微博中特定用户的相似用户发现方法

仲兆满[1),2)] 胡 云[1)] 李存华[1)] 刘宗田[3)]

[1)](淮海工学院计算机工程学院 江苏 连云港 222005)

[2)](江苏金鸽网络科技有限公司软件研发中心 江苏 连云港 222005)

[3)](上海大学计算机工程与科学学院 上海 200072)

摘 要 微博的用户关系分析是近期的研究热点,而用户的相似度计算是微博用户关系分析的基础.已有方法在发现相似用户时,主要面向关注和粉丝群体,用户微博相似度及交互相关性计算对微博的动态特性利用不够.该文提出了新颖的微博特定用户的相似用户发现方法,该方法的创新性主要体现在:(1)发现相似用户时,在关注和粉丝的基础上引入了访客类用户,扩展了已有方法局限于关注和粉丝构建自我网络(Ego Network)的模型,增加了发现相似用户的多样性;(2)根据微博动态社交的特点,提出了用户动态微博的相似度计算和动态交互相关性计算方法,以时间片为动态社交划分的基础,以指数衰减为累加策略,使得微博用户的相似度计算更为合理,发现的相似用户更为准确.以新浪微博为例,选取了学术研究、企业管理、教育、文化、军事 5 个领域的 50 个种子用户,使用 $S@n$(前 n 个用户的得分)为评价指标,进行了相似用户发现的实验分析和比较.结果显示,访客类用户可以扩展相似用户的发现范围,访客在发现的相似用户中的比例为 32%,动态的微博相似度和交互相关性计算方法能够改善用户相似度的计算效果,比已有的最新方法的 $S@n$ 指标提高了 1.3.

关键词 用户关系分析;用户相似度计算;扩展的自我网络;动态微博相似度计算;动态交互相关性计算;社会媒体;社交网络;数据挖掘

中图法分类号 TP301 **DOI 号** 10.11897/SP.J.1016.2016.00765

Discovering Similar Users for Specific User on Microblog

ZHONG Zhao-Man[1),2)] HU Yun[1)] LI Cun-Hua[1)] LIU Zong-Tian[3)]

[1)](*School of Computer Engineering, Huaihai Institute of Technology, Lianyungang, Jiangsu* 222005)

[2)](*Software R&D Center, Jiangsu Jinge Network Technology Co., Ltd., Lianyungang, Jiangsu* 222005)

[3)](*School of Computer Engineering and Science, Shanghai University, Shanghai* 200072)

Abstract Recent studies focused on users' relationship on microblog, while similarity calculation of microblog users is the basis for analysis of users' relationship. Facing the problem of finding similar users, the existing methods mainly centered on followers and fans. Application of microblog dynamic characteristics was not enough when similarity between microblog and correlation among users was calculated. The work proposed a new method on discovering similar users for specific user on microblog. The method has achieved innovative points as follows: (1) Visitors were introduced to develop the Ego Network Model limited to followers and fans, with increased diversity of similar users; (2) Calculation methods were proposed for similarity between dynamic microblog of users, as well as correlation between dynamic interactions of users. It took the time slice as base for dividing dynamic social contact, and exponential damping as the accumulation strategy. It made similarty calculation among microblog users more reasonable, discovering more accurate similar users. With the case study of Sina microblog, we selected 50 seed users in

收稿日期:2015-06-15;在线出版日期:2015-11-09.本课题得到国家自然科学基金(61403156)、江苏省产学研前瞻性联合研究基金(BY2015248)和江苏省六大人才高峰基金资助(XXRJ-013)资助.**仲兆满**,男,1977 年生,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为信息检索、人工智能.E-mail:zhongzhaoman@163.com.**胡 云**,女,1978 年生,博士,副教授,主要研究方向为社交网络分析.**李存华**,男,1963 年生,博士,教授,主要研究领域为数据挖掘.**刘宗田**,男,1946 年生,教授,主要研究领域为人工智能、软件工程.

academic research, business management, education, culture and military. $S@n$ (score of top n users) was used as evaluation index for experimental analysis and comparison among methods discovering similar users. The results showed that visitors can extend the range discovering similar users (the proportion of visitors was 32% in the all mining similar users). Meanwhile, calculation effects of users' similarity can be improved with calculation methods for dynamic topic similarity and correlation of dynamic interaction ($S@n$, comparing to the latest existing methods, has increased by 1.3).

Keywords users' relationship analysis; users' similarity calculation; extended ego network; similarity calculation of dynamic microblog; correlation calculation of dynamic interaction; social media; social networks; data mining

1 引 言

当今,社交媒体被认为是 Web 上最有价值的信息资源之一.微博平台作为众多社交媒体中的一种,由于其传播性强、操作便利,很多用户在微博平台形成了类似于现实社会的交往圈子.传统媒体中用户和话题之间是二部网络,微博平台由于引入了关注关系(follow),使得用户和话题之间变得非常复杂,被认为是多模网络.由于微博平台信息传播性强、具有复杂的网络结构,近几年引起了学术界和产业界的高度重视.

微博中的相似用户是指在微博媒体上具有若干共同属性的用户群,这些属性主要包括用户的背景、关注、粉丝、微博、交互等信息.微博用户相似度度量的基础理论来源于社会学中的"同质性"(Homophily),即有关联的人往往有相似的特征,同时,联系越紧密,相似度越高[1-2].社交媒体上用户的信息总体上分为两类:一类是用户的背景(比如地点、教育、职业、兴趣等)和发表的微博信息(包括原创、转发或者评论);另一类是基于关注和粉丝构建的社交网络.基于这两类信息,已有的用户相似度计算方法大体上可以分为 3 类:(1)基于用户的背景和微博的文本信息方法,简记为 *SUDByText*;(2)基于关注和粉丝的社交网络方法,简记为 *SUDBySN*;(3)混合方法,即对基于文本方法 *SUDByText* 和基于社交网络方法 *SUDBySN* 的融合计算,简记为 *SUDByTSN*.近期出现的一些研究成果多是围绕 *SUDByTSN* 方法展开的,可以认为 *SUDByTSN* 是社交媒体用户相似度计算的主流研究方法.

本文从微博中指定的用户出发,在微博平台上尽量发现多的相似度高的用户.本文的研究内容属于社交网络中的自我网络分析,即站在个体的角度去分析个体本身及个体周围结点;研究方法综合了用户的文本信息和社交网络,属于 *SUDByTSN* 方法的范畴.该项研究的意义在于:(1)特定行业用户的线索发现,比如某个用户有涉恐倾向,在微博中挖掘有相似倾向的用户;(2)个性化推荐,比如为某个用户自动在微博中推荐志同道合的好友,或者厂家自动将广告推送给有相似兴趣的微博用户;(3)是社交整体网络分析的基础,比如通过计算用户的相似度,将微博上的用户划分为若干社群,以进一步研究社群特性.

本文的创新点主要体现在两点:(1)已有相似用户发现方法在微博社交网络关系的利用上,仅考虑了关注和粉丝两类用户.根据访客可以对用户发表的微博进行转发或评论的特点,在相似用户的发现时,引入了访客类用户,提出了扩展的自我网络模型 *EEN*(Extended Ego Network),增加了发现相似用户的全面性和多样性;(2)已有相似用户计算方法在计算用户的微博相似度和交互相关性时,没能体现微博社交的动态性.在用户的微博相似度和交互相关性计算方面,引入了时间的动态划分,能更好的体现微博的动态性,使得发现的相似用户更为准确.

本文第 2 节介绍已有的相关研究工作,包括基于用户的背景和微博的文本信息的方法,基于关注和粉丝的社交网络的方法及混合方法;第 3 节详细地阐述本文所提方法的原理和流程,包括相似用户发现模型、用户相似度计算模型、用户动态微博相似度计算方法及用户动态交互相关性计算方法;第 4 节从发现相似用户的准确性,关注、粉丝及微博的系数权重,时间片的划分,交互相关性的作用,时间衰减累加策略及发现相似用户的分布情况等角度进

行了实验对比分析，以验证本文所提方法的有效性；第 5 节对本文进行总结，探讨该方法的优缺点以及未来的研究方向.

2　相关工作

SUDByText 相关方法是早期研究的重点，Bhattacharyya 等人[3]在计算用户的相似度时，根据用户的基本属性(包括位置、家乡、活动、兴趣、专长等)提取若干关键词，基于语义距离计算关键词的相似性，进而获取用户的相似度. Wang 等人[4]在研究重叠社区发现时，认为用户的关联性(粉丝或关注)过于自由，重点使用了用户的元数据 Metadata(比如标签)计算用户的相似度. Diaby 等人[5]研究社交网络的工作推荐时，重点考虑的是用户的背景信息，对不同的社交媒体，选取了不同的背景信息，主要包括工作、教育、简历、兴趣、职位等信息. 进一步地，利用了用户的朋友(Friends)信息，但结论是背景相似的朋友才有价值. 文献[6-8]在社交推荐系统中也都有朋友信息的利用，但都是基于朋友的背景信息，没有考虑朋友之间的社交信息. Kim 等人[9]从一个用户出发，基于社交标签寻找到他感兴趣的社区. 社区的社交标签通过社区成员的标签提取，包括成员的兴趣、情感、地理位置、时间等. Xie[10]设计了通用的朋友推荐框架，在推荐朋友时，基于他们的地点、时间和内容等信息. 文献[11-12]在研究社区发现时，认为使用用户的背景，共享的图片、视频和标签等信息，既简单又有效.

