软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2012,23(Suppl.(2)):176−185

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

动态多维网络双向链路预测*

王 红[1+], 于晓梅[1], 孙彦燊[2]

[1](山东师范大学 信息科学与工程学院,山东 济南 250014)

[2](浙江大学 理学部,浙江 杭州 310027)

Bi-Direction Link Prediction in Dynamic Multi-Dimension Networks

WANG Hong[1+], YU Xiao-Mei[1], SUN Yan-Shen[2]

[1](School of Information Science and Engineering, Shandong Normal University, Ji'nan 250014, China)

[2](Faculty of Science, Zhejiang University, Hangzhou 310027, China)

+ Corresponding author: E-mail: wanghong106@163.com

Wang H, Yu XM, Sun YS. Bi-Direction link prediction in dynamic multi-dimension networks. ***Journal of Software*, 2012,23(Suppl.(2)):176−185 (in Chinese)**. http://www.jos.org.cn/1000-9825/12037.htm

Abstract: Recently, many researchers have been attracted in link prediction, which is an effective technique \ used in graph based models analysis. By using the link prediction method the study understands associations between nodes. Most of previous works in this area have not explored the prediction of links in dynamic multi-dimension networks and have not explored the prediction of links which could disappear in the future. This paper argues that these kinds of links are important. At least they can serve as a complement for current link prediction processes in order to plan better for the future. This paper proposes a link prediction model, which is capable of predicting bi-direction links that might exist and may disappear in the future in dynamic multi-dimension networks. Firstly, the study presents the definition of multi-dimensional networks, reduction dimension networks, and dynamic networks. Then paper proposes a forward some algorithms which build multi-dimension networks, reduction dimension networks, and dynamic networks. Next, a give bi-direction link prediction algorithms in dynamic multi-dimension weighted networks. At the end, algorithms above are applied in recommendation networks. Experimental results show that the algorithm can improve the link prediction performance in dynamic multi-dimensional weighted networks.

Key words: dynamic network; multi-dimension network; bi-direction link prediction; weight similarity; personalized recommendation

摘 要: 链路预测作为网络分析的有效工具得到许多研究者的关注.链路预测可以使人们更好地了解网络节点之间的内在联系.目前的网络链路预测方法大多是根据已知的网络节点以及网络结构等信息预测网络中尚未产生连边的两个节点之间产生链接的可能性,而且大多是在单关系或静态网络中进行.它们没有综合考虑多维关系动态网络中的链路预测,也忽略了未来将会消失的链接.这些链接的预测非常重要,至少可以作为现有链路预测的必要补充,使人们更准确地预测未来.提出了动态多维网络双向链路预测方法,在动态多维网络中既可以预测将来可能产生的链接,也可以预测现有的而将来可能消失的链接.首先给出多维网络、降维网络和动态网络的定

* 基金项目: 国家自然科学基金(60975081); 山东省科技计划(2012GGB01058); 山东省研究生科技创新计划(SDYY10059)
收稿时间: 2012-05-20; 定稿时间: 2012-09-29

义,然后提出构建多维网络、对多维网络降维以及构建动态网络的算法,再后给出一种动态多维加权网络中双向链路预测算法.实验结果表明,算法能够使多维加权网络中链路预测有更好的效果.

关键词: 动态网络;多维网络;双向链路预测;权重相似度;个性化推荐

随着网络的快速发展,网络的规模越来越大、越来越复杂.如何在已有的网络关系中寻找未知的关系,得到人们想去了解和认知的未来的东西,这方面的研究受到越来越多学者的关注.网络中的链路预测正是解决这方面问题的方法之一.目前大多数的网络链路预测是通过已知的网络节点以及网络结构等信息预测网络中尚未产生链接的两个节点之间产生链接的可能性[1].这种方法包含了对未知链接和未来链接两个方面的预测. 而本文提出的网络链路预测更具有普遍意义,我们称为双向链路预测,即在预测未来可能出现的链接的同时,也预测现在已经出现的链接在未来将消失的可能性.因为在现实世界中这两种可能性同时存在.比如一个现有的企业将来既可能成长也可能衰败,双向链路预测一方面能使企业扩展生产逐步壮大,另一方面也能使企业未雨绸缪,减少亏损,避免失败.因此,这两种链路预测同等重要.

