微博用户的相似性度量及其应用

徐志明 李 栋 刘 挺 李 生 王 刚 袁树仑

(哈尔滨工业大学计算机科学与技术学院 哈尔滨 150006)

摘 要 微博用户的兴趣分析和模型表示是用户关系分析的基础,而用户关系分析又构成了微博社会网络的生成和分析的基础.该文主要讨论微博的用户关系分析技术.作者将微博社会网络视为一个加权无向图,节点表示用户,边表示用户之间的关系,边的权值表示用户之间的关系强度.该文将用户关系强度定义为用户之间的相似度,分别给出了基于各种用户属性信息(背景信息、微博文本、社交信息)的用户相似度计算方法,并通过实验系统性对比了上述方法的优劣.实验结果显示:基于社交信息的用户相似度在用户关系分析方面取得了最好的效果.为了进一步验证上述用户相似度的实际性能,该文将它们应用于用户推荐的相关实验,基于社交信息的用户相似度又取得了最好的推荐效果.最后,该文应用基于社交信息的用户相似度生成了微博的社会网络(称作用户相似性网络),在该社会网络上进行了团体挖掘的实验,实验结果显示了该相似度在团体挖掘上的有效性.

关键词 微博;社会网络;用户相似度;团体挖掘;用户推荐

中图法分类号 TP393 **DOI 号** 10.3724/SP.J.1016.2014.00207

Measuring Similarity between Microblog Users and Its Application

XU Zhi-Ming LI Dong LIU Ting LI Sheng WANG Gang YUAN Shu-Lun

(*School of Computer Science and Technology, Harbin Institute of Technology, Harbin* 150006)

Abstract Analyzing user interest and building user profile is very important for microblog's user relationship analysis, which is the fundamental work for social network formation and analysis. This paper mainly discusses approaches of microblog's user relationship analysis. We view microblog's social network as a weighted undirected graph, where users are treated as nodes linked by edges, and the weights of edges mean the relationship strengths between users. This paper defines user relationship strength as user similarity, and proposes several user similarity estimation approaches by the use of various attribute information of users such as background information, tweets and social information respectively and systematically investigated them by experiments, the experimental results showed that social information-based user similarity achieved the best performance. In addition, we tested them in user recommending experiments, and social information-based user similarity also got the best recommending results. Finally we applied social information-based user similarity to generate microblog's social network, called as user similarity network, on which we conducted community mining experiments, the experimental results showed our approach is of remarkable performance.

Keywords microblog; social network; user similarity; community mining; user recommendation

收稿日期:2011-07-25;最终修改稿收到日期:2013-10-25.本课题得到国家自然科学基金(61173074,60736044)资助.**徐志明**,男,1967 年生,博士,教授,主要研究领域为社会计算、自然语言处理.E-mail:xuzm@hit.edu.cn.**李 栋**,男,1987 年生,博士研究生,主要研究方向为社会计算、信息扩散.**刘 挺**,男,1972 年生,博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为自然语言处理、社会计算.**李 生**,男,1943 年生,博士,教授,主要研究领域为自然语言处理、信息检索.**王 刚**,男,1988 年生,硕士研究生,主要研究方向为社会计算.**袁树仑**,男,1987 年生,硕士研究生,主要研究方向为社会计算.

1 引 言

作为一种在线交互媒体,社会媒体(Social Media)的大量用户组成了虚拟网络社区,允许用户在线交流、发布自身的信息,并支持群体用户协作编辑、分享、传播信息.近年来社会媒体呈现多样化的发展趋势,从早期的论坛、博客、播客、维基到风头正劲的社交网站、微博,各种社会媒体的内在结构呈现了社会网络的特性,社会网络的分析技术(Social Network Analysis)正在成为网络技术研究的热点和趋势[1-2].

作为一种社会媒体,微博媒体完美地仿造了人类的社会结构,将大量的用户组织成社会网络,满足了用户的信息个性化发布、社会性传播和社交的需求.另外,微博媒体具有超强的信息传播特性[3],微博信息可在社会网络上呈网状交叉扩散,随着转发次数的增多,传播速度呈指数性增长,相当于病毒式信息扩散的速度.

对于传统的 Web1.0 的网络媒体,网站编辑人员发布的信息是用户获取信息的主要来源,搜索引擎成为人们查询网络信息的主要工具.而微博媒体正在改变着这一切,搜索引擎不再是人们获取信息的唯一工具,微博媒体正在成为人们发布信息、获取信息的主要工具.在微博媒体上,每个用户节点相当于一个信息频道,他可以自由发布自身信息,也可根据自身兴趣关注一些感兴趣的相关人物,建立起自己的社会关系,据此来接收他所关注的人物频道发布的信息.但是由于微博媒体往往拥有数以亿计的用户节点,当用户在建立自己的社会关系时,会面临着数据过载问题,因此帮助用户在大量的人群节点中发现其感兴趣的人群是非常重要的,而在线社区用户推荐就是一个有效的工具.在线社区用户推荐就是根据社会网络及其网络用户发布的信息,为目标用户推荐其感兴趣的潜在相关用户.在线社区用户推荐具有广阔的应用前景.站在用户的角度看,它可以帮助用户构建起社会关系,使用户获得更多感兴趣的信息;站在社会媒体的角度看,它扩大、增强了用户之间的交互性.社会网络的分析技术是在线社区用户推荐问题的基础,也是本文的研究重点.本文将微博媒体的社会网络分析技术分成以下 3 层:

(1)节点分析.在微博社会网络中,每个节点代表一个用户.节点分析相当于对用户的兴趣分析和模型表示.具体分为几个步骤:① 获取每个用户节点的相关信息;② 从用户的相关信息中,提取用户的兴趣特征;③ 选择合适的模型表示用户,建立用户模型(User Profile)[4-6].

(2)关系分析.微博社会网络可被视为一个加权无向图,每个边表示两个用户之间的关系,边的权值表示它们之间的关系强度.微博用户关系分析的目的是:根据两个用户之间的相关信息,计算它们之间的关系强度[7].

(3)网络分析.微博社会网络的分析对象是网络的拓扑结构,相关的研究集中在:团体挖掘(发现用户的社交圈)[8-9]、人物影响力计算[10-12]、信息传播[12-14]等问题.

针对社会网络的用户关系分析问题[15-16],学者们开展了大量的相关研究.其中,一些学者利用用户相似度、网络拓扑结构的分析技术来计算用户关系强度[17-18],并应用于链接预测问题[16](Link Prediction);Kahanda 等人[19]利用用户之间的交互性来度量用户关系强度;Xiang 等人[7]融合了用户相似度和用户之间的交互性,来计算用户关系强度;另外,用户关系分析经常也被用于好友推荐[20-22].

本文主要研究了微博社会网络中的用户关系分析技术.针对用户信息的模型表示问题,分别给出了用户的几种属性信息(背景信息、微博文本、社交信息、交互信息)的模型表示方法,在此基础上,完成了用户的整体表示;本文将用户关系分析问题视为用户之间的相似度计算问题,并提出了基于各种用户属性信息的用户相似度计算方法,然后通过系统性的实验,考察了它们的性能.

为了更深入地验证本文提出的用户相似度计算方法,本文首先将基于各种用户属性信息的用户相似度进行了用户推荐的实验,系统性地考察了它们在用户推荐问题上的表现;然后,本文应用基于社交信息的用户相似度生成微博的用户相似性网络,在该网络上进行了团体挖掘的实验,实验结果显示了基于社交信息的用户相似度方法的有效性.

本文第 2 节阐述微博用户关系分析的总体研究思路;首先详细描述微博用户信息的采集、兴趣分析与模型表示方法;然后给出基于各种用户属性信息的用户相似度计算方法;第 3 节详细描述各种微博用户相似度算法的实验结果与分析;最后给出本文的结论以及未来的工作.

