杭州电子科技大学

硕 士 学 位 论 文

题　　目：基于社交网络的组织成员识别及其兴趣爱好挖掘方法研究

研 究 生　陈志辉

专　　业　计算机技术

指导教师　万 健 教授

完成日期　2018 年 3 月

杭州电子科技大学硕士学位论文

基于社交网络的组织成员识别及其兴趣爱好挖掘方法研究

研 究 生：陈志辉

指导教师：万健 教授

2018 年 3 月

Dissertation Submitted to Hangzhou Dianzi University

for the Degree of Master

The Identification of Social Network Organization Members and Interests Mining

Candidate: Chen Zhihui

Supervisor: Prof. Wan Jian

March, 2018

杭州电子科技大学
学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。

申请学位论文与资料若有不实之处，本人承担一切相关责任。

论文作者签名：陈志辉　　日期：2018 年 3 月 25 日

学位论文使用授权说明

本人完全了解杭州电子科技大学关于保留和使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属杭州电子科技大学。本人保证毕业离校后，发表论文或使用论文工作成果时署名单位仍然为杭州电子科技大学。学校有权保留送交论文的复印件，允许查阅和借阅论文；学校可以公布论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存论文。（保密论文在解密后遵守此规定）

论文作者签名：陈志辉　　日期：2018 年 3 月 25 日

指导教师签名：[illegible]　　日期：2018 年 3 月 25 日

摘要

随着 Web2.0 时代的发展，社交网络平台已经成为互联网服务中不可或缺的重要组成部分。网民们积极地加入到这种新型的信息交流平台中。社交网络用户数量呈现出急剧增长的态势，同时也导致了社交网络中信息量的爆炸性增长，所以近几年来，越来越多的学者开展了针对社交网络的研究。

用户关系研究中的组织成员识别和用户特征研究中的兴趣爱好挖掘是社交网络领域中很有价值的研究问题。对于组织成员识别的研究还处于起步阶段，现有方法大多数只考查用户与组织之间的粉丝关系，缺乏对用户行为多样性以及不同用户行为对于用户关系判别影响力的研究。对于社交网络用户兴趣爱好挖掘的研究已经有了非常丰富的成果，但现有的研究缺乏对兴趣爱好内在关系的探究，没有将关联关系运用到兴趣爱好挖掘中。本文的主要研究内容如下：

（1）提出了一种基于社交网络用户行为关系的组织成员识别方法。该方法首先根据社交网络 Twitter 的特点及其用户的行为属性，定义了多种识别因子，量化地描述了用户与用户组之间的行为关系，并归纳总结了社交网络用户关系的基本判定规则，将社交网络中用户关系的界定转化为模型计算的过程。然后基于社交网络的真实数据对识别模型进行实证研究，最后详细探讨了多种识别因子在识别过程中的影响程度和最优组合模型.

（2）提出一种基于关联规则的兴趣爱好挖掘方法。基于 LinkedIn 用户的个人档案，首先对兴趣爱好进行建模，将不同形式的兴趣爱好字符串使用统一的兴趣项进行标准化整理，并从中挖掘出高频兴趣项作为研究对象。然后在此基础上建立了兴趣爱好关联分析模型，并基于 LinkedIn 的真实用户数据生成兴趣爱好关联规则。最后使用兴趣爱好关联规则对原始的基于词频的 Twitter 用户兴趣爱好挖掘方法进行了改进。

（3）结合上述两点，作者开发了基于海量社交网络数据的人物属性识别系统，并将组织成员识别与兴趣爱好挖掘方法运用到该系统中。系统能将社交网络用户按照组织为单元进行划分，并根据挖掘出的用户兴趣爱好研究其人物属性。

本文所提出的方法可以在错综复杂的社交网络中识别出属于相同组织的用户，并且准确地挖掘用户的兴趣爱好。研究成果不仅可用于广告投放、好友推荐等商业活动，对社交网络用户行为特征和属性特征的研究也具有借鉴意义。

关键词： 社交网络，用户关系，组织成员，兴趣爱好

Abstract

With the development of Web2.0 era, Social networking platforms have become an integral part of Internet services, Internet users are actively joining this new information exchange platform. The number of social network users is growing rapidly, and it also led to an explosive growth of the amount of user information in social networks. Therefore, in recent years, more and more scholars have carried out research on social networks.

The identification of organization members in the research of user relations and interests mining in the research of user characteristics are a very valuable research issue in the social network field. The research on the identification of organizational members is still in its infancy. Nowadays, most of the identification methods of organizational members only examine fan relationship between users and organizations, and there is a lack of research on the diversity of user behaviors and the influence of different user behaviors on user relationship discrimination. The research on user interests mining in social networking has already had a lot of achievements. However, the existing research lacks the exploration of the inherent relationship of interests and does not apply the association to the mining of interests. The main contents of this paper are as follows:

(1) A method for identifying organization members based on the behavior of social network users is proposed. According to the characteristics of social network Twitter and the behavior attributes of its users, this method defines a variety of identification factors to quantitatively describe the behavioral relationships between users and user groups and summarizes the basic rules for determining social user relationships, so that the definition of human relationships in the social network transforms into the iterative computation of the model. Then the empirical research on the recognition model based on the real data of social network is carried out to discuss the degree of influence and the optimal combination model of multiple recognition factors in the recognition process.

(2) A method of interests mining based on association rules is proposed. First of all, based on the profile of LinkedIn users, interests are modeled, and different types of interest strings are standardized using uniform interest items, and the high frequency interest items are excavated as research objects. Based on this, a correlation analysis

model of interest is established, and association rules of interest are generated based on real user data of LinkedIn. Finally, we use the association rules of interest to improve the original Twitter user interest mining method based on word frequency.

(3) Combining the above two points, the author develops a personal attribute recognition system based on massive social network data and applies the method of mining membership identification and hobby mining to the system. The system divides social network users into organizational units and studies their personal attributes based on the excavated user interests and hobbies.

The method proposed in this paper can identify users belonging to the same organization in a complex social network and accurately excavate the user's interests. The research results can be used not only for commercial activities such as advertisement delivery and friend recommendation, but also for the study of social network users' behavior characteristics and attribute characteristics.

Keywords: Social Network，User Relationships，Organization Members，Hobbies and Interests

目录

第一章 绪论

本章首先介绍课题的研究背景及意义；接着从社交网络用户关系研究和兴趣爱好研究两方面对国内外研究现状进行了介绍；最后，介绍了本文的研究内容并列出了论文的框架结构。

1.1 研究背景及意义

随着社交网络的兴起和快速发展，网络生活变得更加丰富多彩，几乎每个网民都参与到了这种新型的网络组织结构中。社交网络的产生使得用户从传统的网络信息的被动接受者，转变为创造网络信息的主动生产者，社交网络用户不仅可以在虚拟世界中结识新的网络好友，也可以将自己现实生活中的人际关系迁移到其中，在虚拟世界维系现实生活中的人际关系。除此之外，社交网络也是一种基于用户关系的信息共享、信息传播以及信息获取的平台。社交网络已经渗透到我们生活的方方面面，它突破了传统媒体的信息传播方式，正在以前所未有的方式影响和改变着人们的交流方式。用户可以自由地生产自己认为有价值的信息内容，并以短文本、图片或视频的方式与好友共享此信息。与此同时，用户也可以主动选择成为其他用户的好友或者粉丝，随时随地获取自己感兴趣的信息。

组织是指为了实现共同目标而相互合作的集体或群体[1]，例如商业公司、政府部门、教育机构等。从社交网络中可以广泛地挖掘出与组织或团体相关的事件信息，例如事件的发展或分布情况等，从而使得组织或团体能够更及时地制定战略措施，高效地应对突发事件。近几年来，因为有着很高的学术和应用价值，社交网络用户关系挖掘已经成为了新兴的研究热点之一[2]。该研究通常试图将社交网络中用户关系映射为现实世界中的关系网络以便于为社交网络用户提供精准的推荐服务。在同一组织中的成员称为同事，同事关系是用户关系中比较常见的一种类型。对于同事关系的挖掘也称为组织成员识别，现实世界中属于同一组织的成员，无论在工作中还是生活中都存在大量的交集，他们通常有一些相似的特征或者需求，所以组织成员识别结果通常可以较好地应用于商业广告的精准投放和产品或好友的个性化推荐[3, 4]。

兴趣爱好是指个人的心理倾向，希望知道和掌握某些东西，并经常参与这些活动，或是指个人有积极探索某些东西的认知倾向。一般来说，个人的兴趣爱好被视为理解某一特定主题的持久的内在需求，是一种认知和情感诉求，使个人能够保持这种需求[5]。因此，兴趣爱好是承认和参与某些活动的巨大动力，它允许个人自发的或带有积极情绪的对某些事情给予优先的注意[6]。实际上，兴趣爱好

对人格的形成，心理健康，教育和职业发展都有重要的影响。它是心理学和教育学中非常重要的概念。近年来，随着电子商务和社交网络的普及，兴趣爱好推荐系统在实践中得到广泛应用。事实上，根据用户的兴趣和喜好推荐个性化的产品和信息，已经成为产品销售和信息服务的一种非常有效的方法。因此，互联网用户的兴趣爱好建模和挖掘等相关研究已经非常流行。除此之外，兴趣爱好挖掘的研究成果不仅可以应用到推荐系统中，还可以为其他兴趣爱好研究领域提供新的研究思路，如兴趣爱好培养、兴趣爱好建模和兴趣爱好导向教学等。

1.2 国内外研究现状

1.2.1 社交网络用户关系研究

目前，社交网络用户关系研究与应用主要集中在社群发现[7, 8]，也称为社区挖掘。使用或优化图聚类是社群发现问题的主要解决方法之一[9]。吴烨等[10]结合信息论中最小长度原则，基于遗传算法，提出一种高效的属性图聚类方法 GA-ACG，虽然实验表明了该方法在聚类质量和算法的线性复杂度方面都有较好的表现，但其方法仍存在适应度函数不够优化与迭代次数不稳定等问题。Zhang Y 等[11]根据用户的兴趣相似度基于经典的聚类算法对 Twitter 进行社区挖掘，实验表明该方法虽然能有效地进行社区挖掘，但在聚类算法和用户属性特征的选取上还有待提高。Danyllo 等[12]采集了金融机构的用户信用数据，通过信用分析将 Twitter 用户进行聚类，实验表明属于相同社区里的用户有着相似的信用级别。Sotiropoulos 等[13]提出一种基于 LDA(Latent Dirichlet Allocation)模型并采用多目标优化方式的方法，从空间紧密度和话题紧密度两方面对社区挖掘方法进行优化。Ben[14]和 Liu X[15]分别提出一种基于半监督的层次聚类方法进行社区挖掘。但以上这些方法都是将社交网络的用户集合根据用户顶点的结构关联度或属性相似度划分为若干用户集合。

社群发现通常将社交网络的拓扑结构映射为图[16, 17]，然后通过图中顶点拓扑关系或顶点属性的相似计算将其划分为若干个子图，从而将社交网络成员划分为几个社群[18]，但它不会进一步确定用户之间的特殊关系。与社群发现类似，特定组织成员识别问题最终要返回的也是一个虚拟社群，不同的是社群发现重点关注社群内部用户之间的相似度和亲密度，而特定组织成员识别问题关注的重点是返回的社群内部用户与该组织机构的相似度和亲密度。目前社群发现算法可以为特定组织成员识别问题提供一定的借鉴思路，但不能直接应用其中，所以目前针对社交网络中特定组织的组织成员识别方法的研究依然不多。张振华等[19]抓取了特定组织的若干官方账户的两层粉丝的社交网络数据，利用社交网络的拓扑结构计算用户对目标机构的兴趣度，划分了网络中的社交圈子，最后通过定义社区的

R@N 指标来选取相关社区。实验结果表明通过 R@N 指标能够有效的区分出相关社区，但在得到的所有社交圈子中，仍然存在大量噪声。王倩倩[2]提出了基于词激活力模型改进的链接分析方法。基于单词的激活度和亲和度，该方法提出了用户激活度和用户亲密度，并在链接分析中将这两个度量作为节点传递的影响因素。由实验可知，该算法通过改进的链接分析技术提高了用户挖掘的准确率，验证了 WAF 对链接分析技术的改进效果。虽然该算法已经取得了一定的效果，但性能方面还需要改善。

社交网络用户关系挖掘的典型服务是推荐系统。这类研究工作可以进一步的分为好友推荐，内容推荐，兴趣爱好推荐等。最近几年来，许多学者分别提出了基于社交网络用户关系挖掘的好友推荐方法。其中，Yu Z 等[20]应用连接网络的代数连通性来扩展现有的朋友推荐算法，以实现社交网络中的推荐相关性和内容传播。实验结果表明，他们的方法可以在只对朋友推荐成功率进行非常小的牺牲的情况下，显著改善社交网络中的内容传播效率。Huang S 等[21]着重于朋友推荐过程，而 Guo L 等[22]和 Li F 等[23]着重于社交网络中朋友推荐的隐私保护，Chu CH 等[24]关注于对用户的定位 ，Meo 等[25]关注于用户之间的相似性。也有许多学者提出了自己的基于社交网络中用户关系挖掘的内容推荐方法。其中，Chao H 等[26]侧重于开发和应用移动社交互动。Ma H 等[27]尝试在社交网络中使用标签相关和用户社会关系进行微博推荐。Wang Z 等[28]基于用户兴趣矩阵变化的方法，提出了社交网络中用户自制视频的联合社交内容推荐方法，为用户视频条目的推荐提供了基础。此处之外，Yang X[29]系统地讨论了一个基于贝叶斯推理的在线社交网络推荐系统。实验表明，他们提出的基于贝叶斯推理的推荐比现有的基于信任的推荐更好，并且与协同过滤推荐相当。Chen Philip 等[30]和 Wang H[31]等也讨论了用户关系挖掘在社交网络中的其他应用。

总而言之，针对社交网络中特定组织的组织成员识别方法的研究仍然处于起步阶段。如何从结构关系错综复杂的社交网络中识别特定组织的组织成员将会是未来社交网络用户关系挖掘研究的热点问题之一。该问题的研究成果不仅有实际的商业价值，同时也有助于人们对社交网络的群体行为进行分析、研究。

1.2.2 兴趣爱好研究

兴趣爱好是心理学和教育学中非常重要的概念。20 世纪 80 年代以来，来自不同研究领域的学者对兴趣爱好进行了大量的研究。Krapp 等[32]将兴趣爱好分为个人兴趣和情境兴趣。他们认为，个人兴趣相对稳定，随着时间的推移发展，通常与价值，知识和积极情感联系在一起，而情境兴趣则是由一些刺激引发的。Mayer 等[33]阐述兴趣爱好在教育和个人发展中扮演的角色。他们认为，兴趣爱好

是促进学习的重要力量。因此，培养和提高学生的学习兴趣是非常有意义的。Hidi等[34]系统地研究兴趣爱好的培养。他们详细阐述了兴趣爱好培养的四个阶段过程。薛小丽[35]首先从多方面的角度揭示了兴趣与学习的密切关系，提出了兴趣导向的教学理论和模式，将兴趣应用于教育。他们的模式丰富了教学实践，为教学理论的发展提供了有益的途径。在心理学方面，Holtrop[36]，Major[37]和 Houston[6]表明，兴趣在人格的形成与发展，个体心理健康，人格与职业发展中起着重要的作用。此外，Holland[38]认为“职业兴趣是人格的反映”，强调了职业兴趣与人格的关系。近年来，随着互联网的发展，基于兴趣的推荐系统在电子商务和社交网络中得到了广泛的应用。因此，互联网用户的兴趣建模和挖掘已经逐渐展开。例如，Gou等[39]提出了一个新颖的视觉系统，SFViz（社交朋友可视化），以支持用户在感兴趣的情境下交互地探索和发现朋友。在 SFViz 中能够生成社交标签的分层结构，以帮助用户浏览感兴趣的网络。他们的案例研究表明，该系统可以增强用户在不同兴趣情境下的社交网络意识，并可以帮助用户寻找潜在的有着相似兴趣的朋友。Li D 等[40]提出了 Farseer，一种在线社交社区中个性化实时内容推荐和传递系统。他们的方案是识别并利用独特的基于项目的兴趣群集和基于群集的项目评级，以便向个人用户实时推荐新生成的内容项目。Qian X 等[41]基于个人兴趣，个人之间兴趣相似度和人际影响力设计了一个统一的个性化推荐模型。个人兴趣的因素可以使推荐项目满足用户的个性，特别是对于有历史记录的用户。对于冷启动用户来说，人际兴趣相似性和人际影响力可以增强潜在空间中的特征之间的内在联系。他们的实验结果表明所提出的方法优于现有的主要方法。Cao B 等[42]提出了一种基于用户兴趣和服务社交网络向用户推荐混合服务的方法。该方法从用户历史的混合服务中提取用户兴趣，利用目标用户的兴趣和社交网络进行混合服务推荐。大规模的实验表明，他们提出的方法可以有效地向用户推荐合适的用户。此外，Gao H[43]，Yin H[44]，景宁[45]，刘淇[46]和其他学者也在本研究领域提出了自己的方法。Gabriel[47]和 Flinn[48]已经为他们的基于兴趣的推荐系统申请了美国专利。

对于兴趣建模及其应用，Li L 等[49]首先展示了一个关于用户兴趣在现实世界的新闻推荐系统的演变的实验研究，然后提出一个新闻推荐方法，将用户的长期和短期的阅读偏好无缝地整合在推荐新闻项目。他们的实证实验验证了这种方法的有效性。Shen W 等[50]提出了 KAURI，一种基于图形的框架，通过对用户感兴趣的主题建模，将用户发布的所有推文中的所有命名实体集中连接起来。他们认为，每个用户都有一个潜在的主题兴趣分布在各种命名实体中，然后将带有用户兴趣信息的推文集成到一个单一的基于图形的框架。他们的实验结果表明，KAURI 在准确性方面明显优于基准方法。Liu X 等[51]提出并评估了实时兴趣模

型（RIM）的性能，试图识别动态和不断变化的查询级兴趣。除此之外，田军伟[52]从模型表达，用户兴趣收集，模型演化，模型评估四个方面系统地介绍了用户兴趣建模的研究工作。

在兴趣爱好挖掘方面，Deng L 等[53]提出一种基于标签和双向交互的算法来挖掘中国最大的社交服务之一新浪微博的用户的话题兴趣。该算法通过用户交互图的制定，充分地利用了用户之间的相互作用的差异。结果表明，该算法在准确率和召回率方面优于其他方法，能够有效挖掘用户对标签和双向交互的兴趣。Vu T 等[54]构建了一个从 Twitter 消息中提取用户兴趣的系统，该系统使用语言模式提取感兴趣的候选项，并使用四种不同的关键词排序技术对其进行排序：TF-IDF，Text Rank，LDA-Text Rank 和 RI-Rank。结果表明 TF-IDF 和 Text Rank 都适合从推文中提取用户兴趣。Bao H 等[55]提出了一个基于时间和社交概率矩阵分解模型来预测用户在博文中的潜在兴趣。该模型分析了时间信息和用户活动对用户潜在特征空间及其兴趣主题的影响，提供了融合时间信息和社交网络结构的统一方式，以准确预测用户未来的兴趣。除此之外，陈希友[56]，李建廷[57]也提出了自己的兴趣挖掘方法。这些互联网用户兴趣挖掘的方法分别基于访问日志，微博或博客的浏览内容和行为。但是，现有的研究工作很少涉及兴趣爱好本身的内在关系及这些内在关系在兴趣爱好挖掘中的应用。

1.3 论文的研究内容

目前社交网络中用户关系的研究主要集中在社群挖掘问题。社群挖掘方法主要是基于图的聚类。与社群挖掘相比，社交网络的特定组织成员识别具有更直接的目的。组织成员识别，即给定一个组织，试图去发现该组织在社交网络上的成员，为了解决该问题，本文提出并验证了一种基于社交网络用户行为关系的组织成员识别方法。给定一个组织在 Twitter 上的公共账号和一些样本用户（这里的样本用户指的是组织已知的成员，在本文中他们也被称为种子用户），本文的方法可以基于种子用户与 Twitter 候选用户之间的关系挖掘出隶属于该组织的成员。

有了特定组织的组织成员关系信息以后，还需要挖掘组织成员的兴趣爱好，才能进行相应的推荐服务。虽然对兴趣爱好的研究非常广泛，但现有的研究很少基于大数据尝试探索兴趣爱好之间的关系及其应用价值。本文采集了 LinkedIn 中的真实用户数据，并从用户数据中挖掘出用户兴趣爱好，并对兴趣爱好进行标准化，之后进行关联分析并生成了关联规则，最后使用该关联规则对原始的基于词频的 Twitter 用户兴趣爱好挖掘方法进行改进。本文的主要工作总结如下：

（1）基于社交网络 Twitter 特点和用户行为特征，定义了多种识别因子，量化地描述了用户与用户组之间的关系，并根据社交网络用户的关系特点，总结了

用户关系判定规则，将虚拟网络关系的判定转化为数值模型计算的过程。

（2）提出一种基于社交网络用户行为关系的组织成员识别方法。该方法以目标组织的公共账号和种子用户为基础，根据现实生活中人类交际圈的特点，进行迭代的同事关系挖掘。

（3）针对特定组织成员识别方法中的计算模型，基于社交网络 Twitter 真实用户数据进行实证研究实验，将多种识别因子进行组合，探讨最优计算模型。最后通过对比实验，验证了该方法及其计算模型的识别效果。

（4）基于 LinkedIn 用户的个人档案，将不同形式的兴趣爱好字符串使用统一的兴趣项进行标准化整理，并从中挖掘出高频兴趣项作为研究对象，在此基础上建立了兴趣爱好关联分析模型，最后基于真实用户数据生成兴趣爱好关联规则。

（5）基于 Twitter 和 LinkedIn 的实证数据，总结了社交网络 Twitter 用户兴趣爱好的分布特征，并使用兴趣爱好关联规则对原始的基于词频的兴趣爱好挖掘方法进行改进，最后通过对比实验，验证了挖掘效果。

（6）将组织成员识别和兴趣爱好挖掘方法应用在基于海量社交网络数据的人物属性识别系统中，解决了该系统在实际开发过程中遇到的难点。

1.4 论文的组织结构

本文主要包括以下五个部分：

第一章：绪论。介绍了基于社交网络的组织成员识别及其兴趣爱好挖掘的研究背景和意义，详细介绍了国内外研究现状及相关研究成果，最后简述了本文的研究内容和主要工作。

第二章：相关理论和技术。介绍了组织成员识别和兴趣爱好挖掘的基础理论和相关技术，包括本文所涉及的社交网络基本理论与服务平台、MongoDB 非关系型数据存储、关联分析以及关联规则挖掘算法。

第三章：社交网络组织成员识别。首先基于社交网络 Twitter，定义了多种识别因子，量化地描述了用户与用户组之间的关系，并根据社交网络用户的关系特点，总结了用户关系判定规则。然后提出了一种基于社交网络用户行为关系的组织成员识别方法，并对该方法中所使用的计算模型进行最优化实证研究，最后通过实验验证该方法优于现有的方法。

第四章：组织成员兴趣爱好挖掘。首先将 LinkedIn 用户的原始兴趣爱好进行标准化整理，从其中挖掘出高频兴趣项进行关联分析并生成兴趣爱好关联规则。然后分析了 Twitter 用户的兴趣爱好分布特征。最后使用兴趣爱好关联规则对原始的基于词频的兴趣爱好挖掘方法进行改进，并通过实验验证了改进效果。

第五章：基于组织成员与兴趣爱好的应用。首先介绍了基于海量社交网络数

据的人物属性识别系统的应用背景。然后，介绍了组织成员识别与兴趣爱好挖掘方法在系统中的应用情况。最后对系统进行了总结。

第六章：总结和展望。总结了本文的研究工作，指出了本文研究的不足并展望了未来的工作方向。

论文章节结构如图 1.1 所示。

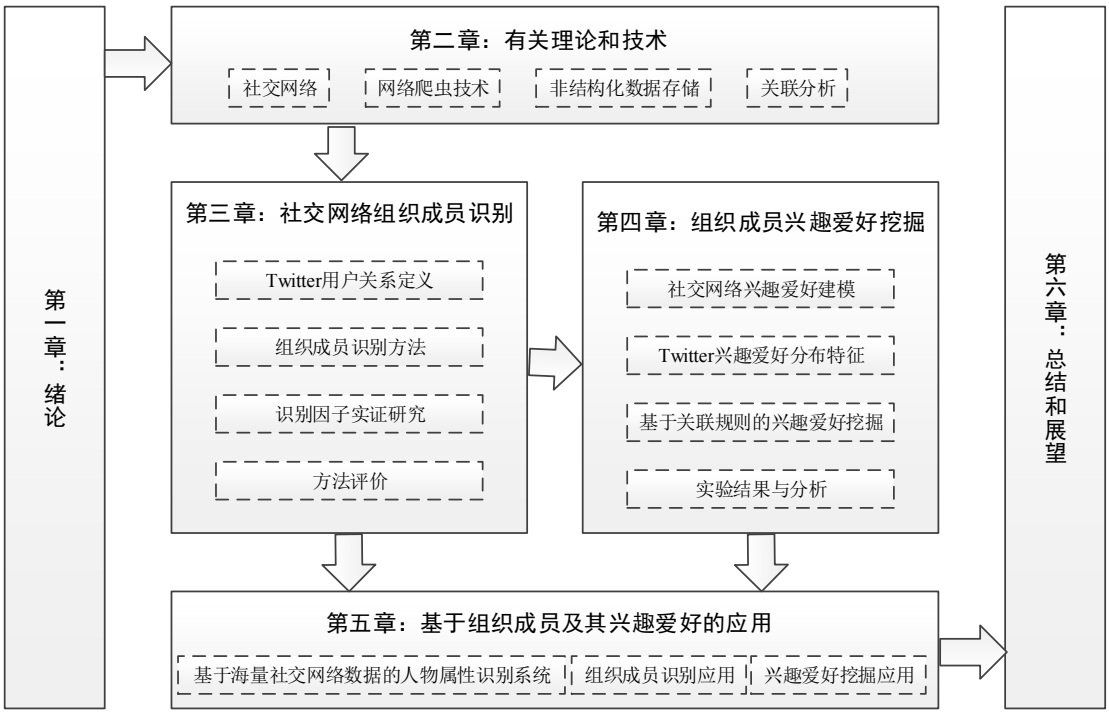

图 1.1 论文章节结构图

第二章 有关理论和技术

随着社交网络的发展与普及，越来越多的人参与其中。社交网络每天都能产生海量的用户原创内容，因此基于社交网络大数据的研究课题也日益增多。基于大数据研究的首要任务是数据的采集，所以网络爬虫是必不可少的技术。其次，因为大数据的特点是“大”而“杂”，所以传统的关系型数据库已经不能满足高效地存储海量异构数据的需求，近几年来非关系型数据库逐渐流行，它的设计思想与功能特性很好地满足了大数据存储的需求。关联分析是研究自然事物之间内在联系的一种机器学习方法，其中的关键步骤是挖掘频繁项集，而 Apriori 和 FP-growth 算法是经典的频繁项集挖掘算法。

2.1 社交网络

2.1.1 社交网络概述

社交网络即社会网络服务。人们在信息上的生活范围由于社交网络的出现得到了极大地扩展，用户可以在社交网络这一新平台上，获取资源、分享信息、表达情感。社交网络一方面为用户提供私人空间，用户可以设置个人信息，例如工作情况，兴趣爱好等，从而可以定位到相关的其他用户；另一方面，社交网络为用户间的信息交流和共享提供了便利。在互联网飞速发展的今天，用户在社交网络中的数据愈显其价值，所以吸引了大批学者对其进行研究。在互联网时代还没有到来之前，就已经有学者对社交网络理论进行了基础研究，邓巴数字和六度分割理论就是这一时期所提出的。这些理论为社交网络快速发展的技术提供了理论支撑，同时，社交网络中的现实数据反过来也验证了这些理论的可靠性。

用户和用户关系构成了社交网络的主体，所以社交网络很适合使用图模型进行描述，图模型中的节点和节点的连接边可以分别代表用户和用户之间的关系。现实中的两个“自然人”之间的二元属性是社交网络关系分析主要研究对象，例如亲属、好友、同事等都属于自然人之间的二元属性关系。基于这种二元属性关系，社交网络分析描述并挖掘了自然人之间错综复杂的关系和这些关系中的隐含的深层信息。

数量庞大的互联网用户及其之间的链接关系构成了社交网络的基础，社交网络可以被看作一种现实的社会关系在虚拟世界中的映射。在研究社交网络时，通常用图$G=(V,E)$形象化地表示社交网络，图中的结点$i\in V$表示一个或几个自然人，边$e\in E$表示自然人之间的二元属性关系。社交关系图可以简单地分为有向