自然界中存在的大量现实系统都可以用复杂网络加以描述[2].如,文献[2]作者及其发表的论文关系形成的网络就是一个典型的复杂网络,其中存在着大量非线性、自组织现象都是复杂网络具有的特征.可以通过链路预测的方法和思想来研究作者所处的错综复杂的社会关系,来研究不同的作者之间合作发表文章的可能性[3].

现实生活中的复杂网络很多时候是一个多维的网络.在多维网络中,对象间的关系是多元的.如,人们在Internet 上的行为一般涉及到多个网络系统,例如影视网络、读书网络等.如果把每类网络看作一个子网络系统,这些子网络的集合就构成了多维网络.同时,多维网络又是动态的,随着时间的推移,网络不断地演化,或成长、或衰退、或变化.网络中出现新的连边表明节点间有新的关系出现[4],网络中已有连边的消失表明节点间不再存在相互关系.若能准确预测两个节点间出现新边、两个节点间连边消失,以及两个节点间消失的连边再重新出现等情况的可能性和时间,对未来的规划无疑十分重要.因此,应该考虑时间因素对多维网络链路预测的影响.

本文提出一种适合在动态多维网络中的链路预测方法.首先建立多维网络模型和降维网络模型,然后提出一种动态多维加权网络中的双向链路预测算法.算法中考虑时间因素和权重因素对于链路预测的影响.本文以个性化推荐网络为例,对用户和产品进行向量空间建模,然后对网络降维,同时考虑时间因素,最后通过相似度比较获得节点间最可能的链接和未来有可能消失的链接.

1 相关工作

最近几年,复杂网络的链路预测问题受到来自不同领域、拥有不同背景的科学家的广泛关注.文献[5]提出了一种利用网络的层次结构进行链路预测的方法.文献[4]基于网络拓扑结构的相似性,分析了若干指标对社会合作网络中链路预测的效果.文献[6]利用随机分块模型预测网络缺失边和错误边.文献[7]利用矩阵和向量的方法对静态网络进行链路预测.文献[8]利用相似性,说明了静态和动态这两种网络链路预测的实现过程.文献[9]采取了一个内部链接和加权映射的方法,预测二分社会网络中未来可能出现的链接.文献[1,10]通过社会网络分析方法对动态社会网络链路预测,证明了链路预测成为准确分析社会网络结构的有力的辅助工具.文献[11]把基于节点(包括性别、年龄等)相似性进行链路预测的方法应用到治疗肿瘤的生物制剂研究中.文献[12]提出了网络医学的概念.该文指出各种疾病的起源和发展并不是孤立的,而是相互联系的,可以使用复杂网络的理论来研究疾病起源构成的网络.文献[13,14]提出了利用二部分图建立用户-产品关联关系,根据用户的矢量模型和权重信息对资源分配矢量模型进行定义和更新,将产品推荐给用户.

总之,这些研究从不同角度、对于网络结构的不同特点进行刻画,得到较好的预测效果,成功地应用到很多领域中.但是,也存在一些缺点,比如:单一考虑二维关系网络或动态网络,而没有综合考虑多维关系动态网络中的链路预测;大多考虑通过加入新连边来预测未来将会产生的链接而忽略了未来将会消失的链接.另外,链路预测的领用领域还可以进一步扩展.因此,本文提出一种动态多维网络中双向链路预测方法.并将其应用到个性化推荐关系网络中.

2 动态多维网络

2.1 相关概念

首先给出相关的一些概念的定义,为进行链路预测提供此研究基础.

定义 1(多维网络):设有一个 n 类节点的集合 $V=\{V_1,V_2,\ldots,V_n\}$,其中 $V_i(1\leqslant i\leqslant n)$ 是一类节点集合.$\forall v_i\in V_i,v_j\in V_j(j=1,2,\ldots,i-1,i+1,\ldots,n)$,设无序偶对 (v_i,v_j) 表示 v_i 与 v_j 之间的连边,$W=\{(v_i,v_j)|v_i\in V_i,v_j\in V_j,j=1,2,\ldots,i-1,i+1,\ldots,n\}$ 为 V_i 中节点 v_i 与 $V_j(1\leqslant j\leqslant n,j\neq i)$ 中节点所有可能连边的集合,则以 V 为节点集合,W 的某个子集为连边集合的网络称为多维网络,表示为 MD.