2 微博用户关系分析

该部分主要阐述微博用户关系分析的技术,包

括可分为几个部分：数据获取、节点分析、关系分析、网络分析、信息推荐. 将它们组合起来，形成一个微博用户关系分析的技术平台（如图 1 所示），本文下面分别详细介绍各个部分的工作原理.

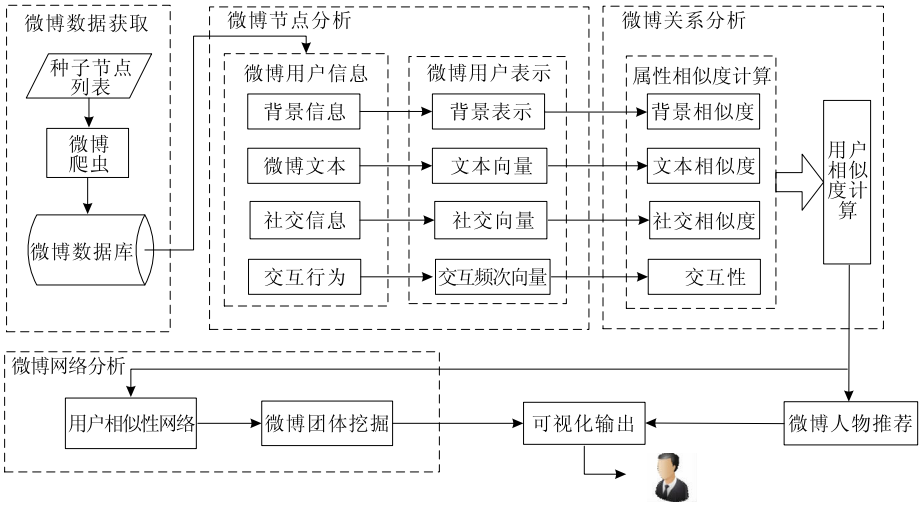


图 1　微博用户关系分析的技术平台

2.1　微博数据获取

该部分根据新浪微博开放平台的 API 接口，设计了一个微博爬虫算法. 它选择一组微博用户作为种子节点，利用雪球采样策略采集一组微博用户的个人数据，作为本文的实验数据. 主要思想如下.

(1) 选择一组微博用户 $\{ID_1, ID_2, \cdots, ID_m\}$ 作为种子节点，加入待爬行节点队列 Q.

(2) 如果 $Q=$NULL 或超过阈值（预设的爬行时间或扩展层数），则退出；否则从 Q 中取出一个用户 ID_k.

(3) 利用新浪微博 API 访问函数，抓取该用户节点 ID_k 的个人信息，将用户的背景信息（位置信息、标签信息、个人描述）、社交信息（关注信息、粉丝信息）、微博文本、交互信息（转发信息、评论信息），分别存入微博用户信息数据库.

(4) 扩展该节点 ID_k，将其邻居节点（ID_k 的关注列表中的全部用户 ID）加入 Q，转到(2).

2.2　微博节点分析

该部分讨论微博用户信息的模型表示方法. 对于给定的一个用户 u，其用户信息包含 4 种属性信息（背景信息、微博文本、社交信息、交互信息），因此 u 的模型表示问题可分解为 4 种属性信息表示问题. 即 $Profile(u)=\{Background(u), \boldsymbol{Tweet}(u), Relation(u), Interaction(u)\}$，具体说明如下：

(1) $Background(u)$：表示 u 的背景信息，包含 u 的 3 种属性信息（位置信息、标签信息、个人描述），均是短文本，可表示为字符串. $Background(u)=\{Place(u), Tag(u), Introduction(u)\}$.

(2) $\boldsymbol{Tweet}(u)$：表示 u 发布的全部微博所拼接成的长文本. 本文将其表示为一个文本向量. 过程如下：

文本预处理. 对 $\boldsymbol{Tweet}(u)$ 进行分词、停用词过滤、词性标注等处理；

特征提取. 采用信息增益的特征选择算法提取 $\boldsymbol{Tweet}(u)$ 的特征词，对文本进行降维处理；

权重计算. $\boldsymbol{Tweet}(u)$ 中的每个特征词 i 的权重 w_i，本文采用 $tf*idf$ 方法来计算，即 $w_i=tf_i(\boldsymbol{Tweet}(u))\times\log N/n_i$，其中 $tf_i(\boldsymbol{Tweet}(u))$ 表示特征词 i 在 $\boldsymbol{Tweet}(u)$ 中的频率，$\log N/n_i$ 为特征词 i 的逆文档频率.

向量表示. $\boldsymbol{Tweet}(u)=(w_1, w_2, \cdots, w_n)$，其中 w_i 为微博文本的某个特征词 i 的权重.

(3) $Relation(u)$：表示 u 的社交信息，包括两种属性信息（关注信息、粉丝信息），本文将它们分别表示为两个向量：关注向量 $\boldsymbol{Followee}(u)$、粉丝向量 $\boldsymbol{Follower}(u)$，则 $Relation(u)=\{\boldsymbol{Followee}(u), \boldsymbol{Follower}(u)\}$. 具体方法是：将所有用户编号 $\{0,1,2,\cdots,n\}$，若用户 u 关注了编号为 i 的用户，则 $\boldsymbol{Followee}(u)$ 的第 i 个分量为 1，否则为 0；同理，如果编号为 i 的用户关注了用户 u，则 $\boldsymbol{Follower}(u)$ 的第 i 个分量为 1，否则为 0.

(4) $Interaction(u)$：表示 u 的交互信息，包括两种属性信息（转发信息、评论信息）. 本文将它们分别表示为两个向量：转发向量 $\boldsymbol{Retweet}(u)$、评论向量 $\boldsymbol{Comment}(u)$. $Interaction(u)=\{\boldsymbol{Retweet}(u), \boldsymbol{Comment}(u)\}$. 具体方法是：将所有用户编号 $\{0,1,$

$2,\cdots,n\}$，$\boldsymbol{Retweet}(u)$的第 i 个分量表示用户 u 转发用户 i 的微博次数；$\boldsymbol{Comment}(u)$的第 i 个分量表示用户 u 评论用户 i 的微博次数.

2.3 微博用户关系分析

这部分分析微博用户之间的关系强度. 对于两个微博用户(u,v)，他们的关系强度 $strength(u,v)$，可以根据他们的相似度 $sim(u,v)$或他们的交互性 $Interaction(u,v)$度量. 其中，$sim(u,v)$更能揭示(u,v)之间潜在的兴趣相似性，而 $Interaction(u,v)$一般只能反映已经建立起关注关系的用户之间的交互程度. 因此从挖掘用户潜在的兴趣相关性的角度看，对于信息推荐来说，$sim(u,v)$更有应用价值. 因此，本文将用户关系强度视为用户之间的相似度，$strength(u,v)=sim(u,v)$，利用 $sim(u,v)$来分析用户关系. 对于(u,v)，$Profile(u)=\{Background(u),\boldsymbol{Tweet}(u),Relation(u),Interaction(u)\}$，$Profile(v)=\{Background(v),\boldsymbol{Tweet}(v),Relation(v),Interaction(v)\}$. 本文根据$(u,v)$之间的前 3 种属性相似度来计算$(u,v)$之间的相似度 $sim(u,v)$. 具体方法是：先求解(u,v)的各个属性相似度，然后对它们进行加权融合，来计算 $sim(u,v)$. 即

$$\begin{aligned}sim(u,v)=&\omega_1 sim(Background(u),Background(v))+\\&\omega_2 sim(\boldsymbol{Tweet}(u),\boldsymbol{Tweet}(v))+\\&\omega_3 sim(Relation(u),Relation(v))\end{aligned}\tag{1}$$

其中，ω_i为各个属性相似度的权重，$\omega_1+\omega_2+\omega_3=1$. 各个属性相似度的计算描述如下.