图和无向图，如图 2.1 和图 2.2 所示。

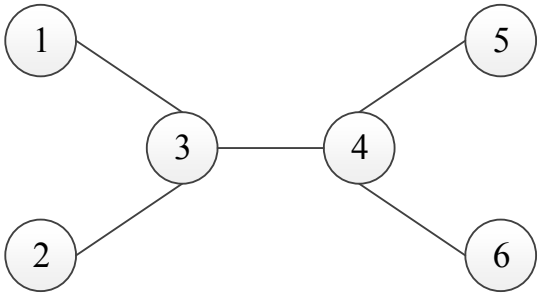

图 2.1 社交关系图中的无向图

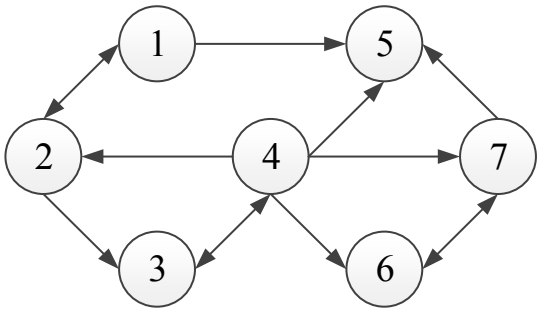

图 2.2 社交关系图中的有向图

2.1.2 Twitter

Twitter（www.twitter.com）是国外的一个社交网络及微博客服务的网站，该网站是由杰克·多西在 2006 年 3 月创办并在当年 7 月启动的，它将即时通讯作为基础功能点，是微博客的典型应用之一，是一个非常流行的在线社交网络服务。Twitter 的通信方式与传统短信通信方式最大的不同在于它不仅仅可以将用户编辑的信息发送给个人，除此之外，还可以分享给用户群组。Twitter 风行于全世界多个国家，是互联网上访问量最大的十个网站之一，根据 Twitter 2017 年第三季度的财务报告显示，Twitter 的月活跃用户达到了 3.3 亿。组织和个人可以在 Twitter 上创建账户，并可以发送和读取 140 字的短消息，称为“推文”。作为一个社交网络，Twitter 围绕着信息分享这一主题[58]。当你选择关注一位 Twitter 用户时，该用户的推文会以时间倒序排列的方式出现在你的 Twitter 主页。用户可以相互关注对方以此便形成了一种关注关系，用户的推文可以通过这种关系传递给关注了自己的用户。为了及时获取自己感兴趣的信息，Twitter 用户通常会关注他们所感兴趣的人。Twitter 用户也可以使用“@”标识符标记用户名的方式将推文定向地推送给一个或多个指定用户[59]。如果一个用户将推文推送给了另一个用户，那他们之间就存在推送关系。Twitter 通过用户之间的关注关系和推送关系来实现信息的传播和共享。

2.1.3 LinkedIn

LinkedIn（www.linkedin.com）是全球最大的职业社交网站，是一家面向商业客户的社交网络，成立于 2002 年 12 月并于 2003 年启动，总部位于美国加利福尼亚州山景城。网站的目的是让网站用户维护他们在商业交往中认识并信任的联系人，俗称“人脉”。用户可以邀请他认识的人成为“关系”圈中的人。LinkedIn 的全球用户数量注册数量已达 5 亿，平均每一秒钟都有一个新会员的加入。LinkedIn 的基本功能允许用户创建个人的资料档案，资料档案主要包含如下部分。第一部分，用户头像。系统允许用户上传一张自己的个人头像，以此增加个人资料档案的可信程度。第二部分，职业概述。综合地展示职业背景、领域、目标与兴趣。第三部分，工作经历。记录了用户过往的所有工作经历与情况，便于让浏览者快速地了解该用户的工作经历。第四部分，教育背景。完善的教育背景将提高个人档案的竞争力，展现全面的自己。第五部分，技能展示。用户可以添加多种个人技能，有利于访客了解用户的工作能力，除此之外，熟人或同事可以为个人技能点赞，以此增加此技能的可信度。在 LinkedIn 的成员通常旨在创建一个专业的个人形象，以此获取商业洞察力，发展专业人脉、寻找更多的就业机会。与其他社交网络相比，LinkedIn 用户可以提供更真实可靠的个人资料信息。

2.2 网络爬虫技术

2.2.1 爬虫技术概述

网络爬虫可以按照自定义规则，自动地抓取网络数据的应用程序或者脚本。网络搜索引擎就广泛地利用了爬虫技术，用户可以根据需求，自定义网络爬虫的访问规则，从而获取这些网站或网页的数据内容。典型的网络爬虫过程可以分解为三个部分，即抓取，加工，存储。

抓取是网络爬虫的基本步骤，爬虫从一个或一些初始的网页出发，获取该网页上的所有外部链接，并根据复杂的过滤条件剔除与目标内容无关的 URL 链接地址，并将剩余可用的 URL 链接地址依次地放入到爬虫队列中。将该页面的数据内容与外部链接信息都访问之后，再从爬虫队列的队首取出新的网页地址，并重复上述过程，直至满足一定停止条件为止。

加工，即对爬虫获取的数据内容进行可用性筛选，例如对网页进行数据爬取，爬虫获取的原始数据大多数都参杂了样式标签等无关内容，这时就需要爬虫程序按照事先设置的过滤规则，将无关内容进行准确的剔除或直接对相关内容进行提取。进行加工操作以后的内容数据才是有效的，可用的。

存储是网络爬虫的最后一步，爬取的内容数据会被合理的存储，并对数据进行进一步地分类过滤处理。如果日后需要对数据内容进行高效地检索，则还应该

为这些数据建立查询索引。传统的网络爬虫系统如图 2.3 所示。总之，网络爬虫面对的问题主要有抓取目标的描述或定义，抓取网站或网页的过滤与存储，以及网页外部链接的搜索策略算法。

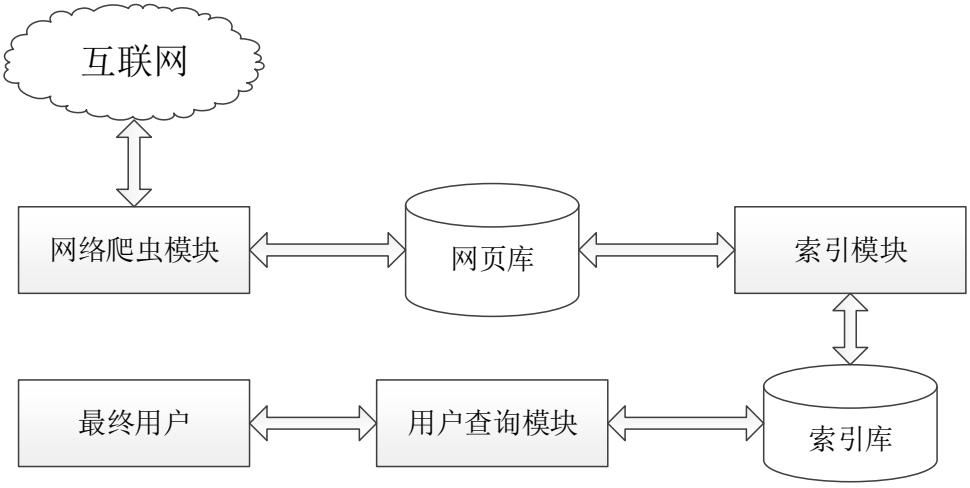

图 2.3 经典的网络爬虫系统

网站爬虫系统的功能是访问网站网页并抓取网页的内容数据，为搜索引擎提供搜索数据来源。网站或网页的数据内容除了访客所浏览的纯文本、多媒体信息之外还存在许多外部链接地址，例如链接到其他网页的 URL 地址，网络爬虫系统正是通过这些错综复杂、相互交错的外部链接地址才可以按规则地在网络中从一个网页跳转到其他网页。

2.2.2 网络爬虫的工作原理

调度器，解析器，数据库三个部分是网络爬虫系统架构的主要构成。调度器的主要工作是负责给不同爬虫线程分配不同的抓取任务。爬虫的基本工作大多数都是由解析器完成的，解析器的主要工作是访问网页并进行网页的分析，将一些网页脚本标签、样式代码内容、空字符、HTML 标签等无意义内容过滤掉。数据库是用来存储处理后的网页资源数据，一般都采用大型的数据库系统存储，如果抓取的数据内容过多，还需要为其建立索引，方便以后的查询操作。以下是对三个主要部分的工作总结：

调度器：调度器是网络爬虫系统的中央控制器，它主要负责根据工作队列分配的 URL 地址，调度并启动爬虫线程进行网站或网页的抓取工作。

解析器：解析器是网络爬虫的“清洁工”，它负责的工作主要有访问并下载网页的原始数据内容，对网页的原始数据内容进行过滤处理，剔除无意义的网页元素，最后按需抓取网页目标信息数据。

数据库：主要是用来保存网页中抓取下来的数据信息的存储仓库，并提供了对抓取数据的查询功能。

网络爬虫系统拥有一个工作队列，工作队列中存放的是准备需要被访问的网

页 URL 地址，每当网络爬虫系统结束对一个页面的抓取工作后，它将从工作队列中取出下一个需要访问的目标对象地址。初始时，网络爬虫系统一般会选择一些搜索权重较高，影响力较大的网站或网页作为起始种子 URL 集合，这样才更有可能访问到其他的高质量网站或网页，互联网中网页之间所组成的结构类似于一个森林，则初始网页就相当于森里中某棵树的根结点。爬虫系统根据自定义规则可以选择广度优先或者深度优先的方式遍历树中的其他结点。广度优先遍历搜索能更广泛的采集靠近网站主页周围的关键信息，而深度优先遍历搜索容易使爬虫系统陷入某个网站的内部，无法访问外部站点的缺陷，所以广度优先遍历搜索算法更适合作为网络爬虫的访问策略算法。网络爬虫系统首先将种子 URL 地址放入爬虫系统的队列中，然后遵循先进先出的原则从队列的队首取出一个 URL 访问其对应的网页。得到网页的数据内容后，经过过滤筛选后将其存储，再经过抓取网页中的外部链接地址可以得到其他目标对象地址，将这些地址放入队尾。然后再取出队首地址，对其进行抓取并解析，周而复始，直到满足停止条件或将整个目标网络访问完毕为止。

2.3 非结构化数据存储

2.3.1 关系型数据库与非关系型数据库

能够支持关系模型数据存储的数据库被称为关系型数据库，从关系型数据库的诞生至今，已经经过了多年的发展。因为其关系模型概念的容易理解，以及关系型数据库软件的操作简单，便于维护等特点，关系型数据库已经被广泛地应用到各个行业的数据管理当中。比较常见的关系型数据库有 MySQL、Oracle 等，这些关系型数据库都有各自的特点及使用领域。

最近几年来，用户自主生成内容的 Web2.0 服务，例如微博、博客、论坛等以交流互动为特点的网站迅速发展，并逐渐成为了主流的网络服务[60]。用户自主生成内容平台的流行使网络信息数据量急剧增长，以往的关系型数据库在存储如此大规模的数据时，面临着许多难以解决的问题。首先是海量数据的写入问题，使用网络服务的用户每天都会产生大量的用户自定义数据，然而传统的关系型数据库在处理这种实时大数据写请求时，其效率会急剧下降甚至导致服务器宕机、丢失数据等严重后果。其次是大规模数据访问请求问题，查询操作是用户最频繁使用的功能，当用户在大数据库表中进行查询操作时，查询操作的效率会显得尤为低下，严重时会达到令用户无法忍受的地步。最后是数据结构扩展的问题，传统的关系型数据库要先定义明确的表结构后才可以进行数据的存储。但由用户自主生成内容网站的用户所生成的用户自定义数据往往都是没有固定数据结构的非结构化数据，所以关系型数据库的扩展能力在面临这种多样化的数据时会遇到

瓶颈。

为了解决上述问题，非关系型数据库开始备受业界的关注。基于不同的数据模型，非关系型数据库可以分为键值存储数据库、列存储数据库、文档型数据库、图形数据库和对象型数据库。非关系数据库最大的特点就是没有固定的表结构约束，用户可以随时修改数据的存储规则，因此在用户自定义数据较多的情况下，非关系数据库的灵活并且易扩展的优势就显得尤其突出。但现阶段大多数的非关系数据库都属于开源项目，没有成熟的商业公司进行安全、稳定地维护，所以出现存储故障的概率要高于成熟、稳定的关系型数据库。

2.3.2 MongoDB

本文中的所有实验数据都以 MongoDB 作为存储平台，同时为了提高实验程序效率以及数据安全，也同时使用了 MongoDB 的分片和副本集机制。

MongoDB 是一种面向文档的非关系型数据库，因为没有采用关系模型的概念，所以非关系型数据库有着更好的扩展性，除此之外，MongoDB 使用的是更为灵活的文档模型，该模型没有类似于传统的关系型数据库的行模型的概念。文档模型的好处在于可以自由的嵌入内层文档或者数组，这样就可以使用一条记录来表示复杂的数据层次关系，这恰好符合面向对象编程语言对数据的定义与理解。

随着计算机行业的迅速发展，应用程序所使用或产生的数据集规模也在迅速地增加，很多应用程序需要存储的数据量已经非常巨大，很多传统的关系型数据库已经不堪重负，难以胜任工作。MongoDB 的设计采用横向扩展的思路。因为需要将巨大的数据量横向得切割到不同的服务器中，MongoDB 采用了面向文档的数据模型来描述数据。除此之外，MongoDB 还有自动处理跨集群的数据和负载，自动重新分配文档，以及对路由请求进行负载均衡等功能。因此，软件开发者能够把大多数的精力放在实现应用程序的功能方面，而不需要考虑底层的数据存储问题。

如果一个集群需要更多的容量来存储新数据，只需要向集群添加新的服务器即可，MongoDB 就会自动将新数据存储到新的服务器中，这就是 MongoDB 的分片机制。分片，也被称为分区，是指将数据拆分，并将其存放到不同机器上的过程。因为每个机器只需要存储一部分的数据，所以对机器的性能要求也就降低了许多，普通的服务器就能达到性能要求。MongoDB 的分片机制对应用程序员来说是透明的，MongoDB 自动完成了对数据的拆分与存储，应用程序员不需要了解其底层实现原理。分片机制允许开发者创建一个包含许多台机器（分片）的集群，将数据子集分散在这些分片中，其中每个分片负责维护着一个数据子集。分片机制达到的目标就是创建一个集群，集群中可以包含任意数量的机器，但对于应用程序员可以将集群看作一台机器来使用。为了对应用程序隐藏底层存储的

实现，在分片操作之前需要执行 mongos 命令进行一次路由操作。该路由服务器维护着一个数据目录，记录了每个分片包含哪些数据内容。应用程序只需要连接到路由服务器，就可以像使用单机服务一样进行操作请求了。

除了分片机制，MongoDB 还支持了复制功能：副本集。使用复制功能可以将一份数据备份到多台机器上，即使一台或多台机器出错，也可以保证数据的安全与完整性。在 MongoDB 中，创建了一个副本集后就可以使用复制功能了。副本集指的是一组服务器，其中有一个主服务器，也称为写服务器，用于处理应用程序的写入请求；还有多个从服务器，也称备份服务器，用于保存主服务器的数据副本。如果主服务器崩溃了，备份服务器会自动选举一个机器变为新的主服务器。使用复制功能时，如果有一台机器宕机了，仍然可以从副本集的其他机器上读取数据。如果某台机器上的数据丢失或损坏了，可以从副本集的其他正常的机器中拷贝一份完整的数据副本。MongoDB 的分片机制与副本集的结构如图 2.4 所示。

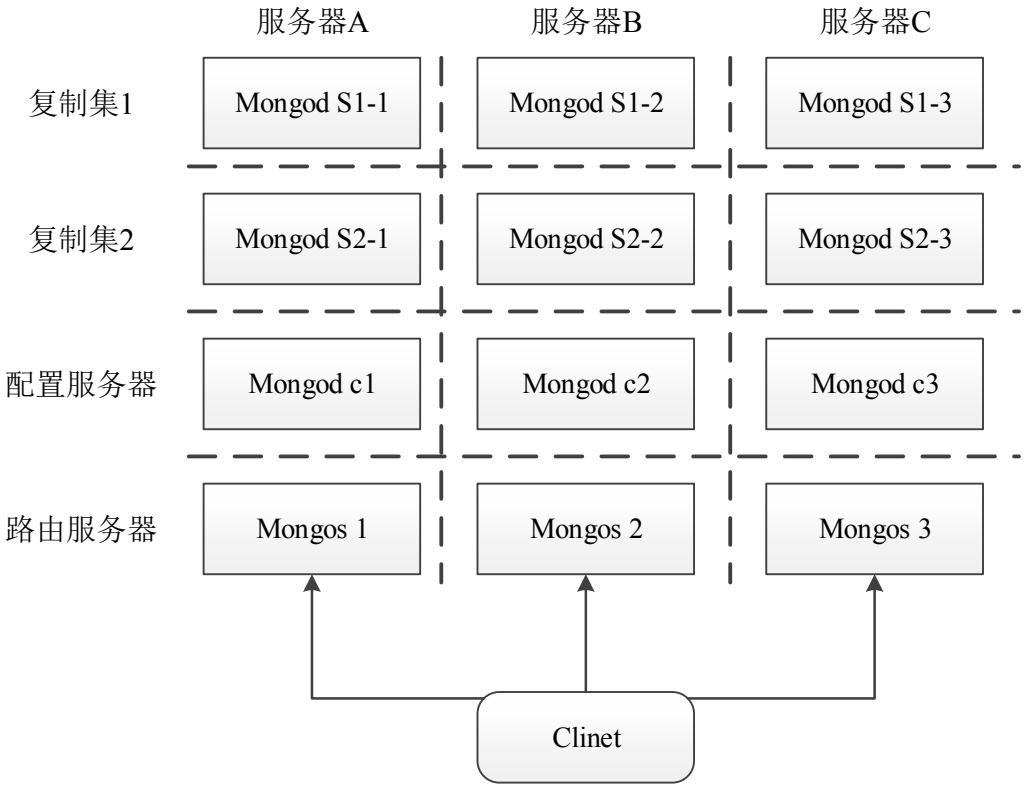

图 2.4 MongoDB 的分片与副本集结构图

2.4 关联分析

2.4.1 关联分析概述

当自然界发生某些事情时，其他的事情将会随之而来，这种关系被称为关联。反映事件之间的依赖性或相关性的知识被称为关联知识。例如，根据体育用品购物单分析，可以发现一些零售规则，例如“购买篮球的顾客中 70％同时购买篮球

运动服”，“40％的顾客同时购买篮球和篮球运动服”。从大规模数据集中寻找物品间的隐含关系被称作关联分析或者关联规则学习。其目的是查找给定数据集中数据项之间的关联规则，并描述数据项之间的紧密程度。关联分析是在大规模数据集中寻找有趣关系的任务。这些关系可以有两种形式：频繁项集和关联规则。关联分析首先找出频繁项集。然后，由它们产生形如 $X \Rightarrow Y$ 的强关联规则，这些规则还满足最小置信度阈值。可以进一步分析关联，挖掘项集 X 和 Y 之间具有统计相关性的相关规则。

关联规则挖掘的数据记录集通常记为 D。$D=\{T_1,T_2,...,T_k,...,T_n\}$，其中 $T_k(k=1,2,...,n)$ 称为一条记录。每一条记录包含一系列项目 $T_k=\{i1,i2,...,im\}$。在关联分析中，关联规则的重要性和价值的度量指标分别是置信度，支持度，期望度和提升度。

置信度：关联规则准确度和强度的度量。数据记录集 D 中的规则 $X \Rightarrow Y$ 的置信度是指在出现 X 的所有记录中出现 Y 的频率，也意味着规则 $X \Rightarrow Y$ 的必然性，表示如下：

$$\text{Confidence}(X \Rightarrow Y)=P(Y|X)=\frac{|\{T: X \cup Y \subseteq T, T \in D\}|}{|\{T: X \subseteq T, T \in D\}|} \times 100\% \quad (2.1)$$

支持度：衡量关联规则的重要性，反映了关联规则的普遍性，并指出关联规则在所有记录集中的出现频繁程度。数据记录集 D 中规则 $X \Rightarrow Y$ 的支持度是指在所有记录中同时出现 X 和 Y 的频率，表示如下：

$$\text{Support}(X \Rightarrow Y)=P(X \cup Y)=\frac{|\{T: X \cup Y \subseteq T, T \in D\}|}{|D|} \times 100\% \quad (2.2)$$

其中 $|D|$ 表示的是数据记录集 D 中所有记录的数量。

期望度：对于 $X \Rightarrow Y$ 规则，它指的是在所有数据记录集中出现 Y 的频率。在规则 $X \Rightarrow Y$ 中，描述了在没有任何影响因素下所有记录集中出现 Y 的频率，表示如下：

$$Expectation(X \Rightarrow Y)=P(Y)=\frac{|\{T: Y \subseteq T, T \in D\}|}{|D|} \times 100\% \quad (2.3)$$

提升度：对于规则 $X \Rightarrow Y$，它描述了 X 的出现如何影响 Y 的出现，它是规则 $X \Rightarrow Y$ 置信度与期望值的比率，表示如下：

$$\text{Lift}(X \Rightarrow Y)=\frac{P(Y|X)}{P(Y)}=\frac{|\{T: X \cup Y \subseteq T, T \in D\}| \times |D|}{|\{T: X \subseteq T, T \in D\}| \times |\{T: Y \subseteq T, T \in D\}|} \times 100\% \quad (2.4)$$

对于可靠的关联规则，其支持度与置信度均应大于设定的阈值。那么，关联分析问题即等价于：对给定的支持度阈值 min_sup、置信度阈值 min_conf，找出所有的满足下列条件的关联规则：（1）支持度 $\geq$ min_sup（2）置信度 $\geq$ min_conf。把支持度大于阈值的项集称为频繁项集。因此，关联规则分析可分为下列两个步

骤：（1）生成频繁项集 $F=X\cup Y$（2）在频繁项集 F 中，找出所有置信度大于最小置信度的关联规则 $X\Rightarrow Y$。

对于包含 K 种项目的数据集共有 2^k-1 种项集组合，关联分析的主要难点在于，寻找物品的不同组合是一项十分耗时的任务，所需的计算代价很高，蛮力搜索方法并不能解决该问题，所以需要用更快捷的方法在合理的时间范围内找到频繁项集。

2.4.2 Apriori 算法

Agrawal 与 Srikant 提出 Apriori 算法[61]，用于做快速的关联规则分析。Apriori 算法是一种较为智能的挖掘项目之间的不同组合的方法，与传统的方法相比，Apriori 算法不需要大量的计算资源就可以在合理的时间内挖掘出频繁项集。Apriori 算法的原理是从只包含单个项目的项集开始，通过添加其他项目形成新的组合，若该组合满足最小支持度的要求则可以生成更大的项集。一个项集在初始数据中的出现频率称为支持度。根据支持度的定义，得到如下的先验定理：

定理（1）：如果一个项集是频繁的，那么其所有的子集也一定是频繁的。该定理比较容易证明，因为某项集的子集的支持度一定不小于该项集。

定理（2）：如果一个项集是非频繁的，那么其所有的超集也一定是非频繁的。定理 2 是上一条定理的逆反定理。

根据定理（2），可以对项集树进行剪枝，如图 2.5 所示。

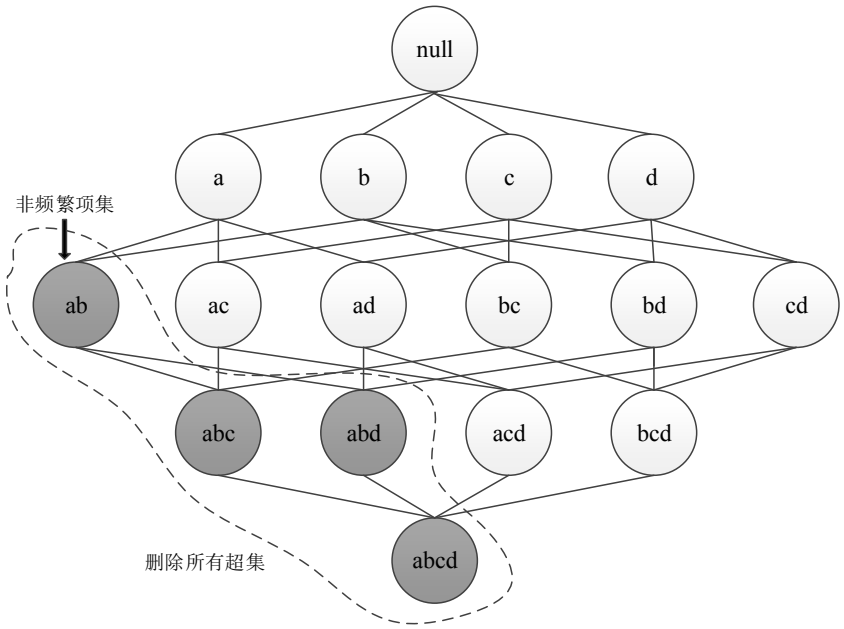

图 2.5 Apriori 算法中对项集树进行剪枝

在 Apriori 算法中，对于计算大小为 K 频繁项集 F_k 给出了两种策略。

（1）$F_k=F_{k-1}\times F_1$ 方法。之所以有时没有选择 F_{k-1} 与所有 1 项集生成 F_k，是因为满足了定理（2）。下图 2.6 给出了由频繁项集 F_2 与 F_1 生成候选项集 C_3。

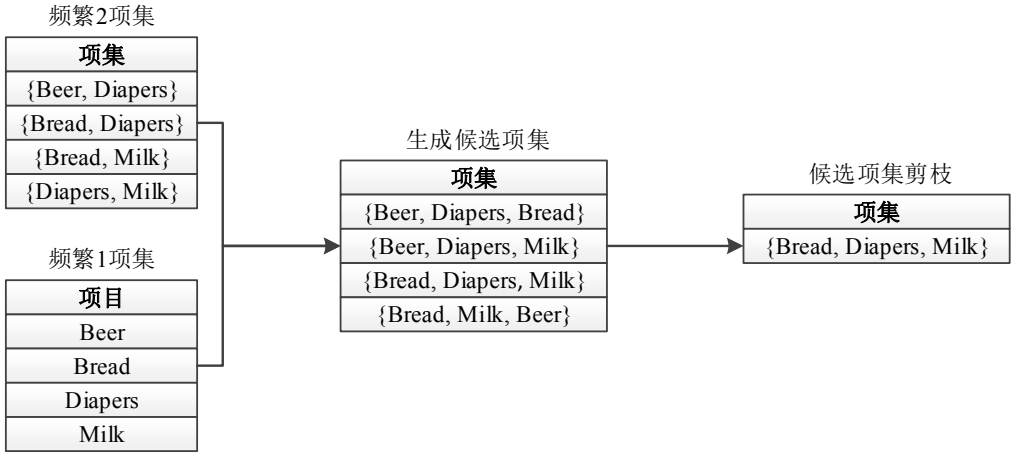

图 2.6 由频繁项集 *F2* 与 *F1* 生成候选项集 *C3*

（2）F_k=F_{k-1}×F_{k-1} 方法。选择前 *k-2* 项均相同的 J_{k-1} 进行合并，生成 F_{k-1}。当然 F_{k-1} 的所有 J_{k-1} 都是有序排序的。之所以要求 *k-2* 项均相同，是为了确保 F_k 的 *k-2* 项都是频繁的。下图 2.7 给出了由两个频繁项集 F_2 生成候选集 C_3。

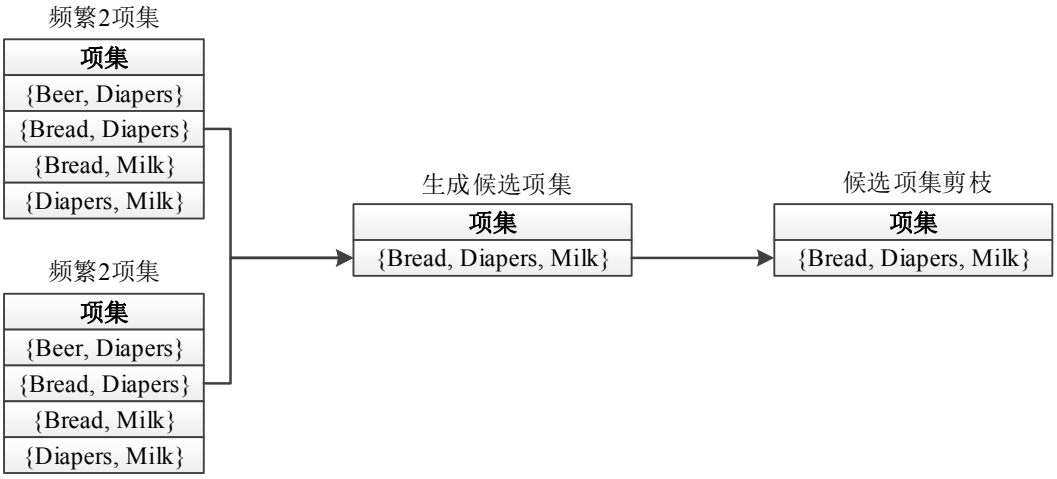

图 2.7 由两个频繁项集 F_2 生成候选集 *C3*

生成频繁项集 F_k 的算法 2.1 如下：

算法 2.1：Apriori 算法生成频繁项集

输入：数据记录集 I，最小支持度阈值 min_sup

输出：频繁项集 F_k

Begin

(1)$k=1$

(2)$F_k=\left\{i\middle|i\in I\wedge\sigma\left(\{i\}\right)\geq N\times\min\_\sup\right\}$. //发现所有1项集

(3)repeat

(4) $k=k+1$

(5) $C_k=apriori-gen(F_{k-1})$. //发现所有候选项集

(6) foreach transaction t∈ T do

(7) $C_t = subset(C_k, t)$. //验证所有属于t的候选项集

(8) foreach candidate itemset c∈ C_t *do*

(9) $\sigma(c) = \sigma(c) + 1$. //支持度计算自增

(10) end for

(11) end for

(12) $F_k = \{c | c \in C_k \wedge \sigma(c) \geq N \times \min\_\sup\}$. //提取频繁$k$项集

(13) until $F_k = \varnothing$

(14) $result = \cup F_k$

End

关联规则是由频繁项集生成的，即对于 F_k，找出项集 h_m，使得规则 $f_k - h_m \Rightarrow h_m$ 的置信度大于最小置信度阈值。同样地，根据置信度定义得到如下定理。

定理（3）：如果规则 $X \Rightarrow Y$ 不满足最小置信度阈值，则对于 X 的子集 X'，规则 $X' \Rightarrow Y$ 也不满足最小置信度阈值。

根据定理（3），可如图 2.8 所示对规则树进行剪枝：

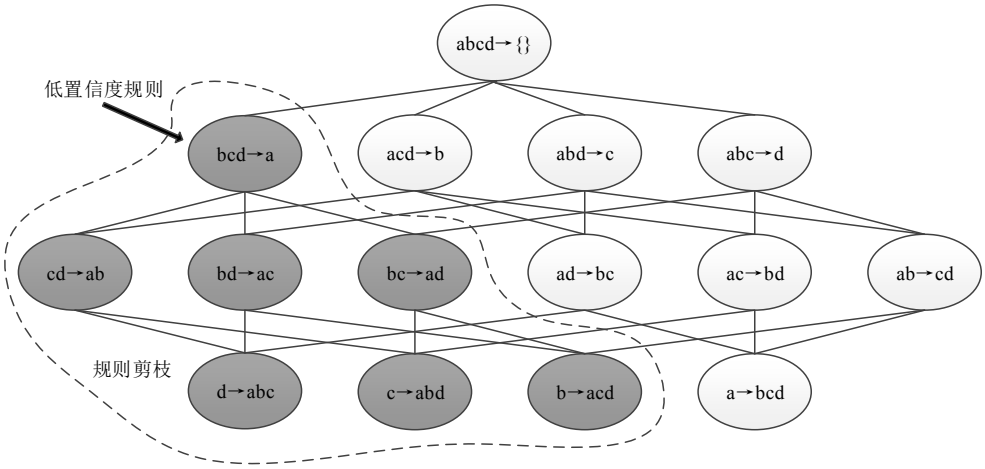

图 2.8 Apriori 算法中对项集树进行剪枝

关联规则的生成算法 2.2 如下：

算法 2.2：Apriori 算法生成关联规则

输入：频繁项集f_k，最小置信度阈值min_conf

输出：关联规则H_m

Begin

(1)$k = |f_k|$ //频繁项集的大小

(2)m $= |H_m|$ //关联规则的大小

(3)if $k > m+1$ then

(4) $H_{m+1} = apriori - gen(H_m)$.

