异质网络中基于节点影响力的相似度度量方法

刘 露[1,2,3,4],胡封晔[4],牛 亮[5],彭 涛[1,2,3]
(1. 吉林大学软件学院,吉林长春 130012;2. 符号计算与知识工程教育部重点实验室(吉林大学),吉林长春 130012;
3. 吉林大学计算机科学与技术学院,吉林长春 130012;4. 吉林大学通信工程学院,吉林长春 130012;
5. 吉林大学第一医院,吉林长春 130012)

摘 要: 异质网络相似度学习,即分析两个不同类型对象间的相关程度. 不同类型对象在异质网络中的重要程度不同,它们在相似度学习过程中的发挥的作用也不同. 针对异质网络,提出了一种基于节点影响力的相似度度量方法 NISim,该模型既考虑了网络中的链接结构,也保留了网络中的语义信息,同时区分不同类型节点对异质网络的作用. 在异质信息网络环境下,通过启发式规则区分并量化不同类型节点的影响力权值,并结合网络链接结构和节点间语义关系,解决了提高相似度学习准确性的问题. 实验结果表明,该方法能够有效地对异质信息网络不同类型节点进行相似度度量,可以应用在网络搜索、推荐系统以及知识图谱构建等不同领域.

关键词: 数据挖掘;异质网络;推荐系统;知识图谱;网络搜索;节点影响力;链接结构;语义关系
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2019)09-1929-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2019.09.016

Node Influence Based Similarity Measure Method in Heterogeneous Network

LIU Lu[1,2,3,4], HU Feng-ye[4], NIU Liang[5], PENG Tao[1,2,3]
(1. *College of Software, Jilin University, Changchun, Jilin* 130012, *China*;
2. *Key Laboratory of Symbolic Computation and Knowledge Engineering (Jilin University), Ministry of Education, Changchun, Jilin* 130012, *China*;
3. *College of Computer Science and Technology, Jilin University, Changchun, Jilin* 130012, *China*;
4. *College of Communication Engineering, Jilin University, Changchun, Jilin* 130012, *China*;
5. *The First Hospital of Jilin University, Changchun, Jilin* 130012, *China*)

Abstract: Heterogeneous network similarity learning is to analyze the degree of correlation between two different types of objects. Different types of objects have different degrees of importance in heterogeneous networks, and play different roles in the similarity learning process. This paper proposes a node influence based similarity measure method (NISim) heterogeneous information network. This method not only considers the link structure in network but also keeps the semantic information in heterogeneous networks. Also, this method distinguishes the effect to heterogeneous network brought by different types of nodes. In heterogeneous network, the heuristic rules are used to distinguish and quantify the influence weight of different types of nodes. In addition, the link structure in network and the semantic relationship are combined to solve the problem of improving similarity learning accuracy. Experimental results show that this method can measure the similarity between different types of nodes effectively. It can be applied in different fields such as network search, recommendation system and knowledge graph construction and so on.

Key words: data mining; heterogeneous network; recommended system; knowledge graph; network search; node influence; link structure; semantic relationship

收稿日期:2018-06-15;修回日期:2019-04-17;责任编辑:蓝红杰
基金项目:国家自然科学基金(No. 61872163, No. 61806084);中国博士后科研基金项目(No. 2018M631872);吉林省教育厅项目(No. JJKH20190160KJ);吉林省科技厅重点科技研发项目(No. 20180201044GX)

1 引言

相似度搜索是异质信息网络挖掘的一项重要任务,在网络搜索[1]、推荐系统[2]、属性标注[3]、以及知识图谱构建[4]等诸多应用领域中都有着重要的作用. 这些应用领域的研究任务大多可以被归结为学习并应用一个特定的相似度函数来计算两种类型实体间的相关程度. 传统的面向同质数据的相似度搜索方法(计算相同类型实体间相关程度)已经不能满足异质信息网络挖掘的需求. 异质信息网络中多种类型的对象和多种类型的关系可以为研究者提供更加丰富的结构信息和语义信息[5]. 由于实体与链接的异质性、网络结构的复杂性以及网络的动态性等因素的存在,所以,在探索网络中复杂的对象和关系时,应同时考虑对象间的链接结构和语义关系.

目前,研究者已经提出了一些相似度计算方法,但多数是针对同质信息网络提出的,主要包括:(1) 基于特征的相似度计算方法[6~8];(2) 基于链接的相似度计算方法[9,10]. 基于特征的相似度计算方法主要通过对象的特征值来度量它们之间的相似度. Mehdi 等人[6]提出了一种基于实例的模式匹配方法,应用谷歌相似度和正则表达式来识别一对一的模式. Zhang 等人[9]提出一种基于链接的相似度度量方法,该方法结合了 P-Rank、SimRank 以及 Personalized PageRank 来进行相似度度量. 同质信息网络中基于链接的相似度计算方法考虑了节点间的链接关系,但并不考虑每条路径的语义信息,也不区分网络中不同类型的节点. 因此,以上同质网络中基于特征和基于链接的相似度计算方法不能直接用于异质信息网络相似度搜索过程.

