



异质信息网络分析及其语义探索

王 锐, 张志强, 石 川

(北京邮电大学智能通信软件与多媒体北京市重点实验室 北京 100876)

摘 要: 由多种类型的实体和关系构成的异质信息网络逐渐成为社会网络分析的研究热点。作为异质信息网络的一个独特属性,元路径包含了丰富的语义信息。实际生活中的许多网络都包含带权值的链接,这使得不考虑链接上权值的传统元路径不能精确地捕捉网络中的语义信息。基于此,描述了异质信息网络的相关概念,并对异质信息网络的应用进行了简要介绍。通过将传统元路径扩展为带权元路径,更精确地描述了带权值的异质信息网络中微妙的语义信息。通过在两个真实数据集上进行实验,说明了带权元路径在推荐、相关性搜索中的应用效果。

关键词: 异质信息网络;元路径;相关性搜索;推荐

doi: 10.11959/j.issn.1000-0801.2015166

Analysis and Semantic Mining in Heterogeneous Information Network

Wang Rui, Zhang Zhiqiang, Shi Chuan

(Beijing Key Lab of Intelligent Telecommunication Software and Multimedia,
Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: Heterogeneous information network (HIN), which is composed of different types of objects and links, has gradually become a hot topic in social network analysis. As a unique characteristic of HIN, meta path contains rich semantic information. Heterogeneous information network with values on links are ubiquitous in real world. Therefore, the traditional meta path, which doesn't consider weight on links, can not exactly capture semantics in many cases. Related concepts of HIN were introduced and a brief introduction of applications of HIN was given. Then subtle semantic information in HIN was explored by extending the traditional meta path to weighted meta path. Experiments on two real data sets demonstrate the applications of the weighted meta path in recommendation, relevance search.

Key words: heterogeneous information network, meta path, relevance search, recommendation

收稿日期:2015-05-25;修回日期:2015-07-06

基金项目: 国家重点基础研究发展计划(“973”计划)基金资助项目(No.2013CB329606), 国家高技术研究发展计划(“863”计划)基金资助项目(No.2015AA050203, No.2015AA050204), 国家自然科学基金资助项目(No.71231002, No.61375058)

Foundation Items: The National Basic Research Program of China (973 Program) (No.2013CB329606), The National High Technology Research and Development Program of China (863 Program) (No.2015AA050203, No.2015AA050204), The National Natural Science Foundation of China (No.71231002, No.61375058)

论文引用格式: 王锐, 张志强, 石川. 异质信息网络分析及其语义探索. 电信科学, 2015166

Wang R, Zhang Z Q, Shi C. Analysis and semantic mining in heterogeneous information network. Telecommunications Science, 2015166



1 引言

社会网络是指个体成员之间存在相互联系和交互的网络。随着作者合作网络、引文关系网络等社会网络的出现,社会网络分析受到了越来越多的关注。当前的社会网络分析主要基于同质信息网络。在同质信息网络中,节点和链接均属于同一类型,各种分析方法对所有节点做同样的处理。

随着在线社交网络的大量出现,对象相互关联形成的复杂网络很难用同质信息网络建模,而是需要建模成异质信息网络。这种网络包含不同类型的对象,对象之间的关系表示了不同的交互语义。相对于同质信息网络,异质信息网络包含更加全面的结构信息和丰富的语义信息,近些年,聚类^[1]、分类^[2]等许多数据挖掘工作在这种网络中展开。

元路径是指实体类型之间的复合关系,作为异质信息网络的一个独特属性,它包含了丰富的语义信息^[2,3]。由于元路径的这个特征,许多工作^[4-6]提出了基于元路径的数据挖掘算法。之前的大多数工作基于不考虑权重的传统元路径,然而,在某些情况下(特别是在推荐系统中)传统元路径不能捕捉到微妙的语义信息。在推荐系统中,用户对于物品可能打分,这个分数代表了用户对物品的喜好程度。如果忽略了这个分数,可能会得到错误的结果。因此,提出一种考虑链接上属性约束的异质信息网络以及在这种异质信息网络中的相似性度量方法是很有必要的。

本文首先描述了异质信息网络中的相关概念,并对异质信息网络的应用进行了简要介绍。之后针对传统元路径没有考虑链接上权重的问题,提出了带权值的异质信息网络和带权元路径的概念,并将已有的基于元路径的相似性度量方法扩展到带权元路径上。最后,通过实验证明了带

权元路径相对于传统元路径的优越性。

2 异质信息网络分析及其应用

异质信息网络是由多种类型的节点和链接构成的信息网络。相对于同质信息网络,异质信息网络的分布更加广泛,并且可以更加全面地表示网络的组成对象和它们之间的关系。因此,异质信息网络分析能够发现更加精细、准确的隐含知识。本节对异质信息网络中的相关概念进行了解释,并介绍了异质信息网络的一些应用。

2.1 相关概念

定义1 (信息网络) 信息网络^[7] $G=(V,E)$ 是一个有向图,其中 V 是节点集合, E 是链接集合,存在一个节点类型的映射函数 $\varphi:V \rightarrow \mathcal{A}$ 和一个链接类型的映射函数 $\psi:E \rightarrow \mathcal{R}$ 。每一个节点 $v \in V$ 都属于一个实体类型 $\varphi(v) \in \mathcal{A}$, 每一个链接 $e \in E$ 都属于一个关系类型 $\psi(e) \in \mathcal{R}$ 。当实体类型数 $|\mathcal{A}| > 1$ 或关系类型数 $|\mathcal{R}| > 1$ 时,该信息网络为异质信息网络;否则为同质信息网络。