2.3.1 微博用户的背景信息相似度计算

对于两个用户(u,v)，他们的背景信息包含 3 种属性信息(位置信息、标签信息、个人描述)，分别表示为 $Background(u)=\{Place(u),Tag(u),Introduction(u)\}$，$Background(v)=\{Place(v),Tag(v),Introduction(v)\}$. 其背景信息相似度 $sim(Background(u),Background(v))$，可以根据$(u,v)$之间的各个属性相似度而计算：

$$\begin{aligned}&sim(Background(u),Background(v))=\\&\omega_1 sim(Place(u),Place(v))+\\&\omega_2 sim(Tag(u),Tag(v))+\\&\omega_3 sim(Introduction(u),Introduction(v))\end{aligned}\tag{2}$$

其中 ω_i为各个属性相似度的权重，$\omega_1+\omega_2+\omega_3=1$. 式(2)中右边各个属性相似度计算方式如下.

(1) $sim(Place(u),Place(v))$：表示(u,v)之间的位置信息相似度. 因为新浪微博的位置信息是“省/市”结构，所以本文采用分层比较的方法来计算 $sim(Place(u),Place(v))$. 若 $Place(u)$和 $Place(v)$的省、市信息均相同，则 $sim(Place(u),Place(v))=1$；若省信息相同，市信息不同，则 $sim(Place(u),Place(v))=2/3$；若省、市信息均不同，则为 0.

(2) $sim(Tag(u),Tag(v))$：表示(u,v)之间的标签相似度，本文采用编辑距离方法计算[23]. 因为，编辑距离常用于评价两个字符串间的相似度，也可用于计算两个自然语言语句的相似度. 编辑距离反映了两个字符串的绝对差异，而相似度以一个$[0,1]$之间的数值反映两个字符串的相似程度，数值越大表示相似程度越高. 令 $Tag(u)$、$Tag(v)$的编辑距离为 $Distance(Tag(u),Tag(v))$，$length(x)$表示 x 的长度，则 $Tag(u)$、$Tag(v)$之间的标签相似度的计算公式如下：

$$\begin{aligned}&sim(Tag(u),Tag(v))=\\&1-Distance(Tag(u),Tag(v))/\\&\max(Length(Tag(u)),Length(Tag(v)))\end{aligned}\tag{3}$$

(3) $sim(Introduction(u),Introduction(v))$：表示$(u,v)$之间的个人描述相似度. 其计算方法与标签相似度相同. 具体的公式为

$$\begin{aligned}&sim(Introduction(u),Introduction(v))=\\&1-Distance(Introduction(u),Introduction(v))/\\&\max(Length(Introduction(u)),\\&Length(Introduction(v)))\end{aligned}\tag{4}$$

2.3.2 微博用户的文本相似度计算

对于两个用户(u,v)，它们的微博文本可分别表示为两个文本特征向量，$\boldsymbol{Tweet}(u)=(w_{u1},w_{u2},\cdots,w_{un})$，$\boldsymbol{Tweet}(v)=(w_{v1},w_{v2},\cdots,w_{vn})$. (u,v)之间的微博文本相似度 $sim(\boldsymbol{Tweet}(u),\boldsymbol{Tweet}(v))$的计算公式如下：

$$sim(\boldsymbol{Tweet}(u),\boldsymbol{Tweet}(v))=\frac{\sum_{i=1}^{n}(w_{ui}*w_{vi})}{\sqrt{\sum_{i=1}^{n}w_{ui}^2}\sqrt{\sum_{i=1}^{n}w_{vi}^2}}\tag{5}$$

2.3.3 微博用户的社交信息相似度计算

对于两个用户(u,v)，他们的社交信息分别表示为 $Relation(u)=\{\boldsymbol{Followee}(u),\boldsymbol{Follower}(u)\}$，$Relation(v)=\{\boldsymbol{Followee}(v),\boldsymbol{Follower}(v)\}$. 因此$(u,v)$之间的社交信息相似度的计算问题可以分解为两种属性信息(关注信息、粉丝信息)的相似度计算问题. 本文先采用了余弦相似度方法来计算(u,v)之间的两种属性信息相似度[24]，(u,v)的关注信息相似度、粉丝信息相似度的计算公式分别为

$$\begin{aligned}&sim(\boldsymbol{Followee}(u),\boldsymbol{Followee}(v))=\\&(\boldsymbol{Followee}(u)\cdot\boldsymbol{Followee}(v))/\end{aligned}$$

$$(\| \boldsymbol{Followee}(u) \| * \| \boldsymbol{Followee}(v) \|) \quad (6)$$

$$sim(\boldsymbol{Follower}(u), \boldsymbol{Follower}(v)) = (\boldsymbol{Follower}(u) \cdot \boldsymbol{Follower}(v)) / (\| \boldsymbol{Follower}(u) \| * \| \boldsymbol{Follower}(v) \|) \quad (7)$$

另外，本文给出了另外一组(u,v)之间的关注信息相似度、粉丝信息相似度的计算公式，分别为

$$sim(\boldsymbol{Followee}(u), \boldsymbol{Followee}(v)) = (\boldsymbol{Followee}(u) \cdot \boldsymbol{Followee}(v)) / (\log \| \boldsymbol{Followee}(u) \|^2 * \log \| \boldsymbol{Followee}(v) \|^2) \quad (8)$$

$$sim(\boldsymbol{Follower}(u), \boldsymbol{Follower}(v)) = (\boldsymbol{Follower}(u) \cdot \boldsymbol{Follower}(v)) / (\log \| \boldsymbol{Follower}(u) \|^2 * \log \| \boldsymbol{Follower}(v) \|^2) \quad (9)$$

为了保证式(8)和(9)的计算结果处在$[0,1]$区间，本文对它们进行了正规化处理.

$$sim(A,B)' = (sim(A,B) - MIN)/(MAX - MIN) \quad (10)$$

其中，MAX是所有用户之间社交信息相似度的最大值，MIN是最小值.

上述的微博用户的关系信息相似度、粉丝信息相似度，即可以单独作为社交信息相似度使用，也可以将它们加权融合，来计算用户之间的社交信息相似度：

$$sim(Relation(u), Relation(v)) = \omega_1 * sim(\boldsymbol{Followee}(u), \boldsymbol{Followee}(v)) + \omega_2 * sim(\boldsymbol{Follower}(u), \boldsymbol{Follower}(v)) \quad (11)$$

其中，ω_i是关注信息相似度、粉丝信息相似度的权值，$\omega_1 + \omega_2 = 1$.

2.3.4　微博用户的交互性计算

对于两个用户(u,v)，他们之间的交互信息分别表示为$Interactive(u) = \{\boldsymbol{Retweet}(u), \boldsymbol{Comment}(u)\}$，$Interactive(v) = \{\boldsymbol{Retweet}(v), \boldsymbol{Comment}(v)\}$. 将所有用户编号$\{0,1,2,\cdots,n\}$，设$(u,v)$的编号分别为$(i,j)$. $\boldsymbol{Retweet}(u)$的第j个分量表示用户u对用户v的微博转发次数，记为$Retreet_num(u->v)$；$\boldsymbol{Comment}(u)$的第j个分量表示用户u对用户v的微博评论次数，记为$Comment_num(u->v)$. 同理，$\boldsymbol{Retweet}(v)$的第i个分量表示用户v对用户u的微博转发次数，记为$Retreet_num(v->u)$；$\boldsymbol{Comment}(v)$的第i个分量表示v对u的微博评论次数$Comment_num(v->u)$. 本文给出了(u,v)之间的交互性计算公式：

$$Interactive(u,v) = \omega_1 * \frac{Retreet\_num(u->v) + Retreet\_num(v->u)}{2} + \omega_2 * \frac{Comment\_num(u->v) + Comment\_num(v->u)}{2} \quad (12)$$

其中，ω_i是转发行为、评论行为的权值，$\omega_1 + \omega_2 = 1$.

3　实验及分析

为了测试本文提出的几种微博用户的相似度计算方法，本文开展了相关的实验. 实验包括3个方面：(1)系统地对比几种用户属性信息的相似度计算方法的优劣；(2)应用基于各种用户属性信息的用户相似度开展了用户推荐的相关实验，以便考察它们在用户推荐方面的性能；(3)应用基于社交信息的用户相似度，生成微博的用户相似性网络，在该网络上进行了团体挖掘的实验；下面给出了详细的实验介绍.

