学校编码：10384　　分类号＿＿＿密级＿＿

学号：X2010230626　　UDC ＿＿＿

厦门大学

工 程 硕 士 学 位 论 文

在线社会网络用户角色识别方法研究与实现

Research and Implementation of Online Social Network User Role Identification Method

胡鹏飞

指 导 教 师： 廖 明 宏 教 授

专 业 名 称： 软 件 工 程

论文提交日期： 2012 年 8 月

论文答辩日期： 2012 年 11 月

学位授予日期： 年 月

答辩委员会主席：＿＿＿＿＿＿

评 阅 人：＿＿＿＿＿＿

2012 年 8 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　√　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘要

随着互联网和信息产业的发展，在线社会网络迅速成长，已经对互联网内容的产生以及用户的行为方式的改变产生了巨大的影响。因此，面向在线社会网络的分析对互联网信息监管、舆论导向以及市场经济领域的相关应用具有重大意义。

在线社会网络中的特定用户群体如意见领袖、网络水军等已成为影响网络舆论发展与社会热点事件走向的重要力量。提出一套行之有效的用户角色识别方法，通过用户的行为特征和网络属性对用户角色进行准确定位，针对特定类型用户采用不同的策略，便于对 Web 社会网络的监控和管理，此项研究具有着重要的意义。

本文完整的研究分析了用户角色识别方法的整个流程，并设计实现了相关的系统。

用户信息获取是整个用户识别工作的基础，本文详细研究了利用新浪微博开放式 API 获取用户信息的相关技术和方法。通过数据获取实验，实现了整个数据获取流程的工作，并且验证了数据获取工作的可行性，为接下来的研究分析奠定了坚实的基础。

按照在线社会网络中信息流传播的特点以及不同用户角色在信息传播过程中所产生的影响，使用一定策略将用户划分为不同的角色。对于信息采集中获取的用户身份特征、网络特征、行为特征等信息，采用相关性分析的方法，研究了不同特征与用户角色之间的相关程度，选择了最能支持用户角色识别的一整套特征子集，并分析了不同用户在特征集中的分布表现。

用户角色识别即是将用户按照角色进行分类的问题，使用数据挖掘中的决策树分类算法对该问题进行研究处理。本文总结介绍了决策树分类算法的原理和特点，研究了决策树算法在解决角色识别过程中的实际应用。

针对所获取的用户信息以及角色识别工作的特点，提出了对原有决策树分类算法进行优化的方法。其中包括将具有超大值域的特征所包含的连续值数据进行离散化处理的数据预处理步骤，针对不同特征在测评用户影响力时所表现出来权重的不同，在分类过程中的分裂属性选择时进行加权处理等。

最后，按照软件工程的相关理论和方法，对用户角色识别系统进行设计与实现。

关键词：在线社会网络；用户角色；数据挖掘

Abstract

With rapid growth of the Internet and development of the information industry, online social networks are making a huge impact on the generation of the Internet contents as well as the changes of user behavior. So the analysis about online social network has a great significance for the regulation of the Internet information, the guidance of public opinion, and the application research of the market economy.

The specific online user groups, such as opinion leaders, network water army, etc., has become an important force on the development of network public opinion and the direction of social hot events. Proposing an effective set of user role identification method, by accurately positioning user behavior characteristics and network properties and using different strategies to specific types of users, can facilitate the monitoring and management of the social networks.

An entire process of methods of the user role identification has been totally analyzed and designed in this paper.

The access of user information is the basis of the entire user identification process. This paper completely analyzes the techniques and methods about the access of user information on Sina open API. The realization of the data acquisition process, and the verification of the feasibility of the data acquisition by the data acquisition experiment, has laid a solid foundation for the following analysis.

According to the characteristics of the transmission of information flow about online social networks and the impact of different user roles in the process of information dissemination, the user can be divided into different roles by using certain strategies. By correlation analysis methods, this paper studies the relation of characteristics and user roles, and selects a subset of features which could support user role identification to analyze the distribution performance of different users in this feature set.

The identification of user role is using the method about decision tree classification algorithm in data-mining research to classify the user role. This paper summarizes the principles and characteristics of the decision tree classification

algorithm, gives several examples about this algorithm, compares the pros and cons of these algorithms, all these lays a theoretical foundation for the realization of the final algorithm.

Meanwhile, this paper gives an optimization of the decision tree classification algorithm by the acquisition of user information and the identification of user role, including the discrete processing about the data which characterized by having a large range of data contains and the weighting processing for the classification which according to the different weights of characteristics in the evaluating process of user influence.

Finally, this paper designs a role identification system in accordance with the theories and methods of software engineering.

Keywords: Online Social Network; User Role; Data Mining.