图1给出了实际生活中异质信息网络的例子。图1(a)是电影推荐系统,其中包含了用户(U)、小组(G)、电影(M)、导演(D)、演员(A)、类型(T)6种实体类型。每部电影与导演、演员、电影类型之间都存在链接,每个用户与用户、小组、电影之间都存在链接,这些链接属于不同的关系类型。图1(b)是文献网络,其中包含了论文(P)、作者(A)、会议(C)、术语(T)4种实体类型。每篇论文与作者、会议、术语之间都存在链接,这些链接也属于不同的关系类型。

在异质信息网络中存在多种类型的节点或者多种类型的链接,用网络模式描述网络中的实体类型以及实体类型之间的关系类型。

定义2 (网络模式) 网络模式^[7] $T_c=(\mathcal{A},\mathcal{R})$ 是定义在实

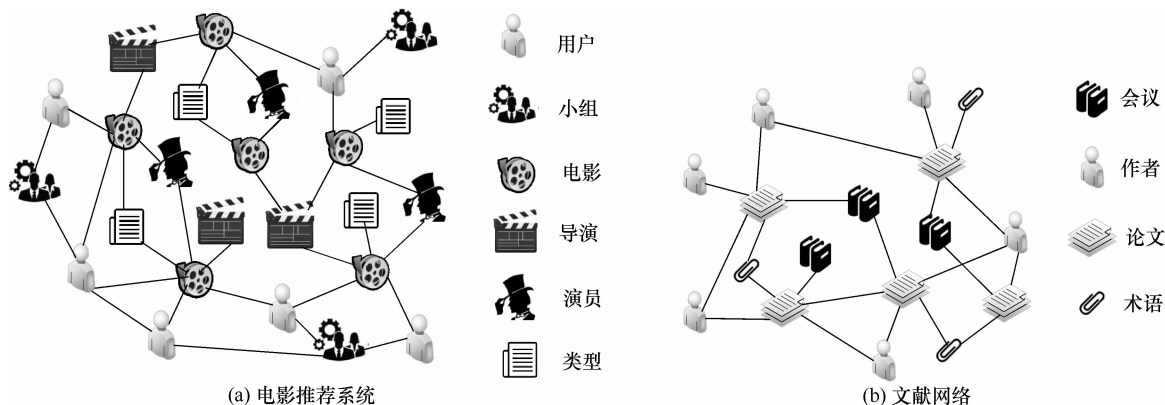


图1 异质信息网络

体类型集合 $\mathcal{A}$ 和实体类型之间的关系集合 $\mathcal{R}$ 上的一个有向图,是节点类型的映射函数为 $\varphi:V \rightarrow \mathcal{A}$ 且链接类型的映射函数为 $\psi:E \rightarrow \mathcal{R}$ 的信息网络 $G=(V,E)$ 的元模板。

图 2 显示了电影推荐系统及文献网络的网络模式。图 2(a) 中,电影推荐系统里用户与用户之间的链接表示社交关系,用户与小组之间的链接表示社团从属关系,用户与电影之间的链接表示影片观看关系,电影与导演之间的链接表示影片执导关系,电影与演员之间的链接表示影片出演关系,电影与类型之间的链接表示电影类型关系。图 2(b) 中,文献网络里论文与论文之间的链接表示论文引用关系,论文与作者之间的链接表示论文写作关系,论文与会议之间的链接表示论文发表关系,论文与术语之间的链接表示术语使用关系。

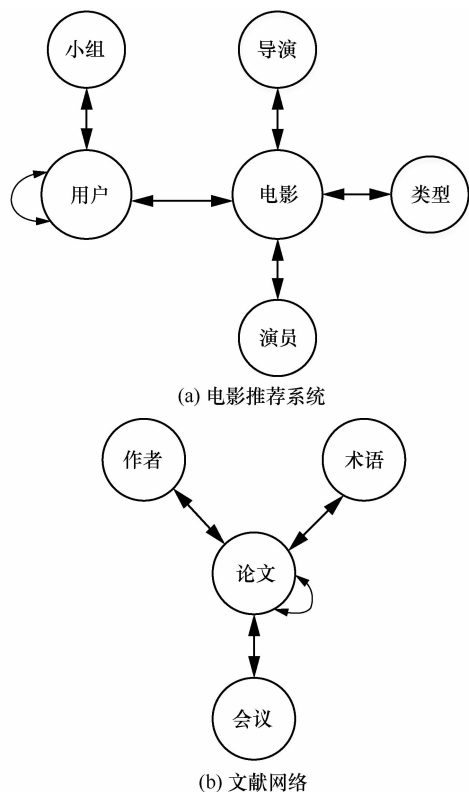


图 2 异质信息网络的网络模式

不同于同质信息网络,异质信息网络中的两个实体类型可以通过不同的路径相连,而这些路径所表达的含义也不尽相同。这些路径称为元路径,在这里作为形式化定义。

定义 3 (元路径)元路径^[4]是定义在网络模式 $S=(\mathcal{A},\mathcal{R})$ 上的路径,其符号表示形式为 $P:A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \dots \xrightarrow{R_l} A_{l+1}$ (简写为 $A_1 A_2 \dots A_{l+1}$)。元路径在实体类型 A_1 和 A_{l+1} 之间定义了一个复合关系 $R=R_1 \circ R_2 \circ \dots \circ R_l$,其中 $\circ$ 代表

关系之间的复合算子。

图 3 显示了电影推荐系统中的元路径实例。其中,用户可以通过元路径“用户—电影—用户(UMU)”相连,也可以通过元路径“用户—电影—导演—电影—用户(UMDMU)”相连。异质信息网络中的元路径包含丰富的语义信息,例如,元路径 UMU 代表用户观看了相同的电影,而元路径 UMDMU 代表用户观看了同一个导演指导的电影。这些不同的语义信息对于数据挖掘任务至关重要,例如,在进行相似性计算时,基于元路径 UMU,曾经观看过相同电影的用户更为相似;基于元路径 UMDMU,观看了相同导演执导的电影的用户更为相似。