(5) foreach $h_{m+1} \in H_{m+1}$ *do*

(6) $conf = \sigma(f_k) / \sigma(f_k - h_{m+1})$.

(7) if $conf \geq \min\_conf$ *then*

(8) output the rule $(f_k - h_{m+1}) \Rightarrow h_{m+1}$

(9) else

(10) delete h_{m+1} from H_{m+1}.

(11) end if

(12) end for

(13) call ap-genrules (f_k, H_{m+1}).

(14) end if

End

2.4.3 FP-growth 算法

FP-growth 算法[62]是韩家炜等人在 2000 年提出的关联分析算法，它采取如下分治策略：将提供频繁项集的数据库压缩到一棵频繁模式树，但仍保留项集关联信息。

FP-growth 算法的功能与 Apriori 算法相同，都是挖掘项目之间的频繁项集，而非挖掘关联规则，FP-growth 算法挖掘频繁项集可以分解为两个阶段首先是构建 FP（Frequent Pattern）树，其次从 FP 树中挖掘频繁项集。相比与 Apriori 算法，FP-growth 算法拥有更高的挖掘效率。FP-growth 算法只需要对数据集进行两次扫描，第一次扫描时统计项目频率，第二次扫描时挖掘频繁项集，并在扫描的过程中将挖掘的数据存储在 FP 树上。

FP 表示的是频繁模式，其通过链接来连接相似元素，被连起来的元素可以看成是一个链表。将事务数据表中的各个事务对应的数据项按照支持度排序后，把每个事务中的数据项按降序依次插入到一棵以 NULL 为根节点的树中，同时在每个结点处记录该结点出现的支持度。

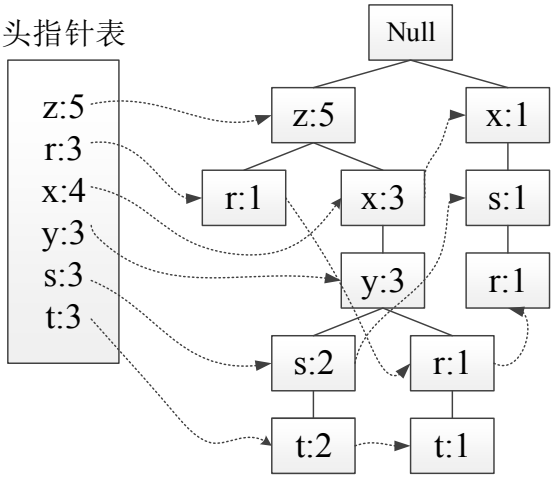

图 2.9 带头指针表的 FP 树

FP-growth 算法还需要一个称为头指针表的数据结构，其实很简单，就是用来记录各个元素项的总出现次数的数组，再附带一个指针指向 FP 树中该元素项的第一个节点。这样每个元素项都构成一条单链表。图 2.9 给出了带头指针表的 FP 树的一个例子。

FP 树会合并相同的频繁项集（或相同的部分）。因此为判断两个项集的相似程度需要对项集中的元素进行排序。排序基于元素项的绝对出现频率来进行。在第二次遍历数据集时，会读入每个项集，去掉不满足最小支持度的元素项，然后对元素进行排序。构造带有头指针表的 FP 树算法 2.3 如下：

算法 2.3：FP-growth 算法创建 FP 树

输入：数据集 data，最小支持度阈值 min_sup

输出：带有头指针表的 FP 树 header

```
Begin
（1）dataSet = set(data)   //去除数据集中的重复元素
（2）foreach k in dataSet:
（3）      header [k] = dat.count(k) //统计各元素出现的次数
（4）foreach k,v in header:
（5）      if v < min_sup:
（6）          del header [k]     //删除小于最小支持度阈值的项
（7）sort header //重排序
（9）return header
End
```

在构建了 FP 树之后，就可以抽取频繁项集了，这里的思想和 Apriori 算法大致类似，首先从单元素项集合开始，然后在此基础上逐步构建更大的集合。大致分为三个步骤：

（1）从 FP 树中获得条件模式基。

（2）利用条件模式基，构建一个条件 FP 树。

（3）迭代重复（1）和（2），直到树包含一个元素项为止。

第一步，获得条件模式基。条件模式基表示的是所查询的项集与根节点路径上的所有内容。为了得到这些路径，结合之前所得到的头指针表，头指针表中包含相同类型元素链表的起始指针，根据每一个元素项都可以上溯到这棵树直到根节点为止。

有了 FP 树和条件 FP 树之后，就可以在前两步的基础上递归得查找频繁项集。查找频繁项集算法 2.4 如下：

算法 2.4：FP-growth 算法查找频繁项集

```
输入：当前数据集的 FP 树 header
输出：频繁项集 freqItemLis
Begin
（1）freqItemList={} //用于存放挖掘出的频繁项集
（2）cnt = 0
（3）foreach list in inDat:
（4）    cnt +=1
（5）    foreach k in header order asc: //按照从小到大的顺序递归 header
（6）        if k ∈    t:     //计算 t 的条件
（7）              k add freqItemList //将新项集加入到 freqItemList
（8）        end if
（9）return freqItemLis
End
```

FP-growth 算法借助 FP 树进行频繁项集的挖掘，虽然 FP-growth 算法同样基于 Apriori 算法的原理，但是 FP-growth 算法在进行挖掘时，只需要对数据集进行两次扫描，所以在相同规模的数据集上，FP-growth 算法的时间复杂度要优于 Apriori 算法。

2.5 本章小结

本章主要介绍了组织成员识别与成员兴趣爱好挖掘过程中所涉及的相关理论知识和技术。首先，对本文所涉及到的社交网络进行了介绍，分析了不同社交网络服务平台的功能特点。其次介绍了网络爬虫技术，使用爬虫技术对本文所需要的数据进行了采集。由于采集的数据量较大且非结构化，所以引入了非关系数据库 MongoDB 作为存储平台，并介绍了其优势和特点。最后介绍了机器学习中

的关联分析技术及其常见的频繁项集挖掘算法。

第三章 社交网络组织成员识别

随着社交网络的兴起和快速发展，网络生活变得更加丰富多彩，几乎每个网民都参与到了这种新型的网络组织结构中。和现实世界一样，虚拟的社交网络中也存在人际关系。从虚拟的社交网络中发掘用户在现实世界中的人际关系有着很高的学术和应用价值，所以社交网络用户关系挖掘已经成为了新兴的研究热点之一。但是当前的研究主要集中于社交网络的社区挖掘，使用或优化图聚类是社区挖掘问题的主要解决方法之一，但这些方法都是将社交网络的用户集合根据用户顶点的结构关联度或属性相似度划分为若干用户集合，而并不能针对某个特定的组织挖掘其相关的成员。

3.1 Twitter 用户关系定义

Twitter 是一个非常流行的在线社交网络服务，用户在 Twitter 上发送的短消息，称为“推文”[63]。作为一个社交网络，Twitter 围绕着信息传播这一主题。当你选择关注另一位 Twitter 用户时，他的推文会以时间倒序排列的方式出现在你的 Twitter 主页。用户可以相互关注对方以此便形成了一种关注关系，用户的推文可以通过这种关系传递给他们的粉丝。所以 Twitter 上的用户通常会关注他们所感兴趣的人。除此之外，Twitter 上的用户也可以使用“@”标识符标记用户名的方式将推文推送给一个或多个指定用户。如果一个用户推送推文给另一个用户，那他们之间就存在推送关系。Twitter 通过用户之间的关注关系和推送关系来实现信息的传播和共享。

因此，给定一个 Twitter 用户，该用户拥有一个主动关注的用户集合和一个被动关注的用户集合，该集合也称为“粉丝集合”。在本文中，若用户在 Twitter 上表示为用户 a，被该用户关注的用户集合表示为 F_a，关注该用户的用户集合表示为 Fed_a，若 $F(a,x)$ 表示 Twitter 上的用户 a 关注用户 x，$F(x,a)$ 表示用户 x 关注用户 a，那么 F_a 和 Fed_a 可以表示为：

$$F_a = \{x | F(a,x)\} \tag{3.1}$$

$$Fed_a = \{x | F(x,a)\} \tag{3.2}$$

类似的，给定一个 Twitter 用户，该用户拥有一个主动推送的用户集合和一个被动推送的用户集合。在本文中，若用户在 Twitter 上表示为用户 a，被该用户推送过推文的用户集合表示为 T_a，给该用户推送过推文的用户集合表示为 Ted_a，若 $T(a,x)$ 表示 Twitter 上的用户 a 向用户 x 推送过推文，$T(x,a)$ 表示 Twitter 上

的用户 x 向用户 a 推送过推文，那么 T_a 和 Ted_a 可以表示为：

$$T_a = \{x | T(a,x)\} \tag{3.3}$$

$$Ted_a = \{x | T(x,a)\} \tag{3.4}$$

若 $M(a,x)$ 表示 Twitter 上用户 a 向用户 x 推送推文的数量，N_a 表示用户 a 主动推送推文的总数量，Ned_a 表示用户 a 被动推送的总数量，那么 N_a 和 Ned_a 可以表示为：

$$N_a = \sum M(a,x) \tag{3.5}$$

$$Ned_a = \sum M(x,a) \tag{3.6}$$

所以，给定一个用户 a 和一个特定的用户集合 S。该方法可以得到 6 种识别因子去量化地描述用户 a 与一组用户之间的关系。6 种识别因子如下所示：

- 用户 a 关注集合 S 中的用户数量记为 G_{as}，称为主动关注：

$$G_{as} = \left| \{x | F(a,x) \cap x, x \in S\} \right| \tag{3.7}$$

- 集合 S 中的用户关注用户 a 的用户数量记为 G_{sa}，称为被动关注：

$$G_{sa} = \left| \{x | F(x,a) \cap x, x \in S\} \right| \tag{3.8}$$

- 用户 a 推送推文给集合 S 中的用户数量记为 T_{as}，称为主动推送人数：

$$T_{as} = \left| \{x | T(a,x) \cap x, x \in S\} \right| \tag{3.9}$$

- 集合 S 中推送推文给用户 a 的用户数量记为 T_{sa}，称为被动推送人数：

$$T_{sa} = \left| \{x | T(x,a) \cap x, x \in S\} \right| \tag{3.10}$$

- 用户 a 推送推文给集合 S 中的推文数量记为 N_{as}，称为主动推送次数：

$$N_{as} = \sum_{x \in S} N(a,x) \tag{3.11}$$

- 集合 S 中推送推文给用户 a 的推文数量记为 N_{sa}，称为被动推送次数：

$$N_{sa} = \sum_{x \in S} N(x,a) \tag{3.12}$$

如图 3.1 所示，根据一个用户与一组用户之间关系的特点属性，例如关注或推送，主动或被动，能系统地划分为 6 种识别因子。

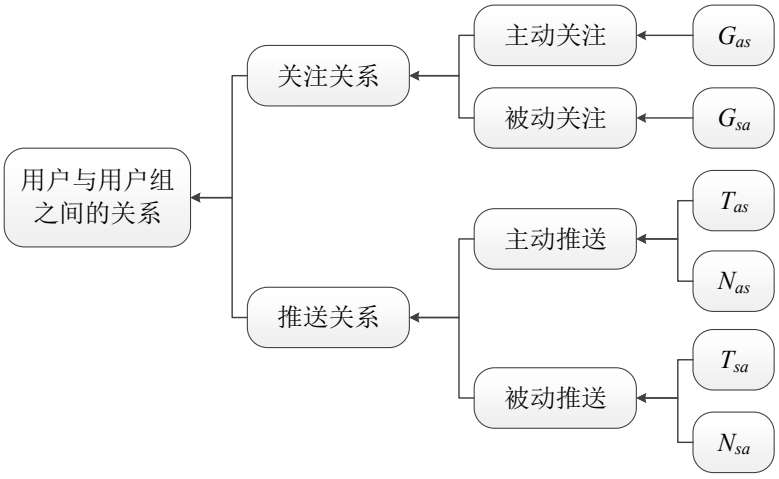

图 3.1 6 种识别因子分类

识别因子 G_{as} 和 G_{sa} 描述了用户 a 与用户集合 S 之间的关注关系，所以它们合称为关注关系因子。并且，G_{as}（用户 a 到集合 S 的关注关系）称为主动关注，G_{sa}（集合 S 到用户 a 的关注关系）称为被动关注。同理地，因子 T_{as}，T_{sa}，N_{as}，N_{sa} 描述用户 a 与集合 S 之间推送关系，所以他们合称为推送关系因子。并且，T_{as} 和 N_{as}（用户 a 到集合 S 的推送关系）称为主动推送，T_{sa} 和 N_{sa}（集合 S 到用户 a 的推送关系）称为被动推送。

为了方便本文中方法的探讨，若 Twitter 上用户 a 是组织 G 的组织成员，表示为 $a \in G$。则用户 a 是组织 G 的组织成员的概率记为 $p(a \in G)$。

3.2 组织成员识别方法

3.2.1 基本规则

（1）在 Twitter 上用户很有可能关注他们所属组织的公共账号。

通常来说，雇员都比较关心他们自己的公司。实际上，一个组织通过 Twitter 发布的推文通常与它的成员或多或少相关。因此，在 Twitter 上用户有关注他们所属组织的公共账号的倾向。所以，Twitter 上某个特定组织的成员有可能在关注该组织公共账户的用户组中找到。

因此若用户 a 关注了组织 G 的公共账号，而用户 b 没有关注，则用户 a 是组织 G 的成员的概率大于用户 b。定义规则（3.1）如下：

规则（3.1）：如果 $F(a,g)$ 并且 $!F(b,g)$，那么 $p(a \in G) > p(b \in G)$。

（2）在 Twitter 属于同一个组织的用户通常很有可能上相互关注对方。

线下熟人是社交网络中朋友的主要组成部分，例如同事，朋友，家人等，他们通常会使用在线的社交网络相互发送消息，所以如果他们都是 Twitter 的用户，则很有可能会相互的关注对方。这就意味着，在 Twitter 上一个用户的被关注集

合和关注集合中可能存在他的同事。

因此若用户 a 比用户 b 关注了更多的属于组织 G 的 Twitter 用户，则用户 a 比用户 b 与该组织有更紧密的关系，这意味着用户 a 是组织 G 的成员的概率大于用户 b。定义规则（3.2）如下：

规则（3.2）： 如果$\left|\left\{x\left|F\left(a,x\right)\cap x,x\in G\right.\right\}\right|>\left|\left\{x\left|F\left(b,x\right)\cap x,x\in G\right.\right\}\right|$，即$G_{as}>G_{bs}$，那么$p\left(a\in G\right)>p\left(b\in G\right)$。

同理，若用户 a 的被关注组中属于组织 G 的用户多于用户 b 的被关注组，则用户 a 比用户 b 与该组织有更紧密的关系，这意味着用户 a 是组织 G 的成员的概率大于用户 b。定义规则（3.3）如下：

规则（3.3）：如果$\left|\left\{x\left|F\left(x,a\right)\cap x,x\in G\right.\right\}\right|>\left|\left\{x\left|F\left(x,b\right)\cap x,x\in G\right.\right\}\right|$，即$G_{sa}>G_{sb}$，那么$p\left(a\in G\right)>p\left(b\in G\right)$。

（3）在 Twitter 上属于同一个组织的用户通常很有可能相互推送推文。

在社交网络 Twitter 上，如果用户使用“@”标识符向其他用户推送推文，那他们通常存在比较紧密的关系，即他们很有可能也是线下熟人。因此，假设在 Twitter 上，若一个用户推送大量的推文给那些属于特定组织 G 的用户，或一个用户收到了大量来自属于特定组织 G 的用户所推送的推文，则该用户与组织 G 有着紧密的关系，该用户很有可能是组织 G 的成员。

因此，若用户 a 相比于用户 b 推送了更多的推文给属于特定组织 G 的成员，则用户 a 比用户 b 与该组织有更紧密的关系，这意味着用户 a 是组织 G 的成员的概率大于用户 b。定义规则（3.4）如下：

规则（3.4）： 如果$\sum M\left(a,x\right)>\sum M\left(b,x\right),x\in G$，即$N_{as}>N_{bs}$，那么$p\left(a\in G\right)>p\left(b\in G\right)$。

除此之外，若用户 a 相比于用户 b 收到更多来自特定组织 G 的成员推送的推文，则用户 a 比用户 b 与该组织有更紧密的关系，这意味着用户 a 是组织 G 的成员的概率大于用户 b。定义规则（3.5）如下：

规则（3.5）： 如果$\sum M\left(x,a\right)>\sum M\left(x,b\right),x\in G$，即$N_{sa}>N_{sb}$，那么$p\left(a\in G\right)>p\left(b\in G\right)$。

类似的，若用户 a 相比于用户 b 给特定组织 G 中更多的成员推送了推文，则用户 a 比用户 b 与该组织有更紧密的关系，这意味着用户 a 是组织 G 的成员的概率大于用户 b。定义规则（3.6）如下：

规则（3.6）：如果$\left|\left\{x\left|T\left(a,x\right)\cap x,x\in G\right.\right\}\right|>\left|\left\{x\left|T\left(b,x\right)\cap x,x\in G\right.\right\}\right|$，即$T_{as}>T_{bs}$，那么$p\left(a\in G\right)>p\left(b\in G\right)$。

同理，若用户 a 相比于用户 b 收到了来自特定组织 G 中更多成员推送的推文，则用户 a 比用户 b 与该组织有更紧密的关系，这意味着用户 a 是组织 G 的

成员的概率大于用户 b。定义规则（3.7）如下：

规则（3.7）：如果$\left|\left\{x\left|T\left(x,a\right)\cap x,x\in G\right.\right\}\right|>\left|\left\{x\left|T\left(x,b\right)\cap x,x\in G\right.\right\}\right|$，即$T_{\mathrm{sa}}>T_{sb}$，那么$p\left(a\in G\right)>p\left(b\in G\right)$。

3.2.2 识别方法

因为 G_{as}, G_{sa}, T_{as}, T_{sa}, N_{as}, N_{sa} 这 6 种识别因子定量地描述用户 a 与用户集合 S 之间的关系，则根据 3.2.1 节的规则（3.1）至（3.7），6 种识别因子的数值越大则用户 a 越有可能与用户集合 S 中的用户存在同事关系。全面的考虑所有情况，基于这 6 种识别因子，本文提出了一种评估模型去评估用户 a 与用户集合 S 中的用户存在同事关系可能性，即用户 a 很有可能也是组织 G 的成员。评估模型定义如下：

$$Score_a = f\left(G_{as},G_{sa},T_{as},T_{sa},N_{as},N_{sa}\right) \tag{3.13}$$

该方法识别的目标用户，即是 $Score_a$ 较高的用户。这些用户更有可能是特定组织的成员。因此，给定一个组织 G 以及它在 Twitter 上的公共账号和一些种子用户。对于 Twitter 上的用户，该方法可以使用基于 6 种识别因子的评估模型,，如式（3.13）所示，计算其 $Score_a$，从而判断该用户是否可能为组织 G 的成员。还需要说明的是，在本文中将种子用户构成的集合称为种子集合 S。

基于规则（3.1），若给定一个组织在 Twitter 上的公共账号，它的成员很有可能出现在集合$\left\{x\left|F\left(g,x\right)\right.\right\}$中，即 F_g。所以基于评估模型，本文设计了一种挖掘特定组织成员的方法。给定一个组织 G 的公共账号和一个由种子用户构成的种子集合 S，在 Twitter 上识别组织 G 的成员的方法其步骤如下：

步骤 1：爬取关注公共账号 g 的 Twitter 用户，其构成了集合$\left\{x\left|F\left(g,x\right)\right.\right\}$，即 F_g。

步骤 2：移除 F_g 中的非候选成员账号，例如其他组织的公共账号，媒体记者账号，使得 F_g 中只包含候选成员账号，得到候选集合 U。

步骤 3：对于候选集合 U 中的每一个用户，爬取其关注列表与被关注列表，即 F_a 与 Fed_a，提取其中的关注关系。

步骤 4：对于候选集合 U 中的每一个用户，爬取其所有推文并提取其中的推送关系。

步骤 5：对于候选集合 U 中的每一个用户，根据种子集合 S，计算 6 种识别因子 G_{as}，G_{sa}，T_{as}，T_{sa}，N_{as}，N_{sa} 的值。

步骤 6：对于候选集合 U 中的每一个用户，基于识别因子 G_{as}，G_{sa}，T_{as}，T_{sa}，N_{as}，N_{sa}，使用评估模型计算每个用户的 $Score_a$。

步骤 7：根据给定的阈值水平从候选集中筛选出 $Score_a$ 大于阈值的用户构

成结果集合 R。

步骤 8：从候选集合 U 中移除本轮产生的结果集合 R，并将新产生的结果集合 R 加入种子集合 S 中。

步骤 9：迭代执行步骤 5 至 8，直到种子集合 S 中成员的数量达到预先设定的期望值。

方法流程图如图 3.2 所示。使用上述方法，该方法可以在 Twitter 上识别出给定组织的潜在成员，所挖掘出来的组织 G 的成员集合就是种子集合 S（除了初始部分的种子用户）。

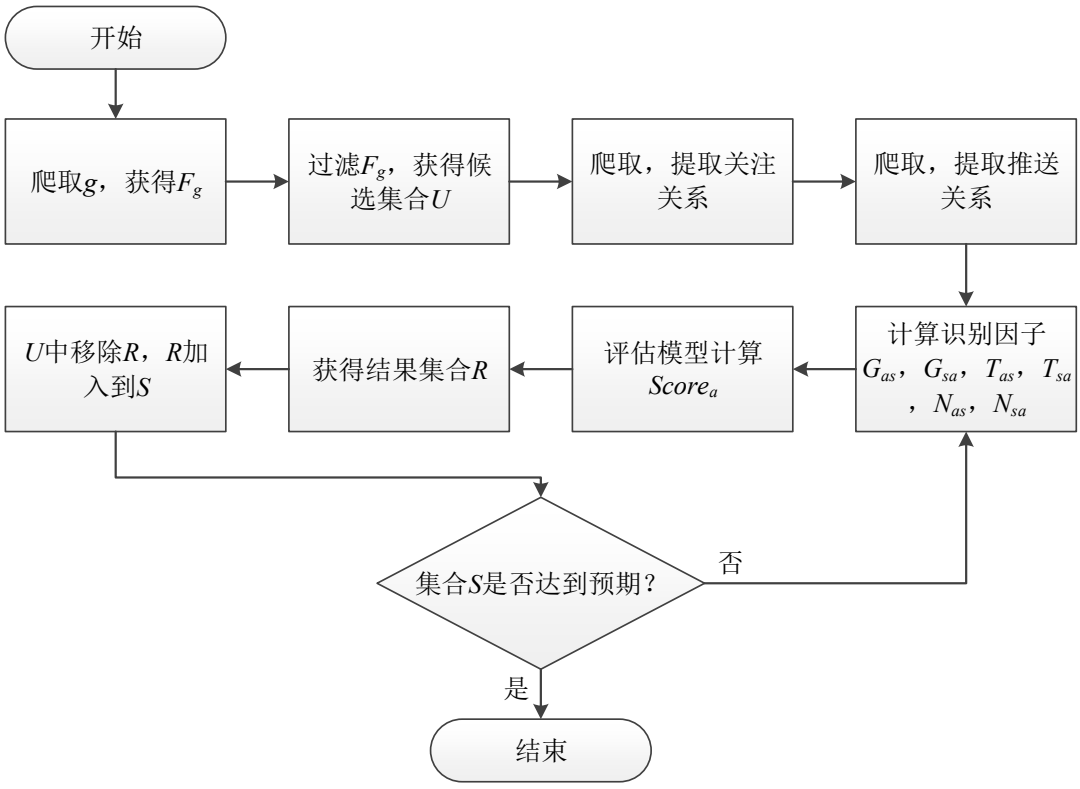

图 3.2 组织成员识别方法流程图

除此之外，在该方法中，应该基于 6 种识别因子设计一种较优的评估模型。较优的评估模型能够尽可能准确地识别 Twitter 上特定组织中的同事关系，所以需要分析 Twitter 中这 6 种识别因子对组织成员识别的影响能力。在 3.3 节中，设计了 5 个实证实验去量化地分析了 6 种识别因子在识别同事关系的影响能力。然后根据实验结果中得出的识别因子影响力，设计了 Twitter 上基于识别因子的组织成员识别方法最优的评估模型。

3.3 识别因子实证研究

3.3.1 实验数据

为了研究特定组织成员识别方法中的 6 种识别因子的影响力，本文抓取了 Twitter 上的某个组织的公共账号，截止到 2016 年 10 月 1 日为止，该公共账号

约有 50 万名粉丝。该组织拥有超过 3000 个雇员，并且他们中有很大一部分是 Twitter 用户。该方法爬取了所有粉丝用户账号从注册时间开始至数据采集截至时间的账号基本信息与其发布的所有推文数据。在本文的实验中，该组织被称为 *G*，它的 Twitter 公共账号称为 *g*。