定义 2(单维网络). 在定义 1 的多维网络 MD 中,考察某个特定的 $V_j(1\leqslant j\leqslant n,j\neq i)$ 设为 V_k,设 $W_k=\{(v_i,v_k):v_i\in V_i,v_k\in V_k\}$ 表示 V_i 中节点与 V_k 中节点所有可能的连边集合.则以 V_i 与 V_k 为节点集合,以 V_k 为连边集合的网络,称为 MD 网络的单维网络,MD 中单维网络的总个数就是 MD 的总维数,即网络的维.

定义 3(降维网络). 在定义 1 的多维网络 MD 中,对于 $\forall p,q\in V_i$,设 $V_j(1\leqslant j\leqslant n,j\neq i)$ 中与 p,q 有连边的节点集合分别为 V_{jp} 和 $V_{jq}(V_{jp}\subseteq V_j,V_{jq}\subseteq V_j)$,则定义 p,q 之间有连边的条件是:当且仅当 $V_{jp}\cap V_{jq}\neq\varnothing$.称以 V_i 为节点集合,按以上连边关系组成的网络称为 MD 的降维网络,表示为 SD.

定义 4(动态网络). 在定义 1 的多维网络 MD 中,考虑时间的影响.设在 T_i 时刻的 MD 表示为 MD_i,T_{i+1} 时刻的 MD 表示为 MD_{i+1},$i=1,2,\ldots$ 依此类推,则 $S=\{(T_i,MD_i)\mid i=1,2,...\}$ 称为一个动态网络,表示为 DD.

图 1 展示了一个二维网络在两个时间点的状态及其投影网络.在这个二维网络中,把其中一类节点定义为对象节点,另一类定义为主体节点.当主体使用对象时,主体节点和对象节点之间就有连边.如果能通过主体与对象的关系获得主体之间的关系,即得到二维网络的降维网络.

图 1 中节点 A 到节点 E 是主体节点,节点 1 到节点 6 是对象节点,T1 时刻主体与对象的关系网络及其降维网络如图 1(a)所示,T2 时刻主体与对象的关系网络及其降维网络如图 1(b)所示.T2 时刻网络状态与 T1 时刻网络状态相比,两个主体节点 B、D 和两个对象节点 5、6 离开,而主体节点 E 和对象节点 8 加入,同时边也发生变化.

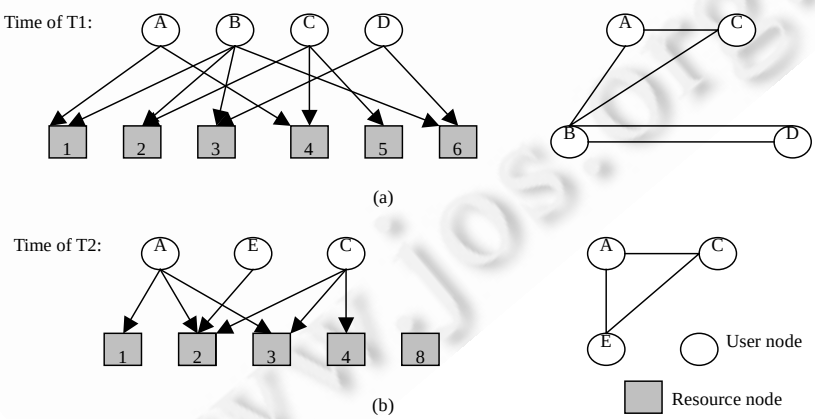

Fig.1 Dynamic two-dimension weighted networks and its reduction dimension networks

图 1 动态二维网络及其降维网络在 T1,T2 两个时刻状态

2.2 构建多维网络

从上节定义可知,多维网络是多个单维网络的有机集合.一个单维网络可以表示为一个二分图 $G=(U,V_i,E_i)$,$i=1,2,3,\ldots,n$,其中 n 是多维网络的维数,U 是多维网络中主体节点的集合,V_i 是第 i 个单维网络中对象节点的集合,E_i 是第 i 个单维网络中边的集合,边上的权重代表关系的强度,为了处理方便,可以归一化使得最大权重为 1.单维网络模型的构建方法是:如果某个主体节点使用了某个对象节点,即这一对节点之间有联系,则在它们之

间添加一条连边,并在边上标出权重.因此多维网络模型的构建算法可以表示如下.

算法 1. 多维网络模型构建算法.

输入:数据集.

