硕士学位论文

面向社会化媒体的社会网络挖掘与分析

SOCIAL NETWORK MINING AND ANALYSIS BASED ON SOCIAL MEDIA

杨方方

哈尔滨工业大学

2011 年 6 月

国内图书分类号：TP315　　学校代码：10213
国际图书分类号：681.3　　密级：公开

工学硕士学位论文

面向社会化媒体的社会网络挖掘与分析

硕士研究生：杨方方
导　　师：王宇颖教授
申请学位：工学硕士
学　　科：计算机科学与技术
所在单位：计算机科学与技术学院
答辩日期：2011 年 6 月
授予学位单位：哈尔滨工业大学

Classified Index: TP315
U.D.C: 681.3

Dissertation for the Master Degree in Engineering

SOCIAL NETWORK MINING AND ANALYSIS BASED ON SOCIAL MEDIA

Candidate: Yang Fangfang
Supervisor: Prof. Wang Yuying
Academic Degree Applied for: Master of Engineering
Speciality: Computer Science and Technology
Affiliation: School of Computer Science and Technology
Date of Defence: June, 2011
Degree-Conferring-Institution: Harbin Institute of Technology

摘 要

社会化媒体在网络信息异常充斥的时代对信息流动和交互起到了非常大的推动作用，基于社会网络构架的机制是社会化媒体之所以能够进行广泛而快速的信息流动和传播的机制和源泉，所以使用数据挖掘的技术对社会化媒体中隐含的社会网络进行挖掘并进行分析显得更为重要。对社会网络的自动挖掘有助于更清楚的了解社会化媒体中真正推动其发展的网络机制，并以良好的可视化方式呈现，对该社会网络的分析有助于人们更方便的了解网络情况，据此来支持信息推荐、信息检索、信息传播、电子商务、广告学等应用。

本文的主要任务是挖掘社会化媒体中的人物相关性网络，并对该网络进行简要的分析，同时给出了一个基于微博的社会网络应用系统。从数据挖掘、网络构建、网络分析及相关应用的角度来开展本文的研究，主要的工作分为以下三个方面：

第一，本文针对维基百科和新浪微博两种社会化媒体，抽取其中的人物实体信息并给出了一种统一的人物建模方法。本文分析了维基百科和新浪微博两种数据的特点，给出了相应的人物实体数据获取方法并对人物实体进行建模。

第二，本文给出了一组人物相关性的计算方法，并给出了人物网络的构建方法。根据数据的特点，选择适合的人物相关性方法进行组合即可得到人物之间的相关性计算结果，并通过网络构建方法构建相关的网络，本文通过分析这些方法在维基百科和新浪微博上的准确性证明了相关性计算方法的正确性。

第三，在相关性网络的基础上开展了团体挖掘的工作，并介绍了面向微博的社会网络应用系统。首先介绍了网络分析的概念，紧接着介绍了应用于团体挖掘的 GN 算法，相关的实验证明了人物相关性网络可以充分表示人物之间的关系。本文还给出了对应于本文工作的基于微博的社会网络应用系统和人物推荐的应用。

关键词：社会化媒体；社会网络；数据挖掘；相关性计算；团体挖掘

Abstract

Social Media plays a very positive role in promoting the flow and intertaction of information in this web information filled time. Social Network framework is the mechanism and source of fast flow and spreading of information.Therefore it is important to mine and analyse the social network hid in social media using data mining technology.And mining the social network automatically helped to reveal the cause driving the development of social media. Nice visualization of the network can give a clear view for the user, and it provides features for the user to support the application in information retrieval, information recommendation, information dissenemination, e-commerce, advertising and so on.

The main task is to mine the social network of people in social media, giving a brief analysis of the network and a microblog data-based social networking application.The paper carries out its research from data mining, network construction, network analysis and related application. The main tasks are as follows:

Firstly, this paper selects Wikipedia and Sina microblog as two samples of social media. We first extract the information of person entities and give a unified notion model of person entities, then give a brief introduction of Wikipedia and Sina microblog and the information acquisition meathod, and finally give out a notion model of person entities suited for these two media and we also give out the related experiments and the result of them.

Secondly, this paper proposes a group of person correlation computing methods and also proposes a network construction meathod. According to the feature of social media, we choose different methods or the combination of some methods to get person correlation, and then get the related network using the network constructing method. This paper uses the precision of these methods applied to Wikipedia and Sina microblog to show that the methods we proposed are correct.

Finally, based on the work of correlation network, this paper do some work on community detection, and give an introduction of the social nerwork application based on microblog. A short introduction of social network analysis and GN algorithm is provided, some experiments of community detection using GN algorithm shows that the correlation network is meaningful to present the relation between people. Besides, two applications are developed, and they are social nerwork application based on microblog and a person recommendation application.

Keywords: Social Media, Social Network, Data Mining, Correlation Computing, Community Detection

目 录

第 1 章 绪 论

1.1 课题的研究背景

随着 Web1.0 发展速度的跌落，社会化媒体（Social Media）渐渐进入人们的视野，人们开始以这种更适应信息传播的方式来表达自己的观点和需求，社会化媒体渐渐成为社会一个主流，在互联网和商业领域受到了高度的关注，但社会化媒体却不是刚刚兴起，也并不是一个新的概念。

社会化媒体并不是伴随着计算机的出现才出现的，早在 20 世纪 50 年代出现的电话就是一种最早的社会化媒体形式；播客和 BBS（Bulletin Board System）是非常典型的社会化媒体形式，而最早出现的播客是语音邮件系，最早出现的 BBS 在 1979 年对公众开放；最早的 BBS 是以一个连接到电话调制解调器上的 PC 机作为服务器的系统，这个时期的 BBS 提供在信息板上的社会化交流，基于社区的文件下载和在线游戏等。随着 1991 年 WWW（World Wide Web）的出现，网络论坛开始流行并开始代替 BBS 等传统的方式；1996 年 Israeli 技术专家提出了即时通信（IM）系统，其在电脑上的应用叫做 ICQ，ICQ 很快被 AOL 收购并变成一个主流，IM 技术成为社会化媒体发展过程中的一个很大的推动力量；Naspter 是在 1999 年被提出来的 P2P 的文件共享应用；然后出现的就是社会网络和社会化新闻网站，第一个出现的社会网络网站是 Sixdegrees，这个网站在 1997 年就允许用户描述自己并和朋友进行交流，这类的可交互的社会化的网站应用类型被广泛称为 Web2.0；紧接着比较有名的是 2001-2003 年的 Friendster，然后是出现在 2003-2006 年的 MySpace，出现在 2007 年的 Facebook；Digg，StumbleUpon，Reddit 都在社会化新闻领域获得了令人咋舌的成绩；Delicious 也因其对静态页面的共享书签而非常著名；2008 年出现的实时状态和基于位置的社会网站，这些网络的最大的特点就是关注的不再是静态的网页，而是一些实时的状态更新的信息流，Twitter 就是一个最典型的应用。

互联网自诞生那一天起就没停止过社会化进程，从简单的 Email 到现在关系复杂的社交网络，互联网越来越接近人类社会结构和交流方式，媒体方式也越来越高级。无论是 Facebook 、MySpace、YouTube 还是 Twitter 都是非常成功的社会化媒体，这也告诉我们互联网社会化是一种必然需求。

大部分社会化媒体的内在结构都呈现社会网络的特性，开展对社会网络的研究至关重要，一方面，社会网络对维持社会化媒体的信息传播、流动、网络

的稳定性和社会化媒体的其他属性是至关重要的，另一方面，社会网络的分析技术可应用于多样性的社会化媒体，支持媒体信息的聚合、推荐、搜索等服务技术。

1.2 社会媒体和社会网络介绍

1.2.1 社会化媒体简介

社会化媒体（Social Media）是一种给与用户极大参与空间的新型在线媒体，博客、维基、播客、论坛、社交网络、内容社区是具体的实例。参考 1.2.2 节中对社会化媒体的分类情况，可知社会化媒体的种类和形式非常繁多，但是无论其具体的形式有何不同，都具有相同的特征：参与性、公开性、交流性、社区化和连通性。

社会化媒体允许用户在网络媒体上发布自身的信息，还支持群体协作编辑、发布、分享、传播信息。近年来，社会化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，并正在成为网络技术发展的热点和趋势。社会化媒体与传统的 Web1.0 媒体的主要区别在于：媒体信息的来源不同，Web1.0 媒体信息是由编辑人员发布，而社会化媒体凝聚大众的群体智慧，由大众发布媒体信息，被称为 Web2.0 媒体；媒体信息的类型不同，Web1.0 媒体信息主要是资源即网页，而社会化媒体信息包括两种实体：用户和资源；媒体信息的网络结构不同，Web1.0 媒体的网络结构是由大量的网页及其链接构成的超链结构，而社会化媒体的网络结构更复杂，是由大量的用户、资源及其关系构成的社会网络；媒体挖掘的关键技术不同，Web1.0 时期，Google 挖掘了 Web1.0 媒体的网络结构，以面向大规模网页的超链分析技术引领了 Web1.0 信息服务的技术潮流。

1.2.2 社会网络简介

人们使用社会网络的概念已经超过一个世纪，通常使用社会网络的概念来记录一个大范围社会系统中成员之间关系的复杂集合，在 1954 年，J.A.Barnes 开始使用系统性的词汇来定义关系之间的模式，包含一些传统中曾经被大众和社会学家使用的概念：有界组（如部落，家庭）和社会范畴（如性别，种族）。

社会网络指社会行动者及社会行动者之间关系的集合[1]。社会网络这个概念强调每个行动者都与其它行动者有一定的关系，但却没有限制关系的类型，关系指的是节点和节点之间的连接，可以从很多不同的方面反映人物之间的关系，比如如果两个人相互认识，则称他们之间具有认识关系，那就可以根据认

识关系建立相关的认识网络，在比如如果两个人有共同的兴趣和属性，就可以说他们之间具有兴趣相似的关系，同时也可据此建立一个人物兴趣的相似性网络。所以人和人之间的关系只有在特定的环境下才会得以明确，比如人之间的好友关系，亲戚关系或联系关系等。

常用的社会网络表示形式是一个图 G =(V , E)，V 节点集合，E 是边集合。本文仍然使用图的形式来对社会网络进行表示，每个节点代表一个人物，而每条边代表两个人物之间的关系，如果关系是有向的，则可使用有向图来进行表示，否则使用无向图表示，如果两个人之间的关系有强弱，则可使用边上的权值来代表关系的强弱，也对应了带权图和无权图。

很多领域的研究者都在研究社会网络，同时社会网络也不再局限于传统社会学家所研究的真实人物网络范畴之内，比如一些物理学家，动力学家和传播舆论家都在讨论复杂社会网络，每个领域的研究都有其不同的侧重点，本文主要研究人物在虚拟社会化媒体中所形成的复杂网络，这对改进社会化媒体的用户体验，研究网络内在的传播规律意义重大。

表 1-1 社会化媒体分类表

类别	社会化媒体：示例
社区交流 资源分享	即时通讯（Instant Message, IM）：MSN、QQ 论坛（Forum）：各种各样的主题论坛 博客（Blog）：Blogger、新浪博客. 微博（Micro Blogging）：Twitter、新浪微博 在线社会网络 (Online Social network)：FaceBook、校内网 信息聚合（Information Aggregating）：Netvibes 资源分享网站：YouTube、Flickr、SlideShare、百度文库
群体协作	维基（Wiki）：维基百科、维基解密 社会化标签（Social Bookmarking）：Delicious 社会化新闻（Social News）：Digg 社会化导航（Social Navigation）：Trapster 社会化问答（Community Q&A）：百度知道、Yahoo! Answers 社会化评论：epinions.com、口碑网 内容管理（Content Management）：Wordpress 文档管理编辑工具（Document Management and Editing Tool）：Google Docs

社会化媒体种类繁多，表 1-1 给出了一个社会化媒体的基本分类，本文把

社会化媒体分为两大类：社区交流、资源分享类媒体和群体协作类媒体，每大类都给出了一些小类别，每个小类别中都包含了对应的几个常用的社会化媒体示例。

1.3 社会网络的研究现状

1.3.1 社会化媒体上的社会网络研究

（1）社区交流、资源分享类媒体

社区交流、资源分享类媒体包括即时通讯、论坛、博客、微博、在线社会网络社区、资源分享网站等，这类社会化媒体是由大量用户组成虚拟的网络社区，实现彼此在线交流。

即时通讯是一种即时性网络通信工具，用户之间通过媒体相互在线交流。在即时通信的社会网络研究方面，Smith 发现即时通信的社会网络呈现小世界特性和无标度特性[2]；Yao 发现即时通信的社会网络的度分布符合幂率分布[3]；Leskovec 研究了微软 MSN 的用户分布、用户的活跃度和用户行为的时间特性等[4]，他具体分析了 MSN 的两个不同的社会网络：活动性网络和好友网络，前者是由用户之间实际的交流行为形成的弱关系网络，后者则是根据好友定义形成的强关系网络。

博客的数据由大量的用户发表的博文、留言评论组成。博客呈现出一定的社会网络特性，用户通过博文阅读、好友定义、留言评论等交互行为，形成用户关系，大量的用户组成了社会网络。博客相关的研究热点是将兴趣相似的用户聚合起来，自动发现博客社区。研究者分别提出了基于博文的内容分析、用户参与度分析以及超链分析等三种不同的方法来发现博客社区。Nardi 融合了前两种方法，分别根据博文的内容、用户评论信息来分析博文的主题和用户参与度[5]；Kumar 采用博客的超链分析技术来发现潜在的博客社区[6]；Christopher 研究了基于标签的博客社区的聚合问题[7]；Chen 采用博客的社会网络分析方法，研究了博客推荐问题[8]；Flora 研究了博客的新颖性和冗余性检测问题[9]。

相比博客，微博媒体组成了更紧密的社会网络，交互性更频繁，信息扩散速度更快、传播范围更广，用户数也呈现爆炸性增长。针对微博信息的新颖性、实时性特点，微软、Google、Twitter 等公司分别研究了实时搜索技术。A. Java 等研究了 Twitter 的用户和微博的地理分布特性[10]；Ye 研究了 Twitter 信息的聚合、权威性计算问题[11]；Weng 发现了 Twitter 的用户兴趣呈现同质性现象，他将用户兴趣相似性和社会网络的结构分析结合起来，提出了一种主题相关的 PageRank 改进算法，称作 TwitterRank 算法[12]，用于计算 Twitter 用户的权威

性；Kwak 在 Twitter 整体数据上考察了 Twitter 的社会网络的拓扑结构，发现用户的关注者（Follower）符合非幂率分布[13]，也考察了基于关注者数目、基于 PageRank 算法、基于回复数目等三种 Twitter 用户权威性排序算法，发现前两种方法排序结果相似，第三种则差异较大。

近年来，大量的 SNS（Social Network Service）站点，例如 FaceBook、Myspace 等崛起，形成了大规模在线社会网络社区（Online Social Network）。Fu[14]、Golder[15]、Ahn[16]、Yuta[17]、Mislove[18]分别研究了校内网、FaceBook、Cyworld、mixi、orkut 等社会网络的拓扑结构和网络测量问题；Gjoka[19]、Ahn[16]分别在 FaceBook、MySpace 上研究了社会网络的数据采样问题，前者提出了一种基于 Metropolis-Hastings 随机游走的均匀采样方法，后者则研究了雪球采样（Snowball Sampling）方法；Chun[20]研究了 Cyworld 中两个不同的社会网络：活动性网络和好友网络，发现两个网络具有很高的相似性。

资源分享网站，如图片类的 Flickr、书签类的 Del.icio.us、视频类的 YouTube 等，是一种支持用户发布、分享、检索多媒体资源的网站。Cha[21]研究了 Flickr 中的网络结构和信息扩散问题，Mislove 系统地比较了 Flickr、LiveJournal 和 You Tube 等三个资源分享网站，发现它们的结构具有很高的相似性。

（2）群体协作类媒体

群体协作类社会化媒体主要包括维基、社会化标签、社会化新闻、社会化导航、社会化问答、社会化评论、内容管理、支持群体的文档管理编辑工具等。这些社会化媒体的共性是凝聚了大规模用户的群体智慧，共同创造内容。Digg 是一种允许大众用户组织、提交新闻的网络媒体，Lerman 研究了它的社会网络和信息过滤技术[22]。社会化问答的研究主要集中在最佳答案的匹配问题上[23]。社会化标签系统（Social Tagging System）支持群体用户对网络资源通过标签进行定义、分享、组织、管理和搜索。标签的研究包括几个方面：①基于标签的用户模型[24-26]，简要的讲是：用户的标签包含用户个性化兴趣信息，资源的标签反映了资源的主题，因此标签被用于建立用户模型、资源模型；②基于标签的相关性计算[27]，标签之间存在着语义关系，根据所使用的标签，可以分析用户之间、资源之间、用户与资源之间的相关性，进行信息推荐；③基于社会化标签的应用，目前标签技术已被应用于媒体信息的搜索[28]、分类[29]、推荐[30]等多个领域。

综上所述，各种社会化媒体的内在结构都呈现了社会网络特性，大部分社会化媒体挖掘的研究都在不同程度上涉及了社会网络的相关技术，这也是本文选择社会化媒体中的社会网络作为研究对象的原因。

1.3.2 社会学方面的社会网络研究

社会网络的研究最初起源于社会学家对人群社会结构的研究，Moreno 利用图论研究了社群的结构，提出了社群图的概念[31]；Lewin 利用拓扑学和集合论分析了由群体构成的社会空间的结构特征[32]，他着重研究了群体及其周围环境的关系；Cartwright 等人进一步研究了群体的凝聚力、合作、权力和领导等模型[33]；Barnes 将整个社会生活看成由代表个人的节点、代表个人关系的边构成的社会网络[34]；哈佛大学的 White 等人研究了社会结构的数学模型[35]，这一成果后来推动了国际社会网络分析网（International Network for Social Network Analysis）的出现；Granovetter 研究了谋职问题中的信息扩散模型[36]，他发现弱关系人群（熟人）比强关系人群（亲属、同事等关系）更容易传递求职信息。综合地看，社会学方面的社会网络研究分析了小规模人群的人际关系和社会结构，他们采用邻接矩阵和社群图表示社会网络，利用一些如密度、中心度、模块度等的网络指标来挖掘网络中的团体、明星节点等。

1.3.3 网络挖掘方面的社会网络研究

早在十几年前就已经出现了各种形式的社会网络抽取方法和系统，近年来，随着图论、概率论和各种几何学的发展和完善，社会网络分析作为一种应用性很强的社会学研究方法越来越受人瞩目。

