(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 106709037 B
(45)授权公告日 2019.11.08

(21)申请号 201611248620.2

(22)申请日 2016.12.29

(65)同一申请的已公布的文献号
申请公布号 CN 106709037 A

(43)申请公布日 2017.05.24

(73)专利权人 武汉大学
地址 430072 湖北省武汉市武昌区珞珈山武汉大学

(72)发明人 余啸 刘进 殷晓飞 崔晓晖 杨威 井溢洋

(74)专利代理机构 武汉科皓知识产权代理事务所(特殊普通合伙) 42222
代理人 鲁力

(51)Int.Cl.
G06F 16/9535(2019.01)
G06F 16/9536(2019.01)
G06F 16/2458(2019.01)

(56)对比文件
CN 104077351 A,2014.10.01,
CN 104063481 A,2014.09.24,
CN 104077351 A,2014.10.01,
US 2013036121 A1,2013.02.07,

审查员 殷飞

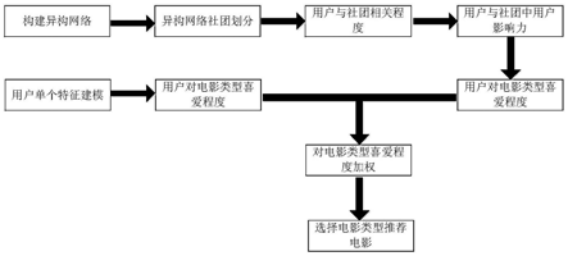
权利要求书3页 说明书10页 附图3页

(54)发明名称

一种基于异构信息网络的电影推荐方法

(57)摘要

一种基于异构信息网络的电影推荐方法,包括链接建模,以用户、电影、电影类型三种类型的对象为结点构建用户-电影异构网络,以电影类型结点为中心,对异构网络进行社团划分,筛选出符合要求的社团,提高其运算速度和效率,利用异构网络中的元路径,计算目标用户与社团中的每个用户之间的影响力,计算目标用户与电影类型的喜爱程度,特征建模,分析目标用户的每一个特征被划分到每种电影类型的概率,综合计算目标用户对每个电影类型的喜爱程度,对链接建模和特征建模的结果进行加权求和,分析目标用户对每一种电影类型的喜爱程度,按照目标用户对每一种电影类型的喜爱程度,选择电影类型,推荐评分高的电影。



CN 106709037 B

1.一种基于异构信息网络的电影推荐方法,其特征在于,包括以下步骤:

步骤1,链接建模,以用户、电影、电影类型三种类型的对象为结点构建用户-电影异构网络,并计算用户与电影类型之间的权值,具体是:

定义用户-电影异构网络模型G＝(V,E,W),其中V＝$V_u \cup V_m \cup V_t$,V_u表示用户集合,V_m表示电影集合,V_t表示电影类型集合,E＝$E_{uu} \cup E_{mm} \cup E_{tt} \cup E_{um} \cup E_{ut} \cup E_{mt}$,其中,$E_{uu}$表示用户与用户之间的链接关系,$E_{mm}$表示电影与电影之间的链接关系,$E_{tt}$表示电影类型与电影类型之间的链接关系,$E_{um}$表示用户与电影之间的链接关系,$E_{ut}$用户与电影类型之间的链接关系,$E_{mt}$表示电影与电影类型之间的链接关系;其中W表示六种链接关系的权重集合,仅考虑用户之间的链接关系、电影类型之间的链接关系以及用户与电影类型之间的链接关系这三种链接关系;

步骤2,以电影类型结点为中心,对异构网络进行社团划分,一种电影类型为一个社团,查找不同社团两两之间的共同结点,并计算电影类型之间的权值;包括:

步骤2.1,利用元路径搜索将异构网络中不同类型的结点以电影类型为中心进行划分,具有相似特征的结点被划分到一个社团当中,划分的结果是社团内结点具有高内聚,社团间结点具有低耦合的特征;划分社团的个数就是电影类型的个数,每一个社团包含该电影类型以及对该电影类型相关的电影进行过评分的用户即喜爱该电影类型相关电影的用户群,一个社团可以包含多个用户,一个用户也可能存在多个社团中,实际实施时用户结点与相邻的电影类型结点的权值可以使用数组a[|Mtype|][|Mtypei|]来存储,|Mtype|表示电影类型种数,|Mtypei|表示属于第i种类型电影中的用户数量,a[i][j]表示第i种电影类型与该类型电影中第j个用户之间的权值,

步骤2.2,基于 $W = \sum_{k=1,m_k \in M_i \cap M_j}^{|M_i \cap M_j|} \sum_{i=1}^{n} g_{k,i} / \sum_{k=1,m_k \in M_i \cup M_j}^{|M_i \cup M_j|} \sum_{j=1}^{m} g_{k,j}$ 计算电影类型之间的权值,其中$M_i \cap M_j$表示这两种电影类型共同的电影集合,m_k代表电影集合中的电影,而相应的$g_{k,i}$代表用户i对于电影m_k的评分;$M_i \cap M_j$表示这两种类型电影的并集;

步骤3,计算目标用户与每个社团即每个电影类型结点之间的相关程度,设置阀值,筛选出符合要求的社团;目标用户与每个社团之间的初始相关程度的计算与Dijkstra算法相似,唯一不同的在于,对于源点到其他结点的所有路径,Dijkstra算法得到的是最短距离,而在此计算的是所有路径长度的和;其中结点之间权值的计算分为两种,一种是用户结点与电影类型结点的权值,一种是电影类型结点之间的权值;

步骤4,对于每一个符合要求的社团,基于目标用户到电影类型结点的初始相关程度,计算目标用户与社团中的每个用户之间的影响力,其中,用户之间的影响力就是用户之间的相似性,其计算步骤为,确定与目标用户存在元路径的用户,计算元路径每一段路径的权值,根据元路径复合规则计算两用户之间的相似性;具体包括:

步骤4.1、两个电影用户之间的相互影响力计算依据用户间的元路径,用户之间的相互影响包括直接影响和间接影响;

所述间接影响为用户之间存在一条路径,用户影响力依靠路径影响彼此相邻结点;

所述直接影响为结点之间有链接关系,即彼此为相邻结点,那么彼此之间存在直接影响力;

两个结点之间的存在多条间接影响路径;对于两结点之间存在直接联系又有间接联

系，他们的相互影响是两种影响的叠加；这种影响力大小定义为公式计算，当i，j为不相邻结点时$sim(i,j)=\sum_{m=1}^{n}\sum_{k\in Nb(i,j)}sim(i,k)\oplus sim(k,j)$，当i，j相邻时，sim(i，j)＝$w_{i,j}$，其中n表示结点i，j之间的路径数目，m表示路径序号，Nb(i，j)表示结点i，j每一条路径的连接点结点集合，符号⊕表示影响力在路径中不同阶段的连接方式，因为结点间随着路径长度加长，影响会逐渐变弱，即路径越长影响力越小，当结点i，j是相邻结点时，使用步骤1中的链接关系的权值，当不是相邻结点时并且元路径较长时，就将元路径以某一结点分为两段，这样反复下去直到元路径只有两个相邻结点；

步骤4.2、元路径将用户之间的链接关系、电影类型之间的链接关系以及用户与电影类型之间的链接关系这三种类型的链接进行构建；在用户-电影异构网络中，对于用户与用户之间的间接关系，起到连接作用的是电影类型结点，当用户都喜爱某一种类型的电影，可以认为彼此之间能够产生影响；用户之间的影响力计算公式使用Sim(u_0,u_{ij})＝l(Mtype$_i$)*a[Mtype$_i$][j]来计算，其中，Mtype$_i$表示第i种电影类型，l(Mtype$_i$)表示目标用户u_0到第i种电影类型的初始相关程度，a[Mtype$_i$][j]表示该电影类型结点到该电影类型所代表的社团中的第j个用户u_{ij}的权值；