对于组织 *G*，首先该方法以人工的方式在 Twitter 上搜寻了 200 名用户（已确定为该组织的成员）作为种子用户，称为种子集合 *S*。然后收集了公共账号 *g* 的所有粉丝作为集合 *T*，并将集合 *T* 中的非候选组织成员的账号剔除，例如其他公共账号，媒体记者账号等，剔除后得到候选集合 *U*。除此之外，还需要收集候选集 *U* 中所有用户的推文。与此同时，该方法也使用了人工的方式从候选集 *U* 中搜寻了 546 个组织 *G* 的成员账号，这 546 名用户构成的集合称为验证集合 *E*，用来进行实验评价。

在实验中，基于包含 200 名种子用户的种子集合 *S*，该方法构建了 8 个种子集合的子集合去评估 6 种识别因子的影响力。这 8 个种子集合的子集合分别拥有不同数量的用户，用户都是从种子集合 *S* 中随机抽取的。8 个种子集合的子集合如表 3.1 所示，例如种子集合的子集合 *S1* 包含了 25 名用户，种子集合的子集合 *S2* 包含了 50 名用户等。

表 3.1 基于种子集合构建的子集合

种子集合	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8
用户数量	25	50	75	100	125	150	175	200

3.3.2 实证研究描述

在实证研究中，首先系统地设计了一系列评估模型，即多种且不同形式的评估模型，用于计算组织 *G* 成员识别方法中的 $Score_a$。对于某种评估模型所识别的结果，若该识别结果相对于其他模型的识别结果包含了更多的来自验证集 *E* 的用户，不失一般性，可以认为该评估模型在 Twitter 的同事关系识别方面更具有效性。对于这一系列的评估模型，本实验分析过程如下：

对于每一种评估模型都使用在 3.2 节中提出的方法进行针对组织 *G* 的成员识别实验。在实验中，方法将迭代执行 2 次。第一轮时，方法将从候选集合 *U* 中选出 $Score_a$ 最高的前 500 名用户作为结果集合 R，把他们从候选集合 U 中剔除并插入到种子集合 S 中。第二轮时，从候选集合 U 中选出 $Score_a$ 最高的前 1000 名用户作为结果集合 R。结合第一轮产生的 500 名用户，最后方法能从候选集合 U 中筛选出 1500 名用户，这就构成了最终识别结果。

为了能够系统性地论证各个识别因子的影响能力，本文设计了如下 5 个实证实验：

实验 1：分析、比较了两种主动推送因子 N_{as} 和 T_{as} 的影响能力。在实验中，

设计了一系列的评估模型只比较了主动推送因子 N_{as} 和 T_{as} 在组织成员识别方面的识别能力，然后探讨了因子 N_{as} 和 T_{as} 组合的最优化计算模型。

实验 2：分析、比较了两种被动推送因子 N_{sa} 和 T_{sa} 的影响能力。在实验中，设计了一系列的评估模型只比较了被动推送因子 N_{sa} 和 T_{sa} 在组织成员识别方面的识别能力，然后探讨了因子 N_{sa} 和 T_{sa} 组合的最优化计算模型。

实验 3：分析、比较了两种推送关系的影响能力。在实验中，基于实验 1 和实验 2 的结果，设计了一系列的评估模型比较了主动推送和被动推送在组织成员识别方面的识别能力，然后探讨了因子 N_{as}，T_{as}，N_{sa}，T_{sa} 组合的最优化计算模型。

实验 4：分析、比较了两种关注因子 G_{as} 和 G_{sa} 的影响能力。在实验中，设计了一系列的评估模型只比较了关注因子 G_{as} 和 G_{sa} 在组织成员识别方面的识别能力，然后探讨了因子 G_{as} 和 G_{sa} 组合的最优化计算模型。

实验 5：分析、比较了推送关系与关注关系的影响能力。在实验中，基于实验 3 和实验 4 的结果，设计了一系列的评估模型比较了推送关系和关注关系在组织成员识别方面的识别能力，然后探讨了 6 种识别因子 G_{as}，G_{sa}，T_{as}，T_{sa}，N_{as}，N_{sa} 组合的最优化计算模型。

3.3.3 实证实验

实验 1：在该实验中分析、比较了两种主动推送因子 N_{as} 和 T_{as}。首先定义了如下的两种评估模型：

$$Score_a = f\left(G_{as}, G_{sa}, T_{as}, T_{sa}, N_{as}, N_{sa}\right) = N_{as} \tag{3.14}$$

$$Score_a = f\left(G_{as}, G_{sa}, T_{as}, T_{sa}, N_{as}, N_{sa}\right) = T_{as} \tag{3.15}$$

基于表 3.1 中的种子集合分别使用模型（3.14）和模型（3.15）针对组织 G 的组织成员进行识别实验。对于每种模型产生的识别结果，统计了其能够从验证集合 E 中所识别出的用户数量，并记录在表 3.2 中。

表 3.2 基于不同种子集模型（3.14）与（3.15）识别的用户数

种子集	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8
模型（3.14）	96	266	267	281	281	312	325	300
模型（3.15）	152	261	293	309	300	330	346	345

图 3.3 展示了基于表 3.1 中不同的种子集合，分别使用模型（3.14）与模型（3.15）进行组织成员识别的识别率。

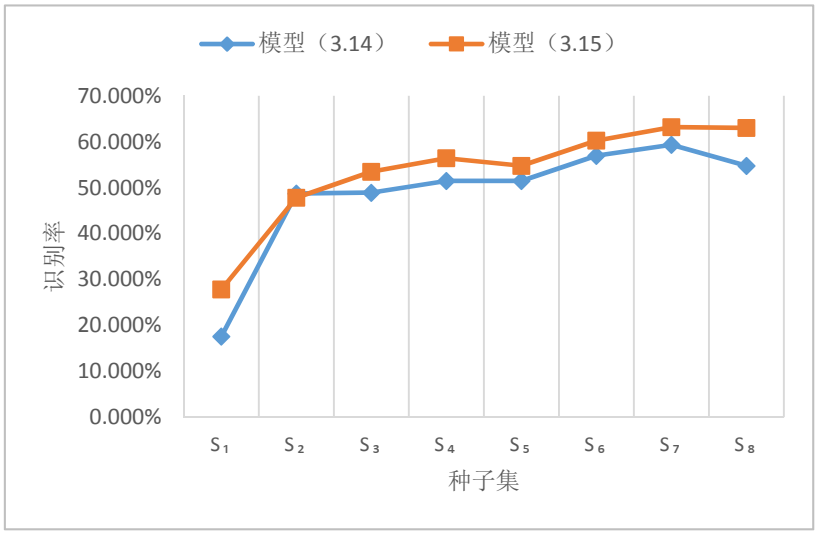

图 3.3 基于不同种子集模型（3.14）与（3.15）的识别率

从图 3.3 中发现模型（3.15）的识别率普遍高于模型（3.14）。基于种子集合 *S8* 时，模型（3.15）的召回率高达 63.003%。所以模型（3.15）相较于模型（3.14），在组织成员识别方面拥有更强的识别能力，因此 T_{as} 相较于 N_{as}，在特定组织的成员识别方面更具影响能力。

为了探究基于主动推送因子 N_{as} 和 T_{as} 组合的最优化模型。不失一般性的，设计了以下基于模型（3.14）和模型（3.15）不同权重组合的评估模型：

$$Score_a = 0.5\times 模型（3.14）+0.5\times 模型（3.15） \quad (3.16)$$

$$Score_a = 0.4\times 模型（3.14）+0.6\times 模型（3.15） \quad (3.17)$$

$$Score_a = 0.3\times 模型（3.14）+0.7\times 模型（3.15） \quad (3.18)$$

$$Score_a = 0.2\times 模型（3.14）+0.8\times 模型（3.15） \quad (3.19)$$

$$Score_a = 0.1\times 模型（3.14）+0.9\times 模型（3.15） \quad (3.20)$$

基于以上所定义的模型和表 3.1 中的种子集合 *S8*，进行针对组织 *G* 的组织成员识别实验。对于以上每种模型产生的识别结果，统计了其能够从验证集合 *E* 中所识别出的用户数量，并记录在表 3.3 中。

表 3.3 基于种子集 *S8* 模型（3.16）至（3.20）识别的用户数

模型	模型（3.16）	模型(3.17)	模型(3.18)	模型(3.19)	模型(3.20)
组合权重	0.5:0.5	0.4:0.6	0.3:0.7	0.2:0.8	0.1:0.9
识别人数	323	326	326	344	348

图 3.4 展示了基于验证集合 *E* 分别使用模型（3.16）至模型（3.20）进行组织成员识别的识别率。

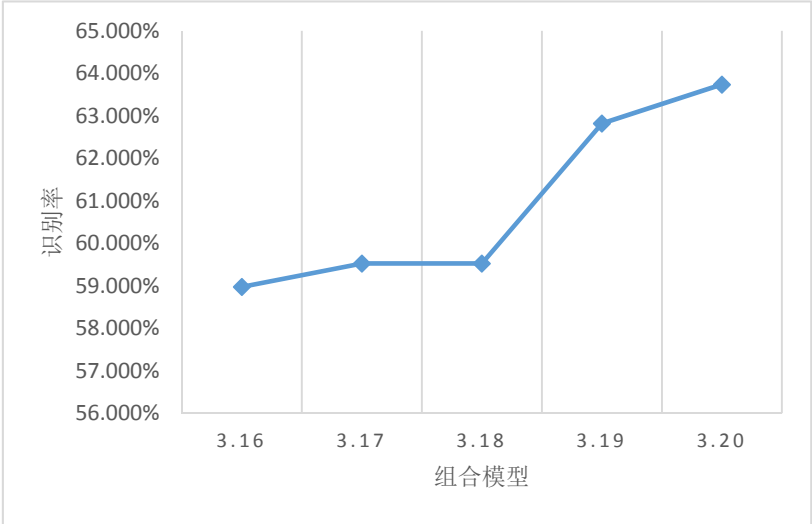

图 3.4 基于种子集 *S8* 模型（3.16）至（3.20）的识别率

从图 3.4 中可以发现模型（3.20），即模型（3.14）与模型（3.15）按照 0.1 与 0.9 的权重组合时，相较于其他组合模型有最高的识别率。基于种子集合 *S8*，模型（3.20）的识别率高达 63.736%，同时该组合模型也优于模型（3.14）和模型（3.15）。因此，综合地考查 2 个主动推送因子 N_{as} 和 T_{as}，得到了组合最优化模型，即模型（3.20），因为它在相同的实验条件下能识别出更多的特定组织成员。

实验 2：在该实验中分析、比较了两种被动推送因子 N_{sa} 和 T_{sa}。首先定义了如下的两种评估模型：

$$Score_a = f\left(G_{as}, G_{sa}, T_{as}, T_{sa}, N_{as}, N_{sa}\right) = N_{sa} \tag{3.21}$$

$$Score_a = f\left(G_{as}, G_{sa}, T_{as}, T_{sa}, N_{as}, N_{sa}\right) = T_{sa} \tag{3.22}$$

基于表 3.1 中的种子集合分别使用模型（3.21）和模型（3.22）进行针对组织 *G* 的组织成员识别实验。对于每种模型产生的识别结果，统计了其能够从验证集合 *E* 中所识别出的用户数量，并记录在表 3.4 中。

表 3.4 基于不同种子集模型（3.21）与（3.22）识别的用户数

种子集	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8
模型（3.21）	237	284	300	310	320	324	331	348
模型（3.22）	237	280	290	318	327	325	339	354

图 3.5 展示了基于表 3.1 中不同的种子集合，分别使用模型（3.21）与模型（3.22）进行组织成员识别的识别率。

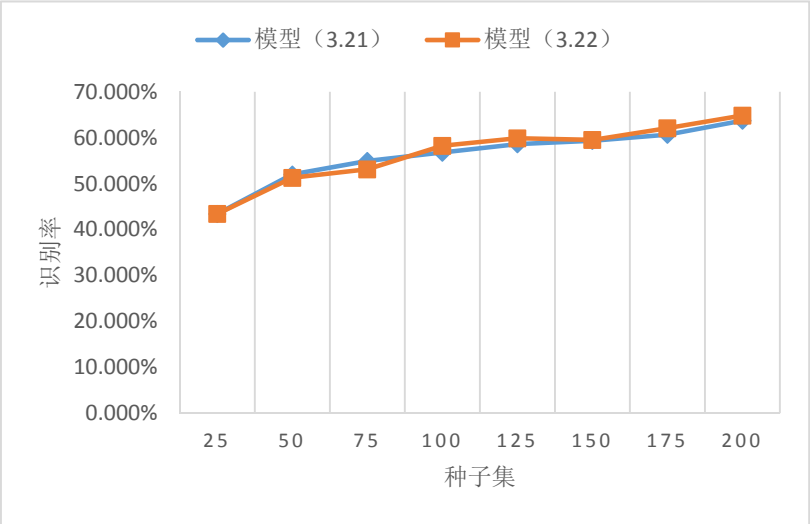

图 3.5 基于不同种子集模型（3.21）与（3.22）的识别率

从图 3.5 中可以发现模型（3.21）与模型（3.22）的识别率没有明显的差距，在某些种子集下，两者的识别率相同，例如基于种子集合 *S1* 时模型（3.21）与模型（3.22）的识别率都是 43.406%。在较小规模的种子集时，模型（3.21）的识别率稍微优于模型（3.22）。在较大规模的种子集时，模型（3.22）的识别率稍微优于模型（3.21）。

为了探究基于主动推送因子 T_{sa} 和 N_{sa} 组合的最优化模型。不失一般性的，设计了以下基于模型（3.21）和模型（3.22）不同权重组合的评估模型：

$$Score_a = 0.9\times 模型（3.21）+0.1\times 模型（3.22） \tag{3.23}$$

$$Score_a = 0.8\times 模型（3.21）+0.2\times 模型（3.22） \tag{3.24}$$

$$Score_a = 0.7\times 模型（3.21）+0.3\times 模型（3.22） \tag{3.25}$$

$$Score_a = 0.6\times 模型（3.21）+0.4\times 模型（3.22） \tag{3.26}$$

$$Score_a = 0.5\times 模型（3.21）+0.5\times 模型（3.22） \tag{3.27}$$

$$Score_a = 0.4\times 模型（3.21）+0.6\times 模型（3.22） \tag{3.28}$$

$$Score_a = 0.3\times 模型（3.21）+0.7\times 模型（3.22） \tag{3.29}$$

$$Score_a = 0.2\times 模型（3.21）+0.8\times 模型（3.22） \tag{3.30}$$

$$Score_a = 0.1\times 模型（3.21）+0.9\times 模型（3.22） \tag{3.31}$$

基于以上所定义的模型和表 3.1 中的种子集合 *S8*，分别进行针对组织 *G* 的组织成员识别实验。对于以上每种模型产生的识别结果，统计了其能够从验证集合 *E* 中所识别出的用户数量，并记录在表 3.5 中。

表 3.5 基于种子集 *S8* 模型（3.23）至（3.31）识别的用户数

模型	模型（3.23）	模型（3.24）	模型（3.25）	模型（3.26）	模型（3.27）
组合权重	0.9/0.1	0.8/0.2	0.7/0.3	0.6/0.4	0.5/0.5
识别人数	348	348	348	348	348
	模型	模型	模型	模型	

	（3.28）	（3.29）	（3.30）	（3.31）
组合权重	0.4/0.6	0.3/0.7	0.2/0.8	0.1/0.9
识别人数	348	348	348	348

从表 3.5 中发现从模型（3.23）至模型（3.31）的每一种模型从验证集合 E 中识别出用户数量都是一样的，均为 348 名，说明无论按照何种方式组合模型(3.21)和模型（3.22）并不能明显提升识别效果。而且在相同种子集条件下，这些组合模型所能识别的用户数量还少于单独的模型（3.22）（表 3.4 中模型（3.22）在基于种子集合 $S8$ 时，能从验证集合 E 中识别出 354 名用户）。因此，就被动推送的两个识别因子而言，模型（3.22）是最优的，因为它在相同的实验条件下能识别出更多的特定组织成员。

实验 3： 在该实验中分析、比较了由实验 1 与实验 2 得出的关于主动推送和被动推送在组织成员识别中的最优化模型。并且为了获得最优化模型，设计一系列基于识别因子 N_{as}，T_{as}， N_{sa}， T_{sa} 的推送关系的评估模型。对于模型（3.20）与模型（3.22），它们是主动推送因子与被动推送因子的最优化模型，所以在该实验中，讨论如何组合这两个模型。

基于表 3.1 中的种子集合分别使用模型（3.20）和模型（3.22）进行针对组织 G 的组织成员识别实验。对于每种模型产生的识别结果，统计了其能够从验证集合 E 中所识别出的用户数量，并记录在表 3.6 中。

表 3.6 基于不同种子集模型（3.20）与（3.22）识别的用户数

种子集	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8
模型（3.20）	160	269	274	297	300	332	346	348
模型（3.22）	237	280	290	318	327	325	339	354

图 3.6 展示了基于表 3.1 中不同的种子集合，分别使用模型（3.20）与模型（3.22）进行组织成员识别的识别率。

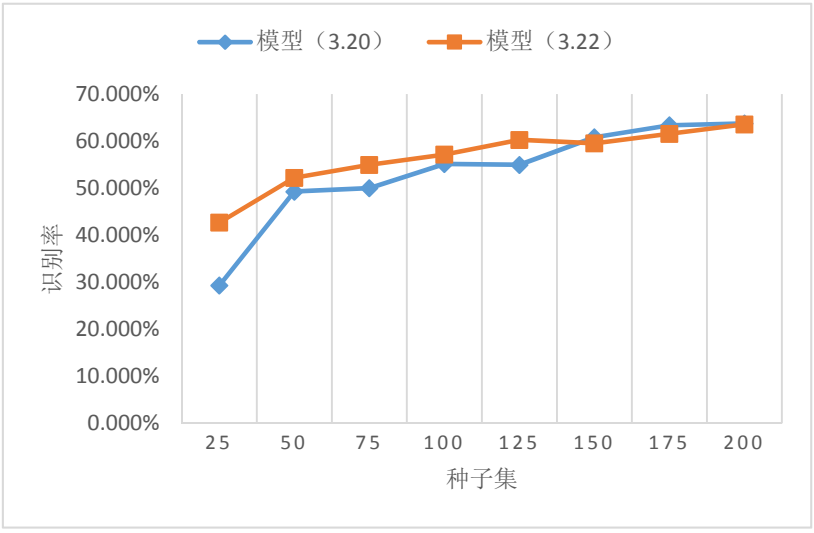

图 3.6 基于不同种子集模型（3.20）与（3.22）的识别率

从图 3.6 中，可以发现在多数种子集的情况下模型（3.22）相较于模型（3.20）有着更好的识别率。所以，模型（3.22）相较于模型（3.20）在组织成员识别方面拥有更强的能力，因此被动推送因子 T_{sa}，N_{sa} 相比于主动推送因子 T_{as}，N_{as} 在特定组织成员识别方面更具有影响力。

为了探究推送因子的最优化模型。不失一般性的，设计了以下基于模型（3.20）和模型（3.22）的不同权重组合的评估模型：

$$Score_a = 0.5\times 模型（3.20）+0.5\times 模型（3.22） \tag{3.32}$$

$$Score_a = 0.4\times 模型（3.20）+0.6\times 模型（3.22） \tag{3.33}$$

$$Score_a = 0.3\times 模型（3.20）+0.7\times 模型（3.22） \tag{3.34}$$

$$Score_a = 0.2\times 模型（3.20）+0.8\times 模型（3.22） \tag{3.35}$$

$$Score_a = 0.1\times 模型（3.20）+0.9\times 模型（3.22） \tag{3.36}$$

基于以上所定义的模型和表 3.1 中的种子集合 *S8*，分别进行针对组织 *G* 的组织成员识别实验。对于每种模型产生的识别结果，统计了其能够从验证集合 *E* 中所识别出的用户数量，并记录在表 3.7 中。

表 3.7 基于种子集 *S8* 模型（3.32）至（3.36）识别的用户数

模型	模型（3.32）	模型(3.33）	模型(3.34）	模型(3.35）	模型(3.36）
组合权重	0.5:0.5	0.4:0.6	0.3:0.7	0.2:0.8	0.1:0.9
识别人数	370	372	373	374	373

图 3.7 展示了基于验证集合 *E* 分别使用模型（3.32）至模型（3.36）进行组织成员识别的识别率。

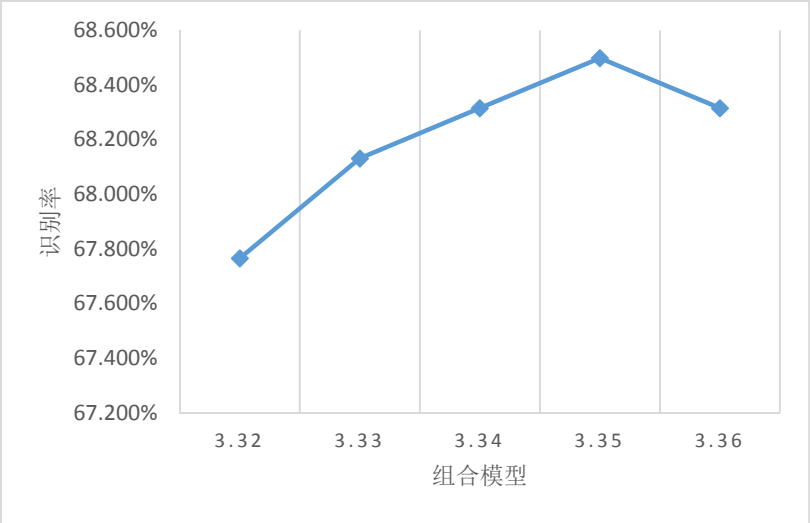

图 3.7 基于种子集 *S8* 模型（3.32）至（3.36）的识别率

从图 3.7 中，可以发现模型（3.35），即模型（3.20）与模型（3.22）按照 0.2

与 0.8 的权重组合时，相较于其他合成模型拥有最高的识别率。基于种子集合 *S8*，模型（3.35）的识别率达到了 68.498%，同时也优于模型（3.20）和模型（3.22）。因此，仅仅考查 4 个推送因子 T_{as}，T_{sa}，N_{as}，N_{sa} 时，模型（3.35）是最优的，它可以在相同的条件下，找到更多的特定组织的用户。

实验 4：在该实验中分析、比较了两种关注关系因子 G_{as} 和 G_{sa}。首先定义了如下的两种评估模型：

$$Score_a = f\left(G_{as}, G_{sa}, T_{as}, T_{sa}, N_{as}, N_{sa}\right) = G_{as} \tag{3.37}$$

$$Score_a = f\left(G_{as}, G_{sa}, T_{as}, T_{sa}, N_{as}, N_{sa}\right) = G_{sa} \tag{3.38}$$

基于表 3.1 中的种子集合分别使用模型（3.37）和模型（3.38）进行针对组织 *G* 的组织成员识别实验。对于每种模型产生的识别结果，统计了其能够从验证集合 *E* 中所识别出的用户数量，并记录在表 3.8 中。

表 3.8 基于不同种子集模型（3.37）与（3.38）识别的用户数

种子集	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8
模型（3.37）	251	313	325	330	338	350	362	370
模型（3.38）	317	354	372	378	386	396	399	404

图 3.8 展示了基于表 3.1 中不同的种子集合，分别使用模型（3.37）与模型（3.38）进行组织成员识别的识别率。

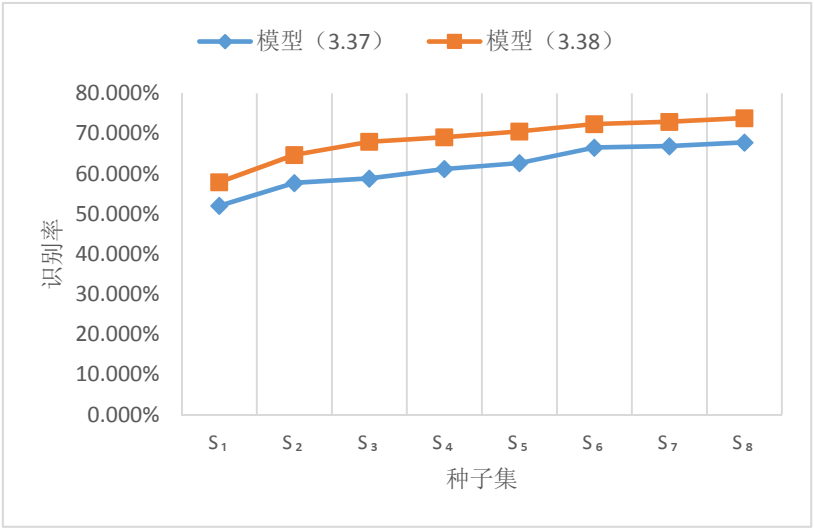

图 3.8 基于不同种子集模型（3.37）与（3.38）的识别率

从图 3.8 中可以发现模型（3.38）的识别率在所有种子集条件下明显高于模型（3.37）。所以，模型（3.38）相较于模型（3.37），在组织成员识别方面拥有更强的能力，因此 G_{sa} 相较于 G_{as}，在特定组织的成员识别方面更具影响能力。

为了探究基于关注关系因子 G_{as} 和 G_{sa} 的最优化模型。不失一般性的，设计了以下基于模型（3.37）和模型（3.38）不同权重组合的评估模型：

$$Score_a = 0.5\times 模型（3.37）+0.5\times 模型（3.38） \quad (3.39)$$

$$Score_a = 0.4\times 模型（3.37）+0.6\times 模型（3.38） \quad (3.40)$$

$$Score_a = 0.3\times 模型（3.37）+0.7\times 模型（3.38） \quad (3.41)$$

$$Score_a = 0.2\times 模型（3.37）+0.8\times 模型（3.38） \quad (3.42)$$

$$Score_a = 0.1\times 模型（3.37）+0.9\times 模型（3.38） \quad (3.43)$$

基于以上所定义的模型和表 3.1 中的种子集合 *S8*，分别进行组织 *G* 的组织成员识别实验。对于每种模型产生的识别结果，统计了其能够从验证集合 *E* 中所识别出的用户数量，并记录在表 3.9 中。

表 3.9 基于种子集 *S8* 模型（3.39）至（3.43）识别的用户数

模型	模型（3.39）	模型（3.40）	模型（3.41）	模型（3.42）	模型（3.43）
组合权重	0.5:0.5	0.4:0.6	0.3:0.7	0.2:0.8	0.1:0.9
识别人数	412	412	415	421	419

图 3.9 展示了基于验证集合 *E* 分别使用模型（3.39）至模型（3.43）进行组织成员识别的识别率。

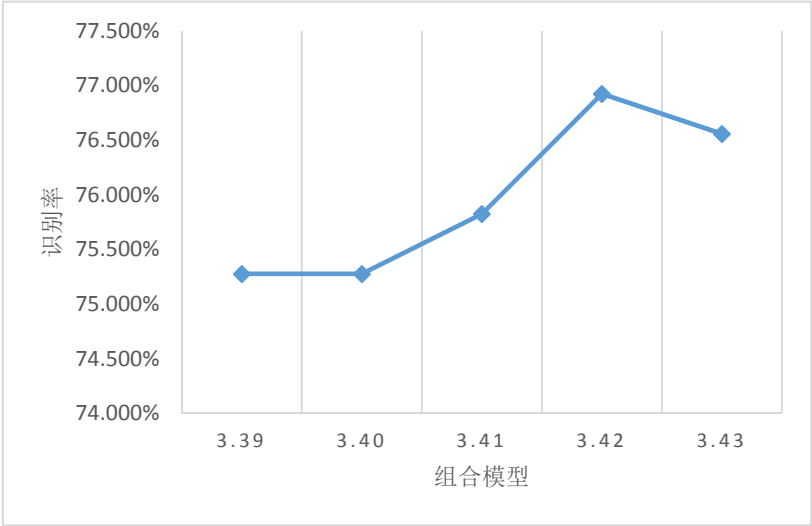

图 3.9 基于种子集 *S8* 模型（3.39）至（3.43）的识别率

从图 3.9 中，可以发现模型（3.42），即模型（3.37）与模型（3.38）按照 0.2 与 0.8 的权重组合时，相较于其他组合模型拥有最高的识别率。基于种子集合 *S8*，模型（3.42）的识别率达到 76.923%，同时也优于模型（3.37）和模型（3.38）。因此，综合地考查两个关注关系因子 G_{as} 和 G_{sa}，得到的最优化模型，即模型(3.42)，它在相同的实验条件下能识别出更多的特定组织成员。

实验 5： 在该实验中讨论了在实验 3 与实验 4 中得到的最优化模型，并评估推送关系与关注关系在组织成员识别方面的能力。然后设计了一系列基于 6 种识别因子的评估模型，并对最优化模型进行了探究。