Kautz et al.[37] 开发了一个叫做 Referral Web 的从网络上抽取信息的社会网络抽取系统。该系统通过使用搜索引擎关注网页中的人名的共现次数来衡量人物之间的紧密度，系统通过在搜索引言中搜索“X and Y”并使用返回条目的次数来估计人物 X 和 Y 的紧密的程度：如果 X 和 Y 的关系比较紧密，那么搜索引擎便返回更多关于 X 和 Y 共同的信息，系统就会自动的建立一个 X 和 Y 之间的边。

P. Mika[38] 开发了一个从语义网络社区抽取、集成和可视化在线社会网络的系统，叫做 Flink。在该系统中，社会网络是通过对网页、邮件信息、出版物和个人日志信息的分析而获得，Flink 中的网络挖掘组件同样使用共现分析的方法。

A. McCallum[39, 40] 和他的团队开发了一个端到端的系统，这个系统可以抽取一个用户的社会网络。该系统首先从邮件信息中识别出一个特定的人物集合并找到这些人物的主页，同时在他们对应的联系人中填上其他人的名字，如此处理之后，主页的所有者和在这个页面中发现的名字就存在对应的联系关系。

1.3.4 复杂网络方面的社会网络研究

近年来，复杂网络的研究逐渐兴起，研究者发现了它的小世界特性和无标度特性。社会网络是一种现实的复杂网络，1967 年哈佛大学的 Milgram 研究了它的小世界网络特性，提出了著名的“六度分离理论”；复杂网络的研究大部分集中在网络的拓扑结构分析、测量，网络测量的指标一般有密度、度分布、聚类系数、平均路径长度、度相关系数、介数、互惠指数、模块度等。研究者发现社会网络具有很强的社区结构，社区内部的成员之间关系紧密，而社区之间关系稀疏。

1.4 课题研究内容及意义

前面介绍了社会化媒体中关于社会网络的研究，社会网络的挖掘、分析，社会学中社会网络的研究和复杂网络等领域中的社会网络的研究情况，综合地看，上述的研究工作提供了很多有价值的社会网络的分析技术和研究思路。但随着社会网络的用户、资源的规模不断增长，面对大规模的社会网络的团体挖掘、节点权威性计算、信息推荐、搜索等任务，这些方法的适用性还有待提高，面对新的社会化媒体，以前的许多方法不再适用。

社会学关于社会网络的研究侧重于分析小规模人群的人际关系和社会结构，社会学家提出的社会网络相关的概念和研究方法，至今还深刻影响着社会网络的研究，但社会学研究的局限性在于缺乏对大规模社会网络的研究经验。

复杂网络的相关研究侧重于社会网络的拓扑结构、演化、动力学等网络特性测量，很少涉及社会网络的内容、结构的深度分析，一般很难用于信息的推荐、搜索等。

社会化媒体上关于社会网络的研究大多涉及了社会网络的分析技术，但是大部分研究是在不同的、单一的社会化媒体上进行的，彼此之间缺乏联系和比较；很多研究是围绕着某种社会化媒体自身的数据特点而展开的，社会网络的通用性分析技术尚缺少系统性的研究。

针对上述存在的问题，本文开展了对基于社会化媒体的社会网络挖掘与分析技术的研究，其中社会网络挖掘包括人物实体信息抽取，人物信息模型化表示，社会网络构建，主要的工作分为四个部分：首先是基于维基百科和微博的人物实体数据获取方案的设计；其次是人物实体模型化表示，给出了统一的人物实体模型化表示方法，并把在维基百科和新浪微博上进行验证；再次是给出了一组计算人物实体之间关系强度的方法，并在人物关系强度的基础之上构建人物之间的网络；最后本文对构建出来的社会网络进行网络分析，主要进行的

网络分析工作是团体挖掘，并介绍了一个面向微博的社会网络系统，该系统是对本文方法的实际应用，并开发了一个可以可视化人物关系网络的工具，以一种良好的方式对各个阶段的不同类型的结果进行展示。

不同于 Web 1.0 以编辑者编辑新闻的信息发布方式，社会化媒体以每一个用户为中心进行信息发布，社会化媒体已经成为人们更喜欢的网络在线方式，所以对社会化媒体的研究变得更加迫切，研究怎么样能够给用户提供更好的网络在线体验，怎么样能够体现每个用户的个性化特征，而不用在阅读大量无用信息之后可以找到自己感兴趣的信息。社会网络作为社会化媒体的骨架，是社会化媒体能够活起来、火起来的源泉，社会网络的一些适合于信息交流和信息传播的特性在社会化媒体中得到了非常好的体现，所以对社会网络中所存在的社会网络的自动挖掘和分析就显得非常重要，社会网络挖掘是指使用一定的数据挖掘方法从社会化媒体中挖掘出其中潜在的社会网络，社会网络分析对该网络进行一定的分析，得到网络特性，把这些分析的结果用来更好的改进社会化媒体的用户体验和灵活的信息传播上。

1.5 内容组织结构

对于本文的工作，本文的组织如下：

第 1 章介绍了本文的研究背景，对社会化媒体和社会网络进行了介绍，关于社会网络的研究现状并对其进行了一定程度的分析，然后介绍了本文的主要研究内容及目的、研究意义，最后给出了本文的组织结构。

第 2 章给出了人物数据的获取方法和人物信息模型化表示的方法。首先分别介绍了维基百科和新浪微博两种社会化媒体并给出了对应的数据获取的方法，然后给出了一种适用于社会化媒体的人物实体建模方法，最后给出了具体的实验结果，并对实验结果进行了简要的分析。

第 3 章主要研究了人物实体关系强度的计算方法和网络构建方法。在简要介绍了相关性的概念之后，分别介绍了本文提出的五种针对不同数据类型的人物相关度计算方法：基于系统相似原理的相关性计算方法、基于余弦相似度的相关性计算方法、基于交互频度的相关性计算方法、基于人物关系相似性的相关性计算和基于线性融合的人物实体相关性计算方法。最后在上述相关性计算结果的基础之上，构建出人物之间的相关性网络，给出了实验结果并进行了详细分析。

第 4 章在相关性网络的基础上，对该网络进行分析，并给出了基于微博的社会网络应用系统。本章进行的社会网络分析主要指团体挖掘分析，给出了该方法分别应用在维基百科数据和新浪微博数据上的实验结果，并对实验结果进

行了评价，最后给出了一个基于微博的社会网络应用系统，介绍了该系统的设计原理，可视化模块中的可视化工作，对人物关系网络，人物相关性网络，人物的模型化表示结果都进行了直观了展示，同时介绍了人物推荐应用的设计原理。

第 2 章 人物实体建模研究

2.1 维基百科和微博简介

（1）维基百科介绍

按照第一章对社会化媒体的分类，维基百科（Wikipedia）属于群体协作类社会化媒体，其主旨思想就是允许多人通过群体协作进行知识共享，维基百科是一个自由、免费、内容开放的百科全书协作计划，参与其内容编辑和管理的人来自世界各地。维基百科是一个基于 wiki 技术的多语言百科全书协作计划，也是一部用不同语言写成的网络百科全书，是一个动态的、可自由访问和编辑的全球知识体，也被称作“人民的百科全书”，维基百科的宗旨是“人人为我，我为人人”。

维基百科由吉米·威尔士等人在 2001 年 1 月 15 日创立，维基百科是多语言的，其条目最多的语言是英文，截至到 2011 年 5 月 31 日共有条目 3,647,446 条，累计编辑次数 464,839,590 次。中文维基百科是维基百科协作计划的中文版本，自 2002 年 10 月 24 日正式成立，由维基媒体基金会负责维持，截至 2011 年 5 月 31 日，条目已达到 357,180 条，累计编辑次数达 16,974,735 次，以每天增加 132 条的趋势在增加（来自维基百科统计信息）。

维基百科中的页面是经过非常多的人物进行合作编辑的，包括对人物的非常全面的介绍，所以信息具有很高的可信度，但是维基百科中所给定的人物词条却不存在任何的显式关系，即没有强关系的制约，所以本文希望能够通过一定的方法提取出维基百科中人物之间的关系网络，对其数据进行挖掘得到的网络是非常真实可信的。

维基百科中包含了丰富的半格式化数据，即维基百科页面中的 Infobox 中提供的模板数据，这些数据很方便抽取且具有良好的准确性，本文研究的对象是人物网络，维基百科中每个人对应的页面中含有丰富的人物正文信息和丰富的词汇链接，图 2-1 是艾伦·麦席森·图灵在维基百科中的部分截图，其中右边框内的是模板信息，左边框内是正文信息，而页面中加粗的字体是一些锚文本，可以直接链接到维基百科中的该概念对应的页面。

在维基百科中，每个人物都对应一个页面，如果出现同名现象，维基百科提供一个重导航（Redirect）页面，这个页面可以导航人物到其准确的页面中去，所以重名问题在维基百科中不需要考虑，这就克服了使用搜索引擎构建网络时遇到的重名问题。

艾伦·麦席森·图灵，OBE， FRS（英语：Alan Mathison Turing；也常翻译成**涂林**或者**杜林**，1912年6月23日－1954年6月7日，英语发音[ˈælən ˈmæθɪsn ˈtjʊərɪŋ]），英国数学家、逻辑学家，他被视为计算机科学之父。

1931年图灵进入剑桥大学国王学院，毕业后到美国普林斯顿大学攻读博士学位，二战爆发后回到剑桥，后曾协助军方破解德国的著名密码系统Enigma，对盟军取得了二战的胜利有一定的帮助。

图灵对于人工智能的发展有诸多贡献，例如图灵曾写过一篇名为《机器会思考吗？》（*Can Machines Think?*）的论文，其中提出了一种用于判定机器是否具有智能的试验方法，即图灵试验。至今，每年都有试验的比赛。此外，图灵提出的著名的图灵机模型为现代计算机的逻辑工作方式奠定了基础。

图灵是著名的同性恋之一，并因为其同性恋倾向而遭到迫害，这使得他的职业生涯尽毁。他亦患有花粉过敏症。

图灵还是一位世界级的长跑运动员。他的马拉松最好成绩是2小时46分3秒，比1948年奥林匹克运动会金牌成绩慢11分钟。1948年的一次跨国赛跑比赛中，他跑赢了同年奥运会银牌得主汤姆·理查兹（Tom Richards）。[1]

目录 [隐藏]

1 孩童和年轻时代
2 大学和可计算性的工作
3 早期的计算机研究：图灵试验
4 图案形成和数理生物学的研究
5 迫害和逝世
5.1 平反
6 参见
7 注释
8 外部链接

孩童和年轻时代 [编辑]

图灵的父亲朱利斯·麦席森·图灵（Julius Mathison Turing）是一名英属印度的公务员。图灵的母亲Ethel1911年在

艾伦·麦席森·图灵

出生	1912年6月23日 英国伦敦
逝世	1954年6月7日 英国
国籍	英国
研究领域	数学，逻辑学，计算机科学
任职	剑桥大学 曼彻斯特大学
母校	剑桥大学 普林斯顿大学
博士导师	阿隆佐·邱奇
著名成就	图灵机，图灵试验

图 2-1 维基百科中关于图灵的页面

（2）微博介绍

微博是目前在国内非常受欢迎的社会化媒体类型，其广泛的用户参与性，简单的操作特点，简洁的页面布局，快速的信息传播方式和良好的网络构架吸引了大批量的用户。微博即微博客（MicroBlog）的简称，是一个基于用户关系的信息分享、传播以及获取平台，用户可以通过 WEB、WAP 以及各种客户端组件个人社区，以 140 字左右的文字更新信息，并实现即时分享。

最早也是最著名的微博是美国的 Twitter，Twitter 是 2006 年 3 月由 blogger.com 的创始人威廉姆斯推出的，英文原意为小鸟的叽叽喳喳声，用户能用如发手机短信的数百种工具更新信息。Twitter 是一个社交网络及微博客服务，用户可以经由 SNS、即时通信、电邮、Twitter 网站或 Twitter 客户端软件输入最多 140 个文字进行更新。Twitter 于 2006 年 3 月 31 日上线，当时注册人数为 100 人，2011 年 4 月，Twitter 注册人数突破 2 亿，新用户数以每天 50 万递增。

国内存在很多微博平台，新浪微博是国内最早出现的微博服务，新浪微博于 2009 年 8 月 28 日上线，其微博开放平台也是开放最早的；腾讯微博是第二大用户群的微博网站，在 2010 年 4 月 1 日上线，其借助于 QQ 用户的粘性，在短期内吸引了大量的用户；网易微博于 2010 年 1 月 20 日上线，搜狐微博在 2009 年 12 月 14 日上线。除了这些主流网站推出的微博服务，还有很多新型的微博系统，比如和讯微博是和讯网推出了具有财经特色的微博产品，可以随时随地的使用和讯微博分享财经世界的各种动态和热门话题；职微博是国内第一家专注于职场交流的微博。

图 2-2 是 bShare 公布的 2010 年 12 月份的微博服务平台的分享量排行榜，可以看出四大主流微博占领了主导地位，而新浪微博是注册用户最多并且使用最为广泛的，也是最早开放其开放平台并允许应用开发者通过 API 使用其数据并建立应用的微博服务，所以本文把新浪微博作为另外一个数据来源，微博的设计理念和结构是都是非常类似的，所以本文针对新浪微博的方法多可以很轻松的移植到其他的微博平台上。

2010年12月份微博类平台分享量排行榜

Rank	平台	%
1	新浪微博	45.88%
2	腾讯微博	13.84%
3	搜狐微博	7.48%
4	网易微博	6.58%
5	Follow5	6.11%
6	嘀咕网	4.58%
7	做啥网	4.51%
8	Twitter	3.17%
9	天涯	2.49%
10	9911微博客	1.71%
11	饭否	0.91%
12	人间网	0.90%
13	139社区	0.77%
14	同学微博	0.47%
15	中金微博	0.40%
16	噗浪网	0.20%
Total		100%

数据来源：www.bShare.cn, 2010.12

图 2-2　bShare 的微博平台分享量排行榜

图 2-3 是李开复的新浪微博页面的截图，个人资料区域包含了其注册信息，如昵称，网址，地理位置，博客地址，称谓，头像等；右上角包含了关注数、粉丝数和微博数，通过点击超链接可得到对应的关注人物列表、粉丝人物列表和微博列表；左侧正文区域包含了他所发布的所有微博（包括文本，多媒体信息如图片视频音频等等）；右侧中部包括个人标签。在新浪微博中人物之间存在两种关系：关注关系、粉丝关系，如果 A 关注 B 则 A 的页面中可以看到所有 B 发布的信息，同时 A 也是 B 的粉丝，B 的主页上不能看到 A 的信息，但是如果在 A 关注 B 的同时 B 也关注 A，则两人都可以看到对方发布的微博。在每个人物发布的微博下面都有转发，评论和收藏按钮，如 B 发布了一条微博，A 可以通过转发动作可以把 B 的微博以自己的名义再发一遍，这样的话 A 的所有粉丝都可以看到这条微博，同时 A 可以对 B 的这条微博进行评论，并且 B 也可以回复 A 的评论，A 也可以通过收藏动作把该微博收藏起来，供以后查看。

图 2-3 李开复在新浪微博页面

2.2 数据获取

（1）维基百科数据获取

维基百科作为全球最大的百科全书，包含了大量的词条信息，维基百科官方对其所有的信息都进行了分类存储，同时在其官方网站进行了公开(http://dumps.wikimedia.org/zhwiki/)，这些信息以不同的文件形式进行存储，并定时更新以保证使用者可以拿到最新的维基百科数据。本文选用中文维基百科网站作为本文的研究对象，首先从维基百科官方网站中下载对应的中文数据文件，本文选择 2011 年 05 月 02 日更新的数据文件，解压后的 XML 文件大小为 1.7G，所以采用 SAX 方式对此 XML 文件进行解析。

下载的数据包中存储着维基百科中文版的所有页面信息，包括各种分类的条目列表，当然也包括所有的人物信息，每个人物都在维基百科中对应一个页面，每个页面都以其人物名称作为区别，提取数据信息时首先给定一个待提取信息的人名列表，值得说明的是，本文选取人名时根据维基百科中的分类列表中的人物分类进行选取的，这么做的好处有两个：一是，维基百科中的页面是根据人名进行区分的，同一个人物可能存在不同的名字，而从维基百科列表中选择则会保证选取的人名一定是和维基百科中条目对应的，不至于出现歧义的

情况；二是，在分类列表中存在的人物多数在维基百科中都存在对应的页面，这样选取人物比较准确，不会出现很多人物实体信息提取失败的情况。

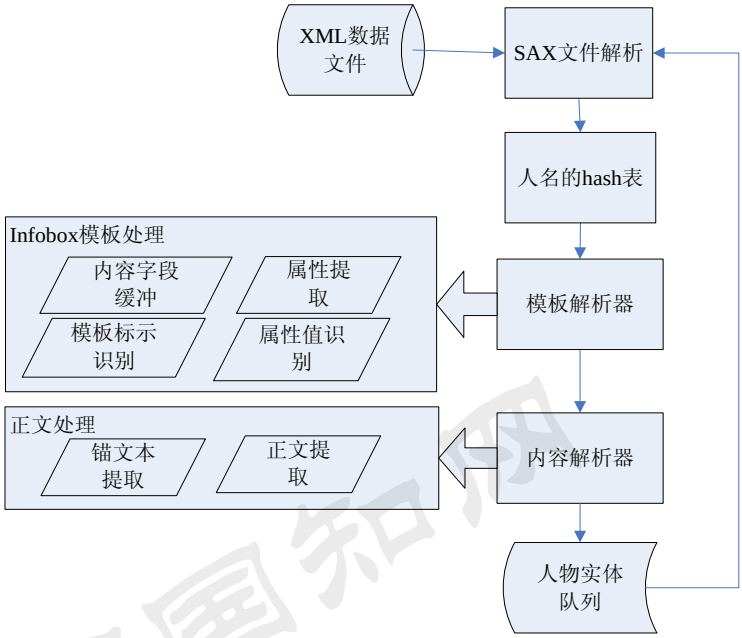


图 2-4 维基百科数据获取示意图

图 2-4 给出了获取维基百科人物数据获取示意图，人物信息获取时，根据给定的人名列表，依次从数据包中找到对应的人物信息，并从该文件中解析出人物对应的信息，解析时需要对人名建立一个 hash 表，以快速的定位人物信息，通过模板解析器和正则表达式的方式从 Infobox 和正文内容中提取出人物页面 Infobox 信息，正文信息和锚文本信息，并把此信息作为人物实体信息存入对应的人物信息队列中，关于维基百科中人物页面的具体语法结构可以参考文献[41]中的介绍。