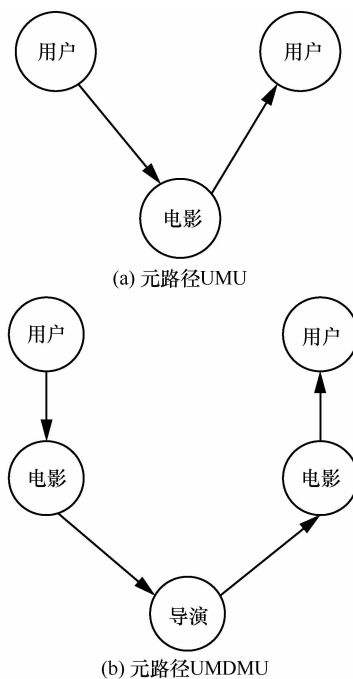


图 3 电影推荐系统中的元路径实例

2.2 异质信息网络的应用

由于异质信息网络包含了更丰富的语义信息,对异质信息网络进行数据挖掘,更有可能产生有意义的知识发现。因此,越来越多的研究开始关注异质信息网络,本小节对异质信息网络的应用进行简要介绍。

相似性计算是许多数据挖掘任务的基础,解决问题的关键在于确定相似性度量标准。许多工作广泛地研究了相似性度量方法,但是这些工作大多数是基于同质信息网络的。两个节点之间基于不同元路径的相似性是不同的,所以异质信息网络中的相似性度量是受元路径约束的。其次,由于多种类型节点的存在,使得可以在异质信息网络中评价不同类型节点间的相关性。近年来,对异质信



息网络中的相似性度量方法进行了很多探究,例如 Lao 和 Cohen^[5]提出了基于路径的随机游走模型(PCRW),用于度量有向图中节点的相似性。Sun 等人^[4]提出了基于对称路径的 PathSim 方法,用于度量相同类型节点的相似性。Shi 等人^[6]提出了具有对称特性的 HeteSim,用于度量异质网络中任意节点对的相似性。

聚类分析是将网络中的对象或边划分成不同的簇,使得同一簇中的对象更相似。传统方法基于对象的特征值进行聚类,例如 k -means。基于网络化数据的聚类(如社团发现)也得到了广泛研究,但是这些方法主要针对同质网络。异质信息网络中聚类的特点为可以利用其他类型节点的信息对目标类型的节点进行聚类。此外,可以通过整合属性信息、用户引导信息等附加信息或其他数据挖掘任务进行聚类。例如,针对异质信息网络中节点属性不完全和多种类型的边,Sun 等人^[8]提出了新颖的概率聚类模型,能够整合不完整的属性信息和异质信息网络的结构信息。由于异质信息网络中聚类结果与选择的元路径相关,Sun 等人^[1]根据用户引导信息进行聚类 and 元路径的自动选择。RankClus^[9]利用排序和聚类相互促进的思想对异质信息网络中的同类对象进行聚类。

分类是通过建立模型或分类器来预测节点的标签。异质信息网络中包含多种类型的节点,因此在对目标节点进行分类时,可以利用其他类型节点的信息。异质信息网络中的链接类型也是不同的,不同的链接对节点之间相似性的影响不同,异质信息网络中的分类主要考虑不同的链接上信息的传播。Angelova 等人^[10]通过随机游走模拟网络中节点之间的相互影响,提出了一个异质信息网络中基于图的多标签分类模型。Jacob 等人^[11]认为两个相连的节点往往有相近的潜在表示,通过计算节点的潜在表示来对节点进行分类。

链接预测是链接挖掘中一个最基础的问题,通过已知的链接和节点属性预测节点之间可能存在的链接。近年来,基于异质信息网络的链接预测成为了一个研究热点。由于实体类型的多样性,异质信息网络中可以预测的链接也是多种多样的。同时,这些链接之间也存在着相关性。Yang 等人^[12]提出了一个基于概率的方法 MRIP,通过模拟异质信息网络中链接的传播,在拥有多种关系类型的异质信息网络中进行链接预测。基于元路径,Sun 等人^[13]提出 PathPredict 方法,通过特征提取和权重学习来进行共同作者关系的预测。Zhang 等人^[14]提出了 TRAIL 方法,利用部分

对齐的 LBSN(location-based social network,基于位置的社会网络)同时预测一个新的 LBSN 中的社会链接和位置链接。

除上述应用之外,异质信息网络还有一些其他的应用,如排序、推荐等。

3 带权值的异质信息网络及带权元路径

传统的异质信息网络不考虑链接上的权重或属性的值,然而在实际生活中,链接含有属性值的网络是很常见的。例如:在电影推荐系统中,用户通常会给看过的电影打出 1~5 分,表示用户对电影的喜好程度;在文献网络中,通常存在作者顺序,表示作者对论文的贡献程度。本节通过考虑链接上的权重对传统的异质信息网络进行了扩展。

