第47卷第6期 四川大学学报(工程科学版) Vol. 47 No. 6
2015年11月 JOURNAL OF SICHUAN UNIVERSITY (ENGINEERING SCIENCE EDITION) Nov. 2015

文章编号:1009-3087(2015)06-0131-08 DOI:10.15961/j.jsuese.2015.06.018

基于概念格和随机游走的社交网朋友推荐算法

李宏涛[1,2],何克清[1,2],王 健[1,2],彭珍连[1,2],田 刚[1,2]
(1.武汉大学 软件工程国家重点实验室,湖北 武汉 430072;2.武汉大学 计算机学院,湖北 武汉 430072)

摘 要: 在社交网络朋友推荐上,现有方法通过用户注册的共同属性或者用户共同邻居来对用户进行朋友推荐,由于缺乏对用户之间关系的深入的挖掘,推荐精度不高。采用概念格从数据中挖掘知识,利用用户特征属性和社交网络图建立概念格,提出了弹性随机游走方法SRWR,并在此基础上用概念格知识指导随机游走,提出了融合概念格和随机游走的FCASRWR方法,度量了用户之间的相似性,算法最终根据相似度进行朋友推荐。实验采用Facebook的真实数据集,采用AUC和精确度评价指标,实验结果表明,该方法比目前主流的方法在指标上有较大提高,验证了方法的准确性。

关键词: 社交网络;概念格;随机游走;朋友推荐
中图分类号: TP301 **文献标志码:** A

A Friends Recommendation Algorithm Based on Formal Concept Analysis and Random Walk in Social Network

LI Hongtao[1,2],*HE Keqing*[1,2],*WANG Jian*[1,2],*PENG Zhenglian*[1,2],*TIAN Gang*[1,2]
(1. State Key Lab. of Software Eng., Wuhan Univ., Wuhan 430072, China;
2. Computer School of Wuhan Univ., Wuhan 430072, China)

Abstract: Formal concept analysis was leveraged to acquire knowledge in data. Two concept lattices were built from the user feature attributes and social networking diagram. The random walk method SRWR was proposed and then the FCASRWR method was put forward with the guidance of concept lattice. The FCASRWR method measured the similarity between users, and recommended friends according to the similarity algorithm to users. The Experiments of using Facebook's real datasets showed that the proposed method has a better performance and proved the accuracy of the method.

Key words: social network; formal concept analysis; random walk; friends recommendation

信息推荐技术[1]往往用来处理信息过载问题,社交网络中的朋友关系推荐通常被转化为复杂网络上的链路预测问题[2]。用户的相似性越大,则越可能成为朋友,这里的相似包括兴趣爱好相似、地域相似、情景相似等。社交网络的兴起和发展,如facebook、twitter、linkedin、新浪微博等,加强了用户间的信息分享和交流,推动了人际交往和社区的演变。人与人之间的关系已经由线下的朋友关系逐步过渡到线上的朋友关系。线上的朋友关系往往已经包含了绝大部分的线下的朋友关系,并且还包含兴趣、爱好、情景等相似的而实际上不认识的朋友关系。作为信息传播平台的社交媒体,其中的用户数量近几年急剧增加,拓展朋友关系,为用户推荐合适的朋友已成为社交网络的重要个性化服务。社交网站注册时要求用户填写学校、性别、年龄、职业、兴趣等信息,可以利用这些用户特征信息进行相似度计算,通过相似度进行朋友预测。但是社交网站的注册信息存在不完整和隐私保护的问题,社交网站通过用户的注册的共同属性或者用户的共同邻居来对用户进行朋友推荐,由于缺乏对用户之间关系的深入的挖掘,推荐精度不高。数据中包含了大量的知识,如何从网络大数据中获取有价值的知识,并充分利用获

收稿日期:2015-03-26
基金项目:国家重点基础研究发展计划资助项目(2014CB340401)
作者简介:李宏涛(1978—),男,博士生.研究方向:面向服务的需求工程研究.

取的知识进行深入的计算和分析,已经成为国内外工业界和学术界研究的热点[3]。

基于社交网络结构拓扑信息进行相似度计算的常用方法有共同邻居 CN(common neighbors)[4]、余弦相似性 Salton[5]、雅克比 Jaccard Coefficient[6]、Sorensen[7]、LHN(leicht-holme-newman)[8]、优先链接 PA(preferential attachment)[9]、AA(Adamic-Adar)[10]、资源分配 RA(resource allocation)[11]等。此外,基于随机游走的方法也吸引了朋友推荐研究学者的注意[12-14],特别是带重启的随机游走 RWR(random walk with restart)[15],它假设随机游走粒子在每走一步的时候都以一定的概念返回初始位置,被应用于推荐系统的算法研究中,并取得较好的推荐效果[16]。相比于前面提到的方法,随机游走的方法具有捕捉用户间多侧面关系的优点。以上方法从用户特征或者网络结构特征的角度来度量用户之间的相似性,进行用户推荐。但是他们都没有引入知识的概念,在大数据时代,知识作为数据中的金子,无疑会对提高推荐性能起到重要作用。如果把人的感性知识作为软知识,那么从数据中抽取出来的知识就是硬知识,是与人的主观性无关的客观的知识,这种客观的知识将会对推荐计算带来帮助。