（2）新浪微博数据获取

新浪微博的数据获取方法和维基百科完全不同，获取新浪微博数据的方式是通过新浪微博开放平台（(http://open.t.sina.com.cn/)）提供接口进行获取，新浪微博开放平台是一个基于新浪微博客系统的开放的信息订阅、分享与交流平台，微博开放平台提供了海量的微博信息、粉丝关系、以及随时随地发生的信息裂变式传播渠道。新浪微博向每个使用者都开放了其开放平台，每个使用者都可以使用新浪微博开放平台向外开发的一组 API 来获取给定格式的数据，这组 API 对应了一组封装良好并且使用方便的操作，这些操作能够让使用者快速的获取新浪微博的用户数据。

新浪微博 API 是获取数据的接口，基本的数据获取方式是 HTTP 请求的方式，默认以 XML 或 JSON 的方式返回用户数据信息。另外，新浪微博在其开放平台提供的基本 API 基础之上，又封装了不同语言的 SDK，主要有 Java、

PHP、Javascript、Flash 等，可根据具体应用的不同，选择不同的开发语言。新浪微博开放平台同时也提供开发文档，文档中详细给出了每个接口的具体使用方法。在数据请求阶段涉及到权限验证、安全及权限控制的问题。

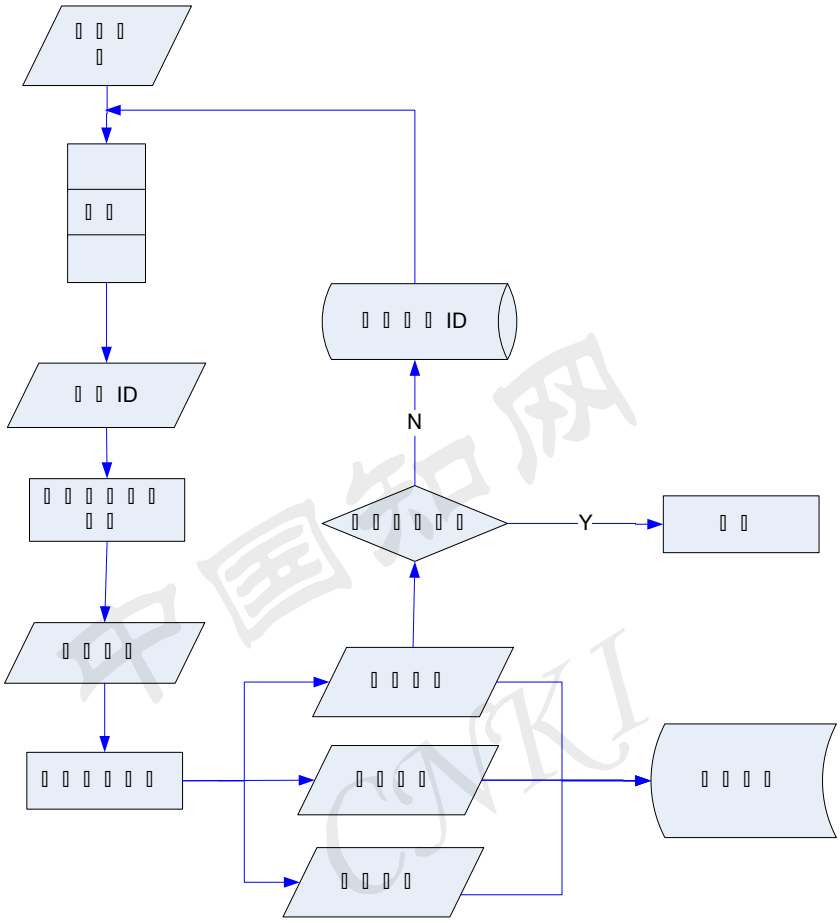


图 2-5 微博人物数据获取示意图

图 2-5 给出了微博爬虫的工作原理示意图，新浪微博中的人物数据并不像维基百科那样以彼此不相关的方式存储，微博中人物之间存在特定的关注和粉丝关系，为了能够最大程度的符合人物的真实微博网络，选取的人物节点更规范，使用雪球取样的方式进行取样，首先给定若干个种子节点，把这些节点放入一个待爬行节点队列中，对队列中的节点依次进行信息爬取，对每一个选定的节点，首先爬取该节点的信息，并把该节点对应的各类信息存入数据库，需爬取的人物信息包括人物的注册信息、人物标签信息、人物关注和粉丝列表、人物所发表、转发和评论的微博信息等，同时把该节点的邻接点（根据人物关注关系确定）放入待爬取队列中，以此类推，直到待爬取队列为空或达到某个控制阈值则停止，此阈值可以设定成很灵活的形式，比如爬行的层数、爬行的节点数或者爬行的时间等。

实验中可以根据需要选定一组初始的种子节点，整个微博爬虫工作的结果

就是得到一些以这些种子节点为中心的局部网络，本文的工作也正是在这样的一个网络基础上进行的。另外，把爬取到的数据存入了数据库中，对数据库的良好设计是非常必要的，本文把人物信息分为三个部分进行存储，分别是个人信息数据、人物好友关系数据、人物微博数据三个部分，这就对 API 返回的 XML 或 JSON 格式的数据进行简要的处理，有助于在数据库中的良好存储。

2.3 人物实体建模方法

本文研究社会化媒体中人物社会网络，人物的信息如何表示就会成为一个非常基础且重要的工作，社会化媒体中人物实体模型化表示的目的是为用户建立相关的描述文件（user profile）。这是后续的实体关系强度计算、社会网络挖掘和分析研究的基础。

本文将用户信息分为显式信息和隐式信息。前者包括静态的用户资料信息和用户关系信息；后者包括用户的内容信息和用户的交互行为信息。对于上述 4 种用户信息，根据它们的数据特点，采用不同的表示策略，最后将它们合成为完整的用户描述文件。

2.3.1 显式信息表示

显示信息是能够很明显的表现出用户的兴趣爱好的信息，显式信息分为静态的用户资料信息和用户关系信息。

（1）用户资料信息

用户资料信息包括用户的注册信息或其他简短的描述用户的信息。资料信息是最直接反映用户兴趣的信息，可以包括：昵称、性别、年龄、职业、学校、地理位置、兴趣爱好等，本文使用字符串集合的方式来表示人物资料信息，把资料信息的每个属性都表示成字符串的形式。

（2）用户关系信息

关系信息就是在给定的社会化媒体中用户之间存在的确定的关系，如好友关系，人人网中的“好友”关系，微博中的“关注”和“粉丝”关系。这里的关系可以是单向的也可以双向的，人人网中的“好友”关系就是双向的，而在微博媒体中“关注”和“粉丝”关系是单向的。

图 2-6 和图 2-7 给出了用户 F 的好友信息，节点表示用户，边是用户之间的好友关系，图 2-6 是一个双向的好友关系，用无向图表示，图 2-7 是有个单向的好友关系，用有向图表示，假设右图是一个微博的好友网络，则边的方向从关注者指向被关注者，此图的含义是用户 A、E 关注了用户 F、用户 F 关注了用户 B、D、C。

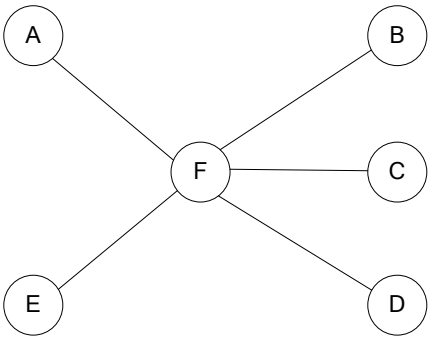


图 2-6 无向关注网络

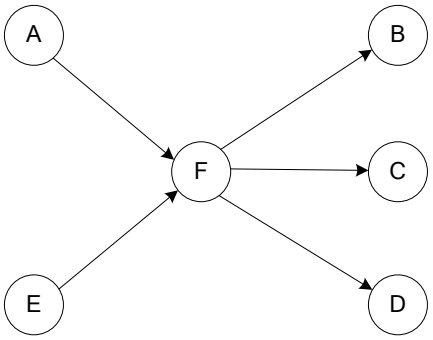


图 2-7 有向关注网络

2.3.2 隐式信息表示

隐式信息指能够间接反映用户兴趣的信息，本文把隐式用户信息分为用户内容信息和用户交互行为信息。

（1）内容信息

内容信息指用户使用社会化媒体所发布或所具有的文本描述信息，能够在一定程度上反应用户在一定时期内的兴趣。如在维基百科中，人物的内容信息就是该人物对应的页面文本内容，而针对微博媒体，内容信息就是用户发布的微博内容。

在社会化媒体中，很多情况下人物实体都可以通过一些文本信息进行表示，传统的文本表示方法很多，最常见的就是使用向量空间模型进行文本表示，本文提出了一种改进的向量空间模型（Improved Vector Space Model，IVSM）来进行文本表示，此模型的主要特点是能给那些对人物描述很重要的词汇更大的权重，以保证对人物描述更加准确。

在微博媒体中，每条微博的长度都是有限的（140 字符左右），不考虑微博中的多媒体信息，只考虑其文本信息，所以首先要将每个用户发布的所有微博拼接成一个长文本，称为用户的微博文档。而在维基百科人物正文描述中，锚文本对人物描述非常重要，维基百科对每个人物都有一个单独的页面，在页面正文中有很多锚文本，文献[42]的统计结果表明，61%的锚文本能够很好的体现链接页面的内容，这些锚文本都是非常重要的描述人物信息的文本，很明显，如果人物 A 和人物 B 包含较多的相同或相似锚文本，则说明他们较为相关，较少则不相关。在其他的社会化媒体中，不局限于锚文本，为了统一表示，以后称这类文本为特殊文本，特殊文本不能使用等同于文本其他普通文本的描述方式，这就是本文的出发点。

董宝力等[43]提出了用于主题网站识别的混合向量空间模型，本文借鉴其混合向量空间模型进行文本表示，对人物页面也分为正文文本和特殊文本，并

提出了改进的向量空间模型的文本表示方法。

在改进的向量空间模型中，人物的特征向量由特殊文本元素 w_i 和普通文本元素 w_j'组成，w_i，w_j'分别是特殊文本特征词 t_i 和普通文本特征词 t_j'的权值。由人物内容信息映射到向量元素，则人物内容信息可以表示为 n 维特征空间的向量。

$$V=(w_1,...,w_i,w_1',...,w_j')=(w_1'',w_2'',...,w_j'') \quad (2\text{-}1)$$

$$w_k''=w_k'+w_s \text{ if } t_s=t_k'$$

其中 i 是特殊文本特征词的数量，j 是普通文本特征词的数量。当 i=0 时退化为传统的向量空间模型。

IVSM 中的特殊文本可以是维基百科中锚文本，也可以是一些典型短语，而在微博中，就没有设定此类特殊文本，所以针对微博数据进行实验时，IVSM 退化为普通的向量空间模型。

下面给出了把人物内容文档表示成其对应的特征向量的过程：

1、文本预处理：对人物文档 d 进行分词、停用词过滤、词性标注等处理。

2、特征提取：选取人物文档 d 中的名词、动词、地名、人名、机构名和一些特殊名词作为特征词，得到普通文本对应的普通特征词和特殊文本对应的特殊特征词。

3、特征词权重计算：对普通特征词和特殊特征词分别采用 TFIDF 方法计算每个特征词的权重，公式为：$w=tf_i(d)\times\log N/n_i$，其中 $tf_i(d)$ 表示特征词 i 在文本 d 中出现的频率，$\log N/n_i$ 为特征词 i 的逆文档频率。

4、特征向量表示：在特征提取和特征词权重计算之后，根据 IVSM 就就可以使用特征向量表示用户的人物文档：$d=(w_1,...,w_i,w_1',...,w_j')$，其中 w_i 为特殊特征词 i 的权重，w_j' 为普通特征词 j 的权重。

以上述的方法就可以把社会化媒体中的人物信息表示为一个向量的形式，并以此向量作为人物内容信息的模型化表示结果。

（2）交互行为信息

在社会化媒体中，用户之间会存在频繁的信息交互动作，这些动作的总和就是一个人的交互行为信息。人人网中的分享行为，微博中的转发评论行为都属于交互行为，这些行为发生的频度可以表明两个人物之间的紧密程度，如果交互越频繁，说明相关性越强，并间接反映两人兴趣的趋同，维基百科中并不存在人物之间的交互行为；在微博中，交互行为包括用户之间的微博转发、评论；现存的很多社会化媒体都存在很多交互行为，这些行为最强烈的反映人物之间的关系紧密性。

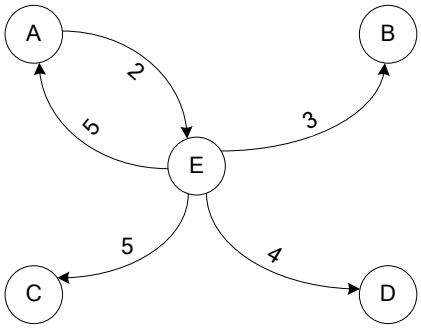


图 2-8 活动性网络

图 2-8 给出了用户 E 和其他人物之间的交互行为网络，称为活动性网络，在图中节点代表用户，边的方向代表行为方向，边上的权值代表用户之间的交互次数。如果此图是微博的活动性网络，则可以解释为用户 A 和用户 E 之间存在相互的交互行为，分别为 2 次和 5 次，而用户 E 对用户 B 的微博存在 3 次交互行为，E 对 C 和 D 分别存在 5 次和 4 次交互行为。

在有些类型的社会化媒体中，并不存在人物之间的交互行为，比如维基百科，其中的人物页面并没有进行交互，所有此信息置空。

综上所述，在人物显示信息和隐式信息都分别以给定的形式表示之后，最终采用 XML 格式对用户的所有信息进行表示。

2.4 实验结果及分析

2.4.1 实验数据

本文采取两个数据来源，分别是维基百科中文版和新浪微博，下面对这两个数据分别进行介绍。

（1）维基百科数据介绍

维基百科数据是维基百科 XML 数据集的中文版，维基百科提供了所有数据的下载，包括各个语言版本。本文采用的维基百科 2011 年 5 月 2 日发布的中文数据集，维基百科会经常对 XML 数据集进行更新，方便用户下载和开放使用。本文选取 376 人，是维基百科人名列表中的“运动员列表”“作家列表”“政治人物列表”中所有人物，称为 3 个大类数据，其中“运动员列表”又分为 3 个小类别，分别是“羽毛球运动员”“篮球运动员”“游泳运动员”，称为 3 个小类数据。

（2）新浪微博数据

新浪微博数据是 2011 年 6 月 2 日止的新浪微博数据，以徐志明、马少平，计算所王斌，刘挺，白硕等人物的微博作为种子节点，选取以这些相关的人物，

得到一个不重复的人物集合，把此集合作为实验人物集合 S，共获取人物 646 人，同时获取 S 中所有人物的资料信息，包括地理位置信息、个人描述、个人标签；获取 S 中人物的微博信息，共 75,408 条；获取 S 中人物的关注和粉丝列表，得到 S 中人物的关系信息，共 1,246,501 个；获取 S 中人物间的交互信息，包括转发和评论信息共 740,406 次。

2.4.2 数据获取结果分析

（1）维基百科数据获取结果

维基百科中每个人物都对应着一个页面，从 XML 数据文件抽取这些人物对应的人物信息，包括半结构化（Infobox）人物属性信息即资料信息、人物页面锚文本信息和人物页面的正文内容信息即内容信息，其中锚文本是特殊文本，正文内容是普通文本。经过信息获取阶段，共获得 302 人的人物信息，经过观察发现未被抽取出信息的 74 人在维基百科中尚未建立页面，所以对应的信息没有被抽取出来，经过对选定的人物信息进行页面对照，发现每个人物的信息都是完全正确的，即抽取正确率为 100%，为下面的实验准备了良好的数据。

（2）新浪微博数据获取结果

每个新浪微博用户在都对应着一个账号，该账号中保存着此用户的各类信息，通过新浪微博开放平台提供的 API，可以得到这些人物对应的注册信息、标签、关注关系，微博信息，交互信息等各种信息。

经过信息获取阶段，成功抽取 646 人的信息，经过人工对比发现实验所抽取人物信息正确，但是也存在一定问题，就是新浪微博 API 在提取人物粉丝时只能提取 5000 个，所以在给出人物关系的时候最多只能存在 5000 个关系，这种限制对那些分析和关注数都没有超过 5000 的人是没有任何影响的，但是对那些超出的将只能得到部分的结果，还有就是新浪微博文本通过 API 只能获得最新的 200 条，也是部分数据，但是考虑到使用 API 获取的方便性，并且本文所关注的网络是人物最新的网络，所以实验就选用了这种数据获取方式。

2.4.3 人物实体建模结果

（1）维基百科人物实体模型化结果

从维基百科数据中获取的人物信息包括半结构化的人物属性信息、人物内容信息，其中人物内容信息包括人物页面锚文本信息和人物页面的正文内容信息，按照人物建模的信息分类，人物半结构化的人物属性信息属于人物资料信息，而锚文本和人物正文属于人物内容信息。然后对人物实体进行建模，维基

百科中的人物只具有资料信息和内容信息，并不具有关系信息和交互行为信息。图 2-9 给出了维基百科中朱德的人物信息模型化表示结果。

```
<person>
  <title>朱德</title>
  <place>[[四川]][[仪陇]]</place>
  <party>{{中国共产党}}</party>
  <Birthdate>1886</Birthdate>
  <past>
    中央人民政府副主席、中国人民解放军总司令
    中共中央纪律检查委员会书记
    中华人民共和国副主席
    授予中华人民共和国元帅军衔、一级八一勋章、一级独立自由勋章、一级解放勋章
    中央书记处书记、中央政治局委员
    中央政治局委员、中央政治局常委
  </past>
  <anchor text>
    全国人民代表大会常务委员会委员长  刘少奇  叶剑英  中国人民解放军  中华人民共和国元帅
    宪兵  上海  孙中山  桂军  陈炯明  陈独秀  德国  哥廷根大学  周恩来  中国
    第二次反围剿战争|第二次  第三次反围剿战争|第三次  国民革命军  周恩来  第四次反围剿战争
    解放军  十大元帅  北京  中共中央  人民大会堂  中共中央副主席|中共中央第一副主席  中华
  </anchor text>
  <content>
  朱德 0.449376 革命 0.093154 史沫特莱 0.0857989 无产阶级 0.0686391 人民 0.0670858 主席 0.0557849
  王伍福 0.0171598 外出 0.0171598 田守尧 0.0171598 特莱 0.0171598 苏联红军 0.0171598 史沫 0.0171598
  红四军 0.011899 攻打 0.011899 工人运动 0.011899 讣告 0.011899 动员 0.011899 东路 0.011899 缔造者
  </content>
</person>
```

