**社交网络分析综述报告**

**摘 要** 社交网络是社会个体成员之间通过社会关系结成的网络体系。个体也称为节点，可以是组织、个人、网络ID 等不同含义的实体或虚拟个体；而个体间的相互关系可以是亲友、动作行为、收发消息等多种关系。而社交网络分析、是指基于信息学、数学、社会学、管理学、心理学等多学科的融合理论和方法，为理解人类各种社交关系的形成、行为特点分析以及信息传播的规律提供的一种可计算的分析方法。本文将从社交网络的结构特性与演化机理、社交网络群体行为形成与互动规律、社交网络信息传播与演化机理、社交网络分析的应用这4个方面，较为详细的阐述社交网络的各个领域研究方向。第一部分社交网络的结构特性与演化机理主要描述了社交网络的结构分析与建模、虚拟社区及其发现技术、虚拟社区演化分析；第二部分社交网络群体行为形成与互动规律主要描述了用户行为分析、社交网络情感分析、个体影响力分析、群体聚集及影响机制分析；第三部分社交网络信息传播与演化机理主要描述了在线社交网络信息检索、社交网络信息传播规律、话题发现与演化、影响力最大化；第四部分社交网络分析的应用主要描述了用户画像、舆情分析、社交推荐、可是化、语言检测、隐私保护。

**关键词** 社交网络； 情感分析； 信息检索； 社交推荐

**Abstract** Social networks are network systems formed by social relationships among individual members of society. An individual is also called a node, and can be an entity or a virtual individual with different meanings such as an organization, an individual, or a network ID; and the mutual relationship between the individuals can be a relationship between friends and relatives, action behaviors, and sending and receiving messages. Social network analysis refers to multi-disciplinary fusion theories and methods based on informatics, mathematics, sociology, management, psychology, etc., which provide for understanding the formation of human social relationships, the analysis of behavior characteristics, and the law of information dissemination. A computable analytical method. This paper will elaborate on the research directions of various fields of social networks from four aspects: the structural characteristics and evolution mechanism of social networks, the formation and interaction rules of social network group behavior, the mechanism of social network information dissemination and evolution, and the application of social network analysis. The first part of the social network's structural characteristics and evolution mechanism mainly describes the social network structure analysis and modeling, virtual community and its discovery technology, virtual community evolution analysis; the second part of the social network group behavior formation and interaction law mainly describes the user behavior Analysis, social network sentiment analysis, individual influence analysis, group aggregation and impact mechanism analysis. The third part of social network information dissemination and evolution mechanism mainly describes online social network information retrieval, social network information dissemination law, topic discovery and evolution, impact Maximization of power; The fourth part of the application of social network analysis mainly describes user portraits, public opinion analysis, social recommendation, customization, language detection, privacy protection.

**Key words** Social network；emotion analysis；Information retrieval；Social reco-mmendation

**章节分工** 钱云冲负责第一章内容**，**范德宝负责第二章内容，蔡静轩负责第三章内容**，**陈凌负责第四章内容。

# 1前言

社交网络在维基百科的定义是“由许多节点构成的一种社会结构。节点通常是指个人或组织，而社交网络代表着各种社会关系。”在互联网诞生前，社交网络分析是社会学和人类学重要的研究分支。早期的社交网络的主要指通过合作关系建立起来的职业网络，如科研合作网络、演员合作网络等。

本文所指的社交网络分析专指在线社交网络分析（Online Social Network Analysis），该门科学的发展是随着在线社交服务（Social Network Service, SNS）的出现而诞生。在线社交服务的种类大致可分为四种：即时消息类应用（QQ、微信、WhatsApp、Skype 等），在线社交类应用（QQ空间、人人网、Facebook、Google+ 等），微博类应用（新浪微博、腾讯微博、Twitter 等），共享空间类应用（论坛、博客、视频分享、评价分享等）。

在线社交网络（下文统称社交网络）有着迅捷性、蔓延性、平等性与自组织性等四大特点。正因为这些特性，其在互联网出现的短短数十年内已经拥有数十亿用户并对现实社会的方方面面产生着影响。在2016年的美国总统大选中，当选总统特朗普就很好地利用了推特作为宣传工具；而在国内，从魏则西事件到和颐酒店事­­­件再到最近的“刺死辱母者”事件，无一不是在社交网络上迅速发酵，并最终对现实社会产生影响。而且这种线上影响线下的趋势越来越明显。

除了社交网络给社会和经济带来许多正面影响之外，也带来了不少负面影响。从Facebook 和 YouTube上的暴力恐怖信息传播到微博微信上大量谣言和假新闻，这些有害信息借助社交网络的特点迅速传播并且往往产生不可控的后果。

为了利用好社交网络的特性，产生价值，消除危害，所以产生了社交网络分析这门科学。它是一种基于信息学、数学、社会学、管理学和心理学等科学的交叉科学。根据社交网络的特性，其主要研究三大内容：结构与演化，群体与互动，信息与传播。

# 2社交网络的结构特性与演化机理

## 2.1社交网络结构分析与建模

### 2.1.1统计特性

社交网络模型许多概念来自于图论，因为社交网络模型本质上是一个由节点（人）和边（社交关系）组成的图。而图模型中有一些非常重要的和社交网络息息相关的概念，例如度数、网络密度等，可以用来具体细节的刻画社交网络的数学特性。

1）度（Degree）

节点的度定义为与该节点相连的边的数目。在有向图中，所有指向某节点的边的数量叫作该节点的入度，所有从该节点出发指向别的节点的边的数量叫作该节点的出度。网络平均度反应了网络的疏密程度，而通过度分布则可以刻画不同节点的重要性。

2）网络密度（Density）

网络密度可以用于刻画节点间相互连边的密集程度，定义为网络中实际存在边数与可容纳边数上限的比值，常用来测量社交网络中社交关系的密集程度及演化趋势。

3）聚类系数（Clustering Coefficient）

用于描述网络中与同一节点相连的节点间也互为相邻节点的程度。其用于刻画社交网络中一个人朋友们之间也互相是朋友的概率，反应了社交网络中的聚集性。

4）介数（Betweeness）

为图中某节点承载整个图所有最短路径的数量，通常用来评价节点的重要程度，比如在连接不同社群之间的中介节点的介数相对于其他节点来说会非常大，也体现了其在社交网络信息传递中的重要程度。

### 2.1.2网络特性

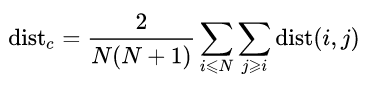
社交网络的网络特性主要有以下2点——小世界现象与无标度特性。

1）小世界现象

小世界现象是指地理位置相距遥远的人可能具有较短的社会关系间隔。早在1967年，哈佛大学心理学教授 Stanley Milgram 通过一个信件投递实验，归纳并提出了“六度分割理论（Six Degrees of Separation）”, 即任意两个都可通过平均五个人熟人相关联起来。1998年，Duncan Watts 和 Steven Strogatz 在《自然》杂志上发表了里程碑式的文章《Collective Dynamics of “Small-World” Networks》，该文章正式提出了小世界网络的概念并建立了小世界模型。

小世界现象在在线社交网络中得到了很好地验证，根据2011年 Facebook 数据分析小组的报告， Facebook 约7.2亿用户中任意两个用户间的平均路径长度仅为4.74，而这一指标在推特中为4.67。可以说，在五步之内，任何两个网络上的个体都可以互相连接。米尔格伦实验、凯文贝肯游戏、埃尔德什数以及一些类似的实验证明了，在现实世界里的一些网络中，尽管节点数量庞大，但从一点出发，其实只需要经过仅仅几步转折，就能到达任一个节点。1998年，美国康奈尔大学的博士生邓肯·瓦茨（Duncan Watts）和他的导师斯蒂文·斯特罗加茨（Steven Strogatz）发表了一篇名为《小世界网络的集体动力学》（Collective dynamics of the 'Small World' networks）的论文。他们把这种现象归类为某一类复杂网络的特性。他们注意到复杂网络可以按两个独立的结构特性分类，就是集聚系数和节点间的平均路径长度。

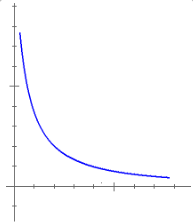
平均路径长度也称为特征路径长度或平均最短路径长度，指的是一个网络中两点之间最短路径长度（或称距离）的平均值。从一个节点sj出发，经过与它相连的节点，逐步“走”到另一个节点sj所经过的路途，称为两点间的路径。其中最短的路径也称为两点间的距离，记作dist(i, j)。而平均路径长度定义为：



集聚系数（也称群聚系数、集群系数）是用来描述图或网络中的顶点（节点）之间结集成团的程度的系数。具体来说，是一个点的邻接点之间相互连接的程度。

2）无标度特性

大多数真实的大规模社交网络都存在着大多数节点有少量边，少数节点有大量边的特点，其网络缺乏一个统一的衡量尺度而呈现出异质性，我们将这种节点度分布不存在有限衡量分布范围的性质称为无标度。无标度网络表现出来的度分布特征为幂律分布，这就是此类网络的无标度特性。



### 2.1.3网络模型

社交网络分析已经是一个相对成熟的领域，因此有非常非常多的网络模型可以用来刻画社交网络，主要可以分为WS模型、BA模型和其他模型。

WS 模型：WS 模型即小世界模型，通过小世界模型生成的小世界网络是从规则网络向随机网络过渡的中间形态。源于1998年Duncan Watts和Steven Strogatz在Nature上联合发表的一篇论文，证明了短路径的存在。W-S模型可以证明两个任意节点产生弱连接的概率很高，抽象地表达了社会网络的基本特征，解释了小世界现象产生的原因。但同时，这个模型并不能体现人们可以很有效地找到这些短路径。

BA 模型：BA模型考虑到现实网络中节点的幂律分布特性，生成无标度网络。源于Albert-László Barabási与Réka Albert在1999年的论文中提出了一个模型来解释复杂网络的无尺度特性。这个模型基于两个假设：