针对社交网络数据,利用用户的注册属性信息和社交图信息,从大量的社交数据中抽取出用户知识,以概念格的形式表达知识,构造了用户属性概念格和用户社交概念格。在带重启的随机游走的基础上提出了弹性重启随机游走方法(SRWR),并在该方法的基础上,融合概念格知识,再次提出了概念格指导下的随机游走方法(FCASRWR),该方法能够有效合理的同时利用用户的属性和用户的社交关系,进行朋友推荐。

2 社交网络中概念格的生成

社交网络用户在注册时会按要求填写出生、教育、家庭、职业、地域、语言等信息,这些信息也叫用户的特征信息,也是用户的属性信息。可以把社交网络中所有的用户所有特征信息集成到一个树型结构中,称为用户信息特征树。图1是特征树的示意图。

特征树中每一个叶子节点就是一个特征,表示社交网络中至少有一个用户具有这样的特征信息。通过特征树,可以把用户的特征信息表示为一个特征向量,向量的个数为特征树中叶子节点的数目,如(1,1,0,1,0,0,1),用户具有的特征在向量中标为1,不具有的特征在向量中标为0. 通过所有用户的

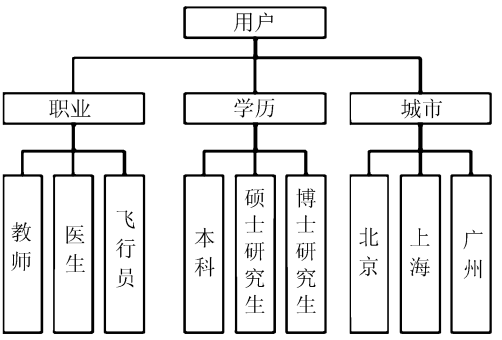

图1 用户信息特征树

Fig. 1 User information feature tree

信息特征向量,可以将社交网络中用户的特征信息转换为概念格中的形式背景。

形式概念分析也叫概念格(FCA),基于哲学中的概念被理解为由外延和内涵所组成的思想单元这一哲学理解,德国数学家 Wille 于1982年首先提出了形式概念分析,用于概念的发现,排序和显示[17]。形式概念分析理论是一种基于概念和概念层次的数学化表达。

定义1[18] 称(U,A,I)为一个形式背景,其中,$U=\{x_1,x_2,\cdots,x_n\}$为对象集,每个$x_i(i\leq n)$称为一个对象;$A=\{a_1,a_2,\cdots,a_m\}$为属性集,每个$a_j(j\leq m)$称为一个属性;I为U和A之间的二元关系,$I\subseteq U\times A$,若$(x,a)\in I$,则称x具有属性a,若$(x,a)\notin I$则称x不具有属性a。

对于形式背景(U,A,I),在对象集$X\subseteq U$和属性集$B\subseteq A$上分别定义运算:

$$X^*=\{a\mid a\in A,\forall x\in X,(x,a)\in I\},$$

$$B^*=\{x\mid x\in U,\forall a\in B,(x,a)\in I\}。$$

式中,X^*为X中所有对象共同具有的属性集合,B^*为具有B中所有属性的对象集合。

定义2[18] 设(U,A,I)为一个形式背景。如果一个二元组(X,B)满足$X^*=B$,且$B^*=X$,则称(X,B)是一个形式概念,简称概念。其中,X称为概念的外延,B称为概念的内涵。

定义3[18] 用$L(U,A,I)$表示形式背景(U,A,I)的全体概念,即概念格,记

$$(X_1,B_1)\leq(X_2,B_2)\Leftrightarrow X_1\subseteq X_2(\Leftrightarrow B_1\subseteq B_2)。$$

则"$\leq$"是$L(U,A,I)$上的偏序关系。其中,(X_1,B_1)叫作(X_2,B_2)的亚概念,(X_2,B_2)叫作(X_1,B_1)的超概念。

社交网络中的用户特征信息可以直接向概念格的形式背景转化。将所有的用户集合作为形式背景的对象集合;将所有的用户特征作为形式背景的属性集;将用户特征向量对应形式背景中的关系。这

样就可以将社交网络中用户的特征信息转换为概念格中的形式背景。用 1 表示$(x,a) \in I$,用 0 表示$(x,a) \notin I$。下面是用户特征信息的形式背景的可视化表格示意图。图 2 中属性列打上“×”标记的表示用户具有这个属性。