仅利用用户之间的社交网络计算用户相似度方法较少，此类方法属于 *SUDBySN* 类型. Kahanda 等人[13]利用用户之间的交互性来度量用户关系强度，比如通信、文件传输等，其中通信也包含相互之间的转发或评论等行为. Hamid 等人[14]认为主流的推荐系统可以分为基于内容的和基于协作的两类类型，提出了基于内聚性(Cohesion)的社交媒体好友推荐方法，内聚性体现在连通性(connectedness)和密集度(density)两个方面. Samanthula 等人[15]计算朋友相似度时，考虑了网络的结构及用户之间真实的信息交互.

综合的利用用户的背景和微博信息及用户的社交网络，是主流的研究方法，即 *SUDByTSN* 类研究方法. 徐志明等人[16]在度量用户的相似度时，考虑了用户的背景信息(位置、标签及个人简介)、微博、社交和交互信息，以 50 个用户作为种子节点，抓取了 1 层关联的粉丝、关注用户，并认为社交信息在计算用户的相似度时最有价值. 背景的位置信息的相似度计算采用了分层比较的方法，标签及个人简介的相似度计算采用了编辑距离的方法. Xiang 等人[17]融合了用户的属性和用户间的交互计算用户关系强度，用户的属性包括学校、工作单位、兴趣组和地理位置等. 彭泽环等人[18]在进行微博用户推荐时，利用了用户的微博、个人信息、交互信息、社交拓扑信息等 4 类因素，面向腾讯微博进行了数据采集与实验对比，结果显示用户的交互信息对相似用户的推荐性能影响最大. Gou 等人[19]提出了使用用户的社交标签及网络的拓扑结构计算用户的相似度. Samanthula 等人[20]在研究私人朋友推荐 PFR 时(PFR 面向的是用户的朋友及社交标签都是隐藏的媒体)，用户相似度计算方法借鉴了文献[19]提出的方法. Modani 等人[21]研究了情趣相同社区(Like-minded communities)的发现，考虑了用户对感兴趣话题的排名，不是简单的计算两个用户话题的交集，此外还使用频繁项集挖掘社区的核心用户. Akcora 等人[22]同样利用了用户的背景信息和网络结构计算用户的相似度. 但不同的是，由于用户背景信息难以全面的获取，提出了从用户朋友已有的数据中，自动挖掘推理出用户的一些可能的背景信息.

综上所述，在社交网络的用户相似度计算方面，已经有一些研究成果，融合用户的文本信息和社交网络的计算方法是目前主流的研究方法. 用户相似度计算是诸多系统的基础，包括好友推荐系统、社区发现、社区划分等. 相关工作没有提及到针对给定的用户，在微博中通过关注、粉丝及访客快速发现相似用户的研究内容. 自我网络的构建是以关注和粉丝为基础，没有提及到访客类用户的利用. 在用户的微博相似度计算及交互相关性计算等方面，已有方法对时间要素的利用不够，缺少微博动态社交的深入研究.

3　特定用户的相似用户发现方法

3.1　基本概念

在介绍本方法之前，先形式化定义几个相关的概念，包括微博网络、微博博文及微博用户.

定义 1.　微博网络. 形式化描述为一个六元组：$MBN=\{U, MBlog, E_{UMB}, E_{UU}, F_{UMB}, C_{UMB}\}$. 其中，$U$ 为微博平台上的注册用户集；$MBlog$ 为用户发表的微博集(含原创、转发或者评论的各类微博)；

$E_{UMB}=\{e=(u_i,mblog_j)\mid u_i\in U,mblog_j\in MBlog\}$ 为用户与其所发表微博的连接边集；$E_{UU}=\{(u_i\rightarrow u_j)\mid u_i \text{ follows } u_j\}$ 为用户通过关注而形成的连接关系集，通过 follow 关系容易得到用户的粉丝关系集；$F_{UMB}=\{(u_i,mblog_j)\mid u_i\in U,u_i \text{ forwarded } mblog_j\}$ 是用户与其所转发的微博的关系集；$C_{UMB}=\{(u_i,mblog_j)\mid u_i\in U,u_i \text{ commented on } mblog_j\}$ 是用户与其所评论的微博的关系集.

定义 2. 微博博文. 形式化描述为一个三元组：$MBlog_i=\{MBlog_i\_body,MBlog_i\_t,MBlog_i\_u\}$. 其中，$MBlog_i\_body$ 为微博主体内容；$MBlog_i\_t$ 为微博发表的时间；$MBlog_i\_u$ 为发表该微博的用户.

定义 3. 微博用户. 形式化描述为一个六元组：$u_i=\{u_i\_Name,u_i\_Bg,u_i\_MBlog,u_i\_Follower,u_i\_Fans,u_i\_Visitor\}$. 其中，$u_i\_Name$ 为微博的用户名，是微博网络中用户的唯一标识符；u_i_Bg 为微博平台上的用户背景信息，不同微博平台背景有所差异；u_i_MBlog 为用户在微博网络上发表的微博集；$u_i\_Follower$ 为用户的关注集；u_i_Fans 为用户的粉丝集；$u_i\_Visitor$ 为用户的访客集，访客类用户指没有与用户 u_i 构建关注和粉丝关系，但与 u_i 进行了微博互动，包括发表微博时的“@”、转发或者评论行为.

依据定义 3，可以容易地获取用户 u_i 的关注数量 $|u_i\_Follower|$、粉丝数量 $|u_i\_Fans|$ 及访客数量 $|u_i\_Visitor|$.

3.2 相似用户发现模型及用户相似度计算模型

给定任意一个用户，从微博平台中尽可能发现多的相似用户是本文研究的主旨. 但由于微博用户的海量性，不可能漫无目的地在用户群体中随机查找. 为减少计算的规模、更快速的发现相似用户，特定用户的相似用户发现方法一般都是从微博用户的某个子集出发，该问题描述如下：

(1) 输入：微博特定用户 $SpecUser$；

(2) 输出：根据 $SpecUser$ 构建用户的自我网络 EN (Ego Network)，自我网络的构建一般是基于用户的关注(followers)和粉丝(fans)，得到的用户集记为 $EN(SpecUser)\subseteq U$，U 为微博平台所有的用户集，再通过某种计算方法从 $EN(SpecUser)$ 中挖掘出与 $SpecUser$ 相似的用户集 $SimUser(SpecUser)$，易知 $SimUser(SpecUser)\subseteq EN(SpecUser)$.

本文提出的特定用户的相似用户发现模型如图 1 所示.

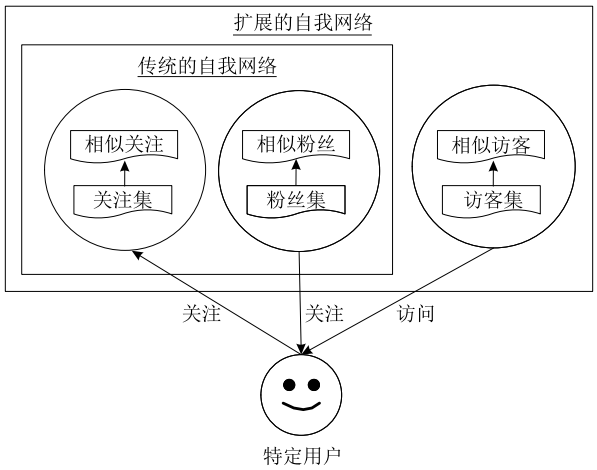

图 1 特定用户的相似用户发现模型

图 1 所示模型中，我们根据发表微博时可以“@”给选定用户及访客(Visitor)可以对用户发表的微博进行评论或者转发的特点，对传统的自我网络 EN(Ego Network)进行扩展. 相似粉丝、相似关注是一种闭合空间的相似用户发现方法，而相似访客的发现是一种类似于随机游走的发现策略，是对 EN 的进一步扩充，是对发现相似用户全面性及多样性有益的补充.

已有的社交网络分析方法，在社交行为分析时，主要基于关注和粉丝构建用户 $SpecUser$ 的自我网络 EN，用于后期分析的自我网络 EN 的用户集表示为 $EN(SpecUser)=FollowerCS(SpecUser)\cup FansCS(SpecUser)$. 其中，$FollowerCS(SpecUser)$ 为特定用户 $SpecUser$ 的关注类用户集，$FansCS(SpecUser)$ 为 $SpecUser$ 的粉丝类用户集. 扩展的自我网络 EEN(Extended Ego Network)扩充了用户的规模，得到用于分析的用户集表示为 $EEN(SpecUser)=FollowerCS(SpecUser)\cup FansCS(SpecUser)\cup VisitorCS(SpecUser)$. 其中，$VisitorCS(SpecUser)$ 为特定用户 $SpecUser$ 的访客集.

实际应用中，根据分析深度的需求，扩展的自我网络 EEN 可以层层扩展，即根据关注、粉丝、访客等用户像滚雪球一样进一步采集到他们关联的用户，不断扩充用户的规模. 可知，最终发现的特定用户 $SpecUser$ 的相似用户集 $SimUser(SpecUser)\subseteq EEN(SpecUser)$.

本文提出的两个微博用户 u_1 和 u_2 之间的相似度计算模型如图 2 所示.