1. 增长模式

不少现实网络是不断扩大不断增长而来的，例如互联网中新网页的诞生，人际网络中新朋友的加入，新的论文的发表，航空网络中新机场的建造等等。

1. 优先连接模式

新的节点在加入时会倾向于与有更多连接的节点相连，例如新网页一般会有到知名的网络站点的连接，新加入社群的人会想与社群中的知名人士结识，新的论文倾向于引用已被广泛引用的著名文献，新机场会优先考虑建立与大机场之间的航线等等。

3）其他模型

森林火灾模型，Kronecker 模型，生产模型。

## 2.2虚拟社区（社团）及发现技术

### 2.2.1定义

虚拟社区基于子图局部性的定义：社区结构是复杂网络节点集合的若干子集，每个子集内部的节点之间的连接相对非常紧密，而不同子集节点之间的连边相对稀疏。

在社交网络中发现虚拟社区有助于理解网络拓扑结构特点，揭示复杂系统内在功能特性，理解社区内个体关系。为信息检索、信息推荐、信息传播控制和公共事件管控提供有力支撑。虚拟社区发现存在着许多经典的算法，这些算法用于挖掘不同规模的虚拟社区，算法在追求高精度的同时力求提高效率（降低时间复杂度）。

### 2.2.2 社区发现算法评价指标

以下评价指标可通过搜索引擎获得详细的介绍：

模块度(Modularity)：通过比较现有网络与基准网络在相同社区划分下的连接密度差来衡量网络社区的优劣。

NMI (Normalized Mutual Information)：利用信息熵来衡量预测社区结构一直社区结构的差异，该值越大，则说明社区结构划分越好，最大值为1时，说明算法划分出的社区结构和一直社区结构一致，算法效果最好。

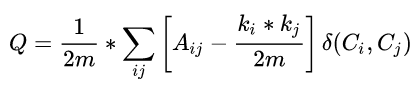
Rand Index：表示在两个划分中都属于同一社区或者都属于不同社区的节点对的数量的比值。

Jaccard Index：Jaccard 系数用来衡量样本之间的差异性，是经典的衡量指标

### 2.2.3 社区静态发现算法

1）模块度最优化算法

Mark Newman 提出了针对模块度的最大化的贪心算法FN[1]。模块度也称模块化度量值，是目前常用的一种衡量网络社区结构强度的方法，最早由Mark NewMan 提出，模块度的定义如下所示。模块度值的大小主要取决于网络中结点的社区分配C，即网络的社区划分情况，可以用来定量的衡量网络社区划分质量，其值越接近1，表示网络划分出的社区结构的强度越强，也就是划分质量越好。因此可以通过最大化模块度Q来获得最优的网络社区划分。



贪心思想的目标是找出目标函数的整体最优值或者近似最优值，它将整体最优化问题分解为局部最优化问题，找出每个局部最优值，最终将局部最优值整合成整体的近似最优值。FN算法将模块度最优化问题分解为模块度局部最优化问题，初始时，算法将网络中的每个结点都看成独立的小社区。然后，考虑所有相连社区两两合并的情况，计算每种合并带来的模块度的增量。基于贪心原则，选取使模块度增长最大或者减小最少的两个社区，将它们合并成一个社区。如此循环迭代，直到所有结点合并成一个社区。随着迭代的进行，网络总的模块度是不断变化的，在模块度的整个变化过程中，其最大值对应网络的社区划分即为近似的最优社区划分。

贪心算法FN具体步骤：

1. 去掉网络中所有的边，网络的每个结点都单独作为一个社区；
2. 网络中的每个连通部分作为一个社区，将还未加入网络的边分别重新加回网络，每次加入一条边，如果加入网络的边连接了两个不同的社区，则合并两个社区，并计算形成新社区划分的模块度增量。选择使模块度增量最大或者减小最少的两个社区进行合并。
3. 如果网络的社区数大于1，则返回步骤（2）继续迭代，否则转到步骤（4）；
4. 遍历每种社区划分对应的模块度值，选取模块度最大的社区划分作为网络的最优划分。

该算法中，需要注意的是，每次加入的边只是影响网络的社区划分，而每次计算网络划分的模块度时，都是在网络完整的拓扑结构上进行，即网络所有的边都存在的拓扑结构上。

为了降低算法的时间复杂度，Vincent Blondel等人提出了另一种层次贪心算法，简称为快速模块度优化算法。该算法包括两个阶段，第一阶段合并社区，算法将每个结点当作一个社区，基于模块度增量最大化标准决定你哪些邻居社区应该被合并。经过一轮扫描后开始第二阶段，算法将第一阶段发现的所有社区重新看成结点，构建新的网络，在新网络上重复进行第一阶段，这两个阶段重复运行，直到网络社区划分的模块度不再增长，得到网络的社区近似最优划分。

这个简单算法具有一下几个优点：首先，算法的步骤比较直观并且易于实现；其次，算法不需要提前设定网络的社区数，并且该算法可以呈现网络的完整的分层社区结构，能够发现在线社交网络的分层的虚拟社区结构，获得不同分辨率的虚拟社区；再次，计算机模拟实验显示，在稀疏网络上，算法是时间复杂度是线性的，在合理的时间内可以处理结点数超过109的网络，因此十分适合在线社交网络这样超大规模的负责网络中虚拟社区的发现。

2）多目标优化算法

目标优化问题一般地就是指通过一定的优化算法获得目标函数的最优化解。当优化的目标函数为一个时称之为单目标优化(Single-objective Optimization Problem, SOP)。当优化的目标函数有两个或两个以上时称为多目标优化(Multi-objective Optimization Problem, MOP) [2]。不同于单目标优化的解为有限解，多目标优化的解通常是一组均衡解。

多目标优化算法归结起来有传统优化算法和智能优化算法两大类。

1. 传统优化算法包括加权法、约束法和线性规划法等，实质上就是将多目标函数转化为单目标函数，通过采用单目标优化的方法达到对多目标函数的求解。

2. 智能优化算法包括进化算法（Evolutionary Algorithm, 简称EA）、粒子群算法（Particle Swarm Optimization, PSO）等。

### 2.2.4 社区动态发现算法

1）派系过滤算法

CPM算法是最早的重叠社区发现算法，它的思想是基于团渗透理论的。算法在二分图、有向图及加权图中均已有所应用。CFinder是G.Palla等基于CPM算法开发的一个自由软件[3]，它不仅能非常有效地定位和可视化处理大规模稀疏网络社群，而且还可用于定量描述社会网络的演变。

算法主要步骤：

1）找到网络中的所有团，构造一个团团重叠矩阵,矩阵是对称的，矩阵的第i行第j列即ccom[i][j]表示第i个团和第j个团的公共节点数。

2）给定参数k，将团团矩阵中非对角线上元素小于k-1，且对角线上元素小于k的所有项置0，其他的元素为1；这样，所有对角线为1的团为k团，而非对角线为1的团i、团j是相邻的。通过这个矩阵可以得到相应的社区。

由于k是个输入参数值，从而k的取值将会影响CPM算法的最终社区发现结果，当k取值越小社区将会越大，且社区结构为稀疏。但是实验证明k的取值影响不是很大，一般值为4到6。然而，由于该算法是基于完全子图，因此比较适用于完全子图比较多的网络，即边密集的网络，对于稀疏网络效率将会很低，且该算法还无法分配完全子图外的顶点。

2）基于相似度的聚合算法

与推荐系统中的最近邻模型类似，基于相似度的聚合算法也需要计算出相似度矩阵，并用聚类方法将用户群体划分[4]。将整个社交网络划分成一个个的子网络，每一个子网都是一个社交圈，从中发现用户所在的子网也就是用户所处的社交圈；再在这个子网中计算用户与其他子网成员的相似度，以此来衡量他们之间成为好友的熟悉程度以及成为好友的可能性，选出相似度高的推荐给用户。

3）标签传播算法

标签传播（LPA）算法是最早的基于标签的一种算法，是所有基于标签的算法的基础，由Zhu等人于2002年提出，它是一种基于图的半监督学习方法，其基本思路是用已标记节点的标签信息去预测未标记节点的标签信息。利用样本间的关系建立关系完全图模型，在完全图中，节点包括已标注和未标注数据，其边表示两个节点的相似度，节点的标签按相似度传递给其他节点[5]。标签数据就像是一个源头，可以对无标签数据进行标注，节点的相似度越大，标签越容易传播。由于该算法简单易实现，算法执行时间短，复杂度低且分类效果好，引起了国内外学者的关注，并将其广泛地应用到多媒体信息分类、虚拟社区挖掘等领域中。标签传播算法最大的特色是简单、高效，缺点是每次迭代结果不稳定，准确率不高。

算法描述如下：

1. 所有节点传播标签一步： Y← TY；

2. 行标准化矩阵 Y 来维持类别的概率；

3. 夹逼标注数据，重复步骤 2 直到 Y 收敛。

步骤 3 可以使得节点标签的类别分布集中在给定的类别。该算法的缺点在于需要预先标注数据，而且需要预先知识知道分类的类别个数。

根据LPA算法基本理论，每个节点的标签按相似度传播给相邻节点，在节点传播的每一步，每个节点根据相邻节点的标签来更新自己的标签，与该节点相似度越大，其相邻节点对其标注的影响权值越大，相似节点的标签越趋于一致，其标签就越容易传播。在标签传播过程中，保持已标注数据的标签不变，使其像一个源头把标签传向未标注数据。最终，当迭代过程结束时，相似节点的概率分布也趋于相似，可以划分到同一个类别中，从而完成标签传播过程。

LPA算法的优点是简单、高效、快速；缺点是每次迭代结果不稳定，准确率不高。在LPA算法中节点的Label有同步更新与异步更新2种更新方法[6]。同步更新方法在二分图中可能出现产生震荡情况。为了避免循环和保证收敛，LPA算法采用异步的策略更新节点的标签，并在每次迭代前对节点重新进行随机排