实验数据：

对于微博用户关系分析问题来说，尚缺少公开的中文标准数据集. 本文的实验数据来自于自行开发的微博爬虫所搜集到的一组新浪微博的用户数据(采集时间是2011年10月20日)，主要是由国内互联网高管领域相关用户的社交圈组成. 采集方法是：首先从新浪微博名人堂互联网高管领域中挑选50个用户作为种子节点，利用雪球采样的爬行策略，顺着种子节点的关系链向外爬行扩展一层，得到本文实验1、2研究的用户集合U(共157 812人)以及它们的个人信息(157 812人)以及它们的个人背景信息(157 812条)、社交信息(11 684 003条)、微博文本(5 712 552条)、转发与评论信息(5 415 110条).

3.1　实验1：微博用户的属性相似度的对比实验

评价策略：

在实验数据的用户集合U上，本文对微博用户之间的各种属性信息(背景信息、微博文本、社交信息)的相似度计算方法进行实验测试. 对于每个待测的用户相似度计算方法的测试方法，具体如下：

(1)对于每个用户$u \in U$，利用用户的交互性计算公式(12)，计算u与$\boldsymbol{Followee}(u)$中每个元素v之间的交互性，然后按着交互性对$\boldsymbol{Followee}(u)$中的元素进行降序排列，得到u的最关注的用户有序集，称作u的关注序列$F(u)$.

(2)对于每个用户$u \in U$，利用待测的用户相似度算法，计算u与$U-\{u\}$中的每个元素v之间的相似度，然后按照相似度对$U-\{u\}$中的所有元素进行降序排列，得到一个与u最相似的用户有序集，称作u的相似序列$S(u)$.

本文将 u 的相似序列 $S(u)$ 视为待测结果，将 u 的关注序列 $F(u)$ 视为标准答案. 通过对比 $F(u)$、$S(u)$ 的差异，来评价所使用的用户相似度算法的性能. 将 $S(u)$、$F(u)$ 的前 N 个元素组成的子集分别记为 $S_1^N(u)$、$F_1^N(u)$. 本文借用信息检索领域的两个评价指标：$P@N$ 和排序准确率(Accuracy)比较两个序列 $S(u)$、$F(u)$ 之间的差异，从而实现对用户相似度算法进行评价，其中 $P@N$ 的计算公式如下：

$$P@N=\frac{1}{|U|}\sum_{u\in U}\frac{S_1^N(u)\bigcap F_1^N(u)}{N} \tag{13}$$

在实验中，N 分别取 3、5、10、15、20、25、30、35、40、45、50. 在 $P@N$ 的计算公式中，没考虑 $S_1^N(u)$、$F_1^N(u)$ 之间的元素的排列次序的因素，本文采用排序准确率(Accuracy)指标对用户相似度算法继续评价，其计算公式如下：

$$Accuracy=\frac{1}{|U|}\sum_{u\in U}\frac{1}{|F(u)|}\sum_{i=1}^{|F(u)|}\frac{1}{1+|S(u)\_rank_i-F(u)\_rank_i|} \tag{14}$$

式(14)中，对于关注序列 $F(u)$ 中的每个用户 i，它在 $F(u)$、$S(u)$ 中出现的次序位置分别记为 $F(u)\_rank_i$、$S(u)\_rank_i$.

实验设置：

本实验利用用户之间的交互性所产生的关注序列，比对各种用户之间的相似度所产生的相似序列，来评价相应的用户相似度计算方法的优劣. 其中，用户(u,v)之间的属性相似度、交互性计算公式中的权重采用层次判别矩阵计算得出，权重设置如表 1 所示. 表 1 中的权值计算所使用的层次判别矩阵如表 2 所示.

表 1 实验中的权重设置

公式/注释	权重分配		
$Interactive(u,v)$ 交互性的计算公式(12)	转发信息 $\omega_1=0.25$	评价信息 $\omega_2=0.75$	
$sim(Relation(u),Relation(v))$ 社交信息相似度的计算公式(11)	关注信息 $\omega_1=0.167$	粉丝信息 $\omega_2=0.833$	
$sim(Background(u),Background(v))$ 背景信息相似度的计算公式(2)	位置信息 $\omega_1=0.297$	标签信息 $\omega_2=0.539$	个人描述 $\omega_3=0.164$
$sim(Profile(u),Profile(v))$ 用户相似度的计算公式(1)	背景信息 $\omega_1=0.139$	社交信息 $\omega_2=0.794$	微博文本 $\omega_3=0.067$

表 2 实验中计算权重的判别矩阵

(a) 交互信息的判别矩阵

交互信息	转发信息	评价信息
转发信息	1	1/3
评论信息	3	1

(b) 社交信息的判别矩阵

社交信息	关注信息	粉丝信息
关注信息	1	1/5
粉丝信息	5	1

(c) 背景信息的判别矩阵

背景信息	位置信息	标签信息	个人描述
地理位置	1	1/2	2
个人标签	2	1	3
个人描述	1/2	1/3	1

(d) 用户信息的判别矩阵

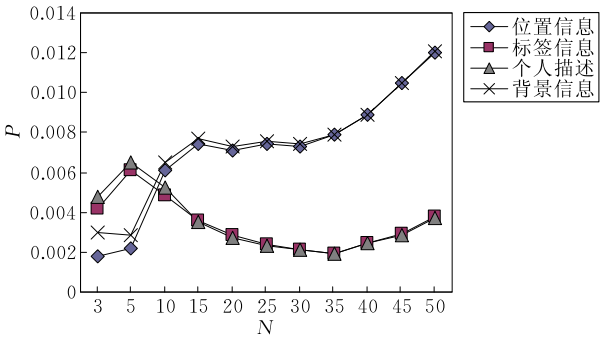
用户信息	背景信息	社交信息	微博文本
资料信息	1	1/9	3
社交信息	9	1	9
微博文本	1/3	1/9	1

(1) 微博用户的背景信息相似度的实验结果.

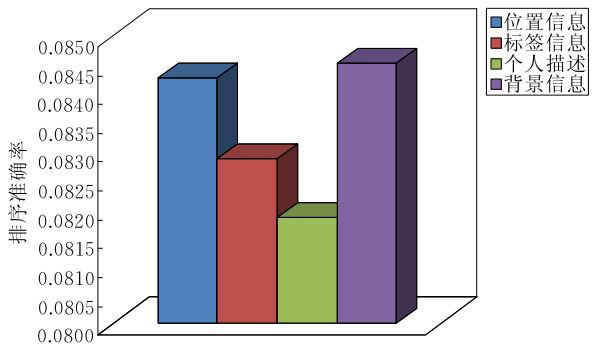
该实验的动机是考察微博用户的背景信息及其 3 种属性信息(位置信息、标签信息、个人描述)对用户相似度的影响. 实验内容：首先分别计算用户的位置信息相似度、标签信息相似度、个人描述相似度，然后将它们加权融合，得到用户的背景信息相似度. 图 2(a)和(b)分别给出了上述各种相似度在 $P@N$ 和排序准确率上的实验结果. 实验结果显示：用户的背景信息的 3 种属性信息相比，位置信息相似度的实验结果最好；综合地看，3 种属性信息相似度加权融合而成的用户的背景信息相似度的实验结果最好.

(2) 微博用户的社交信息相似度的实验结果.