步骤5，计算目标用户与符合要求的社团所代表的电影类型的喜爱程度，用目标用户与社团中每一个用户之间的影响力的均值来表示，社团中各结点对目标用户影响力的平均值作为用户划分到该社团的概率，即目标用户对于该社团所代表的电影类型的喜爱程度，平均值越大，目标用户划分到该社团概率越大，喜爱对应类型电影可能性越高；令目标用户划分到社团c_k的概率为$P_2(c_k|u_0)$，其计算过程如公式$P_2(c_k|u_0)=\sum_{u_k\in M_i}sim(u_0,u_k)\Big/|c_k|$；其中，$u_k$属于社团$c_k$中的用户结点，|$c_k$|表示其中结点的个数，sim($u_0$,$u_k$)表示目标用户$u_0$和$u_k$之间的相互影响力；

步骤6，特征建模，基于经典朴素贝叶斯分类进行用户单个特征属性分类，分析目标用户的每一个特征被划分到每种电影类型的概率；具体包括：

步骤6.1，定义用户各个特征属性之间相互独立，对用户单个特征属性进行建模分类，目的在于分析用户单个特征属于每个电影类型的概率；在用户-电影异构网络中，对特征属性进行建模时，用X_v表示对象V的特征属性，相应地，针对文中的研究对象用户u_i来说，X_{ui}代表其特征属性集合；考虑电影用户有多个特征属性，因此X_{ui}是一个向量，表示为X_{ui}＝{$X_{ui,1}$，$X_{ui,2}$，...，$X_{ui,j}$，$X_{ui,n}$}，其中n＝|X_{ui}|为属性个数；

步骤6.2，目标用户单个特征属性属于某个电影类型的概率用P(c_k|$X_{ui,j}$)表示，P($X_{ui,j}$|c_k)表示在电影类型c_k中用户u_i的第j个属性$X_{ui,j}$所占的比率，$X_{ui,j}$|c_k服从高斯分布，即$X_{ui,j}|c_k\backsim N(u_k,\sigma_k^2)$，P($c_k$)表示属于电影类型$c_k$的电影占所有电影的比例，从数据中可以直接统计获取，P($X_{ui,j}$)表示属性$X_{ui,j}$的概率，定义所有属性概率相同，即P($X_{ui,j}$)＝P($X_{uj,i}$)，i≠j；由贝叶斯定理得两者之间关系如公式P(c_k|$X_{ui,j}$)＝P($X_{ui,j}$|c_k)×P(c_k)/P($X_{ui,j}$)对于单个特征属性$X_{ui,j}$建模完成以后，同样的，用户其他特征属性也相应的得到建模，从而每个用户的每个属性都得到了建模；依据贝叶斯定理可以得出每个用户的每个属性属于某个电影类型的概率；

步骤7，基于目标用户的每一个特征被划分的概率，综合计算目标用户对每个电影类型

的喜爱程度;具体包括:复合特征属性建模,综合考虑所有用户特征属性,从用户自身角度分析用户属于某个电影类型的概率,喜爱某种类型电影的可能性;是将每个用户的每个属性分类结果综合起来考虑,利用对用户属性的分类进而完成对用户的分类,其主要思想是如果用户所有属性同时划分到某一类型的概率比较大,那么目标用户划分到该类型的可能性就比较大,即认为目标用户喜爱对应类型的电影可能性就比较高;用$P(c_k|u_0)$表示u_0划分到电影类型c_k的概率;由用户各个特征属性之间相互独立,得到公式$P(c_k|u_0)=P(c_k|X_{u0})=\prod_{j=1}^{|X_{u0}|}P(c_k|X_{u0,j})$,$X_{u0,j}|c_k$服从高斯分布,即$X_{u0,j}|c_k \backsim N(u_k,\sigma_k^2)$,结合步骤6中目标用户的每一个特征被划分到每种电影类型的概率公式可以得到用户属于每一个电影类型的概率公式$P_1(c_k|u_0)=P(c_k|X_{u0})=\prod_{j=1}^{|X_{u0}|}\frac{P(X_{u0,j}|c_k)\times P(c_k)}{P(X_{u0,j})}=\prod_{j=1}^{|X_{u0}|}(\frac{P(c_k)}{P(X_{u0,j})}\times\frac{1}{\sqrt{2\pi\sigma_k^2}}e^{\frac{-(X_{u0,j}-u_k)^2}{2\sigma_k^2}})$。

步骤8,结合步骤5与步骤7中分别得出的用户对每一种电影类型的喜爱程度,进行加权求和,分析目标用户对每一种电影类型的喜爱程度;

步骤9,按照目标用户对每一种电影类型的喜爱程度,选择电影类型,推荐评分高的电影,具体是,选择用户喜爱程度最高的前k种电影类型,分别从这些电影类型中选择评分高于一定值的前m部电影推荐给目标用户。

2.根据权利要求1所述的一种基于异构信息网络的电影推荐方法,其特征在于,所述步骤6和步骤7中,使用朴素贝叶斯分类,根据用户特征建立了特征模型,计算出目标用户u_0喜爱电影类型c_k的概率子模型$P_1(c_k|u_0)$,步骤1到步骤5中,通过利用元路径和社团划分,对用户-电影异构网络进行了分析,建立了基于链接的模型,得到目标用户u_0喜爱电影类型c_k的概率子模型$P_2(c_k|u_0)$;这两种模型的加权决定了目标用户u_0对于电影类型c_k的喜爱程度;定义两种子模型的权重分别为α和β,得到统一概率模型$P(c_k|u_0)=\alpha*P_1(c_k|u_0)+\beta*P_2(c_k|u_0)$,其中$P_1(c_k|u_0)$和$P_2(c_k|u_0)$分别表示所述步骤6、步骤7和步骤1到步骤5中,通过两种不同方式计算出的目标用户u_0喜爱电影类型c_k的概率,α为非负数,表示特征模型所占权重,β为正常数,表示链接模型所占权重。

一种基于异构信息网络的电影推荐方法

技术领域

[0001] 本发明属于数据挖掘应用中个性化推荐技术领域，特别是涉及一种基于异构信息网络的电影推荐方法。

背景技术

[0002] (1)推荐系统

[0003] 随着互联网的迅速普及，大数据时代已经到来，随之而来的是信息过载问题，如何进行个性化的信息筛选和呈现是各类互联网应用领域中亟待解决的问题。采用科学的方法深度挖掘用户的兴趣并生成个性化推荐即构建个性化推荐系统，是解决这一问题的主要手段。

[0004] 个性化推荐系统基于用户特征、兴趣和历史行为数据构建用户信息模型，使用特定的推荐技术，进而挖掘用户个人偏好，生成对目标用户的推荐。目前，主要的推荐系统有协同过滤推荐，基于内容的推荐，基于网络的推荐等。