相比同质信息网络,异质信息网络相似度搜索方法更关注两个不同类型对象间的相近关系. 现有异质信息网络相似度搜索方法主要包括:(1) 基于路径的相似度搜索方法[11~13];(2) 基于矩阵相乘的相似度搜索方法[14,15];(3) 基于近邻节点的相似度搜索方法[16,17]. 基于路径的相似度搜索方法主要通过链接节点间的路径来传递链接信息及语义信息,进而计算节点间的相似度. Sun 等人[12]提出了一种基于元路径的相似度搜索方法 PathSim,该方法引入了基于元路径的相似度的概念,考虑了网络中不同路径对应的不同语义信息,来度量节点间相关程度. 然而,在许多情况下,我们需要度量不同类型对象的相关程度,如作者与会议间相关程度,用户与电影间相关程度等. 尤其在异质信息网络中,不同类型对象间相似度搜索变得更加重要. Shi 等人[18]提出了一种度量异质信息网络对象间相关程度的方法 HeteSim,该方法既可以计算相同类型对象间的相关程度,也可以计算不同类型对象间的相关程度.

不同类型的对象以及它们之间不同类型的链接关系构成了复杂的异质信息网络. 基于矩阵相乘的相似度计算方法通过邻接矩阵来表示相同类型对象或不同类型对象间的链接权值. Xiong 等人[14]提出了一种基于矩阵相乘的相似度连接方法来根据用户指定的路径按相似程度返回 k 对对象. Zhang 等人[15]提出了一种面向星型网络的相似度搜索方法,根据用户指定的星型网络中心节点查找与该节点相似的中心节点. 也有研究者通过分析目标对象的近邻节点来进行相似度学习. Shams 和 Haratizadeh[16]提出了一种根据近邻节点相似度来进行协同排序的方法,用协同排序的方法来进行合理推荐. Tao 等人[17]提出了一种基于 SimRank 的相似性连接方法,只挖掘 k 对相似度较高的节点对. 该方法通过分析每个节点的邻居节点,根据邻居节点的信息将每个节点用向量表示,将 SimRank 相似度计算方法转化为向量计算,来返回异质网络中最相似的对象对.

异质网络中不同类型节点的作用不同,对其他节点的影响力也应该区别对待,本文提出了一种基于节点影响力的异质网络相似度度量方法. 根据三个启发式规则对异质网络中不同类型节点赋予不同的影响力权值,突出不同节点在相似度计算过程中的不同作用. 在计算异质网络相似度过程中,分别从链接结构相似度和语义关系相似度这两个角度进行建模,并将节点影响力权值加入到计算过程中.

2 问题定义

在本节中,我们给出异质信息网络以及相似度度量涉及到的一些基本定义和概念.

定义 1 异质信息网络[12]. 给定一个有向图 $G=(\mathcal{V},\mathcal{E};\tau,\varphi;\mathcal{A},\mathcal{R})$. $\mathcal{V}$ 代表节点集,$\mathcal{E}$ 代表边集. τ 表示对象类型映射函数. φ 表示关系类型映射函数. $\tau(v)\in\mathcal{A}$ 表示每个对象 $v\in\mathcal{V}$ 都属于一个特定的对象类. $\varphi(e)\in R$ 表示每个关系 $e\in\mathcal{E}$ 都属于一种特定的关系类. 当节点类型数量 $|\mathcal{A}|>1$ 或边的类型数量 $|\mathcal{R}|>1$ 时,这样的信息网络被称为异质信息网络. 反之为同质信息网络.

定义 2 元路径[12]. 给定一个异质信息网络 $G=(\mathcal{V},\mathcal{E};\tau,\varphi;\mathcal{A},\mathcal{R})$,一条元路径 Path 是由对象类型组成的有序序列,由 $\text{Path}=A_1A_2\cdots A_l$ 或 $\text{Path}=A_1\circ A_2\circ\cdots\circ A_l$ 表示. 处于类型 A_1 位置的对象被称为起始节点(starting node),由 s 表示. 处于类型 A_l 位置的节点被称为终止节点(ending node),由 e 表示. 在本文中,我们根据不同元路径度量起始节点与终止节点的相关程度.

定义 3 节点的影响力. 给定一个异质信息网络 G,综合节点类型、节点邻居数量以及节点对所在社区代表程度等不同因素为节点赋予不同权值,该权值即表

示异质网络中节点的影响力.

定义 4　结构语义相似度矩阵. 假设 $\boldsymbol{L}=(l_{ij})$ 和 $\boldsymbol{S}=(s_{ij})$ 均为 $m\times n$ 的矩阵,用来存储不同类型(或相同类型)对象间的链接结构相似度和语义关系相似度. 结构语义相似度矩阵 $\boldsymbol{T}_{LS}=(ls_{ij})_{mn}$. 对于矩阵中每个点对 (i,j) 对应的值 ls_{ij} 为 $\alpha\boldsymbol{L}+\beta\boldsymbol{S}$ 的值,其中,矩阵 $\boldsymbol{L}$ 和矩阵 $\boldsymbol{S}$ 需满足 $\alpha\boldsymbol{L}_{m\times nF}-\beta\boldsymbol{S}_{m\times nF}<\varepsilon$. α 和 β 为调节因子.

例如:$\boldsymbol{L}_{3\times3}=\begin{bmatrix}2&3&7\\9&5&1\\6&4&8\end{bmatrix}$,$\boldsymbol{S}_{3\times3}=\begin{bmatrix}1&5&3\\2&2&12\\1&9&3\end{bmatrix}$($\alpha=\beta=1$),那么,结构语义相似度矩阵为 $\boldsymbol{T}_{LS}=\begin{bmatrix}2+1&3+5&7+3\\9+2&5+2&1+12\\6+1&4+9&8+3\end{bmatrix}=\begin{bmatrix}3&8&10\\11&7&13\\7&13&11\end{bmatrix}$. 其中,$\|\cdot\|_{\mathrm{F}}$ 表示矩阵的 Frobenis 范式,例如:$\|\boldsymbol{A}\|_{\mathrm{F}}=\left(\sum_{m,n}A_{mn}^2\right)^{1/2}$.