图 2 所示模型中，用户属性分为动态属性 $DynamicAttr$ 和静态属性 $StaticAttr$，$DynamicAttr$ 包括微博 $MBlog$ 和交互 $Interaction$，$StaticAttr$ 包

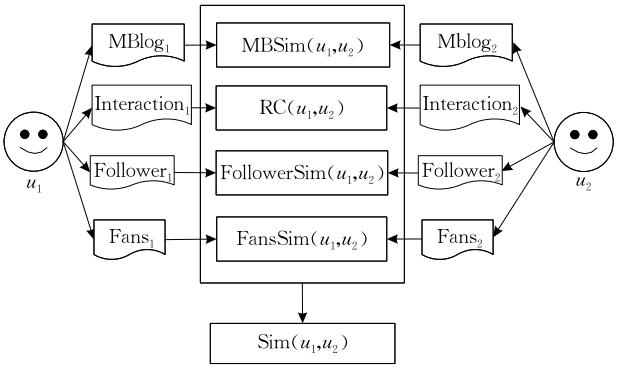

图 2　微博用户相似度计算模型

括关注 *Follower* 和粉丝 *Fans*. 在分别计算用户的动态微博相似度 $MBSim(u_1,u_2)$、动态交互相关性 $RC(u_1,u_2)$、关注相似度 $FollowerSim(u_1,u_2)$ 和粉丝相似度 $FansSim(u_1,u_2)$ 的基础上，最终由这 4 类属性决定两个用户的相似度. 用户文本属性的选取之所以使用用户的微博，这是因为：从隐私的视角而言，用户的很多背景在社交媒体上公开的较少[23-25]，能公开的主要是兴趣爱好、职业等；此外，微博媒体用户的关联，被认为是内容驱动的，即计算用户所发微博的相似度，对用户相关性的计算更有价值[26].

3.3　特定用户的相似用户发现算法

本文在图 1 的特定用户的相似用户发现模型和图 2 的用户相似度计算模型思想的指导下，从微博中发现特定用户的相似用户的算法设计如下.

算法 1.　特定用户的相似用户发现算法.

输入：微博用户 *SpecUser*

输出：相似用户集合 *SimUser*(*SpecUser*)

1. 从给定的微博用户 *SpecUser* 出发，获取如下信息：

1.1. 获取时间片 *TimeSpan* 内用户 *SpecUser* 的所有微博集 *MB-SpecUser*，包括原创、转发、评论的微博；

1.2. 获取 *SpecUser* 的关注集 *FollowerCS*(*SpecUser*) 和粉丝集 *FansCS*(*SpecUser*)；

1.3. 根据 *SpecUser* 的微博集 *MB-SpecUser* 提取访客集 *VisitorCS*(*SpecUser*)，将 3 类用户记为 $EEN(SpecUser)=FollowerCS(SpecUser)\cup FansCS(SpecUser)\cup VisitorCS(SpecUser)$.

2. 获取每个用户 $u_i\in EEN(SpecUser)$ 在时间片 *TimeSpan* 内的微博集 $MB\text{-}u_i$、关注集 $FollowerCS(u_i)$、粉丝集 $FansCS(u_i)$.

3. 计算用户 *SpecUser* 与 $u_i\in EEN(SpecUser)$ 的动态微博相似度，记为 $MBSim(SpecUser,u_i)$.

4. 计算用户 *SpecUser* 与 $u_i\in EEN(SpecUser)$ 的动态交互相关性，记为 $RC(SpecUser,u_i)$.

5. 计算 *SpecUser* 与 $u_i\in EEN(SpecUser)$ 的相似度 $Sim(SpecUser,u_i)$.

6. 按照相似度大小选取 *top* 个用户，得到相似用户集 *SimUser*(*SpecUser*).

该算法包括 6 个步骤，步 1、2 和 6 比较简单，容易实现，步 3 用户动态微博相似度计算、步 4 用户动态交互相关性和步 5 整合各个要素的用户相似度计算是本文研究的重点，在文章的 3.4 节、3.5 节和 3.6 节分别介绍.

特定用户的相似用户发现算法的复杂度由以下 6 个要素组成：

(1) $Time(CrawlMB)$，采集微博时间，包括特定用户、关注、粉丝和访客 4 类用户的博文采集时间；

(2) $Time(CrawlFollower)$，采集用户的关注时间，包括特定用户、关注、粉丝和访客 4 类用户的关注；

(3) $Time(CrawlFans)$，采集用户的粉丝时间，包括特定用户、关注、粉丝和访客 4 类用户的粉丝；

(4) $Time(MBSim(SpecUser,u_i))$，计算用户的动态微博相似度时间；

(5) $Time(RC(SpecUser,u_i))$，计算用户的动态交互相关性时间；

(6) $Time(Sort(EEN(SpecUser)))$，计算用户的动态交互相关性时间.

因此，算法的总复杂度可表示为

$$\begin{aligned}O(TotalTime)=&Time(CrawlMB)+\\&Time(CrawlFollower)+\\&Time(CrawlFans)+\\&Time(MBSim(SpecUser,u_i))+\\&Time(RC(SpecUser,u_i))+\\&Time(Sort(EEN(SpecUser)))\end{aligned}\quad(1)$$

由式(1)可见，算法的复杂度主要由用户的个数决定，即分析的用户个数越多，算法的复杂度越高.

由于微博用户的关系可以层层扩展，因此分析工作量较大，具体实现时，可以对分析用户的层数加以限制. 比如文献[16]选取了 50 个种子用户，采集时仅扩展了 1 层.

3.4　用户动态微博相似度计算

用户的动态微博相似度计算指将用户的微博按时间片进行划分，分别计算每个时间片的微博相似度，再采用一定的衰减策略进行累加.

已有方法在衡量两个用户发表微博相似度时，往往将一段时间范围内(比如 1 个月或者 1 年)的微博作为一个整体. 如果不考虑按时间的动态划分，即按时间片分别计算不同时间片的用户微博相似度，

那么在某个时间片内有较高相似度的微博,在面对整个时间周期时,这种相似度很可能被淹没.

时间片的划分一种是采用固定的时间周期进行均匀划分,另一种是根据用户的活跃程度等指标进行非均匀划分.在计算用户的微博相似度时,采用均匀的划分方法,可能将属于同一话题的微博分到不同的时间片,但这并不影响计算效果,因为在不同的时间片微博的相似度同样可以计算,而且不同时间片的计算结果还需进行累加.因此本文采用了均匀的时间片划分方式.Aaron 等人[27]认为时间窗口受日历等生活周期的影响,我们借鉴此思想,将时间周期定义为天、周及月分别进行实验分析.

微博中用户和微博的动态网络如图 3 所示.

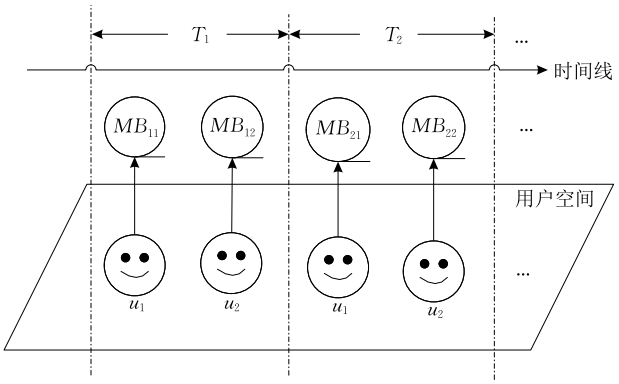

图 3 微博中用户和微博的动态网络

图 3 所示的网络模型和传统二部网络不同,添加了时间轴的动态约束.用户和微博的动态网络表示为 $G=(V,E,T)$.其中,T 为时间片集,$T=\{T_1,T_2,\cdots,T_m\}$;V 为用户和微博集,$V=\{V^{T_1},V^{T_2},\cdots,V^{T_m}\}$,$V^{T_i}=U^{T_i}\cup MB^{T_i}$,其中 U^{T_i},MB^{T_i} 分别为时间片 T_i 的用户集和微博集,满足 $U^{T_i}\cap MB^{T_i}=\varnothing$,$U^{T_1}=U^{T_2}=\cdots=U^{T_m}$;$E$ 为边集,$E=\{E^{T_1},E^{T_2},\cdots,E^{T_m}\}$,是一个映射函数 $E^{T_i}:U^{T_i}\times MB^{T_i}\rightarrow\{0,1\}$,即

$$e_{pq}(u_p^{T_i},MB_q^{T_i})=\begin{cases}1, & u_p^{T_i}\text{在时间片 }T_i\text{发表微博 }MB_q^{T_i}\\ 0, & \text{否则}\end{cases} \tag{2}$$

用户的很多博文过于短小,比如“赞了”、“好的”、“喜欢”、“期待中”等内容.我们对微博中常用的口头语进行了整理,目前整理的微博口头语词库共包括 173 条.基于这些词库对微博内容进行过滤,过滤掉的微博不再参与后期的特征提取及微博的相似度计算,但可以作为用户之间的交互行为,用于用户间的交互相关性的计算.

计算时间片 T_i 内两个用户的微博相似度时,将 T_i 内两个用户的微博分别看作一个整体,相当于是计算两篇文本的相似度.但因为微博样本的特征稀疏,而且话题总是由多个关联性强的词构成的,比如“地震”话题,信息中缺少不了“地震”话题的“时间”、“地点”、“伤亡”等关键词,所以我们采用了互信息的方法从微博中选取最有代表性的若干个特征,用于后期的微博相似度计算.

