基于兴趣图谱的用户兴趣分布分析及专家发现

国 琳[1,2],左万利[1,2]
(1. 吉林大学计算科学与技术学院,吉林长春 130012;2. 吉林大学符号计算与知识工程教育部重点实验室,吉林长春 130012)

摘 要: 尽管用户可自主生成个性化数据以更全面描述个人偏好,但由于用户创建数据不严谨、不可控,导致生成的庞大数据集大多存在质量低、噪声严重的缺陷.因此管理复杂网络信息时,不能仅使用写入性知识,必须重视具有大量领域知识的专家,因为其可为系统提供高质量的信息.本文通过构建和分析用户兴趣分布曲线以发现兴趣领域专家,并提出甄别状态不正常的伪专家算法.由于网络中权威专家数量较少,所以所提供的信息是有限的.因此本文定义的领域专家不仅包含权威专家,而且包含普通用户中对某领域有极高关注的兴趣领域专家.实验证明算法的正确性和高效性,并且较低的复杂度使其可处理海量用户节点信息.

关键词: 专家发现; 兴趣分析; 兴趣图谱; 复杂网络分析
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2015)08-1561-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2015.08.014

Analysis of User Interest Distribution and Expert Finding Based on Interest Graphs

GUO Lin[1,2], ZUO Wan-li[1,2]
(1. *College of Computer Science and Technology of Jilin University*, *Changchun*, *Jilin* 130012;
2. *Symbol Computation and Knowledge Engineer of Ministry of Education of Jilin University*, *Changchun*, *Jilin* 130012)

Abstract: Although users can self-generate personalized data for describing preferences more comprehensively, user-created data is not rigorous and uncontrollable, which leads to enormous data of low quality with serious noise. On managing complex network, more attentions should be placed on high quality information that has been or will be produced by experts who have knowledge in specific fields in order to avoid being restricted to written knowledge. This paper finds experts by constructing and analyzing interest profiles of users and proposes a screening method for detecting abnormal pseudo-experts. Due to the small number of authoritative experts in networks, which provide a limited amount of information, experts defined in this paper not only include authoritative experts, but also ordinary users that have a lot of knowledge in a certain field. Experiments illustrate the correctness and effectiveness of the algorithm, and the low complexity renders it suitable in handling massive user node information.

Key words: expert finding; interest analysis; interest graph; complex network analysis

1 引言

从社会网络用户中发现专家以扩展网络信息储备量,是一种在特定领域中寻求知识量较多用户的信息检索过程.网络信息多样化导致数据噪声和严重的冗余现象,因此知识挖掘和管理不能仅局限于写入性知识(传统互联网页面信息),必须重视拥有大量领域知识的专家所能为系统提供的高质信息.专家发现算法主要包含:基于共词分析和网络特征分析[1],基于关联关系分析或网络节点链接信息[2],基于文本相关度和用户权威度分析[3],发现社区以确定专家[4],基于监督/半监督的机器学习[5],有识别能力的概率模型[6]及多种专家发现策略的融合算法[7].分析现有算法可发现在缺乏训练集或已标记数据集的条件下,从复杂社会网络或模糊数据

收稿日期:2014-03-10;修回日期:2014-10-15;责任编辑:马兰英
基金项目:国家自然科学基金(No.60973040);国家自然科学青年基金(No.61300148);吉林省重点科技攻关项目基金(No.20130206051GX);吉林省科技计划青年科研基金(No.20130522112JH);中国博士后基金项目(No.2012M510879);吉林大学基本科研业务费科学前沿与交叉项目(No.201103129)

中发现专家十分困难,并且大部分算法缺乏排除恶意用户或伪专家用户的能力,通常这些算法选择出的专家仅为活跃的、影响度高的节点,而这些节点不保证一定具备专家特征,因此可能忽略了真正的专家节点.由于网络中权威专家数量较少,所以其为网络提供信息的能力有限,故增加普通专家可有效扩充网络高质量节点数据.本文提出专家发现算法不仅可发现高影响度的权威专家节点,也可发现有较强领域知识的普通节点.

2 兴趣分布分析

2.1 数据构建

由于用户产生的数据质量完全不可控,所以本文提出的兴趣分析算法仅通过用户标注的数据类型分析其关注领域,而不涉及具体数据.