图 2-9 朱德的模型化结果

（2）新浪微博人物实体模型化结果

```
<node id="2049588075">
  <att name="screen_name" value="just___fun_"/>
  <att name="province" value="23"/>
  <att name="city" value="1"/>
  <att name="location" value="黑龙江 哈尔滨"/>
  <att name="tag" value="信息检索 社会计算 计算机应用">
  <Textvector>
    <textvector>小组 0.0299415 思想 0.0299415 生涯 0.0299415 离开 0.0299415 考入 0.029941
  </Textvector>
 <Relation>
      <Follows>
      <edge source="1888740957" target="2049588075" weight="0.0"/>
            <edge source="1869909785" target="2049588075" weight="0.0"/>
      </Follows>
      <Friends>
            <edge source="2049588075" target="1888740957" weight="0.0"/>
      <edge source="2049588075" target="1869909785" weight="0.0"/>
      <edge source="2049588075" target="1648083434" weight="0.0"/>
      </Friends>
 </Relation>
  <Actions>
    <Re_tweet>
      <action actiontype="Re_tweet" source="2049588075" target="1888740957" count="2"/>
      <action actiontype="Re_tweet" source="2049588075" target="1869909785" count="5"/>
    </Re_tweet>
    <comment>
      <action actiontype="Comment" source="2049588075" target="1888740957" count="12"/>
      <action actiontype="Comment" source="2049588075" target="1869909785" count="6"/>
    </comment>
  </Actions>
</node>
```

图 2-10 just___fun_的模型化表示结果、

图 2-10 给出了微博中名为 just___fun_的用户的模型化表示结果，新浪微博实验中的每个用户，可以抽取得到其对应的注册信息、标签、关注关系，微博信息，交互信息，其中注册信息和标签为人物资料信息，微博信息是人物内容信息，关注关系是人物的关系信息，交互信息即是人物的转发和评论信息。

2.5 本章小结

本章的主要内容是介绍实验的数据获取方式和人物实体建模方法，对维基百科和微博分别给出了其人物实体数据的获取流程和对人物信息进行建模的过程，最后本文给出了在这两个社会化媒体上所进行数据抽取和人物实体建模的实验结果，并进行了简要的分析，分析结果表明本文的数据抽取方式能够非常准确的获得人物实体信息，并且能够对人物信息进行准确的建模，使人物信息以一种有助于后续处理的良好结构化形式存在。

第 3 章 人物相关性计算及相关性网络构建

3.1 人物相关性计算

3.1.1 人物相关性概念

相关性是度量人物之间存在关系可能的指标，相关性可以在不同的特征上进行度量，比如位置相关性可以在地理位置上对人物进行关系度量，爱好相关性可以在兴趣爱好上对人物进行度量。人物之间是否具有相关性可以从很多方面体现，如两个人来自同一个城市，是同一个学校同一个年级的学生或来自同一个公司，具有相同的兴趣标识，具有共同的朋友，发布一些相似的言论，经常购买了相同的商品，则这两个人物就具有比较高的相关性。本文使用相关性来度量人物之间的兴趣相关性，即人物之间感兴趣的可能，是一个综合表述人物之间感兴趣可能的指标，也就是说如果两个人是兴趣相似或相关的，则此两人具有较高的相关性。

3.1.2 基于系统相似原理的相关性计算

人物实体信息中的资料信息是直接对人物实体的描述，可从不同的侧面直接表达人物的兴趣，如微博中的注册信息、标签信息、个人描述信息，维基百科 Infobox 中的半结构化信息，资料信息都是由一组属性组成，这些信息用来反映人物背景和兴趣是非常直接和准确的，比如一个用户性别属性值为女，则其应该更对一些女性产品感兴趣；两个来自同一所初中的人可能是很相关，甚至是认识的。

通过上面的分析，计算人物资料信息的相关性的出发点是：如果两个人物的资料信息越类似，则他们之间的相关性越强，本文使用人物资料信息的相似性来度量人物之间的相关性。

设给定用户 A 和 B，把 A 和 B 的资料信息都看作一个系统，使用关毅等提出的系统相似度量（Systematic Similarity Measurement）的方法[44]来计算人物资料信息的整体相似度。将其归结为形式化的计算方法：

给定对象 $O_A(A_1,A_2,\ldots A_m)$，对象 $O_B(B_1,B_2,\ldots,B_n)$，令 A={ $A_1,A_2,\ldots A_m$ }，B={ $B_1,B_2,\ldots B_n$ }，m=|A|，n=|B|；设 $x_i>0$ 表示 $A_i(1\le i\le m)$ 的权重，$y_i>0$ 表示 $B_j(1\le j\le n)$ 的权重，设映射对的个数为 $p(p\le \min\{m,n\})$，分别用 $s_1,s_2,\ldots,s_p\in A\times B$ 表示，不失一般性的表示为 $<A_1,B_1>,<A_2,B_2>,\ldots,<A_p,B_p>$，

且对应的相似度分别为 $\mu_1,\mu_2,...,\mu_p$，则系统 O_A 和 O_B 之间的相似度为公式（3-1）所示：

$$S(O_A,O_B)=\frac{\sum_{i=1}^{p}\mu_i x_i^{\ 2}}{\sqrt{\sum_{i=1}^{m}x_i^{\ 2}}\sqrt{\sum_{i=1}^{p}u_i^{\ 2}x_i^{\ 2}+\sum_{j=p+1}^{n}y_j^{\ 2}}} \tag{3-1}$$

对于同一种媒体中的人物资料信息，人物资料信息都由相同个数的属性组成，人物资料信息都以统一的方式表示，此时可将上述公式进行简化，此时 n=|A|=|B|=p，而且 $x_i=y_i$，公式（3-2）给出了公式(3-1)简化后的结果。

$$S(O_A,O_B)=\frac{\sum_{i=1}^{n}\mu_i x_i^{\ 2}}{\sqrt{\sum_{i=1}^{n}x_i^{\ 2}}\sqrt{\sum_{i=1}^{n}u_i^{\ 2}x_i^{\ 2}}} \tag{3-2}$$

然后对两个人物资料信息对应属性之间进行相似度计算，不同类型社会化媒体人物资料信息所选取的属性不同，而针对同一种社会化媒体中的人物，针对其资料信息不同类型的属性，选择的相似度计算方法是不同的，下面给出按照属性特征进行分类的相似度计算方法。

（1）地理位置信息。人物资料中的地理位置信息通常是简短的地方进行标识，且都比较规格化，比如“黑龙江 哈尔滨 南岗区”这样一个地理位置信息，本文给出计算这种信息的方法是一种层次比较的方法，比如 A 和 B 的省市区均相同，则相似性为 1，若省市同，区不同，则相似性为 2/3，如均不同，相似性为 0，以此类推。

（2）日期信息。一些人物的出生日期或其他的日期信息对计算相似性也是非常重要的，比如都是 80 年代出生的人可能具有更多的共同爱好和兴趣，本文给出了一个计算人物年份相关性的方法，出发点是：如果两个人的日期信息越相近这表明两个人物之间存在的相关性就相对较强。

设 A，B 对应的日期信息（年份）分别是 Y_A，Y_B，公式（3-3）给出了计算日期信息的相关性的方法，Y_A 和 Y_B 越接近，则人物直接的相关性越高，当 $Y_A=Y_B$ 时相关性最高为 1，相关性结果为 0~1 之间的数。

$$\mu=\frac{1}{\log^{|Y_A-Y_B|+1}+1} \tag{3-3}$$

（3）固定描述类短语。人物资料信息中也常存在一些信息是以固定的方式描述的，比如人物的党派信息，这些信息的答案是固定的几个，并且只要是相同的答案其描述方式是相同，本文使用直接比较的方法来计算这类信息的相似

度，如果 A 和 B 的信息相同，则相似度为 1，否则为 0。

（4）短小精悍的字符串组合。一些人物的个人标签，个人经历描述等信息都是一种非常简短的文字进行描述的，这些信息比较准确，并且其中包含了大量人物的兴趣信息，对于此类信息，本文选择使用基于编辑距离[45]的方法来计算其相似度。两个字符串的编辑距离的直观含义是需要多少次编辑的操作（包括插入、删除和替换操作）可以把一个字符串转换为另一个字符串，如果需要的次数越少，则表明两个字符串的编辑距离越小，同时字符串之间的相似性就越高。

给定两个字符串 T[1-m]和 S[1-n]，T 和 S 的编辑距离为 d，本文使用公式（3-4）来计算两个字符串的相似度。

$$\mu = sim(T,S) = 1 - \frac{d}{\max(m,n)} \qquad (3\text{-}4)$$

上面给出了四种人物资料信息不同类型属性的计算方法，对维基百科数据选择四个属性作为人物的资料信息：党派和信仰、出生地、出生年和经历信息，根据各个属性不同的特点，使用不同的相似度计算方法，对应党派使用（3）中直接比较的方法，出生地使用（1）中层次比较的方法，出生年使用（2）中的方法， 经历信息使用（4）中的基于编辑的方法。

对新浪微博中的用户资料信息，选用 3 个信息来作为人物资料信息，分别是：地理位置，个人描述和标签，同样使用不同的方法来计算不同属性的相似性，地理位置使用（1）中层次比较的方法，个人描述和标签都使用（4）中基于编辑距离的相似度计算方法。

计算出来每个属性对之间的相似度之后，对每个属性根据其重要程度赋予不同的权值，根据公式（3-2）即可得到任意两个人物资料信息之间的相似性，并以此相似性来反映人物之间的关系紧密程度。

3.1.3 基于人物关系相似性的相关性计算

人物关系信息的相关性可以反映在人物之间共同关系的多少上，如果两个人关系信息的交集部分越多，也就意味着他们认识更多相同的人或被更多相同的人认识，则这两个人之间的是熟人或者可能感兴趣的可能越大，他们之间的相关性就越高。

最简单的度量人物之间的相关性的方法就是计算两两人物之间的共同关系的个数，这个方法可以最直接的反映人物之间关系的相似性，然而这个方法的缺点是对那些关系比较多的节点比较有利，为了消除这个影响，本文给出如下的方法来计算人物之间的关系信息相关性。

为了便于说明，本文就针对微博上的人物关系信息进行阐述，其他类型的社会化媒体可以参照本文给出的方法。在微博中，人物之间的关系是关注关系和粉丝关系，人物之间的共同关系个数就是人物之间的共同关注和共同粉丝个数，给定两个微博用户 A，B，使用公式（3-5）和公式（3-6）分别来计算 A 和 B 之间的关注信息相关性和粉丝关系相关性。

$$\mathrm{Re}l\_Friend(A,B)=\frac{Friend\_common\_num(A,B)}{\log Friend\_numA*\log Friend\_numB} \quad (3\text{-}5)$$

$$\mathrm{Re}l\_Follow(A,B)=\frac{Follow\_common\_num(A,B)}{\log Follow\_numA*\log Follow\_numB} \quad (3\text{-}6)$$

其中 Friend_common_num(A，B)---------- A 和 B 之间的共同关注人数

Friend_numA，Friend_numB---------- A，B 的分别的关注人数

Follow_common_num(A，B)---------- A 和 B 之间的共同粉丝人数

Follow_numA, Follow_numB---------- A，B 的分别的粉丝人数

为了使公式（3-5）和（3-6）计算得到的相似性在 0-1 之间，使用公式（3-7）对结果进行归一化。

$$Sim(A,B)'=\frac{Sim(A,B)-MIN}{MAX-MIN} \quad (3\text{-}7)$$

其中，MAX 是所有人物之间关系信息相似性的最大值，MIN 是最小值。

综上所述，可将实体 A，B 的关系信息相关性定义为：

$$\mathrm{Re}l(A,B)=w_1*\mathrm{Re}l\_Friend(A,B)+w_2*\mathrm{Re}l\_Follow(A,B) \quad (3\text{-}8)$$

其中 w1 和 w2 是分别赋予给关注信息和粉丝信息的权值，w1 和 w2 的和为 1。

3.1.4 基于文本相似度的相关性计算

对于社会媒体中的人物内容信息，根据上文的模型化表示方法，把内容信息使用改进的向量空间模型表示成为向量的形式。如果两个人物的内容信息相似度越高，则人物之间的相关性就会越强，人物的兴趣越相似，所以此处使用内容相似度来度量人物之间的相关度。

传统的计算内容相似度的方法是余弦相似度，这个方法至今仍然不失为一个很好的计算文本相似度的方法，本文仍然采用余弦相似度来衡量人物内容文本向量的相似度。

给定人物 A 和 B，$p=(w_{p1},w_{p2},...,w_{pn})$ 和 $q=(w_{q1},w_{q2},...,w_{qn})$ 分别为其特征向量，w 为最终的人物词条向量的权重。根据公式(3-9)可以计算任意两个人物内容文本之间的相似性。

$$\mathrm{Re}l(A,B)=\cos(p,q)=\frac{\sum_{i=1}^{n} w_{pi} * w_{qi}}{\sqrt{\sum_{i=1}^{n} {w_{pi}}^2}\sqrt{\sum_{i=1}^{n} {w_{qi}}^2}} \quad (3\text{-}9)$$

3.1.5 基于交互频度的相关性计算

在社会化媒体中，交互行为的主体是一些人物，这些人物是信息的发布者、信息转发者，同时也是信息的接受者，所以在这样的一个信息的交流过程中，人物之间的交互行为频度就能够反应出人物之间的关系强度，比如，在微博中人物之间可以使用的交互行为有转发、评论、私信等，这些行为发生的频率越高，说明人物之间的关系强度越强。本文研究时只考虑了转发和评论两种交互行为，具体的方法如下：

设 A 和 B 是两个微博用户，A 对 B 微博的转发次数是 Z（A，B）次，对 B 的评论次数是 P（A，B）次，本文用转发、评论次数的多少来衡量人物之间转发和评论相关性，记做：$\mathrm{Re}l\_retweet(A,B)$ 和 $\mathrm{Re}l\_comment(A,B)$ 即 $\mathrm{Re}l\_retweet(A,B)=Z(A,B)$，$\mathrm{Re}l\_comment(A,B)=P(A,B)$，转发和评论给人物带来的相关度是不同的，在计算人物所有交互信息的相关性是本文对评论相关性和交互相关性赋予不同的权值以象征其对交互相关性的贡献程度，据此给出了（3-10）的计算人物交互信息相关度的公式。

$$\mathrm{Re}l(A,B)=w_1 * \mathrm{Re}l\_Z(A,B)+w_2 * \mathrm{Re}l\_P(A,B) \quad (3\text{-}10)$$

其中 w1 和 w2 是分别赋予给转发相关性和评论相关性的权值，w1 和 w2 的和为 1。实验时可以通过调节 w1 和 w2 的值对两者不同的权重，以动态调节结果。

3.1.6 基于线性融合的人物实体相关性计算

最后使用公式（3-11）中线性融合的方式，计算实体之间的整体相关性。

$$\mathrm{Re}l(A,B)=\alpha_1 \mathrm{Re}l_1(A,B)+\alpha_2 \mathrm{Re}l_2(A,B)+...+\alpha_n \mathrm{Re}l_n(A,B) \quad (3\text{-}11)$$

其中 α 为比例调节因子，控制组成人物实体整体相关性的几种相关性的权重，即控制不同的人物信息对人物实体相关性网络构建所起的作用，权值可以根据需要进行调整。

对于微博人物实体相关性计算，本文只选用了人物资料信息和人物内容信息，而对于微博人物实体相关性计算，本文选用微博人物实体的资料信息、关系信息和内容信息三者的相关性进行加权计算最终的人物实体相关性，并把人物交互频度相关性作为评价其他相关性的标准答案，原因是：微博人物之间的

交互行为多发生在已经具有关系的人物之间，这对尚未具有关系的人物相关性并不能起到良好的关系分析左右，所以本文使用人物交互信息相关性作为衡量两个已经存在关系的人物之间的关系强度。

本文只是给定了一种普遍的加权方式，具体的选用哪几种相关性进行加权要根据不同媒体人物实体的不同属性决定，这种调整非常灵活。

3.2 相关性网络构建

在得到人物相关性后，任意两个人物之间都可能存在一定的相关性，以人物实体为节点，以人物之间的关系强度为边上的权值建立一个带权图，并以此图来代表人物之间的相关性网络。在这个网络中，一些权值很小的边并不具有衡量其人物相关性的参考价值，所以给定一个阈值 R 来对该网络中不具有实际意义的边进行过滤，即把小于该阈的边上的权值设为 0，删除该边。这样就可以得到一个相对意义良好的人物相关性网络。在此相关性网络中，如果两个人之间有边，则表示两个人之间具有相关性，边上的权值则表示相关性的强度。

本文所研究的关系网络是一个相关性网络，此关系网络可以独立于人物之间的实际关联关系，用人物相关性来度量人物之间的关系，至此，一个社会化媒体中的人物相关性网络已经挖掘出来，这个网络是后续网络分析的基础和数据来源，通过网络分析可以有助于对该网络进行深层次的认识和理解并给一些未来的关于社会化媒体上的技术提供原理支持。

3.3 实验结果与分析

3.3.1 维基百科实验

（1）实验策略

对于维基百科人物数据集，经过人物信息抽取和人物建模之后，根据维基百科数据的特点和本文提出的关系强度计算方法的特点，实验采用两种关系强度计算方法对人物之间的相关度进行了计算，具体情况是：针对维基百科人物资料信息，采用基于系统相似原理的相关性计算方法；针对人物内容信息，采用基于余弦相似度的相关性计算方法。

（2）维基百科实验介绍

最终采用线性融合的人物整体相似度的计算方法对人物整体之间的关系强度计算方法，得到最终的人物实体之间关系强度。在人物资料信息计算时，经过大量的实验统计，当取表 3-1 所示的权重赋值时会得到最佳的相似度计算结果。而在最终的线性融合阶段，使用表 3-2 中的权重设置方法已得到最佳的人

物相关度计算结果。

表 3-1 维基百科资料信息权重

属性名称	党派信息	出生年	出生地	个人经历
权重	0.2	0.1	0.1	0.6

表 3-2 维基百科线性融合权重

方法	基于系统相似原理的方法	基于余弦相似度的相关性计算方法
权重	0.2	0.8

（3）评价指标及评价策略

使用 P@n 来评价实验结果的准确率，选择使用人工评价的方法来评价实验结果好坏，评价策略如下：

首先，对相关性进行分级：1 级为无关，2 级为有点相关，3 级为相关，4 级为很相关，5 级为强相关；

其次，从实验人物集合中随机选定 N 个人评价，给出每个人实验结果中最相关的 n 个人，让评价者进行打分；

第三，针对当前被评价人物 p_i，选定的最相关的 n 个人与 p_i 的相关性进行评价，如果评为 5 级，则为精确率加 score(i)=1 分，若评 4 级，则为精确率加 0.75，以此类推，3 级、2 级和 1 级分别加 0.5、0.25、0 分，则 p_i 的相关性准确率公式（3-12）表示。

$$P(p_i)=\frac{1}{n}\sum_{i=1}^{n} score(i) \qquad (3\text{-}12)$$

最后，N 个待评价人物的准确率使用公式（3-13）来计算。

$$P@n=\frac{1}{N}\sum_{i=1}^{N} P(p_i) \qquad (3\text{-}13)$$

（4）实验结果分析

实验中在进行人工评价时请了 10 个对金庸小说比较熟悉的人进行打分，随机选定了 N=100 个待评价人物，最后使用所有人打分的均值作为最终的 P@n。表 3-3 给出了 n 取 1 到 5 之间值的结果。