一个用户 u_j 在时间片 T_i 发表的所有微博记为 $MB\text{-}u_j^{T_i}$,基于互信息的微博特征词提取步骤如下:

(1)对 $MBS\text{-}u_j^{T_i}$ 进行分词、过滤通用词后(停用动词的过滤参考文献[28]归纳的内容),获取的特征词集合为 $WS_j^{T_i}=\{w_{j1}^{T_i},w_{j2}^{T_i},\cdots,w_{jx}^{T_i}\}$(假设 x 个特征词);

(2)计算两个词的互信息,计算方法如下[29]:

$$MI(w_{ju}^{T_i},w_{jv}^{T_i})=\frac{f(w_{ju}^{T_i},w_{jv}^{T_i})}{f(w_{ju}^{T_i})+f(w_{jv}^{T_i})-f(w_{ju}^{T_i},w_{jv}^{T_i})} \tag{3}$$

其中,$f(w_{ju}^{T_i},w_{jv}^{T_i})$ 为在某个窗口范围内词 $w_{ju}^{T_i}$ 和 $w_{jv}^{T_i}$ 共同出现的次数.由于微博比较短小,本文将窗口定义为每条微博范围内.对 x 个特征词,进行两两计算得到的互信息矩阵 $\boldsymbol{MIM}$(对称矩阵,同一个特征词互信息不做计算,值设为 0)为

$$\begin{bmatrix} & w_{j1}^{T_i} & w_{j2}^{T_i} & \cdots & w_{jx}^{T_i}\\ w_{j1}^{T_i} & 0 & MI(w_{j1}^{T_i},w_{j2}^{T_i}) & \cdots & MI(w_{j1}^{T_i},w_{jx}^{T_i})\\ w_{j2}^{T_i} & \cdots & 0 & \cdots & MI(w_{j2},w_{jx})\\ \cdots & \cdots & \cdots & \cdots & \cdots\\ w_{jx}^{T_i} & \cdots & \cdots & \cdots & 0 \end{bmatrix}.$$

(3)从 $\boldsymbol{MIM}$ 中选取互信息度大的 y 个词作为 $MB\text{-}u_j^{T_i}$ 的最终特征.

用户 $u_j^{T_i}$ 的微博 $MB\text{-}u_j^{T_i}$ 可向量化表示为 $KW\text{-}u_j^{T_i}=\{\langle kw_1\text{-}u_j^{T_i},w_1\text{-}u_j^{T_i}\rangle,\langle kw_2\text{-}u_j^{T_i},w_2\text{-}u_j^{T_i}\rangle,\cdots,\langle kw_y\text{-}u_j^{T_i},w_y\text{-}u_j^{T_i}\rangle\}$.其中,$w_i\text{-}u_j^{T_i}$ 为特征项的权重,使用 $TF\times IDF$ 方式计算.

在时间片 T_i 内两个用户 $u_p^{T_i}$,$u_q^{T_i}$ 的微博相似度计算方法使用经典的余弦相似度计算方法,如式(4)所示:

$$MBSim(u_p^{T_i},u_q^{T_i})=\frac{KW\text{-}u_p^{T_i}\cdot KW\text{-}u_q^{T_i}}{\|KW\text{-}u_p^{T_i}\|\cdot\|KW\text{-}u_q^{T_i}\|} \tag{4}$$

已有研究认为[30-32],微博用户圈子与进化聚类极为相近,存在短时平滑性现象,即动态网络的聚类结构在短时间内的变化是平缓的.短时平滑意味着短期内的历史交互信息(微博的评论、转发或原创)和当前交互信息具有一定相似性,可以将短时历史交互信息和当前交互信息综合作为当前时刻用户交

互相似性，以克服时间窗口划分带来的数据稀疏或观察缺失带来的问题. 综合历史信息和当前信息的实现方法之一是对有限相邻时段历史信息进行衰减累计. 本文使用了指数衰减的方法.

用户 u_p，u_q 微博相似度计算方法如式(5)所示：

$$MBSim(u_p,u_q)=\sum_{i=1}^{m}\lambda\,\mathrm{e}^{-\lambda(T_i-T_1)}MBSim(u_p^{T_i},u_q^{T_i}) \tag{5}$$

其中，T_i-T_1 的计算结果为时间片相差个数；λ 为指数衰减的参数.

3.5　用户动态交互相关性计算

微博媒体中，用户之间通过评论或转发的形式进行交互，这种行为能够体现用户共同的兴趣，反映了用户间的关联强度，即交互性强的，就可能进一步增加用户间的相似性. 用户的交互相关性理解为用户间的交互次数. 微博用户的动态交互网络如图 4 所示.

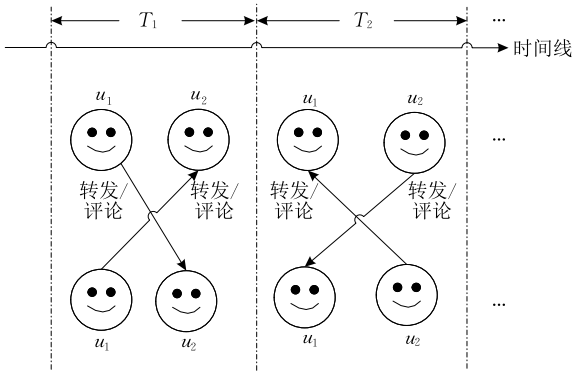

图 4　微博用户的动态交互网络

微博用户的动态交互网络表示为 $G=(V,E,T)$，其中，T 为时间片集，$T=\{T_1,T_2,\cdots,T_m\}$；V 为用户集，$V=\{U^{T_1},U^{T_2},\cdots,U^{T_m}\}$，$U^{T_1}=U^{T_2}=\cdots=U^{T_m}$；$E$ 为交互边集，$E=\{E^{T_1},E^{T_2},\cdots,E^{T_m}\}$，是一个映射函数 $E^{T_i}:U^{T_i}\times U^{T_i}\to\{0,num\}$，$num\geq 1$，即

$$e_{pq}(u_p^{T_i},u_q^{T_i})=\begin{cases}num, & \text{用户 } u_p^{T_i}\text{ 和 }u_q^{T_i}\text{ 在时间片}\\ & T_i\text{ 的交互次数}\\ 0, & \text{用户 } u_p^{T_i}\text{ 和 }u_q^{T_i}\text{ 在时间片}\\ & T_i\text{ 无交互}\end{cases} \tag{6}$$

在时间片 T_i 内两个用户 $u_p^{T_i}$，$u_q^{T_i}$ 的交互相关性为用户间的交互次数，记为 $RC(u_p^{T_i},u_q^{T_i})$，m 个时间片的最大交互次数记为 $RC_{\max}$，以 $RC_{\max}$ 为参考对用户的交互相关性进行归一化：

$$RC(u_p^{T_i},u_q^{T_i})=\frac{RC(u_p^{T_i},u_q^{T_i})}{RC_{\max}} \tag{7}$$

同样的，借鉴文献[30-32]阐述的微博用户圈子的短时平滑性现象，在计算用户动态交互相关性时，引入指数衰减来刻画这种关系.

用户 u_p，u_q 的交互相关性计算方法如式(8)所示：

$$RC(u_p,u_q)=\frac{1}{m}\sum_{i=1}^{m}\lambda\,\mathrm{e}^{-\lambda(T_i-T_1)}\times RC(u_p^{T_i},u_q^{T_i}) \tag{8}$$

3.6　用户相似度计算

考虑到用户的关注、粉丝和微博，计算用户 $SpecUser$ 与 $u_i\in EEN(SpecUser)$ 的相似度 $Sim(SpecUser,u_i)'$ 方法如式(9)所示：

$$\begin{aligned}Sim(SpecUser,u_i)'=&\lambda_1\times FollowerSim(SpecUser,u_i)+\\&\lambda_2\times FansSim(SpecUser,u_i)+\\&\lambda_3\times MBSim(SpecUser,u_i)\end{aligned} \tag{9}$$

其中，关注相似度定义为

$$FollowerSim(SpecUser,u_i)=\frac{|FollowerCS(SpecUser)\cap FollowerCS(u_i)|}{|FollowerCS(SpecUser)\cup FollowerCS(u_i)|} \tag{10}$$

粉丝相似度定义为

$$FansSim(SpecUser,u_i)=\frac{|FansCS(SpecUser)\cap FansCS(u_i)|}{|FansCS(SpecUser)\cup FansCS(u_i)|} \tag{11}$$

微博相似度的计算使用 3.4 节式(4)和式(5)计算.

对于式(9)中 λ_1，λ_2 和 λ_3 3 个系数的确定，已有方法多是经验指导的.

由于微博用户的海量性，发现相似用户的常用评价指标是 $P@n$，即取排名前 n 的相似用户，判断是真正相似用户的比例. 对微博用户而言，由于每个用户涉及的信息较杂，包括关注、粉丝、微博、交互等要素，靠人工判断难度很大. 所以，我们对 $P@n$ 进行改进，提出了 $S@n$ 的评价指标，即计算每种方法(本文使用了 4.2 节介绍的 4 种方法)得到的前 n 个相似用户的得分. 4.2 节的实验评测方法各有侧重点：方法 1 是 $SUDByText$ 使用用户的背景及微博等文本信息；方法 2 是 $SUDBySN$ 使用用户的社交信息，包括关注和粉丝；方法 3 是 $SUDByTSN$ 考虑了用户的文本及社交信息；方法 4 是 $SUDByTSN\text{-}VD$ (取访客 Visitor 及动态 Dynamic 的第 1 个大写字母)为本文提出的方法.

由于 4 种方法在计算用户间相似度时的出发点都各有侧重，因此，如果发现的用户 u_1 能够在 4 种方法的前 n 个用户中出现的比例越高(即不同侧重点的方法都可能发现用户 u_1)，可以相信，用户 u_1 为相似用户的可能性就越大.

假设有 m 种评价方法，方法 $Method_i(1\leq i\leq m)$ 得到的前 n 个相似用户的集合为 $Method_i=\{u_{i1},u_{i2},\cdots,u_{in}\}$，将 u_{i1} 在每种方法得到的相似用户集出现的总次数记为 $Count(u_{i1})$，则方法 $Method_i$ 的 $S@n=\sum_{j=1}^{n}Count(u_{ij})$. 该方法不需要人工干预，容易实现，且相对客观.