## 2.3虚拟社区演化分析

在线社交网络中存在着大量显性或者隐性的虚拟社区结构，这些虚拟社区结构并不是永恒不变的，随着事件变化，社区结构也在不断演变。分析动态的虚拟社区结构演化有助于理解整个社交网络的演化过程，所以有着重要的研究价值。

### 2.3.1虚拟社区的涌现

虚拟社区涌现即在社交网络中虚拟社区从无到有的过程，其最重要的特征是网络聚集现象。

周期闭包：所谓周期闭包，是指网络节点倾向于和自己在网络中邻居的邻居建立连接关系而形成的结构，该机制是导致虚拟社区形成的主要因素。实验表明三元闭包的出现概率随着两个节点之间测地距离的增减呈指数递减[8]。相反地，焦点闭包和测地距离无关，其生成原因是两个节点之间有共同的兴趣或参与共同的活动。

偏好连接：在很多真实网络中，新增加的边并不是随机连接的，而是倾向于和具有较大度数的连接。

### 2.3.2 虚拟社区的演化

在线社交网络虚拟社区演化过程非常复杂，影响因素很多。如何挖掘虚拟社区演化中的关键性因素成为社交网络研究中一个重要而有挑战性的课题， 用户个体的累积效应、结构多样性和结构平衡性三个基本因素对虚拟社区演化都存在影响。

### 2.3.3 演化虚拟社区的发现

演化虚拟社区发现目前已有大量的研究资料，目前有五种是比较成熟的算法模型，它们分别是基于相邻时刻相似度直接比较的演化虚拟社区发现、基于演化聚类分析的演化虚拟社区发现[9]、基于拉普拉斯动力学方法的演化虚拟社区发现、基于派系过滤算法的演化虚拟社区发现、基于节点行为趋势分析的演化虚拟社区发现。

# 3群体行为形成与互动规律

## 3.1用户行为分析

社交网络用户行为是用户对自身需求，社会影响和社交网络技术进行综合评估的基础上做出的使用社交网络服务的意愿，以及由此引起的各种使用活动的总和。用户行为是在线社交网络研究的重要内容。现有研究主要基于如下两种思路展开，一是将在线社交网络作为一种特定的信息技术，研究用户对在线社交网络技术的采纳行为、拒绝行为和用户忠诚；二是将在线社交网络视为提供各种服务和应用的平台，研究用户使用各种服务和应用所表现出的特征与规律。

### 3.1.1用户采纳与忠诚

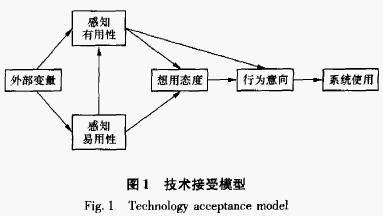
在线社交网络用户采纳是指用户在对自身需求、社会影响和在线社交网络技术进行综合评估的基础上做出的使用在线社交网络服务的意愿或行为，在线社交网络再出现初期能否被尽可能多的用户采纳和试用对于其后续的扩散至关重要。目前已有多种理论被用于揭示在线社交网络用户采纳行为机理。其中，技术接受模型和计划行为理论是研究者们应用最多的两种理论。

在线社交网络用户忠诚是指用户在使用社交网络服务之后，能够继续保持使用的习惯。各种层出不穷的新型网络服务所带来的竞争压力让保持在线社交网络用户忠诚度愈发困难。目前为止，已经有多种理论被用于在线社交网络的用户忠诚研究。其中，期望确认理论和心流体验理论受到较多研究者青睐。

**基于技术接受模型的在线社交网络用户采纳模型**

David Fred 提出技术接受模型是目前信息系统研究领域最经典的模型之一。对模型详细了解可参考：

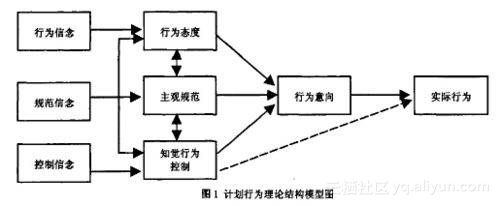
Davis, Fred D. "Perceived usefulness, perceived ease of use, and user acceptance of information technology." MIS quarterly (1989): 319-340.



**基于计划行为理论的在线社交网络用户采纳模型**

Icek Ajzen 提出的计划行为理论已经被广泛用于人类行为研究。对理论详细了解可参考：

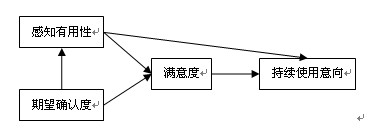
Ajzen, Icek. "From intentions to actions: A theory of planned behavior." Action control. Springer Berlin Heidelberg, 1985. 11-39.



**基于期望确认理论的在线社交网络用户忠诚模型**

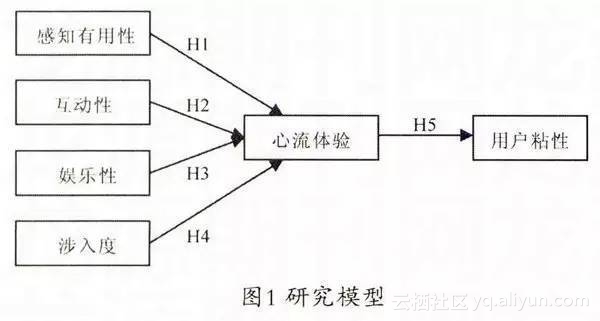
由 Oliver 提出的期望确认理论是研究消费者满意度的基本理论。 Anol Bhattacherjee再该理论的基础上结合信息系统的特点提出了信息系统持续使用的期望确认模型（ECM-ISC）。对模型详细了解可参考：

Bhattacherjee, Anol. "Understanding information systems continuance: an expectation-confirmation model." MIS quarterly (2001): 351-370.



**基于心流体验理论的在线社交网络用户忠诚模型**

Mihaly Csikszentmihalyi等提出的心流体验理论是目前关于用户体验研究的重要理论。对理论详细了解可参考:Csikszentmihalyi, Mihaly. Beyond boredom and anxiety. Jossey-Bass, 2000.



### 3.1.2用户个体使用行为

一般使用行为：用户可以在社交网络上执行各种各样的行为，例如浏览，点击，分享，点赞，收藏等等。具体的分类可参考：

Benevenuto F, Rodrigues T, Cha M, Almeida V. Characterizing User Behavior in Online Social Networks. New York, New York, USA: ACM; 2009:49-62. doi:10.1145/1644893.1644900.

内容创建行为：用户在社交网络通过写博客微博，发帖评论等行为产生内容，对内容创建行为的研究主要研究创建内容的动机、创建内容时的主题选择偏好以及内容创建时的语言表述等。关于主题，可通过搜索引擎搜索 LDA 模型。

内容消费行为：用户在社交网络中通过浏览，分享和评论来满足他们的社交需求，对社交网络内容的消费可分为主动消费和被动消费。被动消费即“浏览”，有研究表明，社交网络中高达92%的行为都是浏览行为。主动消费即社交搜索，例如搜索朋友的信息以及向社交圈内好友提问等等。

### 3.1.3用户群体互动行为

**1)群体互动关系选择**

对群体互动关系的研究主要是识别用户之间的关系，通过制定不同的衡量指标，研究用户之间的关系强弱。

**2)群体互动的内容选择**

社交网络中用户对内容选择与其社交关系密不可分。例如有研究表明两位维基百科编辑在互动前后产生的编辑内容的相似性有所不同。

**3)群体互动的时间规律**

在线社交网络中人类行为的时间特征研究主要集中于分析行为发生的时间间隔分布。研究发现在线社交网络中用户行为时间间隔分布不同于传统的负指数分布，而是呈现幂律分布，即具有“长尾效应”。对群体互动时间规律的研究可以应用到公共管理和决策等场景中。

## 3.2社交网络情感分析

随着互联网技术的迅速发展，网络已经成为人们获取信息，发表意见的主要途径，根据文本内容，我们可以将网络中的文本分为两种，一种是客观描述信息，主要针对事件、产品等进行客观描述，另一种是主观性信息，主要产生与用户对人物、事件、产品进行客观性描述；另一种是主观性信息，主要产生于用户对人物、事件、产品等的评价信息。主观性信息表达了人们的各种情感色彩和情感倾向，如“支持”、“反对”、“中立”等。

情感分析，在此等同于意见挖掘，是针对主观性信息进行分析、处理和归纳过程。情感分析最初起源于自然语言处理领域，主要从语法语义规则方面对文本的情感倾向性进行研判。随着社交网络的兴起与发展，情感分析逐渐涉及多个研究领域，如文本挖掘、Web 数据挖掘等，并延伸至管理学及社会科学等学科，并在产品评论、舆情监控、信息预测等多个领域发挥着重要的作用。

### 3.2.1文本情感分析技术

随着Twitter 、Facebook、新浪微博等社交网络的迅速发展，人们可以随时随地地在网络上发表自己的观点及意见。不同于传统新闻、报道等长文本，社交网络中文本短小、语法不规则性，并含有大量的噪声，针对社交网络中的短文本的情感分析技术具有十分重要的意义。

Go Alec等人在2009年测试了监督学习算法在Twitter短文本上的情感分类效果，如多项式贝叶斯分类、最大熵模型及支持向量机模型。不同于在长文本中依靠人工标注获得的训练集，Go Alec等人采用Twitter中的表情符号获取正面评论和负面评论，从而省去了大量的人工标注成本，同时提升了训练集的规模。

随着微博的兴起，针对短文本的情绪分析成为社交网络分析的一个重点。目前，很多注重于情感分析的评测会议，如自然语言处理与中文计算会议（NLP&CC）、全国信息检索会议（CODE）等，都将情绪分析作为一个重要的部分。Zhang Lumin（张鲁民）等人采用情感向量模型对社交网络中用户的多元化情感进行表示，并基于聚类构造情感向量的层次化结构。

针对短文本话题的情感识别方面，Wang xiaolong（王晓龙）等人针对Twitter中的话题标签（hashtag），通过构造hashtag-graph模型在话题层次上进行情感分析。

