信任网络中的信任实体重要性发现方法①

郭伟鹏, 龚卫华

(浙江工业大学 计算机科学与技术学院, 杭州 310023)

摘 要: 信任网络是开放网络环境下抵御恶意欺诈、降低用户交互风险的有效手段, 现有关于信任网络的研究集中于信任评价的传播和计算策略, 对信任网络中实体重要性没有足够的重视. 本文提出一种信任序列模式挖掘算法 T-Seq, 将信任传播过程作为信任序列, 通过序列挖掘方法有效找出信任网络中的重要信任节点. 实验表明了 T-Seq 算法在信任序列模式挖掘和重要节点发现上的有效性.

关键词: 信任网络; 节点重要性; 信任传播; 信任路径; 序列挖掘

Important Nodes Discovery in Trust Network

GUO Wei-Peng, GONG Wei-Hua

(School of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract: Trust network is an effective means of resisting malicious fraud and reducing the risk of interactions between users. Most of existing researches focus on trust propagation and aggregation strategies, while pay little attention to the importance of agents. This paper presents a sequential pattern mining algorithm which called T-Seq, in which trust propagation process is regarded as trust sequences. It is effective in important nodes discovery through sequences mining. Experiments showed the effectiveness of T-Seq in sequential pattern mining and important nodes discovery.

Key words: trust network; node importance; trust propagation; trust paths; sequential mining

1 引言

近年来, 在线电子商务、社交应用等深刻改变了人们的社会活动和交流方式. 然而, 由于利益的驱使, 大量恶意用户的欺诈行为充斥于这些开放的网络环境中, 所带来的信任危机和安全问题成为阻碍这些应用系统发展的瓶颈. 信任网络结构[1]在抵御恶意欺诈、降低用户交互风险上的突出表现使其备受重视, 通过信任评价信息的传播和聚合计算, 用户能综合判定目标用户的可信程度, 从而决定是否与之进行交互.

信任网络本质上是一个社会网络, 由直接信任关系和间接信任关系构成, 实体的交互决策不仅依靠个体自身的直接交互历史, 还受到其他实体的影响[2]. 现有研究大都集中于间接信任的路径查找分析和推理聚合的计算方式[3-8], 对信任网络中实体的重要性及信任关联关系没有足够的重视, 而这对理解信任传播过程、掌握信任网络结构和帮助抵抗恶意攻击有着重要意义.

目前已有一些判断节点重要性的度量指标如度中心性、特征向量中心性、介数等[9], 但这些指标侧重于从网络拓扑结构角度分析节点重要性, 没有考虑节点之间的关联关系, 不能有效适用于信任网络中实体重要性分析及关键信任路径的发现. 总体上看, 现有信任网络研究存在的不足主要在于在信任传播和信任聚合时对所有信任实体同等对待, 缺乏关于信任实体的重要性分析, 也忽视了信任路径上实体间的关联关系.

针对上述问题, 本文使用序列模式挖掘方法分析信任网络中信任序列结构及节点重要性. 信任网络中节点间的信任传递路径构成了信任序列, 其中一些信任序列成为众多节点对的信任传递频繁经过的公共路径, 因而在信任网络中占有重要地位, 相应地构成这

① 基金项目:国家自然科学基金(61070042);浙江省自然科学基金(LY13F020026,Y1080102)
收稿时间:2013-09-15;收到修改稿时间:2013-10-14

些信任序列的实体即为重要实体. 实验表明本文提出的方法能有效发掘信任网络中的频繁信任序列, 准确体现了信任实体的重要性.