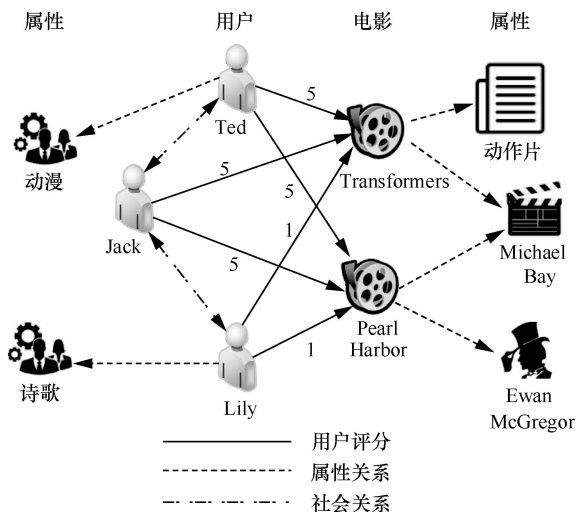
3.1 概念扩展

定义 4 (带权值的异质信息网络)已知网络模式 $S=(\mathcal{A}, \mathcal{R}, \mathcal{W})$, 其中, $\mathcal{A}=\{A\}$ 为实体类型集合, $\mathcal{R}=\{R\}$ 为关系类型集合, $\mathcal{W}=\{W\}$ 是一系列属性值的集合。带权值的异质信息网络(weighted heterogeneous information network, WHIN)被定义为有向图 $G=(V, E, W)$, 其节点类型映射函数为 $\varphi: V \rightarrow \mathcal{A}$, 链接类型映射函数为 $\psi: E \rightarrow \mathcal{R}$, 属性值类型的映射函数为 $\theta: W \rightarrow \mathcal{W}$ 。每一个节点 $v \in V$ 都属于一个实体类型 $\varphi(v) \in \mathcal{A}$, 每一个链接 $e \in E$ 都属于一个关系类型 $\psi(e) \in \mathcal{R}$, 每一个属性值 $w \in W$ 都属于一个属性值类型 $\theta(w) \in \mathcal{W}$ 。当 $|\mathcal{A}| > 1$ 或 $|\mathcal{R}| > 1$ 且属性值的类型数 $|\mathcal{W}| > 0$ 时, 该网络为带权值的异质信息网络; 当 $|\mathcal{A}| > 1$ 或 $|\mathcal{R}| > 1$ 且属性值的类型数 $|\mathcal{W}| = 0$ 时, 该网络为不带权值的异质信息网络。

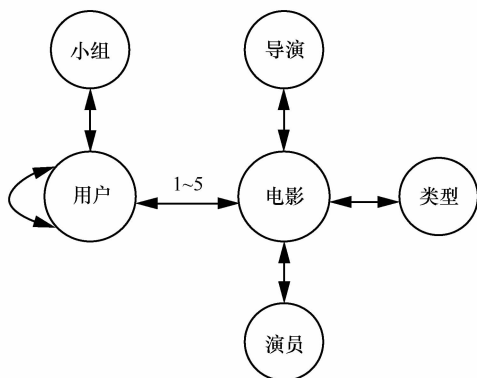
图 4 是带权值的异质信息网络的一个实例。由图 4 可以看出, Lily、Ted 和 Jack 3 人都看过 Transformers 以及 Pearl Harbor 两部电影, 如果基于元路径 UMU 计算用户相似性, Jack 与 Ted 和 Lily 的相似性是相等的。然而, 当考虑用户对电影的评分信息时会发现, Lily 给这两部电影都打了 1 分, 这说明她并不喜欢这两部电影; 而 Ted 和 Jack 都给这两部电影打了 5 分, 这说明他们很喜欢这两部电影。如果考虑采用链接上的评分信息来计算用户相似性, 相对 Lily 而言, Jack 和 Ted 是更为相似的。

在带权值的异质信息网络中, 传统元路径通常不能准确地捕捉到网络中潜在的语义信息。以图 4 为例, 元路径 UMU 不能体现用户对电影的喜好, 因此, 如果基于元路径 UMU 计算用户的相似性, 可能会得到不恰当的结果。为此, 提出了带权值的异质信息网络中带权元路径的概念。

定义 5 (带权元路径)带权元路径是关系上存在属



(a) 电影推荐系统



(b) 网络模式

图 4 电影推荐系统及其网络模式

性值约束的元路径, 其符号表达形式为 $A_1 \xrightarrow{\delta_1(R_1)} A_2 \xrightarrow{\delta_2(R_2)} \dots \xrightarrow{\delta_l(R_l)} A_{l+1} | C$ (简称为 $A_1 (\delta_1(R_1)) A_2 (\delta_2(R_2)) \dots (\delta_l(R_l)) A_{l+1} | C$). 若关系 R 上有属性值, 则关系 R 上的属性值函数 $\delta(R)$ 为 R 的所有属性值的一个非空子集, 否则 $\delta(R)$ 为空。

$A_i \xrightarrow{\delta_i(R_i)} A_{i+1}$ 代表实体 A_i 和 A_{i+1} 之间的关系 R_i 的属性值在 $\delta_i(R_i)$ 中的取值, C 由属性值函数上的一系列关联约束组成。若一条元路径上所有的属性值函数均为空, 则这条元路径为不带权元路径; 否则, 这条元路径为带权元路径。

传统元路径是不带权元路径, 它可以看作带权元路径的特殊情形。在豆瓣电影数据集中, 用户对电影的打分可以为 1~5 分, 带权元路径 $U(i)M(j)Uli=j$ 表示在一些电影上与目标用户有相同打分的用户。元路径 $U(1)M(1)U$ 表示与目标用户一样不喜欢某些电影的用户, 而元路径 $U(5)$

$M(5)U$ 则表示与目标用户一样喜欢某些电影的用户。

定义 6 (原子元路径) 如果带权元路径中的每个属性值函数都取确定的值, 这条元路径称为原子元路径。带权元路径由属性值函数取值满足限制 C 的所有原子元路径组成。

$U(1)M(1)U$ 和 $U(5)M(5)U$ 均为原子元路径, 带权元路径 $U(i)M(j)Uli=j$ 是由 $U(1)M(1)U, U(2)M(2)U, \dots, U(5)M(5)U$ 等一组原子元路径组成的。