而文本情感分析技术主要分为以下三类。

**1)基于语义规则的情感分析技术**

我们将一句话中的带有感情的形容词和副词提取出来构成一个情感词典，这些词语可以代表用户的某种倾向性。基于语义规则的分析技术是计算评价词和情感词典中已经标注倾向性词语的距离，从而达到情感分类的目的。其最经典的算法是 SO-PMI 算法。

**2)基于监督学习的情感分析方法**

基于监督学习的方法是首先通过人工标注文本的情感极性，然后将此作为训练集，通过机器学习的方法对目标文本进行情感分类。常用方法：朴素贝叶斯，支持向量机。

**3)基于话题模型的情感分析技术**

有两个话题模型，PLSA (Probabilistic Latent Semantic Analysis)和 LDA (Latent Dirichlet Allocation) 模型，网络上有大量的学习资料可供读者进一步了解。

### 3.2.2社交网络情感分析技术

面向短文本的情感分析技术：社交网络产生大量的短文本，例如微博和新闻评论，论坛帖子等等，这些短文本不同于新闻报道，其语法不规则，充斥大量噪声，因此对短文本的分析非常重要。

基于群体智能的情感分析技术：用户在社交网络中表达意见会受到其社交关系的影响，情感会沿着社交关系进行传播，因此可以通过研究社交用户之间的关系来提高情感分析的准确度。

社交网络的垃圾意见挖掘技术：社交网络中的垃圾意见，包括水军与广告等信息，通过对垃圾意见的挖掘，能够有效区分有效信息和垃圾信息，从而提高社交网络使用体验。

## 3.3个体影响力分析

随着国内外大量在线社交网络服务的涌现，针对社交网络个体影响力分析的相关研究已经引起了国内外学者的高度关注。在社会学、通信学、经济学、政治科学等领域被广泛研究，在舆情引导与社会运作中起着重要作用。

发现社交网络中的有影响力的个体是社交网络研究中非常重要的研究分支，而且其有着重要的应用价值。例如微博营销，谣言检测，舆情管理等等。

如何在异构、多属性的社交网络中发现高影响力用户，分析社交网络中用户之间的影响强度，是快速变化的网络时代信息决策的一个关键问题。

### 3.3.1基于网络结构的个体影响力计算

基于社交网络的图结构特性，有几个指标用来衡量网络中节点的中心度，即节点的影响力。除了以下三种外还有 PageRank 中心度等度量方法。

**度中心度（Degree Centrality）：**度中心度是指与该节点直接相连的节点的数量。

**接近中心度 (Closeness Centrality)：**指某节点与网络中所有其他节点的最短距离之和。

**介数 (Betweenness Centrality)：**介数用来衡量某节点在社交网络中中介作用大小。网络中某两个节点所有最短路径的数量除以这些路径中经过 A 节点路径的数量便是 A 节点的介数，也叫中间中心度。

### 3.3.2基于行为的个体影响力计算

社交网络中用户的行为决定用户的影响力，以微博为例，用户主要表现的行为是评论、转发、回复、点赞、复制、阅读等等，基于这些行为特征构建多种网络关系图，可通过随机游走等方法发现网络中的影响力个体。

PageRank作为分析用户影响力的传统算法，计算用户影响力时只考虑了用户的粉丝数和关注数，并没有考虑用户本身的行为对影响力传播造成的影响。论文”基于用户行为的实时影响力计算”综合考虑了用户行为中的转发、评论、艾特数，论证了随着时间的变化用户行为符合幂率分布，以此定义了兴趣度模型，用来研究用户行为随时问的变化规律。充分将用户个体的行为融入到传统的影响力评估模型中，定义了活跃度的概念，可剔除不活跃的用户。此外，摒弃了PageRank中将影响力权值平均分配的方法，重新定义了影响力分配因子，对于信息传播过程中影响力贡献较大的用户给予更高的分配比例，使得影响力的评估更加准确。实验结果表明，文章提出的基于用户行为的实时影响力算法(User Real-time Influence)，提高了影响力计算的准确性，更加符合社交网络的特性。

### 3.3.3基于话题的个体影响力计算

在社交网络中用户在不同话题下的影响力不同，可以根据用户的关注网络和用户兴趣相似性来计算用户在每个话题上的影响力。

比如可以按照如下步骤研究微博话题的影响力并基于话题影响力计算个体影响力:1)微博网络主要由用户集合、微博集合、用户之间的关联关系以及用户与微博之间的转发、回复关系构成的，在社会网络中，将用户以及微博都看做节点，对此进行数据有效性处理;2)借鉴信息检索(Information Retrieval)中的查询扩展方法获得要研究的话题的相似话题集合;3)研究话题传播扩散的影响因子，得出话题影响力的计算公式;4)基于话题的影响力计算个体影响力。

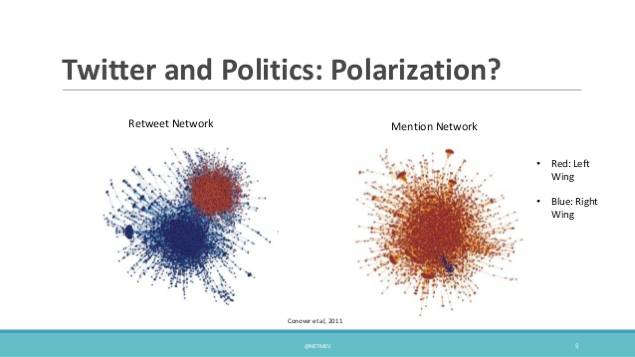
## 3.4群体聚集及影响机制分析

本部分主要介绍群体极化的概念。群体极化是指在群体决策的情境中，个体意见或决定往往会受到群体间的彼此讨论的影响，而产生一个群体性的结果。群体极化往往表现为群体内的个体不经过个人思考而同意大多数人的观点。群体极化是一个社会心理学概念，在社会学名著《乌合之众》中提到的大众心理状态就是群体极化的体现。群体极化产生的条件可概括为四点：第一，必须有激发事件出现；第二，群体内的个人能看到前人的选择；第三，群体信息缺乏；第四，群体有一定的同质性。

在在线社交网络分析中，人们通过建立分析模型和仿真来研究在线社交网络中的群体极化现象。主要的分析模型有基于博弈论和委托—代理理论的从众行为模型，基于信息瀑的群体一致性模型和基于元胞自动机群决策和行为仿真。

几乎在群体极化现象被发现的同时，诸多理论就被提出以解释此行为。其中，最为经典与完整诠释的便为社会比较理论（Social Comparison Theory）。

社会比较理论是美国社会心理学家利昂·费斯廷格（Leon Festinger, 1919-1989年）于1945年提出的构思：当每个个体在缺乏客观评判的情况下，利用他人作为比较的尺度，来进行自我评价，最终导致群体内的趋同行为。费斯廷格指出，社会比较理论解释了人们为什么要模仿传媒中的典范，其一便是为了提高个体自信心，并且成为合理自我完善的基础。



再比如微博，庞大的用户群体以及微博的直播性、共享性、互动性等特点，使微博成为网络群体极化现象的重要“发源地”之一。微博舆论的群体极化现象已屡见不鲜且有愈演愈烈之势。而网络群体极化现象所导致的负面效应，极有可能使对某一事物或某人的声讨转移到现实生活中去，爆发群体性事件，严重影响无辜人群的正常生活。

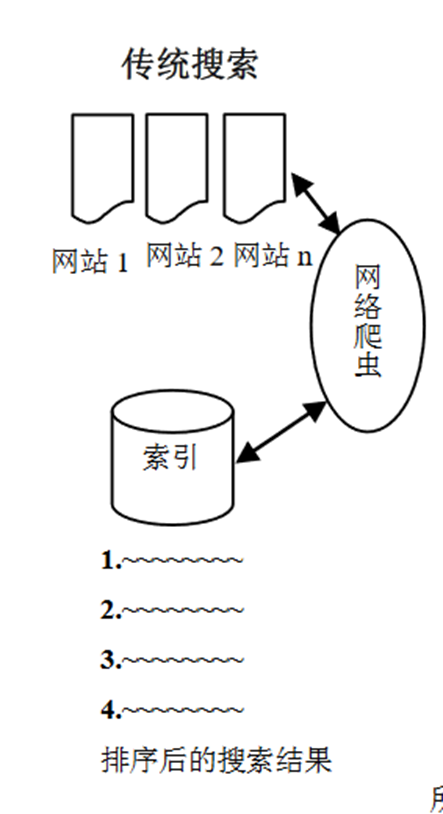
# 4内容搜索与社交信息挖掘部分

## 4.1 社交网络内容搜索

### 4.1.1 简介

内容搜索指的是给定查询，从大量信息中返回相关信息内容的过程，它是信息检索最经典的应用形式。而随着在线社交网络的蓬勃发展，传统的数据挖掘的和搜索方法已经不能完全适用于web2.0时代的社交网络。社交网络具有社交关系复杂、数据量大、动态更新、数据多模态等特点，给数据挖掘和搜索的研究来了巨大的挑战。在社交网络中，内容搜索的需求十分迫切。比如用户在社交上输入“人民币大涨”，希望返回该事件相关的信息。值得一提的是，基于社交网络的内容搜索还可以进行专家搜索或专家定位，比如输入“机器学习”可以从社交网络中返回该领域的相关专家，这是一种特定的社交网络搜索应用。因此，研究基于社交网络的内容搜索新方法成为学术界和工业界的一项新任务。

### 4.1.2 传统搜索引擎的原理

传统的搜索引擎一般采用经典的倒排索引，然后对索引结果根据相关算法进行重排序，典型的算法有PageRank、HITS等。右图显示了传统搜索的一般步骤：

传统搜索引擎通过网络爬虫，从互联网上自动搜集网页，自动访问互联网，并沿着网页中所有的

URL爬到其他网页，跟踪网络链接，通过不断地重复该动作来搜集所有爬过的网页，并将之存储到数据库中。抓取网页是整个搜索引擎工作的开始。最简单的抓取网页的路径是按照超链接的拓扑顺序进行的。整个过程就是首先爬虫会拥有一个初始的URL列表，访问到对应的网页中，分析该网页中的＜a href=...＞结构，获取新的URL，并将之插入到原有队列中。同时根据需要也可以获取＜mg src=...＞等结构，抓取相应的其它类型的文件。重复地进行这个过程，直到抓取到指定数量的网页为止。将抓取到的页面文件进行分析、分解，按照一定的算法，通常根据网页中关键词的匹配程度，出现的位置/频次与链接质量等——计算出各网页的相关度及排名等级，然后根据关联度高低，建立索引，按顺序将这些网页链接返回给用户。