数据分为两类:(1)项目类,其表达形式为 item:<itemID,item>,其中 itemID 表示项目的唯一标识编码,item 表示具体项目;(2)用户-项目类,其表达形式为 user-item:<userID,itemID,score>,其中 userID 表示用户编码,itemID 表示项目编码,score 表示用户对项目的评分或者网页停留时间等信息.根据兴趣间的相关性将 item 分类以分析用户对某一兴趣类别关注的强度,以此增强算法处理问题的能力和提高分析结果的准确度.最终 item 的表达形式为 <itemID,class>,其中 class 表示兴趣类别.以此推导,user-item 矩阵转换为 user-class 矩阵:<userID,class,count>,其中 count 表示用户与类别关联次数(下文中,user-class 矩阵记为 $\boldsymbol{T}$).

$$\boldsymbol{T}=\begin{pmatrix}4&2&3&3&5&3&2\\7&5&0&0&1&2&0\\6&0&0&1&1&0&0\\1&1&1&1&1&1&1\end{pmatrix}$$

假设 $\boldsymbol{T}$ 包含 4 位用户对 7 个兴趣类别的关注程度信息.计算每个用户对不同兴趣的关注程度的过程等价于对矩阵 $\boldsymbol{T}$ 的分析和处理.由于不同用户的兴趣关注程度和关注范围不同,因此需要统一度量标准,使用户间可横向比较.

规范化公式为:

$$\boldsymbol{\xi}_{(k)}\in\boldsymbol{T},\ x_i\in\boldsymbol{\xi}_{(k)} \tag{1}$$

$$\boldsymbol{T'}_i^k=\frac{x_i}{\sum_{Z_i\in\xi_{(k)}}Z_i} \tag{2}$$

$$\sum_{i=1}^{n}\boldsymbol{T'}_i^k=1 \tag{3}$$

$\boldsymbol{\xi}_{(k)}$表示 $\boldsymbol{T}$ 的第 k 行向量.x_i和 Z_i表示 $\boldsymbol{\xi}_{(k)}$中第 i 参数,其描述用户 k 与兴趣 i 的关联程度.$\boldsymbol{T'}_i^k$ 表示兴趣 i 在用户 k 的所有被关注兴趣中的比重,其体现用户 k 对兴趣 i 的关注程度.

数据标准化处理后的 $\boldsymbol{T'}$ 为:

$$\boldsymbol{T'}=\begin{pmatrix}0.183&0.091&0.136&0.136&0.227&0.136&0.091\\0.467&0.333&0&0&0.067&0.133&0\\0.750&0&0&0.125&0.125&0&0\\0.143&0.143&0.143&0.143&0.143&0.143&0.143\end{pmatrix}$$

如果直接运用矩阵 $\boldsymbol{T'}$ 分析用户兴趣分布,则无法执行数据拟合.因此需要对每一行参数按照特定顺序排序,以此体现数据变化趋势.本文采用升序排序策略以显示兴趣关注递增的变化程度,排序后的 $\boldsymbol{T'}$ 记为 $\boldsymbol{T}_{\text{order}}$.

$$\boldsymbol{T}_{\text{order}}=\begin{pmatrix}0.091&0.091&0.136&0.136&0.136&0.183&0.227\\0&0&0&0.067&0.133&0.333&0.467\\0&0&0&0&0.125&0.125&0.750\\0.143&0.143&0.143&0.143&0.143&0.143&0.143\end{pmatrix}$$

$\boldsymbol{T}_{\text{order}}$每行数据中排序最后的数据对应该用户关注程度最高的兴趣领域,从后向前对应的兴趣关注程度依次减弱.虽然 $\boldsymbol{T}_{\text{order}}$有能力分析用户对不同兴趣领域的关注程度,但对数据变化趋势描述不明确,因此需要改进 $\boldsymbol{T}_{\text{order}}$以扩大兴趣类别间的差异,更便于数据比较和专家发现.

$\boldsymbol{T}_{\text{order}}$改进公式:

$$\boldsymbol{\eta}_{(k)}^{\mathrm{T}}\in\boldsymbol{T}_{\text{order}} \tag{4}$$

$$\boldsymbol{\zeta}_{(k)}^{\mathrm{T}}=\sum_{i=1}^{k}\boldsymbol{\eta}_{(i)}^{\mathrm{T}} \tag{5}$$

$\boldsymbol{\eta}_{(k)}^{\mathrm{T}}$表示 $\boldsymbol{T}_{\text{order}}$中第 k 列向量.$\boldsymbol{\zeta}_{(k)}^{\mathrm{T}}$表示从 $\boldsymbol{\eta}_{(1)}^{\mathrm{T}}$到 $\boldsymbol{\eta}_{(k)}^{\mathrm{T}}$相应位置元素的加和.由 $\boldsymbol{\zeta}$ 构成的矩阵记作 $\boldsymbol{T}_{\text{sp}}$,其描述用户兴趣分布趋势,位于行向量偏前位置的数据对应用户关注度较低的兴趣类别,位于偏后位置的数据对应用户关注度较高的兴趣类别.$\boldsymbol{T}_{\text{sp}}$扩大不同兴趣间关注程度的差异性,增加描述数据变化的曲线的弧度,以便于发现领域专家.