表 3-3 M，M1，M2 的 P@n

Top_n	M_P@n	M1_ P@n	M2_P@n
n=1	0. 97057	0.97999	0.90092
n=2	0. 89069	0.87324	0.89708
n=3	0. 88950	0.86992	0.87249
n=4	0. 86058	0.84328	0.81670
n=5	0.84671	0.82805	0.78779

表 3-3 中 M_P@n 是按照表 3-2 的线性加权赋值得到的结果，为了给出对比

结果，表中同时给出了 M1 和 M2 方法的对比结果，M1_P@n 是只对人物内容信息计算相关性得到的结果，M2_P@n 是只取人物资料信息计算相关性得到的结果。可以看出使用 M 模型要比分别使用 M1 和 M2 得到的准确率要高，M 模型在 n 取 1~5 时都可以得到令人满意的结果，n=5 时仍可得到 84.67%的准确率，这说明维基百科的人物的两类信息的相关性计算结果都可以从一定的侧面反映人物相关性，最终使用组合的方式要比单独使用任意一个方式要好。

3.3.2 微博实验

（1）实验策略

对选择的微博人物集合，经过信息抽取和人物建模之后，同样针对不同的实体信息选用不同的关系强度计算方法来获取人物两两之间的相关性，对于人物资料信息，使用基于系统相似原理的相关性计算方法；对于人物的关系信息，使用基于人物关系相似性的相关度计算方法；而针对人物内容信息，使用基于余弦相似度的相关度计算方法，此处需要说明的是，在微博文本中锚文本对人物表示所起到的作用并不大，所以此处改进的向量空间模型退化为普通的向量空间模型进行文本表示；而针对人物的交互行为信息，使用基于交互频度的相关性计算方法。

（2）评价指标及评价策略

标准答案：关注人物集合和交互信息相关性计算结果。

待评价结果：人物资料信息相关性计算结果，人物关系信息相关性计算结果和人物内容信息相关性计算结果。

评价指标：P@N 和排序准确率。

在经过计算得到给定人物集合上两两人物之间的相关性后，对于每个人物都得到与其他人物的相关性排序列表，借助于传统信息检索中的评价方法，选择 P@n 来评价其准确率，并把人物的关注人物集合作为每个人的相关人物。用人物集合中所有人物的平均准确率来表示最终的准确率。

用户 A 的 P@n(A)可以通过公式(3-14)计算得到。

$$P@n(A)=\frac{Com\_num(R,F)}{n} \tag{3-14}$$

其中：R 为与 A 最相关的前 n 个组成的人物集合，F 是 A 关注的人物集合，Num(F)为集合 F 的大小，Com_num(R,F)为 R 和 F 中的公共人物个数。

P@n 跟顺序没有关系的，并且把人物的关注集合作为标准答案，为了对本文方法做更全面的评价，在实验结果上再通过排序准确率来进行评价，把人物之间的交互紧密度排序作为标准答案，通过对比人物各种信息的相关性计算结

果排序和标准排序来评价结果的准确性。本文选择使用排序准确率[46]来对排序结果进行评价，公式（3-15）给出了排序准确率的定义。

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{1+|rank_i - ideal\_rank_i|} \quad (3\text{-}15)$$

式中 N----------N 是标准相关人物个数，不同的人 N 值不同

$rank_i$ ----------待评价结果中人物 i 的相关性排序的位置

$ideal\_rank_i$ ------标准答案中人物 i 的相关性排序的位置

（3）微博实验及结果分析

在本文研究的集合中，有些人物节点与集合中其他节点完全没有产生交互行为(评论和转发)；有些节点资料信息(地理位置，个人描述，个人标签)为空；有些节点的在集合 S 中的关注关系非常少，在 S 所组成的网络中处于边缘位置，这些节点都是噪音节点，不进行分析，本文对剩余的 351 个节点进行了评价分析。

交互行为相关性是评测计算结果的标准答案，下面首先给出计算交互行为相关性的权重赋值，这是经过多次的实验给出的最佳赋值，表 3-4 给出的权重分配方式来对转发行为和交互行为的相关性赋值，计算得到人物交互信息相关性结果，并将其作为后面评价其他相关性计算结果的标准答案。

表 3-4 交互行为权重

行为名称	转发行为相关性	评价行为相关性
权重	0.25	0.75

1） 资料信息相关性计算结果分析

对于选定的三种人物资料信息属性（地理位置，个人标签，个人描述）的计算结果分别使用 P@N 和排序准确率进行评价，结果如图 3-1 和 3-2 所示，图 3-1 分别给出了 n 取 3、5、10、15、20、25、30、35、40、45、50 的准确率，从图中可以看出，资料信息的三个属性（地理位置，个人描述，个人标签）中，标签信息相关性计算结果准确率最高，据此下表给出了表 3-5 中的系统相似计算时权重赋值策略，来计算人物资料信息的相似性，把计算结果也放在了图 3-1 和图 3-2 中。

表 3-5 资料信息权重

属性名称	地理位置	人物标签	个人描述
权重	0.2	0.6	0.15

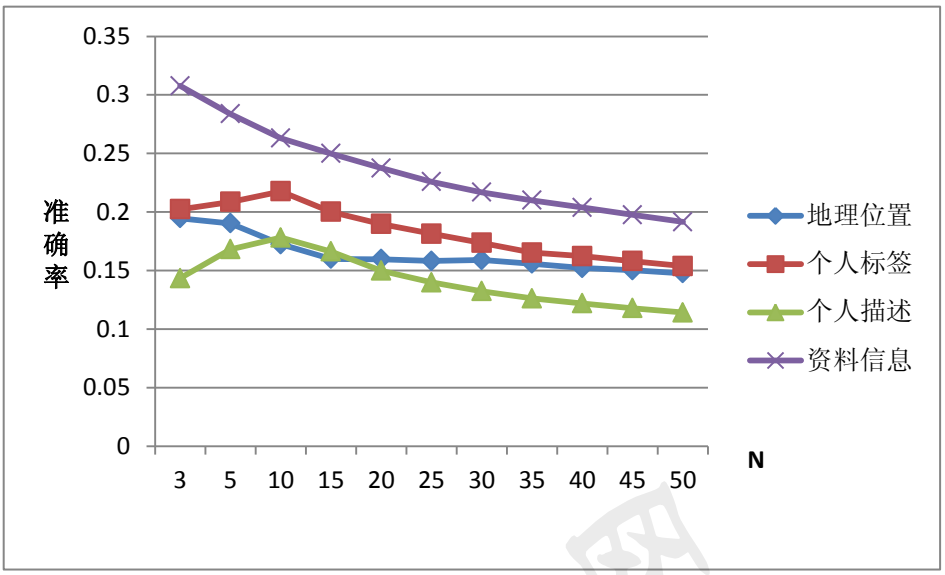


图 3-1 人物资料信息的 P@N 对比结果

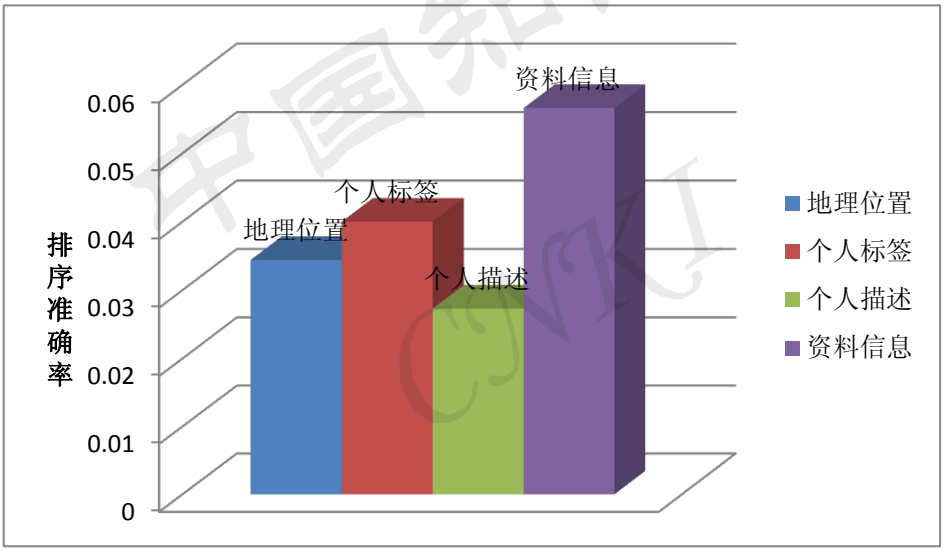


图 3-2 人物资料信息的排序准确率对比结果

从以上两图的对比结果中可以看出使用系统相似原理计算出的人物资料总体相关性结果准确率和排序准确率都是最高的，这说明资料信息的丰富对人物相关性计算结果的准确性有提升，本文工作时鉴于微博中人物的资料信息很多人是不全的，所以选择了三个填写最多并对人物的兴趣最具表达性的属性。

2） 关系信息相关性计算结果分析

关系信息是微博和维基百科很明显的不同之处，微博中的关系数据指用户之间的关注和粉丝关系，而维基百科中却不存在显式的关系信息。对于微博数据，根据上文介绍的关系信息的计算方法，关注信息相关性和粉丝信息相关性都给出了对应的计算方法，而对于关注和粉丝关系结合的形式经过大量实验给出了如表 3-6 的权重赋值方式，最终得到关系信息整体之间的相关性。同样使

用 P@N 和排序准确率对关系信息的结果进行评价，图 3-3 和图 3-4 分别给出了单独使用关注关系、粉丝关系和同时使用二者的结果对比。从图中可以看出单独使用粉丝关系得到的准确率在 n 取不同值都是最好的，当 n=3 是可以达到 70.09%，而使用二者结合得到的关系信息可以达到 65.34%，关系信息的排序准确率则稍好，这说明关注关系在一定程度上可以辅助粉丝关系调整人物之间的关系紧密度。

表 3-6 关系信息权重

关系名称	关注信息	粉丝信息
权重	0.3	0.7

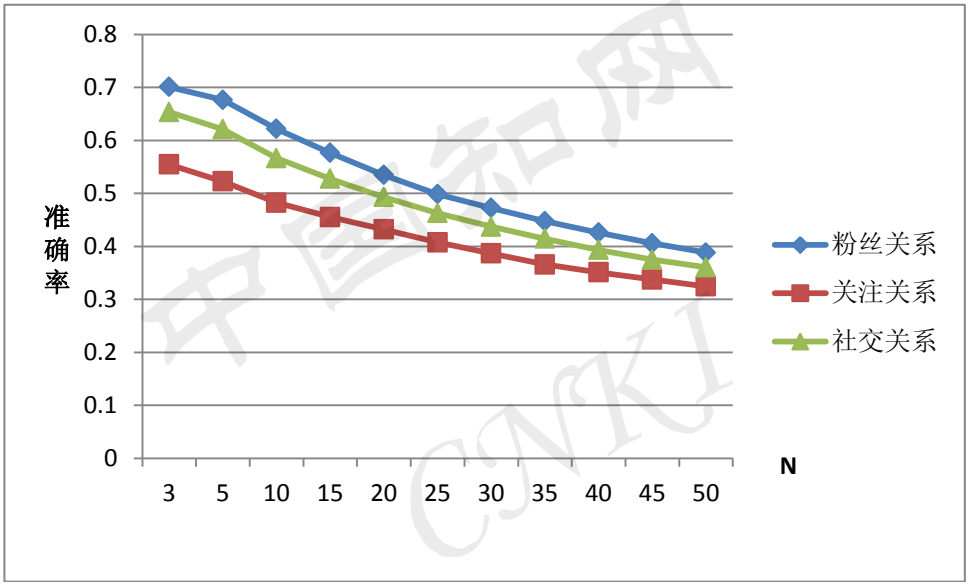


图 3-3 关系信息的 P@N 对比结果

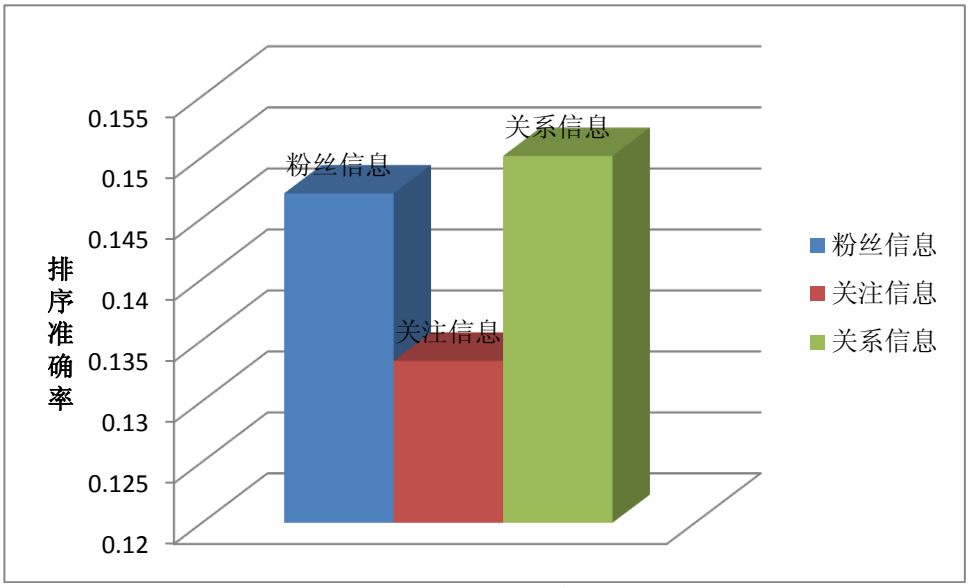


图 3-4 关系信息的排序准确率对比结果

3） 内容信息结果分析

针对人物实体的内容信息，同样使用 P@N 和排序准确率进行评价，结果在图 3-5 中可以看到，人物内容信息的准确率比较低，内容信息是指人物微博文本，微博文本实时性强，短文本，语言不规范的特点应该是造成准确率低的原因。

4） 总体分析

上面分别对人物实体的三种信息（资料信息，关系信息和内容信息）都进行了分析，最终使用基于线性融合的方法来计算人物实体整体的相关性结果，并且每个类型的信息都选取最佳结果进行计算，从图 3-1 和图 3-2 中可以清楚的看到，对于资料信息的相关性计算结果，把资料信息看作一个系统并采用系统相似的原理计算得到的相关性是准确率最高;从图 3-3 和图 3-4 中可以看到，对于人物关系信息则是单独采用粉丝信息得到的人物相关性准确率较高，而取关注关系和粉丝关系的加权结果的排序准确率较好，此处为了得到更好的排序准确率用来推荐，本文选用加权后的关系信息。为了得到最佳的人物实体相关性计算结果，经过大量实验，采用表 3-7 中的权重进行赋值，并把 P@N 和排序准确率的结果表示在图 3-5 和 3-6 中。

表 3-7 线性融合权重

名称	资料信息	关系信息	内容信息
权重	0.2	0.75	0.05

从图中可以看出，使用线性融合的方式得到的人物相关性准确率比单独使用任何一个信息的准确率都要高，而使用关系信息得到的排序准确率确较好，可以得出结论，关系信息可以最有效的反应人物相关性。

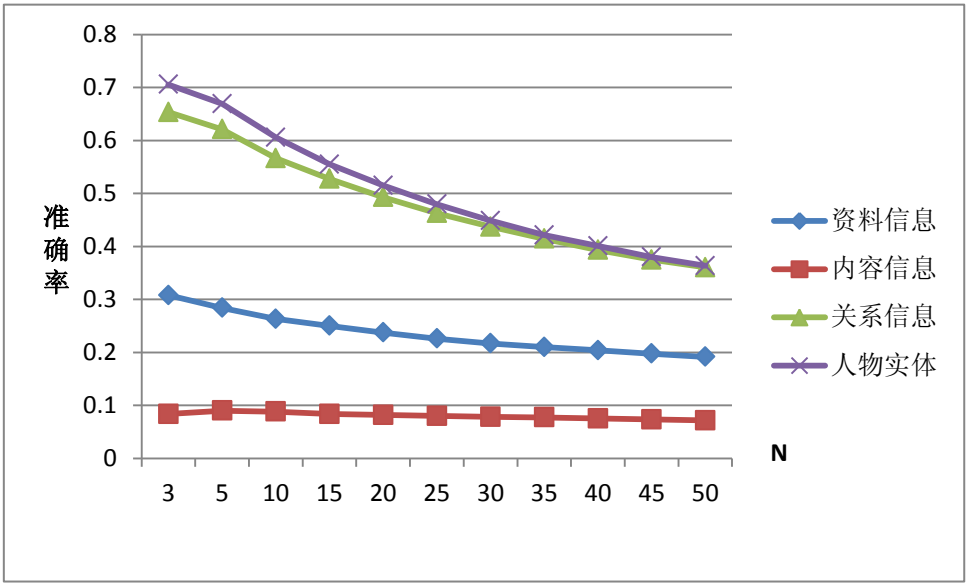


图 3-5 人物实体的 P@N 对比结果

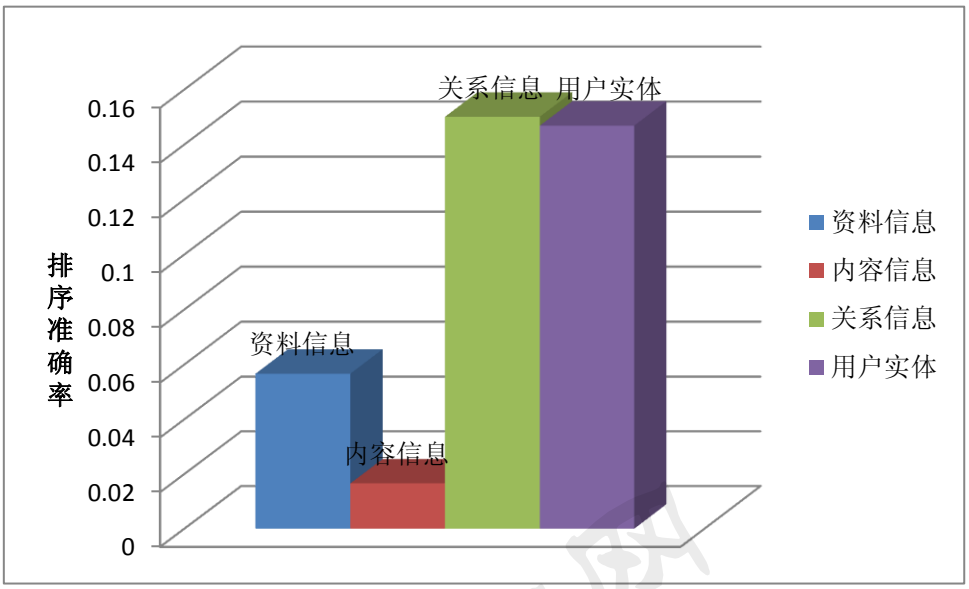


图 3-6 人物实体的排序准确率对比结果

综上所述，在微博媒体上，丰富的人物资料信息对更准确计算人物之间的相关性是有帮助的，而人物的关系信息是计算人物之间的相关性的决定性因素；人物的内容信息对反映人物相关性并没有起到正面的作用，原因是微博内容信息是由一些短文本组成的，这些短文本实时性强并且变动比较大，对反应人物兴趣产生一定的噪音；而人物的资料信息则对人物相关性起到了一定的补充作用，可以使用资料信息的相关性来对关系信息相关性的结果进行部分重排序，已达到最好的推荐作用。