### 4.1.3 基于社交网络的内容搜索

随着社交网络的发展，搜索引擎提供商也开发了基于社交网络的内容搜索功能，而且社交网络服务公司也提供了基于本平台的社交搜索引擎。例如国内的搜狗公司提供了基于微信、知乎等在线社交网络的搜索服务，以提供更加适应于社交网络的搜索技术。特别地，在社交平台搜索方面，FaceBook开发了备受欢迎的知识图谱搜索技术［64］，通过该技术可以对用户发布的图片、内容、地点等进行社会化的搜索和推荐，相比传统的搜索引擎，Twitter提供的搜索完全是基于社交网络的搜索，提供了基于内容的搜索、社交关系搜索、事件的情感搜索、用户等搜索功能、热点事件搜索等。国内的新浪微博也提供基于社交网络搜索功能，与Twitter相比新浪微博的搜索比较简单，仅提供了相关话题推荐、热点事件推荐、用户搜索、内容搜索等功能。

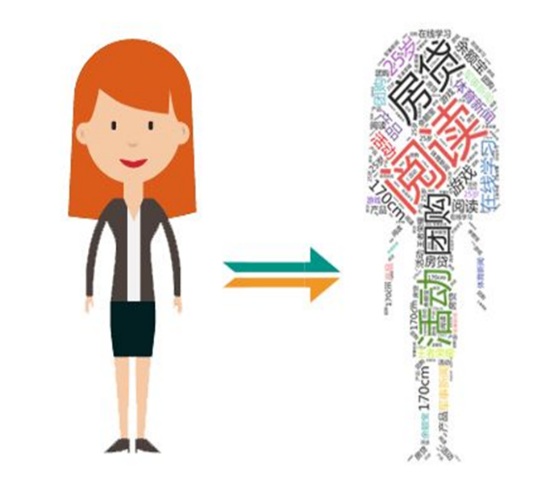
基于社交网络的内容搜索对排序算法的研究主要是对传统排序算法的改进和引入情感计算、社会影响力、用户参与度等因素，提高搜索排序算法的精度和准确性。一种算法主要是基于多特征的微博搜索排序，包括三个维度的特征：一是时间特征，即当一个重要事件发生时，用户会发布很多博客。由于大多数情况下用户搜索查询都是基于值得注意的事件，因此用户发送时间和查询时间之间的时间差越小，查询和推文就越相关；二是用户特征：是指发送该博客的用户的相关特征，例如粉丝数量，近期发送博客的频繁程度；三是文本特征，是指给定查询和目标推文之间的文本相似性。第二种算法主要是基于用户参与度的搜索结果排序算法，用户参与度主要是指根据用户回复和转发数进行评分，回复和转发数高的结果评分越高，为了降低名人微博回复和转发数对评分结果的影响，该算法还引入了基于参与用户时间窗的异常检测方法，把用户的原始参与引入评分算法，很好地解决了名人评分问题。此外还可以运用LDA推断微博的主题分布和用户的兴趣取向，通过实时TOP-K响应，实时微博查询，然后比较用户取向和微博主题分布之间的相似度判断用户的兴趣。 该方法是基于社交网络搜索算法基础上的 一个用户个性化的搜索算法，可以应用在社交网络个性化推荐中。还有的算法将把深度学习方法引入社交网络搜索，建立了基于卷积神经网络的短文本对的重排序方法，取得了较好的效果。

上述搜索和排序算法大部分都是基于传统的搜索和排序机制上的改进和扩展，尽管提高了搜索的效率，但是没有从用户角度去感知用户的搜索目的和实现智能化的社交搜索。同时搜索内容都是基于社交网络文本或事件的搜索，而缺少对社交网络中跨媒体内容的索引和搜索机制，这也是目前需要进行研究的地方。

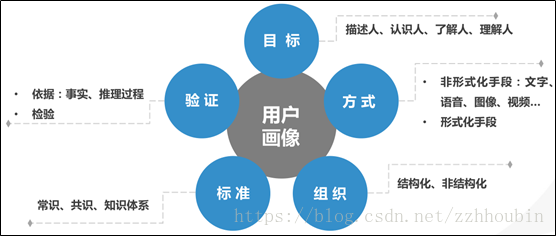
## 4.2 个体用户画像挖掘

### 4.2.1 用户画像的定义

用户画像又称用户角色，作为一种勾画目标用户、联系用户诉求与设计方向的

有效工具，用户画像在各领域得到了广泛的应用。我们在实际操作的过程中往往会以最为浅显和贴近生活的话语将用户的属性、行为与期待联结起来。作为实际用户的虚拟代表，用户画像所形成的用户角色并不是脱离产品和市场之外所构建出来的，形成的用户角色需要有代表性能代表产品的主要受众和目标群体。

用户画像是对现实世界中用户的建模，用户画像应该包含目标，方式，组织，标准，验证这5个方面。目标：指的是描述人，认识人，了解人，理解人。方式又分为非形式化手段，如使用文字、语言、图像、视频等方式描述，人形式化手段，即使用数据的方式来刻画人物的画像。组织：指的是结构化、非结构化的组织形式。标准：指的是使用常识、共识、知识体系的渐进过程来刻画人物，认识了解用户。验证：依据侧重说明了用户画像应该来源事实、经得起推理和检验。

举一个比较简单的案例就是，如果你经常购买一些玩偶玩具，那么电商网站即可根据玩具购买的情况替你打上标签“有孩子”，甚至还可以判断出你孩子大概的年龄，贴上“有5-10岁的孩子”这样更为具体的标签，而这些所有给你贴的标签统一在一起，就成了你的用户画像，因此，也可以说用户画像就是判断一个人是什么样的人

### 4.2.2 用户画像的作用

在互联网、电商领域用户画像常用来作为精准营销、推荐系统的基础性工作，其作用总体包括：

1. 精准营销：根据历史用户特征，分析产品的潜在用户和用户的潜在需求，针对特定群体，利用短信、邮件等方式进行营销。
2. 用户统计：根据用户的属性、行为特征对用户进行分类后，统计不同特征下的用户数量、分布；分析不同用户画像群体的分布特征。
3. 数据挖掘：以用户画像为基础构建推荐系统、搜索引擎、广告投放系统，提升服务精准度。
4. 服务产品：对产品进行用户画像，对产品进行受众分析，更透彻地理解用户使用产品的心理动机和行为习惯，完善产品运营，提升服务质量。
5. 行业报告&用户研究：通过用户画像分析可以了解行业动态，比如人群消费习惯、消费偏好分析、不同地域品类消费差异分析。

根据用户画像的作用可以看出，用户画像的使用场景较多，用户画像可以用来挖掘用户兴趣、偏好、人口统计学特征，主要目的是提升营销精准度、推荐匹配度，终极目的是提升产品服务，起到提升企业利润。用户画像适合于各个产品周期：从新用户的引流到潜在用户的挖掘、从老用户的培养到流失用户的回流等。总结来说，用户画像必须从实际业务场景出发，解决实际的业务问题，之所以进行用户画像，要么是获取新用户，要么是提升用户体验、或者挽回流失用户等具有明确的业务目标。另外关于用户画像数据维度的问题，并不是说数据维度越丰富越好，总之，画像维度的设计同样需要紧跟业务实际情况进行开展。

### http://bbs.itheima.com/data/attachment/forum/201808/14/144129pxhktmh1uf9f6rhz.png.thumb.jpg4.2.3 用户画像的构建流程

构建一个用户画像，包括数据源端数据收集、数据预处理、行为建模、构建用户画像。

1. 数据源的准备与分析

用户的数据收集主要包括四类，一类是网络行为数据：是指活跃人数、页面浏览量、访问时长、激活率、外部触点、社交数据等；然后是服务内行为数据：是指浏览路径、页面停留时间、访问深度、唯一页面浏览次数等；然后是内容偏好数据：是指浏览／收藏内容、评论内容、互动内容、生活形态偏好、品牌偏好等；最后是用户交易数据：贡献率、客单价、连带率、回头率、流失率等。

1. 用户分类

通过对用户静态数据的分析，可以进行基本的用户分类，比如男、女用户，城市分布、年龄段划分、用户来源划分、收入水平划分、职业化分等。静态数据获取后，需要对人群进行因子和聚类分析，不同的目的分类依据不同：如对于产品设计来说，按照使用动机或使用行为划分是最为常见的方式，而对于营销类媒体来说，依据消费形态来区分人群是最为直接的分类方式。

1. 标签体系

用户画像的标签分为三个层面，事实标签是指通过对于原始数据库的数据进行统计分析而来的，比如用户投诉次数，是基于用户一段时间内实际投诉的行为做的统计。模型标签是指以事实标签为基础，通过构建事实标签与业务问题之间的模型，进行模型分析得到的标签。预测标签是指在模型的基础上做预测，比如针对投诉倾向类型结构的变化，预测平台舆情风险指数。标签体系的建设一要便于使用，二要区分度明显。结合具体产品而言，在不同的场景下，对这两点的要求重点是不同的。最终在产品中选择哪些标签并没有明确的依据，还是需要充分了解到底是什么驱动用户使用产品。有效的标签体系，要能反应用户决定买什么、不买什么的逻辑与依据。