[0005] 1)协同过滤推荐

[0006] 协同过滤推荐的思想是，根据目标用户的历史偏好，为目标用户或项目找到相似项，根据相似项对项目进行评分并推荐给目标用户，即最近邻技术。协同过滤推荐是推荐系统中运用最成功的推荐技术之一，在各种环境下被广泛运用，许多不同领域算法和技术都相继用于改进和优化其性能。按照分析对象分类，协同过滤推荐有基于用户的最近邻推荐和基于项目的最近邻推荐。基于用户的最近邻推荐的基本思想：第一步，根据一个用户对项目的评分数据集，寻找与目标用户有相似偏好的其他用户，这些用户被称为最近邻；第二步，如果目标用户没有对某一项目评分，则依据最近邻过去对该项目的评分来预测目标用户对改项目的评分。基于项目的最近邻推荐的思想与此类似。协同过滤算法简单、高效且准确率高，然而协同过滤推荐是从用户历史数据出发的，对于没有评分过的用户和没有被评分过的项目，则永远无法被推荐，即冷启动问题，这就是协同过滤算法的固有缺点。另外，实际情况下，数据库中的评分数据往往是稀疏的，所以，也存在数据稀疏的问题。

[0007] 2)基于内容的推荐

[0008] 基于内容的推荐主要运用于信息过滤，与协同过滤不同的是，基于内容的推荐不以用户对项目的评分为依据，而是基于用户本身的特征来判断用户的偏好，匹配与用户偏好相似度高的项目，进行推荐。基于内容的推荐只需要项目和用户的特征信息，不需要大量的用户数量和历史评分数据，只需要对目标用户的特征信息进行提取分析，就可以进行项目匹配推荐，有效避免了由于数据稀松导致的误差，也避免了新用户或者新项目在没有历史评分记录的情况下存在的冷启动问题，另外，由于推荐是基于用户或者项目的特征产生的，方便向用户解释推荐理由。然而，现有的技术对于内容的分析也就是对于用户或者项目的特征的提取仅限于一些简单的文本内容，较为复杂的内容分析还存在一定的困难，所以基于内容的推荐对用户兴趣的挖掘深度有限，推荐准确度有限。而且用户对于项目的爱好会随着时间发生改变，基于内容的推荐很难向用户推荐一些新的可能会感兴趣的项目。

[0009] 3)基于网络的推荐

[0010] 典型复杂网络近年来在各个学科领域被广泛研究,逐渐成为一个独立的研究方向。随着复杂网络的研究技术逐渐走向成熟,越来越多的研究者正在尝试将复杂网络运用到推荐系统中,基于网络的推荐也就发展起来。相对于基于内容的推荐,基于网络的推荐不用用户或者项目的特征信息,而是使用用户和项目作为结点,使用用户和项目之间的关系作为边,构建一个网络图。一般而言,基于网络的推荐都是通过一些算法来挖掘网络路径中潜在的用户偏好,比如使用随即游走算法来计算用户之间的相似度,用户与用户之间的路径数量和路径长度就代表着用户之间的影响力。

[0011] (2)异构网络

[0012] 由单一研究对象构成的网络称为同构网络,而与之相对应的则是由多种研究对象构成的异构网络。异构网络符合现实世界的关系模型,更容易包含用户之间、用户与项目之间的一些潜在的信息,因此,异构网络成为数据挖掘领域个性化推荐新兴的一种挖掘技术,尤其是面对多种研究对象的研究。一般,网络的表示形式是G=(V,E,W),其中V代表研究对象的集合,E代表研究对象之间的链接关系的集合,W代表研究对象之间链接关系的权重的集合。对于异构网络,|V|>1或者|E|>1,表示网络中有多种类型的对象或者多种链接关系。与同构网络相比,异构网络不仅可以体现同种类型对象之间的关系,也可以体现不同类型对象之间的关系,如果同种类型之间并无直接联系,还可以通过其他类型对象得到同种类型之间的间接联系,这是异构网络最大的特点,也是其被广泛运用于现实世界中的聚类、分类、预测等研究的根本原因。因此,如何充分利用异构网络的优势,深度挖掘出异构网络中潜在的用户偏好,提高异构网络运用时的计算速度和效率,是将异构网络运用于数据挖掘领域个性化推荐时面临的主要问题。

发明内容

[0013] 针对现有的个性化推荐系统中普遍存在的冷启动、数据稀疏和文本特征分析技术有限等问题,本发明对现有的基于网络的推荐方法进行改进,基于用户特征信息建立子模型,结合异构网络元路径中潜在的信息,实现一种基于异构信息网络的电影推荐方法。在此基础之上,对异构网络进行社团划分和筛选,提高其运算速度和效率,充分发挥异构网络的优势。

[0014] 本发明提供的技术方案是一种基于异构信息网络的电影推荐方法,包括以下步骤:

[0015] 一种基于异构信息网络的电影推荐方法,其特征在于,包括以下步骤:

[0016] 步骤1,链接建模,以用户、电影、电影类型三种类型的对象为结点构建用户-电影异构网络,并计算用户与电影类型之间的权值,具体是:

[0017] 定义用户-电影异构网络模型G=(V,E,W),其中$V=V_u \cup V_m \cup V_t$,V_u表示用户集合,V_m表示电影集合,V_t表示电影类型集合,$E=E_{uu} \cup E_{mm} \cup E_{tt} \cup E_{um} \cup E_{ut} \cup E_{mt}$,其中,$E_{uu}$表示用户与用户之间的链接关系,$E_{mm}$表示电影与电影之间的链接关系,$E_{tt}$表示电影类型与电影类型之间的链接关系,$E_{um}$表示用户与电影之间的链接关系,$E_{ut}$用户与电影类型之间的链接关系,$E_{mt}$表示电影与电影类型之间的链接关系;W表示所有链接关系的权重集合,所述链接关系分别为用户与用户之间的链接关系、电影类型之间的链接关系、用户与电影之间的链接关系。

[0018] 步骤2,以电影类型结点为中心,对异构网络进行社团划分,一种电影类型为一个社团,查找不同社团两两之间的共同结点,并计算电影类型之间的权值;具有包括:

[0019] 步骤2.1,利用元路径搜索将异构网络中不同类型的结点以电影类型为中心进行划分,具有相似特征的结点被划分到一个社团当中,划分的结果是社团内结点具有高内聚,社团间结点具有低耦合的特征。划分社团的个数就是电影类型的个数,每一个社团包含该电影类型以及对该电影类型相关的电影进行过评分的用户即喜爱改电影类型相关电影的用户群,一个社团可以包含多个用户,一个用户也可能存在多个社团中,实际实施时用户结点与相邻的电影类型结点的权值可以使用数组a[|Mtype|][|Mtypei|]来存储,|Mtype|表示电影类型种数,|Mtypei|表示属于第i中类型的用户的数量,a[i][j]表示第i种电影类型与该类型电影中第j个用户之间的权值,

[0020] 步骤2.2,基于$W=\sum_{k=1,m_k\in M_i\cap M_j}^{|M_i\cap M_j|}\sum_{i=1}^{n}g_{k,i}\Big/\sum_{k=1,m_k\in M_i\cup M_j}^{|M_i\cup M_j|}\sum_{j=1}^{m}g_{k,j}$计算电影类型之间的权值,其中$M_i\cap M_j$表示这两种电影类型共同的电影集合,$m_k$代表电影集合中的电影,而相应的$g_{k,i}$代表用户i对于电影$m_k$的评分;$M_i\cap M_j$表示这两种类型电影的并集。

[0021] 步骤3,计算目标用户与每个社团即每个电影类型结点之间的相关程度,设置阀值,筛选出符合要求的社团;目标用户与每个社团之间的初始相关程度的计算与Dijkstra算法相似,唯一不同的在于,对于源点到其他结点的所有路径,Dijkstra算法得到的是最短距离,而在此计算的是所有路径长度的和。其中结点之间权值的计算分为两种,一种是用户结点与电影类型结点的权值,一种是电影类型结点之间的权值。