2 相关工作

目前关于信任网络的研究大多关注的是间接(推荐)信任关系的推理及其聚合计算. 如 Sunny[3]基于所有路径估计目标节点的可信程度; 而 Verbiest 等[5]则认为路径越长其信任度越低, 因而提出了一种限制路径长度的动态路径搜索算法, 即在聚合计算时仅考虑长度不超过一定阈值的路径; Li 等[6]分析了影响信任评价的多个因子, 使用加权移动平均(WMA)和有序加权平均(OWA)算法动态分配多因子的权重. 此外, 间接信任的计算方法还有主观逻辑[7]、D-S 证据理论[8]等, 这些方法往往仅考虑信任强度和路径长度对信任传递的影响, 忽视了地位不同的节点对信任传递的影响存在较大差别这一事实, 且间接信任的推理和计算过程还存在一些信息损失, 因而导致其信任评价不够可靠.

社会网络中度量节点重要性的常用指标有度、介数、接近中心性、特征向量中心性等[9], 这些指标从网络拓扑结构出发, 以不同角度刻画节点重要性. 例如, 基于度的节点重要性度量方法强调节点的邻居节点的数量, 但无法反映节点所处复杂环境对其重要性的影响; 介数是基于经过节点的最短路径数体现其对网络中信息流的控制能力, 但实际上多数网络中信息的流动并不仅仅依赖最短路径. 上述指标各有其优缺点, 但在复杂的信任网络中很难依靠单个指标判断节点重要性.

另外一些研究则基于链接分析提出度量节点重要性的方法, 其中最为经典的是 PageRank 算法[10]和 HITS 算法[11]. PageRank 算法认为一个节点的重要性取决于指向它的其他节点的重要性, 每个节点的 PR 值可由指向它的节点的 PR 值按出度大小加权计算得到. 而 HITS 算法为每个网页节点分配两个指标: Hub 值和 Authority 值. 节点的 Hub 值为它所指向的节点的 Authority 值之和, Authority 值则为指向它的节点的 Hub 值之和, 二者是相互增强的关系. 这两种算法分别依据节点的 PR 值或 Authority 值进行节点重要性排序, 在计算过程中都没有区分链接的重要性.

综上所述, 信任网络中实体间的信任关系由于受多种因素影响而更加复杂, 上述节点重要性评价指标并不适用于信任网络中的重要节点发现, 而从目前国内外研究现状的调研情况看, 尚无信任网络重要实体发现的相关研究. 因此, 本文提出一种信任序列模式挖掘算法, 该算法从节点间的信任关系出发, 分析和挖掘网络中所有节点对之间的信任路径中的频繁信任序列, 并结合节点关联特征, 有效找出信任网络中的重要节点.

3 信任实体重要性分析

信任网络中陌生实体间信任关系的建立依赖于中间实体的信任传递,这些相互关联和作用的实体形成有向的信任路径, 频繁的公共信任路径及相关节点在信任传递过程中起着枢纽作用. 挖掘和发现重要信任实体对深入分析信任网络结构具有重要意义, 为此本文提出一种信任序列模式挖掘算法 T-Seq, 以找出频繁信任路径序列和相应的重要信任实体.

3.1 信任序列定义及表示

本文首先给出信任网络的形式化定义如下:

定义 1. 信任网络(Trust Network): 由代表实体的节点以及它们之间的信任关系组成一个有向加权图, 表示为 TN=(V,E), 其中 V 为信任节点集, E 为具有信任值的有向边集.

从社会关系角度看, 信任网络主要包含两类关系: 直接信任和间接信任, 其中边集 E 表示了直接信任关系, 而陌生节点间的间接信任关系建立依赖于中间节点的信任传递, 这些相互关联和作用的中间节点通过若干个直接信任边连接形成有向的路径序列, 每一条路径序列就是一个信任传播过程, 称之为信任序列.