	男	女	北京	上海	广州	医生	飞行员	教师
Obj 1	×		×			×		
Obj 2	×			×			×	
Obj 3		×			×			×
Obj 4	×		×			×		
Obj 5								
Obj 6	×		×				×	
Obj 7				×		×		

图 2　用户属性形式背景

Fig. 2　Formal context of user information

社交网络图中有边相连的用户互相为朋友关系,可以将用户的朋友关系作为用户的社交属性,将社交网络图转换为概念格的形式背景。社交网络图可以表示为 $G=(V,E)$,其中,V 为顶点(用户)集合,E 为无向边(朋友关系)集合。仅当两个顶点 v_i,v_j 间的无向边$(v_i,v_j) \in E$ 时,$v_j(v_i)$ 被称为 $v_i(v_j)$ 的邻接。这样社会图能够表示为邻接矩阵 $\boldsymbol{M}=(m_{ij}) \in E$,如果 v_i 和 v_j 为朋友,则 $m_{ij}=1$,否则为 0。对于一个无向图,邻接矩阵是对称的。图 3 是一个局部社交网络图,边代表两个节点之间的朋友关系。

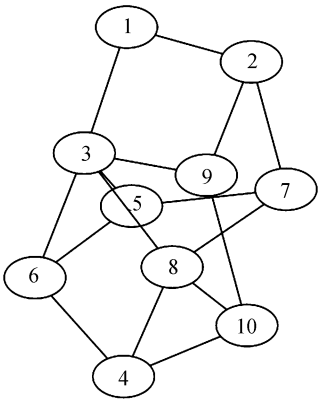

图 3　社交网络图

Fig. 3　Social network graph

对于一个用邻接矩阵 $\boldsymbol{M}=(m_{ij})$ 表示的社交图 $G=(V,E)$,可以通过映射方法将其转换为概念格的形式背景(U,A,I)。

算法 1　社交图生成概念格形式背景算法

输入: 社交图 $G=(V,E)$,邻接矩阵 $\boldsymbol{M}=(m_{ij})$

输出: 概念格形式背景(U,A,I)

```
BEGIN
  U ←∅; A ←∅;
  FOR 每个 v 属于 V
     将 v 添加到 U 中;
     将 v 添加到 A 中;
  ENDFOR
  FOR 每个用户 i,每个用户 j
    IF m_ij = 1  THEN
      (u_i,a_j) ∈ I
    ENDIF
  ENDFOR
END
```

算法中首先将社交图中各顶点(用户)加入到概念格形式背景的 U 和 A 中,如果两个用户为朋友关系,则将两个用户的关系加入到 I 中,遍历所有的用户,即可生成形式背景(U,A,I)。

在生成用户属性形式背景和用户社交形式背景后,可以采用建格算法生成形式背景相对应的概念格,即属性概念格和社交概念格。从形式背景中生成概念格的过程实际上是概念聚类的过程,生成概念格方法可以分为 2 类,批处理算法和增量算法。考虑到社交网中会不断有新用户加入的情况,作者采用增量方法 Godin 算法[9] 来构建概念格。当有新的对象插入后,概念格在原有的基础上进行扩展。格中所有的节点分为 3 类: 不变节点、更新节点和新增节点。不变节点是新格 L' 中所保留的 L 中的节点,这些节点内涵和新对象的内涵没有交集; 更新节点是对原来格 L 中的节点更新后的节点,这些节点的内涵包含在新对象的内涵中,因此只需将其外延更新包括新对象即可; 新增节点是所要插入的节点的内涵与原来格 L 中某个节点的内涵交所产生的集合在格中没有出现过。从社交形式背景生成的概念格如图 4 所示。节点上两排数字分别表示概念的内涵和外延,即概念的属性和包括这些属性的对象。

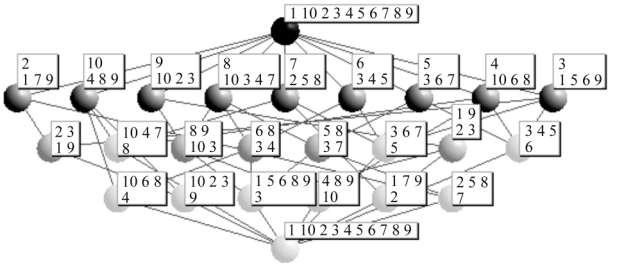

图 4　生成概念格图

Fig. 4　Concept lattice

3　知识指导的随机游走

社交网络通过一系列转换算法可以生成用户属性概念格和用户社交概念格。这 2 个概念格将用来指导随机游走。引入概念格知识后,随机游走者应该沿着更有可能成为朋友的路线进行带权重的随机