输出:多维加权网络模型.

```
MultiDimentionModel()
{
    forEach(node1 in U)
    {
        for(int i=1;i≤n;i++)
        {
            forEach(node2 in Vi)
            {
                if(node1 uses node2)
                {
                    add an edge between node1 and node2 labeled with a weight value;
                    normalize the weight value;
                }
            }
        }
    }
}
```

2.3 多维网络降维

上一节定义的多维网络表示了主体节点参与的多个关系.要预测出主体节点之间未来可能发生的新联系或预测出主体节点之间现有的但将来可能消失的联系,发现隐藏其中的重要关系,则需要构造主体节点关系网络,因此多个单维网络都需要做投影,最终获得主体节点构成的降维网络.但是在降维过程中不能将这些单维网络进行简单的一种叠加,这样必定造成重要信息的流失,这是因为每个单维网络的重要性是不一样的,而且单维网络之间不是独立的,而是相互作用相互影响的.这一点对于多维网络链路预测结果有着重要意义.

我们将多维网络降维分为 3 步进行.

第 1 步,每个单维网络向主体节点集合做投影,得到 n(n 是多维网络的维数)个投影网络;

第 2 步,对这 n 个投影网络做相关性分析,剔除原始关系集合中的冗余信息,挑选最有效的关系来进行分析;

第 3 步,用第 2 步得到的相关性最小的 L 个关系来表示多维网络,得到主体节点构成的降维网络.

第 1 步单维网络向主体节点集合做投影的算法表示如下.

算法 2. 多维网络投影网络算法.

输入:多维网络 $G=\{G_1,G_2,\ldots,G_n\}$(n 是多维网络的维数),其中,$G_i=(U,V_i,E_i)$是单维网络(i=1,2,3,…,n);

输出:n 个单维网络投影网络的集合.

```
NetProjection()
{
    forEach (node1 in U)
    {
        forEach(node2 in U)
        {
            forEach(edge in Ei)
            {
                if(there are N1 neighbour nodes between node1and node2)
                {
                    add an edge between node1 and node2 labeled with N1;
                }
            }
        }
    }
}
```

为了描述方便,将算法 2 得到的这 n 个投影网络的集合表示为 $P=\{P_1,P_2,\ldots,P_n\}$,其中 $P_i=(U,E_i)$是投影网络(i=1,2,3,…,n).令 C_i 是 P_i 对应的权值矩阵,它是一个 $k\times k$ 的方阵,k 是主体节点集合 U 中所含节点的个数.对 C_i 作归一化处理,使矩阵 C_i 的元素值均在[0,1]之间.又因为 C_i 是一个对称矩阵,所以 C_i 可以用一个 $k(k–1)/2$ 维的向量 V_i 来表示.

下面对这 n 个投影网络特征做相关性分析,剔除原始关系集合中的冗余信息.选择出的特征间相关性越少,就越可能使用尽尽量少的特征来获得较好的分类效果.

文献[15]中给出 N 个向量中任意两个的相关系数公式为

$$\rho_{xy}=\frac{\frac{1}{N}\sum_{i=1}^{n}(x_i-u_x)(y_i-u_y)}{\sqrt{\frac{1}{N}\sum_{i=1}^{n}(x_i-u_x)^2}\sqrt{\frac{1}{N}\sum_{i=1}^{n}(y_i-u_y)^2}} \tag{1}$$

其中,x_i,y_i 分别为 x,y 向量的第 i 个特征分量;u_x,u_y 分别为 x,y 向量的特征均值.相关系数矩阵中中相关系数值为 1 或者接近 1 的向量完全相关或者近似完全相关.在原始向量集中增加或者删除特征相关的向量,不会影响原始向量集的分类能力[16].在实际情况中,原始向量集中完全相关或者近似相关的情况并不多,大多数情况下向量间只存在着一定的相关性,那么应该尽量选择相关系数小的向量来表示原始网络.因此多维网络降维过程的第 2 步,基于相关性分析的关系选择算法的基本流程见算法 3.

算法 3. 基于相关性分析的关系选择算法.