定义 2. 信任序列(Trust Sequence): 信任网络中从源节点 i 经 k 跳到达目标节点 j 的 k 条直接信任边相连接所构成的一条有向路径序列, 可表示为:

$$S_{ij}^{k}=\{<i\to\cdots\to j>\left|W_{ij}'\geq\omega\right.\} \tag{1}$$

其中 W'_{ij} 是信任序列 S_{ij}^{k} 从源点 i 到终点 j 的间接信任值, 取决于 k 条直接信任边上的信任权重 W, 采用式((2)计算, ω 为信任阈值, 满足 $0<\omega<1$, 设置信任阈值限制是为了过滤掉信任度偏低的序列. 包含 k 条边的信任序列称之为 k-信任序列.

$$W_{ij}'=\prod_{w_{ik}\in P_{ij}}W_{ik} \tag{2}$$

定义 3. 频繁信任序列: 给定包含 n 个信任序列的集合 $P=\{P_1,...,P_n\}$ 和支持度阈值 σ，如果存在一个信任序列 S^k 在 P 中出现次数满足支持度 σ 条件，即 $\exists S^k,\exists P_i,S^k\subseteq P_i\wedge\left|S^k\right|\geq\sigma$，则序列 S^k 是频繁 k-信任序列，又称为 k-序列模式.

传统的频繁项定义只需满足集合特性，而本文所定义的频繁信任序列是项集的有序列表，序列中的节点都保持有序性和关联性. 频繁 k-信任序列其具体意义在于: 作为信任传播过程中的骨干，该序列上所有节点共同为网络中较多节点提供推荐信任评价. 相应地构成频繁信任序列的节点也是网络中的重要节点.

3.2 频繁信任序列挖掘算法

本文通过信任序列模式挖掘方法获取信任网络中的重要实体，该方法主要分两步完成: (1)遍历信任网络获取所有信任路径集合; (2)通过序列挖掘算法在信任路径集合中筛选频繁信任序列集. 由于节点间传递信任评价信息构成了信任路径序列，为分析节点对整个网络的信任传播影响程度，故需要首先获取网络中的所有信任路径. 本文基于深度优先搜索(DFS)获取网络中的所有信任路径集合，其过程描述如下.

算法 1. 信任路径集获取算法 GetPaths

输入: 信任网络 TN=(V,E)，信任阈值 ω

输出: 信任网络 TN 中 n 条信任序列集合

P={P_1,⋯,P_n}

① for each 节点 i∈V do

② i 入栈 seq, visited[i]=true;

③ repeat

④ 取栈 seq 的栈顶节点 u，当前路径 P_k 信任值 W_u=1;

⑤ if∃v∈OutNeighbors(u)且!visited[v]

⑥ 当前路径 P_k 的信任值 $W_k=W_k*w_{uv}$;

⑦ if $W_k>\omega$ then

⑧ 将边 e_{uv} 加入当前路径 P_k;

⑨ visited[v]=true, v 压入栈 seq;

⑩ else

⑪ 将 u 从栈 seq 弹出, $P=P\cup P_k$;

⑫ until stack=Φ;

⑬ 将所有节点的访问标记 visited 重新置为 false

⑭ end for

⑮ 输出 P={P_1,⋯,P_n}

由算法 1 获得整个网络的信任路径集后，本文进一步使用类 Apriori 性质查找频繁信任序列. 频繁信任序列以最小支持度阈值 σ 为约束条件，即信任序列在信任传播过程中作为公共信任路径的频次最少要满足 σ，其体现了信任序列的重要程度，因此构成频繁信任序列的节点集就是对信任传播有较大影响的重要节点集. 首先统计每条边即 1-信任序列在信任路径集中出现的频次，找出所有满足支持度阈值 σ 的频繁 1-序列，然后，通过合并频繁 1-序列生成新的长度加 1 的候选序列，再判断其是否频繁. 这样依次类推，由频繁 1 序列合并生成频繁 k 序列的过程就是逐渐发现信任路径集中所有频繁信任序列的过程，具体算法描述如算法 2 所示.

算法 2. 信任序列挖掘算法 T-Seq

输入: 信任网络 TN 中 n 条信任序列集合

P={P_1,⋯,P_n}

输出: 频繁信任序列集

TS={$F^1\cup F^2\cup\cdots\cup F^k$}={ S_1^1,⋯, S_m^i,⋯, S_n^k }.