游走,用户之间的随机游走权重按照用户之间的概念格相似度进行分配。

3.1 概念格相似度计算

定义4(概念相似度) 概念格 $L(U,A,I)$ 中概念 (X_1,B_1) 和 (X_2,B_2) 的相似度为:

$$SimCon((X_1,B_1),(X_2,B_2)) = \frac{|X_1 \cap X_2|}{\max(|X_1|,|X_2|)} \cdot a + \frac{|B_1 \cap B_2|}{\max(|B_1|,|B_2|)} \cdot b \tag{1}$$

其中,a,b 为参数,$a>0$,$b>0$,且 $a+b=1$,根据概念格的对偶原理[20],概念的对象和属性具有同等的地位,可以取 $a=0.5$,$b=0.5$。

定义5(对象的概念格相似度) 在概念格 $L(U,A,I)$ 中,若对象 x_1 涉及到的概念集合为 C,对象 x_2 涉及到的概念集合为 D,则对象 x_1 和 x_2 的概念格相似度为:

$$SimLa(L,x_1,x_2) = \frac{\sum_{c_i \in C, d_j \in D} Sim(c_i,d_j)}{|C| \cdot |D|} \tag{2}$$

概念集合 C 中的每个概念的外延都包含对象 x_1,概念集合 D 中的每个概念的外延都包含对象 x_2,$|C|$ 为概念集合 C 中概念的个数,$|D|$ 为概念集合 D 中的概念个数。

定义6(社交网中用户的概念格相似度) 社交网 G 中生成的用户特征信息概念格为 $L1(U,A1,I1)$,社交图概念格为 $L2(U,A2,I2)$,则用户 u_1 和 u_2 的概念格相似度为:

$$SimG(u_1,u_2) = \alpha \cdot SimLa(L1,u_1,u_2) + \beta \cdot SimLa(L2,u_1,u_2) \tag{3}$$

其中,α,β 为参数,$\alpha>0$,$\beta>0$,且 $\alpha+\beta=1$,表示两个概念格在相似度计算中的权重。

通过上述一系列定义,可以计算得到用户之间的概念格相似度矩阵。对于新增用户,采用增量计算相似度,只计算在概念格中节点变化的用户,对无变化的用户之间的相似度无需重新计算。

3.2 带重启的随机游走(RWR)

带重启的随机游走(RWR)[15],假设随机游走者在每走一步的时候都可以以一定的概率返回初始位置,设粒子的返回概率为 $1-c$,$\boldsymbol{P}$ 为网络的马尔可夫概率转移矩阵,粒子在 $t+1$ 时刻到达各节点的概率为:

$$\boldsymbol{q}_x(t+1) = (1-c) \cdot \boldsymbol{e}_x + c \cdot \boldsymbol{P}^{\mathrm{T}} \boldsymbol{q}_x(t) \tag{4}$$

不断随机游走,到达稳态后,可以得到 x 到其他顶点的概率,从而得到节点的相似性:

$$S_{xy}^{RWR} = q_{xy} + q_{yx}。$$

其中,$\boldsymbol{q}_x = (1-c)(\boldsymbol{I} - c\boldsymbol{P}^{\mathrm{T}})^{-1}\boldsymbol{e}_x$,$\boldsymbol{e}_x$ 为初始向量。

3.3 弹性重启随机游走(SRWR)

弹性重启随机游走(SRWR)是在带重启随机游走方法(RWR)上扩展而成。在用邻接矩阵 $\boldsymbol{M}$ 描述的社交图 $G(V,E)$ 中,对任意一对顶点 $x,y \in V$,如果 x、y 有边相连,则 $m_{xy}=1$,如果 x、y 没有边相连,则 $m_{xy}=0$。随机游走是一个描述随机游走者访问顶点序列的马尔可夫链[21]。这个过程可以用概率转换矩阵 $\boldsymbol{P}$ 来描述:

$$p_{xy} = m_{xy}/k_x \tag{5}$$

式中,p_{xy} 为随机游走者在 x 顶点处下步到达 y 顶点的概率,$[m_{xy}]$ 为社交图的邻接矩阵 $\boldsymbol{M}$,k_x 为顶点 x 的度。