[0022] 步骤4,对于每一个符合要求的社团,基于目标用户到电影类型结点的初始相关程度,计算目标用户与社团中的每个用户之间的影响力,其中,用户之间的影响力就是用户之间的相似性,其计算步骤为,确定与目标用户存在元路径的用户,计算元路径每一段路径的权值,根据元路径复合规则计算两用户之间的相似性。

[0023] 步骤5,计算目标用户与符合要求的社团所代表的电影类型的喜爱程度,用目标用户与社团中每一个用户之间的影响力的均值来表示,社团中各结点对目标用户影响力的平均值作为用户划分到该社团的概率,即目标用户对于改社团所代表的电影类型的喜爱程度,平均值越大,目标用户划分到该社团概率越大,喜爱对应类型电影可能性越高。令目标用户划分到社团c_k的概率为$P_2(c_k|u_0)$,其计算过程如公式$P_2(c_k|u_0)=\sum_{u_k\in M_i}sim(u_0,u_k)\big/|c_k|$。其中,$u_k$属于社团$c_k$中的用户结点,$|c_k|$表示其中结点的个数,$sim(u_0,u_k)$表示目标用户$u_0$和$u_k$之间的相互影响力。

[0024] 步骤6,特征建模,基于经典朴树贝叶斯分类进行用户单个特征属性分类,分析目标用户的每一个特征被划分到每种电影类型的概率;

[0025] 步骤7,基于目标用户的每一个特征被划分的概率,综合计算目标用户对每个电影类型的喜爱程度;

[0026] 步骤8,结合步骤5与步骤7中分别得出的用户对每一种电影类型的喜爱程度,进行加权求和,分析目标用户对每一种电影类型的喜爱程度;

[0027] 步骤9,按照目标用户对每一种电影类型的喜爱程度,选择电影类型,推荐评分高

的电影，具体是，选择用户喜爱程度最高的前k种电影类型，分别从这些电影类型中选择评分高于一定值的前m部电影推荐给目标用户。

[0028] 在上述的一种基于异构信息网络的电影推荐方法，所述步骤4具体包括：

[0029] 步骤4.1、两个电影用户之间的相互影响力计算依据用户间的元路径，用户之间的相互影响包括直接影响和间接影响。

[0030] 所述间接影响为用户之间存在一条路径，用户影响力依靠路径影响彼此相邻结点。

[0031] 所述直接影响为结点之间有链接关系，即彼此为相邻结点，那么彼此之间存在直接影响力。

[0032] 两个结点之间的存在多条间接影响路径。对于两结点之间存在直接联系又有间接联系，他们的相互影响是两种影响的叠加。这种影响力大小定义为公式计算，当i，j为不相邻结点时$sim(i,j)=\sum_{m=1}^{n}\sum_{k\in Nb(i,j)}sim(i,k)\oplus sim(k,j)$，当i，j相邻时，sim(i，j)＝$w_{i,j}$，其中n表示结点i，j之间的路径数目，m表示路径序号，Nb(i，j)表示结点i，j每一条路径的连接点结点集合，符号⊕表示影响力在路径中不同阶段的连接方式，因为结点间随着路径长度加长，影响会逐渐变弱，即路径越长影响力越小，当结点i，j是相邻结点时，使用步骤1中的链接关系的权值，当不是相邻结点时并且元路径较长时，就将元路径以某一结点分为两段，这样反复下去直到元路径只有两个相邻结点。

[0033] 步骤4.2、元路径将依据上述三种类型的链接进行构建。在用户-电影异构网络中，对于用户与用户之间的间接关系，起到连接作用的是电影类型结点，当用户都喜爱某一种类型的电影，可以认为彼此之间能够产生影响。用户之间的影响力计算公式使用Sim(u_0，u_{ij})＝l(Mtype$_i$)*a[Mtype$_i$][j]来计算，其中，Mtype$_i$表示第i种电影类型，l(Mtype$_i$)表示目标用户u_0到第i种电影类型的初始相关程度，a[Mtype$_i$][j]表示该电影类型结点到该电影类型所代表的社团中的第j个用户u_{ij}的权值。

[0034] 在上述的一种基于异构信息网络的电影推荐方法，所述步骤6具体包括：

[0035] 步骤1，定义用户各个特征属性之间相互独立，对用户单个特征属性进行建模分类，目的在于分析用户单个特征属于每个电影类型的概率。在用户-电影异构网络中，对象特征属性进行建模时，用X_v表示对象V的特征信息属性集，相应地，针对文中的研究对象用户u_i来说，X_{ui}代表其特征属性集合。考虑电影用户有多个特征属性，因此X_{ui}是一个向量，表示为X_{ui}＝{$X_{ui,1}$，$X_{ui,2}$，...，$X_{ui,j}$，$X_{ui,n}$}，其中n＝|X_{ui}|为属性个数。

[0036] 步骤2，目标用户单个特征属性属于某个电影类型的概率用P(c_k|$X_{ui,j}$)表示，P($X_{ui,j}$|c_k)表示在电影类型c_k中用户u_i的第j个属性$X_{ui,j}$所占的比率，$X_{ui,j}$|c_k服从高斯分布，即$X_{ui,j}|c_k \backsim N(u_k,\sigma_k^2)$，P($c_k$)表示属于电影类型$c_k$的电影占所有电影的比例，从数据中可以直接统计获取，P($X_{ui,j}$)表示属性$X_{ui,j}$的概率，定义所有属性概率相同，即P($X_{ui,j}$)＝P($X_{uj,i}$)，i≠j。由贝叶斯定理得两者之间关系如公式P(c_k|$X_{ui,j}$)＝P($X_{ui,j}$|c_k)×P(c_k)/P($X_{ui,j}$)对于单个特征属性$X_{ui,j}$建模完成以后，同样的，用户其他特征属性也相应的得到建模，从而每个用户的每个属性都得到了建模。依据贝叶斯定理可以得出每个用户的每个属性属于某个电影类型的概率。

[0037] 在上述的一种基于异构信息网络的电影推荐方法，所述步骤7具体包括：复合特征

属性建模，综合考虑所有用户特征属性，从用户自身角度分析用户属于某个电影类型的概率，喜爱某种类型电影的可能性。是将每个用户的每个属性分类结果综合起来考虑，利用对用户属性的分类进而完成对用户的分类，其主要思想是如果用户所有属性同时划分到某一类型的概率比较大，那么目标用户划分到该类型的可能性就比较大，即认为目标用户喜爱对应类型的电影可能性就比较高。用$P(c_k|u_0)$表示u_0划分到电影类型c_k的概率。由用户各个特征属性之间相互独立，得到公式$P(c_k|u_0)=P(c_k|X_{u0})=\prod_{j=1}^{|X_{u0}|}P(c_k|X_{u0,j})$，$X_{u0,j}|c_k$服从高斯分布，即$X_{u0,j}|c_k \backsim N(u_k,\sigma_k^2)$，结合步骤6中每一个特征属性的概率公式可以得到用户属于每一个电影类型的概率公式

$$P_1(c_k|u_0)=P(c_k|X_{u0})=\prod_{j=1}^{|X_{u0}|}\frac{P(X_{u0,j}|c_k)\times P(c_k)}{P(X_{u0,j})}=\prod_{j=1}^{|X_{u0}|}\left(\frac{P(c_k)}{P(X_{u0,j})}\times\frac{1}{\sqrt{2\pi\sigma_k^2}}e^{\frac{-(X_{u0,j}-u_k)^2}{2\sigma_k^2}}\right)。$$