3.2 带权元路径中的相似性度量

相似性度量是社会网络分析中的基础工作, 它可以应用到推荐、聚类分析等许多领域。随着对异质信息网络研究的发展, 许多基于元路径的相似性度量方法^[4-6]被提出。然而, 这些相似性度量方法只适用于传统的异质信息网络和传统元路径。本文将经典的相似性度量方法扩展到带权值的异质信息网络中的带权元路径中。

带权元路径 $P_l: A_1 \xrightarrow{\delta_1(R_1)} A_2 \xrightarrow{\delta_2(R_2)} \dots \xrightarrow{\delta_l(R_l)} A_{l+1} | C$ 是由各个属性值函数取所有可能的取值, 组成的一组原子元路径 (记作 P_a) 的集合。带权元路径的相似性度量分两个步骤: 首先, 由于每条原子元路径 P_a 的属性值函数都有确定的取值, 因此可以直接应用传统的相似性度量方法计算基于该原子元路径的相似性; 然后, 把基于不同原子元路径的相似性累加起来, 即可得到基于带权元路径 P_l 的相似性。需要注意的是, 对于那些采用归一化处理的相似性度量方法, 在累加原子元路径相似性之后还要进行相应的归一化。下面以 PathCount 和 PathSim 为例, 说明带权元路径的相似性度量方法。

PathCount 是基于元路径的最基本的相似性度量方法, 通过计算两个节点之间路径实例的个数来衡量节点之间的相似性。连接节点 x 和节点 y 的带权元路径的路径实例的条数等于两个节点之间所有原子元路径的路径实例个数之和, 其形式化定义为:

$$S(x, y | P_l) = \sum_{P_a \in P_l} |\{p_{x \rightarrow y} : p_{x \rightarrow y} \in P_a\}| \quad (1)$$

PathSim 是基于 PathCount 的相似性度量方法, 是 PathCount 的归一化版本, 它可以找到相似的对等实体。其形式化定义为:

$$S(x, y | P_l) = \frac{2 \times \sum_{P_a \in P_l} |\{p_{x \rightarrow y} : p_{x \rightarrow y} \in P_a\}|}{\sum_{P_a \in P_l} |\{p_{x \rightarrow x} : p_{x \rightarrow x} \in P_a\}| + \sum_{P_a \in P_l} |\{p_{y \rightarrow y} : p_{y \rightarrow y} \in P_a\}|} \quad (2)$$



4 带权元路径的应用

本节通过带权元路径在推荐系统以及相关性搜索领域的应用实验,说明了带权元路径在异质信息网络语义理解中的优越性。

4.1 带权元路径在推荐系统中的应用

基于元路径的推荐,可分为两个步骤:首先,基于指定的元路径可以计算出在该条元路径下与目标用户相似的其他用户;然后根据用户相似性,向目标用户推荐与其相似的用户偏好的电影。下面给出实验的具体方法。

给定一条元路径 P_l , S 是用户相似性矩阵,其中 $S_{u,v}^{(l)}$ 是在给定的元路径 P_l 下,用户 u 和 v 之间的相似性; R 是打分矩阵,其中 $R_{u,i}$ 是用户 u 对电影 i 的打分,用户对电影的打分取值范围为从 $1 \sim N$ 的离散值。

根据用户对电影的打分与和该用户相似的用户对该电影的打分相关的原理,首先按照式(3)计算打分强度矩阵 Q ,其中 $Q_{u,i}^{(l)}$ 代表在给定的元路径 P_l 下用户 u 给电影 i 打 r 分的打分强度。 $I_{v,i,r}$ 表示用户 v 是否给电影 i 打了 r 分。

$$Q_{u,i}^{(l)} = \sum_v S_{u,v}^{(l)} \times I_{v,i,r} \quad (3)$$

$$I_{v,i,r} = \begin{cases} 1, & R_{v,i} = r \\ 0, & \text{其他} \end{cases} \quad (4)$$

之后按照式(5),以归一化后的打分强度为分值加权来得到该用户对该电影的平均打分,以此作为用户对电影的预测评分。其中 $\hat{R}^{(l)}$ 为预测打分矩阵。

$$\hat{R}_{u,i}^{(l)} = \sum_{r=1}^N r \times \frac{Q_{u,i,r}^{(l)}}{\sum_{k=1}^N Q_{u,i,k}^{(l)}} \quad (5)$$

本实验所用的数据集是从中国著名的社交媒体网络——豆瓣电影上爬取得来的,该数据集包括 13 367 个用户、12 677 部电影以及 1 068 278 个评分项。本实验采用推荐系统中评分预测准确度的常用度量指标均方根误差 (root mean square error, RMSE) 和平均绝对误差 (mean absolute error, MAE) 来评价预测的准确度, RMSE 和 MAE 的值越小,推荐效果越好。

为了证明使用带权元路径得到的预测准确度更高,使用两条带权元路径与传统元路径进行对比,该实验使用的元路径见表 1。针对每一条元路径均使用 PathCount、PathSim、PCRW 和 HeteSim 4 种相似性度量方法计算用户之间的相似性。为了比较在不同的数据稀疏程度下的预测效

表 1 评分预测使用的元路径

元路径	路径语义
UMU	观看了相同电影的用户
$\{U(i)M(j)U\}_{i=j}$	给相同的电影打了相同分数的用户
$\{U(i)M(j)U\}_{ i-j \leq 1}$	给相同的电影打分差距不超过 1 的用户