给定一个从顶点 x 开始的随机游走者,$\boldsymbol{\pi}_x(t)$ 表示经过 t 步后,到达其他各顶点的概率向量。为了使游走者经过 t 步后不至于与出发点的距离越走越远,游走方法是带重启的随机游走,可以类似的想象在游走者身上加上弹簧,走的步数越多,则回到起始点的弹簧拉力越大,游走者重启的概率就越大。SRWR 与 RWR 的区别是二者返回初始点的概率不一样,RWR 为一个固定的返回概率值,而 SRWR 的返回概率是指数可变的。因此,提出的改进的带重启的弹性局部随机游走方法(SRWR)的游走规则是:以 $e^{-\gamma t}$ 的概率向其他节点游走,以 $1-e^{-\gamma t}$ 的概率回到出发点,t 是游走的步数,γ 为弹性拉力系数。这样可以得 SRWR 模型为:

$$\boldsymbol{\pi}_x(t+1) = (1-e^{-\gamma t}) \cdot \boldsymbol{\pi}_x(0) + e^{-\gamma t} \cdot \boldsymbol{P}^{\mathrm{T}} \boldsymbol{\pi}_x(t) \tag{6}$$

其中,$\boldsymbol{\pi}_x(0)$ 为一个 $N \times 1$ 的列向量,第 x 个元素值为 1,其他元素为 0,$\boldsymbol{P}^{\mathrm{T}}$ 为转换矩阵 $\boldsymbol{P}$ 的转置。通过 t 步的随机游走,可以度量出从一个顶点 x 到达另一个顶点 y 的概率 $\pi_{xy}(t)$,以该概率作为顶点之间的相似度可以进行朋友推荐。

3.4 概念格指导的弹性重启随机游走(FCASRWR)

随机游走加入概念格知识后,随机游走的概率转换矩阵 $\boldsymbol{P}$ 调整为带权重的随机游走,权重按照用户的概念格相似度来进行分配。社交图 $G(V,E)$ 中,随机游走者下一步从顶点 x 到顶点 y 的转移概率为:

$$p_{xy} = \begin{cases} 0, y \notin NB; \\ \dfrac{SimG(x,y)}{\sum_{nb_i \in NB} SimG(x,nb_i)}, y \in NB \end{cases} \tag{7}$$

其中,NB 为顶点 x 在图 G 中的所有邻居的集合,

$[p_{xy}]$ 为概率转换矩阵 P。接下来随机访问者游走的方法采取 SRWR 方法。因此 FCASRWR 方法与 SRWR 方法的区别是转换矩阵 P 不一样。

概念格指导下的弹性重启随机游走方法的用户 x 和 y 的最终相似度为:

$$Sim_{xy} = (\pi_{xy}(t) + \pi_{yx}(t))/2 \quad (8)$$

显然,$Sim_{xy} = Sim_{yx}$。社会网分析指出,人群之间交往遵循 6 度分隔和 3 度影响力原则,6 度分隔是指通过 6 次连接可以与任何人建立连接。但能建立连接并不意味着两者有可能成为朋友,而 3 度影响力是指用户所作所说的任何事情,都会在网络上泛起涟漪,影响用户的朋友(1 度),用户朋友的朋友(2 度),甚至用户朋友的朋友的朋友(3 度)。3 度以外,影响力基本消失。因此在影响力的作用下 3 度以内的连接更容易成为朋友,故对用户推荐朋友的范围限制在用户的 3 度朋友范围内。以概念格知识为指导的随机游走推荐算法(FCASRWR)如下:

算法 2　FCASRWR 推荐算法

输入: 社交网 $G(V,E)$,所有顶点的用户信息特征向量

输出: 用户之间相似度,推荐朋友集合

BEGIN

1　将用户信息特征和社交网 $G(V,E)$ 分别生成形式背景 (U,A_1,I_1) 和 (U,A_2,I_2);

2　$(U,A_1,I_1) \rightarrow L1(U,A,I)$;　$(U,A_2,I_2) \rightarrow L2(U,A,I)$;

3　FOR $\forall u_1 \in U, u_2 \in U$:

4　　$SimG(u_1,u_2) = \alpha \cdot SimLa(L1,u_1,u_2) + \beta \cdot SimLa(L2,u_1,u_2)$;

5　ENDFOR

6　计算转换矩阵 $P = (p_{xy})$;

7　FOR　$\forall x \in U, y \in U$:

8　计算 t 步随机游走后用户 x 和 y 的相似度: $S_{xy} = (\pi_{xy}(t) + \pi_{yx}(t))/2$

9　ENDFOR

10　排除用户 x 以及 x 的朋友以及 x 的 3 度朋友以外的用户;

11　按照相似度排序,取 top K 对用户 x 进行朋友推荐;

12　输出用户的相似度矩阵 S 以及推荐的朋友集合。

END

算法首先从社交网络的用户信息和网络拓扑信息建立各自的形式背景,从形式背景中抽取概念格;对任意两格用户,计算用户间的概念格相似度,在此基础上得到随机游走的转换矩阵;经过有限步游走,得到用户间的相似度,并利用该相似度进行排序,对用户进行 top K 朋友推荐。

4　实验分析

4.1　实验数据

实验采用 Facebook 的真实社交数据作为数据集,包含 4 039 个用户,63 个用户属性。在算法比较时还采用了链路预测里面常用的数据集 USAir、PB、NetScience、Yeast。