$$\boldsymbol{T}_{\text{sp}}=\begin{pmatrix}0.091&0.182&0.318&0.454&0.590&0.773&1\\0&0&0&0.067&0.200&0.533&1\\0&0&0&0&0.125&0.250&1\\0.143&0.276&0.429&0.572&0.715&0.858&1\end{pmatrix}$$

2.2 幂函数曲线拟合

上一章节中,矩阵的一系列变形工作的目的在于使数据满足洛仑兹曲线(Lorenz curve)[8]要求.本文运用洛仑兹曲线分析用户兴趣分布的不平等程度.不同兴趣的关注差异度越大,说明用户对某些领域有更多的关注,更有可能成为领域专家,反之说明用户的兴趣分布均匀,成为领域专家的概率较小.

对于任一 $\alpha_t\in\boldsymbol{T}_{\text{sp}}$表示 $\boldsymbol{T}_{\text{sp}}$中第 t 行向量.α_t为递增序列,其数据满足洛仑兹曲线要求,通过幂函数拟合生成的曲线称为兴趣分布曲线,设该曲线函数为 $y=ax^b$.幂函数等式两边执行对数运算,将函数曲线拟合问题转换为线性回归问题,以降低函数计算量.因此问题转

换为优化方程 $lgy = lga + b \times lgx$,其依据公式为:

$$\boldsymbol{T}_{\text{order}} = \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} \tag{6}$$

$$\boldsymbol{R}(\boldsymbol{U},\boldsymbol{V}) = \frac{1}{2}(\|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2) \tag{7}$$

$$\underset{a,b}{\operatorname{argmin}} \sum_{i=1}^{m}\{(\lg a + b\lg x_i) - \lg cz_i\}^2 + \boldsymbol{R}(\boldsymbol{U},\boldsymbol{V}) \tag{8}$$

公式(8)引入正规化参数 $\boldsymbol{R}(\boldsymbol{U},\boldsymbol{V})$相当于明确了一个拟合函数允许的误差范围,从而防止过度拟合出现.其中 $\|\cdot\|_F$表示矩阵的 $\boldsymbol{F}$ 范数,z_i表示实际观测数值.根据上述公式合理预测参数 a 和 b 取值,由此计算每个用户对应的 $y = ax^b$函数,通过分析幂函数可获得用户兴趣分布信息.

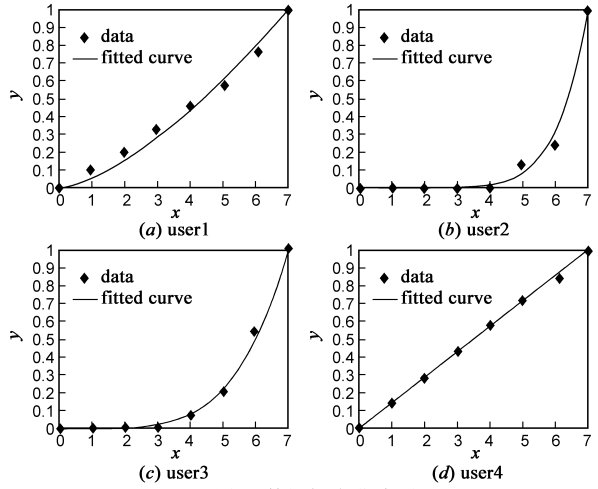


图1　数据拟合曲线图

分析曲线图可知,user1 对多个兴趣分配了几乎相同程度的关注,即无法划分出重点关注领域,因此 user1 趋于归纳为普通用户.user2 和 user3 的曲线偏右部分出现突然上升趋势,说明用户对某几个兴趣有极高关注度,因此 user2 和 user3 可能为兴趣领域专家.user4 为一种极端情况,该用户对所有涉及到的兴趣赋予完全相同的关注度,因此该图形呈直线形状.

2.3　兴趣相似度分析

由于部分兴趣间具有相似性,所以用户在关注某个兴趣的同时可能对与其类似或相关的兴趣亦有一定关注,由此导致幂函数曲线变得平滑,以至于难以识别用户高度关注的兴趣领域.为明显区分普通用户和专家用户,根据兴趣间相似度确定不同兴趣间的关系,合并相似兴趣以突出用户重点关注领域,使专家用户的兴趣曲线图更加明确.由于不同兴趣间不能完全保证相互独立,因此采用 Co-visiting 系数度量兴趣相似度:

$$\text{Co}-\text{visiting}(v_i, v_j) = \frac{c_{ij}}{f(v_i, v_j)} \tag{9}$$

$$f(v_i, v_j) = \text{count}(v_i) + \text{count}(v_j) \tag{10}$$

v_i和 v_j分别表示兴趣 i 和兴趣 j. C_{ij}表示 v_i和 v_j出现在同一用户兴趣列表中的次数. $f(v_i, v_j)$表示 v_i和 v_j受欢迎程度,本文以用户拥有量作为其度量值. count(v)表示兴趣 v 出现于不同用户兴趣列表中的次数.当 Co-visiting 系数大于阈值 Φ 时,定义 v_i与 v_j相关,否则不相关.数据的稀疏性可能导致无法获取与某些兴趣有关的足够数据,因此无法准确计算某些兴趣之间的关联关系,所以需要辅助算法来预测缺失的关联值.由于矩阵 $\boldsymbol{T}$ 仅描述用户与兴趣的关联情况,因此需要建立矩阵 $\boldsymbol{P}$ 以分析兴趣与兴趣的关联程度,过程如下:

$$p_{ij} = \begin{cases} 1, & T_{ij} \neq 0 \\ 0, & T_{ij} = 0 \end{cases} \tag{11}$$

式(11)描述借助矩阵 $\boldsymbol{T}$ 构建矩阵 $\boldsymbol{P}$ 的过程,即如果用户 i 关注兴趣 j($T_{ij} \neq 0$),则 P_{ij}赋值为 1,否则 P_{ij}赋值为 0.矩阵分解获得 $\boldsymbol{P} = \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}$,其中 $\boldsymbol{U}$ 体现用户信息,$\boldsymbol{V}$ 体现兴趣类别信息.根据分解获得的矩阵 $\boldsymbol{V}$ 构建 $\boldsymbol{Z} = \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}$,其中 $\boldsymbol{Z}$ 描述兴趣间的关联度的信息.虽然 $\boldsymbol{Z}$ 也可通过人工分析获得,但是时间消耗较大,同时数据搜集困难.因此采用矩阵分解方式预测 $\boldsymbol{Z}$ 更加符合实际要求.当 $Z\_ij \geq \Delta$ 时,说明兴趣 i 和兴趣 j 相关,当 $Z\_ij < \Delta$ 时,说明兴趣 i 和兴趣 j 无关.

由于 Co-visiting 系数和矩阵 $\boldsymbol{Z}$ 的度量标准不一致,所以无法简单融合为唯一度量标准.因此衡量兴趣相关度时,以 Co-visiting 系数作为第一判定标准.当 Co-visiting 系数缺失时,以矩阵 $\boldsymbol{Z}$ 中的预测系数作为度量参数,成为第二判定标准.

假设上例中 user2 的第二个和第六个兴趣相关,则矩阵 $\boldsymbol{T}$ 中对应 user2 的记录由 < 0.467, 0.333, 0, 0, 0.067, 0.133, 0 > 变成 < 0.467, 0.466, 0, 0, 0.067, 0 > ,此时 user2 数据变化前后对应的拟合曲线如图 2 所示.由图像分析可知,经过兴趣相似度分析处理后的曲线坡度变得更加陡峭,使得用户关注度集中的兴趣领域更加突出.

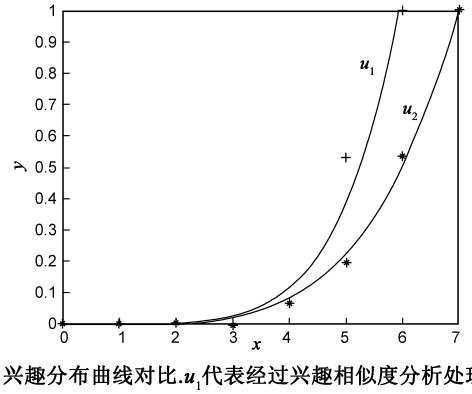


图2　兴趣分布曲线对比.u_1代表经过兴趣相似度分析处理后的曲线,u_2代表未经过兴趣相似度分析处理的曲线

3 专家发现

定义(临界点):连续曲线斜率变化的分界点.

发现幂函数曲线的临界点并计算其斜率可帮助分析用户关注度集中的兴趣领域.由于临界点为一个相对概念,因此设定阈值 θ 以区别普通点和临界点.已知 $y=ax^b$ 为递增函数,所以导数 y' 必定大于 0.曲线中某一点的右临近点导数记为 y_+',左临近点导数为 y_-'.当 $y_+'-y_-'>\theta$ 时,该点为候选临界点,否则为普通点.由于可能出现在以 a 为中心的邻域 $U(a,\delta)=\{x\mid a-\delta<x<a+\delta\}$ 内的点都满足或大部分满足阈值 θ 要求的情况,所以选取候选临界点中导数最大的点为临界点.如果曲线不存在临界点,则说明该用户兴趣分布平均,定义为普通用户.如果曲线存在临界点,则仅能说明用户的兴趣关注度集中,但却不一定为专家节点.为了准确抓取专家节点,在此引入活跃度的概念.