对于模型（3.35）与模型（3.42），它们是推送关系因子与关注关系因子的最优化模型，所以在该实验中，讨论如何组合这两个模型。基于表 3.1 中的种子集合分别使用模型（3.35）和模型（3.42）进行组织 G 的组织成员识别实验。对于每种模型产生的识别结果，实验统计了其能够从验证集合 E 中所识别出的用户数量，并记录在表 3.10 中。

表 3.10 基于不同种子集模型（3.35）与（3.42）识别的用户数

种子集	S1	S2	S3	S4	S5	S6	S7	S8
模型（3.35）	279	310	330	342	347	348	366	374
模型（3.42）	356	389	403	399	412	415	419	421

图 3.10 展示了基于表 3.1 中不同的种子集合，分别使用模型（3.35）与模型（3.42）进行组织成员识别的识别率。

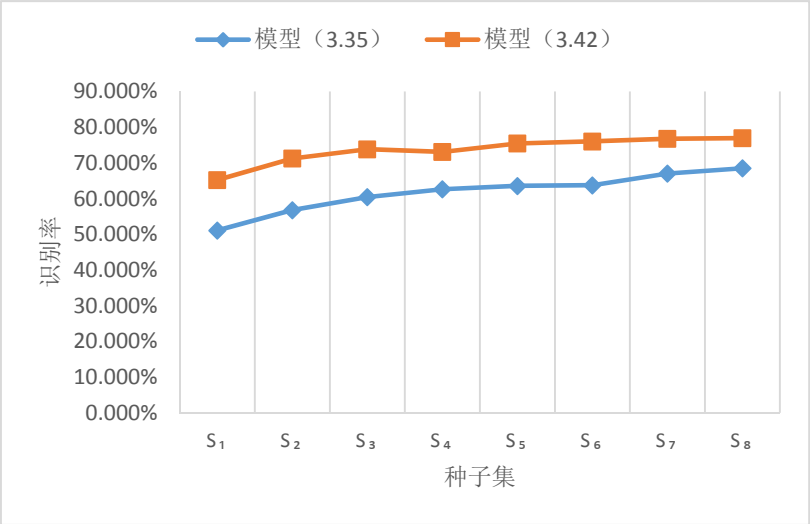

图 3.10 基于不同种子集模型（3.35）与（3.42）的识别率

从图 3.10 中可以发现模型（3.42）相较于模型（3.35）有着更好的识别率。所以模型（3.42）相较于模型（3.35）在组织成员识别方面拥有更强的能力，因此关注关系因子相比于推送关系因子在特定组织成员识别方面更具有影响力。

为了探究最优化模型。不失一般性的，设计了以下基于模型（3.35）和模型（3.42）的不同权重组合的评估模型：

$$Score_a = 0.5\times\text{模型（3.35）}+0.5\times\text{模型（3.42）} \tag{3.44}$$

$$Score_a = 0.4\times\text{模型}\ (3.35)+0.6\times\text{模型}\ (3.42) \quad (3.45)$$

$$Score_a = 0.3\times\text{模型}\ (3.35)+0.7\times\text{模型}\ (3.42) \quad (3.46)$$

$$Score_a = 0.2\times\text{模型}\ (3.35)+0.8\times\text{模型}\ (3.42) \quad (3.47)$$

$$Score_a = 0.1\times\text{模型}\ (3.35)+0.9\times\text{模型}\ (3.42) \quad (3.48)$$

基于以上所定义的模型和表 3.1 中的种子集合 S_8，分别进行组织 G 的组织成员识别实验。对于每种模型产生的识别结果，统计其能够从验证集合 E 中所识别出的用户数量，并记录在表 3.11 中。

表 3.11 基于种子集 *S8* 模型（3.44）至（3.48）识别的用户数

模型	模型（3.44）	模型（3.45）	模型（3.46）	模型（3.47）	模型（3.48）
组合权重	0.5:0.5	0.4:0.6	0.3:0.7	0.2:0.8	0.1:0.9
识别人数	423	422	422	422	422

图 3.11 展示了基于验证集合 E 分别使用模型（3.44）至模型（3.48）进行组织成员识别的识别率。

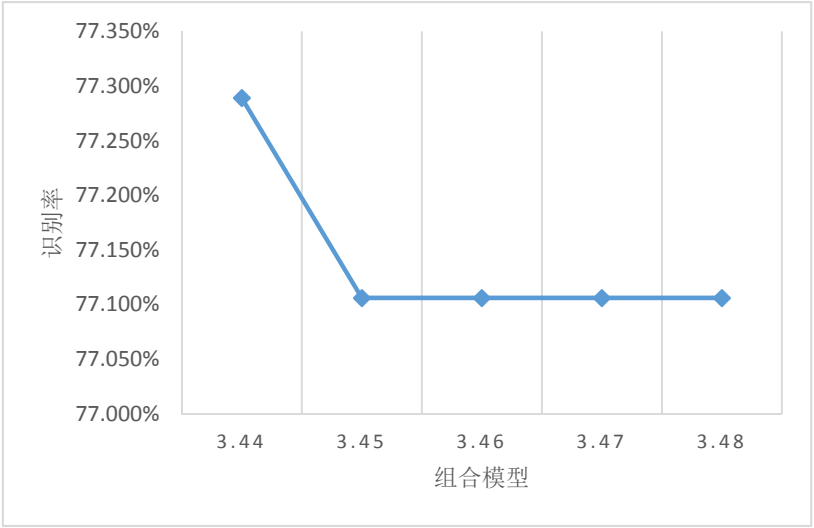

图 3.11 基于种子集 *S8* 模型（3.44）至（3.48）的识别率

从图 3.11 中可以发现模型（3.44），即模型（3.35）与模型（3.42）按照 0.5 与 0.5 的权重组合时，相较于其他组合模型拥有最高的识别率。基于种子集合 *S8*，模型（3.44）的识别率达到 77.289%，同时也优于模型（3.35）和模型（3.42）。因此，综合地考查了推送关系因子与关注关系因子 T_{as}, T_{sa}, N_{as}, N_{sa}, G_{as}, G_{sa} 得到了最优化模型，即模型（3.44），它在相同的实验条件下能识别出更多的特定组织成员。

综上所述，基于实验 1 至 5 的实证研究表明模型（3.44）是最优化计算模型，即将它作为组织成员识别方法中的评估模型。

$$Score_a = 0.5\times\text{模型}\ (3.35)+0.5\times\text{模型}\ (3.42)$$

$$
\begin{aligned}
&= 0.5\times\left(0.2\times\text{模型（3.20）}+0.8\times\text{模型（3.22）}\right)+0.5\times\left(0.2\times G_{as}+0.8\times G_{sa}\right)\\
&= 0.5\times\left(0.2\times\left(0.1\times N_{as}+0.9\times T_{as}\right)+0.8\times T_{sa}\right)+0.5\times\left(0.2\times G_{as}+0.8\times G_{sa}\right)\\
&= 0.01\times N_{as}+0.09\times T_{as}+0.4\times T_{sa}+0.1\times G_{as}+0.4\times G_{sa}
\end{aligned}
\tag{3.49}
$$

3.3.4 模型分析

经过 3.3.3 节的实证研究，得到了本文提出的组织成员识别方法中的最优化模型，如式（3.49）所示。通过分配给每个识别因子的权重可以发现各个识别因子在组织成员识别方面的影响能力。T_{sa} 和 G_{sa} 对于识别同事关系的影响是最大的，它们分别表示的是被动推送人数和被动关注人数。其次是 G_{as}，它表示是主动关注人数。而 T_{as}，N_{as}，N_{sa}，它们分别表示的是主动推送人数，主动推送次数和被动推送次数，它们的影响能力都比较微弱。

特别的，就主动关系和被动关系而言，被动关系的影响力明显高于主动关系，即从一组用户指向一名用户的关系更能说明用户与用户组之间关系的紧密程度。

除此之外，与人数有关的识别因子，例如被动推送人数和被动关注人数，也具有很强的识别能力。这与现实生活中的情况是相符的，即一名用户与一组用户中的多名用户存在互动关系，比一名用户只与一组用户中的个别用户存在多次互动关系更能说明用户与用户组之间存在紧密的关系。

3.4 实验结果与分析

3.4.1 实验平台

实验环境采用实验室 MongoDB 集群，由 4 台相同配置的普通 PC 机搭建而成。其中 3 台为 MongoDB 的分片服务器，1 台为 MongoDB 的路由服务器。具体软硬件配置如表 4.8 所示。

表 3.12 实验环境软硬件配置

硬件	CPU	Intel i5 双核 3.2GHz
	硬盘	500GB SATA 机械硬盘
	网络	100Mbps 局域网
	内存	8GB DDR3
软件	操作系统	Windows 7
	数据库	MongoDB 3.0.6
	开发语言	Python 2.7.11
	集成环境	PyCharm

3.4.2 实验设置

与本文提出的组织成员识别方法（识别方法 M1）进行对比的是文献[19]中提出的方法（识别方法 M2），识别方法 M2 的实验步骤是，首先抓取了实验组织的若干官方账户的两层粉丝的社交网络数据，利用社交网络的拓扑结构计算用户对

目标机构的兴趣度，并在该数据集上进行去噪、压缩处理后，应用基于模度值最大化算法划分出网络中的社交圈子，最后通过定义社区的 *R@N* 指标来选取相关社区，社区中的成员即为组织成员识别的结果。识别方法 M1 的实验步骤如 3.2.2 节所示，最后对比识别方法 M1 与识别方法 M2 的识别结果中包含的组织成员数量。

3.4.3 数据集

本章对比实验所涉及到的数据包括了实验集与验证集。验证集使用的是以人工筛选的方式选取出的 Twitter 上某组织 *H* 的 500 名组织成员账户，这些用户已经确定为组织 *G* 的组织成员。识别方法 M1 所使用的实验集是组织 *H* 在 Twitter 上的官方账号 *a* 的第一层粉丝和 200 名种子用户（组织已知的成员）以及这些用户的关注列表与推文信息。识别方法 M2 所使用的实验集是组织 *H* 在 Twitter 上官方账号 *a*、*b*、*c*、*d*、*e* 的两层粉丝，不同官方账号所拥有的粉丝数量均不相同，具体的两层粉丝数量如表 3.13 所示。

表 3.13 组织 *H* 官方账号的两层粉丝数量

组织 *G* 的公共账号	第一层粉丝数量	第二层粉丝数量
a	522062	61342285
b	33507	3620708
c	4498	685907
d	9806	1336767
e	88185	9118329

3.4.4 实验分析

表 3.14 识别方法 M1 与识别方法 M2 的识别人数

识别方法	M1	M2
识别人数	387	290

识别方法 M1 与识别方法 M2 的识别人数如表 3.14 所示。识别方法 M1 可以在验证集的 500 名组织成员中识别出 387 人，而识别方法 M2 只能识别出 290 人。识别方法 M1 的识别率达到了 77.4%，远远超过了识别方法 M2 的识别率 58%。识别方法 M1 与识别方法 M2 的识别率如图 3.12 所示。

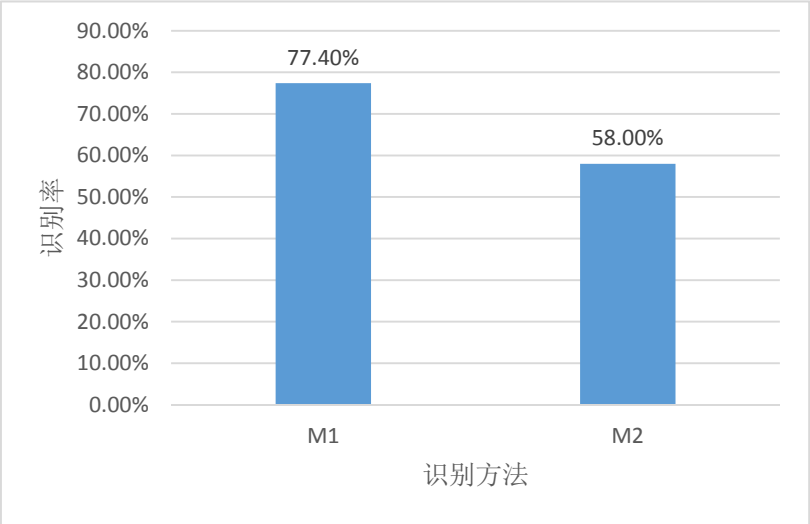

图 3.12 识别方法 M1 与识别方法 M2 的识别率

识别方法 M2 的缺陷在于只单一地考查了社交网络用户与组织公共账号之间的粉丝关系，而没有充分地利用社交网络用户之间其他的行为关系。识别方法 M1 关注的重点是社交网络用户与组织用户组之间的行为关系，根据社交网络的交互特定，识别方法 M1 不仅考查了用户与用户组之间的关注关系，还考查了用户与用户组之间的推送关系。除此之外，用户与用户组之间的行为发起方向，即主动与被动行为也是有很大区别的，所以识别方法 M1 也将每种用户行为关系也进行了主动与被动的细分，最后通过实证研究探究了不同行为识别因子对用户关系判定的影响能力，得到了基于多种识别因子的最优组合计算模型。通过对比实验的结果可以得出识别方法 M1 的识别效果明显的优于识别方法 M2，这也是符合实验预期的。

3.5 本章小结

本章提出了一种基于社交网络用户行为关系的组织成员识别方法。本章首先介绍了 6 种用户行为识别因子，使用识别因子去量化地描述 Twitter 上用户与用户组之间的关系。然后，总结了社交网络中用户关系判定的基本规则，并介绍了 6 种识别因子在基本规则中的运用。随后通过实证研究，系统地分析了 6 种识别因子在识别 Twitter 组织成员关系方面的影响能力，由此得出了最优评估模型。只要给定一个组织在 Twitter 上的公共账号和一些种子用户，该方法可以挖掘出属于该组织的其他成员。最后，本文也对该方法进行了对比实验，实验的结果验证了该方法比现有的基于粉丝关系的方法具有更高的识别率。

第四章 组织成员兴趣爱好挖掘

兴趣爱好是心理学和教育学的一个重要概念。一般来说，个人的兴趣爱好被视为理解某一特定主题的持久的内在需求，是一种认知和情感诉求，使个人能够保持这种需求。兴趣爱好对人格的形成，心理健康，教育和职业发展都有重要的影响，在许多领域得到广泛的研究。特别是近几年来，基于兴趣爱好的推荐系统的广泛应用极大促进了兴趣爱好建模和社交网络兴趣爱好挖掘的研究。但是，现有的研究很少探究社交网络中用户兴趣爱好的分布特征和兴趣爱好之间的关联关系以及兴趣爱好关联关系的应用价值。本章从 LinkedIn 用户的个人档案中挖掘出兴趣爱好的关联规则，并将这些规则用于 Twitter 用户的兴趣爱好挖掘工作中。

4.1 社交网络兴趣爱好建模

4.1.1 兴趣爱好数采集

LinkedIn 是一个非常受欢迎的创业和就业导向的社交网络服务。截至 2016 年 9 月，LinkedIn 已有超过 4.67 亿个账户。LinkedIn 的基本功能允许用户创建个人档案。一个典型的档案描述自己的工作经验，教育和培训，兴趣爱好，以及他们的照片。在 LinkedIn 的成员通常旨在创建一个专业的个人形象，获取商业洞察力，发展专业人脉、寻找更多的就业机会。与其他社交网络相比，LinkedIn 用户可以提供更真实可靠的个人资料[64]。

LinkedIn 用户通常会在个人档案中列出他们的兴趣。一些兴趣爱好总是出现在同一个档案上，这表明这些兴趣爱好是有内在联系的。例如，“read”和“travel”通常同时出现，它们之间必然有密切的关系。因此，可以收集带有兴趣爱好的 LinkedIn 职业资料，以分析兴趣爱好的相关特征。数据采集阶段，首先使用 LinkedIn 网络爬虫，随机收集 44,623 个 LinkedIn 个人资料，其中 10,028 个填写了他们的兴趣爱好。

4.1.2 兴趣爱好识别与标准化

当 LinkedIn 用户创建个人档案时，LinkedIn 并没有准备一组兴趣爱好以供选择，这样的做法对用户的兴趣爱好而言是很开放，LinkedIn 用户可以使用任何词汇，自由地编辑他们的兴趣爱好。所以，LinkedIn 用户所填写的兴趣爱好并不规范。在 LinkedIn 个人档案的兴趣爱好列表中，不同兴趣之间没有固定的分隔符。有些用户使用单词“and”，有些使用逗号“,”，有些用分号“;”，有些用户

直接用行来划分不同的兴趣。例如，有些用户的兴趣爱好列表是这样的“Movies and walking”，有些却是这样的“Yoga; hiking; singing; reading; poetry; art; music; Kids!”。此外，LinkedIn 用户可以使用不同的词汇表达相同的兴趣爱好。在自然语言中，相同的兴趣往往有各种不同的表达。例如，“ski”也可以写为“skiing”，“book”也可以写为“books”等。在本文中，收集处理兴趣爱好数据的过程如下。

首先设计了一种分词算法，可以准确地分离 LinkedIn 成员的兴趣爱好列表，以便将兴趣词识别为每个用户的集合。从 10,028 个有兴趣爱好的个人档案中，可以发现 25,913 个兴趣词，它们分别代表一种兴趣爱好。毫无疑问的，有一些兴趣爱好词汇是同义词，例如“ski”和“skiing”，“book”和“books”这些都代表了相同的兴趣爱好，只是使用了不同的词汇表达。然后进行同义词的识别并将其聚合成相同的兴趣项。经过人工地校对后，得到了所有兴趣词的 19430 个同义词集。毫无疑问的，兴趣词的同义词集合对应于一个兴趣项。为了便于描述工作，本文使用同义词集合中最常用的兴趣词命名相对应的兴趣项。根据同义词集合，将用户个人档案中的由用户自己填写兴趣词替换为兴趣项。然后，针对每个兴趣项，计算出其在 10,028 份个人资料中出现的频率及出现次数与所有兴趣项的总出现次数的百分比，这表明了该兴趣项的普遍性。部分结果如表 4.1 所示。

表 4.1 LinkedIn 中部分兴趣项的出现频率

兴趣项	频次	占比	兴趣项	频次	占比
travel	1689	3.12%	cycling	582	1.08%
music	1266	2.34%	running	551	1.02%
read	1140	2.11%	business	542	1.00%
technology	1073	1.98%	sport	524	0.97%
photography	811	1.50%	cooking	491	0.91%
movie	772	1.43%	art	473	0.87%
ski	742	1.37%	design	445	0.82%
golf	674	1.25%	family	427	0.79%

兴趣项出现频率按降序排列，如图 4.1 所示，前十名兴趣项的累计百分比达到 17.02％，前 50 名达到 37.69％，前 100 名达到 46.63％，而 前 210 名达 52.36％。因此，可以发现，19430 个兴趣爱好的频率分布是非常不平衡的，很少有兴趣爱好的频率很高，大多数兴趣爱好的频率很低。

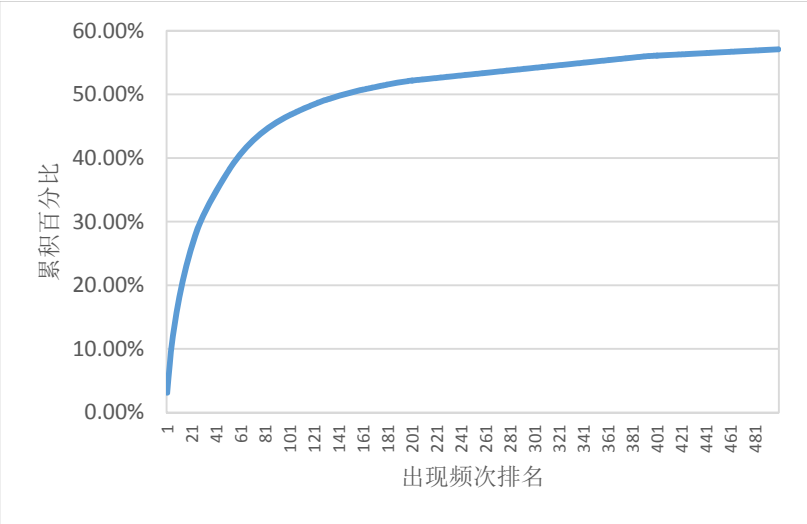

图 4.1 前 494 名兴趣项的累计百分比

毫无疑问，兴趣项的出现频率越高，表示兴趣项越受欢迎，分析价值也就越高。因此对出现频率较低的低频兴趣项进行了删除处理，并保留 210 个高频兴趣项作为研究对象。在实验数据中，10,028 份个人档案中有 8,675 份包含 210 个兴趣项中的至少一个兴趣项。因此，对于每个 LinkedIn 个人档案，只需将属于 210 个兴趣项的兴趣爱好进行标准化，就可以获得他的兴趣项的标准化表现。一些例子如表 4.2 所示。

表 4.2 LinkedIn 档案兴趣爱好列表标准化样例

用户 ID	用户兴趣爱好列表	标准化后的兴趣爱好集合
168915697	New technology\nSciences\nLanguages	technology; language; science
27428582	Wine, food and good music!	music; food; wine
113724463	Rugby; Golf; Travel and adventures	travel; golf; rugby; adventure
7735645	Theater & Improvisation; Tango dancing.	dance; movie
135927166	Sports; Drama class; Philosophy.	sport; philosophy
182769915	Rugby\nBoxing\nSocial media	rugby; media; boxing
145915690	Travelling; Football and Fishing	travel; fishing; football
134641445	Reading; writing; music; eating; cooking; traveling; camping; hiking	read; music; cooking; hike; travel; writing; camping; eating
60091	Music; Writing & Great Food	music; writing; food
13715016	music: piano and guitar; photography (b&w); skiing; badminton	music; photography; ski; guitar; badminton; piano

4.2 Twitter 兴趣爱好分布特征

4.2.1 分布特征假设

Twitter 是一个在线社交网络平台。用户可以在 Twitter 上创建账户，发布和阅读简短的 140 字信息，称为“推文”。用户的推文能传播给他的粉丝。目前 Twitter 是一个非常受欢迎的用户信息发布平台，拥有超过 5 亿用户。毫无疑问，用户可能会发布一些他们感兴趣的推文。所以可以做如下假设。

假设 1:可以表达 Twitter 用户兴趣爱好的词语通常出现在他的推文中。换句话说，Twitter 用户的推文中提到的兴趣爱好可能也是 Twitter 用户的兴趣爱好。

假设 2: Twitter 用户的推特中出现频次越高的兴趣爱好，就越有可能是用户的真正兴趣爱好。

4.2.2 分布特征验证

为了验证 4.2.1 节中的假设，首先从 Twitter 上收集 930 名 Twitter 用户的推文，他们都是 LinkedIn 上的成员，并且已经在 LinkedIn 上提供了他们的真实兴趣爱好。然后，针对给定 Twitter 用户，可以挖掘出他所有的推文中提到的兴趣项及其出现频次，其中的兴趣项已经在 4.1 节中做了叙述。毫无疑问，从用户的所有推文中挖掘出来的这些兴趣项并不一定是他的真正兴趣爱好，但他的真正兴趣爱好可能在其中。例如，在用户 Melgallant 在 Twitter 上的所有推文中，挖掘出了 128 个兴趣项和它们相应的频次。把这些兴趣项按照频次降序排列，同时也将他在 LinkedIn 上的真实兴趣爱好列表进行 4.1.2 节中标准化，得到其真实兴趣项集合，结果在表 4.3 中展示。

表 4.3 Twitter 用户 Melgallant 兴趣爱好挖掘样例

用户昵称	Melgallant
LinkedIn 真实兴趣项集合	media; editing; international relations; writing
来自 Twitter 的有序兴趣项列表	***<media,311>***; <art,261>; <read,104>; <leader,103>; <performance,71>; <Canada,58>; <coffee,49>; <video,43>; <culture,43>; <marketing,43>; <ski,37>; <business,37>; <health,30>; <rock,27>; <internet,27>; <food,25>; <building,25>; <movie,24>; <kids,24>; <design,23>; <UK,20>; <communication,17>; <sport,15>; <surf,14>; <eating,14>; <family,14>; <planning,13>; <talent management,13>; <music,13>; <wine,13>; <dog,13>; <dance,12>; <technology,12>; ***<writing,11>***; <drink,11>; <hockey,10>; <law,9>; <bridge,8>; ***<editing,8>***; <skate,8>; <analytics,7>; <shopping,7>; <mentor,7>; <nature,7>; <recruitment,6>; <science,6>; <rowing,6>; <sales,6>;

	<gas,6>; <blogging,6>; <research,6>; <friends,6>; <travel,6>; <innovation,5>; <yoga,5>; <speaking,5>; <shooting,5>; <painting,4>; <security,4>; <startup,4>; <acting,4>; <fashion,4>; <running,4>; <bigdata,3>; <android,3>; <motivation,3>; <fitness,3>; <philanthropy,3>; <camping,3>; <china,3>; <marathon,2>; <risk,2>; <golf,2>; <fishing,2>; ***<international relations,2>***; <environment,2>; <opera,2>; <driving,2>; <drums,2>; <cooking,2>; <Asia,2>; <singing,2>; <museum,2>; <India,2>; <oil,2>; <bass,2>; <cars,2>; <garden,2>; <volunteering,2>; <walking,1>; <spirituality,1>; <soccer,1>; <finance,1>; <photography,1>; <architecture,1>; <bowling,1>; <branding,1>; <computer,1>; <customerservice,1>; <psychology,1>; <entertainment,1>; <Europe,1>; <baseball,1>; <acquisitions,1>; <elearning,1>; <hunting,1>; <language,1>; <programming,1>; <advertising,1>; <retail,1>; <consulting,1>; <creativity,1>; <WordPress,1>; <assessment,1>; <nutrition,1>; <television,1>; <history,1>; <computing,1>; <politics,1>; <education,1>; <economy,1>; <gaming,1>; <adventure,1>; <crafts,1>;
查全率	100.0%
查准率	3.91%
F1 值	51.95%

在本文中，查全率是指挖掘出的兴趣项与真实兴趣项的百分比，查准率指真实兴趣项与挖掘出的所有兴趣项的百分比，而 F1 值是指查全率和查准率的平均值。来自 Twitter 的有序兴趣项列表指的是为 Twitter 用户挖掘的兴趣项列表，并按照其频次按降序排列。通过表 4.3 可以发现，用户 Melgallant 的真实兴趣项都出现在他的推文中，所以他的兴趣项查全率是 100％。但是他的推文中出现的大量兴趣项并不是他的真正兴趣项，所以他的兴趣项查准率是 3.91％。此外，把已知 LinkedIn 账户的 930 名 Twitter 用户作为实验样本。以相同方式处理后，查全率高达 95％以上的用户有 313 个，占 33.66％，召回率达到 50％以上的用户有 672 个，占 72.25％。具体数据记录在表 4.4 中。

表 4.4 所有实验样本的查全率、查准率、F1 值的统计结果

百分比区间	查全率		查准率		F1 值
	用户数量	百分比	用户数量	百分比	百分比
[100%-95%]	313	33.66%	1	0.11%	0.00%
(95%-90%]	5	0.54%	0	0.00%	0.00%
(90%-85%]	19	2.04%	0	0.00%	0.00%
(85%-80%]	55	5.91%	0	0.00%	0.00%
(80%-75%]	60	6.45%	0	0.00%	0.00%

(75%-70%]	8	0.86%	0	0.00%	0.00%
(70%-65%]	68	7.31%	1	0.11%	0.00%
(65%-60%]	36	3.87%	0	0.00%	0.00%
(60%-55%]	8	0.86%	0	0.00%	0.11%
(55%-50%]	100	10.75%	4	0.43%	0.00%
(50%-45%]	2	0.22%	0	0.00%	0.00%
(45%-40%]	19	2.04%	2	0.22%	0.54%
(40%-35%]	6	0.65%	0	0.00%	0.22%
(35%-30%]	48	5.16%	0	0.00%	0.32%
(30%-25%]	35	3.76%	4	0.43%	2.04%
(25%-20%]	25	2.69%	8	0.86%	4.19%
(20%-15%]	11	1.18%	17	1.83%	8.60%
(15%-10%]	14	1.51%	82	8.82%	20.97%
(10%-5%]	0	0.00%	280	30.11%	35.27%
(5%-0%]	98	10.54%	531	57.10%	27.74%