定义 5　出度受限集合. 对于图 G 中任意节点 $v\in\mathcal{V}$,当节点 v' 满足下列两个性质时,节点 v' 将被加入到节点 v 关于类型 A_i 的出度受限集合中,记作 $\mathrm{COut}(v)_{\tau(v)\to A_i}$.

(1)$v'\in\mathcal{V}$ 且 $(v,v')\in\varepsilon$;

(2)$\tau(v')\in A_i$ 且 $A_i\in\mathcal{A}$.

其中,当 $(v,v')\in\varepsilon$ 时,v 为前驱节点,v' 为后继节点. 反之,当 $(v',v)\in\varepsilon$ 时,v' 为前驱节点,v 为后继节点.

3　异质信息网络节点影响力计算过程

给定一个异质信息网络 $G=(\mathcal{V},\varepsilon)$,对于网络中任意一个目标节点 $v\in\mathcal{V}$,根据以下 3 条启发式规则来计算不同类型节点的影响力.

规则 1　目标节点和与其相同类型节点的度数越大,该目标节点的影响力就越大.

例如:文献信息网络中,论文和论文之间存在引用关系,他引数量越高,该论文的影响力越大;社交网络中,用户之间存在关注关系,用户被关注人数越多,该用户的影响力越大. 当异质网络为有向图时,我们将区分节点的出度与入度. 目标节点 i 基于规则 1 的节点影响力权值被表示为

$$\lambda_i^1=\frac{d_i}{m-1}\tag{1}$$

其中,m 为节点 i 所在类型的节点总数,d_i 为节点 i 的度数.

规则 2　对于给定路径,通过目标节点的路径数越多,该目标节点的影响力越大.

例如:给定一条路径“作者—论文—出版物名称(APC)”,若目标节点为“作者”,出版物名称为“KDD”,那么,在 KDD 上发表论文较多的作者,与 KDD 的相关程度更高,该节点的影响力也较大. 目标节点 i 基于规则 2 的节点影响力权值被表示为

$$\lambda_i^2=\frac{p_i}{\mathrm{pNum}}\tag{2}$$

其中,pNum 为给定节点类型对应的路径总数,p_i 为对应路径中通过节点 i 的路径数量.

规则 3　在异质网络中,某一类型节点越重要,该类型节点对异质网络影响力越大.

异质网络中包含不同类型的节点,不同类型节点对异质网络具有不同的影响力,比如,社交网络中包含用户、用户年龄、用户性别、发表的内容、图片、音频等不同类型的对象,显然,在进行知识发掘的过程中,发表的内容、图片及音频这些类型对象的重要程度要高于用户年龄和性别的重要程度,那么,重要程度高的节点对异质网络的影响力也越大. 基于规则 3 的节点影响力权值需满足一个条件,即所有节点类型对应的影响力权值之和为 1.

我们应用以上 3 条规则来进行异质信息网络链接结构相似度度量,得到 λ^1,λ^2 和 λ^3 后,需要对同一类型节点的影响力权值进行归一化[19],使 λ^1,λ^2 和 λ^3 的取值范围为[0,1]. 在异质信息网络中,节点最终的影响力权值为 λ^1,λ^2 及 λ^3 的和.

4　基于节点影响力的相似度度量方法

4.1　基于节点影响力的链接结构相似度模型

给定一个异质网络 G,m 个 A_1 类型的对象以及 n 个 A_2 类型的对象. 矩阵 $\boldsymbol{L}_{m\times n}$ 中存储 $m\times n$ 个对象间的链接结构相似度. 例如,在文献信息网络中,我们可以计算作者和研究方向(关键词)间的相关程度,也可以计算作者与作者间的相关程度;在社交网络中,我们可以计算用户和主题间的相关程度,也可以计算用户和用户间的相关程度. 两个不同类型对象间的相关程度存储在矩阵 $\boldsymbol{L}$ 中,表示方法如下:

$$\boldsymbol{L}_{m\times n}=\begin{bmatrix}\boldsymbol{l}_1\\\boldsymbol{l}_2\\\vdots\\\boldsymbol{l}_m\end{bmatrix}=\begin{bmatrix}l_{11}&l_{12}&\cdots&l_{1n}\\l_{21}&l_{22}&\cdots&l_{2n}\\\vdots&\vdots&\ddots&\vdots\\l_{m1}&l_{m2}&\cdots&l_{mn}\end{bmatrix}$$

其中,

$$l_{ij}=\begin{cases}\sum_{a=1}^{k}\sum_{b=1}^{k}\frac{1}{a\times b}\cdot\frac{\mathrm{SumInf}(\mathrm{Dist}_a(i)\cap\mathrm{Dist}_b(j))}{\mathrm{SumInf}(\mathrm{Dist}_a(i)\cup\mathrm{Dist}_b(j))} & ,i\neq j\ \mathrm{and}(i,j)\notin\varepsilon\\ 1 & ,i=j\ \mathrm{or}\ (i,j)\in\varepsilon\end{cases}\tag{3}$$

$\mathrm{Dist}_a(i)$表示和节点 i 最短距离为 a 的邻居节点集合. $\mathrm{SumInf}(T)$表示集合 T 中节点的影响力权值之和. 举一个具体的例子对$\mathrm{Dist}_a(i)$进行说明. 图 1 为一个具有 10 个节点的局部网络,此时,我们忽略节点的类型. 节点 A(至节点 J)最短距离为 1 ~4 的邻居节点集合如表 1 所示. 基于节点影响力的链接结构相似度算法描述如算法 1 所示.