基于 $S@n$ 的计算思想，对每个用户而言，目标是 $S@n$ 最大，该问题的描述如式(12)所示：

$$\begin{aligned}&\max \quad S@n\\&\text{s.t.} \quad 0\leq\lambda_1\leq 1,\ 0\leq\lambda_2\leq 1,\ 0\leq\lambda_3\leq 1,\\&\qquad\lambda_1+\lambda_2+\lambda_3=1\end{aligned}\tag{12}$$

本文选取了学术研究、企业管理、教育、文化、军事 5 个领域的 50 个种子用户，在实验的基础上考察 λ_1、λ_2 和 λ_3 3 个系数的取值范围，详见 4.4 节的论述.

进一步地，考虑到用户交互相关性对用户相似度的影响，最终的用户 $SpecUser$ 与 $u_i\in EEN(SpecUser)$ 的相似度计算如式(13)所示：

$$Sim(SpecUser,u_i)=log_2^{2+RC(SpecUser,u_i)}\times Sim(SpecUser,u_i)'\tag{13}$$

其中 $RC(SpecUser,u_i)$ 为用户 $SpecUser$ 和 u_i 动态交互相关性，计算方法使用 3.5 节式(7)和式(8)，取对数为了体现交互对最终相似度计算影响的平滑性.

4 实验及分析

4.1 实验数据

目前，没有用于微博用户相似度计算的公开语料，研究者大多是根据需求，自行从指定的微博上采集相关语料. 比如，文献[16]以互联网高管领域 50 个用户为种子，爬行了 1 层，得到了关注、粉丝用户集及他们的背景信息、交互信息和微博集. 胡云等人[33]使用了数据堂公司提供的 5000 个用户(该数据是发布者随机抽取的)进行了实验，但这些数据仅包含用户名、关注数、粉丝数及微博数等基本信息，实验时还需进一步采集. Hamid 等人[14]面向 Facebook，从 1 个用户的社交网络出发，随机选取了 20 个用户进行了数据采集与分析. 彭泽环等人[18]面向腾讯微博，选取了两个时间点收集部分数据.

本文以新浪微博为例，选取了学术研究、企业管理、教育、文化、军事 5 个领域的 50 个种子用户进行实验数据的采集与分析.

在新浪微博搜索框中输入领域关键字进行检索，然后点击“找人”按钮，选取了“个人认证”及“普通用户”两类用户，使用 HtmlUnit 进行采集. 有些领域用户的关注或者粉丝过多，超过几万、甚至是上百万. 为了分析的方便，对获取的用户进行了筛选，关注及粉丝数限定在 5000 以内. 从每个领域中随机选取 10 个种子用户进行实验分析，微博的采集时间限定在 2015 年 1 月 1 日至 2015 年 5 月 28 日，共计 5 个月. 5 个领域获取的认证及普通用户情况见表 1 所示.

表 1 实验选用的 5 个领域①

序号	领域	关键字	认证及普通用户数
1	学术研究	信息检索	490
2	企业管理	互联网高管	45
3	教育	幼儿教育	6049
4	文化	谍战	876
5	军事	歼 20	728

目前，新浪微博为防止他人获取用户的关注、粉丝进行恶意关注或广告骚扰，对非本人的关注、粉丝的访问量进行了限制，只能获取前 5 页内容，大约 100 个关注、100 个粉丝. 从统计分析的角度而言，抽取 100 个关注和 100 个粉丝样本进行统计分析也是有代表性的.

5 个领域 50 个用户的关注、粉丝、访客及微博数量如表 2 所示.

表 2 50 个用户的关注、粉丝、访客及微博数

序号	领域	关注数	粉丝数	访客数	微博数
1	学术研究	967	1000	2807	1267
2	企业管理	884	1000	3402	1431
3	教育	1000	1000	1883	836
4	文化	982	1000	2291	1236
5	军事	976	1000	1936	1041
共计		4809	5000	12319	5811

为了计算特定用户与每个关注、粉丝及访客的相似度，需扩展下一层采集关注、粉丝及访客 3 类用户的关注、粉丝及微博. 同样的，每个用户的关注和粉丝的个数都为 100，采集微博的时间限定在 2015 年 1 月 1 日至 2015 年 5 月 28 日.

用户的微博内容一方面是原创的，另一方面是转发/评论的，将转发/评论的微博同样作为用户的微博内容，但转发/评论同一微博多次时仅算 1 次.

最终获取的用于实验分析的关注类用户总数为 2157843、粉丝类用户总数为 2086613、微博总数为 932531.

① 2015 年 5 月 28 日执行完采集.

4.2 几种实验方法

实验选用的 6 种方法介绍如下：

(1)方法 1：*SUDByText*，基于用户的背景和微博计算用户的相似度，类似于文献[5-6,10,12]介绍的方法. 根据新浪微博的特点，选取的用户背景信息包括简介、标签、教育、职业信息. 背景信息的相似度计算采用了 Jaccard 方法，和式(10)、(11)相同. 微博的相似度计算和式(3)相同，没有按时间片划分考虑微博的动态性. 背景及微博相似度线性整合时的取值参考文献[5-6,12]，分别为 0.3 和 0.7.

(2)方法 2：*SUDBySN*，基于关注和粉丝的社交网络计算用户的相似度，类似于文献[13,15]介绍的方法. 社交网络构建时仅利用了关注和粉丝，没有考虑访客. 关注和粉丝的相似度计算采用了 Jaccard 方法，和式(10)、(11)相同. 最终的相似度对关注和粉丝的相似度进行了线性整合，参考文献[13,15]的取值，关注相似度的权重为 0.6，粉丝相似度的权重为 0.4. 用户间的交互相关性计算和式(8)相同，但没有考虑交互的动态性.

(3)方法 3：*SUDByTSN*，已有的混合方法，基于用户的文本信息和社交网络计算用户的相似度，类似于文献[16-18] 介绍的方法. 文本信息包括微博及简介、标签、教育、职业信息等背景信息，社交网络仅利用了关注和粉丝，没有考虑访客. 背景信息及微博的相似度计算和方法 *SUDByText* 相同；关注和粉丝的相似度计算和方法 *SUDBySN* 相同.

(4)方法 4：*SUDByTSN-VD*，本文提出的混合方法. 该方法仅选取了用户的微博信息，社交网络构建时利用了关注、粉丝和访客 3 类用户. 为了减少统计分析量，对用户扩展的自我网络涉及的 3 类用户，第 1 层扩展时采集了关注和粉丝的基本信息及其在 2015 年 1 月至 5 月发表的所有微博，并从微博中提取出访客，第 2 层扩展时仅采集了关注和粉丝的用户名. 时间片的指数衰减参数 λ 的取值参考文献[34]，$\lambda=0.3$. 计算用户相似度时，关注、粉丝及微博相似度的权重分别为 $\lambda_1=0.5$，$\lambda_2=0.2$，$\lambda_3=0.3$，3 个参数的取值对实验结果的影响参见 4.4 节的论述. 时间片按周划分，不同的时间片划分(天、周及月)对实验结果的影响参见 4.5 节的论述.

进一步地，为验证用户间的交互性、时间片划分以及时间衰减累加策略对本文所提方法 *SUDByTSN-VD* 的影响，对 *SUDByTSN-VD* 进行修改，得到方法 5 和方法 6：

(5)方法 5：$SUDByTSN\text{-}VD_1$，和方法 4 不同的是没有考虑用户之间的交互性，即用户相似度的计算使用了关注、粉丝和微博 3 个指标，实验参数的设置(包括指数衰减参数 λ，关注、粉丝及微博的 3 个系数 λ_1，λ_2 和 λ_3)和方法 4 相同.

(6)方法 6：$SUDByTSN\text{-}VD_2$，和方法 4 不同的是微博相似度及交互相关性计算时没有考虑按时间的衰减累加策略，关注、粉丝及微博的 3 个系数 λ_1，λ_2 和 λ_3 的取值和方法 4 相同.

4.3 发现相似用户的准确性比较

如 3.6 节所述，我们采用 $S@n$ 评价前 4 种方法在计算用户相似度时的优劣.

要注意的是，本文提出的方法 *SUDByTSN-VD* 的相似用户由于扩展到了访客类，而这些访客在方法 1、方法 2 和方法 3 中是无法得到的. 因此，在计算 *SUDByTSN-VD* 的 $S@n$ 指标时，对访客进行如下处理：对其他 3 种方法，分别计算访客与指定用户的相似度，对方法 $Method_j \in \{SUDByText, SUDBySN, SUDByTSN\}$ 而言，如果访客 $visitor_i$ 的相似度值可以进入前 n，则认为 $visitor_i$ 存在于 $Method_j$ 的相似用户集中.

对学术研究领域(信息检索)用户“王利锋 Fandy”，本文提出的方法 *SUDByTSN-VD* 得到的前 10 个相似用户如表 3 所示. 对 50 个微博用户，4 种方法得到的平均 $S@n$ 如表 4 所示，趋势如图 5 所示.