果,训练集所占的比例分别设为 50%、60%、70%、80%、90%,其中训练集数据所占比例为 90%,意味着从所有的评分项中随机选择 90% 作为训练集来预测剩余 10% 的评分项。该实验重复进行了 10 次,评分预测的平均结果如图 5、图 6 所示。

由图 5、图 6 可知,带权元路径在所有情况下的预测效果都比传统元路径好,这说明本文提出的带权元路径可以捕捉到更微妙的语义信息。在训练集取不同的稀疏程度时,3 条元路径获得了不同的效果。当训练集数据较少时,带权元路径的预测优势相对较小,因为缺少链接数据对带权元路径的影响更大。而当训练集数据较多时,两条带权元路径的预测优势相对于传统元路径更大,并且路径 $\{U(i)M(j)U\}_{i=j}$ 的预测准确性比路径 $\{U(i)M(j)U\}_{|i-j| \leq 1}$ 更高。这说明在训练集数据较多时,链接的权重对相似性计算的影响更大。另一个很有趣的现象是带权元路径相对于传统元路径的优势,在评价指标 MAE 上体现得更为明显。原因可能是 RMSE 使用的是平方项的惩罚,它加大了对预测不准的用户电影评分的惩罚,相对于 MAE 而言, RMSE 对推荐结果的评测更加苛刻。而带权元路径比传统的元路径更稀疏,因此会有更多随机的预测,这导致了 RMSE 对带权元路径的惩罚更大。

4.2 带权元路径在相关性搜索中的应用

为了评价带权元路径在相关性搜索中的表现,爬取了 2003–2013 年 20 个顶级会议 (AAAI、CIKM、CVPR、ECIR、ECML、EDBT、ICDE、ICDM、ICML、IJCAI、KDD、PAKDD、PKDD、PODS、SDM、SIGIR、SIGMOD、VLDB、WSDM、WWW) 的所有论文。删去了共同作者超过 10 名的论文之后,得到了一个 DBLP 数据集的子数据集作为实验的主要数据集。这个数据集包括 23 661 篇论文、26 741 名作者、20 个会议以及作者和论文之间的 73 603 个链接。

本实验使用基于带权元路径、传统元路径以及基于皮尔逊相关系数的相似性计算方法进行对比,使用的元路径见表 2。基于元路径的方法使用 PathSim 作为相似性度量的方法,以美国东北大学的助理教授 Yizhou Sun 为例,分析不同相似性计算方法的实验效果。

