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　15006）

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

**Measuring Similarity between Microblog Users and Its Application**

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

(School of Computer Science and Technology, Harbin Insitute of Technology, Harbin 15006)

**Abstract** Analyzing user interest and building user profile is very important for microblog’s userrelationship analysis, which is the fundamental work for social network fotmation and analysis. This paper mainly discusses approaches of microblog’s user relationship analysis. We viewmicroblog’s social network as a weighted undirected grah, where users are treated as nodeslinked by edges, and the weights of edges mean the relationship strengths between users. This paper defines user relationship strength as user similarity, and propose several user similarity esimation 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 acheieved the best performance. In addition, we tested them in user recommending experiments, and social information-based user similarity also got the best recommending results. Fominally 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

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 Netwotk 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]（L ink Prediction）；Kahand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

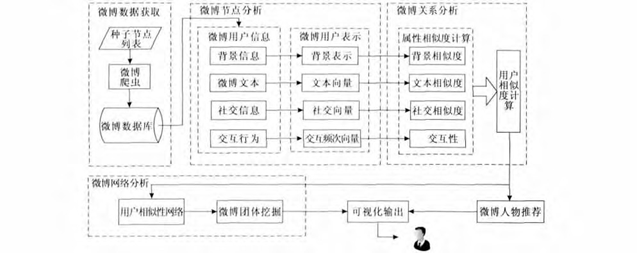
该部分主要阐述微博用户关系分析的技术包括可分为几个部分：数据获取、节点分析、关系分析、网络分析、信息推荐．将它们组合起来，形成一个微博用户关系分析的技术平台（如图1所示 ），本文下面分别详细介绍各个部分的工作原理. 

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｛ID1，ID2，…，IDm ｝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中出一个用户IDK．

(3)利用新浪微博API访问函数，抓取该用户节点IDK 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IDK，将其邻居节点(IDK的关注列表中的全部用户ID）加入Q，转到(2)．

**个人感悟**：科技论文的确是一个繁琐又仔细的工作，一篇优秀的科技论文不仅需要里面有足够的知识含量，还需要我们在写论文的时候，要仔细，认真，注意格式要求，并且注意美观，或许过程会有些繁琐，枯燥，但是这是我们必须要做到的。通过这次的模拟投稿，学会了很多，学习了论文的格式，明白了一篇稿件的发布过程，同时也把信息技术中学的word使用方法亲自的实践了一下，更加熟练地掌握了word的应用，也锻炼了自己的耐心，以及查找文章错误错误的能力，还提高了打字的速度，学习了一些计算机方面的理论知识，收获很多。俗话说：熟能生巧，实践出真知，确实应该通过实践才能找出自己所学的不足，并多实践才能熟练地掌握，在这次的过程中，我知道了英文跟数学公式的复制中容易出现乱码，我们需要多加注意，仔细核对，在必要的时候应该不怕麻烦，手打出来。