表 3　方法 *SUDByTSN-VD* 得到的与用户“王利锋 Fandy”相似的前 10 个用户

排序	用户名	关注数	粉丝数	5 个月的微博数	相似度	关系
1	LR 机器学习计算机视觉	895	4205	547	0.1741	Follower
2	七月算法问答	119	3471	735	0.1330	Follower, Fans
3	winsty	263	3832	409	0.1173	Follower
4	YJYJ_Focus	453	226	386	0.0832	Visitor
5	张大奎	1701	5771	11	0.0755	Visitor
6	ansj	564	1569	56	0.0751	Follower
7	李亚超 NLP	737	1147	401	0.0658	Visitor
8	韧在百度	263	3771	18	0.0574	Follower
9	小村长 zack	358	202	95	0.0503	Fans
10	赵家平 USC	759	2459	32	0.0498	Follower

表 4 4 种方法得到的 50 个微博用户的 *S@n*

领域	*S@n*			
	SUDByText	*SUDBySN*	*SUDByTSN*	*SUDByTSN-VD*
学术研究	31.2	32.20	35.4	37.1
企业管理	29.4	30.30	33.2	33.9
教育	27.6	31.40	32.6	33.8
文化	28.5	29.40	32.1	32.5
军事	30.3	32.10	34.2	36.7
Average S@n	29.4	31.08	33.5	34.8

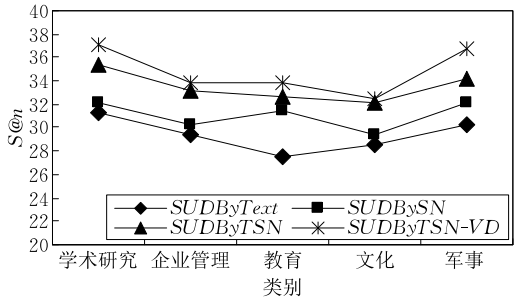

图 5 4 种方法得到的 50 个用户的平均 *S@n*

由表 4 和图 5 可见，对 50 个用户的平均 *S@n* 而言，方法 *SUDByTSN-VD* 得分最高，为 34.8；方法 *SUDByTSN* 的得分其次，为 33.5；方法 *SUDByText* 得分最低，为 29.4. 在 4 种方法中，*SUDByTSN* 与 *SUDByTSN-VD* 的得分都比较高，这进一步验证了，混合型社交网络分析的优势所在. 方法 *SUDByTSN-VD* 的 *S@n* 得分高于方法 *SUDByTSN*，这是因为 *SUDByTSN-VD* 引入时间的动态约束，使得发现的用户更为准确. 方法 *SUDByText* 仅利用了用户的背景和微博信息，方法 *SUDBySN* 仅使用了微博的社交网络信息，包括关注及粉丝，这两种方法都有一定的缺陷. 就 *SUDByText* 和 *SUDBySN* 而言，方法 *SUDBySN* 要优于 *SUDByText*，这也进一步验证了用户的社交信息比用户的其他信息更有利用价值.

对 5 个领域而言，“学术研究”和“军事”两个领域得分较高，主要原因是获取该领域用户时使用了“信息检索”、“歼 20”进行搜索，关键词的范围限定比较具体，得到种子用户的朋友圈比较窄小，所发表的微博比较专业，每个用户的相似用户得分比较平稳. 而对于另外 3 个领域(“企业管理”、“教育”和“文化”)的用户而言，他们的朋友圈往往过大，粉丝都可能达到几十万人，日常所发微博也比较发散，对相似用户的计算干扰较大. 这说明，用户所属的领域范围越是狭小、专业化程度高，在发现相似用户时的效果越好.

此外，我们对 50 个用户发现的 500 个相似用户(每个用户取排名靠前的 10 个相似用户)的活跃性进行统计，发现在 5 个月的时间段内，500 个用户中，95%以上的用户都有 100 次以上的转发、评论或者发表微博的行为，只有 5%的用户不太活跃. 不太活跃的用户之所以排名靠前，原因是计算相似度时的关注、粉丝指标得分较高.

微博媒体中存在的“冷启动”用户可以分为两种情况：(1) 新用户；(2) 不活跃用户. 对新用户而言，由于关注、粉丝、微博及交互 4 类信息几乎没有，因此本文所提方法难以发现此类“冷启动”用户. 对不活跃用户而言，这类用户发表的微博、与其他用户间的交互较少，这两个属性得分较低，如果粉丝和关注的信息量较大，即使微博、交互较弱，本文所提方法也可能发现此类用户. 但明显的是，本文所提方法更有利于发现微博中的活跃用户，即经常发表微博、有一定量的关注和粉丝、与其他用户互动性强的用户.

4.4 关注、粉丝及微博的系数取值

基于 3.6 节介绍的参数优化思想，我们对 5 个领域的 50 个用户在确保 *S@10* 最大的前提下，对关注、粉丝及微博 3 个系数的取值范围进行了实验.

过程 1. 计算 *S@10* 的 3 个核心步骤：

(1) *Sort*. 按照计算的相似度对用户集中的用户排序；

(2) *Select*. 从排序的用户集中，选取相似度最高的前 10 个用户；

(3) *Count*. 计算 10 个用户在各种方法中(本文使用了 4.2 节介绍的 4 种方法)的得分(具体的得分计算论述参见 3.6 节内容).

计算用户相似度时的 3 个参数 λ_1，λ_2，λ_3 取值分为[0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]，共 11 种情况. 以 *S@10* 最大值为参考，值相差范围不大时(本文取 5%)，认为 λ_1，λ_2，λ_3 的取值都是合理的. 时间片的选取以周为单位.

50 个用户的 λ_1，λ_2，λ_3 3 个参数取值范围，统计结果如图 6 所示.

从图 6 可见，λ_1 的取值位于区间[0.3, 0.6]居多，占 70%，其中最多的是区间[0.4, 0.5]；λ_2 的取值位于区间[0.2, 0.4]居多，占 78%，其中最多的是区间[0.2, 0.3]；λ_3 的取值位于区间[0.3, 0.5]居多，占 64%，最多的是区间[0.3, 0.4]. 这说明，在计算用户相似度时，对关注、粉丝、微博 3 个要素而言，最重要的是关注，其次是微博，最后是粉丝. 在实际应用中，在保证 $\lambda_1+\lambda_2+\lambda_3=1$ 的情况下，λ_1，λ_2，λ_3 的取值只要是落在这些区间，可以认为就是合理的.

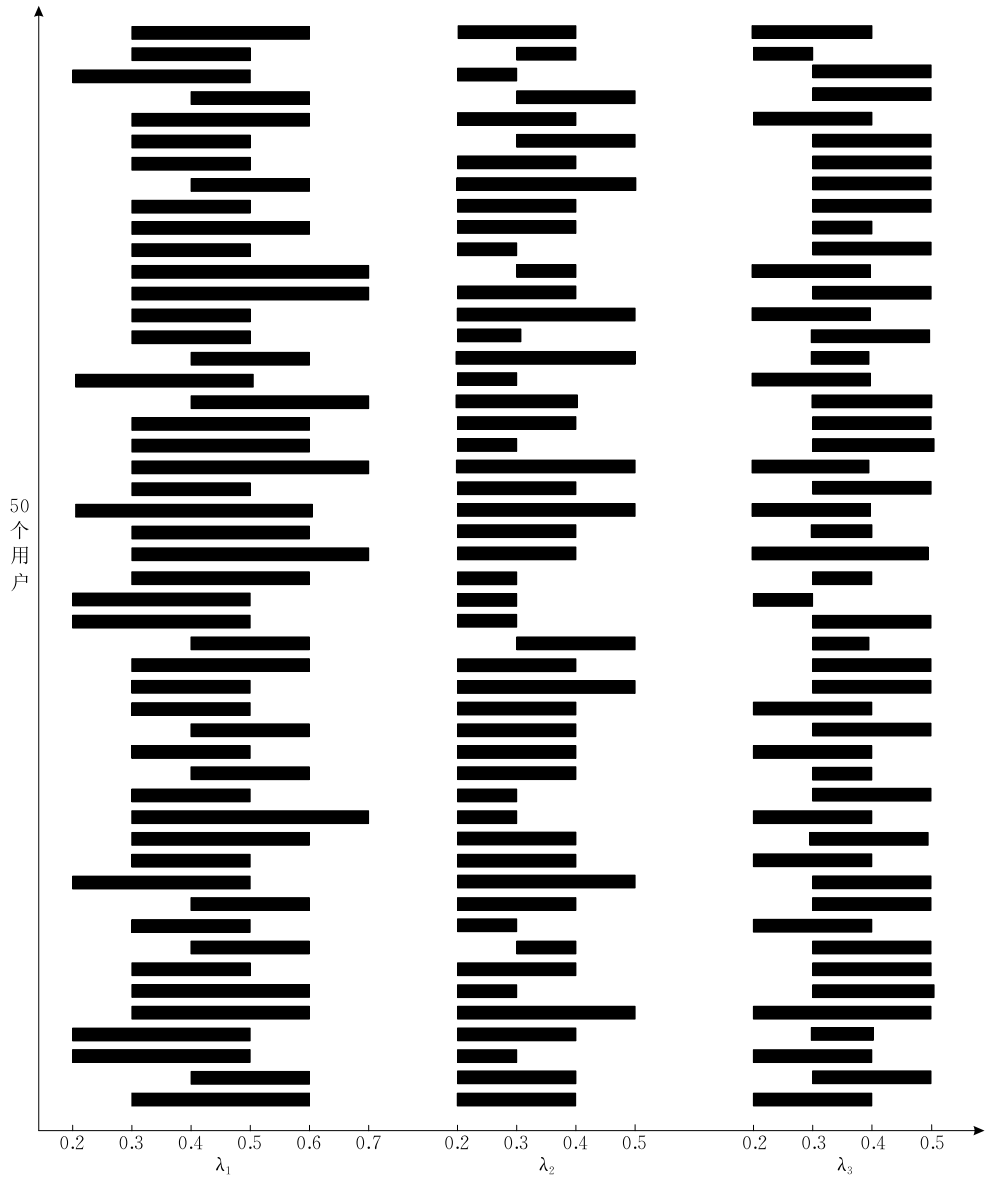

图 6　50 个用户的 λ_1，λ_2，λ_3 3 个参数取值范围

4.5　时间片划分对发现相似用户准确性的影响

对获取的 2015 年 1 月至 5 月的共 5 个月的微博，在用户动态微博相似度计算及用户动态交互相关性计算时，参考工作周期的划分原理，我们分别以“天”、“周”和“月”进行时间片划分实验.