4.2　评价指标

利用曲线下面积 AUC(area under the receiver operating characteristic curve)、精确度(*Precision*)作为朋友推荐算法的评价指标。AUC 是最常用的一种评价指标,它从整体上衡量算法的精确度。Precision 只考虑排在前 L 位的边是否预测准确。

把数据集分为训练集和测试集。任意两顶点的边如果不在数据集中,那么将该边划入不存在的边的集合中。AUC 的评价算法指标的过程是,每次从测试集和不存在的边集合中各任取一条边,比较这两条边的分数(推荐算法确定的值)的大小,从测试集抽取的边的分数大则得 1 分,等于得 0.5 分,重复进行这样的抽样 n 次,令大于的次数是 n_1,等于的次数是 n_2,则 *AUC* 指标为:

$$AUC = \frac{n_1 + 0.5n_2}{n}。$$

AUC 越大,则预测效果越好,它反映了算法多大程度上好于随机选择。

精确度定义为在前 L 个预测边中预测准确的比例。如果有 m 个预测准确,即根据出现的可能性的值从大到小排序,排在前 L 的边中有 m 在测试集中,那么精确度定义为:

$$Precision = \frac{m}{L}。$$

显然,此值的大小和 L 有关。对于给定的 L,*Precision* 越大预测越准确。

4.3　比较算法介绍

为了评估新提出的朋友推荐算法 FCASRWR 的有效性,将 FCASRWR 和其它一些朋友推荐方法进行比较。这些方法都是用户相似度计算方法,根据相似度的大小对用户进行排序和朋友推荐。欲比较的算法如下:

CN: 共同邻居，对于网络中的节点 v_x，定义其邻居集合为 $\Gamma(x)$，则2个节点 v_x 和 v_y 的相似性就定义为它们共同的邻居数，即

$$S_{xy}^{\mathrm{CN}} = |\Gamma(x) \cap \Gamma(y)|。$$

Salton: 又称余弦相似性，定义如下，其中，k_x 和 k_y 为节点 x 和 y 的度

$$S_{xy}^{\mathrm{Salton}} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\sqrt{k_x k_y}}。$$

Jaccard: 雅克比相似性，定义为:

$$S_{xy}^{\mathrm{Jac}} = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}。$$

AA: 其思想是度小的共同邻居节点的贡献大于度大的共同邻居节点。根据共同邻居节点的度为每个节点赋予一个权重值，为度的对数分之一，其定义为:

$$S_{xy}^{\mathrm{AA}} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\lg k_z}。$$

RA: 资源分配，从网络中的节点 v_x 可以传递一些资源到 v_y，它们的共同邻居就成为传递的媒介，每个媒介都有一个单位的资源并平均的分配到它的邻居，根据收到的资源数定义相似度

$$S_{xy}^{\mathrm{RA}} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z}。$$

LRW: 局部随机游走，粒子 t 时刻从节点 v_x 出发，$\pi_{xy}(t+1)$ 为 $t+1$ 时刻 v_x 到达 v_y 的概率，那么系统演化方程为:

$$\boldsymbol{\pi}_x(t+1) = \boldsymbol{P}^{\mathrm{T}} \boldsymbol{\pi}_x(t)。$$

$\boldsymbol{P}$ 为转移矩阵，设各点的初始资源分布为 $q_x = k_x/M$，M 为网络所有节点度的总和，相似性定义为:

$$S_{xy}^{\mathrm{LRW}} = q_x \cdot \pi_{xy}(t) + q_y \cdot \pi_{yx}(t)。$$

4.4 实验结果

实验将数据集按边随机划分为 90% 训练，10% 测试，然后进行各算法的计算，重复 100 次，取这 100 次结果的统计值作为最终的比较结果。

实验 1: 验证提出的弹性重启的随机游走方法 SRWR 准确性。

实验先不引入概念格相似度权重，直接验证提出的随机游走的方法，采用 Facebook 数据集以及链路预测中常用的数据集 USAir、PB、NetScience、Yeast。在对各算法独立做完 100 次实验后，取平均值进行比较。

在参数选择上，SRWR 算法选取弹性拉力系数 $\gamma = 0.1$，游走步数 $l = 12$。

各算法的 AUC 结果如表 1 所示。

表 1 各算法预测精确度 AUC 指标

Tab. 1 Accuracy comparisonof algorithms

方法	Facebook	NS	PB	Yeast	USAir
CN	0.914 0	0.970 6	0.923 3	0.915 4	0.953 6
Salton	0.897 7	0.872 3	0.877 6	0.914 3	0.924 6
Jaccard	0.875 7	0.977 6	0.876 2	0.914 4	0.914 2
AA	0.926 9	0.983 2	0.926 1	0.915 8	0.965 9
RA	0.937 4	0.983 7	0.928 1	0.916 0	0.972 2
LRW	0.928 7	0.987 3	0.945 6	0.966 4	0.956 3
RWR	0.890 8	0.982 3	0.928 9	0.978 8	0.952 2
SRWR	0.938 4	0.988 9	0.956 2	0.978 3	0.973 4

