,4,10,1)。图2(c)中对图2(b)的结果进行归一化处理,使得每一个用户的所有的特征向量的值的和为1,具体做法是首先将每一个用户的新的兴趣向量求和,然后将各个维度除以求和所得的结果,得到归一化之后的兴趣向量,例如经过上述计算得到u1的兴趣向量(0.374,0.250,0.214,0.162);u2的兴趣向量(0.288,0.276,0.248,0.188)等等。图2(d)通过将图2(c)中各个向量的相应的维度相加得到了用户u1的最终的修正兴趣向量,因为此时u1保留的兴趣向量的维度为(4,9,2,3),u2为(3,4,10,1),u3为(4,6,1,9),u4为(1,4,3,2),u6为(8,3,1,9),所以最终的u1的修正的兴趣向量的维度为(4,9,2,3)∪(3,4,10,1)∪(4,6,1,9)∪(1,4,3,2)∪(8,3,1,9)=(1,2,3,4,6,8,9,10),并且各个维度相应的值为(1.007,0.35,0.919,1.348,0.25,0.309,0.569,0.248)。通过分析可以得到:假设修正后的兴趣向量的维度为n,那么n的取值范围为:10≥n≥4。

[0033] 步骤3.重构用户兴趣向量;

[0034] 在得到了目标用户u1的修正兴趣向量之后,需要对目标用户u1和用户u5的兴趣向量进行重构,如图2(d)所示,通过前面的分析得到目标用户u1的修正的兴趣向量为(1.007,0.35,0.919,1.348,0.25,0.309,0.569,0.248),该兴趣向量对应的维度为(1,2,3,4,6,8,9,10)。从图2(e)中可以看到用户u5的原始兴趣向量为(0.21,0.003,0.05,0.11,0.02,0.134,0.112,0.11,0.131,0.12),针对u5用户,使用步骤2中方法,首先选取值最大的4个维度是(1,6,9,10),其对应的值为(0.21,0.134,0.131,0.12)。然后考虑u1的修正兴趣向量和u5的Top-4兴趣向量的维度的并集,即,(1,2,3,4,6,8,9,10)∪(1,6,9,10)=(1,2,3,4,6,8,9,10)。并重新分别构造u1和u5的兴趣向量,若用户没有某一个维度上的特征,则使用0补齐,如图2(f)所示,重构之后的u1和u5的兴趣向量分别为(0.21,0,0,0,0.134,0,0.131,0.12)和(1.007,0.35,0.919,1.348,0.25,0.309,0.569,0.248),分别记这两个向量为$\vec{A}$和

$\vec{B}$。这两个兴趣向量涉及到的特征的维度都是(1,2,3,4,6,8,9,10)

[0035] 步骤4.预测社会关系；

[0036] 通过上面的分析得到u1的最终的兴趣向量为$\vec{A}$=(0.21, 0, 0, 0, 0.134, 0, 0.131, 0.12),u5的最终的兴趣向量为$\vec{B}$=(1.007, 0.35, 0.919, 1.348, 0.25, 0.309, 0.569, 0.248)。通过计算$\vec{A}$和$\vec{B}$的内积可以得到用户u1和u5的兴趣相似性。另外我们也考虑了$\vec{A}$和$\vec{B}$中重叠出现的兴趣分量的个数,通过该指标可以反映用户兴趣点的联系率,这在一定的程度上弥补了内积只衡量相似性的强度的缺点。除了兴趣相似性,在对连边进行预测时候,桥接点的作用也被考虑了进来。因为桥节点的个数越多,信息越有可能从u5传到u1。综合上述的考虑,u1关注u5的概率P_{15}可以通过如下的计算公式给出,这里给出更为一般的计算公式,即,用户i关注用户j的概率的计算公式,在本例中,i即为用户u1,j即为用户u5：

[0037] $$P_{ij}=\sum_{k\in s_{if}}\left(I_{kj}\times(\vec{A}\cdot\vec{B})\right)+\frac{I_{\vec{A}}\cdot I_{\vec{B}}}{K}$$