通过表 4.4 可以得出，绝大多数用户的查全率很高。这意味着绝大多数用户的大部分的真实兴趣爱好都出现在他们自己的推文中。因此，可以相信假设 1 是正确的，也就是说，可以表达 Twitter 用户兴趣爱好的词语通常出现在他的推文中。然后进一步分析这些数据，发现对于实验样本，有 17.42％的用户，他们出现频次第一的兴趣项就是其真实的兴趣爱好。有 15.32％的用户，他们出现频次第二的兴趣项就是其真实的兴趣爱好，以此类推。数据的部分如表 4.5 所示。表 4.5 中的用户数量指的是至少可以从他们自己的推文中挖掘出兴趣项的数量。例如，在表 4.5 中的第十行中，用户数量 871 指的是在实验样本中有 871 个用户，至少能从他们的推文中挖掘出 10 个兴趣项。并且命中数 60 指的是在 871 个 Twitter 用户中有 60 个用户，他们的第十个兴趣项是他们真实的兴趣爱好。

表 4.5 前 20 高频兴趣项的命中率

序号	用户数量	命中数量	命中率
1	930	162	17.42%
2	927	142	15.32%
3	921	129	14.00%
4	915	120	13.11%
5	911	103	11.30%
6	905	85	9.40%
7	902	83	9.20%
8	895	76	8.50%
9	884	79	8.90%
10	871	60	6.90%
11	867	67	7.70%
12	857	74	8.60%
13	850	54	6.40%
14	842	51	6.00%

15	836	64	7.70%
16	829	44	5.30%
17	815	57	7.00%
18	802	43	5.30%
19	794	44	5.50%
20	786	40	5.10%

从表 4.5 可以发现，在 Twitter 用户的推文中，一个兴趣项出现的频率越高，那么通常该兴趣项是该用户真实兴趣爱好的概率就越大。图 4.2 展示了前 30 个最高出现频次兴趣项与其命中率的趋势变化。

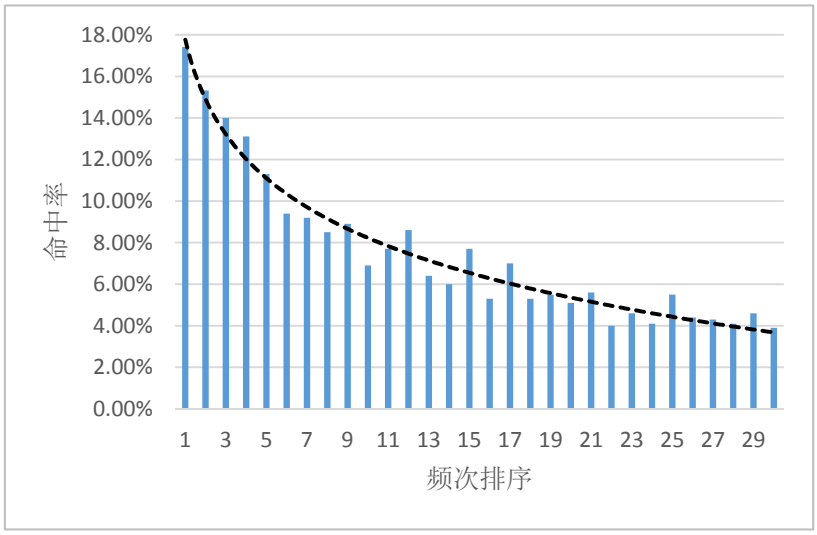

图 4.2 高频兴趣项命中率变化趋势

图 4.2 表明了，就整体而言，命中率整体曲线随着频次的递减逐渐出现下降的趋势，所以频次较高的兴趣项通常更可能是用户的真正兴趣。所以可以相信，假设 2 是正确的，也就是说，Twitter 用户的推文中出现频率越高的兴趣项，就越有可能是他的真实兴趣爱好。此外，在不失一般性的情况下，可以进一步认为，对于 Twitter 用户来说，他的推文中最高频率的兴趣项有 17.42％的可能性是他的真正兴趣，第二高的可能性是 15.32％ ,以此类推。

4.3 基于关联规则的兴趣爱好挖掘

4.3.1 兴趣爱好关联分析

根据 2.4 节关联分析的介绍可知，在关联规则的挖掘过程中，有必要设置最小置信度阈值和最小支持度阈值。满足阈值的关联规则是一个强有力的关联规则。Apriori 是经典的挖掘强关联规则的算法之一。基于 4.1 节收集到的 LinkedIn 用户档案中的真实兴趣项数据，将 Apriori 算法应用于该数据集，进行强关联规则挖掘。根据不同的最小阈值所挖掘的强关联规则的数量如表 4.6 所示。

表 4.6 基于不同的最小阈值所挖掘出关联规则数量

最小置信度阈值	最小支持度阈值									
	0.2%	0.4%	0.6%	0.8%	1%	1.2%	1.4%	1.6%	1.8%	2.0%
10%	1751	588	309	195	127	86	67	52	38	30
20%	857	286	133	93	66	43	37	28	19	15
30%	421	124	58	35	21	14	13	10	8	5
40%	180	42	17	8	4	2	2	1	1	1
50%	86	18	4	2	1	1	1	0	0	0
60%	28	7	1	0	0	0	0	0	0	0
70%	12	3	1	0	0	0	0	0	0	0
80%	3	1	0	0	0	0	0	0	0	0
90%	1	1	0	0	0	0	0	0	0	0

通过表 4.6 可以得出，根据不同的最小置信度阈值和最小支持度阈值，可以挖掘出一定数量的兴趣关联规则。因此，针对预期应用的具体要求，可以通过设置不同的最小阈值来获得一组强关联规则。另外，表 4.7 列出了一些被挖掘出的关联规则。

表 4.7 挖掘出的关联规则样例

编号	前件	后件	置信度	提升度	支持度	期望值
1	friends	family	59.63%	983.49%	1.50%	6.06%
2	culture	travel	48.33%	232.77%	1.16%	20.76%
3	food	travel	46.67%	224.78%	1.13%	20.76%
4	marketing	media	32.55%	592.02%	1.60%	5.50%
5	read; music	movie	31.05%	343.53%	0.99%	9.04%
6	read; photography	travel	53.24%	256.43%	0.85%	20.76%
7	read; cooking	travel	48.18%	232.05%	0.76%	20.76%
8	read; movie	music	39.09%	252.88%	0.99%	15.46%
9	sport; music	travel	38.01%	183.09%	0.75%	20.76%
10	read; movie	travel	35.00%	168.59%	0.89%	20.76%
11	movie; travel	music	34.47%	222.98%	0.93%	15.46%
12	cooking; marathon	travel; running	76.92%	3475.49%	0.12%	2.21%
13	planning; marathon	Travel; running	76.92%	3475.49%	0.12%	2.21%
14	photography; marathon	running; travel	76.92%	3475.49%	0.12%	2.21%
15	communication; sales	marketing; business	75.00%	5333.14%	0.10%	1.41%

通过表 4.7 可以得出，人类兴趣爱好之间存在很强的相关性。例如，关联规则“culture⇒travel”，置信度高达 48.33％，支持度高达 1.16％，提升度高达 232.77％。这表明，在人类兴趣爱好上，“culture”和“travel”是高度相关的。另一些例子，例如关联规则“read; photography⇒travel”，置信度达到 53.24％，提升度达到 256.43%。“communication; sales⇒marketing; business”，置信度达到 75.00%，提升度高达 5333.14%。

因此，通过实证相关分析发现，人类兴趣爱好之间存在着很多的关联关系，一些关联规则具有很高的置信度，提升度和支持度。这表明人类兴趣爱好之间有

一些固有的内在联系。所以它们可以应用于社交网络用户的兴趣爱好挖掘。

4.3.2 挖掘方法

尽管已经证实了 4.2.1 节的假设，即可以表达 Twitter 用户兴趣爱好的词语可能出现在他的推文中，并且用户的推文中兴趣项的出现频率越高，则更有可能成为他的真实兴趣爱好。但是依然无法直接从他的推文中区分出该用户真实的兴趣爱好，因为通常可以从 Twitter 用户的推文中挖掘出大量的兴趣项。此外，从表 4.4 也可以看到，基于词频的兴趣挖掘方法其查准率非常低。

实际上，根据兴趣爱好关联规则的性质，如果用户有一个兴趣爱好，他也可能有一个与此兴趣爱好存在关联关系的兴趣爱好。因此，该方法运用从 LinkedIn 中分析出的兴趣爱好关联规则对从 Twitter 中挖掘出的每个用户的兴趣项列表进行重新排序，以使他们的真实兴趣项尽可能地出现在来自 Twitter 的有序兴趣项列表的前面。所以可以把最初的几个兴趣项作为用户真实的兴趣项提取出来，因为根据假设 2，它们很可能是用户真实的兴趣项。

在不失一般性的情况下，该方法把兴趣项的出现频次视为权重。对于 Twitter 用户的有序兴趣项列表（例如表 4.3 所示），可以根据 4.3.1 节挖掘出的兴趣爱好关联规则改变它们的权重，然后根据他们的权重按降序重新排序。因此，经过综合考虑，设计了以下的方法来将兴趣爱好关联规则应用于 Twitter 用户的兴趣爱好挖掘，其步骤如下。

步骤 1：给定一个 Twitter 用户，从 Twitter 中收集他的所有推文。

步骤 2：根据 4.1.2 节整理出的 210 个高频兴趣项，挖掘该用户所有推文中涉及到的兴趣项和出现频次。

步骤 3：根据兴趣项的出现频次，按降序排序兴趣项，作为该用户的原始有序兴趣项列表，记为 *ittsList*。

步骤 4：将原始有序兴趣项列表中的所有兴趣项提出，作为该用户的原始兴趣项集合，记为 *ittsSet*。

步骤 5：根据 4.3.1 节中挖掘出的兴趣爱好关联规则，选择其中一组关联规则作为关联规则集合 *ruleSet*。

步骤 6：从 *ittsSet* 中逐个取出每个兴趣项，称为 *irt*。如果在 *ruleSet* 中存在这样一条关联规则，其前件兴趣项是 *irt*，且该规则的后件兴趣项也存在于 *ittsSet* 中，则增加该后件兴趣项在 *ittsList* 的权重 *W*。

步骤 7：将 *ittsSet* 中每个兴趣项处理之后，按照新的权重大小重新按降序排序这些兴趣项，得到结果列表 *rsltList*。

步骤 8：根据实际需要，从 *rsltList* 中取出前 n 个权重较高的兴趣项作为对用户进行兴趣爱好挖掘的结果。

基于关联规则的兴趣爱好挖掘方法流程图如图 4.3 所示。在该方法中，权重 W 被设置为 $w+k\times r$，其中参数 w 为初始权重，即该兴趣项的出现频次。在该公式中，参数 k 是设定关联规则对于兴趣爱好挖掘的影响能力的常数。k 的值越大，关联规则对兴趣爱好挖掘的影响就越大。此处之外，参数 r 指的是兴趣项 irt 是用户真实的兴趣项的概率，其值对应于表 4.5 中的命中率。该参数确保了如果某个兴趣项是用户真实兴趣爱好的概率越大，则其引入的关联规则的作用也越大。

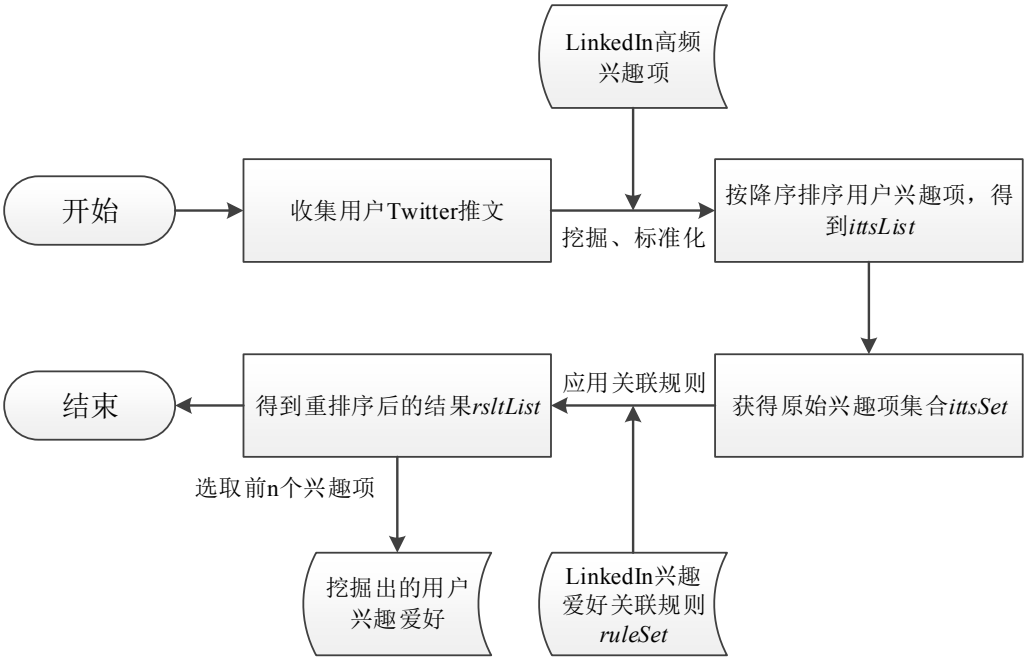

图 4.3 基于关联规则的兴趣爱好挖掘方法流程图

社交网络用户的兴趣爱好并非一成不变，用户的兴趣爱好可能随着时间的推移发生变化。该方法中存储兴趣项的集合底层采用哈希表进行数据存储。若需要关注某一个用户兴趣爱好变化的过程，只要将新发现的兴趣项加入原始兴趣项集合，并再次运用关联规则进行重排序即可实现增量挖掘兴趣爱好。

4.4 实验结果与分析

4.4.1 实验平台

实验环境采用实验室 MongoDB 集群，由 4 台相同配置的普通 PC 机搭建而成。其中 3 台为 MongoDB 的分片服务器，1 台为 MongoDB 的路由服务器。具体软硬件配置如表 4.8 所示。

表 4.8 实验环境软硬件配置

硬件	CPU	Intel i5 双核 3.2GHz
	硬盘	500GB SATA 机械硬盘
	网络	100Mbps 局域网
	内存	8GB DDR3
软件	操作系统	Windows 7

	数据库	MongoDB 3.0.6
	开发语言	Python 2.7.11
	集成环境	PyCharm

4.4.2 实验设置

为了验证关联规则对于挖掘兴趣爱好的影响，从 4.3.1 节中挖掘的兴趣爱好关联规则中选取了一组强关联规则进行兴趣爱好的挖掘实验。最后，通过比较和分析实验结果的命中率、查全率、查准率和 F1 值，综合评价关联规则对于挖掘兴趣爱好的影响程度。

将最小置信度阈值和最小支持度阈值都设置为较大的值，这样虽然生成的关联规则数量较少，但生成的关联关系都较强，称为强关联规则。如表 4.6 所示，如果最小支持度和置信度阈值分别设为 0.4%和 20%，则根据在 4.3.1 节中讨论的实验数据，可以挖掘出 286 个关联规则。此外，对于该实验，应用 4.3.2 节所提出挖掘方法来处理实验样本，实验样本即在第 4.4.2 节中讨论的 930 个社交网络用户。

不失一般性，在考查命中率的实验中，分别设定挖掘方法中的参数 k 为 0,7,14,21,38,35 和 42，并进行 7 组实验。实际上，当参数 k 被设置为 0 时，关联规则不起作用，方法只是返回原始基于词频的有序兴趣列表，其中兴趣项的顺序是基于它们在用户推文中的出现频次。在考查查全率、查准率和 F1 值的实验中，将参数 k 的值设置为 0，7 和 42，其分别表示了不使用关联强度进行基于词频的挖掘，使用关联规则但弱化关联规则影响力的挖掘和使用关联规则但强化关联规则影响力的挖掘。该实验将从三个维度，即查全率，查准率，F1 值讨论其实验结果。

4.4.3 数据集

本章的实验数据与 4.4.2 节使用的数据一致。首先在 LinkedIn 随机抽取了 930 名用户，这些用户在 LinkedIn 的个人档案都准确地填写了兴趣爱好列表以及 Twitter 个人主页地址。兴趣爱好列表是由用户亲自填写，所以经过 4.1.2 节的标准化处理后，可以得到每个用户的真实兴趣项集合，该集合在实验中作为验证集，验证不同挖掘方法的挖掘效果。除此之外，通过 LinkedIn 用户所填写的其 Twitter 个人主页地址，能够快速定位到其个人 Twitter 并使用爬虫程序抓取该用户的所有推文，供给兴趣爱好挖掘使用。

综上所述，数据集为社交网络 LinkedIn 中随机的 930 名用户的真实兴趣爱好以及其在 Twitter 上的所有推文。

4.4.4 评价指标

基于关联规则的兴趣爱好挖掘实验使用了如表 4.5 中所描述的命中率，即将所有实验用户通过 4.3.2 节的方法对其原始的，基于词频的用户兴趣项列表进行重排序后，得到的结果列表，结果列表中第 n 项是对应用户的真实兴趣爱好的数量与至少含有 n 个兴趣项用户数量的比值。除此之外，实验中也使用了查准率、查全率和 F1 值作为模型的评估指标，它们的定义分别如下所示：

$$查全率 = \frac{TP}{TP + FP} \times 100\% \tag{4.1}$$

$$查准率 = \frac{TP}{TP + FN} \times 100\% \tag{4.2}$$

$$F1值 = \frac{2 \times 查全率 \times 查准率}{查全率 + 查准率} \times 100\% \tag{4.3}$$

其中，*TP* 表示挖掘出的兴趣项是用户真实兴趣项的数量，*FN* 表示挖掘出的兴趣项不是用户真实兴趣项的数量，*FP* 表示未挖掘出的用户真实兴趣项的数量。

4.4.5 实验分析

完成 7 组命中率实验之后。对于每组实验的结果，计算用户的有序兴趣项列表在重排序之后第 n 个兴趣项是用户的真实兴趣爱好的概率，即第 n 个兴趣项的命中率。例如，对于每个用户的有序兴趣列表 *rsltList* 的第一个兴趣项，当参数 *k* 分别设置为 0,7,14,21,28,35 和 42 时，对应的命中率分别是 17.42%，19.25%，21.61%，21.72% 23.23%，23.44%和 23.76%。而对于用户的第二个兴趣项，相应的比例分别为 15.32%，16.40%，16.18%，17.69%，17.04%，17.26%和 16.94%。图 4.4 直观地展示了 7 组实验中前 10 个兴趣项的命中率。

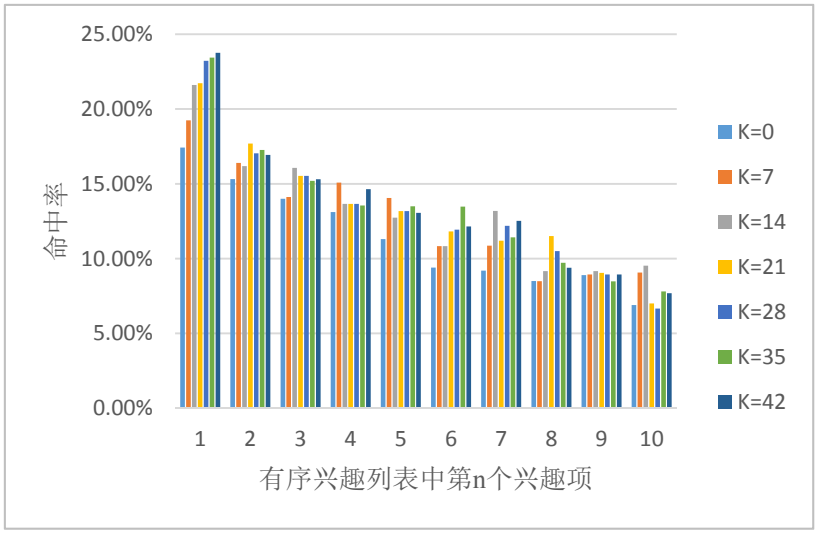

图 4.4 基于不同 k 值的 rsltList 前 10 兴趣项的命中率

通过图 4.4 可以得出，一旦关联规则起作用，即参数 *k* 没有被设置为 0 时，则前 10 个兴趣项的命中率均有一定的增加。在某些情况下，关联规则的效果非

常明显。例如，对于用户的第一个兴趣项，当参数 k 设置为 42 时命中率达到了 23.76%，而当参数 k 设置为 0 时，命中率却只有 17.42%。另外一个例子，对于用户的第二个兴趣项，当参数 k 被设置为 21 时命中率是 17.69%，高于当参数 k 被设置为 0 时的 15.32%。这意味着兴趣爱好关联规则的应用大大提高了该方法挖掘用户列表 *rsltList* 中的前几个兴趣项是他真实的兴趣爱好的可能性。这表明兴趣爱好关联规则在该方法中起到了显著作用。

对于每组实验的结果，给定一个用户，首先可以得到他的有序兴趣列表中的前 10 个兴趣项并计算他的查全率，然后计算其查全率大于某一给定值时的用户数量占比。当参数 k 分别为 0，7 和 42 时，查全率大于 70%的用户比例分别为 6.77%, 7.96%和 8.17%。查全率大于 30%的用户比例分别为 40.22%, 45.59%和 48.49%。图 4.5 直观地展示了当参数 k 分别为 0，7 和 42 时的查全率比例图。

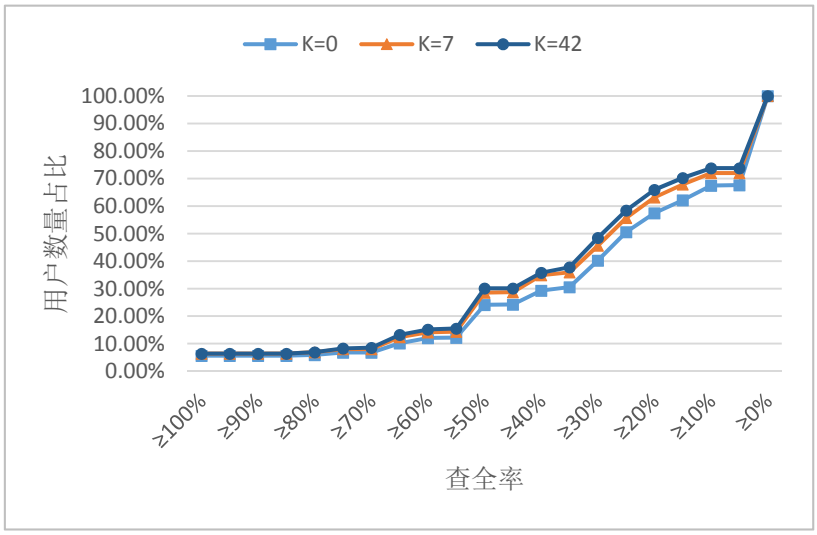

图 4.5 基于不同 k 值的兴趣爱好挖掘的查全率比例

通过图 4.5 可以得出，如果参数 k 的值为 0，则对应的曲线最低，即挖掘效果最差。这意味着在相同的条件下，查全率是最低的。在实验中，如果参数 k 的值不被设置为 0，其相应的曲线都比将参数 k 设置为 0 时高。这里可以看到关联规则明显提高了不同权重下的查全率。他们对兴趣爱好挖掘有较好的提升作用。除此之外，通过该实验中可以得出关联规则的权重设置越大，其挖掘效果往往越好。

同理地，对于每组实验的结果，给定一个用户，同样可以得到他的有序兴趣项列表中的前 10 个兴趣项的查准率和 F1 值，然后计算其查准率和 F1 值大于某一给定值时的用户数量占比。图 4.6 和图 4.7 分别展示了在参数 k 值取 0，7 和 42 时查准率和 F1 值的比例图。

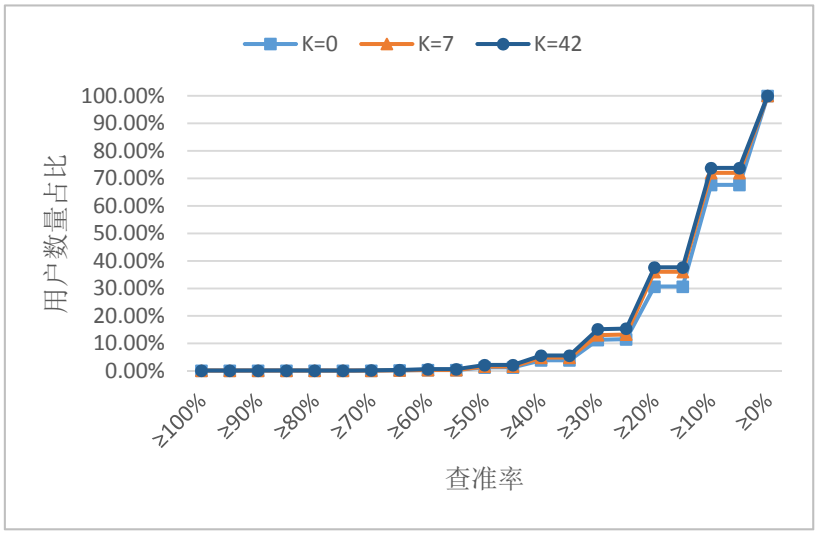

图 4.6 基于不同 k 值的兴趣爱好挖掘的查准率比例

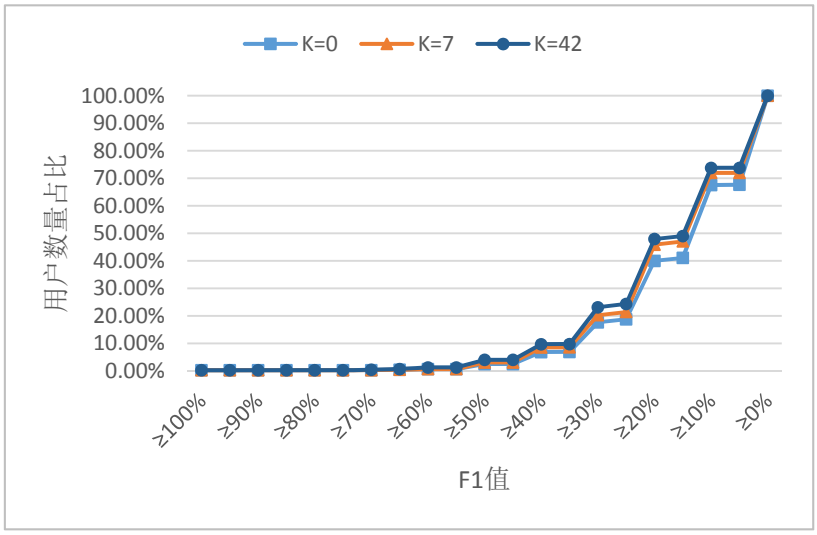

图 4.7 基于不同 k 值的兴趣爱好挖掘的 F1 值比例

通过图 4.6 和图 4.7 可以得出，参数 k 的值为 0 时，其对应的曲线不如其他曲线。而对应于参数 k 的值为 42 的曲线则是最好的。这也意味着关联规则对于兴趣爱好挖掘是有提升效果的。

通过该实验及其结果，发现兴趣爱好关联规则对于基于词频的兴趣项挖掘确实起到了提升的作用。事实上，该结论应该是合理的。由于关联规则反映的是自然事物之间的关系，从兴趣爱好的角度来讲，某人有某个兴趣爱好，在一定程度上，这意味着他应该也有与其相关的其他兴趣爱好。

该实验的结果也可以验证：基于社交网络数据所挖掘出的兴趣爱好关联规则是可靠的。因为它们对于基于词频的兴趣爱好挖掘是有提升的。当参数 k 被设置为不同的值时，查全率、查准率和 F1 值是不同的。而且，参数 k 的值设置越大，关联规则的效果就越好。但参数 k 的值从 0 开始增大时，相应的挖掘效果区别还是非常明显的，但当参数 k 的值设置达到一定值后，例如分别设置为 28，35 和

42 时，相应的挖掘效果则已经非常接近了。

4.5 本章小结

兴趣爱好是心理学和教育学的一个重要概念，在许多领域得到广泛的研究。但是，现有的研究很少探究社交网络上兴趣爱好的分布特征以及其内在的关联关系。为了解决此问题，本章首先采集了社交网络用户的兴趣爱好，并进行了建模，挖掘出 210 种高频兴趣项。然后分析社交网络 Twitter 上用户兴趣爱好的分布特征。之后，根据 LinkedIn 的真实用户数据，挖掘了兴趣爱好之间的关联规则。最后，基于 LinkedIn 上的兴趣关联规则和 Twitter 上的用户兴趣爱好分布情况，设计了一个方法来挖掘 Twitter 用户的兴趣爱好，并进行了实验来系统地展示该挖掘方法的有效性。根据本章研究，可以发现在人的兴趣爱好之间存在大量的关联规则，这些规则可以提升基于词频的兴趣爱好挖掘方法的挖掘效果。本章研究工作不仅为社交网络上的兴趣爱好挖掘提供了一个新的思路，而且揭示了兴趣爱好内在关联关系以及应用价值。该研究工作具有一定的理论和实践价值。