该实验的动机是考察微博用户的社交信息及其两种属性信息(关注信息、粉丝信息)在计算用户相似度上的性能. 实验内容：M_1 表示一组面向社交信息的对数相似度(式(8)、(9))，而 M_2 表示一组面向社交信息的余弦相似度(式(6)、(7)). 首先分别利用 M_1 和 M_2 计算用户的关注信息相似度、粉丝信息相似度，得到 M_1 的关注信息相似度、M_1 的粉丝信息相似度、M_2 的关注信息相似度、M_2 的粉丝信息相似度；采用式(11)分别将 M_1(或 M_2)计算的关注信息

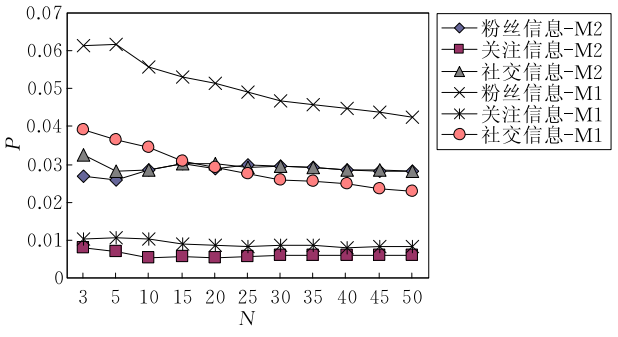


(a) $P@N$的实验结果

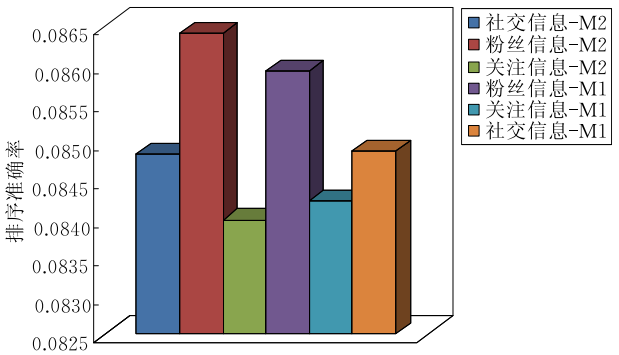


(b) 排序准确率的实验结果

图 2　用户背景信息相似度的实验结果



(a) $P@N$的实验结果



(b) 排序准确率的实验结果

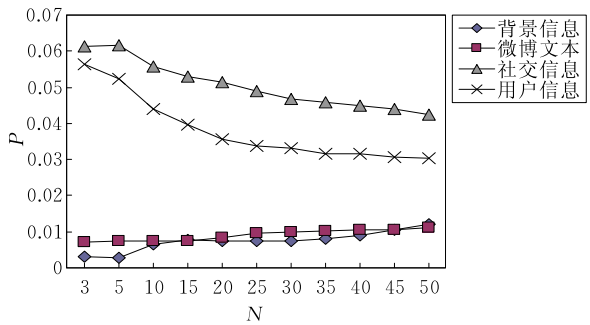
图 3　社交信息相似度的实验结果

相似度、粉丝信息相似度加权融合，得到用户的 M_1-社交信息相似度(或 M_2-社交信息相似度). 图 3(a)和(b)分别给出了上述各种相似度在 $P@N$ 和排序准确率指标上的实验结果. 实验结果显示：对于用户相似度来说，M_1 的粉丝信息相似度的 $P@N$ 最好，而 M_2 的粉丝信息相似度的排序准确率最高.

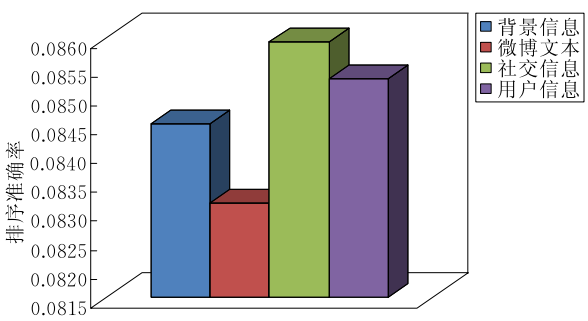
(3) 微博用户相似度的实验结果.

该实验的动机是考察用户信息及其 3 种属性信息(背景信息、微博文本、社交信息)在计算用户相似度上的性能. 实验内容：首先计算微博文本相似度、社交信息相似度、背景信息相似度，然后对它们加权融合，得到用户的整体相似度. 图 4(a)和(b)给出了上述各种相似度在 $P@N$ 和排序准确率上的实验结果. 实验结果显示：对于用户相似度来说，用户信息的 3 种属性信息相比，社交信息相似度的 $P@N$ 和排序准确率实验结果均最好，而用户的整体相似度均次之.

通过上述实验结果的观察，可见当计算两个用户之间的相似度时，用户的结构性信息(社交信息)所起的作用较大，而用户的文本性信息(背景信息、微博文本)所起的作用较小. 可能的原因是对于一个微博人物来说，他的个人信息(背景信息、社交信息、微博文本、交互信息)可分为两类：结构性信息(社交信息)与文本性信息(背景信息、微博文本、交互信息)①.



(a) $P@N$的实验结果



(b) 排序准确率的实验结果

图 4　用户相似度的实验结果

(1) 用户的文本性信息的特点.

① 用户的背景信息的不完整性. 因为用户的背

① 用户的背景信息、微博文本、交互信息内部，也带有一定的结构性，但它们最主要的特征还是短文本. 其中：对于交互性信息不在讨论之中，因为本文将其用于计算用户之间的交互性，没有用来计算用户之间的相似度.

景信息涉及用户隐私，很多用户没填写完整的背景信息，这影响了它们刻画用户特征的能力.

② 用户发布的微博常常带有很大的随意性，含有大量的噪音信息，不能完全精确地反映用户的兴趣意图.

③ 上述的背景信息、微博文本均为短文本信息，抽取用户兴趣特征有一定难度.

(2) 用户的结构性信息的特点.

① 与用户的文本性信息相反，用户的结构性社交信息反映了用户的社会关系，决定着用户们愿意接收的人物频道信息，因此被用户精心地维护、选择，因此社交信息几乎没有噪音信息，可以更准确地反映用户的兴趣.

② 其中，对于用户的社交信息来说，粉丝信息与关注信息相比，前者对度量用户之间的相似度所起的作用更大的原因在于：(i) 用户的关注信息反映了他的兴趣多样性，而不能反映他最本质的兴趣特征，而用户的粉丝信息则更能反映用户最本质的兴趣；(ii) 关注信息往往比粉丝信息少得多，粉丝信息往往比关注信息带有更多的信息量. 基于上述两种原因，两者相比，粉丝信息更能反映人物之间的相似性.

3.2 实验 2：微博用户相似度的用户推荐实验

评价策略：

在实验数据的用户集合 U 上，本文利用上述微博用户相似度计算方法进行了用户推荐的实验，测试这些算法对用户推荐的能力. 对于每个待测的用户相似度计算方法的测试方法如下：

(1) 对于每个用户 $u\in U$，从 ***Followee***(u) 中随机删除 50% 的元素，被删除的元素组成一个集合 $T(u)$.

(2) 利用用户的交互性计算公式(12)，计算 u 与 $T(u)$ 中每个元素 v 之间的交互性，然后按交互性对 $T(u)$ 中的元素进行降序排列，得到 u 的按交互性排序的、被删除的用户有序集，称作 u 的删除序列 $D(u)$.

(3) 对于每个用户 $u\in U$，利用待测的用户相似度算法，计算 u 与 $U-\{u\}-$***Followee***(u) 中的每个元素 v 之间的相似度，然后对 $U-\{u\}-$***Followee***(u) 中的所有元素按该用户相似度进行降序排列，得到与 u 最相似的用户有序集，称作 u 的相似序列 $S(u)$.