[0038] 在上述的一种基于异构信息网络的电影推荐方法，如果以目标用户u_0为例，所述步骤6和步骤7中，使用朴素贝叶斯分类，根据用户特征建立了特征模型，可以计算出目标用户u_0喜爱电影类型c_k的概率子模型$P_1(c_k|u_0)$，步骤1到步骤5中，通过利用元路径和社团划分，对用户-电影异构网络进行了分析，建立了基于链接的模型，得到目标用户u_0喜爱电影类型c_k的概率子模型$P_2(c_k|u_0)$。这两种模型的加权决定了目标用户u_0对于电影类型c_k的喜爱程度。定义两种子模型的权重分别为α和β，得到统一概率模型$P(c_k|u_0)=\alpha*P_1(c_k|u_0)+\beta*P_2(c_k|u_0)$，其中$P_1(c_k|u_0)$和$P_2(c_k|u_0)$分别表示所述步骤6、步骤7和步骤1到步骤5中，通过两种不同方式计算出的目标用户u_0喜爱电影类型c_k的概率，α为非负数，表示特征模型所占权重，β为正常数，表示链接模型所占权重。

[0039] 本发明具有如下优点：本发明从异构网络的角度出发，利用用户基本信息，进行用户特征属性建模，从用户自身的角度分析用户属于某个电影类型的概率，喜爱某种电影类型的可能性，避免了使用复杂技术对用户兴趣爱好进行分析，简化了用户特征属性的获取、分析和用户划分的过程，本发明还进行链接建模，对异构网络进行社团划分，利用异构网络中元路径潜在的用户之间的影响力，分析用户属于某一个电影类型的概率，更准确地挖掘异构网络中用户的偏好，在此基础之上，对于用户之间相似度的计算，还提出设置阀值筛选符合要求的社团，简化计算并提高运行效率，避免了传统相似度计算中需要计算目标用户与每一个用户的相似度。本发明的技术方案具有简单、高效和高准确率的特点，能够很好地解决现有推荐系统中的固有问题，并能较好地运用于电影推荐系统中。

附图说明

[0040] 图1本发明实施例的流程图。

[0041] 图2本发明的用户-电影异构网络图示意图。

[0042] 图3本发明实施例的用户-电影异构网络的社团划分示意图。

[0043] 图4本发明实施例的用户-电影异构网络元路径示意图。

[0044] 图5本发明实施例中用户之间的相互影响力在元路径中的传播示意图。

具体实施方式

[0045] 下面通过实施例，并结合附图，对本发明的技术方案作进一步具体的说明。

[0046] 实施例

[0047] 以电影推荐系统为例，实施例具体实现过程如下：

[0048] 步骤1，链接建模，以用户、电影、电影类型三种类型的对象为结点构建用户-电影异构网络，并计算用户与电影类型之间的权值。

[0049] 根据网络的表示形式，定义用户-电影异构网络模型G＝(V,E,W)，如图1所示，其中$V=V_u\cup V_m\cup V_t$，V_u表示用户集合，V_m表示电影集合，V_t表示电影类型集合，$E=E_{uu}\cup E_{mm}\cup E_{tt}\cup E_{um}\cup E_{ut}\cup E_{mt}$，其中，$E_{uu}$表示用户与用户之间的链接关系，$E_{mm}$表示电影与电影之间的链接关系，$E_{tt}$表示电影类型与电影类型之间的链接关系，$E_{um}$表示用户与电影之间的链接关系，$E_{ut}$用户与电影类型之间的链接关系，$E_{mt}$表示电影与电影类型之间的链接关系。其中W表示六种链接关系的权重集合，他们的计算方式不尽相同。本技术主要考虑用户之间的链接关系、电影类型之间的链接关系以及用户与电影类型之间的链接关系。用户对于某个电影类型的喜爱程度即E_{ut}用该电影类型中所有被用户评分过的电影的平均分来量化，用户之间的链接关系E_{uu}用他们直接的相互影响力来量化，电影类型之间的链接关系用不同电影类型之间共同的电影来进行量化，即边的权值。

[0050] 异构网络中的元路径是两个结点通过不同链接关系建立的一条可达路径，传递着用户之间的影响力，如图2所示，在用户-电影异构网络中，有6种链接关系，但是在利用元路径计算时，所使用的链接关系主要有两种，一种是用户与电影类型之间的链接关系，用L_1表示，电影类型与电影类型之间的链接关系，用L_2表示。

[0051] 对于用户结点与电影类型结点之间，由于电影类型对应着多部电影，它们之间是一对多的关系，以用户对于特定电影类型的所有电影的平均评分作为权值，即用户与该电影类型的初始相关程度。特别地，为了整个推荐算法计算，需要将所有权值进行归一化处理，使其属于0～1，归一化使用函数公式$f(x)=(x-X_{min})/(X_{max}-X_{min})$，其中，x为归一化之前的值，f(x)为归一化后的值，X_{min}为数据集合X中的最小值，X_{max}为数据集合X中的最大值。令元路径权值为w，用户i对于电影类型M_j中电影m_k评分为$g_{k,i}$，那么可以得到归一化之前的权值W，计算方法如公式$W'=(\sum_{k=1,m_k\in M_j}^{|M_j|}\sum_{i=1}^{n}g_{k,i})/|M_j|$，其中$|M_j|$为电影类型为$M_j$的电影数量，结合上述两式，最终归一化后的权值W为公式

$$W=f(W')=((\sum_{k=1,m_k\in M_j}^{|M_j|}\sum_{i=1}^{n}g_{k,i})/|M_j|-X_{\min})/(X_{max}-X_{min})$$

即L_1链接关系的权值计算方式。

[0052] 步骤2，以电影类型为中心，对异构网络进行社团划分，一种电影类型为一个社团，查找不同社团两两之间的共同结点，并计算电影类型之间的权值。

[0053] 利用元路径搜索将异构网络中不同类型的结点以电影类型为中心进行划分，具有相似特征的结点被划分到一个社团当中，划分的结果是社团内结点具有高内聚，社团间结点具有低耦合的特征。划分社团的个数就是电影类型的个数，每一个社团包含该电影类型以及对该电影类型相关的电影进行过评分的用户即喜爱改电影类型相关电影的用户群，一个社团可以包含多个用户，一个用户也可能存在多个社团中，实际实施时用户结点与相邻的电影类型结点的权值可以使用数组a[|Mtype|][|Mtypei|]来存储，|Mtype|表示电影类

型种数，|Mtypei|表示属于第i中类型的用户的数量，a[i][j]表示第i种电影类型与该类型电影中第j个用户之间的权值，其权值计算方式如步骤1中的L_1链接关系的去权值计算。

[0054] 社团所代表的电影类型之间没有之间链接关系，但因为他们有共同的电影爱好者，所以可以使用共同爱好者的数量来量化两种电影类型结点之间的链接权值。社团间共同用户的查找有如下步骤：第一步，对原始数据进行分解，用户与看过的电影是一一对应的关系，但是一部电影可以对应多种电影类型，所以将用户对一部电影的评分记录分解为多条记录，使一条记录对应一种电影类型，具体实施时可以构建一个Urecord数据结构，包含UID、Nuser、sex、age、occupation、MID、Mname、Mtype、goal和数组a，其中UID表示电影用户ID，Nuser表示用户记录分解以后每个UID对应的记录数，sex表示用户性别：0表示男性、1表示女性，age表示用户年龄，occupation代表用户职业用1-21来表示，对应21种职业，MID表示电影ID，Mname表示电影名字，Mtype代表电影类型用1-18表示表示，对应18种电影类型，goal表示电影用户对电影的评分，数组a表示用户是否对每个类型的电影进行过评分，每个元素为0或者1，假设原数据中电影类型共18个，因此数组a大小为18。a[i]＝0表示该用户没有对第i个电影类型进行过评分，a[i]＝1表示该用户对第i个电影类型进行过评分。第二步，统计分解以后的记录个数N，具体实施时用Vector<Urecord>来存储所有记录。第三步，遍历所有记录，将相同电影类型的记录，存放在一个集合中，具体实施时，可以使用HashSet存储同一电影类型的所以记录，有多少中电影类型，就使用多少个HashSet。第四步，对记录不同电影类型的集合取交集。第五步，检索交集中每一条记录所对应的用户，相同用户只记录一次，那么，所得用户的集合就是两个社团之间共同的结点。