3.3.3 相关性网络构建展示

经过人物相关性计算之后，即可得到人物之间的相关性网络，网络构建时需要选定一个 R 值来对网络进行简化，过略掉一些没有意义的边，R 可以根据需求进行设定，本文设定 R 的标准是得到比较准确的反映网络关系，并且人物之间的边不至于过度稠密，并且有利于网络分析的进行。

针对微博数据，本文使用人物资料信息和内容信息综合计算得到人物之间的相关性网络，图 3-7 给出当 R=0.27 时的维基百科人物相关性网络，经过 R 的过滤，一些孤立的节点在图中并没有显示出来，这些节点与其他人物之间的关系非常的弱，分析原因是这些人的描述信息非常少或者其资料信息填写不完整造成的。

针对新浪微博数据，本文选定加权的人物实体相似度计算结果来构造人物之间的相关性网络，图 3-8 给出了当 R=0.4 时的微博人物相关性网络，同样也过滤掉了一些孤立节点，这些人物与微博中的其他人物之间的关系非常少，并

且资料信息填写不全，发布的微博非常少。

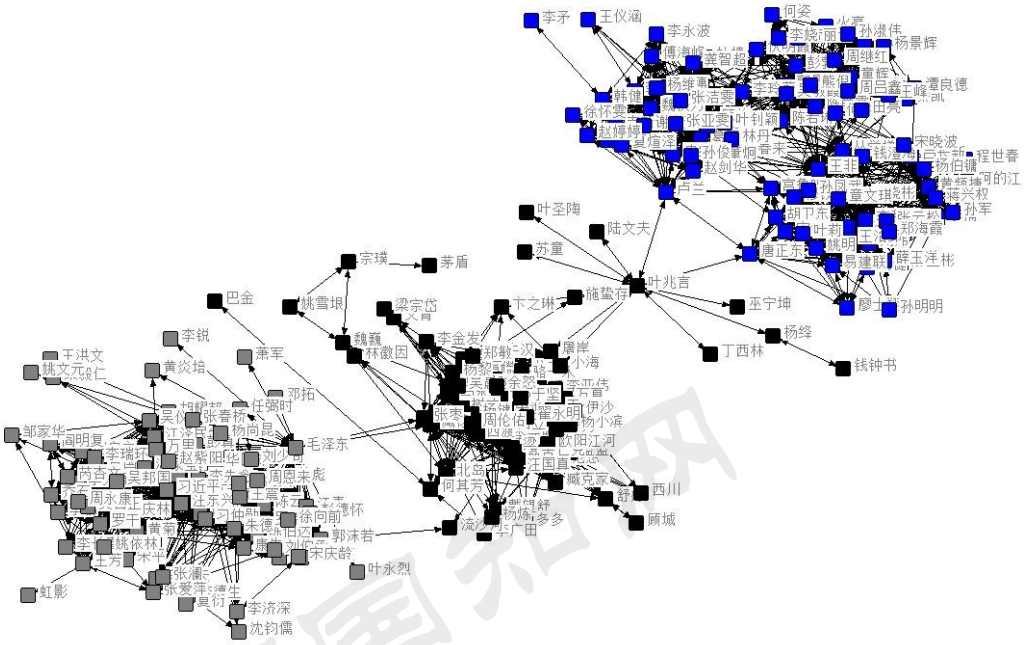


图 3-7 R=0.27 时的维基百科人物网络

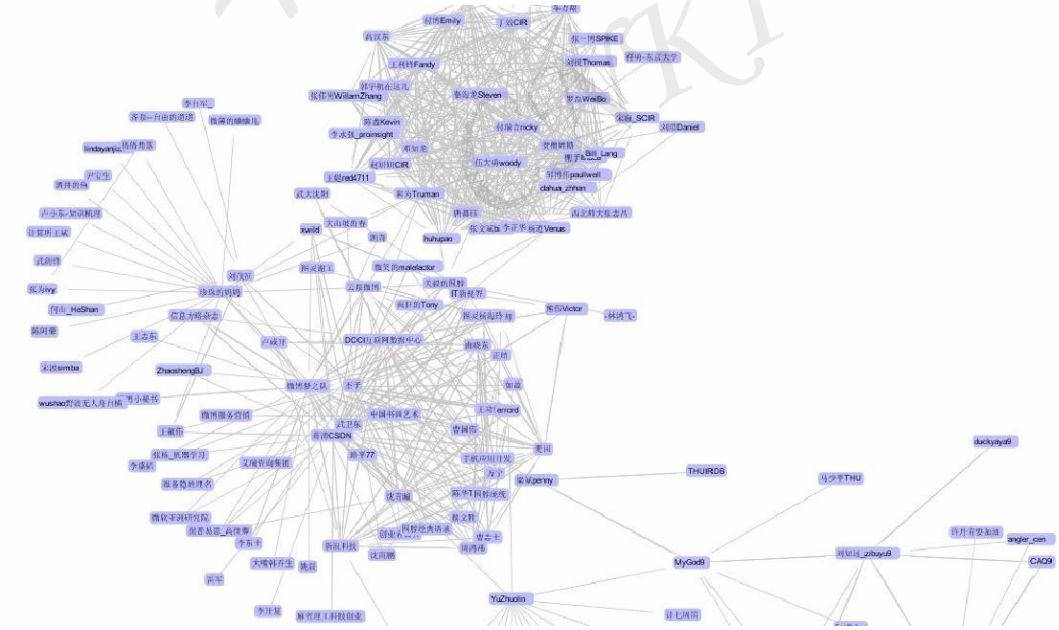


图 3-8 R=0.4 时的微博人物网络

3.4 本章小结

本文首先介绍了人物之间相关性的概念，并根据人物信息的分类提出了五种计算人物实体不同维度属性信息相关性的方法，并给出了详细的使用过程和

原理介绍，然后给出了通过人物相关性构建人物网络的方法，并在本章最后对相关实验进行了介绍，实验分别从维基百科和微博两个数据方面进行了介绍，针对每种数据源，都给出了相应的实验策略，实验方法，实验评价指标和实验结果分析，最后根据人物相关性计算的结果给出了 R 取一定值时的网络结构图示。

第 4 章 社会网络分析及相关应用系统

4.1 基于相关性网络的团体挖掘分析

4.1.1 社会网络分析概念

社会网络分析（Social Network Analysis）是研究社会结构和社会关系的一种方法。实践中，社会网络分析可用于反映人与人之间或部门间重要知识关系，因此特别有助于提高组织中的协作、知识创新和知识传播，研究个人在整个网络中的作用，整体网络的特点，以及如何改进网络组织结构等都有重要价值。

根据上文对社会网络的表示，节点表示人物，边表示人物之间的关系，这样构成了由一组节点和节点之间的边构成的图形。社会网络中的节点是社会学分析社会网络的基本单位，普通意义上的节点可以表示很多类型的，如人物、社区等，而边则根据节点类型的不同而变化，比如若节点是人物，则边可以是相似性，相关性，亲戚、好友关系等，如果节点是社区，边可以是信息流，可以是社会关系等，总之节点之间如果存在边，表明节点之间存在某种关系。

本文开发了一个社会网络的应用系统，这个应用系统提供一个良好的、动态的人物网络展示界面。并在此网络的基础上进行团体挖掘，得到的团体内部人物联系紧密，团体之间人物联系稀疏，同一个团体内部的人物应为其关系而聚合在一起，比如本文研究的人物相关性网络，通过团体挖掘得到一个团体，则此团体内部的人物相关性很强，而与其他的团体的人物联系比较弱。

4.1.2 基于 GN 算法的团体挖掘方法介绍

团体挖掘的方法常分为基于凝聚的方法和基于分裂的方法，基于凝聚的方法的主要思想简单的说就是加边，逐渐向结点个数为 n 变数为 0 的初始网络中加入节点间关系比较紧密的边。基于分裂的团体挖掘方法的主要思想是删点，从网络着手，找出网络中关系最稀松的节点对，并从网络中将这些节点删除，最终得到一组社团。

基于凝聚的团体挖掘方法的不足之处在于不能良好的处理边缘节点，因为边缘节点相比于核心节点之间的关系比较稀松，并且本文主要研究关于所有节点之间的团体关系，边缘节点同样需要被处理，所以本文选用基于分裂的团体挖掘方法，GN 算法是一种经典的基于分裂的团体挖掘方法，GN 算法是 Girvan 和 Newnan[47]提出的一个基于边介数的社团发现算法，使用 GN 算法便可以从相

关性网络中挖掘出比较感兴趣的相关性比较紧密的团体。为了更好的介绍 GN 算法，首先给出边介数的概念。

边介数：网络中经过每条边的最短路径数目，即所有最短路径通过该边的次数之和。边介数可以很有效的度量团体的内部边和连接团体之间的边，因为连接社团之间的边一般具有比社团内部边更大的边介数。

GN 算法是一种分裂方法，该算法思想是根据边介数把不属于任何社团的边逐步删除。通过逐步移去这些边介数较高的边就能够把它们连接的社团分割开来。

GN 算法的基本步骤如下：

(1)计算网络中所有边的边介数；

(2)找到边介数最高的边同时把从网络中将其删除；

(3)对网络中剩余的节点重新计算边介数；

(4)从第 2 步开始重复执行。

事实证明第 3 步的重新计算边介数的过程是 GN 算法非常重要的部分，直到计算出满意的结果。为了更好的衡量得到的社团挖掘的好坏， Newman 等人引进了模块度的概念，可以使用模块度（Q）来衡量挖掘出来的社团个数是否最优，模块度的物理意义是：网络中连接同一社区内部两个节点之间的边的比例，减去在同样的社区结构下任意连接两个节点的边的比例的期望值。模块度的取值是 0 到 1 之间的值，越接近 1，则表示社区划分结果越好。实际上，模块度在一般网络上的值分布在 0.3~0.7 之间，更高的值是比较少出现的。

4.2 团体挖掘结果与分析

4.2.1 维基百科团体挖掘实验结果与分析

1）实验介绍

在维基百科数据挖掘出的相关性网络上，本文使用上述 GN 算法进行团体挖掘，挖掘出的团体具有团体内部关系紧密而团体之间联系稀松的特点，下面进行详细介绍。

首先设定一个 R 值来过滤网络中意义比较小的边，经过多次的实验发现，当阈值 R 取 0.11 时，得到的网络是最佳的，图 4-1 是模块度 Q 和团体个数之间的关系。从上图可以看出当团体数 3 时模块度取最大值 0.643，从模块度的介绍中可知取值接近 0.7，已经是一种非常理想的划分结果。根据模块度的含义，可以得到当从上文所得到的相关度网络中挖掘出 3 个网络是最合理的，图 4-2 给出了当阈值为 0.11 时，模块度 Q 取最大值时的团体挖掘结果图示。

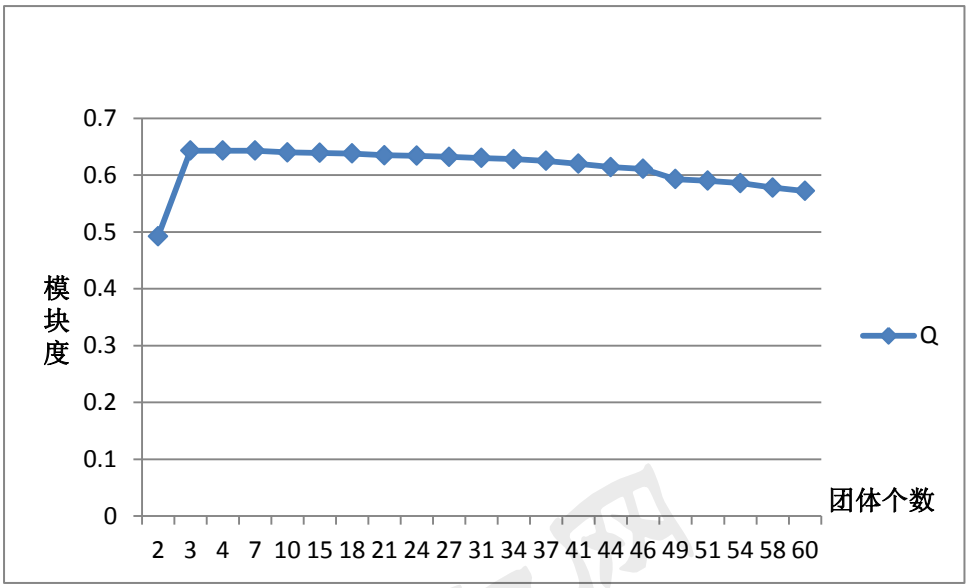


图 4-1 团体数-模块度关系

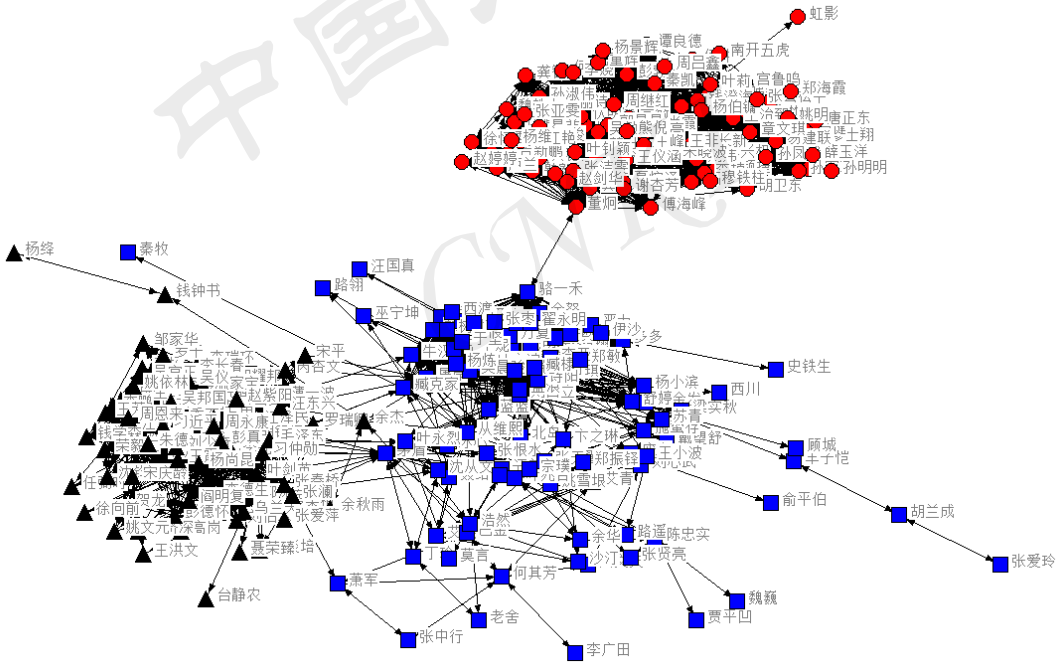


图 4-2 维基百科团体挖掘结果图示

图中不同的节点形状和颜色代表不同的团体，每个节点密集区域表示一个团体，不同团体内部的节点由不同的形状表示；本文在选取数据时，运动员列表的人物又分为了三个类别，分别是“羽毛球运动员”、“篮球运动员”、“游泳运动员”，针对这个子网络，经过大量的实验发现，当阈值为 0.36 时，能得到最佳的团体挖掘结果。图 4-3 给出了这个子网络在阈值为 0.36 时的团体数和模块度 Q 值的关系，可以看出，针对此网络的分析，当团体个数为 3 时模块度取