对本文提出的方法 *SUDByTSN-VD*，选用了 3 个不同时间片，得到的实验结果 $S@n$ 如表 5 所示.

表 5　不同时间片划分得到的 $S@n$

领域	天	周	月
学术研究	137.00	142.00	127.00
企业管理	143.00	144.00	132.00
教育	155.00	151.00	145.00
文化	149.00	147.00	138.00
军事	143.00	150.00	139.00
平均	29.08	29.36	27.24

由表 5 可见，时间片按照“天”划分得到的 $S@n$ 为 29.08，时间片按照“周”划分得到的 $S@n$ 为 29.36，结果相差很小，仅为 0.28. 而按照“月”划分得到的 $S@n$ 为 27.24，由于时间周期偏长，微博的动态相似度及交互相关性计算效果体现不够明显，实验结果 $S@n$ 差一些. 考虑到微博相似度计算及交互相关性计算的时间消耗，实际应用中建议时间片选取“周”进行划分比较合理.

4.6　交互相关性对发现相似用户准确性的影响

与方法 *SUDByTSN-VD* 相比，*SUDByTSN-VD*$_1$ 在计算用户相似度时没有考虑用户之间的交互性，仅利用微博、关注和粉丝 3 个属性. 2 种方法得到 50 个微博用户的 $S@n$ 如表 6 所示.

表 6 不考虑交互相关性时得到的 $S@n$

领域	*SUDByTSN-VD*	$SUDByTSN\text{-}VD_1$
学术研究	37.1	35.50
企业管理	33.9	32.20
教育	33.8	32.10
文化	32.5	30.60
军事	36.7	35.70
平均	34.8	33.22

由表 6 可见，$SUDByTSN\text{-}VD_1$ 得到的 $S@n$ 为 33.22，*SUDByTSN-VD* 的 $S@n$ 为 34.8，结果较大，为 1.58. 同时发现，$SUDByTSN\text{-}VD_1$ 不如 *SUDByTSN* 的得分高，*SUDByTSN* 的得分为 33.5. 在计算微博用户的相似度时，用户之间的交互信息对改善相似度计算的效果是有较大帮助的.

4.7 时间衰减对发现相似用户准确性的影响

与方法 *SUDByTSN-VD* 相比，$SUDByTSN\text{-}VD_2$ 在计算用户相似度时没有采用按照时间片而进行衰减累加的策略，用户动态微博相似度计算方法如式(14)所示：

$$MBSim(u_p,u_q)=\sum_{i=1}^{m}MBSim(u_p^{T_i},u_q^{T_i}) \quad (14)$$

用户动态交互相关性计算如式(15)所示：

$$RC(u_p,u_q)=\frac{1}{m}\sum_{i=1}^{m}\times RC(u_p^{T_i},u_q^{T_i}) \quad (15)$$

两种方法得到 50 个微博用户的 $S@n$ 如表 7 所示.

表 7 不考虑时间衰减累加时得到的 $S@n$

领域	*SUDByTSN-VD*	$SUDByTSN\text{-}VD_2$
学术研究	37.1	36.70
企业管理	33.9	33.70
教育	33.8	32.80
文化	32.5	31.50
军事	36.7	36.40
平均	34.8	34.22

由表 7 可见，$SUDByTSN\text{-}VD_2$ 得到的 $S@n$ 为 34.22，*SUDByTSN-VD* 的 $S@n$ 为 34.8，相差 0.58，效果有所下降. 这说明，按照时间片进行指数衰减累加的策略对用户相似度计算是合理的.

4.8 发现相似用户的分布比较

已有方法发现的相似用户仅分布于关注和粉丝两类，本文提出的方法 *SUDByTSN-VD* 发现的相似用户分布于关注、粉丝和访客 3 类.

相似用户的分布评价包括：

(1) 关注比例

$$p_{\text{follower}}=\frac{10\text{ 个相似用户中关注的个数}}{\text{前 10 个相似用户}}\times 100\%;$$

(2) 粉丝比例

$$p_{\text{fans}}=\frac{10\text{ 个相似用户中粉丝的个数}}{\text{前 10 个相似用户}}\times 100\%;$$

(3) 访客比例

$$p_{\text{visitor}}=\frac{10\text{ 个相似用户中访客的个数}}{\text{前 10 个相似用户}}\times 100\%.$$

对 5 个领域的用户，4 种方法得到的 p_{follower}、p_{fans} 和 p_{visitor} 结果如表 8 所示. 表 8 中，发现的相似用户可能同时属于多类用户，比如同时属于关注和粉丝，计算指标得分时，需重复统计. 假设一个相似用户既是关注，又是粉丝，在统计关注和粉丝的分布比例时，需各自计算 1 次.

表 8 4 种方法得到的相似用户的分布情况

领域	*SUDByText*		*SUDBySN*		*SUDByTSN*		*SUDByTSN-VD*		
	p_{follower}	p_{fans}	p_{follower}	p_{fans}	p_{follower}	p_{fans}	p_{follower}	p_{fans}	p_{visitor}
学术研究	78	38	74	40	82	32	56	34	36
企业管理	76	36	80	32	80	38	56	26	28
教育	74	32	76	32	78	36	58	26	30
文化	74	34	74	36	76	42	60	28	28
军事	68	36	72	42	76	40	52	36	38
Average	74	35	75	36	78	38	56	30	32

由表 8 可见，方法 *SUDByTSN-VD* 通过扩展传统的自我网络，引入了访客类用户，增加了获取的相似用户的多样性. 同时，由于获取的用户都是按照相似度排名的，引入访客后，获取到了更加相似的用户. 对 4 种方法而言，p_{follower} 普遍较大，*SUDByText* 的平均 $p_{\text{follower}}=74\%$，*SUDBySN* 的平均 $p_{\text{follower}}=75\%$，*SUDByTSN* 的平均 $p_{\text{follower}}=78\%$，*SUDByTSN-VD* 的平均 $p_{\text{follower}}=56\%$，这说明了微博的相似用户在关注类用户中比例最大. 对方法 *SUDByTSN-VD* 而言，访客的比例(32%)稍大于粉丝的比例(30%). 实验的过程中，我们发现访客类用户的相似度之所以能够排在前面，主要是用户间的微博相似度比较大，有很多用户对某个用户 u_i 的微博进行了转发或者评论，但这些用户其实并不是用户 u_i 的关注或者粉丝. 这也进一步说明了在相似用户发现的过程中，访客类用户利用的优势. 再加上有些微博(比如新浪)开始限制用户获取非本人的关注和粉丝的个数，借助访客发现相似用户的思路更是值得

借鉴的.

对 5 个领域发现的相似粉丝类用户及访客类用户而言，由于“学术研究”和“军事”选取的用户领域比较狭小，“学术研究”和“军事”的粉丝类相似用户的比例分别为 34%和 36%，访客类相似用户的比例分别为 36%和 38%. 这同样说明了，对于领域范围较窄的用户，粉丝/访客既然对某用户进行了转发/评论，表明该粉丝/访客在朋友圈或者微博话题方面与此用户有较高的相似度.

5 总　结

社会网络是由作为节点的社会行动者及他们之间的关系构成的集合，其研究领域总体上分为个体网及整体网两大类别. 整体网可以根据节点或者节点之间的联系分为多种类型，比如朋友关系网、组织关系网、城市网、战略同盟网、产业链网等等. 个体网强调的是从单个节点或者连接边出发，分析节点或者连接边本身及周围的情况. 比如研究一个用户的情感倾向、兴趣爱好、关注、好友关系的分布、密切交互关系的年龄特征等等.

本文从个体网的角度出发，面向微博社交媒体，研究了微博中任意给定一个用户，从中挖掘出与其相似用户的群体，以作为线索发现、社区划分、好友推荐等工作的基础. 本文在研究过程中，所提的两个核心创新点：访客用户的利用、动态微博相似度及交互相关性计算，通过与已有方法的对比，在相似用户发现的多样性及准确性方面有了显著的改善.

对于该问题的研究，我们认为如下内容还需进一步加深：(1) 用户微博的相似度计算，可以考虑基于话题的微博相似度计算方法，其核心问题是微博媒体的话题提取，已有的 LDA、动量模型、速度增长、有意义串、分层次聚类等技术，需通过大规模的实验比较它们在微博话题提取中的应用效果；(2) 探寻更合理的微博用户的动态计算方法，研究基于非均匀划分、滑动窗口等策略的微博相似度及交互相关性计算技术；(3) 结合应用场景，研究可调整的微博用户相似度计算模型. 根据实际问题，在用户的属性选取、权重系数设置等方面设计不同的模型. 比如，挖掘同一研究领域的用户，用户的背景、发表的微博等属性就比较重要；研究关系密切的用户，用户间发表的微博、交互行为等属性就比较重要；(4) 基于大数据分析技术，对微博的自我网络进行更深层次的扩展，以发现更多的相似用户，进一步对数据进行统计显著性分析.