从实验结果可以看出，在所采用的 5 组数据集中，提出的 SRWR 弹性重启动随机游走方法在其中的 4 组数据集中结果都是最好的，而在 Yeast 数据集中结果也和最好的结果相差无几，可以看出 SRWR 方法的有效性。

在社交网 Facebook 数据集中，各算法的 AUC 值箱线图如图 5 所示。

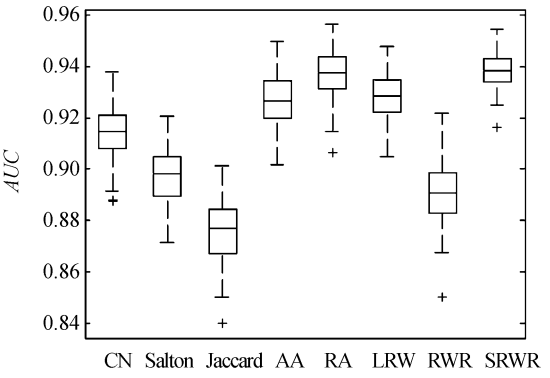

图 5 算法的 AUC 值箱线图

Fig. 5 Boxplot for AUC values

箱线图利用数据中的 5 个特征值: 最小值、第一四分位点、中值、第三四分位点、最大值来描述数据，能体现数据的统计特征，从图形中可以看出 SRWR 方法的表现是最好的。

实验 2: 验证提出的融合概念格和弹性重启随机游走的 FCASRWR 推荐方法的准确性。

由于链路预测中常用的数据集没有节点属性特征，概念格就用不上，因此该实验只能用 Facebook 的数据集。实验将 Facebook 数据集随机分为 6 个组 FB1 到 FB6，在每个组独立验证加入概念格后算法的性能，采用数据 75% 训练，数据 25% 测试。在参数选择上，算法选取弹性拉力系数 $\gamma = 0.1$，游走步数 $l = 12$，用户特征信息概念格和社交图概念格的权重系数分别取 $\alpha = 0.6$，$\beta = 0.4$。对每组数据中

的每个用户,按照算法的用户间的相似度排序,选取相似度靠前的 3 个且不构成朋友关系的用户进行朋友推荐,采取精确度 Precision 作为衡量指标。FCASRWR 和 SRWR 算法的精度 Precision 如图 6 所示。

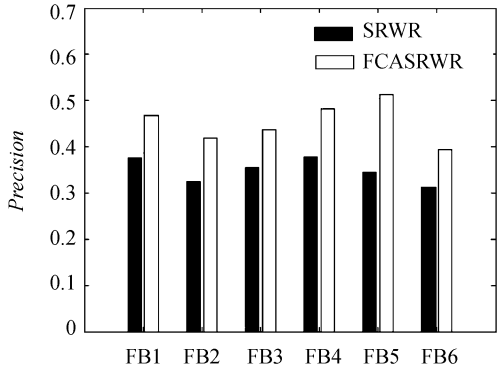

图 6　算法加入概念格后的准确率比较

Fig. 6　Comparison of the accuracy of the algorithm

从图 6 中可以看出,将弹性重启动随机游走方法加入概念格后,准确率均有较大提高,说明了 FCASRWR 算法的有效性。在参数上,用户属性概念格和社交概念格的权重系数分别取 α、β 对准确率有一定影响,由于 $\alpha+\beta=1$,因此只需考虑 α 对准确率的影响,图 7 显示了权重系数 α 与对应的准确率的关系。随着用户特征信息概念格的权重系数的增加,准确率也在慢慢增加,当 $\alpha=0.6$ 时准确率达到最高点,然后随着权重系数的增加,准确率逐步降低,这体现了用户特征信息概念格和社交图概念格在概念格相似度计算中所占的比重。