3.1 活跃度分析

根据节点的指入指针和指出指针数量确定用户的权威度:

$$\text{density}=\frac{E_{\text{bi-directly}}}{E_{\max}} \tag{12}$$

$$E_{\max}=\frac{n(n-1)}{2} \tag{13}$$

$$\text{reciprocity}=\frac{E_{\text{followback}}}{E_{\text{followto}}} \tag{14}$$

公式中,density 为基于密度计算的参数,reciprocity 为基于回复率计算的参数.$E_{\text{bi-directly}}$ 表示朋友圈中双向边(节点间存在双向关联关系)的数量.$E_{\max}$ 表示朋友圈中边数量的最大值,n 表示朋友圈中节点数量.E_{followto} 表示该用户关注其他用户数量,$E_{\text{followback}}$ 表示该用户的朋友圈中相互关注的用户数量.

低密度且低回复率的用户为权威用户,高密度且高回复率的用户为普通用户[9].如果某用户既是权威用户又是专家用户,则将其直接加入到该兴趣领域对应的活跃专家用户群.如果用户既是普通用户又是专家用户,则不能直接判定为专家用户,因为普通专家分为两种类型:活跃专家用户和伪活跃专家用户.伪活跃用户在网络中大部分或全部的动作为浏览网页,而提问、评价等动作极少发生,为网络产生的数据较为有限.因此伪活跃用户不增加兴趣网络知识和信息存储量,故将其排除以降低兴趣网络复杂程度.

活跃用户判定公式:

$$\text{activity}=\frac{\text{dataflow}}{\text{time\_online}} \tag{15}$$

dataflow 表示用户在时间 t 内产生的信息量或发布信息次数.time_online 表示用户浏览网络的单位时间 t.当 activity 大于阈值 θ_{ac} 时,则该用户为活跃用户,否则为非活跃用户.

阈值 θ_{ac} 的计算过程:

用户信息发布次数(记为 x)符合泊松分布,并且 x 与用户使用社会媒体次数和兴趣领域产生信息量有关[10],因此 $x\sim\text{poisson}(\lambda)$,$\lambda=ev\times n$,其中 ev 表示时间 t 内用户使用社会媒体次数,n 表示时间 t 内某兴趣领域中产生的数据总量.由此推出公式如下:

信息发布次数的概率密度公式:

$$p(x)=\mathrm{e}^{-ev\times n}\frac{ev\times n^x}{x!} \tag{16}$$

函数两边执行对数运算,结果为:

$$\log p(x)=\sum_{i=1}^{n}\{x_i\log(ev\times n)-ev\times n\} \tag{17}$$

运用最大似然估计预测参数 ev:

$$\arg\min_{ev}-\sum_{i=1}^{n}\{x_i\log(ev\times n)-ev\times n\}+\delta \tag{18}$$

δ 为常量,表示算法允许的误差范围.n 由统计网络中发布的信息量获得.ev 为未知变量.根据已观测到的数据,最小化目标函数,进而计算最大似然估计获得 ev,并使用 ev 预测 θ_{ac}.这里根据 θ_{ac} 和 ev 的不同定义,设 $\theta_{ac}=ev/t$.当待定专家用户产生的平均数据量超过某领域普通用户数据产量均值 θ_{ac} 时,定义此用户为该兴趣领域的活跃用户,并将其加入到该兴趣领域专家列表中.

3.2 专家排序

假设用户兴趣分布曲线如图 3 所示,两条曲线的临界点处的导数和其左右相邻点导数差值虽然都相等,但呈现的图型完全不同.u_1 拥有大量高关注度兴趣,u_2 拥有少量高关注度兴趣.兴趣分析时,倾向选择拥有少量高关注度兴趣的用户作为领域专家,因为这样的用户对高关注度兴趣更加专注.根据专家在兴趣领域中的可信度对专家排序,即根据临界点位置进行排序.临界点位于曲线偏右位置的视为可信度较高的专家,位于曲线偏左位置的视为可信度低的专家.

专家排序依据原则:

(1)临界点处斜率越大,专家用户权威度越高.

(2)临界点处斜率相等时,临界点的 x 坐标值越大,专家用户权威度越高.

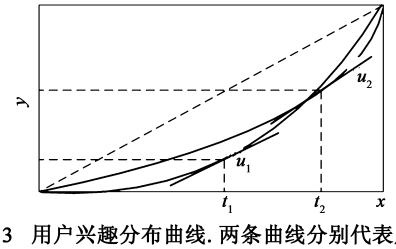


图3 用户兴趣分布曲线.两条曲线分别代表用户 u_1 和用户 u_2 的兴趣分布

综合上述分析逻辑,提出领域专家发现模型.