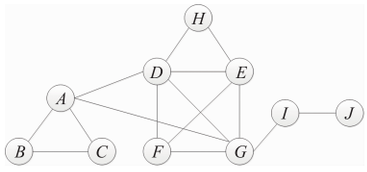

图1 由10个节点构成的局部网络

表 1 与图 1 中节点最短距离分别为 1 ~4 的邻居节点集合

	Dist_1	Dist_2	Dist_3	Dist_4
A	$\{B,C,D,G\}$	$\{E,F,H,I\}$	$\{J\}$	
B	$\{A,C\}$	$\{D,G\}$	$\{E,F,H,I\}$	$\{J\}$
C	$\{A,B\}$	$\{D,G\}$	$\{E,F,H,I\}$	$\{J\}$
D	$\{A,E,F,G,H\}$	$\{B,C,I\}$	$\{J\}$	
E	$\{D,F,G,H\}$	$\{A,I\}$	$\{B,C,J\}$	
F	$\{D,E,G\}$	$\{A,H,I\}$	$\{B,C,J\}$	
G	$\{A,D,E,F,I\}$	$\{B,C,H,J\}$		
H	$\{D,E\}$	$\{A,F,G\}$	$\{B,C,I\}$	$\{J\}$
I	$\{G,J\}$	$\{A,D,E,F\}$	$\{B,C,H\}$	
J	$\{I\}$	$\{G\}$	$\{A,D,E,F\}$	$\{B,C,H\}$

算法 1 基于节点影响力的链接结构相似度计算

输入:异质网络 G,给定路径 Path,m 个 A_1类型的对象,n 个 A_2类型的对象;

输出:m 个 A_1类型的对象与 n 个 A_2类型的对象间链接结构相似度,存储在矩阵 $\boldsymbol{L}$ 中.

```
for each nodei in Path
    λ_i^1 ← d_i/(m−1) /* d_i 为节点 i 的度数 */
    λ_i^2 ← p_i/pNum
end for
for each node i in Path
    for m = 1 to 4
        Dist_m^i ← getShortestDis(i,m) /* 将与节点 i 最短距离为 m 的节点加入集合Dist_m^i 中 */
    end for
end for
calculating l_ij using Eq. 3
```

4.2 基于节点影响力的语义关系相似度模型

仅考虑异质网络中的链接结构来度量节点间的相近关系是不准确且不全面的. 我们应用异质网络中的元路径,考虑节点间不同类型关系蕴含的不同语义,在区分不同类型节点影响力的基础上构建语义关系相似度模型. 和链接结构相似度矩阵 $\boldsymbol{L}$ 类似,矩阵 $\boldsymbol{S}$ 用来存储与 $\boldsymbol{L}$ 矩阵相同的 $m\times n$ 个对象间的语义关系相似度,表示方法如下:

$$\boldsymbol{S}_{m\times n}=\begin{bmatrix}\boldsymbol{s}_1\\ \boldsymbol{s}_2\\ \vdots\\ \boldsymbol{s}_m\end{bmatrix}=\begin{bmatrix}s_{11} & s_{12} & \cdots & s_{1n}\\ s_{21} & s_{22} & \cdots & s_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ s_{m1} & s_{m2} & \cdots & s_{mn}\end{bmatrix}$$

其中,

$$s_{ij}=\frac{2h(i,j|\mathrm{Path})\cdot h(j,i|\mathrm{Path}^{-1})}{h(i,j|\mathrm{Path})+h(j,i|\mathrm{Path}^{-1})}\tag{4}$$

$h(i,j|\mathrm{Path})$表示起始节点 i 到终止节点 j 在路径 Path 上的可达概率. 起始节点 i 到终止节点 j 在路径 Path 上的可达概率和终止节点 j 到起始节点 i 的可达概率的调和平均值为节点 i 和节点 j 最终的语义关系相似度,被存储在矩阵 S 中. $h(i,j|\mathrm{Path})$的计算方法如下:

$$\begin{aligned}h(i,j|\ \mathrm{Path})&=h(i,j|\ A_1\circ A_2\circ\cdots\circ A_l)\\&=\frac{\sum_{a=1}^{|\mathrm{COut}(i)_{\tau(i)\to A_2}|}\left[\frac{(\lambda_i^1+\lambda_i^2+\lambda_i^3)}{3}+1\right]\cdot h(\mathrm{COut}(i)_{\tau(i)\to A_2},j|\ A_2\circ\cdots\circ A_l)}{(|\ \mathrm{COut}(i)_{\tau(i)\to A_2}|+1)}\end{aligned}\tag{5}$$

其中,$\mathrm{COut}(i)_{\tau(i)\to A_2}$表示节点 i 相对于类型 A_2 的出度受限集合. $\lambda_i^1+\lambda_i^2+\lambda_i^3$ 表示节点 i 的影响力权值总和. 由于公式迭代进行语义关系相似度计算,下一步迭代时节点 i 由类型 A_2 的节点替代,类型 A_2 由类型 A_3 替代. 基于节点影响力的语义相似度模型的计算终止条件为:若起始节点与终止节点相遇,则相似度乘以 1,否则,该路径语义相似度为 0,即起始节点不能与终止节点相遇.

异质信息网络中存在多种类型的路径,蕴含不同的语义关系. 但是仅通过计算和分析路径可达实体间的相关程度是远远不够的,得到的不同类型节点间的相关程度也是不够准确的. 为此,我们应用协同相似

度[20]来计算同类型节点间的相似性,并将与路径中节点有语义相关的所有节点都加入到语义关系相似度计算当中. 算法 2 描述了基于节点影响力的语义关系相似度计算过程.