目录

Contents

第一章 绪论

1.1 研究背景及意义

人类是群体动物，社会性是其基本属性之一。人类社会由数以亿计的人和他们之间的社会联系，如亲人、同学、朋友、邻居等所构成。社会网络是描述人类社会的理论之一，社会网络理论于 20 世纪 50-60 年代开始出现。在互联网出现的早期，就有很多尝试工作通过计算机网络建立社会化网络服务，比如 BBS、Email 等。近年来，随着互联网技术的发展，以 Web2.0 类应用为典型案例的在线社会网络得到了迅猛发展。根据中国互联网络信息中心 2012 年 7 月发布的《中国互联网络发展状况调查统计报告》[1]显示，包括即时通信、博客/个人空间、微博、电子邮件、社交网站在内的各类广义的在线社会网络应用位居网民使用率排行榜的前列。特别是新型的在线社会网络应用——微博在“2011 年上半年爆发式增长”，目前用户增长势头逐渐趋稳，“截至 2012 年 6 月底，我国微博用户数达到 2.74 亿，网民使用率为 50.9%”。

社会化网络使人们之间连接的成本大为降低，甚至使人们可以方便的与世界各地的朋友联系，也使人们更容易的扩大自己的社交圈。社会化网络上的朋友关系可以分为两类：强联系和弱联系。强联系是指两者在现实中本身存在关系，如校友、同事、同学等。弱联系是指两者通过社会化网络的服务成为朋友，而并不在现实中存在关系。通常社会化网络服务中这两种朋友关系并存，两者都能强化社会化网络服务的效果。

社会网络具有小世界和无标度的特性。哈佛大学的社会学家米尔格兰姆（S. Milgram）于 1967 年通过一个传递信件的实验对现实社会网络进行了小世界研究，得出“六度分割”推论：世界上任何两人的平均距离不超过 6[2]。社会网络被最广泛传播的定义是米切尔（J. Mitchell）在 1969 年提出的“社会网络是一群特定的个人之间的一组独特的联系”[3]。在当今的社会网络分析的研究中，社会网络已经超越了描述人类之间的社会关系的概念。作为社会网络中实体的行动者，可以是单个个体或者是团体、社会单元，如群体中的人、公司中的部门、公共服务机构、民族及国家。而作为将社会网络中各个活动者连接在一起的联系，其范围

和种类更加宽泛，可以是固有的关系（如朋友关系、亲属关系、性关系、网络联系等），也可以是彼此之间的金钱交易等流动关系。

社会网络是分析和描述一个系统中个体之间流动性事物如何传播的基础模型。现实社会中某种事物的传播规律（信息、病毒等），通常都可以基于社会网络进行刻画和建模，如舆论主题的产生、服装时尚的流行、感冒病毒的传播等。社会网络中最重要的因素是个体之间相互的行为而对彼此产生的影响，这种影响力对社会网络结构的形成以及在此结构下事物进行传播的规律是占主导地位的。因此，社会网络分析的重点也集中在对团体中成员之间社会关系的挖掘和分析上。主要包括社会关系的定义、社会关系的挖掘方法、社会结构层次的发现、社会网络模型等方面。

针对传统社会网络（即真实社会中的社会网络），社会学研究界提出了多种不同理论。例如：针对信息传播的二级传播理论，利用社会网络结构对个性进行分析的结构洞理论。

二级传播理论，即指意见从媒介到舆论领袖到受众，再从受众到媒介的过程。拉扎斯菲尔德（P. Lazarsfeld）在其著作《人民的选择》中提出了二级传播理论的概念[4]。指出意见通常从传统媒介传播给意见领袖，再从意见领袖传播给人群中不太活跃的部分。二级传播理论描述了意见领袖是一类在信息传递和人际互动过程中少数具有影响力、活动力的人。二级传播理论很好的揭示了信息传播规律，描述了信息传播的模式，具有非常重要的意义。

在线社会网络是社会网络的一种，同时具有社会网络的小世界、无标度特性。随着互联网技术的迅速发展以及互联网概念的广泛普及，在线社会网络得到了迅速发展。各种形式应用的涌现为用户提供了一个信息传播与共享、意见表达、思想交汇、情感交流和经济往来的平台，促进了社会行为向 Web 行为、现实社会关系向网络社交关系以及社会信息向 Web 信息的转化进程，使得人类的活动正以前所未有的广度和深度发生着变化。作为现实社会网络在万维网中的映射与扩展，在线社会网络重建了社会连接与纽带，重新划定了社会边界。

在线社会网络中，内容信息以及相应的传播行为对国家安全和社会稳定有着重大的影响，因此面向在线社会网络的分析对互联网信息监管、舆论导向以及市场经济领域的相关应用具有重大意义。通过近年来的实际情况看，在线社会网络