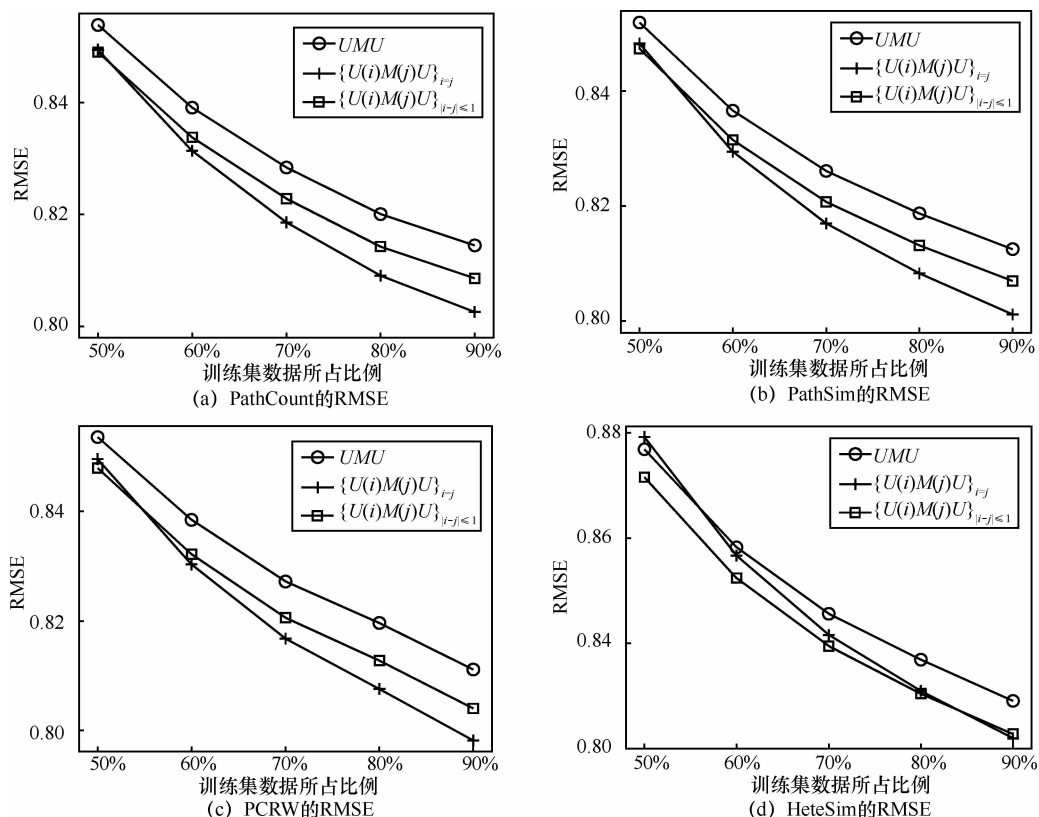


图5 不同元路径上不同相似性度量方法的 RMSE 值对比

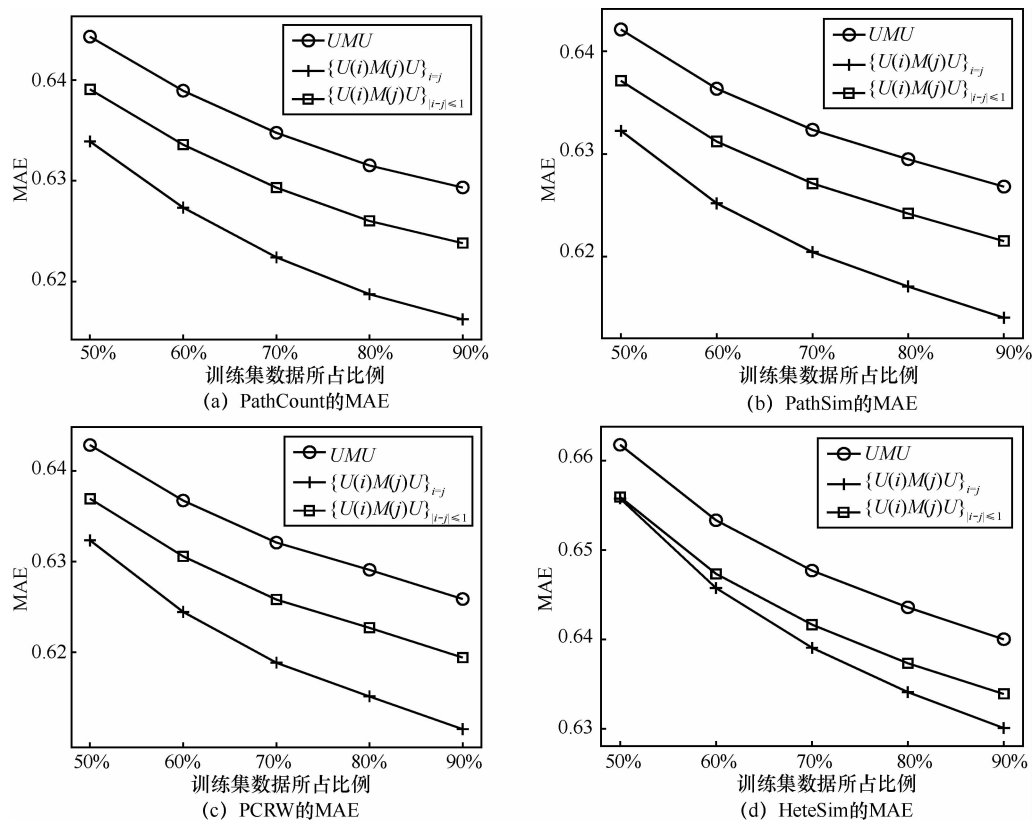


图6 不同元路径上不同相似性度量方法的 MAE 值对比



表 2 相关性搜索使用的元路径

元路径	路径语义
APCPA	在同一个会议上发表论文的作者
$\{A(i)PCP(j)A\}_{i=j}$	在同一个会议上以相同的作者顺序发表论文的作者

表 3 DBLP 数据集上,不同相似性度量方法下与“Yizhou Sun”最相似的 5 位作者

APCPA	相似性	$\{A(i)PCP(j)A\}_{i=j}$	相似性	皮尔逊相关系数	相似性
Yizhou Sun	1	Yizhou Sun	1	Yizhou Sun	1
Wilfred Ng	0.909 1	Rainer Gemulla	0.6	Tianyi Wu	0.333
Yannis Throdoridis	0.864 9	Daniele Braga	0.444 4	Yintao Yu	0.288 4
Ira Assent	0.839 3	Yiping Ke	0.44	Jiawei Han	0.285 7
Panayiotis Tsaparas	0.833 3	Gagan Agrawal	0.439	Yaguang Liu	0.235 6

从 Yizhou Sun 的个人主页可以看出,近十年中她大多是作为第一、第二作者发表论文,这说明她是一个主要的研究人员。为了显示带权元路径可以捕捉到更微妙的语义信息,分别使用这 3 种相似性计算方法寻找与 Yizhou Sun 最相似的 5 名作者,实验结果见表 3。

由表 3 可以看出,基于这 3 种相似性计算方法进行的相关性搜索得到的结果不同。在元路径 APCPA 下,与 Yizhou Sun 最相似的作者是 Wilfred Ng 和 Yannis Throdoridis,这是因为他们在 Yizhou Sun 参加的会议中发表过许多论文。而在元路径 $\{A(i)PCP(j)A\}_{i=j}$ 下,与 Yizhou Sun 最相似的作者是 Rainer Gemulla 和 Daniele Braga。Rainer Gemulla 和 Daniele Braga 不仅在 Yizhou Sun 参加的会议中发表过许多论文,并且他们在自己的大多数论文中是第一、第二作者,这说明他们也像 Yizhou Sun 一样是论文的主要执笔者。而与此同时,Wilfred Ng 和 Yannis Throdoridis 在他们过去十年的大多数论文中作为第三或者更次要的作者,这说明他们更多的是作为一名指导者的身份参与论文的写作。基于皮尔逊相关系数的相似性度量方法得出与 Yizhou Sun 最相似的作者是 Tianyi Wu 和 Yintao Yu,他们都是与 Yizhou Sun 同时期的 Jiawei Han 教授的学生,并且都有一些论文是与 Yizhou Sun 一起完成的,但在这些论文中他们大多不是主要作者。

从实验结果可以看出,基于皮尔逊相关系数的相似性度量方法只能发现“共同作者”这一关系,而基于元路径的相似性度量方法则能够发现某一学术领域中相似的作者。相对于传统元路径,带权元路径能够捕捉到“论文作者”这一关系中更为微妙的语义信息。

5 结束语

本文介绍了异质信息网络的相关概念及应用。在此基

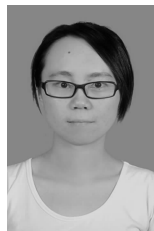
础上,提出了带权异质信息网络和带权元路径的概念,并将已有的基于元路径的相似性搜索算法扩展到带权元路径上。通过在两个真实数据集上的实验说明,在推荐、相关性搜索等任务中,提出的带权元路径能够捕捉到更微妙的语义信息,并取得更好的实验效果。