算法 1　PreferEF

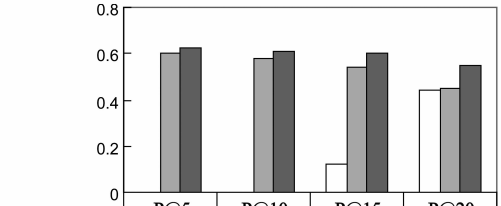
INPUT:item 矩阵,user-item 矩阵
OUTPUT:领域专家列表 L
(1)根据 item 和 user-item 生成 user-class 矩阵
(2)user-class 数据标准化处理,生成矩阵 $\boldsymbol{T}$
(3)对 $\boldsymbol{T}$ 排序(降序或升序),生成 $\boldsymbol{T}_{sp}$矩阵
(4)构建兴趣相关度预测矩阵 $\boldsymbol{Z}$
(5)FOREACH $\alpha_t \in \boldsymbol{T}_{sp}$ DO:
(6)　FOREACH $v_i \in \alpha_t$ && $v_j \in \alpha_t$ && $i \neq j$ DO:
(7)　　计算 Co-visiting 系数
(8)　根据 Co-visiting 系数矩阵和预测矩阵 $\boldsymbol{Z}$,合并相似兴趣
(9)　计算 α_t的兴趣曲线 y_t
(10)　确定临界点 a(邻域$\{x \mid a-\delta < x < a+\delta\}$内):
　　$y_{t+}(a)' - y_{t-}(a)' > \theta$ && $y_t(a)' = \mathrm{MAX}(y_t(x)')$
(11)　IF 临界点 a 存在 THEN 判断 α_t是否为领域专家:
　　(a)IF density $< \theta_d$ && reciprocity $< \theta_r$
　　THENα_t作为权威用户加入领域专家列表 L
　　(b)IF density $> \theta_d$ && reciprocity $> \theta_r$ && activity $> \theta_{ac}$
　　THEN α_t作为普通用户加入领域专家列表 L
(12)领域专家列表 L 排序
(13)返回 L

4　实验

实验采用 DBLP* 和 MovieLens***** 数据集.实验 1 采用 DBLP 数据集,并根据 Arnetminer**** 提供领域专家名单进行算法评估.由于 DBLP 的数据格式不完全适合 PreferEF 算法,因此需要为每条数据人工添加一个类别属性,所以只能随机选取部分数据执行对比实验.实验 2 使用 MovieLens 数据集,可通过评分信息将用户和电影类别进行关联,进而确定用户对不同类别的电影的感兴趣程度,所以 MovieLen 数据集更适合 PreferEF.由于这个数据集没有固定的专家列表,因此无法判断 PreferEF 找出的专家的正确性,所以将识别出的大量专家节点和普通节点的差别进行可视化.综上,实验 1 和实验 2 相互弥补不同实验的不足,综合说明 PreferEF 的执行效果.

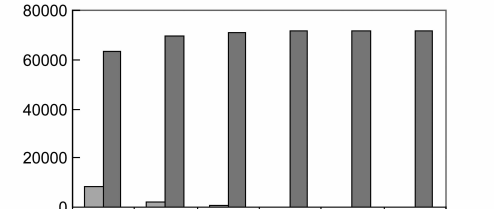
实验 1　将 PreferEF 分别与基于分析网络结构的专家发现算法 ExpertiseRank[11] 和基于文本结构分析的专家发现算法 Document-model[12] 比较,其对比结果如图 4 所示.从实验结果可知,PreferEF 优于其余两个对比算法,并且在被测评的专家节点数量增加时,其优势更加明显.

实验 2　图 5 表示设定不同的斜率阈值 θ 对专家用户类和普通用户类的影响.实验发现当 $\theta=0.9$ 时,由于阈值条件设定偏高导致无法发现领域专家节点.当 $\theta=0.8$ 时,计算出的 39 个专家节点中存在 37 个节点状态不正常(可能由关注兴趣类型单一、网络活动不频繁等因素造成极高斜率),因此这种节点应从专家类别中删除.θ 设定在[0.5,0.7]区间比较合理,发现的专家数量既满足网络需求又保证很高的节点质量.