最大值，最大值为 0.638，也是非常理想的，图 4-4 给出了当阈值取 0.36，团体个数为 3 时的团体挖掘结果图示。

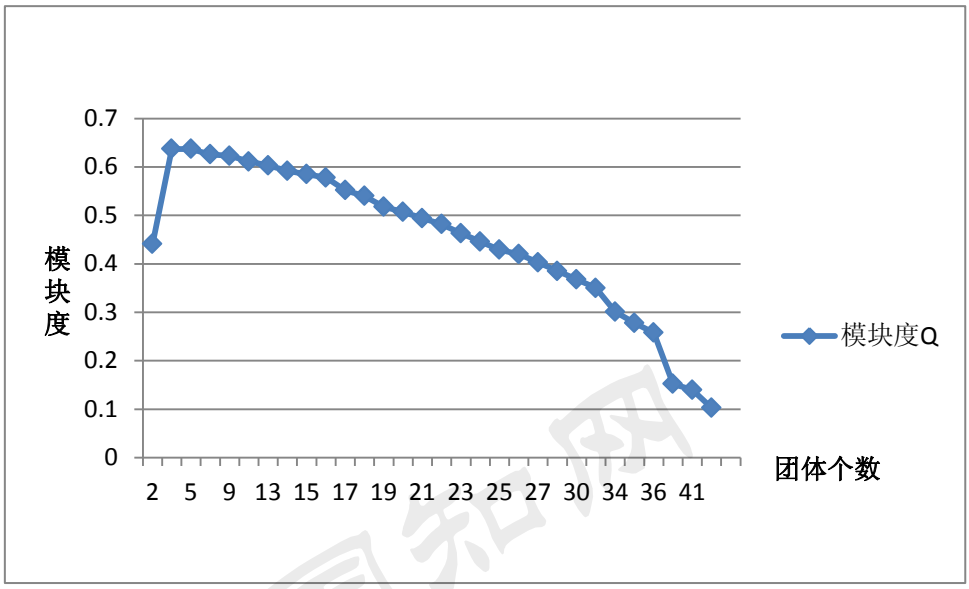


图 4-3 团体数-模块度关系图

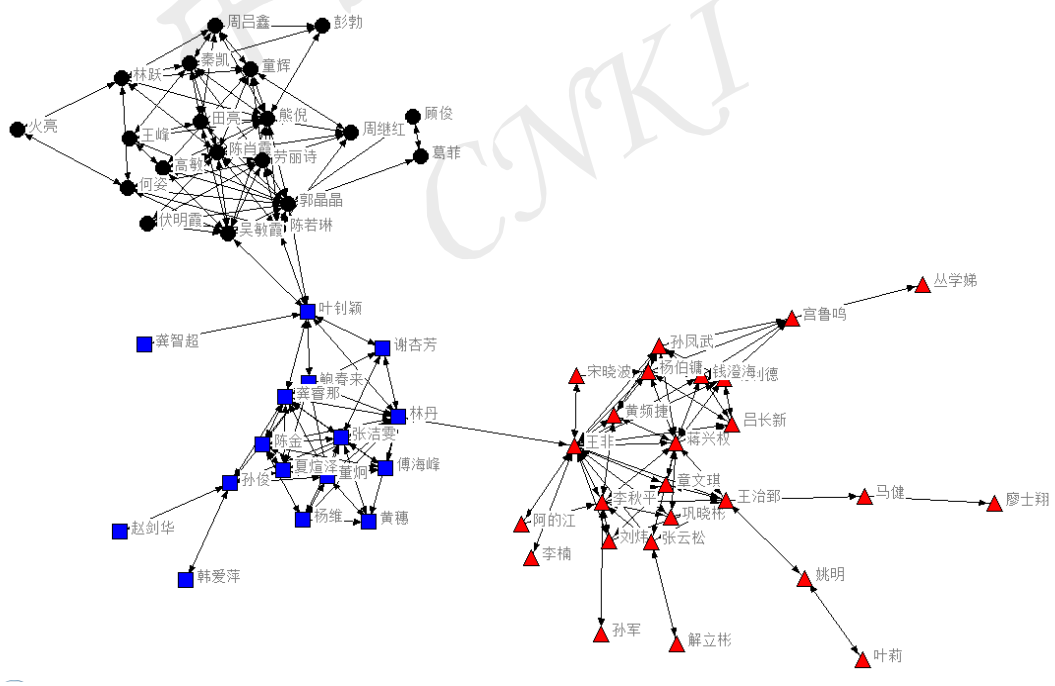


图 4-4 团体挖掘结果

2）结果分析

把团体挖掘的结果和标准的团体分类进行对比，即可得到团体挖掘结果的正确率。针对实验给定的 376 个人物，经过人物信息提取，人物关系强度计算和社区发现之后得到如上图所示的三个团体（三个形状代表不同的团体），把这

三个团体和人物选定时的 3 个人物类比进行比较，即把维基百科分类列表中的分类作为标准的团体结果，使用准确率、召回率和 F 值来衡量方法的好坏，P，R，F1 值分别由公式（4-1），（4-2）和（4-3）定义：

$$P = \frac{Cnum}{Rnum} \tag{4-1}$$

$$\mathrm{R} = \frac{Cnum}{Pnum} \tag{4-2}$$

$$\mathrm{F1} = \frac{2*P*R}{P+R} \tag{4-3}$$

其中 Cnum 是实验中正确计算关系强度的关系节点个数，Wnum 是实验中未正确计算关系强度的关系节点个数，Rnum 是我们计算出关系强度的结点个数，Pnum 是指原始网络中的结点个数。

表 4-1 给出使用选定的评价指标得到的结果，分别给出了 M、M2 的结果，表中 Data1 代表维基百科中的 3 个大类人物，Data2 为维基百科中选定的三个小类人物。M 型是使用了所有人物属性的整体人物相关性网络，并在此网络上进行了团体挖掘的结果，而 M2 是只使用内容信息得到人物相关性网络，并在此网络基础上进行团体挖掘的结果，表中的结果可以定量的评价使用 GN 算法挖掘出来的社团的准确性，同时也可以间接的测量相关性计算方法的好坏。

表 4-1 M1、M2 在 Data1 和 Data2 上的 P、R 和 F1 指标结果

数据集	类别/模型	P	R	F1
Data1	政治人物/M	100%	100%	100%
	政治人物/M2	98.57%	97.18%	97.87%
	运动员/M	100%	96.63%	98.29%
	运动员/M2	100%	100%	100%
	作家/M	93.46%	70.42%	80.32%
	作家/M2	91.15%	72.54%	80.79%
Data2	篮球类/M	100%	68.57%	81.35%
	篮球类/M2	100%	71.43%	83.33%
	羽毛球类/M	100%	55.17%	71.11%
	羽毛球类/M2	100%	55.17%	71.11%
	游泳类/M	100%	80%	88.89%
	游泳类/M2	95%	76%	84.44%

为了综合分析，表 4-2 给出每个数据集在每个模型下的平均指标结果。从表 4-2 可以看出，M 在不同的数据集上的指标除了 Data1 上的召回率之外其他

指标都比 M2 要好，这说明对人物实体的属性信息的丰富可以增加人物相关性构建网络的正确性。同时，M2 模型在不同数据集上的准确率也达到 96%~98%，这同时也说明了，在维基百科数据集上构建人物相关性网络时人物内容信息的重要性。

表 4-2 平均 P，R，F 指标结果

数据集/模型	P 平均	R 平均	F 平均
Data1/M	97.82%	89.02%	93.66%
Data2/M	100%	67.91%	80.89%
Data2/M2	96.57%	89.91%	92.85%
Data2/M2	98.33%	67.53%	80.07%

根据挖掘出的团体的准确性，说明本文通过相关性构造人物关系网络的方法的正确性，从而在推荐任务中可以根据相关性网络进行推荐。

4.2.2 微博团体挖掘实验结果

与维基百科上的相关性网络的处理方法相同，本文使用准确率最高的相关性网络即人物关系信息得到的相关性网络，首先对该网络进行无意义边过滤，针对本实验中微博数据集计算得到的网络，经过大量的实验发现，当 R 取 0.3 时得到的网络是最佳的，并且通过分析知当团体个数为 5 时取得最大的模块度 0.466。图 4-5 给出了当 R 为 0.3，网络划分成 5 个团体时的团体挖掘结果图。

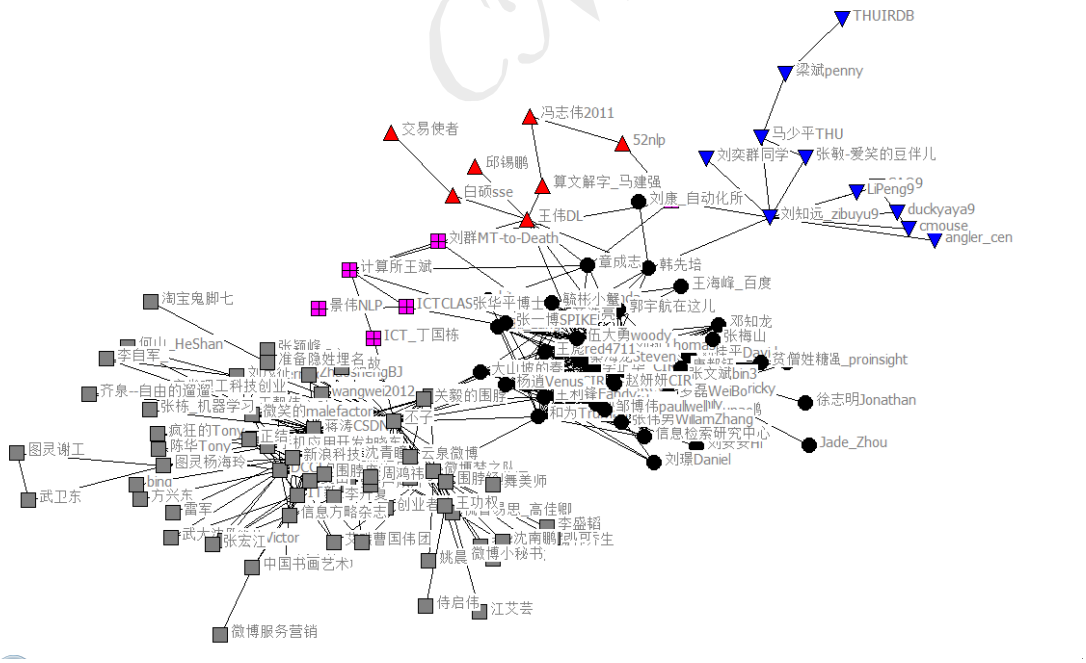


图 4-5 微博团体挖掘结果图

图中每个团体使用不同的形状表示，其中田字格形状的团体表示国内自然

语言处理领域比较著名的几个人，圆形团体主要包含哈尔滨工业大学、信息检索和自然语言处理比较相关的人物的社交圈。

4.3 面向微博的社会网络应用系统

面向微博的社会网络应用系统是本文工作对应微博数据的系统，该系统的设计思想是：使用微博数据，建立一个从数据抽取，节点表示，关系计算，网络构建及网络分析的系统。该系统是本文研究工作的集中体现，本节介绍该系统的基本设计方法和思想。

系统的主要功能是针对微博数据，首先建立一个微博数据爬虫，爬取人物原始信息并存储到数据库中，然后针对人物的原始信息使用上文的方法对人物实体进行建模，进行人物相关性计算，网络构建，对该网络进行网络分析，把分析结果应用到人物推荐的应用中。本系统中数据爬虫阶段是依赖微博的数据获取方法，如果更换数据源，需要改为对应的爬虫，而人物模型化表示、人物相关性计算、网络构建和网络分析等部分适合具体的数据类型和数据来源无关的，适合于各种类型的社会化媒体应用。应用设计时考虑到以后模块的可重用性，设计了专门的数据转换接口，这样根据这些接口可以把不同类型社会化媒体数据都转化为本系统可以接受的数据格式，这就保证了模块的完全独立性。

下面分别详细介绍了本应用的整体框架和可视化模块，其他模块的思想上文都已经介绍，此处不再赘述，同时也给出了基于相关性网络的人物推荐应用程序。

4.3.1 整体框架介绍

图 4-6 是本系统的整体构架图，从图中也可以看出，该系统分为数据层，分析层和应用层三个部分，每个部分都有其不同的功能，数据层部分要完成的工作是数据下载、数据分析、数据标准化等工作，对于不同的社会化媒体，其数据格式、数据来源和数据获取方式是截然不同的，如在维基百科中，其数据来源于维基百科网站提供的数据包，本系统面向微博媒体，需要从其官方网站提供的 API 接口，通过 HTTP 请求获取数据，然后进行分析和标准化。

分析层要完成的工作是根据数据层提供的标准化人物信息对人物实体进行建模，称为节点层；然后根据节点建模结果对人物进行关系计算，关系计算的方法如 3.1 节中介绍，并在关系计算之后通过关系构建网络，此处成为关系层；最后对该社会网络进行分析，主要是指社区发现和人物影响力计算（此部分不是本文的内容，故文中没有给出详细说明）两个部分，此处称为分析层；分析层是一个独立的部分，可以适应任意的社会化媒体数据，针对给定的数据，对

其进行节点建模，关系计算，网络构建和网络分析等工作。

应用层的主要任务是根据前面的计算结果，基于相关性网络的人物推荐，主要功能是针对每个微博用户，能够根据其关系网络，对其推荐其可能感兴趣的人。

在分析层和应用层都用了一个共同的网络可视化模块，该模块的主要功能是以一种动态的可视化效果良好的方式对人物网络进行可视化，以便于使用该系统的人能够快速的看到网络的基本情况，人物的分布情况，人物关系的分布等。

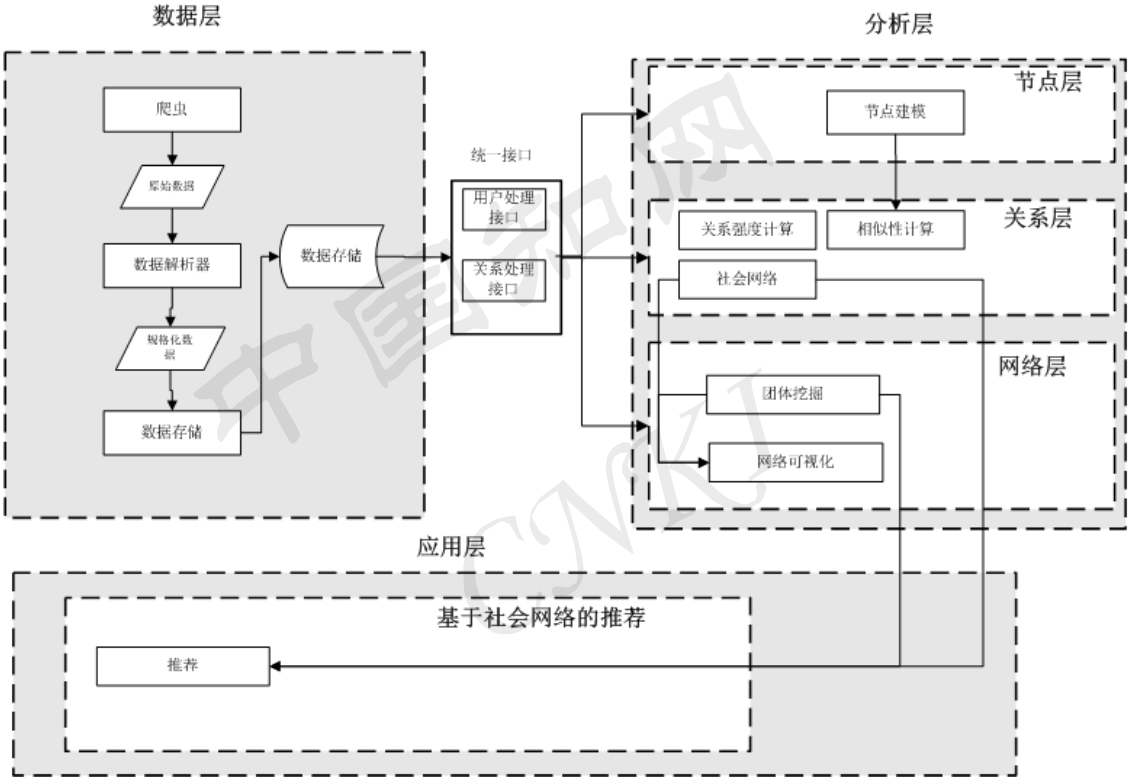


图 4-6 系统架构图

4.3.2 网络可视化

网络可视化模块是一个非常重要的模块，可视化的重要性在于以非常人性化和简洁的图形界面的形式给出人物之间的关系网络，良好的可视化有助于看清楚每个人物在其所在的网络中所处的位置和网络的结构。网络分析的结果也可用可视化的方式表达，比如对于团体挖掘的结果，可以把同一个团体的几点放在一起或显示成相同的颜色形状等，对于节点影响力的计算结果可以用用户节点的大小来表示，节点越大则表示其影响力越大。

比较常用的网络可视化工具是一些社会学研究的成果，这些工具都比较简

单易用，但同时也存在很多问题，首先是都只能针对非常小的网络节点数，比如 Netdraw 在处理 300 人形成的网络时已经非常吃力，并且其显示的界面也不够优美，所以本文自行设计了一个用来对关系网络进行可视化的工具 Graph Show。

Graph Show 使用 java 进行设计的，可以实现节点的自动位置计算，根据力学原理保证节点分布比较清晰并且可以动态调整，只要拖动其中任意一个节点，其他节点的位置都会随之变化，最终达到一个平衡的状态，并且双击任意一个节点，都会展开其在微博中的对应的网络，通过这种方式可以找到想要找到的节点；为了能够更清晰的显示节点，当显示比较多的节点时，可以通过鼠标进行缩放，并且还可以进行搜索已进行快速节点定位，搜索的方式是使用前缀匹配的方法；还可以通过菜单调节画质，比较清晰的画质比较占内存，当节点比较多时建议选择非清晰画质，另外还有一个人物信息展示面板，当双击一个人物时可以显示该人物对应的信息。

使用 Graph Show 对微博中的人物真实关注网络图，图 4-7 是以人物 just___fun_为初始节点，向外扩 3 层得到的关注网络，本文采用的微博数据采样方式是雪球采样，从图中可以清晰的看出数据来源的层次扩展特征。

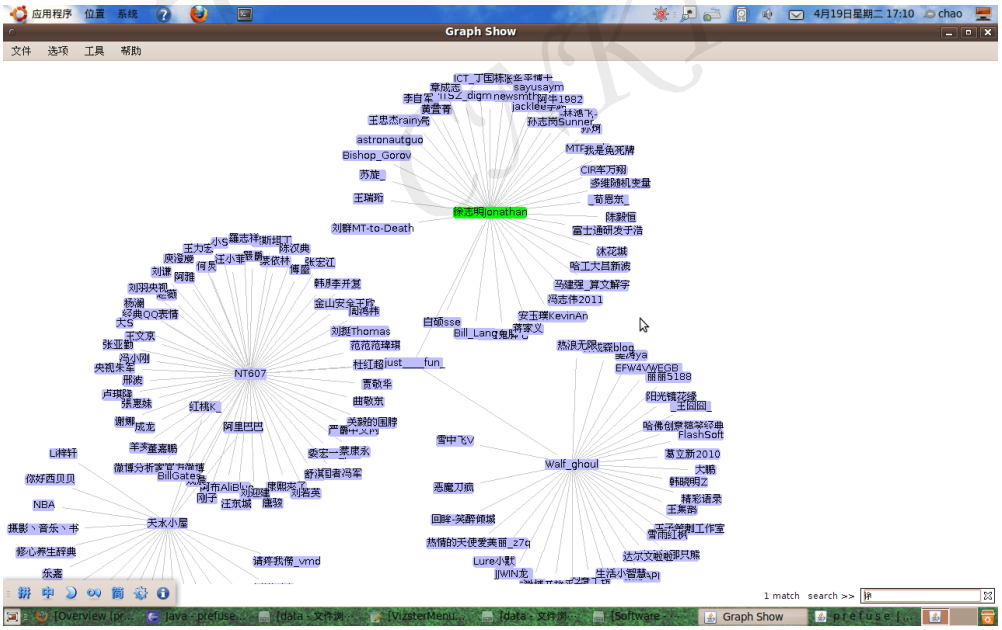


图 4-7 关注网络示例

Graph Show 也可以对人物建模结果进行展示，通过给出人物建模结果的信息面板，人物内容信息特征提取的关键词列表，团体的挖掘结果，人物影响力的计算结果，和信息传播路线图等。

本实验配置的运行该可视化工具的 pc 内存为 2G，可以保证能够同时在一

个屏幕内显示 1000 个节点。

4.3.3 人物推荐应用

应用层的主要任务是应用前两个层次的计算结果，实现一些简单的应用，本文是现实的基于相关性网络的推荐应用，该应用的主要功能是对微博用户，能够根据其相关性网络，推荐那些他可能感兴趣的人物。