本文将 u 的相似序列 $S(u)$ 视为待测结果，将 u 的删除序列 $D(u)$ 视为标准答案. 通过对比 $S(u)$、$D(u)$ 的差异，来评价所使用的用户相似度算法在用户推荐上的性能. 将 $S(u)$、$D(u)$ 的前 N 个元素组成的子集分别记为 $S_1^N(u)$、$D_1^N(u)$. 本文采用两个评价指标：$P@N$ 和 MAP 比较两个序列 $S(u)$、$D(u)$ 之间的差异，从而实现对用户相似度算法进行评价，其中 $P@N$ 的计算公式如下：

$$P@N=\frac{1}{|U|}\sum_{u\in U}\frac{S_1^N(u)\bigcap D_1^N(u)}{N} \tag{15}$$

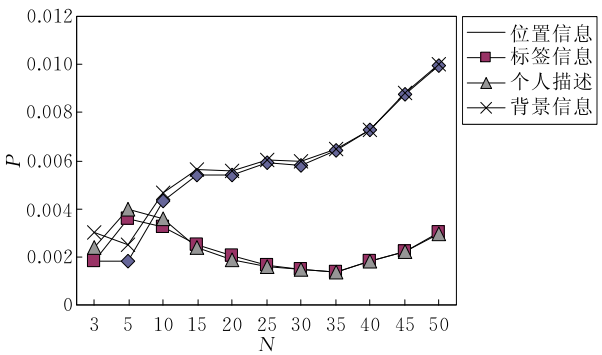
在实验中，N 分别取 3、5、10、15、20、25、30、35、40、45、50. MAP 的计算公式如下：

$$MAP=\frac{1}{|U|}\sum_{u\in U}\left[\frac{1}{|D(u)|}\sum_{i=1}^{|D(u)|}\frac{i}{S(u)\_rank_i}\right] \tag{16}$$

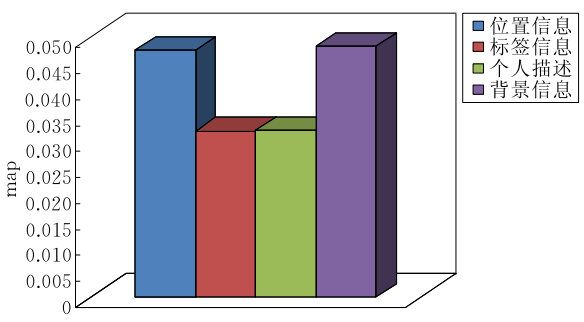
式(16)中，对于删除序列 $D(u)$ 中的每个用户 i，它在 $S(u)$ 中出现的次序位置记为 $S(u)\_rank_i$.

(1) 基于背景信息相似度的用户推荐实验结果

该实验的动机是考察用户的背景信息及其 3 种属性信息(位置信息、标签信息、个人描述)的相似度在用户推荐上的性能. 实验内容：分别应用上述各种相似度，进行用户推荐. 图 5(a)和(b)给出了上述各种相似度在 $P@N$ 和 MAP 上的实验结果. 实验结果显示：对于用户信息推荐来说，用户背景信息的 3 种属性信息相比，位置信息相似度的用户推荐效果最好；综合地看，3 种属性信息加权融合而成的用户背景信息相似度的用户推荐效果最好.



(a) $P@N$的实验结果



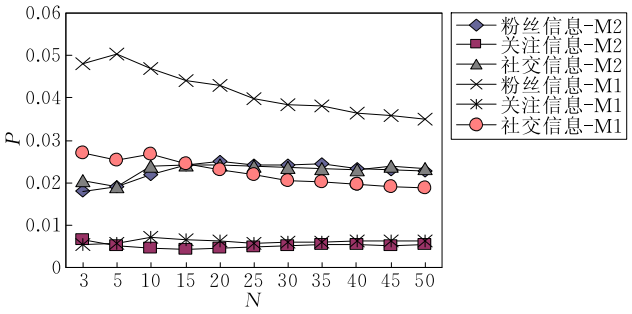
(b) MAP的实验结果

图 5 基于背景信息相似度的用户推荐结果

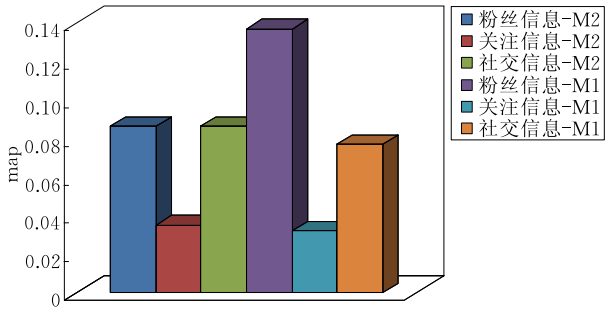
(2) 基于社交信息相似度的用户推荐实验结果

该实验的动机是考察用户的社交信息及其两种

属性信息(关注信息、粉丝信息)的相似度在用户推荐上的性能. 实验内容:分别应用上述各种相似度,进行用户推荐. 图 6(a)和(b)给出了上述相似度在 $P@N$ 和 MAP 指标上的实验结果. 实验结果显示:在 $P@N$ 和 MAP 指标上,M_1 的粉丝信息相似度的用户推荐效果最好.



(a) $P@N$的实验结果



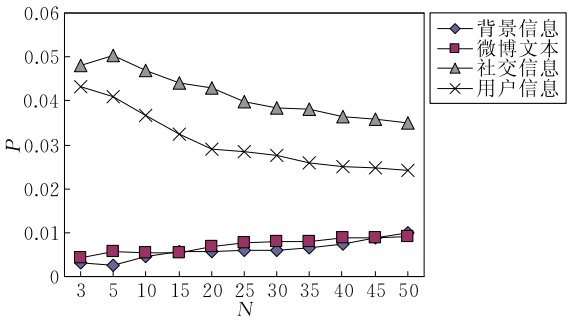
(b) MAP的实验结果

图 6　基于社交信息相似度的用户推荐结果

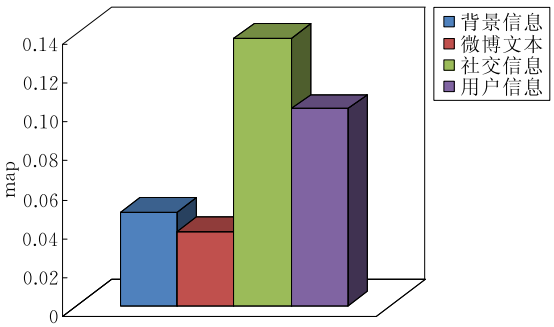
(3) 基于用户相似度的用户推荐实验结果

该实验的动机是考察用户信息及其各个属性信息(背景信息、社交信息、微博文本)的相似度在用户推荐上的性能. 实验内容:分别应用上述各种相似度,进行用户推荐. 图 7(a)和(b)给出了上述各种相似度在用户推荐上的实验结果. 实验结果显示:对于用户推荐来说,用户信息的 3 种属性信息相比,社交信息相似度取得了最好的推荐效果.

上述实验考察了各种用户相似度方法在用户推荐上的性能. 根据实验结果,我们观察到:在用户的各种信息中,用户的结构性社交信息对用户推荐所起的作用较大,而用户的文本性背景信息、微博文本所起的作用较小. 换言之,用户的各种信息对于用户之间的相似度计算问题以及由此开展的用户推荐问题所起的作用具有较为一致的规律,即结构性的社交信息更好地刻画了用户之间的相似特征. 本文根据 M_1 式(9)的用户相似度开发了一个微博用户推荐的应用程序,举例:图 8 是给白硕老师推荐的用户好友,图中深色的边表示推荐的用户,深色边的粗细度表示推荐力度.



(a) $P@N$的实验结果



(b) MAP的实验结果

图 7　基于用户相似度的用户推荐结果

3.3　实验 3:用户相似性网络初步的团体挖掘实验

考虑到实验 1、2 所使用的数据集涉及到的人物节点数目较大,我们难以将它们的团体分析的实验结果在论文中清晰展现. 因此我们选取了新浪微博的 NLP 领域人群作为本文的团体挖掘的实验数据. 最初随意选定了 5 个 NLP 领域的用户(@白硕 sse、@马少平 THU、@计算所王斌、@刘挺 Thomas、@徐志明 Jonathan)作为种子节点,利用雪球采样的爬行策略,顺着种子节点的关系链向外爬行扩展一层,得到实验 3 的用户集合 U(共 646 人).