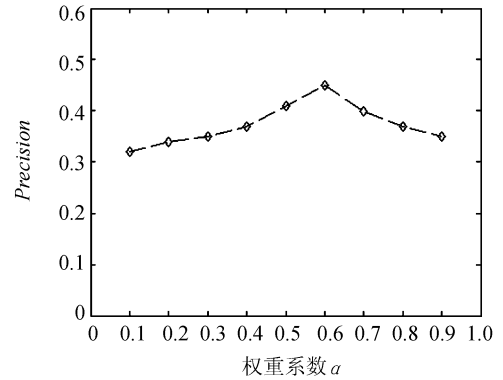

图 7　权重系数 α 对准确率的影响

Fig. 7　Fluency of weighting coefficients α on accuracy

5　结束语

利用社交网络的顶点特征属性和社交图信息,通过映射,分别构建了用户特征信息概念格和社交图概念格,形成社交网络的概念格知识。在传统的带重启随机游走方法的基础上,提出了弹性重启随机游走 SRWR 方法,并融合概念格知识,提出了基于概念格和随机游走的 FCASRWR 朋友推荐算法。实验采用真实的数据集,实验结果表明方法的性能比主流的朋友推荐算法有提高,不足之处是算法还是应用在离线数据的环境中,对在线环境中考虑不周全,特别是新增用户出现后,只能在概念格建格和概念格相似度计算方面采用增量算法,未来将在随机游走的增量计算上进行深入研究。

参考文献:

[1] Wang Yuanzhuo, Jiayan Tao, Liu Dawei, et al. Information retrieval and data mining based on open network knowledge [J]. Journal of Computer Research and Development, 2014, 52(2): 456 – 474. [王元卓,贾岩涛,刘大伟,等. 基于开放网络知识的信息检索与数据挖掘 [J]. 计算机研究与发展,2014,52(2):456 – 474.]

[2] Dimicco J, Millen D R, Geyer W, et al. Motivations for social networking at work [C] //Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work. New York: ACM, 2008: 711 – 720.

[3] Guy I, Ronen I, Wilcox E. Do you know? Recommending people to invite into your social network [C] //Proceedings of the 14th International Conference on Intelligent User Interfaces. New York: ACM, 2009: 77 – 86.

[4] NewManme J. Clustering and preferential attachment in growing networks [J]. Physical Review E, 2001, 64(2): 025102.

[5] Carmi S, Havlin S, Kirkpatrick S, et al. A model of Internet topology using k-shell decomposition [J]. Proceedings of the National Academy of Sciences, 2007, 104(27): 11150 – 11154.

[6] Murata T, Moriyasu S. Link prediction of social networks based on weighted proximity measures [C] //IEEE/WIC/ACM International Conference on Web Intelligence. Silicon Valley, California, USA: IEEE Computer Society, 2007: 85 – 88.

[7] Srensen T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the begetation on Danish commons [J]. Biologiske Skrifter, 1948, 5(4): 1 – 34.

[8] Leicht E A, Holme P, Newman M E J. Vertex similarity in networks [J]. Physical Review E, 2006, 73(2): 026120.

[9] Chowdhury G. Introduction to modern information retrieval [M]. London: Facet Publishing, 2010.

[10] Adamic L A, Adar E. Friends and neighbors on the web [J]. Social Networks, 2003, 25(3): 211 – 230.

[11] Lu L, Zhou T. Link prediction in complex networks: A survey [J]. Physica A: Statistical Mechanics and its Applications, 2011, 390(6): 1150 – 1170.

[12] Yin Z, Gupta M, Weninger T, et al. A unified framework for link recommendation using random walks [C]//2010 International Conference on Advances in Social Networks Analysis and Mining (ASONAM). Washington DC, USA: IEEE, 2010: 152 – 159.

[13] Backstrom L, Leskovec J. Supervised random walks predicting and recommending links in social networks [C]//Proceedings of the 4th ACM International Conference on Web Search and Data Mining. New York: ACM, 2011: 635 – 644.

[14] Xia J, Caragea D, Hsu W H. Bi-relational network analysis using a fast random walk with restart [C]//Proceedings of the 9th IEEE International Conference on Data Mining. Washington D C, USA: IEEE, 2009: 1052 – 1057.

[15] Tong H, Faloutsos C, Pan J Y. Fast random walk with restart and its applications [C]// International Conference on Data Mining. Washington DC, USA: IEEE, 2006: 613 – 622.

[16] Pongnumkul S, Motohashi K. Random walk-based recommendation with restart using social information and bayesian transition matrices [J]. International Journal of Computer Applications, 2015, 114(9): 32 – 38.

[17] Wille R. Restructuring lattice theory: An approach based on hierarchies of concepts [M]. Berlin: Springer, 2009.

[18] Ganter B, Wille R. Formal concept analysis: Mathematical foundations [M]. Berlin: Springer Science & Business Media, 2012.

[19] Godin R. Incremental concept formation algorithm based on Galois (concept) lattices [J]. Computational Intelligence, 1995, 11(2): 246 – 267.

[20] Liu W, Lu L. Link prediction based on local random walk [J]. Europhysics Letters, 2010, 89(5): 58007.

[21] Vincent M, Liu J, Liu C. Strong functional dependencies and a redundancy free normal form for xml [D]. Orlando, FL, USA: Int Inst Informatics and Systemics, 2003.

(编辑 张 琼)