① 初始化: k=1; $F^1=\Phi$;

② for each 候选 $e_{ij}\in P_i$ do

③ 统计其在集合 P 中频次$|e_{ij}|$;//即节点 i 推荐节点 j 的信任评价信息的频次，次数越多越重要

④ if $|e_{ij}|\geq\sigma$ then 加入集合 $F^1=\{S_i^1\,|\,S_i^1=e_{ij}\wedge|e_{ij}|\geq\sigma\}$中;//筛选满足支持度 σ 的 1-序列

⑤ end for

⑥ 删除集合 P 中未出现 F^1 中 1-序列的信任序列;//削减不频繁即不重要的信任序列集

⑦ repeat

⑧ k=k+1;

⑨ $F^k=F^{k-1}\infty F^{k-1}$; //频繁的 k-1 序列经自然连接生成候选的 k-序列集

⑩ for each $S^k\in F^k$ do

⑪ 统计 S^k 在 P 中的频次并计算支持度$|S^k|$;

⑫ if $|S^k|<\sigma$ then $F^k=F^k-S^k$; //验证 S^k-候选序列是否频繁

⑬ end for

⑭ 删除集合 P 中未出现 F^k 中 k-序列的路径序列; //削减不频繁的路径序列集

⑮ until $F^1=\Phi$ or $|P|<\sigma$;

⑯ TS= {$F^1\cup F^2\cup\cdots\cup F^k$ }. //输出所有的频繁 k-序列集，即频繁信任序列集 TS

步骤⑨中频繁序列集合 F^{k-1} 中任意两个信任序列

S_i^{k-1}，S_j^{k-1}合并生成候选的信任序列 S^k 所需满足的条件是：频繁序列 S_i^{k-1} 和 S_j^{k-1} 间的交集存在相同且连续的 k-2 子序列，这样两个序列自然连接的结果 S^k 应是由 S_i^{k-1} 与 S_j^{k-1} 的最后一条边连接而成(假设序列的前 k-2 子序列是 S_i^{k-1} 的 S_j^{k-1} 子集)，生成的候选序列 S^k 需进一步在路径序列集合中验证是否频繁.

频繁信任序列集构成了信任传播过程中的骨干，相应地其中的节点也是影响信任传播的重要节点. 由于每一条频繁信任序列都是由边构成，可以构成一个信任子图 $TN_{sub}=(V',E')$，其中 $V'\subset TS\subset V$，$E'\subset TS\subset E$. 将信任子图中的节点按其度大小进行排序，即可得到一个重要性递减的重要节点集.

获取重要节点集的整个过程可以总结为：将节点间信任传递构成的有向信任路径作为信任序列，首先获取整个网络的所有信任序列，对之进行序列挖掘以获取对整个网络信任传播有较大影响的频繁信任序列集，然后利用频繁序列集构建信任子图，从而得到重要节点集.

重要节点间存在一定的关联关系，这种关系体现在频繁信任序列上，以一条最长频繁序列 F^k 进行说明，$F^k=\{i\to\cdots k\to\cdots\to j\}$中的重要节点 i, …, k, …, j 都是相关联的，即在一个信任传播过程中这些节点共同提供了推荐信任评价信息. 在以后的信任网络研究中，需要着重考虑重要节点提供的推荐信任评价，优化信任传播的过程以提高效率，针对重要节点集建立保护和监测机制，防止重要节点被攻击对整个网络造成恶劣影响.

4　实验及结果分析

本文采用 Advogato 数据集构建信任网络进行算法实验与分析，该数据集包含 7184 个节点和 51516 条有向边，同时节点对上有向边的权重表示信任评价，共分 4 个级别：Observer、Apprentice、Journeyer 和 Master，为便于处理，本文将之映射到[0,1]区间的离散值，即 Observer=0.4、Apprentice=0.6、Journeyer=0.8、Master=1.0.