在此数据集上,首先根据微博人物之间的社交信息相似度式(9),计算任意两个人物之间的相似度,据此生成了人物相似性矩阵,称之为用户相似性网络(它是社会网络的一种特例. 广义上讲,我们可以根据各种人物关系而定义不同的社会网络,例如:根据人物之间的关注关系、交互关系、相似性关系等,可分别构造出关注网络、交互性网络、相似性网络等). 然后通过一个阈值 r 过滤网络中关系较弱的边(在本文中,$r=0.05$),得到一个关系更紧密的用户相似性网络(包括 269 个用户节点、3607 条边). 这样,我们就以微博用户为节点、以用户之间的相似度为边的权值,构造出用户相似性网络.

具体地讲,我们将用户相似性网络表示为一个 XML 数据文件,它定义了用户相似性网络的图数据,包括图的节点属性、边的属性以及图的节点集

合、边的集合等.然后将其输入到 NodeXL(微软、斯坦福大学等多所大学联合研发的一个开源社会网络分析软件).该软件读入用户相似性网络的 XML 数据文件,然后调用它自带的 CNM 团体挖掘算法对该网络进行团体挖掘、可视化展示[25],得到了图 9.图中可见:该团体挖掘方法把本文研究的人物集合分为 18 个团体,左边 3 个团体是 NLP 领域的社交圈,其中标记为左 1 的团体代表 NLP 企业界社交圈,而标记为左 2 上的团体代表 NLP 教师的社交圈,标记为左 2 下的团体代表 NLP 学生的社交圈.图中连接 NLP 教师社交圈与学生社交圈的那些节点恰好是那些刚刚由学生身份变成教师身份的用户节点,该类节点既与学生交互频繁,同时也和教师交互密切;右边 15 个团体基本上和 NLP 领域不相关,

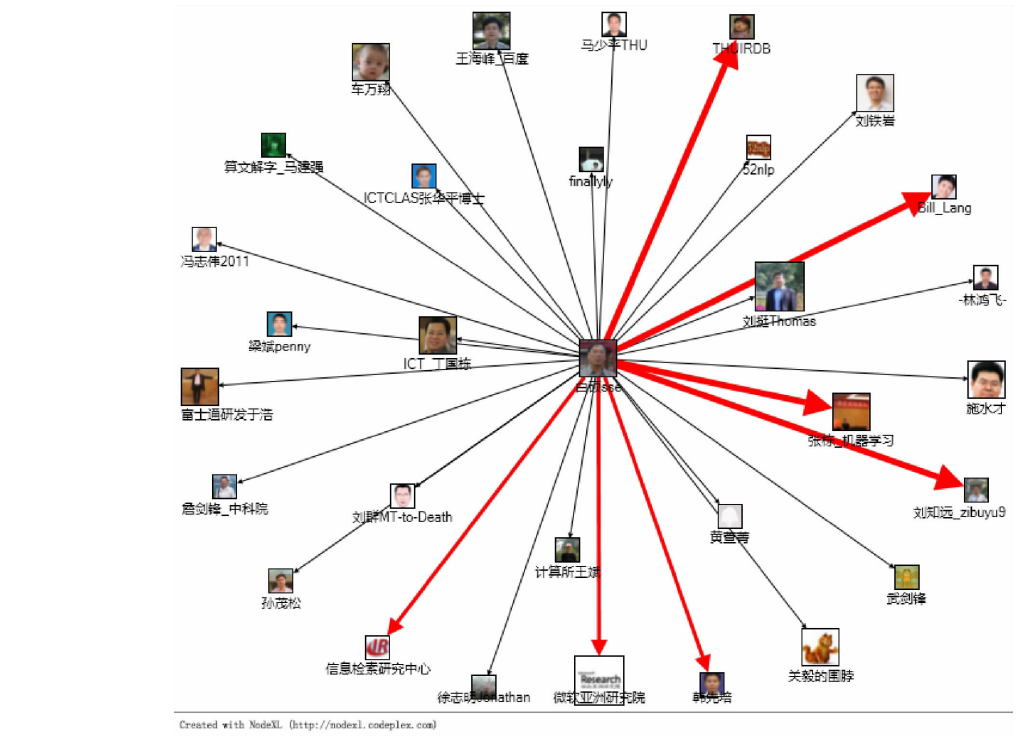


图 8 “@白硕 sse”的好友推荐

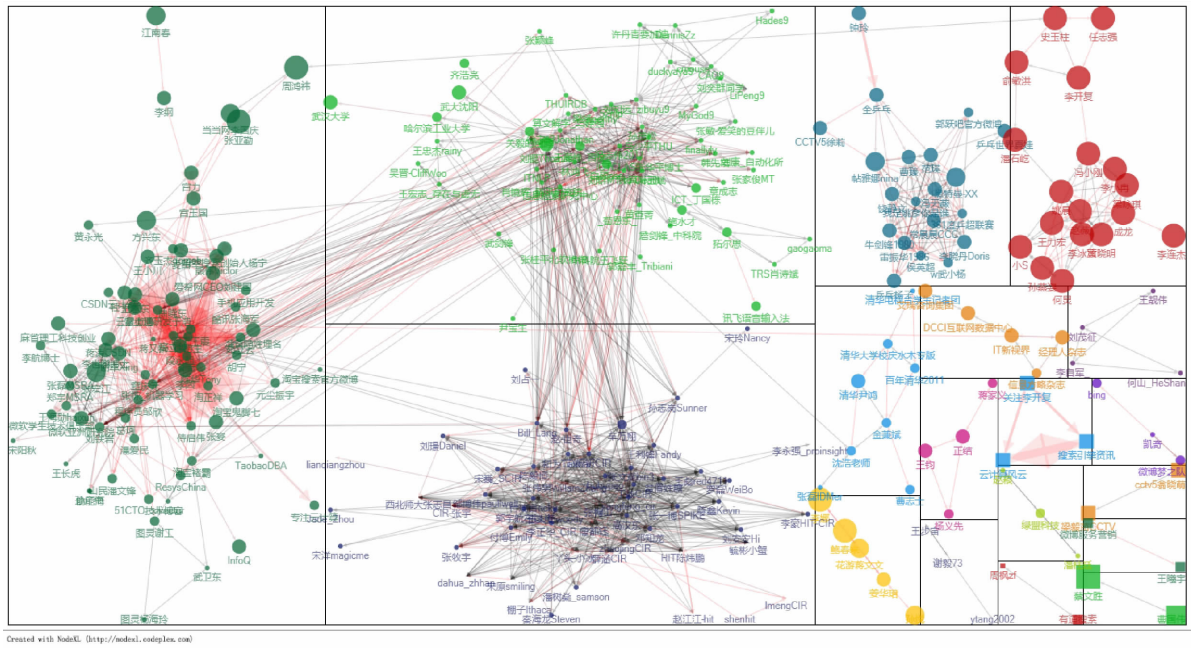


图 9 $r=0.05$ 的人物相关性网络

根据它们每个团体也被良好地划分，比如标记为右 1 上的团体代表微博的一些明星的社交圈，标记为右 2 上的团体代表一些运动员的社交圈. 由此可以看出：本文提出的基于社交信息的微博用户相似度计算方法以及由此生成的用户相似性网络，在团体分析方面有着较好的表现力. 同时，该实验也从侧面验证了本文提出的微博用户关系分析的正确性.

4　结束语

本文利用微博用户数据，开展了微博用户关系分析的研究工作. 针对微博用户信息的特点，分别给出了用户属性信息的模型表示、用户相似度的计算方法. 系统性考察了基于 3 种不同类型的用户属性信息（背景信息、社交信息、微博文本）的用户相似度的性能，实验结果表明：社交信息对用户的关系分析最有价值，基于社交信息的用户相似度对用户的关系分析能力最强. 为了进一步考察基于各种用户属性信息的用户相似度的实际性能，我们开展了用户推荐的实验，实验结果显示：基于社交信息的用户相似度对用户推荐的效果最好. 最后，为了验证基于社交信息的用户相似度对团体挖掘的性能，我们利用基于社交信息的用户相似度，计算了用户关系强度，据此构造了微博的用户相似性网络，并在该网络上开展了团体挖掘的初步实验，实验结果证明了该方法的有效性. 本文的局限在于：本文只是给出了人物相似性网络的团体分析的初步实验结果，尚未展开系统性的团体分析算法的研究，下一步准备系统地研究该网络的团体分析问题.