算法 2　基于节点影响力的语义关系相似度计算

输入:异质网络 G,给定路径 Path,m 个 A_1 类型的对象,n 个 A_2 类型的对象;

输出:m 个 A_1 类型的对象与 n 个 A_2 类型的对象间语义关系相似度,存储在矩阵 $\boldsymbol{S}$ 中.

1. for each nodei in Path
2. $\lambda_i^1 \leftarrow \frac{d_i}{m-1}$ /* d_i 为节点 i 的度数 */
3. $\lambda_i^2 \leftarrow \frac{p_i}{\text{pNum}}$
4. end for
5. for each node i in Path
6. add its semantically related nodes in local network /* 加入语义相关节点 */
7. end for
8. calculatingh$(i,j|\text{Path})$ using Eq. 5
9. calculating s_{ij} using Eq. 4 /* 计算语义关系相似度 */

4.3　结合结构、语义以及节点影响力的相似度计算过程

本节将根据链接结构相似度和语义关系相似度,结合不同类型节点影响力,来详细介绍异质网络相似度学习过程. 我们对 4.1 节构建的链接结构相似度矩阵 $\boldsymbol{L}$ 与 4.2 节构建的语义关系相似度矩阵 $\boldsymbol{S}$ 中的每个向量进行线性变换,结合二者得到最终的节点相似度矩阵,如式(6)所示. 为了使链接结构相似度和语义关系相似度具有相同的数量级,矩阵 $\boldsymbol{L}$ 和矩阵 $\boldsymbol{S}$ 分别乘以标量 α 和 β,使 $\alpha\boldsymbol{L}_{m\times n}$ 和 $\beta\boldsymbol{S}_{m\times n}$ 的 Frobenis 范式差值小于一个常量 ε.

$$\begin{aligned}\boldsymbol{T}_{LS} &= \begin{bmatrix} \alpha l_{11}+\beta s_{11} & \alpha l_{12}+\beta s_{12} & \cdots & \alpha l_{1n}+\beta s_{1n} \\ \alpha l_{21}+\beta s_{21} & \alpha l_{22}+\beta s_{22} & \cdots & \alpha l_{2n}+\beta s_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ \alpha l_{m1}+\beta s_{m1} & \alpha l_{m2}+\beta s_{m2} & \cdots & \alpha l_{mn}+\beta s_{mn} \end{bmatrix} \\ &= \alpha\boldsymbol{L}_{m\times n}+\beta\boldsymbol{S}_{m\times n} \qquad (6) \\ &\text{s.t. } \|\alpha\boldsymbol{L}_{m\times n}\|_{\mathrm{F}} - \|\beta\boldsymbol{S}_{m\times n}\|_{\mathrm{F}} < \varepsilon\end{aligned}$$

约束条件 $\|\alpha\boldsymbol{L}_{m\times n}\|_{\mathrm{F}} - \|\beta\boldsymbol{S}_{m\times n}\|_{\mathrm{F}} < \varepsilon$ 确保了链接结构相似度和语义关系相似度具有相同的数量级. α 和 β 被用来调节 $\boldsymbol{L}$ 和 $\boldsymbol{S}$ 的 Frobenis 范式值,初始值均为 1. 若 $\|\boldsymbol{L}\|_{\mathrm{F}}$ 的值远大于 $\|\boldsymbol{S}\|_{\mathrm{F}}$,则 α 保持不变,β 增加 0.1. 相反,若 $\|\boldsymbol{S}\|_{\mathrm{F}}$ 的值远大于 $\|\boldsymbol{L}\|_{\mathrm{F}}$,则 β 保持不变,α 增加 0.1. 继续迭代计算是否满足条件 $\|\alpha\boldsymbol{L}_{m\times n}\|_{\mathrm{F}} - \|\beta\boldsymbol{S}_{m\times n}\|_{\mathrm{F}} < \varepsilon$,若满足条件,算法终止,否则,继续变化 α 和 β 的值直到满足终止条件. 算法结束时,记录此时 α 和 β 的值,得到链接语义相似度矩阵 $\boldsymbol{T}_{LS}$. 结合结构、语义以及节点影响力的相似度计算过程如算法 3 所示. 基于节点影响力的相似度度量方法的整体框架如图 2 所示. 通过启发式规则计算节点影响力权值,从结构和语义的角度,分别构建不同类型节点链接结构相似度模型和语义关系相似度模型,并将节点影响力权值应用到二者当中,来计算不同类型节点间的相似度.

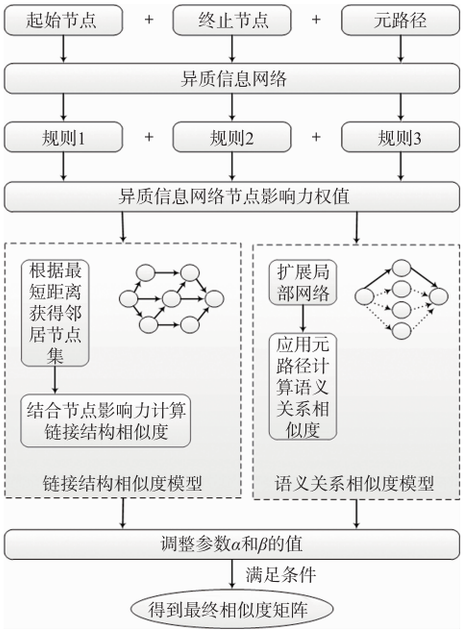

图2　基于节点影响力相似度度量方法整体框架

算法 3　结合结构、语义以及节点影响力的相似度计算

输入:链接结构相似度矩阵 $\boldsymbol{L}$,语义关系相似度矩阵 $\boldsymbol{S}$;

输出:链接语义相似度矩阵 $\boldsymbol{T}_{LS}$.