[0038] 其中,S_{if}表示用户i关注的人,在本例中即为u1的关注的人u2,u3,u4和u6。如果这些用户也关注了用户u5,那么I_{kj}=1;否则I_{kj}=0，$I_{\vec{A}}$是值为0或者是1的二元向量。如果某个维度上的权值为正,那么$I_{\vec{A}}$在这个维度上的值为1;否则为0。所以$I_{\vec{A}}\cdot I_{\vec{B}}$表示用户u1和用户u5的兴趣点重叠的个数。最终的计算结果Pij即为用户u1关注u5的概率的大小。

[0039] 本说明书实施例所述的内容仅仅是对发明构思的实现形式的列举,本发明的保护范围不应当被视为仅限于实施例所陈述的具体形式,本发明的保护范围也及于本领域技术人员根据本发明构思所能够想到的等同技术手段。

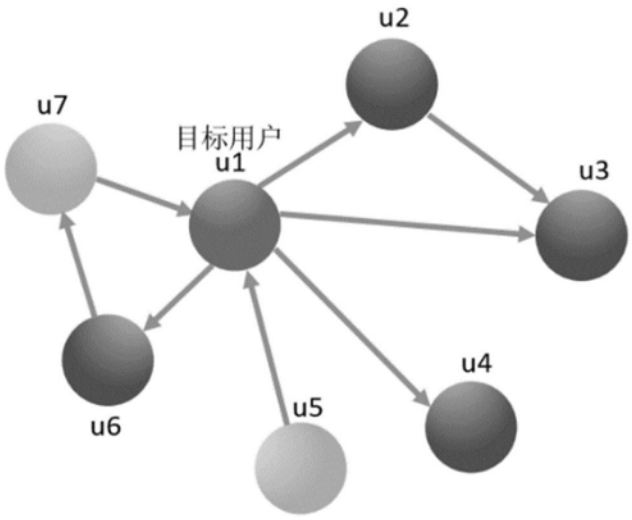

u2
u7
目标用户
u1
u3
u6
u5
u4

图1

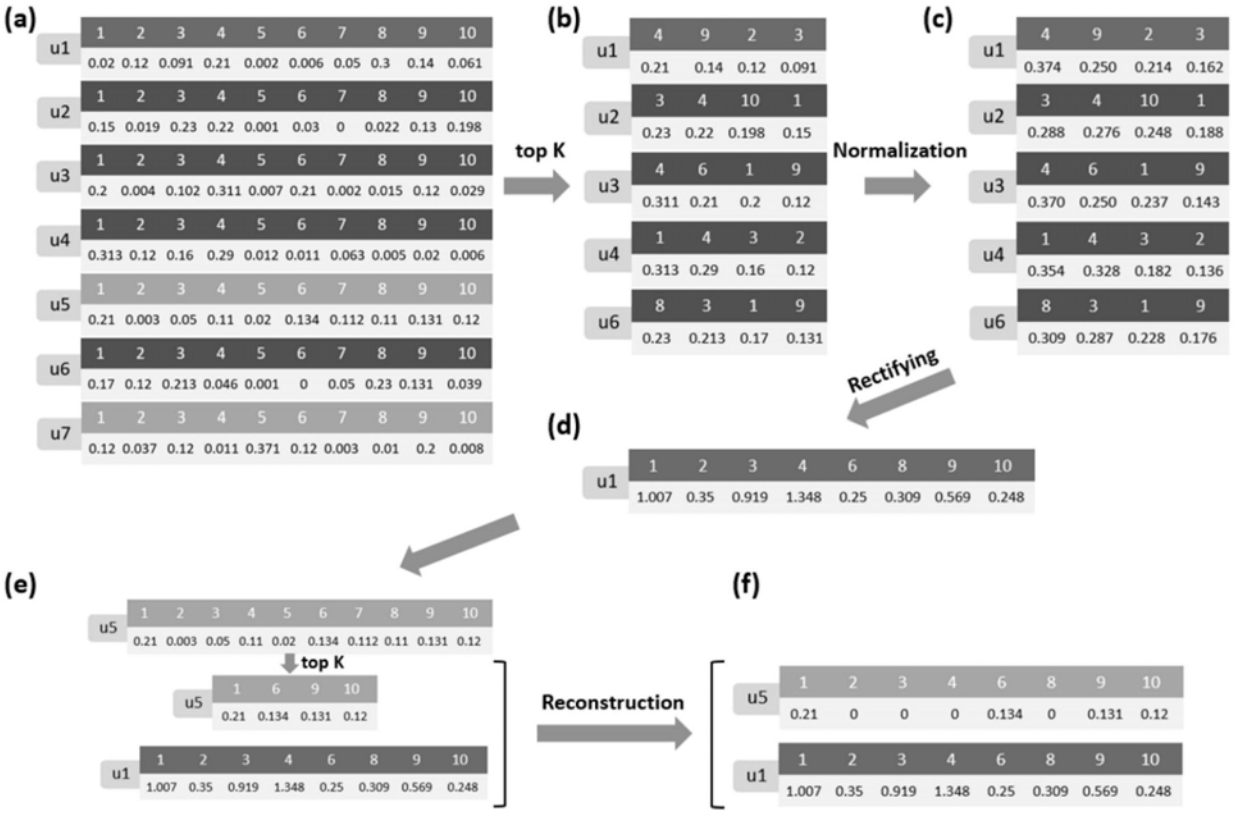

(a)
u1 1 2 3 4 5 6 7 8 9 10
0.02 0.12 0.091 0.21 0.002 0.006 0.05 0.3 0.14 0.061
u2 1 2 3 4 5 6 7 8 9 10
0.15 0.019 0.23 0.22 0.001 0.03 0 0.022 0.13 0.198
u3 1 2 3 4 5 6 7 8 9 10
0.2 0.004 0.102 0.311 0.007 0.21 0.002 0.015 0.12 0.029
u4 1 2 3 4 5 6 7 8 9 10
0.313 0.12 0.16 0.29 0.012 0.011 0.063 0.005 0.02 0.006
u5 1 2 3 4 5 6 7 8 9 10
0.21 0.003 0.05 0.11 0.02 0.134 0.112 0.11 0.131 0.12
u6 1 2 3 4 5 6 7 8 9 10
0.17 0.12 0.213 0.046 0.001 0 0.05 0.23 0.131 0.039