图 1 显示了 Advogato 数据集所构成的信任网络图，从图中可以看出，节点间关系错综复杂，没有呈现出明确的结构特征和规律性，也无法识别重要节点.

在图 1 基础上运行算法 T-Seq 获得支持度阈值为 25 时频繁信任序列集如图 2 所示. 图中用红色的有向边标记了总共 15 条序列长度 k≥10 的频繁信任序列集，橙色节点表示该集合所包含的重要节点，这些序列路径边数占该图中总边数的 3.89%. 与图 1 相比，图 2 中结果表明本文所提出的算法 T-Seq 能有效发现信任网络中频繁信任序列所包含的重要节点，呈现出非常清晰的信任网络结构特征和规律性.

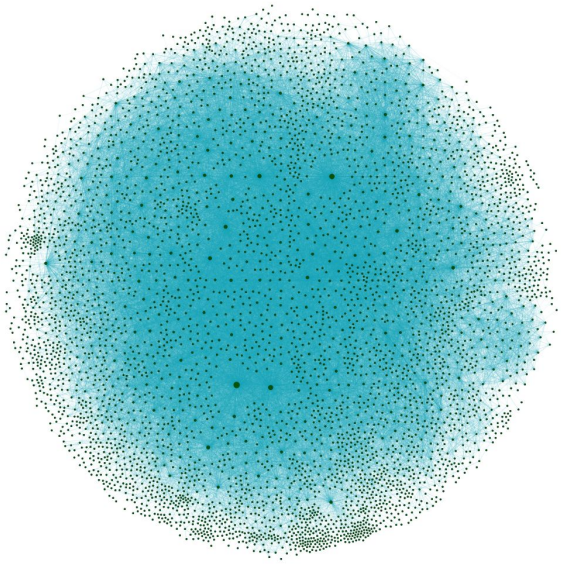

图 1　信任网络图

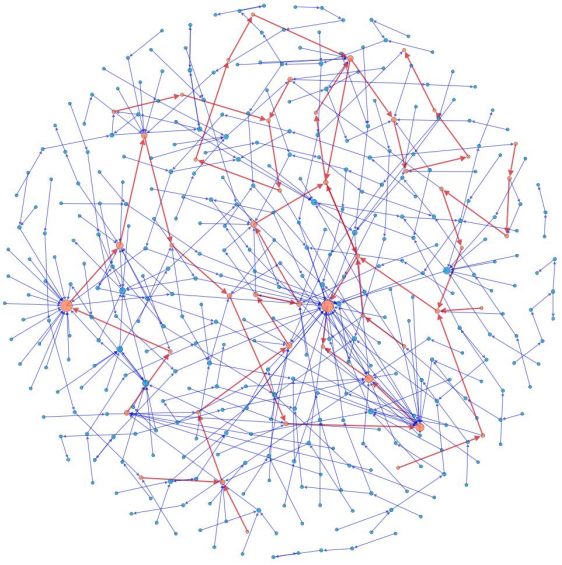

图 2　挖掘出的多维信任序列结构

另外，本文将 T-Seq 算法与经典的 PageRank 算法和 HITS 算法的排名结果进行比较，当信任度阈值从 0.2 递增到 0.8 时，这三种算法在 Advogato 数据集中挖

掘到的重要节点集的平均绝对误差(MAE)指标如图 3 所示.测试中 T-Seq 设置支持度阈值为 15, 而 PageRank 和 HITS 算法分别依据排序的 PR 值和 Authority 值选择与 T-Seq 挖掘结果相同数量的节点集进行对比. 从图 3 可以看出, HITS 算法挖掘结果的 MAE 值最大并且随着信任度的增加而逐渐增大, 因信任度变化导致节点集内链接变化很大从而直接影响到节点排名. 而 PageRank 的 MAE 值比较稳定, 一直维持在 0.35 左右, 这是由于 PageRank 算法对网络结构变化有一定的鲁棒性, 节点 PR 值的计算仅与其出入度链接相关而不受信任度变化影响. 三者之中, T-Seq 算法挖掘结果最好, 其不同信任度阈值下的 MAE 值都低于 PageRank 和 HITS, 这是因为 T-Seq 挖掘的重要节点集是在关联性基础上考虑节点的重要性, 更能准确体现节点在网络中的重要地位.