1. Initialize $\alpha\leftarrow 1$,$\beta\leftarrow 1$ and Calculate $\|\boldsymbol{L}\|_{\mathrm{F}}$ and $\|\boldsymbol{S}\|_{\mathrm{F}}$
2. if $\|\boldsymbol{L}\|_{\mathrm{F}} \geq \|\boldsymbol{S}\|_{\mathrm{F}}$ then
3. while $\|\alpha\boldsymbol{L}\|_{\mathrm{F}} - \|\beta\boldsymbol{S}\|_{\mathrm{F}} \geq \varepsilon$
4. $\beta\leftarrow\beta+0.1$ /* 调整矩阵 $\boldsymbol{S}$ 的 Frobenis 范式值 */
5. end while
6. else
7. while $\|\alpha\boldsymbol{L}\|_{\mathrm{F}} - \|\beta\boldsymbol{S}\|_{\mathrm{F}} \geq \varepsilon$
8. $\alpha\leftarrow\alpha+0.1$
9. end while
10. end if
11. calculating $\boldsymbol{T}_{LS}$ using Eq. 6.

5　实验与结果

在本节中,我们基于不同类型节点影响力,结合节点链接结构相似度和语义关系相似度构建了一个面向异质信息网络的相似度度量模型,并且在 2 个数据集上进行了测试. 在 5.1 节中,我们介绍了聚类度量标准以评估我们提出的方法. 5.2 节对 2 个数据集进行了介

绍. 实验结果和分析在 5.3 节中给出.

5.1 度量标准

我们将采用兰德指数[21]和 NMI[22]这两个度量标准来评估基于节点影响力相似度度量方法的性能.

兰德指数 RI 作为一种聚类的度量标准,描述了聚类算法对无标注数据的区分程度,其计算公式如下:

$$\mathrm{RI}=\frac{A+B}{C_n^2} \tag{7}$$

其中,A 表示在得到的聚簇集合与数据真实聚簇集合中都为同一类别的元素对数,B 表示在得到的聚簇集合与数据真实聚簇集合中都为不同类别的元素对数,$C_n^2=\frac{n!}{2!\ (n-2)!}$.

标准化互信息 NMI,作为另一种聚类度量标准,也被应用于本次实验中. NMI 的计算公式如下:

$$\mathrm{NMI}(U,V)=\frac{\sum_{h=1}^{|U|}\sum_{l=1}^{|V|}p(h,l)\log\left(\frac{p(h,l)}{p(h)p(l)}\right)}{\frac{1}{2}\left(-\sum_{h=1}^{|U|}p(h)\log(p(h))-\sum_{l=1}^{|V|}p(l)\log(p(l))\right)} \tag{8}$$

其中,U 和 V 表示两个聚簇,$p(h)$ 和 $p(l)$ 分别为 h 和 l 在 U 和 V 中的概率函数. $p(h,l)$ 为 h 和 l 的联合概率分布. NMI 的取值范围为[0,1]. NMI 的值越高,聚类效果越好.

5.2 数据集

本文将在 DBLP 和 AMiner 2 个数据集上测试我们提出的方法. 下面我们从数据集大小和数据特点等方面分别介绍这 2 个数据集.

DBLP(http://dblp.uni-trier.de/xml/)是异质网络中常用的实验数据,数据集是以引文的形式存储在 xml 文件中,其中包括出版物名称、作者、发表时间等等. 我们根据关键词搜索得到 4 类(社交网络、文本挖掘、信息检索、机器学习)2000 篇论文及其对应的相关信息,构成本文实验数据集.

AMiner 数据集[23]是一个异质的文献信息网络. 它由 3 个部分构成,分别是 AMiner-Author, AMiner-Paper, AMiner-Coauthor. 这 3 部分共包括 1712433 个作者和 2092356 篇论文的信息,这些节点和节点之间的关系构成了涵盖计算机不同领域的异质信息网络. 同样,我们根据关键词提取 4 类(社交网络、文本挖掘、信息检索、机器学习)共 2000 篇论文及其对应的相关信息,构成本文实验数据集.

5.3 结果及分析

在本节中,我们通过 3 个实验对本文提出的方法进行评估. 在第 1 个实验中,分析不同的 α 和 β 值对聚类结果的影响,来确定不同数据集中的最佳参数选择. 当矩阵 $\boldsymbol{L}$ 的 Frobenius 范式值大于矩阵 $\boldsymbol{S}$ 的 Frobenius 范式值时,参数 β 的值增加 0.1,反之,参数 α 的值增加 0.1,直到矩阵 $\boldsymbol{L}$ 和矩阵 $\boldsymbol{S}$ 的 Frobenius 范式差值小于阈值 ε. 表 2 和表 3 给出了在 DBLP 和 AMiner 数据集中,随着参数 α 和 β 的变化,对应的兰德指数和 NMI. 此时,我们不考虑不同类型节点的影响力权值. 从表 2 和表 3 中可以看出,在不考虑节点影响力权值的情况下,在 DBLP 数据集中,当 α 和 β 的取值分别为 2.7 和 1 时,RI 和 NMI 达到峰值,此时聚类效果最好,即相似度度量模型效果达到最优. 类似地,在 AMiner 数据集中,当 α 和 β 的取值分别为 1.9 和 1 时,RI 和 NMI 达到峰值. 在接下来的实验中,我们将依据实验 1 的结果来选取相应的 α 和 β 值.