u7 1 2 3 4 5 6 7 8 9 10
0.12 0.037 0.12 0.011 0.371 0.12 0.003 0.01 0.2 0.008
top K
(b)
u1 4 9 2 3
0.21 0.14 0.12 0.091
u2 3 4 10 1
0.23 0.22 0.198 0.15
u3 4 6 1 9
0.311 0.21 0.2 0.12
u4 1 4 3 2
0.313 0.29 0.16 0.12
u6 8 3 1 9
0.23 0.213 0.17 0.131
Normalization
(c)
u1 4 9 2 3
0.374 0.250 0.214 0.162
u2 3 4 10 1
0.288 0.276 0.248 0.188
u3 4 6 1 9
0.370 0.250 0.237 0.143
u4 1 4 3 2
0.354 0.328 0.182 0.136
u6 8 3 1 9
0.309 0.287 0.228 0.176
Rectifying
(d)
u1 1 2 3 4 6 8 9 10
1.007 0.35 0.919 1.348 0.25 0.309 0.569 0.248
(e)
u5 1 2 3 4 5 6 7 8 9 10
0.21 0.003 0.05 0.11 0.02 0.134 0.112 0.11 0.131 0.12
top K
u5 1 6 9 10
0.21 0.134 0.131 0.12
u1 1 2 3 4 6 8 9 10
1.007 0.35 0.919 1.348 0.25 0.309 0.569 0.248
Reconstruction
(f)
u5 1 2 3 4 6 8 9 10
0.21 0 0 0 0.134 0 0.131 0.12
u1 1 2 3 4 6 8 9 10
1.007 0.35 0.919 1.348 0.25 0.309 0.569 0.248

图2

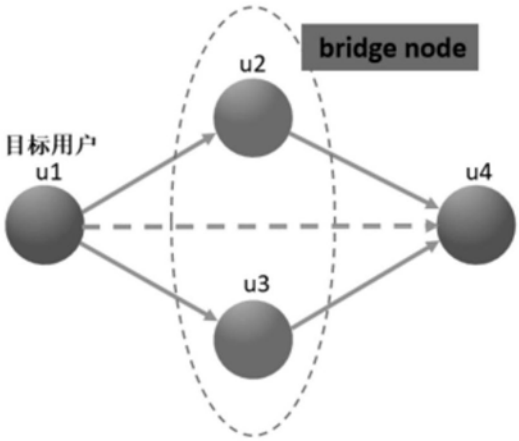


图3