	P@5	P@10	P@15	P@20
Document-model	0	0	0.1185	0.442
ExpertiseRank	0.6033	0.576	0.544	0.4503
PreferEF	0.6225	0.6117	0.604	0.547

图4　对比实验.P@k=$r(k)$/k,$r(k)$为前k个输出结果中计算正确的数量



	0.4	0.5	0.6	0.7	0.8	0.9
E	8370	1777	502	164	39	0
unE	63197	69790	71065	71403	71528	71567

图5　θ影响分析

分析每个用户的兴趣分布情况、构建拟合曲线并计算最大斜率值,不同用户的兴趣曲线斜率分布情况如图 6 所示.为将专家节点和普通节点明显区分,设定 $\theta=0.60$. 图 6 中虚线以上的节点为兴趣领域专家节点,虚线以下为普通用户节点.

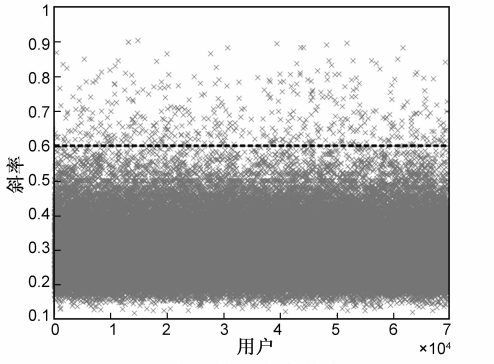


图6　发现专家节点

*　http://arnetminer.org/citation
**　http://movielens.org/login
***　http://www.datatang.com/datares/detail.aspx? id=44295
****　http://arnetminer.org/lab-datasets/expertfinding/

图 7 描述专家用户和普通用户的分布情况,观察发现领域专家节点通常为兴趣网络中的中心节点,其原因是普通用户与领域专家用户通过共同兴趣在兴趣网络中建立连接,导致领域专家节点与大量普通节点存在关联关系,因此专家节点在用户-用户关系图中成为中心节点.

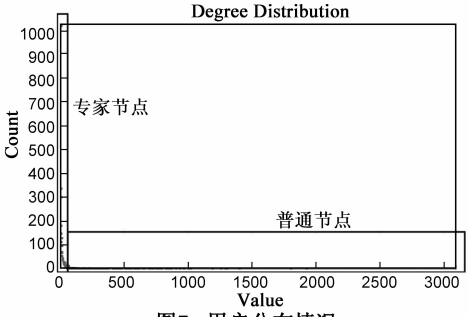


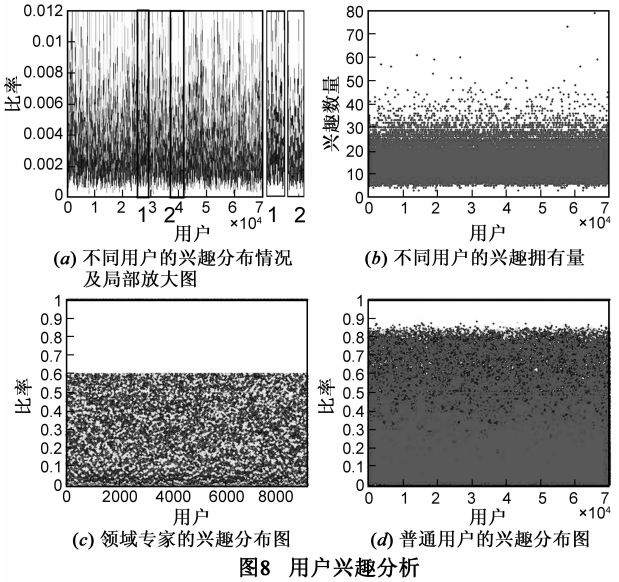
图7 用户分布情况

从 MovieLens 分析结果中随机抽取 300 个和 1000 个被算法分别归类为专家节点和普通节点的用户信息,并人工分析其划分结果的正确性,如表 1 所示.

表 1 对比数据(E 表示专家用户类,unE 表示普通用户类)

	E	unE	P(%)	R(%)	F(%)
E	25	5	66.67	86.67	75.37
unE	13	87	94.56	87.00	95.90

图 8(*a*)中每一列代表一个用户的兴趣分布,其中每种灰度表示一种兴趣,位于图像下部分区域的兴趣关注度低,位于图片上部分区域的兴趣关注度高.图像顶端白色区域代表针对单一用户关注度最高的兴趣,白色区域越大说明用户成为领域专家的概率越大.



(*a*) 不同用户的兴趣分布情况及局部放大图 (*b*) 不同用户的兴趣拥有量

(*c*) 领域专家的兴趣分布图 (*d*) 普通用户的兴趣分布图

图8 用户兴趣分析

如果用户拥有极少或单一兴趣,将导致图 8(*a*)中某用户对应图像中出现大面积白色区域,该用户可能由于网络行为不活跃造成出现类似专家图型,因此不能将其归为领域专家,需通过判定活跃用户以准确删减不满足条件的伪专家用户.图 8(*b*)为用户涉及兴趣领域数量,分析 MovieLens 网络数据发现用户兴趣关注数量位于[10,20]区间的用户最多.不同的网络数据特征和用户行为导致不同网络中用户拥有兴趣均值不同,因此处理网络数据需要分析其固有特征以确定阈值设定方法.