[0055] 对于电影类型之间的权值主要利用不同电影类型的共同电影进行计算，计算公式如 $W=\sum_{k=1,m_k\in M_i\cap M_j}^{|M_i\cap M_j|}\sum_{i=1}^{n}g_{k,i}\Big/\sum_{k=1,m_k\in M_i\cup M_j}^{|M_i\cup M_j|}\sum_{j=1}^{m}g_{k,j}$ 即L_2连接关系的计算方式，其中$M_i\cap M_j$表示这两种电影类型共同的电影集合，m_k代表电影集合中的电影，而相应的$g_{k,i}$代表用户i对于电影m_k的评分；$M_i\cap M_j$表示这两种类型电影的并集。

[0056] 步骤3，计算目标用户与每个社团即每个电影类型结点之间的相关程度，设置阀值，筛选出符合要求的社团。

[0057] 基于目标用户结点、电影类型结点集合以及权值邻接矩阵，使用Dijkstra算法计算路径长度之和，实现计算目标用户结点与每一个电影类型结点的影响力，即与该社团的相关程度。并通过设置阈值，筛选出阈值之上的社团，从而节省计算，提高效率。

[0058] 步骤4，对于每一个符合要求的社团，基于目标用户到电影类型结点的初始相关程度，计算目标用户与社团中的每个用户之间的影响力。

[0059] 两个电影用户之间的相互影响力计算依据用户间的元路径，用户之间的相互影响分为直接影响和间接影响。间接影响反映了用户之间存在一条路径，那么用户影响力就会依靠路径影响彼此相邻结点。如果结点之间有链接关系，即彼此为相邻结点，那么彼此之间存在直接影响力。在网路图中，两个结点之间的可能存在多条间接影响路径。特别地，对于两结点之间存在直接联系又有间接联系，他们的相互影响是两种影响的叠加。这种影响力大小可以定义为公式计算，当i，j为不相邻结点时

$sim(i,j)=\sum_{m=1}^{n}\sum_{k\in Nb(i,j)}sim(i,k)\oplus sim(k,j)$，当i，j相邻时，sim(i,j)＝$w_{i,j}$，其中n表示结点i，j之间的路径数目，m表示路径序号，Nb(i,j)表示结点i，j每一条路径的连接点结点集合，符号⊕表示影响力在路径中不同阶段的连接方式，因为结点间随着路径长度加长，影响会逐渐变弱，即路径越长影响力越小，为了反映这种特性，在本文中，连接方式用相乘的方式。公式原理是分治与递推思想，当结点i，j是相邻结点时，使用步骤1中的链接关系的权值，当不是相邻结点时并且元路径较长时，就将元路径以某一结点分为两段，这样反复下去直到元路径只有两个相邻结点。

[0060] 用户-电影异构网络中，一共存在6种类型的链接，主要研究三种类型的链接关系，分别为用户与用户之间的链接关系、电影类型之间的链接关系、用户与电影之间的链接关系。用户之间的相互影响可能会涉及到上述三种类型的链接。在用户-电影异构网络中，不同链接关系对于影响程度作用会不一样，而异构网络中特有的的元路径正好可以区分不同类型的链接关系，因此本文将根据用户之间的元路径来研究用户之间的相互影响。本文中，元路径将依据上述三种类型的链接进行构建。在用户-电影异构网络中，对于用户与用户之间的间接关系，起到连接作用的是电影类型结点，当用户都喜爱某一种类型的电影，可以认为彼此之间能够产生影响。用户之间的影响力计算公式可以使用$Sim(u_0,u_{ij})=l(Mtype_i)*a[Mtype_i][j]$来计算，其中，$Mtype_i$表示第i种电影类型，$l(Mtype_i)$表示目标用户$u_0$到第i种电影类型的初始相关程度，$a[Mtype_i][j]$表示改电影类型结点到改电影类型所代表的社团中的第j个用户u_{ij}的权值。

[0061] 步骤5，计算目标用户与符合要求的社团所代表的电影类型的喜爱程度。

[0062] 社团中各结点对目标用户影响力的平均值作为用户划分到该社团的概率，即目标用户对于改社团所代表的电影类型的喜爱程度，平均值越大，目标用户划分到该社团概率越大，喜爱对应类型电影可能性越高。令目标用户划分到社团c_k的概率为$P_2(c_k|u_0)$，其计算过程如公式$P_2(c_k|u_0)=\sum_{u_k\in M_i}sim(u_0,u_k)\big/|c_k|$。其中，$u_k$属于社团$c_k$中的用户结点，$|c_k|$表示其中结点的个数，$sim(u_0,u_k)$表示目标用户$u_0$和$u_k$之间的相互影响力。

[0063] 步骤6，特征建模，用户单个特征属性分类，分析目标用户的每一个特征被划分到每种电影类型的概率。

[0064] 假设用户各个特征属性之间相互独立，对用户单个特征属性进行建模分类，目的在于分析用户单个特征属于每个电影类型的概率。在用户-电影异构网络中，对象特征属性进行建模时，用X_v表示对象V的特征信息属性集，相应地，针对文中的研究对象用户u_i来说，X_{ui}代表其特征属性集合。考虑电影用户有多个特征属性，因此X_{ui}是一个向量，表示为$X_{ui}=\{X_{ui,1},X_{ui,2},...,X_{ui,j},X_{ui,n}\}$，其中n＝$|X_{ui}|$为属性个数。

[0065] 目标用户单个特征属性属于某个电影类型的概率用$P(c_k|X_{ui,j})$表示，$P(X_{ui,j}|c_k)$表示在电影类型c_k中用户u_i的第j个属性$X_{ui,j}$所占的比率，一般来讲，$X_{ui,j}|c_k$是服从高斯分布的，即$X_{ui,j}|c_k\backsim N(u_k,\sigma_k^2)$，$P(c_k)$表示属于电影类型$c_k$的电影占所有电影的比例，从数据中可以直接统计获取，$P(X_{ui,j})$表示属性$X_{ui,j}$的概率，在本文中认为所有属性概率相同，即$P(X_{ui,j})=P(X_{uj,i})$，i≠j。由贝叶斯定理得两者之间关系如公式$P(c_k|X_{ui,j})=P(X_{ui,j}|c_k)\times P(c_k)/P(X_{ui,j})$对于单个特征属性$X_{ui,j}$建模完成以后，同样的，用户其他特征属性也相应的得