输入:n 个单维网络投影网络的权值矩阵所对应的关系特征向量$\{V_1,V_2,\ldots,V_n\}$;

输出:L 个相关系数最小的关系特征向量的集合.

```
SelectRelation()
{
    for(int i=1;i≤n;i++ )
    {
        计算 Vi 的特征均值 ui
    }
    从 n 个关系特征向量中随机抽取 k 个,得到 n!/(k!(n−k)!) 个向量集合;
    对每个向量集合中的的任意 2 个向量按照公式 1 计算相关系数;
    for(int x=1;x≤n;x++)
    {
        求 Σ_{y=1}^{n} |ρxy|
    }
    将 Σ_{y=1}^{n} |ρxy| 从小到大排序(x=1,2,…,n);
    取排序中前 L 个元素对应的关系向量构成的集合;
}
```

在算法 3 执行后,需要将 L 个关系进行组合,最终形成降维网络.因为在实际网络中,不同关系在多维关系网络产生的作用并不相同,因此考虑权重因素,对不同的关系设置相应的权重,这样能更好的体现出主体节点之间的关系.多维网络降维的第 3 步描述如下.

算法 4. 降维网络构成算法.

输入:L 个相关系数最小的关系特征向量的集合;

输出:降维网络 SD.

```
DimensionReduction()
{
    获得由 L 个相关系数最小的关系特征向量集合对应的投影网络集合 P={P1,P2,…,PL};
```

确定由 L 个[0,1]之间的数构成的权重集合$\Lambda=\{\alpha_1,\alpha_2,\alpha_3,\ldots,\alpha_L\}$;
forEach(x,y in U)
{
initial the weight of the edge between x and y called W_{xy};
for(int i=1;$i\leqslant L$;i++)
{
$W_{xy}=W_{xy}+W[P_i]_{xy}\times\alpha_i$;
}
SD is a weighted network with nodes in U;
}

2.4 多维网络动态模型

在多维网络中,主体节点间的关系不仅是多元的,而且是动态变化的.如新节点的加入、旧节点的退出,以及节点间关系的变化,这些变化会给网络的链路预测带来新的影响,在构造多维网络的过程中应该充分考虑这样的影响.因此,在多维网络中再考虑动态变化因素.如果用 S_i 代表多维网络在时间片 t_i 的网络图,其中包括节点集 V_i、边集 E_i 和边权重的集合 W_i,则多维网络在 t_i 时刻的状态可以表示为一个二元组$(t_i,S_i(V_i,E_i,W_i))$.因此,动态多维网络可以表示为二元组序列 $G=\{(t_1,S_1(V_1,E_1,W_1)),(t_2,S_2(V_2,E_2,W_2)),\ldots,(t_i,S_i(V_i,E_i,W_i)),\ldots\}$.这样可以描述多维网络状态的变化,有效描述多维网络的动态特性.下面给出构建动态网络的算法.

算法 5. 动态网络构建算法.

输入:多维网络和时间序列 $T=\{t_1,t_2,\ldots,t_i,\ldots,t_n\}$,其中,$t_1<t_2<\ldots t_i<\ldots<t_n$;

输出:动态网络.

DynimicReductionNetwork()
{
forEach (x in T)
{
MultiDimentionModel();
NetProjection();
SelectRelation();
DimensionReduction();
}
Get $G=\{(t_1,S_1(V_1,E_1,W_1)),(t_2,S_2(V_2,E_2,W_2)),\ldots,(t_n,S_n(V_n,E_n,W_n))\}$
}

在动态多维网络链路分析中需要体现两个特点:

1) 在同一个时刻,各维网络状态对整个网络链路预测的重要性不同.如,参加朋友酒会与给病人看病对反映医生群体的工作关系有不同程度的重要性.

2) 网络动态变化情形下,时间较远的网络状态对反映当前网络状态或预测以后可能网络状态的贡献较小,而时间较近的网络状态对此的贡献较大.如,两位作者早期有合作,共同发表过论文,但近年却一直没有合作,我们将考虑他们今后发表论文的可能性会降低;但如果两位作者近期也有合作,则显然他们再次合作的可能性会增大.

对于第 1 个特点,在多维关系网络降维过程(见算法 4)中,已经通过边的权重体现出来,对多维关系网络中不同关系的连边设置了不同的权重,这样能在降维网络中更好地体现出节点之间的关系亲疏程度.