1. 标签属性

标签属性可以理解为针对标签进行的再标注，这一环节的工作主要目的是帮助内部理解标签赋值的来源，进而理解指标的含义。如图所示，可以总结为5种来源：

1. 固有属性：是指这些指标的赋值体现的是用户生而有之或者事实存在的，不以外界条件或者自身认知的改变而改变的属性。比如：性别、年龄、是否生育等。
2. 推导属性：由其他属性推导而来的属性，比如星座，我们可以通过用户的生日推导，比如用户的品类偏好，则可以通过日常购买来推导。
3. 行为属性：产品内外实际发生的行为被记录后形成的赋值，比如用户的登陆时间，页面停留时长等。
4. 态度属性：用户自我表达的态度和意愿。比如说我们通过一份问卷向用户询问一些问题，并形成标签，如询问用户：是否愿意结婚，是否喜欢某个品牌等。当然在大数据的需求背景下，利用问卷收集用户标签的方法效率显得过低，更多的是利用产品中相关的模块做了用户态度信息收集。
5. 测试属性：测试属性是指来自用户的态度表达，但并不是用户直接表达的内容，而是通过分析用户的表达，结构化处理后，得出的测试结论。比如，用户填答了一系列的态度问卷，推导出用户的价值观类型等。
6. 标签建模

用户画像的核心是标签的建立，用户画像标签建立的各个阶段使用的模型和算法如下图所示:



1. 用户画像验证

用户画像验证主要包括两个方面的内容，第一个是模型是否反映了现实，即逻辑上是否可靠，机器学习模型的准确率是否达到了要求，搜索的查准率是否合格等；第二个是显示是否在模型中有所体现，主要考察逻辑是否完全，机器学习算法的召回率以及搜索的查全率等。

## 4.3 社交圈挖掘

### 4.3.1 研究背景

本文所讨论的社交圈挖掘是在微博环境下，如何自动挖掘某个微博用户的社交圈子或者兴趣圈子。如果能够对于某个用户在微博上体现的社交关系进行准确的挖掘，对于很多具体应用来说都有很好的作用，比如可以更好的对用户的兴趣进行挖掘或者能够推荐用户还未关注的社交圈子成员等，或者根据其社交圈子更准确的对用户进行个性化建模，为其它基于用户个性化模型的推荐或者广告推送等提供基础服务。

举一个案例就是如用户的某个同事，你们并不是直接的好友关系，但腾讯会知道这期间的潜在关系，或自动分到同事分组并同时加上备注。或者在添加某个新好友的时候，会自动显示备注。

### 4.3.2 社交圈挖掘算法的要求

在设计微博社交圈挖掘算法的时候，通常希望它满足以下几个条件：

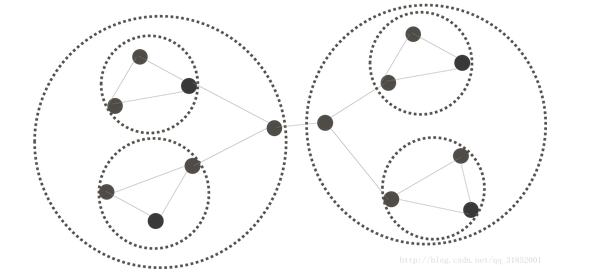
1. 对于某个微博用户A来说，可以挖掘出其所属的多种社交圈子，比如用户既有同事关系圈，也有所属的专业兴趣圈。
2. 同时对于另外一个用户B来说，可能同时属于用户A的不同社交圈子，比如B既是A的大学同学，也是A的某公司同事，那么B应该同时出现在用户A的两个不同兴趣圈里。
3. 不使用用户隐私数据，出于保护用户隐私的目的，我们希望算法只使用用户公开行为和信息，所以HipHop算法只使用了互动关系这种公众完全可见的公开信息。
4. 社交圈可解释，即可以通过简洁的方式描述社交圈子的性质或者特点，目前是通过给每个圈子打上不同的标签来进行区分。

### 4.3.3 GN算法

边介数指的是在网络中任意两个节点通过此边的最短路径的数目。在一个网络中，通过社区内部的边的最短路径相对较少（边介数小），而通过社区之间的边的最短路径数目的数目则相对较多（边介数大）。GN算法是一个基于删除边的算法，本质上是基于聚类中的分裂思想，在原理上是使用边介数作为相似度的度量方法。在GN算法中，每次都会选择边介数高的边删除。连接社区之间的边介数大，社区内部的边介数小。通过删除边来发现社区。

GN算法的步骤如下：

1. 计算每一条边的边介数；
2. 删除边介数最大的那条边；
3. 重新计算网络中剩下的边的边介数；
4. 重复上述步骤，直到网络中的任一顶点作为一个社区为止



GN算法是有效的算法，但是这是一种“单社群”发现方法，就是说，对于图中某个节点，只能属于固定的一个社群，不可能同时属于多个社群，这个与实际应用场景需求是有较大差异的，形成了该算法的局限。

# 5社交网络的应用

社交网络分析主要研究人在社会网络中的位置及其信息在整个社会网络中传播形式等内容，通过关系分析，发现网络系统背后潜在的信息及价值，辅助决策者进行决策。随着社交媒体的出现，SNA技术变得越来越成熟，越来越完善。如今，它已经被应用到各个领域。比如，在社交媒体领域，服务提供商通过社交网络分析技术分析用户间好友或关注网络，可以发现网络中具有影响力的用户，进行广告投放，或者发现网络中的群体进行特定主题的内容推荐。在电子商务领域，利用用户的消费信息，可以得到用户与购买物品间的购物网络，进而抽取用户间的共同购买网络。在共同购买网络上，利用社交网络分析方法可以发现特定用户群体，从而推荐相应的商品。

## 5.1 社交网络在公共安全领域的应用

公安领域利用网络分析进行侦查办案的历史由来已久。传统的方式是将涉案人员以图钉和连线的方式在黑板上展示出来。近年来，社交网络分析技术的发展，不仅将传统的方式数字化，利用计算机就能快速分析案情，而且使得干警能够处理和分析更大规模网络，从而提高办案的效率和质量。在实际的警务工作中，公安干警对案件的研判会利用犯罪嫌疑人既有的复杂社会关系，对犯罪嫌疑人的可能动向、嫌疑性大小进行分析。与社交网络分析相同，整个研判过程可以使用网络结构进行建模。网络中的一个节点可以表示一个人、一个公司、一个组织，而边则代表着二者之间不同的信息交互类型，例如买卖双方的毒品交易、恐怖分子之间的通信联系、歹徒与受害者之间的联系等。由此构建的社交网络图，首先通过可视化技术，既可在视觉上直观地体现出犯罪嫌疑人之间、犯罪嫌疑人与受害人之间、犯罪嫌疑人与其亲人朋友之间的联系，又将抽象的案件分析具体化、数据化;然后，利用社交网络分析技术的定量分析，能够刻画出案件中的关键人物、团伙间的协作流程，为警员的案件侦破或者复盘提供丰富多角度的信息。

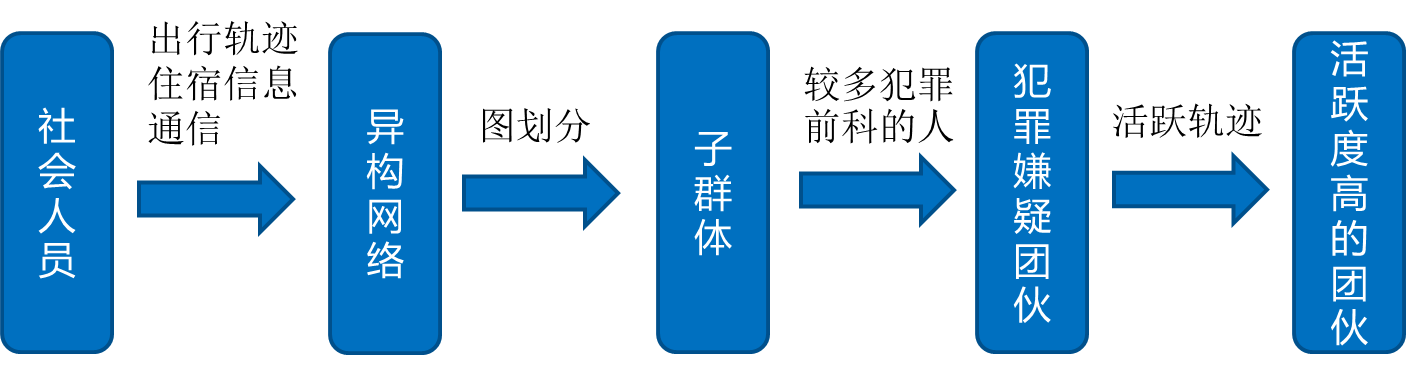
本节将通过三个具体应用场景来介绍社交网络分析技术在公安领域的应用。

### 5.1.1犯罪团伙挖掘

随着社会的发展，犯罪手段和技术也在不断地演进，呈现出团伙作案的趋势。在一个犯罪团伙内部，成员们分工明确，通过相互协作实现隐蔽作案的目的。比如，电信网络咋骗团伙内部，有人负责编撰故事，有人负责跟受骗者沟通，有人负责洗钱;在涉毒团伙中，有人负责购买原材料，有人专门制毒，有人负责贩毒，还有人进行洗钱;在贩婴团伙中，有人负责寻找婴儿，有人运输，有人找买家，有人收账。总之，团伙往往分工明确，在办案过程中仅仅打掉其中一个环节或某几个环节，效果都不尽如人意，团伙仍有可能重新组织人手继续犯罪。因此，明确团伙成员、分析清楚团伙内部成员的地位等级，将给执法办案提供极大的帮助，同时也是需要解决的一大难题。

社交网络分析为上述这一难题的解决带来了希望。前文提到，社区发现是社交网络分析的一项基本内容。而基于以往的经验了解到，犯罪团伙间都存在或多或少的联系，团伙内部的关系与团伙外部的关系有着显著的差异，比如，团伙的结构相对稀疏，但外围的人员与外界关系频繁等。根据上述经验，设置合理的模块优化目标可以利用社区发现的思想得到一个粗略的群体，然后再利用精细的分类方法，比如基于图核函数的聚类算法，从群体中分离出小团体。