参考文献

- 1 Sun Y, Norick B, Han J, *et al.* Integrating meta-path selection with user-guided object clustering in heterogeneous information networks. Proceedings of the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Beijing, China, 2012: 723~724
- 2 Ji M, Han J, Danilevsky M. Ranking-based classification of heterogeneous information networks. Proceedings of the 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, San Diego, California, USA, 2011: 1298~1306
- 3 Yu X, Ren X, Sun Y, *et al.* Personalized entity recommendation: a heterogeneous information network approach. Proceedings of the 7th ACM WSDM Conference, New York, USA, 2014: 283~292
- 4 Sun Y, Han J, Yan X, *et al.* Pathsim: meta path-based top-k similarity search in heterogeneous information networks. Proceedings of the 37th International Conference on Very Large Data Bases, Seattle, WA, USA, 2011: 992~1003
- 5 Lao N, Cohen W W. Fast query execution for retrieval models based on path-constrained random walks. Proceedings of the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington DC, USA, 2010: 881~888
- 6 Shi C, Kong X, Huang Y, *et al.* HeteSim: a general framework for relevance measure in heterogeneous networks. IEEE Transactions on Knowledge & Data Engineering, 2014(10): 2479~2492
- 7 Sun Y, Han J. Mining heterogeneous information networks: a structural analysis approach. ACM SIGKDD Explorations Newsletter, 2013, 14(2): 20~28

- 8 Sun Y, Aggarwal C, Han J. Relation strength-aware clustering of heterogeneous information networks with incomplete attributes. Proceedings of the 38th International Conference on Very Large Data Bases, Istanbul, Turkey, 2012: 394~405
- 9 Sun Y, Han J, Zhao P, *et al.* RankClus: integrating clustering with ranking for heterogeneous information network analysis. Proceedings of EDBT, Saint-Petersburg, Russia, 2009: 565~576
- 10 Angelova R, Kasneci G, Weikum G. Graffiti: graph-based classification in heterogeneous networks. Proceedings of International Conference of World Wide Web, Lyon, France, 2012: 139~170
- 11 Jacob Y, Denoyer L, Gallinari P. Learning latent representations of nodes for classifying in heterogeneous social networks. Proceedings of the 7th ACM WSDM Conference, New York, USA, 2014: 373~382
- 12 Yang Y, Chawla N V, Sun Y, *et al.* Predicting links in multi-relational and heterogeneous networks. Proceeding of the 12th IEEE International Conference on Data Mining, Brussels, Belgium, 2012: 755~764
- 13 Sun Y, Barber R, Gupta M, *et al.* Co-author relationship prediction in heterogeneous bibliographic networks. Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining, Kaohsiung, Taiwan, China, 2011: 121~128
- 14 Zhang J, Kong X, Yu P S. Transferring heterogeneous links across location-based social networks. Proceedings of the 7th ACM WSDM Conference, New York, USA, 2014: 303~312

[作者简介]



王锐,女,北京邮电大学硕士生,主要研究方向为数据挖掘与机器学习。

张志强,男,北京邮电大学硕士生,主要研究方向为数据挖掘与机器学习。

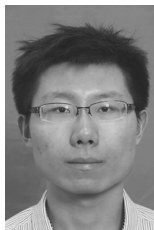
石川,男,北京邮电大学教授、博士生导师,IEEE/ACM/CCF会员,主要研究方向为机器学习、数据挖掘和演化计算。

2015166-9

(上接 2015180-9)

- Peng X M, He Y X, Zhu B J. Application of ant colony algorithm in web services composition. Computer Engineering, 2009, 35(10): 182~187
- 23 夏亚梅,程渤,陈俊亮等.基于改进蚁群算法的服务组合优化.计算机学报,2012,35(2): 2270~2281
- Xia Y M, Cheng B, Chen J L, *et al.* Optimizing services composition based on improved ant colony algorithm. Chinese Journal of Computers, 2012, 35(2): 2270~2281
- 24 Zhao S, Ma L, Wang L, *et al.* An improved ant colony optimization algorithm for QoS-aware dynamic web service composition. Proceedings of the International Conference on Industrial Control and Electronics Engineering, Xi'an, China, 2012
- 25 Wei Z, Chang C K, Taiming F, *et al.* QoS-based dynamic web service composition with ant colony optimization. Proceedings of the IEEE 34th Annual Computer Software and Applications Conference, Seoul, Korea, 2010
- 26 Zhang L J, Li B. Requirements driven dynamic services composition for web services and grid solutions. Journal of Grid Computing, 2004, 2(2): 121~140
- 27 Dorigo M, Gambardella L M. Ant colony system: a cooperative learning approach to the traveling salesman problem. IEEE Transactions on Evolutionary Computation, 1997, 1(1): 53~66
- 28 Dorigo M, Maniezzo V, Colnari A. Ant system: optimization by a colony of cooperating agents. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 1996, 26 (1): 29~41

[作者简介]



孙明瑞,男,哈尔滨工业大学计算机科学与技术学院博士生,主要研究方向为服务计算与服务网络、大数据分析算法。

臧天仪,男,博士,哈尔滨工业大学计算机科学与技术学院教授、博士生导师,主要研究方向为大数据计算与算法、服务计算与服务网络、云计算、复杂系统建模及其验证分析、e-Healthcare。

刘偲俣,男,北京航空航天大学自动化科学与电气工程学院在读,主要研究方向为建模和仿真实验。

2015180-10