第五章 基于组织成员与兴趣爱好的应用

社交网络用户的人物属性包含了基本属性、行为属性以及情感属性等内容。情感属性又包含了用户的性格属性，例如，冒险精神、内向程度、决断力等。现有的用户情感属性挖掘方法主要通过对用户发布的自定义文本内容进行语义分析从而获得用户的情感属性。结合兴趣爱好进行用户的情感属性挖掘是一种新的分析思路。挖掘组织成员的人物属性对于企业的人事管理工作，员工的个人发展规划都很有帮助。本章基于海量社交网络数据开发了一套人物属性识别系统，该系统可以对社交网络 Twitter 中的数据进行深度挖掘和统计分析，准确地获取出组织机构的成员在 Twitter 中的注册账号，分析其基本属性、行为属性和情感属性。但在系统开发过程中，如何从错综复杂的社交网络人际关系网中对某组织机构的成员进行识别以及如何准确地挖掘这些成员的兴趣爱好是该系统的关键环节，也是难点所在。

5.1 基于海量社交网络数据的人物属性识别系统

5.1.1 项目背景

社会化的网络应用服务被称为社交网络服务，是 Web 2.0 下的典型应用。它通过 SNS（Social Network Service）平台为用户提供一个以关系连接和信息产生、分享为主的社交服务[65]。类似于现实世界中的人际关系网络，社交网络平台中也存在好友关系。在此平台上的用户可以通过用户关系传递个性化的信息内容。社交网络平台上的这种用户关系大多数都由现实世界中的好友关系或自然人之间相同的兴趣爱好所产生，并随着人际关系圈不断的延伸和展开。

社交网络平台上的用户网络，实际上是基于真实世界用户关系和用户网络的映射，在此平台上的信息传播是真实世界人际关系网络信息传播的反映。社交网络的应用非常广泛，积累了大量的数据，这些数据通常能够反映一个人的兴趣、爱好、政治立场、习性及其基本的个人属性。典型的社交网络应用有 Twitter、Facebook、LinkedIn 等，它们在国外非常流行。比如，2015 年 Twitter 公布的注册用户就已经超过 10.17 亿，当月活跃用户 2.88 亿，普及率非常高。

因此可以基于海量社交网络数据的分析，挖掘个人属性，在现实生活中对社交网络用户进行定位；挖掘其行为属性，包括发帖、转发、评论关注的时间和频率等，得出用户在现实生活中的作息规律、行为轨迹并进一步构成用户的行为特征；对用户的网络言论进行潜在语义分析和计算，分析其人格特征、价值观取向、

自我认知状态、社会需求以及政治倾向等，这些信息带有强烈的个人色彩，反映了用户的情感特征和内在的心理状态[66]。比如，在选举期间，各候选人可以对社交网络上选民的留言评论进行深入地挖掘研究，从而针对态度摇摆不定的选民实施特别的政策，提高胜选概率。

5.1.2 系统整体介绍

基于海量社交网络数据的人物属性识别系统总体由七部分组成，分别是爬虫子系统、身份识别子系统、人物属性判别子系统、分析和预测子系统、No-SQL型语义数据库管理子系统、有效性验证子系统和多维结果可视化子系统，系统总体架构如图 5.1 所示。

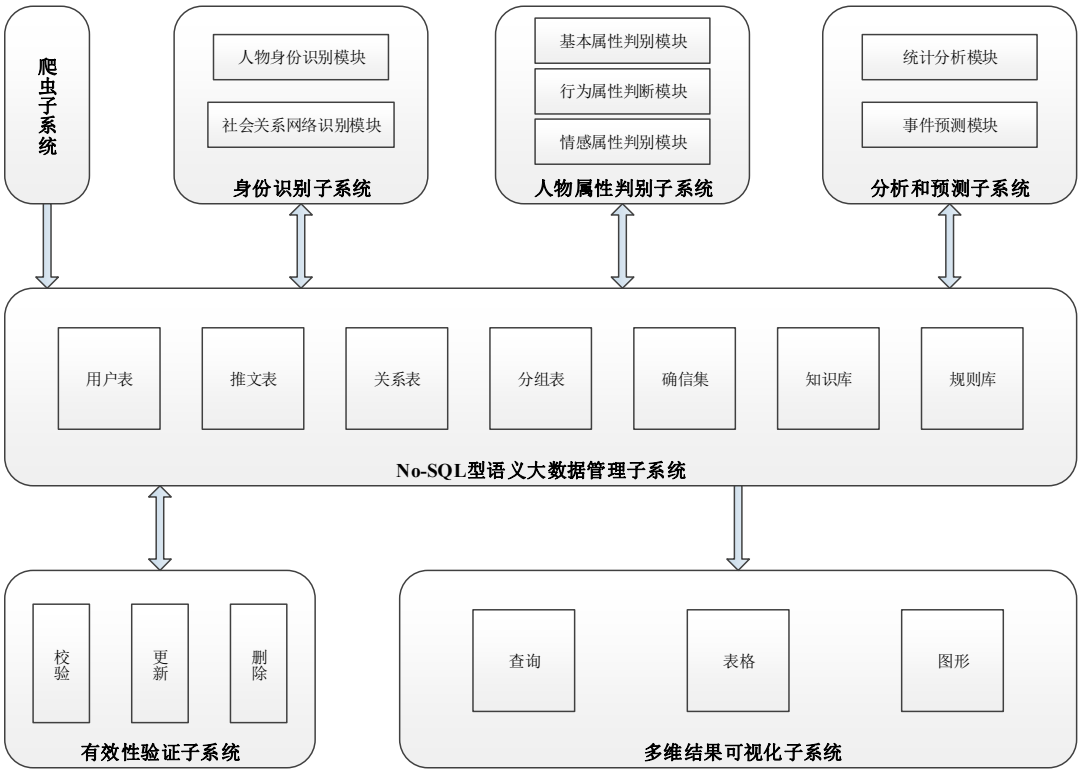

图 5.1 系统总体架构

（1）爬虫子系统：获取 Twitter 上关注某组织机构的所有粉丝（第一层核心用户）及关注这些粉丝的粉丝（第二层核心用户）的所有静态信息和动态信息，生成用户表、推文表、关系表等。

（2）身份识别子系统：定向地搜索出某组织机构工作人员在 Twitter 上注册的用户账号，并部分地重现其在现实世界中的主要社会关系，形成社会关系层次图。

（3）人物属性判别子系统：组织机构工作人员的人物属性可以分为三类：

1）基本属性：包括其真实姓名、性别、职位、居住地、个人主页等。

2）行为属性：由用户在 Twitter 上的行为，包括发帖、转发、评论关注事件

和频率等，推测其在现实生活中的作息规律、行为轨迹等行为属性。

3）情感属性：对用户生成和转发的推文进行深度语义分析，对他政治倾向、人格特征、价值观取向以及社会需求等情感属性进行推测。

（4）分析和预测子系统：统计分析模块根据已有的历史记录，对工作人员整体按时间、性别、机构进行多层次多粒度的信息统计和分析；事件预测模块根据上下文环境，对工作人员在未来某热点事件中可能具有的行为做出预测。

（5） No-SQL 语义大数据管理子系统：对从 Twitter 获取的数亿条海量元数据、推文、各识别和分析子系统产生的中间结果、知识库、用于验证和推理的启发式规则库进行统一管理。

（6）有效性验证子系统：采用半自动化的方式，利用启发式规则和约束，对人物属性识别的结果进行真实性检验，以保证结果完全可靠。

（7）多维结果可视化子系统：对输出结果以查询、表格和图形等多种形式进行展现和人机交互，使得本系统的使用更为高效便捷。

其中爬虫子系统、身份识别子系统、人物属性判别子系统是整个系统的基础组成部分。身份识别子系统与人物属性判别子系统又可以归纳为三个部分，目标成员分析、社会特征分析与个人特征分析。该部分的系统工作原理如图 5.2 所示。

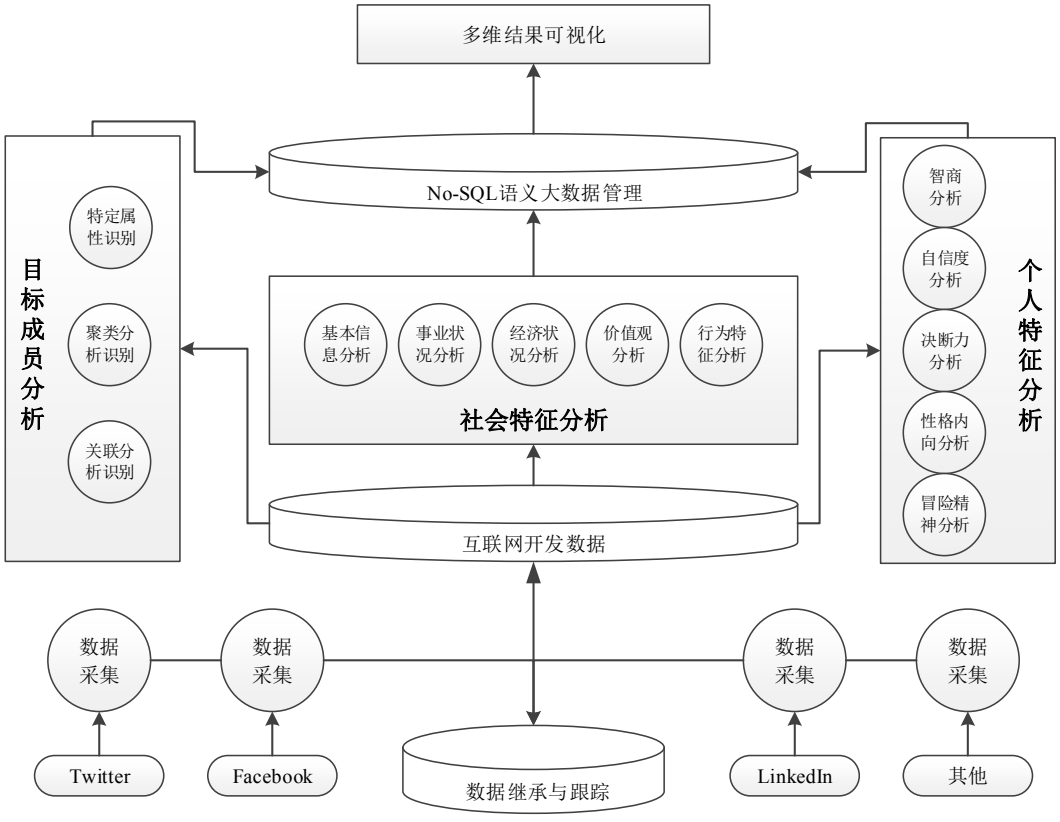

图 5.2 部分的系统工作原理

目标成员分析即对特定的组织机构进行成员识别，可采用的方法有基于特定属性的识别、基于聚类分析的识别与基于关联分析的识别。社会特征分析包括了

基本信息分析、事业状况分析、经济状况分析、价值观分析、行为特征分析。个人特征分析包括了智商分析、自信度分析、决断力分析、性格内向分析、冒险精神分析。这三大部分是爬虫子系统所采集到的互联网数据的主要输入入口，经过一系列的处理分析后，分析结果将输出到多维结果可视化子系统进行展示。

5.1.3 系统难点

在基于海量社交网络数据的人物属性识别系统中，特定组织机构的组织成员是系统的基本研究对象，系统是以组织机构为基本单元划分组织成员的，即系统将根据用户需求，从社交网络 Twitter 中对特定组织机构的成员进行识别，将识别出的用户集合作为该组织机构的组织成员，然后对组织机构中的个体进行人物属性分析。系统的首页也是根据不同的组织机构划分模块入口，如图 5.3 所示。

图 5.3 基于海量社交网络数据的人物属性识别系统模块入口

如何从用户数量巨大，人际关系网错综复杂的 Twitter 中尽可能多地，且准确地识别出特定组织机构的组织成员是系统开发过程中面对的一个难点。

除此之外，在对组织成员的人物属性分析过程中，个人特征分析是系统的主要功能之一。由 5.1.2 节中可知个人特征分析包括了智商分析、自信度分析、决断力分析、性格内向分析、冒险精神分析。这些个人特征的分析可以通过直接挖掘与个人特征紧密相关的推文内容，但这类型的推文数量比较稀少，不足以支撑对个人特征的分析。因此系统设计了一种基于兴趣爱好的个人特征分析方法，即先通过挖掘组织成员的兴趣爱好，再结合心理学知识，对各项兴趣爱好在不同个人特征方面的表征程度进行综合评价，最后分析组织成员的个人特征。因此如何准确地挖掘出组织成员的兴趣爱好成为个人特征分析的关键环节，也是系统开发过程中面对的另一个难点。

5.2 成果应用

5.2.1 组织成员识别的应用

由 5.1.3 节可知，组织成员识别是基于海量社交网络数据的人物属性识别系统中一个难点。系统采用了第三章中所提出的社交网站组织成员方法作为解决方案，即首选使用人工筛选的方式找出一些种子用户（已经确定为特征组织机构的组织成员），然后通过爬虫子系统采集特定组织机构在 Twitter 上的公共账号的粉丝用户（作为候选用户）以及他们的所有推文，经过 MongoDB 的 MapReduce 脚本处理后生成用户表、关注表、推送表并存入 No-SQL 型语义数据库管理子系统。数据结果如图 5.4 至 5.6 所示。

_id	created_time	user_id	screen_name	following	nick_name	site	lists	tweets	join_date	followers	location	favori...	desc
595a3eec8a4f7646c85a6f23	2017/7/1 12:11:...	20148882	Seftonbabe	6	Faye	[NULL]	0	2	上午5:47 - 2009年2月5日	5		1	[NULL]
595a3eec8a4f7646c85a6f24	2017/7/1 12:11:...	16231640	TomMillerUK	5309	Cllr Tom Miller	http://tommill...	10	42800	下午6:23 - 2008年9月10日	4851	WillesdenGreen	2073	cllr for willesden green | cab...
595a3eec8a4f7646c85a6f26	2017/7/1 12:11:...	18537981	johnhanna	146	John Hanna	[NULL]	0	951	下午2:28 - 2009年1月1日	127		255	[NULL]
595a3eec8a4f7646c85a6f27	2017/7/1 12:11:...	19766969	cottrells	494	Steve Cottrell	http://about....	22	4183	上午5:28 - 2009年1月30日	326	Bournemouth,UK	127	Easy going bean counter, sta...
595a3eec8a4f7646c85a6f28	2017/7/1 12:11:...	20674976	tinkythree	51	Mike Taylor	[NULL]	0	332	上午5:01 - 2009年2月12日	29	London	9	eCommunications, cricket a...
595a3eec8a4f7646c85a6f29	2017/7/1 12:11:...	83815278	ABrassett	846	A Brassett LTD	http://www.ab...	0	12100	上午3:26 - 2009年10月20日	2010	SEENGLAND	2	A.Brassett - Appliances... Plu...
595a3eec8a4f7646c85a6f2a	2017/7/1 12:11:...	16907069	derekdraper	2334	CDP Consultants	http://www.cd...	4	3443	上午8:28 - 2008年10月22日	5476		43	I am the founder & MD of C...
595a3eec8a4f7646c85a6f2b	2017/7/1 12:11:...	15650774	UK_Law_Tutor	1963	Brian Risman	http://www.th...	6	435	下午1:14 - 2008年7月29日	1609	LondonUK	3	Law Tutor, Publisher The La...
595a3eec8a4f7646c85a6f2c	2017/7/1 12:11:...	20770886	dmontfort	4054	PETER FLEMING	http://www.se...	12	8155	上午5:48 - 2009年2月13日	3220		4315	Leader multi award winning ...
595a3eec8a4f7646c85a6f2d	2017/7/1 12:11:...	20985544	GarethU	30	Gareth Underwood	[NULL]	0	49	上午5:43 - 2009年2月16日	10		0	[NULL]
595a3eec8a4f7646c85a6f2e	2017/7/1 12:11:...	16925469	Bendoo133	238	Ben Wilson	[NULL]	0	89	上午5:31 - 2008年10月23日	100	London	0	Payments, financial services ...
595a3eec8a4f7646c85a6f2f	2017/7/1 12:11:...	7169162	libbymiller	2197	Libby Miller	http://planb.ni...	2	25099	上午3:05 - 2007年6月30日	2089	MostlyBristol	4960	Foaf / Beeb / Radiodan / Ha...
595a3eec8a4f7646c85a6f30	2017/7/1 12:11:...	17688089	joolians	446	j	[NULL]	0	5515	下午12:33 - 2008年11月27日	458		6525	[NULL]
595a3eec8a4f7646c85a6f32	2017/7/1 12:11:...	18452477	LondonSummit	130	LondonSummit	http://www.lo...	0	258	上午8:16 - 2008年12月29日	343	London	0	A Twitter in support of the t...
595a3eec8a4f7646c85a6f33	2017/7/1 12:11:...	16528767	carlabutler	518	Carla Butler	[NULL]	0	359	上午8:09 - 2008年9月30日	388	Sussex	60	[NULL]
595a3eec8a4f7646c85a6f34	2017/7/1 12:11:...	20432761	alistairford	1860	Alistair Ford	[NULL]	1	84	上午4:53 - 2009年2月9日	438	London	0	Husband, father and therefo...
595a3eec8a4f7646c85a6f35	2017/7/1 12:11:...	19064397	tsingleton	569	Tony Singleton OBE	https://www.li...	9	7658	上午4:37 - 2009年1月16日	2341		1136	Turning ideas into reality. C...
595a3eec8a4f7646c85a6f36	2017/7/1 12:11:...	20596957	IA_Forum	1974	Int'l Affairs Forum	http://www.ia-...	0	10500	上午7:37 - 2009年2月11日	6187		69	International Affairs Forum i...
595a3eec8a4f7646c85a6f37	2017/7/1 12:11:...	10079442	crowsond	917	David Crowson	[NULL]	7	5879	下午1:28 - 2007年11月8日	830	StowHill,Newp...	120	[NULL]
595a3eed8a4f7646c85a6f38	2017/7/1 12:11:...	20299544	TopJobsInLon...	3533	Jobsonica In London	https://twitter....	6	2530...	上午12:30 - 2009年2月7日	3720	Jobsonica.com:...	142	Get all your top jobs in Lond...
595a3eed8a4f7646c85a6f39	2017/7/1 12:11:...	18999424	danielatkinson	354	Daniel Atkinson	[NULL]	12	1803	下午2:32 - 2009年1月14日	381	London,Greater...	1434	Government digital manage...
595a3eed8a4f7646c85a6f3a	2017/7/1 12:11:...	14494377	jen725	1847	jen725	[NULL]	3	80900	上午7:29 - 2008年4月23日	1361	London	1773	toddler725, athletics, centre l...
595a3eed8a4f7646c85a6f3b	2017/7/1 12:11:...	20772941	vickyandre1	395	VickyAndre	[NULL]	0	8872	上午6:20 - 2009年2月13日	88	London	216	Mum, knitter, lawyer. Part ti...
595a3eed8a4f7646c85a6f3c	2017/7/1 12:11:...	12542672	ShaneMcC	1899	Shane McCracken	http://galloma...	8	13200	上午7:11 - 2008年1月22日	1894	Bath,UK	322	Wearer of PINK jacket at con...
595a3eed8a4f7646c85a6f3d	2017/7/1 12:11:...	19066503	oldshep	1304	David Shepherd	[NULL]	33	6568	上午5:41 - 2009年1月16日	4048	Highbury,Lond...	105	Employment law,

图 5.4 部分用户表

_id	value
000dcp	Array[47]
027Sweet	Array[180]
0723411929	Array[13]
101Goldstein	Array[11]
11209964252	Array[15]
121Almuola	Array[7]
Gdbxhwdbv	Array[1]
0102Ch	Array[17]
1951Colin	Array[2]
95390040	["hmtreasury", "QPRFC"]
ADYBABES666	Array[1]
00004Marie	Array[230]
00012859dn	Array[1972]
007Hugh	Array[529]
007Monkey	Array[281]

图 5.5 部分关注表

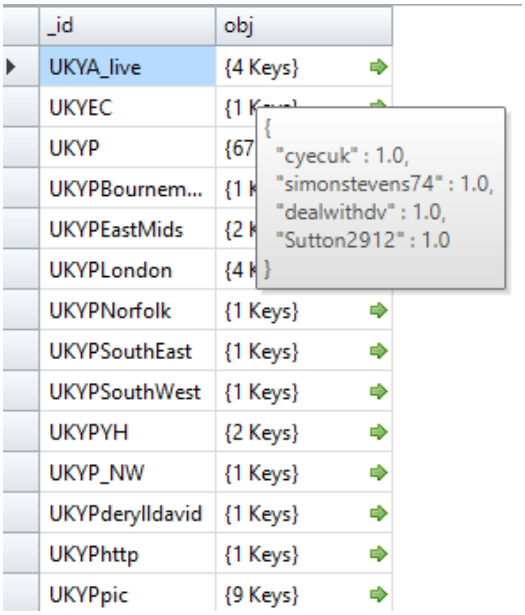

图 5.6 部分推送表

基于组织成员识别方法中每个用户的 6 种识别因子的最优计算模型，可以计算出每个候选用户与种子用户集合之间的亲密程度，根据按照亲密程度从高到低排序，可以筛选出前 500 名候选用户视为第一层组织成员识别的结果，并且将第一层组织成员识别的结果加入到种子用户集合。同理地，进行第二层组织成员的识别，筛选出前 1000 名候选用户视为第二层组织成员识别的结果。最后将这两层的组织成员识别结果作为该方法挖掘出的特定组织机构的组织成员，得到组织成员列表如图 5.7 所示。

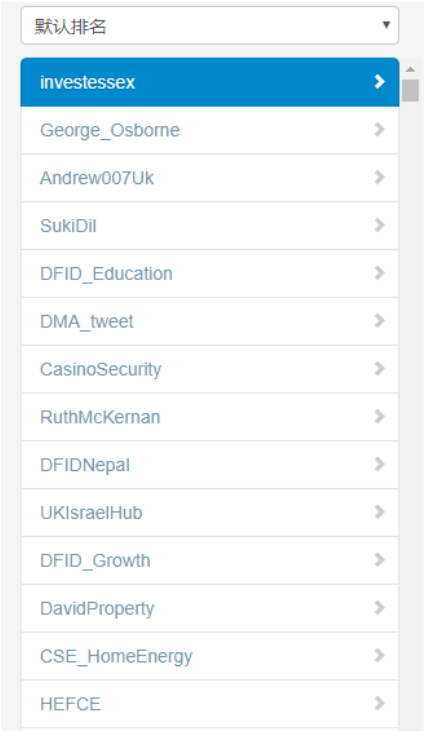

图 5.7 部分组织成员列表

除此之外，将组织成员识别过程中的中间数据结果进行整理后，还可以生成相对应的用户关系表和用户关系图，分别如图 5.8、5.9 所示，并在组织成员的个人信息主页上进行展示，方便用户查看和梳理组织成员在其组织中的人际关系结构，以及通过人工判断其关系的紧密程度决定是否该用户就是特定组织机构的组织成员。

人物关系表

关系人物	关注	被关注	@次数	被@次数
Novak_Trevor	是	否	0	0
UKTINigeria	是	是	0	0
britchamberESP	是	否	0	0
UKTI_Greece	是	否	0	0
UKTradeMinister	是	否	1	0
BritishinFrance	是	否	0	0
ukinportugal	是	否	0	0

图 5.8 某用户的部分用户关系表

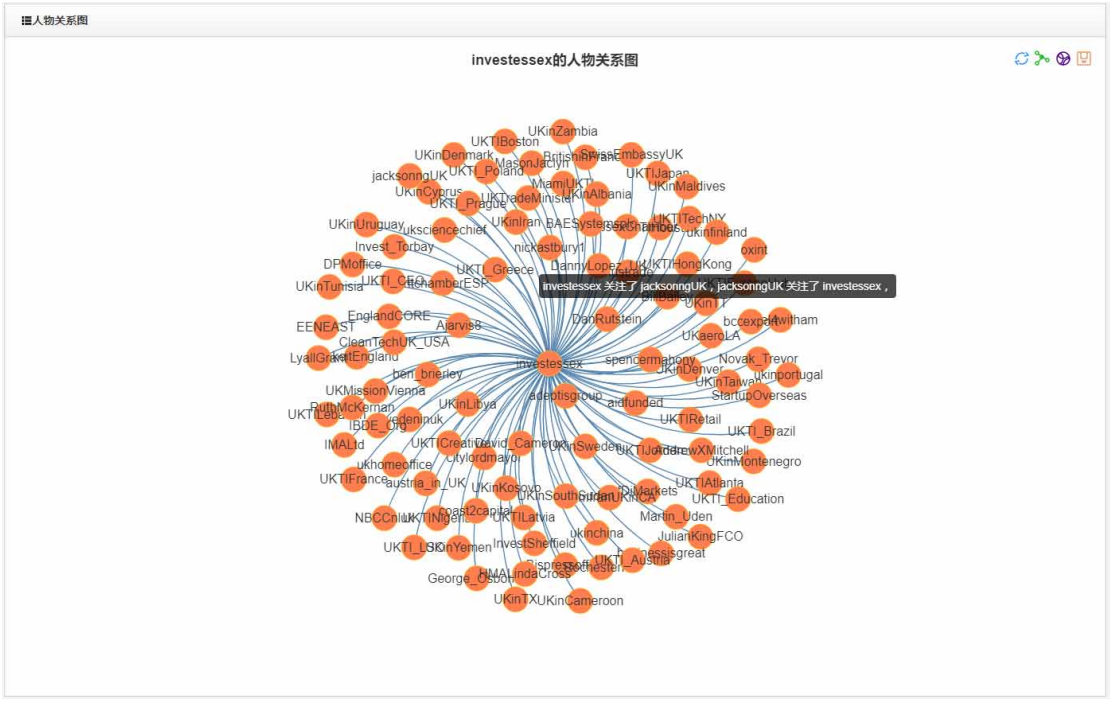

图 5.9 某用户的用户关系图

5.2.2 兴趣爱好挖掘的应用

由 5.1.3 节可知，兴趣爱好挖掘是基于海量社交网络数据的人物属性识别系统中一个难点。系统采用了第四章中所提出的基于关联规则的兴趣爱好挖掘方法作为解决方案。首先通过爬虫子系统，采集组织成员在 Twitter 上所发表的所有推文，然后使用 4.1.2 节中归纳整理的 210 个高频兴趣项，对组织成员在推文中涉及到的兴趣爱好进行挖掘，可以获得基于词频的原始兴趣项列表。之后，使用 4.3.1 节所挖掘出的兴趣爱好关联规则表，如图 5.10 所示。

_id	prefix	lift	imbalance_ratio	conf	conseq	expected_conf	kulc	sup_xy
0	Array[1]	2.258012	0.737952	0.358491	Array[1]	0.158764	0.211124	0.010123
1	Array[1]	4.794372	0.728395	0.37037	Array[1]	0.077251	0.219668	0.005328
2	A	5722	0.83731	0.369565	Array[1]	0.230155	0.204459	0.009057
3	Array[1]	4.438998	0.154321	0.24359	Array[1]	0.054875	0.214028	0.010123
4	Array[2]	1.77494	0.828947	0.266667	Array[1]	0.15024	0.147518	0.004262
5	Array[1]	1.685474	0.466258	0.22449	Array[1]	0.133191	0.156245	0.011721
6	Array[2]	2.539477	0.806723	0.296296	Array[1]	0.116676	0.166413	0.004262
7	Array[2]	2.606954	0.659574	0.222222	Array[1]	0.085242	0.136111	0.004262
8	Array[2]	1.511268	0.914989	0.347826	Array[1]	0.230155	0.183172	0.004262
9	Array[2]	2.861296	0.702128	0.25	Array[1]	0.087373	0.14939	0.004262
10	Array[1]	2.762618	0.069343	0.213415	Array[1]	0.077251	0.227397	0.018647
11	Array[1]	2.762627	0.069343	0.241379	Array[1]	0.087373	0.227397	0.018647
12	Array[1]	1.432829	0.494083	0.2	Array[1]	0.139584	0.13626	0.010123

["literature"]

图 5.10 部分关联规则表

根据第四章中所提出的基于关联规则的兴趣爱好挖掘方法对原始兴趣项列表中的权值 w 进行重新计算，并根据权值 w 从大到小序列重新排列所挖掘出的兴趣项，最后得到排序后的兴趣项列表，根据项目的需要，选取前 10 个兴趣项作为该用户的兴趣爱好。除此之外，根据用户的兴趣爱好并结合心理学知识可以对用户的个人特征做初步的分析，用户的兴趣爱好挖掘结果以及个人特征分析如图 5.11 所示。