根据上文人物相关性计算微博的实验分析结果可知，关系信息对人物相关性的影响很大，本文推荐的过程就简化为，把关系信息相关性很高的人物作为推荐人物。

具体的说，给定人物相关性网络，其中包含 N 个用户，设当前用户为 A，相关性网络中与 A 具有相关性的人物有 n 个，按照这 n 个人物与 A 的相关性从大到小排序得到集合 $R=\{R_1,R_2,...,R_n\}$，设用户在微博网络中已经关注的 k 个人物集合为 $F=\{F_1,F_2,...,F_k\}$， $S=R-F=\{S_1,S_2,...,S_n\}$ 表示 R 中尚未被 A 关注的 n 个人物，应为这些人物已经按照与用户 A 的相关性进行排序了，所以按照 S 中的顺序进行推荐，图 4-8 给出了用户名为“NT607”的用户的推荐页面。

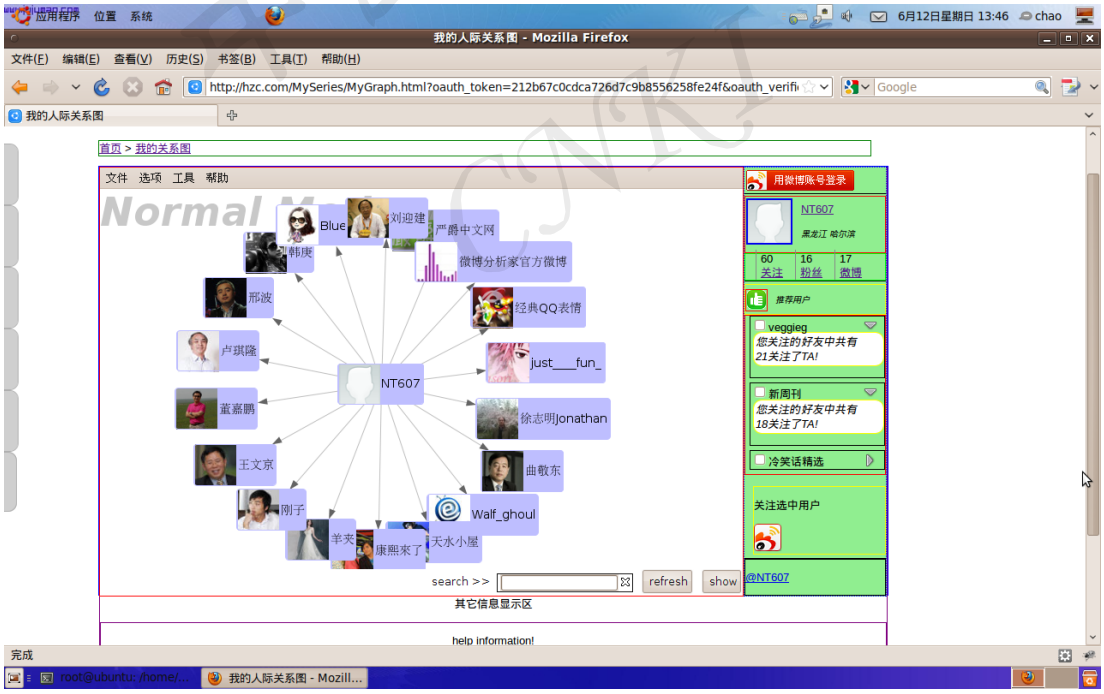


图 4-8 人物推荐应用页面

左侧面板给出了此人的关注网络，显示此人到登录应用为止已经关注的人物，每个节点表示一个人物，用人物头像和昵称表示，箭头表示关注方向，此处可通过设置选择关注网络或粉丝网络或同时显示，右侧上方首先给出了此用户的基本信息包括关注数、粉丝数和发布微博数，下面则给出了推给用户的人

物列表，并显示了一些推荐理由信息，每个被推荐用户前有一个选择框，如果勾选此用户则可以在下面的关注此用户中关注被选中的用户。

人物推荐应用是社会网络在微博数据上的一个应用，新浪微博开放平台提供开发者开发接口，开发者可以把自己开发的应用程序放到其平台上进行展示。

4.4 本章小结

本章首先介绍了社会网络分析的基本概念，然后对选择进行团体挖掘的 GN 算法进行了原理说明，在选定的两种数据分别给出了团体挖掘结果，并对该结果进行了比较分析。最后介绍了本章的另外一个重点，面向微博的社会网络应用系统的设计过程，分别从应用系统的整体框架和网络的可视化输出进行了介绍，同时也给出了一个基于人物相关性网络的人物推荐应用。

结 论

社会化媒体的飞速发展给计算机技术带来了很大挑战和很多机遇，同时也为社会网络提供了更多施展其才能的空间，计算机技术的运用将可以大大的推动社会化媒体的发展，同时对社会化媒体中社会网络的挖掘与分析对更深入的了解社会化媒体的信息传播原理，对社会化媒体的更进一步发展至关重要。

本文研究社会化媒体中的社会网络挖掘和分析技术，以人物之间的相关性作为社会网络的关系，分别以维基百科和新浪微博作为研究对象，抽取人物实体信息，并对人物实体建模，并给出一套完整的面向社会化媒体人物实体的相关性计算方法，在此基础之上构建人物之间的相关性网络，并对网络进行了分析，最后介绍了面向微博的社会网络应用系统，每个阶段都给出了对应的实验介绍和结果分析，实验结果表明本文的社会化媒体的社会网络挖掘与分析方法都能取得良好的效果。

本文的主要工作和创新点如下：

第一，从维基百科和新浪微博中提取出人物实体对应的真实信息，并对人物实体建模。维基百科和新浪微博是两种比较典型的社会化媒体，使用这两种媒体的数据作为研究对象具有代表意义；从新浪微博中抽取数据的方式和传统的 Web1.0 网页的方式不同的，同时本文首次给出了统一的社会化媒体人物实体模型化表示方法，该方法适合于各种社会化媒体数据，以维基百科和新浪微博数据作为实验数据进行了测试，表示本文的方法是可行的。

第二，提出了一组完整的人物相关性计算方法。本文创新性点是根据本文选定的数据特点，针对社会化媒体人物模型化表示的结果，使用不同的相关性计算方法对人物不同维度的信息计算相关性，最终使用线性融合的方法计算人物实体之间整体的相关度，并以此相似度构建人物之间的相关性网络。

第三，网络分析和相关系统设计。本文在构造了人物相关性网络之后，对该网络进行了团体挖掘的分析，该分析使用了比较常用的 GN 算法，最后本文给出了基于微博的社会网络应用系统，该系统创新的使用微博数据构建人物之间的关系网络，并把本文前述的一些思想封装到系统中，并给出了一个良好的可视化界面，可以对人物网络进行动态多方位的展示，同时给出了一个人物推荐应用的设计过程和思想。

虽然本文取得了阶段性的研究成果，但是仍然存在一些需要继续努力的研究点，在以后的研究中，还有以下几个方法值得进一步研究：

（1）在本文人物相关性计算实验中，对应维基百科和新浪微博数据两种数据来源，人物内容信息在计算人物相关性所做的贡献差别很大，分析可知，导致这个问题的原因是新浪微博中每个人物的信息是其所发布的微博所合成的到文本，这个文本由一些很短小的句子组成，并且由于微博的语言特点，造成使用人物内容信息表达一个人物不够准确，下一步的人物就是寻找一个适合于微博这类媒体的内容处理方法。

（2）社会网络分析的工作需要进一步拓展，本文只开展了团体挖掘这一个类型的网络分析，还可以开展网络信息传播模型、节点影响力计算、舆情控制等方面研究。

（3）寻找更合适的评价体系，本文存在人工评价的工作，这种评价方法的说服力都不够强，寻找一个适应于社会化媒体的评价机制是一个非常重要的研究方向。

参考文献

[1] M.E.J.Newman. The Structure of Scientific Collaboration Networks [J]. Proceedings of the National Academy of Science. 2001, 98(2):403-409.

[2] Smith R D.Instant messaging as a scale-free network [EB]. ArXiv: cond-mat/ 0206378. 2002.

[3] Yao Yuanyuan.Internet topology study and its application in IM network modeling[R].Dissertation, China, Zhengzhou University, 2006.

[4] J.Leskovec and E.Horvitz. Worldwide buzz: Planetary-scale views on an instant-messaging network[R].Technical report, Microsoft Research, June 2007.

[5] Nardi.B.A. Why we blog [J].Communications of the ACM, 2004, 47 (12): 41-46.

[6] Kumar, R. et al. Trawling the Web for emerging cyber communities[C]. Computer Networks, Amsterdam, Netherlands, 1999:1481-1493.

[7] Christopher H. Brooks, Nancy Montanez. Improved annotation of the Blogosphere via autotagging and hierarchical clustering[C]. Proceedings of the 15th international conference on World Wide Web (WWW2006). Edinburgh, Scotland, 2006: 624- 632.

[8] Y. Chen, F.S. Tsai and K.L. Chan. Machine learning techniques for business Blog search and mining[J]. Expert Systems and Applications, 2008, 35 (3):581-590.

[9] Flora S. Tsai, Kap Luk Chan. Redundancy and novelty mining in the business blogosphere[J]. Learning Organization, 2010, 17(6): 490- 499.

[10] A. Java, X. Song, T. Finin, and B. Tseng. Why We Twitter: Understanding microblogging usage and communities[C]. In Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, New York, NY, USA, 2007: 56 – 65.

[11] Shaozhi Ye, S. Felix Wu, Measuring message propagation and social influence on Twitter.com[C]. Proceedings of the Second international conference on Social informatics, Laxenburg, Austria, 2010: 216-231.

[12] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers[C]. In Proc. of the third ACM international conference on Web search and data mining, New York, NY, USA, 2010: 261-270.

[13] Haewoon Kwak, Changhyun Lee, Hosung Park, Sue Moon. What is Twitter, a social network or a news media[C]? Proceedings of the 19th international

conference on World wide web, Raleigh, North Carolina, USA, 2010: 591-600.
[14] Fu F, Chen X, Liu L, et al. Social dilemmas in an online social network: The structure and evolution of cooperation [J]. Physics Letters A, 2007, 371: 58-64.
[15] Golder S, Wilkinson D, Huberman B.Rhythms of social interaction: messaging within a massive online network[C].Proceedings of the Third International Conference on Communication and Technologies, 2007:41-46.
[16] Ahn Y Y,Han S,Kwak H,et al.Analysis of topological characteristics of huge online social networking services[C].Proceedings of the 16th international conference on World Wide Web, New York, 2007: 834-844.
[17] Yuta K, Ono N, Fujiwara Y. A gap in the community-size distribution of a large-scale social networking site [EB].arXiv: physics/0701168v2, 2007.
[18] Mislove A, Koppula H S, Gummadi K P, et al.Growth of the Flickr social network[C]. Proceedings of the first workshop on online social networks, New York, 2008:24-30.
[19] Gjoka M, Kurant M, Butts C T, et al.A walk in Facebook: Uniform sampling of users in online social networks [EB].arXiv:0906.0060, 2009.
[20] Chun H,Kwak H,Eom Y H,et al.Comparison of online social relations in terms of volume vs.interaction:A case study of Cyworld[C].Proceedings of the 8th ACM SIGCOMM conference on Internet measurement,New York,2008: 57-70.
[21] Cha M, Mislove A, Gummadi K P.A measurement-driven analysis of information propagation in the Flickr[C].Proceedings of the 18th international conference on World Wide Web, New York, 2009: 721-730.
[22] Lerman K.Social networks and social information filtering on Digg [EB].arXiv: cs/0612046v1, 2007.
[23] Gao Cong, Long Wang, Chin-Yew Lin, Young-In Song, and Yueheng Sun. Finding question-answer pairs from online forums[C]. Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98), New York, NY, 2008: 467–474.
[24] Ames, M., Naaman, M. Why we tag: Motivations for annotation in mobile and online media[C]. In Proceedings of the SIGCHI conference on Human factors in computing systems, 2007:971-980.
[25] Li, X., Guo, L., Zhao, Y.E. Tag-based social interest discovery[C]. In: Proc. 19th Int. World Wide Web Conf. (WWW), Beijing, China, 2008: 674-684.
[26] Martin Szomszor, Harith Alani, Ivan Cantador, Kieron O'Hara, and Nigel Shadbolt. Semantic Modelling of User Interests Based on Cross-Folksonomy Analysis[C]. Proceedings of the 7th International Conference on The Semantic Web, 2008: 632–648.

[27] Ciro Cattuto, Dominik Benz, Andreas Hotho, and Gerd Stumme. Semantic Grounding of Tag Relatedness in Social Bookmarking Systems[C].In International Semantic Web Conference, 2008:615–631.

[28] Shengliang Xu， Shenghua Bao， Ben Fei. Exploring Folksonomy for Personalized Search[C].In Proceedings of 31st annual international ACM SIGIR conference on Research and development in information retrieval SIGIR’08, Singapore, 2008:154-162

[29] Zhijun Yin, Rui Li, Qiaozhu Mei, and Jiawei Han. Exploring Social Tagging Graph for Web Object Classification[C]. KDD’09, Paris, France, June 28–July 1, 2009, pp: 957-965.

[30] Robert Jäschke, Leandro Marinho, Andreas Hotho, Lars Schmidt-Thieme and Gerd Stumme[C]. Tag recommendations in social bookmarking systems. AI Communications. Vol.21, 2008, pp: 231–247.

[31] Moreno. Who Shall Survive [M]? New York: Beacon Press, 1934.

[32] Lewin. Principles of Topological Psychology [M]. New York: McGraw Hill, 1936.

[33] Cartwright, D and Zander. Group Dynamics: research and theory [M]. Londen: Tavist, 1956.

[34] Barnes J.A. and Fank Harary.Graph Theory in Network Analysis [J].Social Networks, 1983, 5(2):234-244.

[35] White, H.C., et al. Social Structure from Multiple Networks, I: Blockmodels of Roles and Position [J]. American Journal of Sociology. 1976, 81(4): 730-738.

[36] Granovetter M S.The Strength of Weak Ties [J].American Journal of Sociology, 1973, 78 (6): 1360-1380.

[37] H.Kautz, B.Selman, and M.Shah.The hidden Web [J].AI magazine, 1997, 18(2):27–35.

[38] P.Mika.Flink: Semantic web technology for extraction and analysis of social network [J].Journal of Web Semantics, 2005, 3(2):211-223.

[39] A. Culotta, R. Bekkerman, A. McCallum. Extracting social networks and contact information from email and the Web[C]. In: Proceedings of the CEAS-1, 2004.

[40] R. Bekkerman, A. McCallum. Disambiguating Web appearances of people in a social network[C]. In: Proceedings of the 14th international conference on WWW, 2005:463-470.

[41] 卢克.基于 Wikipedia 的社会网络挖掘[D].哈尔滨：哈尔滨工业大学硕士学位论文.2009.

[42] Hodgson J. Do HTML Tags Flag Semantic Content [J]? IEEE Internet

Computing, 2001, 5(1):20-25.

[43] 董宝力, 祁国宁, 顾新建. 基于混合向量空间模型的主题网站识别[J]. 清华大学学报(自然科学版) , 2005 45:1794-1801.

[44] Yi Guan, Xiaolong Wang, Qiang Wang. A New Measurement of Systematic Similarity [J]. IEEE Transactions on Systems, Man, and Cybernetics, 2008, 38(4):743-758.

[45] E. S. Ristad, P. N. Yianilos. Learning String-edit Distance [J]. IEEE PAMI. 1998, 20(5):522-532.

[46] 冯子威. 用户兴趣建模的研究[D]. 哈尔滨：哈尔滨工业大学硕士学位论文.2010.

[47] Newman M.E.J, Girvan M. Finding and Evaluating Community Structure in Networks[C]. Physical Review E, 2004, 69 (2): 026113-1-026113-15.

攻读硕士学位期间发表的论文及其它成果

（一）发表的学术论文

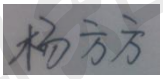
[1] Fangfang Yang, Zhiming Xu, Sheng Li, Zhikai Xu.Social Network Mining Based on Wikipedia.2010 International Conference on Asian Language Processing. Dec,28-30,2010.pp:223-226.(EI 收录号: 20110613644373)

[2] Fangfang Yang, Zhiming Xu, Sheng Li, Xiukun Li.Social Network Mining Based on Improved Vector Space Model. The Second International Conference on Internet Multimedia Computing and Service. Dec, 30-31, 2010. pp:118-121. (EI 收录号: 20111113748657)

哈尔滨工业大学学位论文原创性声明及使用授权说明

学位论文原创性声明

本人郑重声明：此处所提交的学位论文《面向社会化媒体的社会网络挖掘与分析》，是本人在导师指导下，在哈尔滨工业大学攻读学位期间独立进行研究工作所取得的成果。据本人所知，论文中除已注明部分外不包含他人已发表或撰写过的研究成果。对本文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。本声明的法律结果将完全由本人承担。

作者签名：  日期：2011 年 06 月 27 日

致 谢

值此论文即将完成之际和研究生生活即将结束之时，借此机会向那些在学术上、生活上和精神上帮助、鼓励过我的人表示最衷心的感谢。

首先要对我的导师王宇颖老师表示最衷心的感谢，硕士两年内的她对学问孜孜不倦的追求对我产生了非常大的影响，这将是一生的财富，她为人和蔼，知识渊博，她严谨的治学态度和让我受益匪浅，感谢她在我硕士阶段的指导和教诲，我以成为她的学生而感到骄傲。

非常感谢徐志明老师，徐老师渊博的知识、真诚的为人和规格严格功夫到家的思想深深的影响了我，是我十几年的学习生涯中对我影响最大的老师。本文的撰写主要得益于徐老师的亲自指导，对此我深感荣幸并对徐老师的关心和指导表示在再次的感谢。

感谢实验室所有研二的成员，大家在一种和谐的气氛中工作学习，并在我需要帮助的时候无私的帮助我，在我遇到困难时给予我强有力的支持，他们是：许志凯、黄运豪、杨忠宝、徐雷洋。感谢寝室同学给予我生活和工作上的支持，他们分别是王静、付金莹和张红芳。

感谢实验室的其他的成员，师兄师姐在我刚来实验室时给予了莫大的帮助和指导，他们分别是：李栋、王付刚、冯子威、丛帅、袁吕；师弟们在论文和项目进行时给予了很多的支持，对他们表示感谢，他们分别是：王刚、袁树仑、王亮、王文胜，胡智超、夏志海。

感谢在研究生生活中对我有过影响的老师，他们做学问和做人的态度都给我留下了深刻的印象，同时也时刻教诲着我要努力进取，不能落后。他们是李生老师、关毅老师、杨沐昀老师等。

感谢我的父母，他们给予了我无限了宽容、理解和支持，我爱你们。

再次衷心感谢所有曾经关心过、帮助过、鼓励过我的人。