为了体现对于第 2 个特点,在动态模型中需要再次对边的权重进行修正,即此时算法 5 中的 W_i 应该是算法 4 中 W_{xy} 在不同时间段的组合.因此,将不同时刻的节点权值赋予不同的因子,早期的节点权值赋予较低的因子,而近期节点的权值赋予较高的因子.比如,将动态网络按照时间先后顺序分成 3 个时间段:$T_1,T_2,T_3,(T_1<T_2<T_3)$,设某对节点$(x,y)$在这给 3 个时间段的权值分别为 W_1,W_2,W_3,给这 3 个权值赋予不同的因子$\alpha_1,\alpha_2,\alpha_3,(\alpha_1<\alpha_2<\alpha_3)$,则

该对节点最终的权值 $W=W_1\times\alpha_1+W_2\times\alpha_2+W_3\times\alpha_3$.

3 动态多维加权网络双向链路预测

本节对该模型进行双向链路预测,即在预测未来可能出现的链接的同时,也预测现在已经出现的链接在未来将消失的可能性.本节提出基于相似性指标的链路预测算法.

3.1 修正相似度指标

相似性度量是链路预测算法中的重要指标.在计算主体节点之间相似性时,我们使用修正的余弦相似性(adjusted cosine)计算方法[17].对于主体节点集 $U=\{u_1,u_2,...,u_n\}$,节点间的相似度公式为

$$sim(i,j)=\left[\sum_{k\in I_{ij}}(R_{i,k}-\overline{R_i})(R_{j,k}-\overline{R_j})\right]\Big/\left[\sqrt{\sum_{k\in I_i}(R_{i,k}-\overline{R_i})^2}\sqrt{\sum_{k\in I_j}(R_{j,k}-\overline{R_j})^2}\right] \quad (2)$$

其中,I_{ij} 表示主体 i 和主体 j 共同的邻居节点集合,I_i 和 I_j 分别表示主体 i 和主体 j 邻居节点集合,这里节点的邻居节点定义为与该节点直接相连的节点.R_{ij} 是节点 i 和节点 j 之间边上的权重.

本文采用的共同邻居相似性指标直观易懂、容易进行数据处理.但是,本文需要对相似性指标作调整,以适合动态多维网络模型.

本文考虑体现在相似度指标中.将不同时刻的节点相似度赋予不同的权重,早期的节点相似度赋予较低的权值,而近期的节点相似度赋予较高的权值.比如,将动态网络按照时间先后顺序分成 3 个时间段:$T_1,T_2,T_3,(T_1<T_2<T_3)$,设某节点在这给 3 个时间段的相似度分别为 S_1,S_2,S_3,给这 3 个相似度赋予不同的权值α_1,$\alpha_2,\alpha_3(\alpha_1<\alpha_2<\alpha_3)$,则该节点最终的相似度 $S=S_1\times\alpha_1+S_2\times\alpha_2+S_3\times\alpha_3$.此时的节点相似度公式调整为

$$sim(i,j)=\sum_{k=1}^{n}\alpha_k\times sim(i,j)_k,(i,j)\in E_k \quad (3)$$

3.2 动态多维网络正向链路预测算法

相似度指标调整后,就可以利用它进行动态多维网络正向链路预测算法,见算法 6.

算法 6. 动态多维网络正向链路预测算法.

输入:多维网络和时间序列 $T=\{t_1,t_2,\ldots,t_n,\ldots\}$,其中,$t_1<t_2<\ldots t_n<\ldots$;

输出:主体节点未来可能发生的链接序列.

```
LinkPrediction()
{
    MultiDimentionModel();
    NetProjection();
    SelectRelation();
    DimensionReduction();
    forEach (int x=1; x<n; x++)
    {
        Compute similarity for each pair of nodes according to formula 3;
    }
    Put these similarity numbers in a list and sort them descending;
    Get the first L … in the list;
}
```

3.3 动态多维网络反向链路预测算法

进行反向链路预测的目的是得到现有的链接中哪些将来有可能消失.其思想是将降维网络中的权值改变,即原来权值大的现在变成权值小的,反之亦然.在此基础上进行正向链路预测,这样预测到的结果就是未来最可

能消失的链接,见算法 7.

算法 7. 动态多维网络反向链路预测算法.