_id	screen_name	interest	luxuries	decisiveness	intelligence	adventure	comsumption
100011309	mukeshkapila	{7 Keys}	{1 Keys}	1	3	0	2
1001807802	CommonsIDC	{7 Keys}	{1 Keys}	2	4	1	0
1003537712	NicholasHopton	{7 Keys}	{2 Keys}	0	2	0	1
1003843194	Crisis_Action	{7 Keys}	{1 Keys}	0	2	1	1
1005309654	G8	{0 Keys}	{1 Keys}	0	0	0	0
1006698170	Kathy_Settle	{7 Keys}	{1 Keys}	0	3	0	0
1008373598	BRCC_UK_RO	{10 Keys}	{1 Keys}	3	2	3	2
1009149996	WhatTimTweets		Keys}	1	3	0	2
1009722420	_iamsunny_		Keys}	0	3	0	1
101108995	MichaelBrianLaw		Keys}	0	3	0	0
1011249092	andrewkpike		Keys}	0	2	0	1
1012828891	franmendez74		Keys}	0	2	1	0
1014468602	Neilmckillop84		Keys}	0	2	0	1
1014922284	hannahconnagha		Keys}	0	3	0	0
101503567	marshallreport		Keys}	1	2	0	1
1015216418	robaway123		Keys}	0	2	0	0

{
"news" : 3.0,
"markets" : 7.0,
"business" : 75.0,
"socialising" : 76.0,
"investment" : 16.0,
"europe" : 18.0,
"food" : 19.0,
"internet" : 8.0,
"research" : 3.0,
"energy" : 6.0
}

图 5.11 部分用户的兴趣爱好与个人特征分数表

5.3 本章小结

本章主要介绍了基于海量社交网络数据的人物属性识别系统。首先介绍了该系统的项目背景，然后对系统的总体架构与部分工作原理进行了详细说明。接下来，阐述了系统在实际开发过程中的关键环节以及所遇到的重要难点。最后介绍

了本文中所提出的组织成员识别方法和基于关联规则的兴趣爱好挖掘方法在该系统的实际应用，并展示了在系统中的部分应用效果。

第六章 总结和展望

6.1 工作总结

随着互联网的普及与发展，社交网络的流行趋势也达到了高峰期，各大社交网络服务也越来越多样化。众多网民积极参与到这种新型的信息交流平台，社交网络用户数量呈现出急剧增长的态势，同时也带来了社交网络中用户信息的爆炸性增长。因为社交网络与人们的生活越来越密切，所以针对社交网络领域的研究也越来越得到学者们的重视。

社交网络中的信息主要分为两类。一类是用户关系信息，即存在于虚拟的社交网络中的人际关系圈所隐含的信息，针对该信息进行的研究也称为社交网络用户关系研究。目前社交网络中用户关系的研究主要集中在社群发现领域，与社群发现相比，社交网络的特定组织成员识别研究具有更直接的目的，对于组织成员识别，给定一个组织，试图去发现该组织在社交网络上的成员。对该问题进行研究具有很高的学术与应用价值，但目前针对该问题的研究还处于初始阶段，大部分的学者提出的组织成员识别方法依然局限于单纯地考查社交网络用户与组织机构公共账号之间的关注关系，并没有深入地研究组织机构中组织成员之间的关注关系。除此之外，对用户行为关系的研究也只涉及到了关注关系，对其他潜在的用户行为关系没有研究。另一类是消息信息，包括社交网络用户自己发布或者转发的推文、图片、音视频信息等。通过对消息信息进行深入挖掘，可以得到社交网络用户的个人属性特征。在用户属性特征中，兴趣爱好又是学者们关注的重点，兴趣爱好是指个人的心理倾向，希望知道和掌握某些东西，并经常参与这些活动，或是指个人有积极探索某些东西的认知倾向。社交网络中基于兴趣爱好的推荐系统无论是在好友推荐或是商品营销领域，都能起到良好的推荐效果。目前已经有非常多的学者对社交网络用户兴趣爱好挖掘开展了研究，但大多数的学者只关注于对消息信息的直接挖掘，而忽略了兴趣爱好之间存在的关联关系。

针对当前社交网络中组织成员识别与兴趣爱好挖掘所存在的不足与问题，本文结合基于海量社交网络数据的人物属性识别系统的实际开发需求，提出了基于社交网络用户行为关系的组织成员识别方法和基于关联规则的兴趣爱好挖掘方法。本文主要包括以下成果和创新点：

（1）提出了一种基于社交网络用户行为关系的组织成员识别方法。

该方法首先根据社交网络 Twitter 的特点以及其用户的行为属性，定义了多种识别因子量化地描述了用户与用户组之间的行为关系，并归纳总结了社交网络

用户关系的基本判定规则，将虚拟网络中人际关系的界定转化为模型迭代计算的过程。然后基于社交网络的真实数据对识别模型进行实证研究，探讨了多种识别因子在识别过程中的影响程度和最优组合模型。

最后通过对比实验验证了该方法比现有的基于公共账号粉丝关系的方法具有更高的组织成员识别率。

（2）提出一种基于关联规则的兴趣爱好挖掘方法。

首先基于 LinkedIn 用户的个人档案，对兴趣爱好进行建模，将不同形式的兴趣爱好字符串使用统一的兴趣项进行标准化处理，并从中挖掘出高频兴趣项作为研究对象。在此基础上建立了兴趣爱好关联分析模型，并基于 LinkedIn 的真实用户数据生成兴趣爱好关联规则。然后使用兴趣爱好关联规则对原始的基于词频的兴趣爱好挖掘方法进行了改进。

最后通过实验结果验证了合理地利用兴趣爱好关联规则可以显著提升了兴趣爱好挖掘的效果。

（3）开发了基于海量社交网络数据的人物属性识别系统

基于海量社交网络数据，并结合本文中提出的组织成员识别方法与兴趣爱好挖掘方法，作者开发了基于海量社交网络数据的人物属性识别系统。该系统能将社交网络用户按照组织为单元进行划分，并根据挖掘出的用户兴趣爱好分析用户的人物属性。

最后展示了组织成员识别方法与兴趣爱好挖掘方法在该系统中应用的实际效果。

6.2 未来工作展望

本文主要针对社交网络组织成员识别与兴趣爱好挖掘进行了研究，通过对该领域国内外现状的研究，总结了现有的组织成员识别与兴趣爱好挖掘方法的一些缺陷，随后根据存在的缺陷，对该问题的方法进行了探索研究和创新，完成了一部分工作。由于作者在社交网络信息的研究上只是一个新人，科研能力较弱，个人水平和时间精力的限制，尽管本文提出的方法相较于现有方法已经有所进步，但仍然存在一些不足需要继续改进。在未来的工作中，仍然可以围绕以下几个方面进行更加深入的研究。

（1）在组织成员识别方法中，本文提出的方法只考查了主动、被动情景下的关注关系与推送关系。在后续的研究中，可以考虑引入更多的社交网络用户行为属性，例如点赞关系，转发关系等。充分地探究并利用这些用户行为属性在用户关系研究中隐含的影响能力可以显著地提升组织成员识别的效果。

（2）在组织成员识别方法中所使用的实验数据虽然已经采集了较长的时间。但对于未来新产生的用户信息数据没有进行充分地利用。所以可以在后续的工作

中继续采集一段时间的用户数据进行验证实验，以此证明该方法的时间有效性。

（3）在基于关联规则的兴趣爱好挖掘方法中，实验中采用了最小支持度和置信度阈值分别设为 0.4%和 20%的规则集，该规则集总共有 286 条关联规则，属于较强的关联规则。但对于其他最小支持度和置信度阈值设置下的规则集，没有进行考查与探究。如何选择最优的规则集，以使基于关联规则的兴趣爱好挖掘的效果进一步提升也是未来工作中值得重点研究的问题。

最后，由于作者在社交网络信息研究领域的水平有限，还有很多需要学习提升的部分，论文中难免存在一些不足和错误，在此恳请各位学者、专家、同行朋友们批评指正，本人将不胜感激。

致谢

转眼间即将毕业，我将离开熟悉的大学校园，开始人生的另一个阶段。2011年我第一次离开家乡，来到杭州这美丽而又陌生的城市，当年入学时的画面还历历在目。在杭电的四年本科生活和两年半硕士研究生生活是我人生中永远值得回忆的时光。感谢母校，让我在自由温暖的大学环境中度过了青春最美好的时光。

正如罗大佑的《光阴的故事》中所唱的那样“流水它带走光阴的故事，改变了一个人”。经历了多年的大学生活，我也得到了蜕变。回首过去，有成功时的喜悦，也有失败时的失落。虽然一路摸爬滚打，磕磕绊绊也最终到达了旅程的终点，感谢这一路上给予我帮助的老师们，同学们，朋友们。

首先感谢我的硕士导师万健教授。我第一次见到万老师是在大学本科二年级的 C++程序设计课程上，他是我们的任课老师，万老师在课堂上亲自编写代码，并耐心教导我们调试程序，他的课程让我深深地体会到了编程之美。后来我报考了万老师的硕士研究生，也很荣幸地再次成为他的学生。虽然在硕士研究生期间，因为万老师工作调度的原因，我平时不能频繁地和他见面交流，但为数不多的几次交流已经让我受益匪浅。再次感谢万老师在科研项目上的耐心指导以及在生活中的关心。

感谢我的小导师司华友老师。从研究生一年级开始，我就在杭州健港科技有限公司担任实习生，跟随着司老师参加各种各样的科研工程项目。司老师对科研工作的认真负责，让我印象深刻。每当我在科研项目中遇到困难时，司老师总会不厌其烦的给予我帮助，在他的悉心教导下我克服了许多难题，最终都顺利的完成了科研项目。除此之外，在学术论文方面，司老师也给我提供了巨大的帮助，从论文的选题，实验设置以及写作定稿，司老师都给予了极好的引导和鼓励。在此谨向司华友老师致以最诚挚的谢意和最崇高的敬意。

感谢云技术研究中心的张纪林老师、殷昱煜老师、蒋从锋老师、任祖杰老师、张伟老师、周仁杰老师、贾刚勇老师、黄杰老师、沈静老师、李尤慧子老师在我的研究生生涯中给予我的关怀和帮助。

感谢实验室的师兄师姐们。感谢宋爱华学姐在我考研时对我的帮助和指导。感谢任迪、信桂龙师兄在工程项目中对我帮助和指导。感谢梁敏军、张炫师兄在招聘求职中对我的帮助和指导。需要感谢的师兄师姐还有太多太多，这里没有办法一一列举。总之有幸认识这些优秀的师兄师姐，是我人生宝贵的财富，因为你无私地帮助，我才能快速成长。

感谢和我一起入学，陪伴我一起渡过研究生生活的同学们。感谢“生活百科全书”金厅同学对我学习和生活的帮助，感谢“算法大神”陈彬彬同学对我算法学习的指导，他们毕业之后都将前往北京工作，祝愿他们在帝都能够工作顺利，生活幸福。感谢“文艺青年”潘可同学为我的学术研究铺路，帮我完成了一些初步实验工作。感谢学弟吴浩鹏同学帮我采集了大量的社交网络实验数据。在健港实习的时光里，能和你们一起共事，我感到很开心。感谢“胖嘟嘟”孙可嘉同学给我的生活带了许多欢乐，研究生阶段的业余时间他是我最好的玩伴，他也通过我认识了人生的另一半高敏同学，希望他们能幸福圆满，白头偕老。除此之外，还要感谢实验室的其他同学，是你们让我的研究生生活变得绚丽多彩。

感谢我的父母家人，感谢你们的付出，感谢你们的无私，感谢你们的支持理解，我虽然不在家乡学习与工作，但我内心时刻惦记着你们。

最后，以最诚挚的谢意，向这些年来所有关心过我，帮助过我的人们表示真挚的感谢，并向百忙之中抽出宝贵时间评审本文的专家、学者和教授们致以最真挚的谢意！

参考文献

[1] Scott W R. Institutions and organizations: ideas, interests, and identities[M]. Sage, 2013:9-15.

[2] 王倩倩. 特定组织结构的微博相关用户挖掘算法研究[D]. 北京邮电大学, 2014.

[3] Hazratzadeh S, Navimipour N J. Colleague recommender system in the expert cloud using features matrix[J]. Kybernetes, 2016,45(9):1342-1357.

[4] Luo Q, Zhong D. Using social network analysis to explain communication characteristics of travel-related electronic word-of-mouth on social networking sites[J]. Tourism Management, 2015,46:274-282.

[5] 仲兆满, 管燕, 胡云, 李存华. 基于背景和内容的微博用户兴趣挖掘[J]. 软件学报, 2017, 28(2): 278-291.

[6] Houston J M, Harris P B, Howansky K, et al. Winning at work: Trait competitiveness, personality types, and occupational interests[J]. Personality and Individual Differences, 2015,76:49-51.

[7] 邓钟晟. 社交网络中基于关系强度的用户群体发现研究[D]. 东华大学, 2015.

[8] 李洋, 陈毅恒, 刘挺. 微博信息传播预测研究综述[J].软件学报,2016, 27(2): 247-263.

[9] 李桃陶, 周斌, 王忠振. 基于社交网络的图数据挖掘应用研究[J]. 计算机技术与发展, 2014(10):6-11.

[10] 吴烨, 钟志农, 熊伟, 等. 一种高效的属性图聚类方法[J]. 计算机学报, 2013,36(8):1704-1713.

[11] Zhang Y, Wu Y, Yang Q. Community Discovery in Twitter Based on User Interests[J]. Journal of Computational Information Systems, 2012,8(3):2012.

[12] Danyllo W A, Alisson B V, Alexandre D N, et al. Identifying Relevant Users and Groups in the Context of Credit Analysis Based on Data from Twitter[C]. 2013 International Conference on Cloud and Green Computing, Karlsruhe, Germany. IEEE, 2013: 587-592.

[13] Sotiropoulos D N, Kounavis C D, Giaglis G M. Semantically meaningful group detection within sub-communities of Twitter blogosphere: a topic oriented multi-objective clustering approach[C]. Proceedings of the 2013 IEEE/ACM

International Conference on Advances in Social Networks Analysis and Mining, Niagara, Ontario, Canada. ACM, 2013:734-738.

[14] Ben Ahmed E, Nabli A, Gargouri F. Group extraction from professional social network using a new semi-supervised hierarchical clustering[J]. Knowledge and Information Systems, 2014,40(1):29-47.

[15] Liu X, Song M, Tao D, et al. Random Forest Construction with Robust Semisupervised Node Splitting[J]. IEEE Transactions on Image Processing, 2015,24(1):471-483.

[16] 刘冰玉, 王翠荣, 王聪, 王军伟, 王兴伟, 黄敏. 基于动态主题模型融合多维数据的微博社区发现算法[J]. 软件学报, 2017, 28(2): 246-261.

[17] 孙怡帆, 李赛. 基于相似度的微博社交网络的社区发现方法[J]. 计算机研究与发展, 2014,51(12):2797-2807.

[18] 范田. 基于主题和结构的微博社区挖掘方法研究[D]. 中国科学技术大学, 2014.

[19] 张振华, 刘瑞芳. 微博社交网络中面向机构的用户挖掘[J]. 软件, 2013,34(1):121-124.

[20] Yu Z, Wang C, Bu J, et al. Friend recommendation with content spread enhancement in social networks[J]. Information Sciences, 2015,309:102-118.

[21] Huang S, Zhang J, Schonfeld D, et al. Two-Stage Friend Recommendation Based on Network Alignment and Series Expansion of Probabilistic Topic Model[J]. IEEE Transactions on Multimedia, 2017,19(6):1314-1326.

[22] Guo L, Zhang C, Fang Y. A Trust-Based Privacy-Preserving Friend Recommendation Scheme for Online Social Networks[J]. IEEE Transactions on Dependable and Secure Computing, 2015,12(4):413-427.

[23] Li F, He Y, Niu B, et al. Small-world: Secure friend matching over physical world and social networks[J]. Information Sciences, 2017,387:205-220.

[24] Chu H C, Wu C W, Wang C C, et al. Friend Recommendation for Location-Based Mobile Social Networks[C]. 2013 Seventh International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, Taichung, Taiwan, 2013:365-370.

[25] De Meo P, Nocera A, Terracina G, et al. Recommendation of similar users, resources and social networks in a Social Internetworking Scenario[J]. Information Sciences, 2011,181(7):1285-1305.

[26] Chao C H, Lai F C, Chen Y S, et al. A M-Learning Content Recommendation

Service by Exploiting Mobile Social Interactions[J]. IEEE Transactions on Learning Technologies, 2014,7(3):221-230.

[27] Ma H, Jia M, Zhang D, et al. Combining tag correlation and user social relation for microblog recommendation[J]. Information Sciences, 2017, 86:325-337.

[28] Wang Z, Sun L, Zhu W, et al. Joint Social and Content Recommendation for User-Generated Videos in Online Social Network[J]. IEEE Transactions on Multimedia , 2013,15(3):698-709.

[29] Yang X, Guo Y, Liu Y. Bayesian-Inference-Based Recommendation in Online Social Networks[J]. IEEE Transactions on Parallel and Distributed Systems, 2013,24(4):642-651.

[30] Philip Chen C L, Zhang C. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data[J]. Information Sciences, 2014,275:314-347.

[31] Wang H, Xu Z, Fujita H, et al. Towards felicitous decision making: An overview on challenges and trends of Big Data[J]. Information Sciences, 2016,367:747-765.

[32] Krapp A, Prenzel M. Research on Interest in Science: Theories, methods, and findings[J]. International Journal of Science Education, 2011,33(1):27-50.

[33] Mayer R E. The Role of Interest in Learning and Development. Psychology Press[J]. The American Journal of Psychology, 1994,107(2):319-323.

[34] Hidi S, Renninger K A. The Four-Phase Model of Interest Development[J]. Educational Psychologist, 2006,41(2):111-127.

[35] 薛小丽. 西方近现代兴趣教学思想研究——兼论当代教学论的重建[M]. 西南大学, 2008:11-17.

[36] Holtrop D, Born M P, de Vries R E. Relating the Spherical representation of vocational interests to the HEXACO personality model[J]. Journal of Vocational Behavior, 2015,89:10-20.

[37] Major J, Johnson W, Deary I. Trait complexes of cognitive abilities and interests and their relations to realized occupation[J]. Personality and Individual Differences, 2014,60: S46.

[38] Holland J L. Making Vocational Choices: A Theory of Vocational Personalities and Work Environment[J]. British Journal of Guidance & Counselling, 1995(1):153-154.

[39] Gou L, You F, Guo J, et al. SFViz: interest-based friends exploration and recommendation in social networks[C]. Proceedings of the 2011 Visual Information Communication - International Symposium, Hong Kong, China.

ACM,2011:1-10.

[40] Li D, Lv Q, Xie X, et al. Interest-based real-time content recommendation in online social communities[J]. Knowledge-Based Systems, 2012,28:1-12.

[41] Qian X, Feng H, Zhao G, et al. Personalized Recommendation Combining User Interest and Social Circle[J]. IEEE Transactions on Knowledge and Data Engineering, 2014,26(7):1763-1777.

[42] Cao B, Liu J, Tang M, et al. Mashup Service Recommendation Based on User Interest and Social Network[C]. 2013 IEEE 20th International Conference on Web Services, Santa Clara, CA, USA. 2013:99-106.

[43] Gao H, Tang J, Hu X, et al. Content-aware point of interest recommendation on location-based social networks[C]. Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin Texas, USA. 2015: 1721-1727.

[44] Yin H, Cui B, Huang Z, et al. Joint Modeling of Users' Interests and Mobility Patterns for Point-of-Interest Recommendation[C]. Proceedings of the 23rd ACM international conference on Multimedia, Brisbane, Australia. ACM, 2015: 819-822.

[45] 景宁, 王跃华, 钟志农, 等. 地理社交网络位置推荐[J]. 国防科技大学学报, 2015(5):1-8.

[46] 刘淇. 基于用户兴趣建模的推荐方法及应用研究[D]. 中国科学技术大学, 2013.

[47] Gabriel R, Gabriel F. Automatic topic and interest-based content recommendation system for mobile devices: US, US20150262069A1[P]. 2015-09-17.

[48] Flinn S D, Moneypenny N F. Mutual expressions of interest-based people matching system and method:US, US9026488B2[P]. 2015-05-05.

[49] Li L, Zheng L, Yang F, et al. Modeling and broadening temporal user interest in personalized news recommendation[J]. Expert Systems with Applications, 2014,41(7):3168-3177.

[50] Shen W, Wang J, Luo P, et al. Linking named entities in Tweets with knowledge base via user interest modeling[C]. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, Chicago, Illinois, USA. ACM, 2013:68-76.

[51] Liu X, Turtle H. Real-time user interest modeling for real-time ranking[J]. Journal of the American Society for Information Science & Technology, 2013,64(8):1557-1576.

[52] 田军伟. 基于社会网络的用户兴趣模型研究[D]. 电子科技大学, 2010.

[53] Deng L, Jia Y, Zhou B, et al. User interest mining via tags and bidirectional interactions on Sina Weibo[J]. World Wide Web, 2017.

[54] Vu T, Perez V. Interest mining from user tweets[C]. Proceedings of the 22nd ACM international conference on Information & Knowledge Management, San Francisco, California, USA. ACM, 2013:1869-1872.

[55] Bao H, Li Q, Liao S S, et al. A new temporal and social PMF-based method to predict users' interests in micro-blogging[J]. Decision Support Systems, 2013,55(3):698-709.

[56] 陈希友. 基于web日志挖掘的用户访问预测研究[D]. 厦门大学, 2009.

[57] 李建廷, 郭晔, 汤志军. 基于用户浏览行为分析的用户兴趣度计算[J]. 计算机工程与设计, 2012,33(3):968-972.

[58] Russell M A. Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More[M]. O'Reilly Media, Inc., 2013:3-5.

[59] Kwak H, Lee C, Park H, et al. 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. ACM, 2010: 591-600.

[60] 赵文硕. 关系型与非关系型数据库的应用研究[D]. 华北电力大学(北京), 2016.

[61] Borgelt C, Kruse R. Induction of Association Rules: Apriori Implementation[M]// Härdle W, Rönz B. Compstat: Proceedings in Computational Statistics. Heidelberg: Physica-Verlag HD, 2002:395-400.

[62] Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation[J]. SIGMOD Rec., 2000,29(2):1-12.

[63] Fire M, Puzis R. Organization Mining Using Online Social Networks[J]. Networks and Spatial Economics, 2016,16(2):545-578.

[64] Ellison N B, Gibbs J L, Weber M S. The Use of Enterprise Social Network Sites for Knowledge Sharing in Distributed Organizations: The Role of Organizational Affordances[J]. American Behavioral Scientist, 2015,59(1):103-123.

[65] Serrat O. Social Network Analysis[M]//Knowledge Solutions: Tools, Methods, and Approaches to Drive Organizational Performance. Singapore: Springer Singapore, 2017:39-43.

[66] Watanabe M, Olson K, Falci C. Social isolation, survey nonresponse, and nonresponse bias: An empirical evaluation using social network data within an organization[J]. Social Science Research, 2017,63:324-338.

附录

作者在读期间发表的学术论文及参加的科研项目

一、硕士研究生期间发表或录用的论文

[1] Zhihui Chen, Huayou Si*, Feiwei Qin, Jian Wan. An Approach to Member Recognition for Specific Organizations Based on User Interaction on Twitter[C]. 2017 9th International Conference on Intelligent Human-Machine Systems and Cybernetics, Hangzhou, China. IEEE, 2017: 97-100.（EI，已发表）

[2] Huayou Si, Zhihui Chen, Wei Zhang, Jian Wan, Jilin Zhang, Naixue Xiong*. A Member Recognition Approach for Specific Organizations Based on Relationships among Users in Social Networking Twitter[J]. Future Generation Computer Systems. (SCI，已录用)

[3] Huayou Si, Zhihui Chen, Jian Wan, Naixue Xiong*, Wei Zhang. Association Rules Mining among Interests and Application to Interest Mining for Users on Social Network [J]. Future Generation Computer Systems. (SCI，已投稿)

[4] Huayou Si, Zhihui Chen, Jian Wan*, Naixue Xiong*, Wei Zhang. OWL Axioms Publication and Retrieving on Demand Based on Structured P2P System for Knowledge Sharing in Semantic Sensor Network[J]. Sensor (SCI，已投稿)

二、 硕士研究生期间申请的软件著作权

[1] MyCMS 内容管理系统 v1.0

三、 硕士研究生期间参加的科技竞赛

[1] 2017 华为软件精英挑战赛杭厦赛区二等奖

四、 硕士研究生期间参加的科研项目

[1] 国家自然科学基金：基于用户语义生成的 Web 资源自组织与共享方法研究 (61472112) 2015-2018

[2] 舟山同博科技有限公司：北斗船舶定位搜救系统 2015-2016

[3] 杭州经济技术开发区高层次人才创业创新项目：基于云存储的大数据分析系统及其医疗行业智能分析与服务平台

杭州电子科技大学

硕 士 学 位 论 文
详 细 摘 要

题　　目：基于社交网络的组织成员识别及其兴趣爱好挖掘方法研究

研 究 生　陈志辉

专　　业　计 算 机 技 术

指导教师　万 健 教授

完成日期　2018 年 3 月

基于社交网络的组织成员识别及其兴趣爱好挖掘方法研究

详细摘要

随着 Web2.0 时代的发展，社交网络平台已经成为互联网服务中不可或缺的重要组成部分。网民们积极地加入到这种新型的信息交流平台中。社交网络用户数量呈现出急剧增长的态势，同时也导致了社交网络中信息量的爆炸性增长，所以近几年来，越来越多的学者开展了针对社交网络的研究。

社交网络中的信息主要分为两类。一类是用户关系信息，即存在于虚拟的社交网络中的人际关系圈所隐含的信息，针对该信息进行的研究也称为社交网络用户关系研究。目前社交网络中用户关系的研究主要集中在社群发现领域，与社群发现相比，社交网络的特定组织成员识别研究具有更直接的目的，对于组织成员识别，给定一个组织，试图去发现该组织在社交网络上的成员。对该问题进行研究具有很高的学术与应用价值，但目前针对该问题的研究还处于初始阶段，大部分的学者提出的组织成员识别方法依然局限于单纯地考查社交网络用户与组织机构公共账号之间的关注关系，并没有深入地研究组织机构中组织成员之间的关注关系。除此之外，对用户行为关系的研究也只涉及到了关注关系，对其他潜在的用户行为关系没有研究。另一类是消息信息，包括社交网络用户自己发布或者转发的推文、图片、音视频信息等。通过对消息信息进行深入挖掘，可以得到社交网络用户的个人属性特征。在用户属性特征中，兴趣爱好又是学者们关注的重点，兴趣爱好是指个人的心理倾向，希望知道和掌握某些东西，并经常参与这些活动，或是指个人有积极探索某些东西的认知倾向。社交网络中基于兴趣爱好的推荐系统无论是在好友推荐或是商品营销领域，都能起到良好的推荐效果。目前已经有非常多的学者对社交网络用户兴趣爱好挖掘开展了研究，比较常见的方法有基于标签双向交互的挖掘算法、基于关键字排序技术的挖掘算法等。但大多数的学者只关注于对消息信息的直接挖掘，而忽略了兴趣爱好之间存在的关联关系，没有将关联分析技术运用到兴趣爱好挖掘的工作中：

（1）提出了一种基于社交网络用户行为关系的组织成员识别方法。该方法首先根据社交网络 Twitter 的特点及其用户的行为属性，定义了多种识别因子，量化地描述了用户与用户组之间的行为关系，并归纳总结了社交网络用户关系的基本判定规则，将社交网络中用户关系的界定转化为模型计算的过程。然后基于社交网络的真实数据对识别模型进行实证研究，最后详细探讨了多种识别因子在识别过程中的影响程度和最优组合模型.

（2）提出一种基于关联规则的兴趣爱好挖掘方法。基于 LinkedIn 用户的个人档案，首先对兴趣爱好进行建模，将不同形式的兴趣爱好字符串使用统一的兴趣项进行标准化整理，并从中挖掘出高频兴趣项作为研究对象。然后在此基础上建立了兴趣爱好关联分析模型，并基于 LinkedIn 的真实用户数据生成兴趣爱好关联规则。最后使用兴趣爱好关联规则对原始的基于词频的 Twitter 用户兴趣爱好挖掘方法进行了改进。

（3）结合上述两点，作者开发了基于海量社交网络数据的人物属性识别系统，并将组织成员识别与兴趣爱好挖掘方法运用到该系统中。系统能将社交网络用户按照组织为单元进行划分，并根据挖掘出的用户兴趣爱好研究其人物属性。

本文所提出的方法可以在错综复杂的社交网络中识别出属于相同组织的用户，并且准确地挖掘用户的兴趣爱好。研究成果不仅可用于广告投放、好友推荐等商业活动，对社交网络用户行为特征和属性特征的研究也具有借鉴意义。

关键词： 社交网络，用户关系，组织成员，兴趣爱好