由 PreferEF 发现 MovieLens 中的专家节点和普通节点分别如图 8(*c*)和图 8(*d*)所示,图中比率为 1 的点表示用户关注度最高的兴趣,其他节点为关注度较弱的兴趣,经过比较图 8(*c*)和图 8(*d*)可知领域专家的不同兴趣关注差异度较大,关注度高的兴趣与关注度低的兴趣被明显区分.普通用户兴趣分布相对平均,不能发现兴趣关注度间的明显差异,因此不存在高关注度兴趣领域.

参考文献

[1] Ping Liu, Dan He, Kan Liu. Construction of experts network based on co-word analysis[A]. Proceedings of the IEEE International Conference on Computer Science and Service System [C]. Nanjing, China: IEEE Press, 2011. 2163 – 2166.

[2] Kai-Hsiang Yang, Chun-Yu Chen, Hahn-Ming Lee et al. EFS: Expert finding system based on Wikipedia link pattern analysis [A]. Proceedings of the IEEE International Conference on Systems, Man and Cybernetics[C]. Singapore: IEEE Press, 2008. 631 – 635.

[3] G Alan Wang, Jian Jiao, Alan Abrahams, et al. ExpertRank: A topic-aware expertfinding algorithm for online knowledge communities[J]. Decision Support Systems, 2013, 54(3): 1442 – 1451.

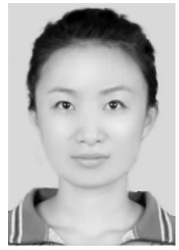
[4] Clara Higuera, Gonzalo Pajares, Javier Tamames et al. Expert system for clustering prokaryotic species by their metabolic features[J]. Expert Systems with Applications, 2013, 40(15): 6185 – 6194.

[5] Farnoush Farhadi, Elham Hoseini, Sattar Hashemi et al. TeamFinder: A co-clustering based framework for finding an effective team of experts in social networks[A]. Proceedings of the IEEE 12th International Conference on Data Mining Workshops[C]. Brussels, Belgium: IEEE Press, 2012. 107 – 114.

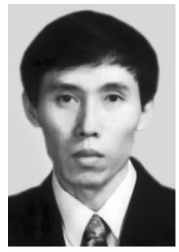
[6] Yi Fang, Luo Si, Aditya P Mathur. Discriminative models of integrating document evidence and document-candidate associations for expert search[A]. Proceedings of the 33rd international conference on research and development in information retrieval (SIGIR'10)[C]. Geneva, Switzerland: SIGIR, 2010. 683

– 690.
[7] Catarina Moreira, Andreas Wichert. Finding academic experts on a multisensor approach using Shannon's entropy[J]. Expert Systems with Applications, 2013, 40(14): 5740 – 5754.
[8] Max O. Lorenz. Methods of measuring the concentration of wealth[J]. Publications of the American Statistical Association, 1905, 9(70): 209 – 219.
[9] Lujun Fang, Alex Fabrikant, Kristen LeFevre. Look Who I Found: Understanding the effects of sharing curated friend groups[A]. Proceedings of the 3rd Annual ACM Web Science Conference[C]. Evanston, USA: ACM, 2012. 95 – 104.
[10] Jun-Ming Xu, Aniruddha Bhargava, Robert Nowak et al. Socioscope: spatio-temporal signal recovery from social media (extended abstract)[A]. Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI)[C]. Beijing, China, 2013. 3096 – 3100.
[11] J Zhang, M S Ackerman, L Adamic. Expertise networks in online communities: structure and algorithms[A]. Proceedings of the 16th International World Wide Web Conference[C]. Banff, Canada: WWW, 2007. 221 – 225.
[12] K Balog, L Azzopardi, M de Rijke. Formal models for expertfinding in enterprise corpora[A]. Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. Seattle, USA: SIGIR, 2006. 43 – 50.

作者简介



国　琳　女, 1987 年出生, 吉林吉林人, 2013 年至今于吉林大学计算机学院攻读博士学位, 从事社会化网络、知识挖据及搜索引擎有关研究.
E-mail: gl_personal@sina.com



左万利(通信作者)　男, 1957 年出生, 吉林吉林人, 博士, 现为吉林大学计算机科学与技术学院教授、博士生导师, ACM 职业会员, 从事数据库、Web 智能、网络搜索引擎、自然语言处理等有关研究.
E-mail: wanli@jlu.edu.cn