输入:多维网络和时间序列 $T=\{t_1,t_2,\ldots,t_n,\ldots\}$,其中 $t_1<t_2<\ldots t_n<\ldots$;

输出:主体节点未来可能消失的链接序列.

```
ReverseLinkPrediction()
{
    MultiDimentionModel();
    NetProjection();
    SelectRelation();
    DimensionReduction();
    Get reduction demention networks called SD=(U,E,W);
    forEach (w_xy∈W)
    {
        w_xy=1−w_xy;
    }
    LinkPrediction();
    Put these similarity numbers in a list and sort them descending;
    Get the first L … in the list;
}
```

4 实验结果及分析

4.1 实验数据集

实验数据来自于某综合视频网站最近 3 个月的后台访问日志记录,网站提供的资源很多,其中有电影、电视剧、音乐,本文提出的推荐算法将会在这 3 个维度的网络中进行.

经过数据预处理,选取了其中 318 位用户,电影为 511 部,电视剧为 397 部,歌曲 633 首.对于电影、电视剧以及歌曲的评分数据集,分别有 1 496,1 253,1 800 条数据.经处理后,每维实验数据是二维表格结构的评分表,表结构见表 1.

Table 1 Evaluation table structure of user

表 1 用户评分表结构

标识符	含义
userId	用户标识
itemNum	项目类别标识
itemId	项目标识
rating	评分
timestamp	时间戳

实验中,每维网络对应一种项目类型.因为用户的一致性,将评分表按 userId 作为外键连接起来,就生成了针对每位用户的三维数据表.

4.2 评价标准

本文采用衡量链路预测算法的精确度指标作为评价算法指标.精确度 Accuracy 是系统产生的正确预测与所有预测的百分比.其公式见式(4),其中 n_cp 代表正确的预测数, n_tp 代表所有的预测数目.

$$Accuracy=n\_cp/n\_tp\times 100\% \tag{4}$$

4.3 实验结果

实验中用到的用户和资源数据集随机分成训练集和测试集,二者分别占有 80%和 20%的比例.训练集用来

产生预测结果,测试集用来评价推荐结果的优劣.通过训练集预测可能产生的连边,并将可能连边的相似度由大到小排列,从中抽取前 L 相似度最大连边.

本文设置了 2 组对比实验.首先以维度为考量因素,将本文提出的多维网络双向链路预测方法简称为 DDMLP 算法,将其与经典的决策树方法和贝叶斯方法进行比较.被测试用户集合分成 10 组,对特定的目标用户分别进行 3 种方法的链路预测,得到准确率为各组用户结果的平均值,实验结果如图 2 所示.

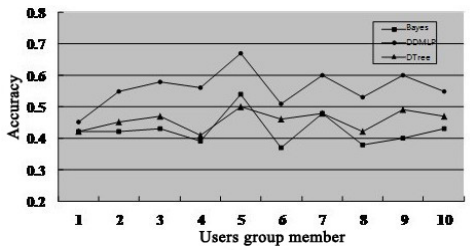

Fig.2 Accuracy comparison

图 2 精确度实验结果

实验结果可以得出,动态多维社会网络中的双向链路预测,相比于决策树方法和贝叶斯方法,精确率较高,说明 DDMRS 算法有比较准确的预测结果.需要注意的是,算法运行时间要长于其他两种算法,分析原因应是算法中在迭代 3 种网络中的用户数据构建多维加权网络时耗时较多.

其次,以动态为考察因素,对比了动态多维网络环境下的双向链路预测算法(DDMLP_dyc)与未实行动态演化的多维加权推荐算法(DDMLP),同样随机抽取 10 组用户,计算机出结果取平均值.得到准确率的平均值和时间消耗平均值,实验结果如图 3 所示.

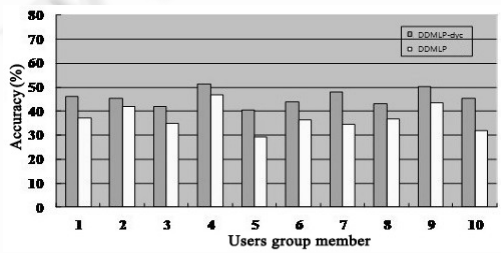

Fig.3 Accuracy comparison

图 3 精确度实验结果

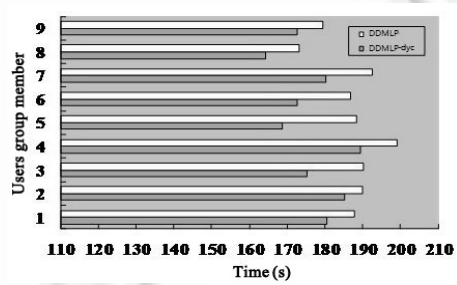

Fig.4 Time comparison

图 4 时间实验结果

实验结果表明,DDMLP_dyc 的精确率 DDMLP 高.另外,由实验结果可以看到,DDMLP_dyc 时间消耗要普遍少于 DDMLP 算法,在某些用户组中表现还十分明显.因为在 DDMLP 中,形成降维网络后,在用户更新时需要重构网络,增加了数据冗余.而两方面的实验数据可以得到,动态多维网络链路预测算法,获取了良好的推荐质量和效率.