举个具体的例子，在实际工作中，可以结合社会人员的出行轨迹、通信情况、住宿记录等信息建立多种关系，并形成一张异构网络。针对得到的异构网络，利用经典的图划分算法得到一系列高内聚性的群体。对于群体内部有较多(如大于3人)犯罪前科的人，则可以初步判定此群体为嫌疑犯罪团伙。针对感兴趣的犯罪团伙，再通过提取顶点特征，利用k-mean:算法对群体内顶点进行分类，形成精细的团伙。其中，顶点特征可以包括最近轨迹行为时间、通信频繁程度等。最后，根据团伙最新的活跃轨迹筛选出活跃度较高的群体，供业务人员进行研判分析。以上过程如下图。



### 5.1.2 犯罪团伙中的核心成员识别

在分析出犯罪团伙以后，需要进一步识别团伙中各个成员的角色和地位。比如处于聚簇边缘、与外界联系相对较紧密(与其他聚簇相连)的节点，可能是联络人，也可能是新人伙的人;在聚簇中心的可能是组织中的核心人物。所谓“擒贼先擒王”，首先识别出网络中的核心人员，警方再根据分析结果，选择特定的突破口，对这一组织实施打击粉碎行动。这里利用节点重要指标可以迅速定位挖掘犯罪团伙中的核心成员。

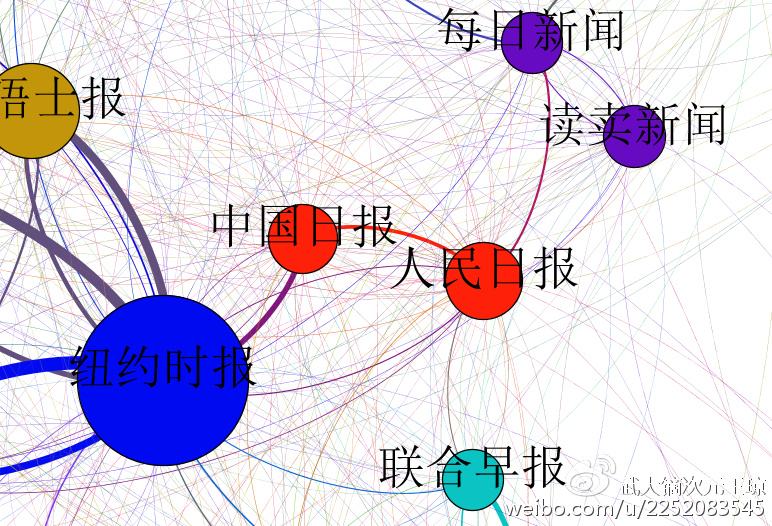
### 5.1.3 串并案挖掘

串并案是公安机关提高办案效率的一个有效途径，它是指把多个可能相关的案件一起进行研判分析。这里的相关包括同一伙犯罪成员或者是相似的作案手段等。通过对多个案件的串并分析，不仅能起到信息互补的作用，而且能解决不少冤假错案。社交网络分析有利于将同一犯罪团伙所为的案件进行串并。若独立地分析多个由同一犯罪团伙所为的各个案件，警员无法有效洞察其背后千丝万缕的联系。结合社交网络分析技术同样能给案件侦破提供线索。

以美国某州的警局针对该地区一系列便利店的抢劫案为例，介绍如何利用社交网络分析进行并案，从而抓捕了多个之前未被发现的犯罪嫌疑。他们首先把单独的案件中涉及的人、事、物映射到一张网络上，然后在调查中发现某个正在被调查的人A与其他辖区的有案底的人B存在联系。于是，他们以这两人（A和B）的姓名为种子，利用类似标签传播的算法，构造了一张未知网络，即网络中出现了多个未被关注的人员。对这个未知网络进一步分析，发现该有案底人员A又与多个青少年有频繁的联系，于是就把之前未被人视为同一伙人作案的抢劫案给串联起来了。最后，通过可视化展现的方式，对多张网络进行浏览和管理，为干警破案提供了快速有效的线索。

## 5.2 社交网络在舆情分析上的应用

互联网的诞生给人们的生活方式带来极大冲击，人们随时随地进行着信息共享与沟通交流，大量移动互联网用户创造的内容成为重要的信息来源，这也催生着越来越多的社交网络平台。移动终端的普及和4G网络的成熟运用，也使得社交网络迅速成为新的研究热点。目前的社交网络主要利用各种终端设备作为社交活动的媒介。如Facebook, Twitte:这样的社交网站，己经提供了移动社交网络的应用产品。据Facebook官方网站宣布截至2015年3月31日，它的活跃用户己达11亿1000万，而其中移动用户为7亿5100万，并且以每年23%的增速继续增长。国外专家学者对于网络舆情的相关研究主要是从网络舆情的影响因素及对网络舆情的产生、传播及演化等整个生命周期进行切入的。Savigny和Heather认为，目前网络舆情总体上是由大众媒体促进形成，并通过各种网络媒介跟踪报道相继创造出来。移动社交网络作为诞生不久的一种沟通媒介，它与传统的互联网相比对舆情有更重要的影响作用。Christen 等人对于地方政府与国家政府及相关媒体对公共网络舆情的影响因素进行实证研究。Sakaki 等人通过实时监控移动社交网络平台Twitter用户的状态更新来进行地震探测，并成功构建了一个地震探测系统。Naohiro, Matsumura等学者对日本知名的移动社交网络平台Z Channel论坛进行研究后提出了移动社交网络舆情的影响因素扩散模型; Henry M.Kim 等则建立一个舆情早期分析预警系统，并通过该系统对公共卫生进行监测和分析;此外，国外各国政府还有相应的法律法规对网络言论、信息及一些商业行为进行管理规范，如美国在9.11后颁布的《爱国者法案》和法国的《信息社会法》等。



舆情分析在互联网出现之前就被广泛应用在政府公共管理，商业竞争情报搜集等领域。在社交媒体出现之前，舆情分析主要是线下的报纸，还有线上门户网站的新闻稿件，这些信息的特点是相对专业准确，而且易于分析和管理；但随着社交媒体出现，舆情事件第一策源地已经不是人民日报、新华社这样的大媒体，而是某一个名不见经传的微博用户，一个个人微信公众号。他们的特点是信息非常新鲜，缺点是真实度较低且传播十分迅速，难以控制。所以在社交网络下的舆情分析是一门新的学问。举几个例子，几年前的和颐酒店、北京地铁骂人事件这类急性舆情事件最早就是在微博上爆出，而且在短时间内迅速传播。“北京房价”这类民生话题，也是在微信公众号逐渐发酵。当然，在新形势下的舆情应对，也已经有新的工具，通过百度“舆情分析平台”或者“舆情分析软件”可以找出一大堆。比较有名的有蚁坊、红麦、清博、知微、新榜等等。一些传统的舆情分析机构开始转型做“大数据”的舆情分析，也有近年来完全基于社交媒体的舆情平台，比如基于微信的新榜和基于微博的知微 。除此之外，BAT 等大型平台有自己舆情分析工具，可以私人订制，也有开放的指数（百度指数、微信指数）。

## 5.3 社交网络在个性化推荐服务的应用

个性化推荐服务最早应用在商业领域中，由于工业的快速发展，产品的质量和数量都在不断提升，在电子商务领域中，各大服务商根据用户的属性，或者用户之前的购买历史记录，通过个性化推荐服务来预测用户可能想要购买的产品，从而在无形中提升了广告投放的转化率并提升了服务商的收益。这样的个性化推荐服务不仅提升了用户体验和产品的转化率，还在发展中无形的带动了个性化推荐服务自身的发展与商业应用，使个性化推荐服务变得更加成熟。

目前，个性化推荐技术也应用于社交网络领域之中，这些推荐技术根据推荐的信息不同可以分为两大类，第一种是为用户推荐可能感兴趣的人。

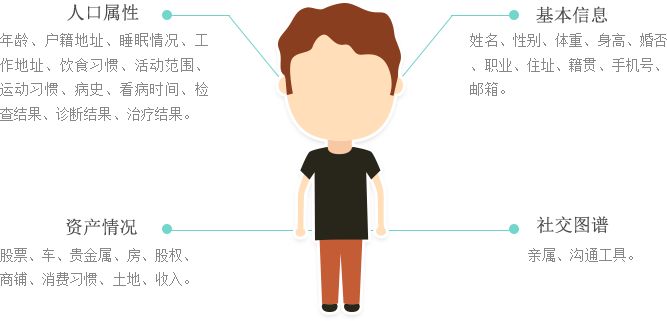
Liben-Nowell D等人在《The link-prediction problem for social networks》一书中认为，社交网络的好友推荐算法就是一种“链路预测”问题，也就是研究用户和用户之间的拓扑结构问题。但是在真实的社交网络中，相近用户的兴趣爱好也不一定相同，为解决传统社交网络个性化推荐算法中好友推荐依靠好友关系，而不是共同兴趣的问题，各位专家学者提出了自己的见解。YANG等人使用社会标签，将好友推荐问题转化为链路预测问题，将拓扑结构网络和用户兴趣网络进行了有机结合，从而进行社交网络中的个性化好友推荐;GUO等人提出了一种新的可视化系统，该系统能够与用户交互，帮助用户探索和寻找存在共同兴趣的好友，并且不仅仅只利用了社交网络的拓扑结构，还同时利用了活动数据的语义结构，之后产生具有层次结构的社会性标签，系统通过这些标签来表现用户的兴趣，并将用户和用户之间的相似度多角度的融合在其中，为用户在社交网络中寻找到可能感兴趣的好友提供了支持;Hannon J等人利用了Twitte:社交网站中的用户结构，并结合了用户信息，提出了一种新的个性化推荐算法，该算法可以在不同范围内进行分析，较为有效地提升了推荐系统推荐好友的质量，提升了算法的准确率;王帅等人提出了一种混合推荐算法，该方法在利用用户之间的好友关系来构建显性社交网络的同时，使用用户标签来构建隐性社交网络;吴不晓等人提出了另一种解决方案FRBT算法，该方法使用TF-IDF(词频一逆向文件频率)对用户标签进行聚类，将相似的标签聚类成话题，并在话题的基础上进行用户相似度计算，在计算的同时融合了话题相似度和物品相似度来产生相似用户列表，从而为用户产生推荐;汤颖等人在对用户进行推荐的时候结合了用户评分信息，首先使用余弦相似度进行用户之间关注关系的相似度，接着建立概率模型，得出用户聚类相似度矩阵，依据该矩阵推算出用户兴趣相似度，最后将两者结合得到推荐列表进行推荐。