致　谢　审稿专家对本文提出了细致、富有建设性的修改建议；江苏金鸽网络科技有限公司为本研究提供了实验数据集. 在此一并致谢！

参　考　文　献

[1] Granovetter M. The strength of weak ties: A network theory revisited. Sociological Theory, 1983, 1(1): 201-233

[2] McPherson M, Smith-Lovin L, Cook J. Birds of a feather: Homophily in social networks. Annual Review of Sociology, 2001, 27(1): 415-444

[3] Bhattacharyya P, Garg A, Wu S H. Analysis of user keyword similarity in online social networks. Social Network Analysis and Mining, 2011, 1(3): 143-158

[4] Wang X F, Tang L, Gao H J, Liu H. Discovering overlapping groups in social media//Proceedings of the 10th IEEE International Conference on Data Mining. Sydney, Australia, 2010: 569-578

[5] Diaby M, Viennet E, Launay T. Exploration of methodologies to improve job recommender systems on social networks. Social Network Analysis and Mining, 2014, 4(1): 227

[6] Ma H, Zhou D, Liu C, et al. Recommender systems with social regularization//Proceedings of the 4th ACM International Conference on Web Search and Data Mining. New York, USA, 2011: 287-296

[7] Kantor P B, Ricci F, Rokach L, Shapira B. Recommender Systems Handbook. New York: Springer, 2010

[8] Walter F E, Battiston S, Schweitzer F. A model of a trust-based recommendation system on a social network. Autonomous Agents and Multi-Agent Systems, 2008, 16(1): 57-74

[9] Kim H N, Saddik A E. Exploring social tagging for personalized community recommendations. User Modeling and User-Adapted Interaction, 2013, 23(2): 249-285

[10] Xie X. Potential friend recommendation in online social network //Proceedings of the IEEE/ACM International Conference on Cyber, Physical and Social Computing and International Conference on Green Computing and Communications. Stockholm, Sweden, 2010: 831-835

[11] Cruz J D, Bothorel C, Poulet F. Entropy based community detection in augmented social networks//Proceedings of the International Conference on Computational Aspects of Social Networks. Salamanca, Spain, 2011: 163-168

[12] Qi G J, Aggarwal C C, Huang T. Community detection with edge content in social media networks//Proceedings of the 28th International Conference on Data Engineering. Washington, USA, 2012: 534-545

[13] Kahanda I, Neville J. Using transactional information to predict link strength in online social networks//Proceedings of the 3rd International Conference on Weblogs and Social Media. San Jose, USA, 2009: 74-81

[14] Hamid M N, Naser M A, Hasan M K, Mahmud H. A cohesion-based friend-recommendation system. Social Network Analysis and Mining, 2014, 4(1): 175-185

[15] Samanthula B K, Jiang W. A Randomized approach for structural and message based private friend recommendation in online docial networks. Lecture Notes in Social Networks, 2014: 1-34

[16] Xu Zhi-Ming, Li Dong, Liu Ting, et al. Measuring similarity between microblog users and its application. Chinese Journal of Computers, 2014, 37(1): 207-218(in Chinese)
(徐志明, 李栋, 刘挺等. 微博用户的相似性度量及其应用. 计算机学报, 2014, 37(1): 207-218)

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

[18] Peng Ze-Huan, Sun Le, Han Xian-Pei, Shi Bei. Micro-blog user recommendation using learning to rank. Journal of Chinese Information Processing, 2013, 27(4): 96-102 (in Chinese)
(彭泽环, 孙乐, 韩先培, 石贝. 基于排序学习的微博用户推荐. 中文信息学报, 2013, 27(4): 96-102)

[19] Gou L, You F, Guo J, Wu L, Zhang X L. SFViz: Interest-based friends exploration and recommendation in social networks //Proceedings of the Visual Information Communication-International Symposium. Hong Kong, China, 2011: 1-10

[20] Samanthula B K, Jiang W. Interest-driven private friend recommendation. Knowledge and Information Systems, 2015, 42(3): 663-687

[21] Modani N, Nagar S, Saswata S, et al. Like-minded communities: Bringing the familiarity and similarity together. World Wide Web, 2014, 17(5): 899-919

[22] Akcora C G, Carminati B, Ferrari E. User similarities on social networks. Social Network Analysis and Mining, 2013, 3(3): 475-495

[23] Korolova A, Motwani R, Nabar S U, Xu Y. Link privacy in social networks//Proceedings of the 17th ACM Conference on Information and Knowledge Management. Napa Valley, USA, 2008: 289-298

[24] Gao H, Hu J, Huang T, et al. Security issues in online social networks. IEEE Internet Computing, 2011, 15(4): 56-63

[25] Cutillo L A, Molva R, Onen M. Analysis of privacy in online social networks from the graph theory perspective//Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM 2011). Houston, USA, 2011: 1-5

[26] Aiello L M, Barrat A, Cattuto C, et al. Link creation and information spreading over social and communication ties in an interest-based online social network. European Physical Journal Data Science, 2012, 1(12): 1-31

[27] Aaron C, Nathan E. Persistence and periodicity in a dynamic proximity network//Proceedings of the DIMACS Workshop on Computational Methods for Dynamic Interaction Networks. Piscataway, USA, 2007: 1-5

[28] Zhong Zhao-Man, Zhu Ping, Li Cun-Hua, et al. Research on event-oriented query expansion based on local analysis. Journal of the China Society for Scientific and Technical Information, 2012, 31(2): 151-159(in Chinese)
(仲兆满, 朱平, 李存华等. 一种基于局部分析面向事件的查询扩展方法. 情报学报, 2012, 31(2): 151-159)

[29] Zhang J, Gao J F, Zhou M. Extraction of Chinese compound words: An experimental study on a very large corpus//Proceedings of the 2nd Workshop on Chinese Language Processing: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics. Hong Kong, China, 2000: 132-139

[30] Chakrabarti D, Kumar R, Tomkins A. Evolutionary clustering//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, USA, 2006: 554-560

[31] Chi Y, Song X, Zhou D, et al. Evolutionary spectral clustering by incorporating temporal smoothness//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, USA, 2007: 153-162

[32] Wang Li, Cheng Xue-Qi. Dynamic community in online social networks. Chinese Journal of Computers, 2015, 38(2): 219-237(in Chinese)
(王莉, 程学旗. 在线社会网络的动态社区发现及演化. 计算机学报, 2015, 38(2): 219-237)

[33] Hu Yun, Wang Chong-Jun, Wu Jun, et al. Overlapping community discovery and global representation on microblog network. Journal of Software, 2014, 25(12): 2824-2836(in Chinese)
(胡云, 王崇骏, 吴俊等. 微博网络上的重叠社群发现与全局表示. 软件学报, 2014, 25(12): 2824-2836)

[34] Wei Bing-Jie, Wang Bin. Time-aware mixed language model for microblog search. Chinese Journal of Computers, 2014, 37(1): 229-237(in Chinese)
(卫冰洁, 王斌. 面向微博搜索的时间感知的混合语言模型. 计算机学报, 2014, 37(1): 229-237)

ZHONG Zhao-Man, born in 1977, Ph. D., associate professor. His research interests include information retrieval and artificial intelligence.

HU Yun, born in 1978, Ph. D., associate professor. Her research interest is social network analysis.

LI Cun-Hua, born in 1963, Ph. D., professor. His research interest is data mining.

LIU Zong-Tian, born in 1946, professor. His research interests include artificial intelligence and software engineering.

Background

Currently, social media is considered as one of the most valuable information resources on the Web. Microblog is a social media with strong communicating ability and convenient operation, and users have established social circle similar to the reality on microblog. Microblog has attracted great attention of academia and industry circle due to strong communicating ability of information and complex network structure.

Existing methods of users' similarity calculation can be divided into three categories: (1) method based on the users' setting and microblog posts; (2) method based on social network with followers and fans; (3) mixed method, which has become the mainstream research method in the recent literatures. However, visitors have not been used to discover similar microblog users in relevant literatures, and time has not been fully considered in calculation of microblog similarity and interaction correlation, lacking further research on microblog dynamic social contact.

Innovative points of our work are mainly reflected in following aspects. Firstly, existing methods for discovering similar users only consider followers and fans, based on the use of social network on microblog. Visitors can reply and forward on articles on microblog of users. Therefore, visitors were introduced in the discovering method of similar users, with the proposal of Extended Ego Network to increase comprehensiveness and diversity of similar users. Secondly, in calculation of topic similarity and interaction correlation, existing calculation methods for users' similarity fail to reflect social dynamics of microblog social contact. Thus, dynamic constraints of time were introduced in the calculation method, better reflecting dynamics of microblog to discover more accurate similar users.

This research is supported by the National Natural Science Foundation of China (No. 61403156), the Prospective Joint Research of University-Industry Cooperation of Jiangsu (No. BY2015248) and the Six Talent Peaks Project of Jiangsu (No. XXRJ-013). The first aims to detect overlapping community on social network, the second is emphasized on event understanding and monitoring on the Internet, and the third focuses on event extraction and ontology construction. This work belongs to SNA part in these projects, especially users' similarity calculation.

Centering on these projects, our research team is mainly engaged in social network analysis, event representation, event ontology model and construction, event extraction, event retrieval and event crawling from 2007. Our research results are mainly published in the following journals: Journal of Software (Web news oriented event multi-elements retrieval), Chinese Journal of Electronics (Method of multi-topic crawling based on search strategy), Frontiers of Computer Sciences (Efficient multi-event monitoring using built-in search engines), International Journal of Machine Learning and Cybernetics (Event ontology reasoning based on event class influence factors), Journal of Chinese Information Processing (Model of event relation representation), etc.