在第 2 个实验中,我们通过对比应用节点影响力权值前后,相似度度量方法对聚类效果的影响来验证节点影响力权值的有效性. 我们选取路径 APCPA 和 CPAPC 分别对作者和出版物名称进行聚类,表 4 给出了在 DBLP 和 AMiner 数据集中,应用节点影响力权值前后,RI 及 NMI 的变化. 从实验结果可以看出,在同一数据集中,将考虑节点影响力的相似度度量方法可以获得更好的聚类效果.

表 2 在 DBLP 中,随着 α 和 β 的变化,NISim 在聚类中对应的 RI 和 NMI

α	β	RI	NMI
2.2	1	0.791	0.735
2.3	1	0.804	0.742
2.4	1	0.810	0.751
2.5	1	0.817	0.760
2.6	1	0.825	0.773
2.7	1	0.836	0.787
2.8	1	0.832	0.781
2.9	1	0.827	0.773
3.0	1	0.821	0.767
3.1	1	0.812	0.761

表 3 在 AMiner 中,随着 α 和 β 的变化,NISim 在聚类中对应的 RI 和 NMI

α	β	RI	NMI
1.4	1	0.783	0.724
1.5	1	0.789	0.735
1.6	1	0.795	0.747
1.7	1	0.802	0.760
1.8	1	0.816	0.771
1.9	1	0.825	0.780
2.0	1	0.818	0.774
2.1	1	0.807	0.767
2.2	1	0.791	0.760
2.3	1	0.786	0.753

表 4　在 DBLP 和 AMiner 数据集中,应用节点影响力权值前后,NISim 方法的 RI 及 NMI 值

DBLP					
APCPA	不考虑节点影响力	考虑节点影响力	CPAPC	不考虑节点影响力	考虑节点影响力
RI	0.836	0.853	RI	0.867	0.899
NMI	0.787	0.824	NMI	0.875	0.903
AMiner					
APCPA	不考虑节点影响力	考虑节点影响力	CPAPC	不考虑节点影响力	考虑节点影响力
RI	0.825	0.841	RI	0.850	0.884
NMI	0.780	0.818	NMI	0.862	0.890

在实验 3 中,我们通过指定出版物名称,按作者相关程度排序,判断 Top-100 的作者是否与该出版物处于同一聚簇,进而计算相似度度量方法的准确率. 算法初始时,我们对作者与出版物所在聚簇进行类别标记,若二者处于相同聚簇,则认为它们是相似的. 我们选取路径 APC 来度量作者与给定出版物之间的相似程度,进行 10 次实验取平均值. 将提出的 NISim 算法与 HeteSim[18]、AvgSim[24] 和 PathSim[12] 算法进行比较,图 3 描述了在 DBLP 和 AMiner 数据集上,4 种相似度算法的准确率. 实验结果表明,本文提出的 NISim 算法在相似度计算效果上明显优于 HeteSim、AvgSim 和 PathSim 算法.

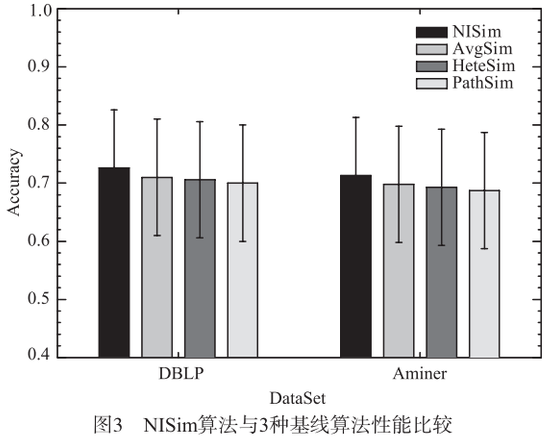

图3　NISim算法与3种基线算法性能比较

6　总结

本文提出了一种基于节点影响力的相似度度量方法. 异质信息网络中不同类型的节点对整个网络的作用不同,在相似度计算过程中应区别对待. 将节点影响力权值应用到异质网络相似度度量模型当中也是一个新的尝试,本文提出的相似度度量方法包括链接结构相似度和语义关系相似度,在考虑了不同类型节点间链接结构的同时也利用了节点间蕴含的语义信息. 通过直接及间接相连的节点度量结构相似度,通过元路径及语义相关节点度量语义相似度,同时将节点影响力权值应用到二者当中. 根据相似度度量方法在聚类中的效果来评估本文提出方法的有效性. 实验结果表明,考虑不同节点在异质网络中的影响力能够提高相似度学习的准确性. 本文提出的方法是可行、有效的.

参考文献

[1] WEI Y, SONG Y Q, ZHEN Y, LIU B, YANG Q. Heterogeneous translated hashing: A scalable solution towards multi-modal similarity search [J]. ACM Transactions on Knowledge Discovery from Data, 2016, 10(4): article no. 36.

[2] NGUYEN P, WANG J, HILARIO M, KALOUSIS A. Learning heterogeneous similarity measures for hybrid recommendations in meta-mining [A]. Proceedings of the 12th IEEE International Conference on Data Mining [C]. Washington, DC, USA: IEEE Computer Society, 2012. 1026 – 1031.

[3] WU J, SHEN H, LI Y D, XIAO Z B, LU M Y, WANG C L. Learning a hybrid similarity measure for image retrieval [J]. Pattern Recognition, 2013, 46(11): 2927 – 2939.