5 结束语

本文提出了动态多维网络双向链路预测方法,在动态多维网络环境中既可以预测将来可能产生的链接,也可以预测现有的而将来可能消失的链接.我们认为本文提出的动态多维网络双向链路预测方法非常重要,它适合现实世界中很多场景,可以作为现有链路预测的必要补充.实验结果表明,它能够使多维加权网络中链路预测有更好的效果.今后应该做一些细致的工作,研究调整和细化相应的参数.另外,应该优化算法,提高算法的效率.

References:

[1] Getoorl L, Diehl CP. Link mining: A survey. ACM SIGKDD Explorations Newsletter, 2005,7(2):3−12.

[2] Thurner S, Biely C. Two statistical mechanics aspects of complex networks. Physica A, 2002,74(2):346−353.

[3] Gallagher B, Tong HH, Eliassi-Rad T, Faloutsos C. Using ghost edges for classification in sparsely labeled networks. In: Proc. of the ACM SIGKDD 2008. 2008. 256−264.

[4] Liben-Nowell D, Kleinberg J. The link prediction problem for social networks. In: Proc. of the 12th Int'l Conf. on Information and Knowledge Management. ACM, 2003. 556−559.

[5] Clauset A, Moore C, Newman MEJ. Hierarchical structure and the prediction of missing links in networks. Nature, 2008,45(3): 98−101.

[6] Holland PW, Laskey KB, Leinhard S. Stochastic block models: First steps. Social Networks, 1983,5:109−137.

[7] Dunlavy DM, Kolda TG. Temporal link prediction using matrix and tensor factorizations. ACM Trans on Knowl Discov Data (TKDD), 2011,5(2):10−18.

[8] 王林,商超.无标度网络中的链路预测问题研究.计算机工程,2012,38(3):67−70.

[9] Allali O, Magnien C, Latapy M. Link prediction in bipartite graphs using internal links and weighted projection. In: Proc. of the IEEE Conf. on Computer Communications Workshops. 2011. 936–941.

[10] Tylenda T, Angelova R, Bedathur S. Towards time-aware link prediction in evolving social networks. In: Proc. of the 3rd SNA-KDD Workshop on Social Network Mining and Analysis. ACM, 2009. 1–10.

[11] Naji G, Nagi M, Elsheikh AM, Gao S, Kianmehr K, Özyer T, Rokne J, Demetrick D, Ridley M, Alhajj R. Effectiveness of social networks for studying biological agents and identifying cancer biomarkers. In: Proc. of the Counterterrorism and Open Source Intelligence. 2011. 285–313.

[12] 孙继佳,蒋健,严广乐,李季明,苏式兵,朱蕾蕾,高月求.复杂网络理论及其在中医学研究中的应用.复杂系统与复杂性科学,2008, 5(3):55−61.

[13] Zhou T, Ren J, Medo M, Zhang YC. Bipartite network projection and personal recommendation. Phys. Rev. E, 2007, 76–115.

[14] Zhou T, Jiang LL, Su RQ, Zhang YC. Effect of initial configuration on network-based recommendation. Europhy. Lett., 2008, 81–85.

[15] Ordonez C, Omiecinski E. Efficient disk-based *K*-means clustering for relational databases. IEEE Trans. on Knowledge and Data Engineering, 2004,16(8):909–921.

[16] Han JW, Kamber M. Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers, 2000.

[17] 邓爱林.电子商务推荐系统关键技术研究[博士学位论文].上海:复旦大学,2003.

王红(1966－),女,天津人,教授,博士生导师,主要研究领域为移动社会性软件,复杂网络,工作流.

于晓梅(1972－),女,副教授,主要研究领域为智能计算,网理论及应用.

孙彦燊(1993－),女,主要研究领域为优化算法.