到建模，从而每个用户的每个属性都得到了建模。依据贝叶斯定理可以得出每个用户的每个属性属于某个电影类型的概率

[0066] 步骤7，基于目标用户的每一个特征被划分的概率，综合计算目标用户对每个电影类型的喜爱程度。

[0067] 复合特征属性建模，综合考虑所有用户特征属性，从用户自身角度分析用户属于某个电影类型的概率，喜爱某种类型电影的可能性。是将每个用户的每个属性分类结果综合起来考虑，利用对用户属性的分类进而完成对用户的分类，其主要思想是如果用户所有属性同时划分到某一类型的概率比较大，那么目标用户划分到该类型的可能性就比较大，即认为目标用户喜爱对应类型的电影可能性就比较高。用$P(c_k|u_0)$表示u_0划分到电影类型c_k的概率。由于步骤6中已经假设用户各个特征属性之间相互独立的所以得公式

$$P(c_k|u_0)=P(c_k|X_{u0})=\prod_{j=1}^{|X_{u0}|}P(c_k|X_{u0,j})$$

[0068] 因为$X_{u0,j}|c_k$是服从高斯分布的，结合步骤6中每一个特征属性的概率公式可以得到用户属于每一个电影类型的概率公式

$$P_1(c_k|u_0)=P(c_k|X_{u0})=\prod_{j=1}^{|X_{u0}|}\frac{P(X_{u0,j}|c_k)\times P(c_k)}{P(X_{u0,j})}=\prod_{j=1}^{|X_{u0}|}\left(\frac{P(c_k)}{P(X_{u0,j})}\times\frac{1}{\sqrt{2\pi\sigma_k^2}}e^{\frac{-(X_{u0,j}-u_k)^2}{2\sigma_k^2}}\right)$$

[0069] 步骤8，结合步骤5与步骤7中分别得出的用户对每一种电影类型的喜爱程度，进行加权求和，分析目标用户对每一种电影类型的喜爱程度。

[0070] 以目标用户u_0为例，步骤6和步骤7，使用朴素贝叶斯分类，根据用户特征建立了特征模型，可以计算出目标用户u_0喜爱电影类型c_k的概率子模型$P_1(c_k|u_0)$，步骤1到步骤5中，通过利用元路径和社团划分，对用户-电影异构网络进行了分析，建立了基于链接的模型，得到目标用户u_0喜爱电影类型c_k的概率子模型$P_2(c_k|u_0)$。这两种模型的加权决定了目标用户u_0对于电影类型c_k的喜爱程度。设两种子模型的权重分别为α和β，得到统一概率模型$P(c_k|u_0)=\alpha*P_1(c_k|u_0)+\beta*P_2(c_k|u_0)$，其中$P_1(c_k|u_0)$和$P_2(c_k|u_0)$分别表示所述步骤6、步骤7和步骤1到步骤5中，通过两种不同方式计算出的目标用户u_0喜爱电影类型c_k的概率，α为非负数，表示特征模型所占权重，β为正常数，表示链接模型所占权重。

[0071] 步骤9，按照目标用户对每一种电影类型的喜爱程度，选择电影类型，推荐评分高的电影电影。

[0072] 结合目标用户自身特征属性和其他具有相同爱好的用户对其的影响力两方面来预测目标用户对于每种类型的电影的喜爱程度，喜爱程度越高，被推荐的可能性越大。因此推荐策略就是取前k种喜爱程度高的电影类型作为推荐目标，再分别从每种电影类型类型中取评分在3(评分为1-5，3分及其以上的认为是喜欢)以上的前m部电影推送给目标用户。

[0073] 现有推荐技术普遍存在数据稀松、冷启动等问题，本发明充分利用了异构网络存在潜在用户兴趣的优势，结合用户特征属性，从两个角度挖掘用户偏好，很好的解决了数据稀松和冷启动问题，在此基础之上，还提出在异构网络中进行社团划分以后，进行阈值筛选，有效的提高了算法的执行效率。

[0074] 本文中所描述的具体实施例仅仅是对本发明精神作举例说明。本发明所属技术领域的技术人员可以对所描述的具体实施例做各种各样的修改或补充或采用类似的方式替

代,但并不会偏离本发明的精神或者超越所附权利要求书所定义的范围。

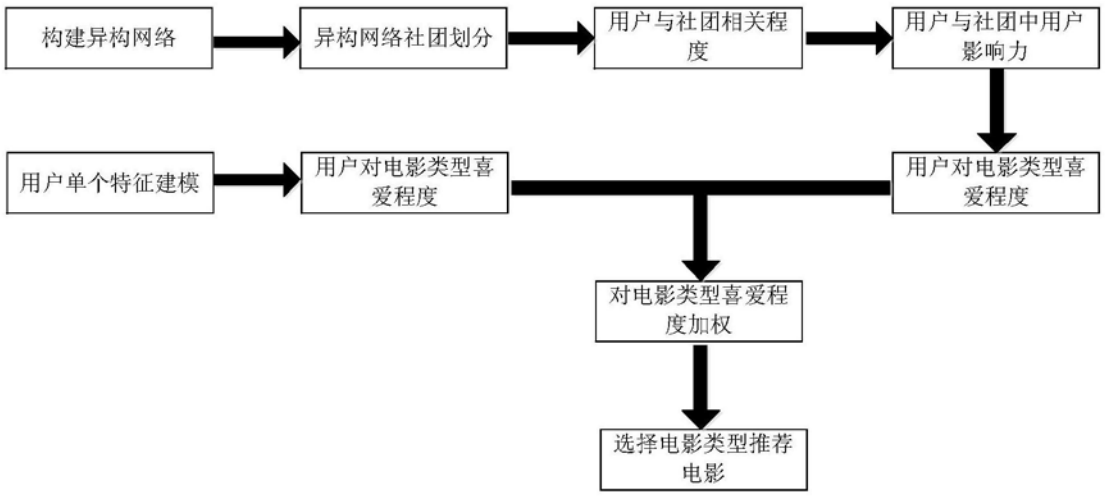

构建异构网络
异构网络社团划分
用户与社团相关程度
用户与社团中用户影响力
用户单个特征建模
用户对电影类型喜爱程度
用户对电影类型喜爱程度
对电影类型喜爱程度加权
选择电影类型推荐电影