用户特别是意见领袖、网络水军等特定类型网络用户的行为对于互联网中的信息产生与传播具有重要的影响力，已成为影响网络舆论发展与社会热点事件走向的重要力量。提出一套行之有效的用户角色识别方法，通过用户的行为特征和网络属性对用户角色进行准确定位，针对特定类型用户采用不同的策略，便于对 Web 社会网络的监控和管理，具有重要的实际意义。通过研究分析可以深入理解网络拓扑结构以及网络的时态演变对用户角色的形成和变化的作用，便于了解信息在网络中的传播过程，以及用户关系的建立、巩固和减弱、消除，有助于网络资源的优化配置。

1.2　国内外研究现状

目前，面向在线社会网络的相关研究工作主要集中在社会网络挖掘与分析、网络结构性质分析与建模、信息流挖掘与传播规律分析等方面。在面向互联网的社会网络挖掘方面，研究者主要利用数据挖掘的相关技术，针对 Web 的社会化新特征，面向特定社会媒体抽取社会网络。

巴拉巴斯（A-L. Barabasi）和阿尔伯特(R. Albert)在《科学（Science）》杂志上发表的文献[5]分析了现实中电影演员网络的结构特性，推动了对复杂网络变化进行建模的研究工作。对于作为基于复杂网络理论的在线社会网络研究也具有很重要的指导意义。

Web2.0 的迅速发展推动了社会化媒体的迅速发展，以 Flickr 为代表的图片媒体、以 Youtube 为代表的视频媒体等在线社会网络应用服务极大的改变了传统媒体内容的产生和传播模式。在线社会网络中的用户高度参与并且在彼此之间积极互动，用户产生内容（User Generated Content，UGC）改变了传统的由主流媒体发布的内容产生形式，基于网络结构以及用户行为模式的细胞分裂式信息传播模式也改变了传统的广播式信息传播模式。研究人员对新型媒体的产生给予了高度重视与广泛研究，不同学者从不同角度研究了 Flickr 和 Youtube，揭示了蕴含于此类新型网络媒体服务中的在线社会网络的特征和基本原理。Cha 等[6]研究了 Flickr 中受欢迎图片的传播轨迹，通过借鉴流行病学的理论框架阐述了社会级联关系是信息传播过程中的重要因素，并且展示了在线社会网络下信息传播的特征。Lerman 等[7]通过研究 Flickr 中用户的浏览、评论行为以及基于通讯录或朋友列表

的社会关系的分析，发现用户使用其所提供的社会化浏览模式便于发现新的照片，很好地符合了信息觅食理论。Cha 等[8]通过分析 Youtube 中视频的流行周期以及视频年龄与请求特征之间的关系，解释用户对 UGC 内容访问时的偏好特征，10%的视频内容吸引了 80%的用户点击。研究结果对于在线社会网络中的市场营销、广告投放等具有指导意义，同时有助于互联网服务供应商、网站管理员、内容拥有者优化网络服务和营销模式。马延妮[9]利用标签传播算法检测了存在于 Youtube 在线社会网络中的团结构，使用并优化相关算法对其中的团特性进行了分析。郑皓[10]分析了基于自动化方法和社会化方法对 Web2.0 环境下所产生的网络内容建立标注层次结构，用量化指标刻画了社会网络的相关性质。

唐晋韬[3]针对面向互联网的社会网络挖掘和信息传播分析两个方面展开了相关关键问题的研究，其中包括面向社会媒体的社会网络挖掘问题、社会媒体中社会网络结构信息传播的影响等问题。胡海波等[11]应用复杂网络理论通过对整体网络拓扑性质，以及通过度值分布、聚类系数、平均路径长度、度相关性、介数、社团结构等方法对在线社会网络结构进行了分析，总结和印证了在线社会网络除了无标度特性、小世界特性以外的三大共性：高聚类系数、正的度同配指数、强的社团性。同时还指出连通子图规模分布、度分布和社团规模分布呈现出锯齿状以及度异配性等特性。刘志明等[12]以微博中的网络舆情为研究背景，概括了意见领袖的定义，建立了以用户影响力、用户活跃度为量化标准的意见领袖识别指标体系，结合 AHP 得出了不同指标的权重，并且提取了意见领袖决策规则集。文中还对意见领袖的跨主题性进行了一定的研究。文中对意见领袖识别指标体系的建立以及对用户角色的识别对本文的研究具有一定的借鉴意义，但其仅对意见领袖这一具有重要地位的角色进行了分析，忽略了其他角色对网络信息传播与整个网络的拓扑结构的影响。王钰等[13]研究了网络论坛中意见领袖的发现，文中选取了与论坛信息传播拓扑结构以及行为模式相关的特征建立用户影响力模型，采用基于 EM 算法的用户聚类算法进行用户类别的划分，并从中选取网络论坛中的意见领袖。

Twitter 创办于 2006 年，目前已经成为全球最受欢迎的微博服务商。近年来，Twitter 吸引了大量的学者对其进行研究。其中就包括对 Twitter 用户的影响力进行了不同的定义，并且使用不同的模型和方法对用户的影响力进行了相关的分析