[4] LIN Y K, LIU Z Y, SUN M S, LIU Y, ZHU X. Learning entity and relation embeddings for knowledge graph completion [A]. Proceedings of the 29th AAAI Conference on Artificial Intelligence [C]. Austin, Texas, USA: AAAI Press, 2015. 2181 – 2187.

[5] WANG C G, SONG Y Q, LI H R, ZHANG M, HAN J W. KnowSim: A document similarity measure on structured heterogeneous information networks [A]. Proceedings of IEEE International Conference on Data Mining [C]. Washington, DC, USA: IEEE Computer Society, 2015. 1015 – 1020.

[6] MEHDI O, IBRAHIM H, AFFENDEY L. An approach for instance based schema matching with Google similarity and regular expression [J]. International Arab Journal of Information Technology, 2017, 14(5): 755 – 763.

[7] JUNEJO K N, KARIM A, HASSAN M T, JEON M. Terms-based discriminative information space for robust text classification [J]. Information Sciences, 2016, 372: 518 – 538.

[8] TRIPATHY A, ANAND A, RATH S. K. Document-level sentiment classification using hybrid machine learning approach [J]. Knowledge and Information Systems, 2017, 53(3): 805 – 831.

[9] ZHANG Y L, LI C P, XIE C W, CHEN H. Accuracy estimation of link-based similarity measures and its application [J]. Frontiers of Computer Science, 2016, 10(1): 113 – 123.

[10] HAMEDANI M R, KIM S W. JacSim: An accurate and efficient link-based similarity measure in graphs[J]. Information Sciences, 2017, 414: 203 – 224.
[11] LIU H P, JIN C Q, YANG B, ZHOU A Y. Finding top-k shortest paths with diversity[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(3): 488 – 502.
[12] SUN Y Z, HAN J W, YAN X F, YU P S, WU T Y. PathSim: meta path-based top-k similarity search in heterogeneous information networks[A]. Proceedings of the VLDB Endowment[C]. Seattle, Washington, USA: ACM Press, 2011. 4(11): 992 – 1003.
[13] GAO X J, MU T T, GOULERMAS J Y, WANG M. Topic driven multimodal similarity learning with multi-view voted convolutional features[J]. Pattern Recognition, 2018, 75: 223 – 234.
[14] XIONG Y, ZHU Y Y, YU P S. Top-k similarity join in heterogeneous information networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(6): 1710 – 1723.
[15] ZHANG M X, HU H, HE Z Y, WANG W. Top-k similarity search in heterogeneous information networks with x-star network schema[J]. Expert Systems with Applications, 2015, 42(2): 699 – 712.
[16] SHAMS B, HARATIZADEH S. IteRank: A iterative network-oriented approach to neighbor-based collaborative ranking[J]. Knowledge-based Systems, 2017, 128: 102 – 114.
[17] TAO W B, YU M H, LI G L. Efficient top-k simrank-based similarity join[A]. Proceedings of the VLDB Endowment[C]. Hangzhou, China: ACM Press, 2014. 8(3): 317 – 328.
[18] SHI C, KONG X, HUANG Y, YU P S, WU B. HeteSim: A general framework for relevance measure in heterogeneous information networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(10): 2479 – 2492.
[19] JAIN Y K, BHANDARE S K. Min max normalization based data perturbation method for privacy protection[J]. International Journal of Computer & Communication Technology, 2011, 2(8): 45 – 50.
[20] 刘露. 异质信息网络中离群点检测方法研究[D]. 长春: 吉林大学, 2017.
LIULu. Research on Outlier Detection Methods in Heterogeneous Networks[D]. Changchun: Jilin University, 2017 (in Chinese)
[21] 王岩, 彭涛, 韩佳育, 等. 一种基于密度的分布式聚类方法[J]. 软件学报, 2017, 28(11): 2836 – 2850.
WANGYan, PENG Tao, HAN Jia-yu, et al. Density-based distributed clustering method[J]. Journal of Software, 2017, 28(11): 2836 – 2850. (in Chinese)
[22] HUANG X H, YE Y M, XIONG L Y, WANG S K, YANG X F. Clustering time-stamped data using multiple nonnegative matrices factorization[J]. Knowledge-based Systems, 2016, 114: 88 – 98.
[23] TANG J, ZHANG J, YAO L M, LI J Z, ZHANG L, SU Z. ArnetMiner: Extraction and mining of academic social networks[A]. Proceedings of the 14th ACM SIGKDD International Conference of Knowledge Discovery and Data Mining[C]. Las Vegas, Nevada, USA: ACM Press, 2008. 990 – 998.
[24] MENG X F, SHI C, LI Y T, ZHANG L, WU B. Relevance measure in large-scale heterogeneous networks[A]. Asia-Pacific Web Conference[C]. Germany: Springer, 2014. 636 – 643.

作者简介

刘 露 女, 1989 年生于辽宁大连, 博士、讲师. 主要研究方向为数据挖掘、机器学习、自然语言处理、异质信息网络挖掘.
E-mail: liulu@jlu.edu.cn

胡封晔 男, 1974 年生于河南原阳, 博士、教授. 主要研究方向为信号处理, 无线体域网、认知无线电网络、空时通信和无线定位等.
E-mail: fufy@jlu.edu.cn.

牛 亮 女, 1980 年生于吉林长春, 硕士, 主要研究方向为社会网络.
E-mail: 379879995@qq.com.

彭 涛(通信作者) 男, 1977 年生于吉林松原, 博士、教授. 主要研究方向为数据挖掘、机器学习、信息检索、自然语言处理.
E-mail: tpeng@jlu.edu.cn