图1

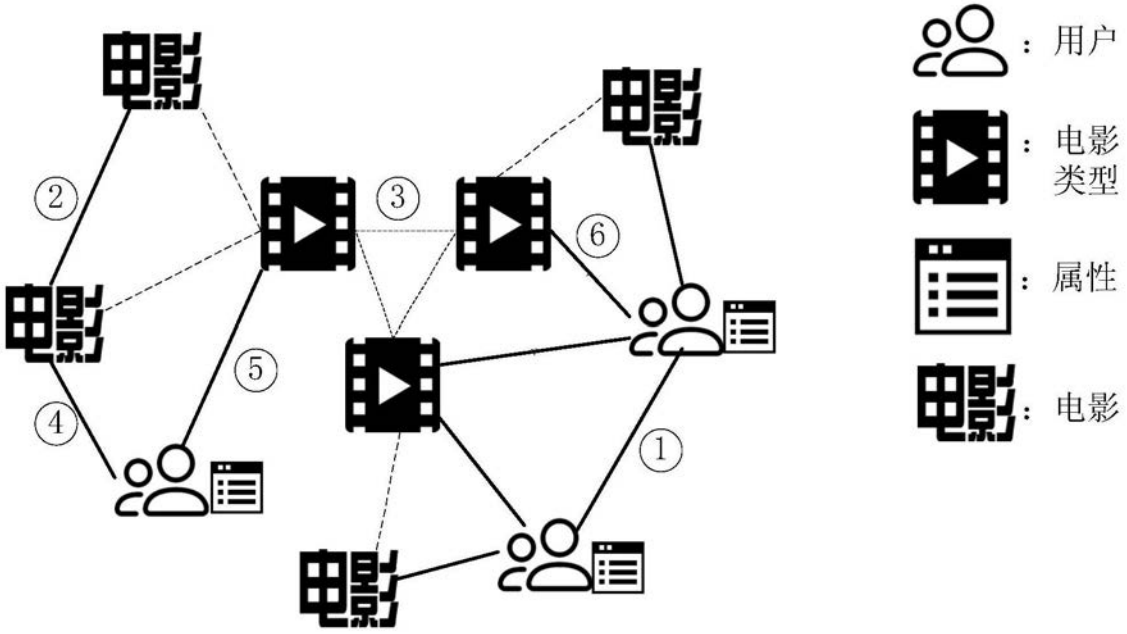

电影
电影
电影
电影
电影
①
②
③
④
⑤
⑥
：用户
：电影类型
：属性
：电影

图2

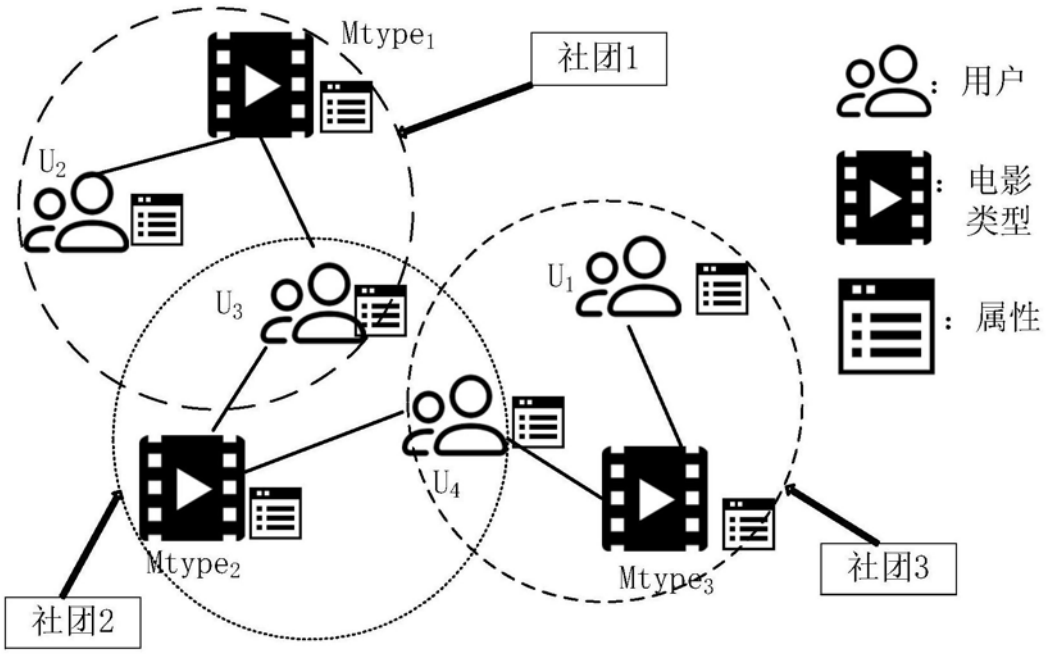

Mtype1
社团1
用户
U2
电影
类型
U3
U1
属性
U4
Mtype2
Mtype3
社团3
社团2

图3

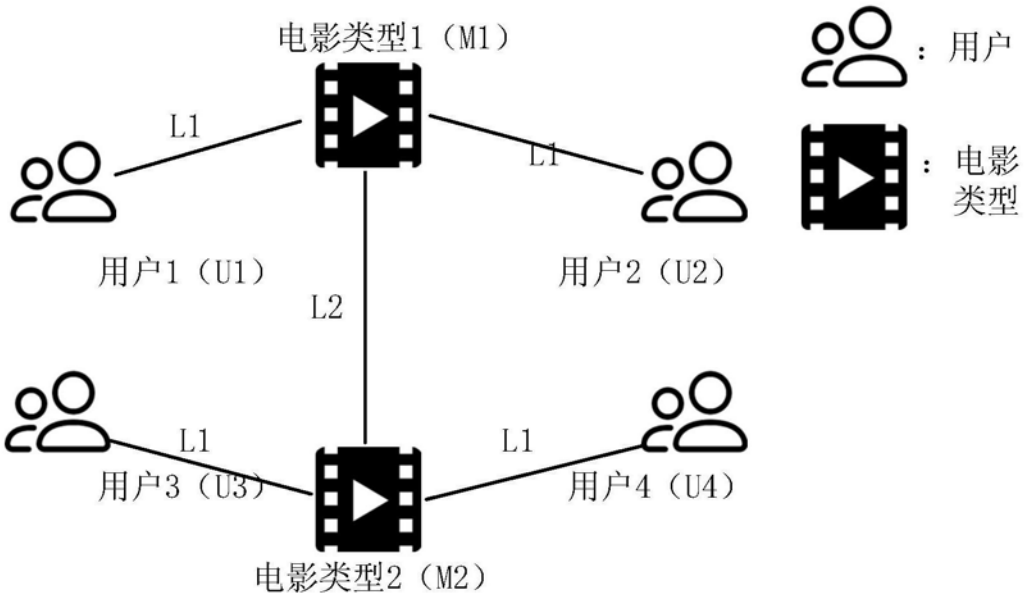

电影类型1（M1）
用户
L1
L1
电影
类型
用户1（U1）
用户2（U2）
L2
L1
L1
用户3（U3）
用户4（U4）
电影类型2（M2）

图4

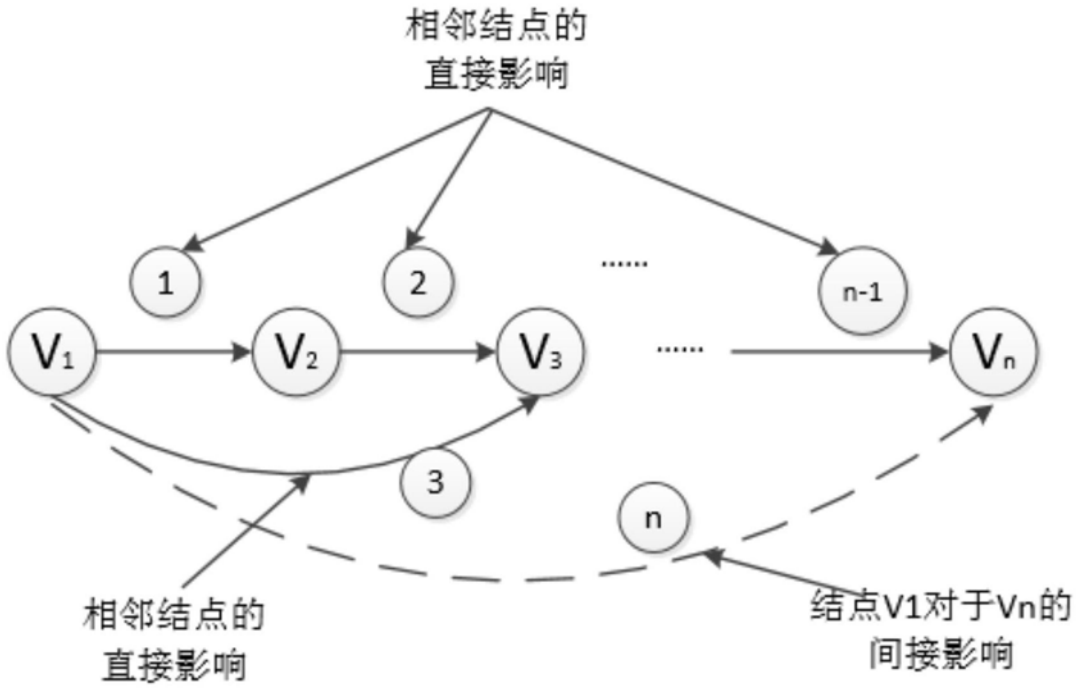


图5