参　考　文　献

[1] Hannon J, Bennett M, Smyth B. Recommending twitter users to follow using content and collaborative filtering approaches//Proceedings of the 4th ACM Conference on Recommender Systems (RecSys'2010). New York, USA, 2010: 99-206

[2] Tang Jie, Sun Jimeng, Wang Chi, Yang Zi. Social influence analysis in large-scale networks//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France, 2009: 807-816

[3] Wu S, Hofman J M, Mason W A, Watts D J. Who says what to whom on twitter//Proceedings of the 20th International Conference on World Wide Web (WWW'11). Hyderabad, 2011: 705-714

[4] Lin Hong-Fei, Yang Yuan-Sheng. The representation and update mechanism for user profile. Journal of Computer Research and Development, 2002, 39(7): 844-846(in Chinese)
(林鸿飞，杨元生. 用户兴趣模型的表示和更新机制. 计算机研究与发展，2002，39(7)：844-846)

[5] Guo Yan, Bai Shuo, Yang Zhi-Feng, Zhang Kai. Analyzing scale of Web logs and mining users' interests. Chinese Journal of Computers, 2005, 28(9): 7-10(in Chinese)
(郭岩，白硕，杨志峰，张凯. 网络规模日志分析和用户兴趣挖掘. 计算机学报，2005，28(9)：7-10)

[6] Jeon H, Kim T, Choi J. Adaptive user profiling for personalized information retrieval//Proceedings of the 2008 International Conference on Convergence and Hybrid Information Technology. Busan, 2008: 836-841

[7] Xiang Rongjing, Neville J, Rogati M. Modeling relationship strength in online social networks//Proceedings of the WWW2010. Raleigh, North Carolina, USA, 2010: 981-990

[8] Clauset A, Newman M E J, Moore C. Finding community structure in very large networks. Physical Review E, 2004, 70(6): 066111

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

[10] Crandall D, Cosley D, Huttenlocher D, et al. Feedback effects between similarity and social influence in online communities//Proceedings of the KDD'08. Las Vegas, Nevada, USA, 2008: 160-168

[11] Weng J, Lim E-P, Jiang J, He Q. Twitterrank: Finding topic-sensitive influential twitterers//Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. New York, USA, 2010: 261-270

[12] Kwak H, Lee C, Park H, Moon S B. What is twitter, a social network or a new media?//Proceedings of the 19th International Conference on World Wide Web (WWW'10). Raleigh, 2010: 591-600

[13] Wang Dashun, Wen Zhen, Tong Hanghang, Lin Ching-Yung. Information spreading in context//Proceedings of the WWW2011. Hyderabad, India, 2011: 735-744

[14] Zhao Jichang, Wu Junjie, Xu Ke. Weak ties: Subtle role of information diffusion in online social networks. Physical Review, 2010, 82(1): 016105

[15] Gilbert E, Karahalios K. Predicting tie strength with social media//Proceedings of the CHI 2009. Boston, Massachusetts, USA, 2009: 211-220

[16] LesKovec J, Huttenlocher D, Kleinberg J. Predicting positive and negative links in online social networks//Proceedings of the 19th International Conference on World Wide Web (WWW'10). Raleigh, 2010: 641-650

[17] Liben-Nowell D, Kleinberg J. The link prediction problem for social networks//Proceedings of the 2003 ACM CIKM International Conference on Information and Knowledge Management (CIKM'03). New Orleans, Louisiana, USA, 2003: 556-559

[18] Hasan M, Chaoji V, Salem S, Zaki M. Link prediction using supervised learning//Proceedings of the Workshop on Link Discovery: Issues, Approaches and Applications. Chicago, Illinois, USA, 2005

[19] Kahanda I, Neville J. Using transactional information to predict link strength in online social networks//Proceedings of the ICWSM'09. San Jose, USA, 2009

[20] Chen Jilin, Geyer W, Dugan C, et al. "Make new friends, but keep the old"—Recommending people on social networking sites//Proceedings of the 27th International Conference on Human Factors in Computing Systems. New York, USA, 2009: 201-210

[21] Guy I, Ronen I, Wilcox E. Do you know? Recommending people to invite into your social network//Proceedings of the 13th International Conference on Intelligent User Interfaces. New York, USA, 2009: 77-86

[22] Hannon J, McCarthy K, Smyth B. Finding useful users on twitter: Twittomender the followee recommender//Proceedings of the 33rd European Conference on IR Research (ECIR 2011). Dublin. LNCS 6611. 2011: 784-787

[23] Ristad E S, Yianilos P N. Learning string-edit distance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(5): 522-532

[24] Wang Xiao-Yu, Xiong Fang, Ling Bo, Zhou Aoying. A similarity-based algorithm for topic exploration and distillation. Journal of Software, 2003, 14(9): 1578-1585(in Chinese)
(王晓宇, 熊方, 凌波, 周傲英. 一种基于相似度分析的主题提取和发现算法. 软件学报, 2003, 14(9): 1578-1585)

[25] Smith M A, Shneiderman B, Frayling N M, et al. Analyzing (social media) networks with NodeXL//Proceedings of the 4th International Conference on Communities and Technologies. University Park, Pennsylvania, 2009: 255-264

XU Zhi-Ming, born in 1967, Ph. D., professor. His research interests include social computing and natural language processing.

LI Dong, born in 1987, Ph. D. candidate. His research interests include social computing and information diffusion.

LIU Ting, born in 1972, Ph. D., professor. His research interests include natural language processing and social computing.

LI Sheng, born in 1942, Ph. D., professor. His research interests include natural language processing and information retrieval.

WANG Gang, born in 1988, M. S. candidate. His research interest is social computing.

YUAN Shu-Lun, born in 1987, M. S. candidate. His research interest is social computing.

Background

Along with the vast development of social media, many Web2.0 applications such as FaceBook, Twitter, and Sina Weibo and so on, are attracting more and more users to communicate interactively in various web communities. Web2.0 media are viewed as a group of Internet-based applications that allow the creation and exchange of user-generated content. Web2.0 social media have massive user-generated information, which should be deeply analyzed by social network analysis approaches(SNA). So some researchers pay attention to SNA techniques.

In general SNA can be viewed as a three-level analysis procedure. On the node-level phase, researchers are interested in how to analyze user interest and construct user profile. On the relation-level phase, relative research work is emphasized on how to estimate tie strength between users. On the network-level phase, researchers often use topology analysis approaches to mine communities, estimate user's social influence, and analyze information diffusion. In this paper, the authors mainly discuss how to compute user similarity, and use it to measure user relation strength. They firstly investigate various user attribute information' effects on user relation analysis problem. Their experiment data are downloaded from Sina weibo site by use of a Sina weibo crawler we designed; several user similarity estimation approaches are proposed according to various kinds of user attribution information. Then they tested them on some experiments. The experimental results showed that social information-based user similarity achieved the best performance. In addition, the authors conducted user recommendation experiments on them, and social information-based user similarity also achieved the best performance among of them. Finally they use social information-based user similarity approach to generate a user similarity network, on which they conducted a community detection experiments, the experimental results showed our social information-based user similarity approach is of remarkable performance.

This work is supported by the Program of National Science Foundation of China (Grant No. 61173074) and the Key Program of National Science Foundation of China (Grant No. 60736044). The former aims to large-scale SNA techniques on Microblog media, the latter is emphasized on personal information retrieval techniques. This work belongs to SNA part in these programs. The main research fields of our research team is social computing, the main purpose of our research team is to develope fundamental analysis tools for a Chinese microblog named Sina Weibo.