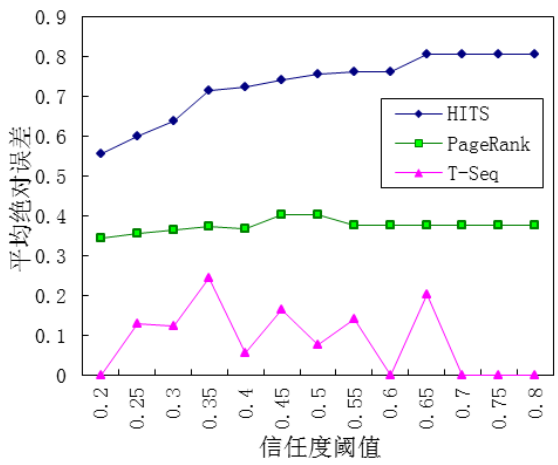

图 3 三种算法的平均绝对误差比较

5 小结

针对目前信任网络研究缺乏重要节点发现方法的问题, 本文提出了一种信任序列模式挖掘算法 T-Seq, 该算法根据信任传递的特点, 通过信任序列模式挖掘, 结合节点关联特征有效地找出信任网络中的重要节点. 在今后的信任网络研究中, 需要着重考虑重要节点提供的推荐信任评价, 以优化信任传播过程, 同时针对重要节点集建立保护和监测机制, 防止重要节点被攻击对整个网络造成恶劣影响.

参考文献

1 Li X, Liu L. PeerTrust: Supporting reputation-based trust for peer-to-peer electronic communities. IEEE Trans. on Knowledge Data Engineering, 2004,16(7): 843–857.

2 郭零兵,罗新星,朱名勋.移动商务信任的演化博弈及动态仿真.计算机系统应用,2013,22(7):1–6.

3 Kuter U, Golbeck J. Sunny: A new algorithm for trust inference in social networks using probabilistic confidence models. Proc. of 22nd AAAI conf on Artificial Intelligence. Vancouver, AAAI Press. 2007. 1377–1382.

4 Richters O, Peixoto TP. Trust transitivity in social networks. PloS One, 2011, 6(4): 1–14.

5 Verbiest N, Cornelis C, Victor P, Herrera-Viedma E. Trust and distrust aggregation enhanced with path length incorporation. Fuzzy Sets and Systems, 2012, 202: 61–74.

6 Li X, Zhou F, Yang X. A multi-dimensional trust evaluation model for large-scale P2P computing. Journal of Parallel and Distributed Computing, 2011, 71(6): 837–847.

7 Sensoy M, Pan JZ, Fokoue A, Srivatsa M, Meneguzzi F. Using subjective logic to handle uncertainty and conflicts. Proc of 11th Int Conf on Trust, Security and Privacy in Computing and Communications (TrustCom). Liverpool, IEEE Computer Society. 2012. 1323–1326.

8 Jiang L, Xu J, Zhang K, Zhang H. A new evidential trust model for open distributed systems. Expert Systems with Applications, 2012, 39(3): 3772–3782.

9 Costa LF, Rodrigues FA, Travieso G, Villas Boas PR. Characterization of complex networks: A survey of measurements. Advances in Physics, 2007, 56(1): 167–242.

10 Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 1998, 30(1): 107–117.

11 Kleinberg JM. Authoritative sources in a hyperlinked environment. Journal of the ACM, 1999, 46(5): 604–632.