第二种是为用户推荐可能感兴趣的文本内容，如高明等人为解决微博社交网络中推荐实时性差的问题，提出了一种实时个性化推荐系统，该系统使用LDA主题模型推断用户兴趣和微博主题分布，从而为用户产生推荐;肖晓丽等人为解决社交网络个性化推荐系统中矩阵稀疏问题，提出了一种基于用户兴趣和社交信任的聚类推荐算法，该算法首先根据用户评分，将用户进行聚类，并建立基于用户兴趣相近的邻居集合，其次引入了信用机制，将社交网络转化为信任网络，最后通过加权的方式将两者结合，从而为用户产生推荐。

由于社交网络中用户数量庞大，以上推荐系统都需要依据一定的标准将用户进行聚类，形成社区之后，在社区中进行用户相似度的计算，来产生推荐列表，进行推荐。使用了MovieLens经典数据集进行实验，该数据集中包括了943个用户历史数据。根据文章的实验结果，当阂值为0.8时，最大社区人数为42人，最小社区人数仅为5人。此时如果想为最小社区中的某位用户产生数量为10的推荐列表显然不可能，因此影响了推荐算法的质量。为了解决该问题，文中提出了两种解决方案。第一种方法是降低阂值，放宽用户融入社区的条件，进而增加每个社区中用户的数量，从而使社区中的人数能够达到所需人数。第二种方法也是目前采用较多的方法，即考虑用户的重叠兴趣，进行重叠社区的划分。在现实生活中，人的兴趣是多种多样的，这类解决方法通常将一个用户划分到多个社区中，如果用户在某类兴趣社区中的相似用户较少，则在其他重叠社区中为该用户进行推荐。

## 5.4 社交网络在用户画像上的应用

在大数据时代，社交网络数据是大数据的主要来源之一。随着社交网络的蓬勃发展，社交网络吸引了大量的用户，社交网络已经成为了当前时代人与人交往的重要途径之一。正是由于社交网络的普遍流行，使得在社交网络中积累了大量的个人用户信息，这些信息主要包括用户的属性信息(例如住址、学校、性别、年龄、工作背景、兴趣偏好等等)。这些用户的属性信息在工业界扮演着重要的角色，并被用在很多智能化的场景之中。以新闻推荐为例，一个从事数据挖掘工作的算法工程师可能对机器学习技术，数据分析工具(例如python)更感兴趣，对娱乐八卦、时尚前沿可能就并不关心。在这些广泛的应用中，很多算法要行之有效，必须立足于用户的私有属性。然而，一个不可避免的现实问题是，社交网络中的用户属性信息往往不够完整。有一些用户是出于隐私的考虑，会避免填写一些敏感信息；有一些用户则是懒于填写，认为填写这些用户属性信息是一件很麻烦的事情。对缺失的用户属性信息做有效的预测和推断是一个亚待解决的问题，也是一个重大的挑战。利用机器学习的方法对已有的数据建模进而对缺失的用户属性信息进行精确推断的方法统称为用户画像分析。用户画像分析吸引了大量的学者进行研究，由于其所蕴含的巨大的商业价值和重大的学术价值。



用户画像是指最能代表一个人的一系列特征(例如年龄、性别、居住地、作地点等等)、用户画像分析是指推断用户特征的过程、手段和方法。因此，用户画像分析是对某个个人贴上精确的标签的有效手段。在用户画像分析中，所用到的数据一般包括两大类:一类是社交网络的结构信息;一类是社交网络中和用户相关的信息，这些信息主要包括用户的部分已知的属性信息。社交网络用户画像分析就是在社交网络中所有用户社交关系已知，部分用户的已知的属性信息可以当做用户的标签(Label)的情况下，然后去推断未知属性的用户的属性信息。

用户画像问题通常被当做一个标记预测的问题，即在已知部分用户的属性信息(Label)以及用户的社交关系。去预测未知属性用户(Unlabeled)的属性信息。目前在国内外的研究论文中，用户画像分析方法根据其所使用的数据类型主要可以分为两大类:第一类研究方法重点关注如何通过分析用户自产生的数据去显式的估计用户画像信息:第二类研究方法重点关注如何通过分析用户的社交关系，从而建立合理的用户关系，进而去构建用户画像模型。

第一类方法所涉及的用户自产生的数据一般包括用户的文本信息，图像信息以及互联网收集的用户的其他使用信息(例如GPS信息、浏览日志、搜索日志等等)。由于文本信息和图像信息本身蕴含了丰富的语义信息，所以可以很自然的将用户画像分析问题转换为文本分析问题或者图像分析问题。由于在自然语言处理领域和图像分析领域已经有了比较成熟完整的技术体系，所以利用这些领域的经典学习方法进行相应的改进和修改使之适用于用户画像场景下的用户自产生的多媒体数据成了很多用户画像分析的主要思路之一。除此以外，我们也很容易从用户的其他使用信息中提取特征，然后结合经典的分类或者回归模型进行用户画像分析。

考虑到用户自产生的数据本身传递了大量的可供挖掘的有用信息，并且很多信息具有丰富的语义表达能力、因此在研究过程中，这类方法往往不会过多考虑社交信息。这类方法有很多的例子，例如基于博客数据的工作，研究人员在这个土作中通过对博客中的文本数据使用了一些自然语言处理的方法进行了语义分析获得文本特征，通过对博客用户的使用位置以及使用时间、语言等一些列使用信息做了详细的统计分析获得了统计特征。最终将统计特征以及文本特征结合在一起、结合经典的回归模型和分类模型对用户的年龄做了预测。又例如基于推特数据的工作、研究人员在这个工作中通过对推特数据进行预处理，构造了一个适用于弱监督学习的数据集(所谓的弱监督学习数据集是指没有明确的标记信息，但是这个数据集中包含数据和标签之间的某种关系，例如，我的家在陕西，这句话包含了我和陕西的某种关系.但是没有对“我”进行人为的打上确定的标签)，然后通过直接对这个数据集进行建模分析，最终实现了用户的工作类别、居住地、婚否等信息的预测。另一个基于推特数据的工作、研究人员在这个工作试图将用户的属性特征、文本特征和用户的行为特征收集到一起，然后使用梯度决策树(GBDT)对用户的政治倾向做回归分析。这类方法对用户自产生的数据有较大的依赖，需要对数据集进行标注或者进行比较精巧的构造。一方面，这类方法会在不同程度上受到标注数据中的误标数据的影响;另一方面，在特定的数据集上精巧设计的基于用户自产生数据的特征很难在不同数据集上进行迁移。

不同于第一类方法，第二类方法关注的重点在于社交关系本身。基于物以类聚，人与群分这一普遍的社会现象。在社交网络分析领域有一些常见的假设、例如有相似属性的用户更容易成为好友，有好友关系的用户之间往往具有相似的属性。正是有了这些基本的假设，用户画像研究人员试图通过理解社交关系本身，去挖掘人与人之间的潜在联系，解释好友关系形成的潜在因素(例如，相似的教育背景，相近的地理位置，相似的兴趣爱好，同一所学校等等)，进而根据这些潜在的联系做用户画像分析。

在这类方法依赖的主流的社交网络分技术手段中，社区发现算法(Commu-pity detection)是最常用的算法之一。社区发现(Community detection)是社交网络分析中的一个重要问题，旨在揭示网络中潜在的社区结构、是理解网络性质的重要途径和方法。在这个领域中有很多现有的算法，其中不得不提的一类算法是凝聚类算法((Agglomeration Algorithms)，这类算法是通过构建一个将小的类簇合并为大的类簇的层次树得到的多层聚类算法。其中非常著名的是提出的CNM算法，这个算法是基于NFGA困ewman Fast Greedy Algorithm)，算法过程可以描述为从单个点组成的类簇开始，根据极大化Modularity合并类簇，最终迭代生成一棵最优的合并树。由于这个算法的有效性，之后的很多研究人员对这个算法做了很多扩展研究及应用。除此以外，基于标签传播的社区发现算法在这个领域也占有重要的位置、LPA算法采用了一种异步更新策略，每个节点是否加入某个社区，取决于其邻居节点的选择。HANP算法基于间隔越远的节点相互影响力越弱以及节点偏好这两个因素、建立了一个更加稳定和鲁棒的算法。

在有了社交网络分析方面的理论支撑之后，许多研究学者基于这些算法在不同的用户画像场景之下设计了行之有效的算法。例如，基于某个知名大学的学生在Facebook上的社交信息。去推断学生所在的学院。文章作者假定在同一个学院的学生更容易成为好友并且会形成一个社区(Community)，在此假定之下，对CNM算法中的Modularity重新做了定义使之更加适合数据和问题场景，并利用改进后的算法对实验数据集中的学生用户做社区发现。基于社区发现的结果，根据社区中多数学生的投票结果去推断未知学院信息的学生属于哪个学院。

**参考文献**

[1]Zhao, Yuxin, et al. "Acellular learning automata based algorithm for detecting community structure incomplex networks." Neurocomputing 151 (2015): 1216-1226.

[2] Du, Jingfei, Jianyang Lai,and Chuan Shi. "Multi-Objective Optimization for Overlapping CommunityDetection." International Conference on Advanced Data Mining andApplications. Springer, Berlin, Heidelberg, 2013.

[3] Newman, Mark EJ, andElizabeth A. Leicht. "Mixture models and exploratory analysis innetworks." Proceedings of the National Academy of Sciences104.23(2007): 9564-9569.

[4] Ren,Wei, et al. "Simple probabilistic algorithm for detecting communitystructure." Physical Review E 79.3 (2009): 036111.

[5] Palla, Gergely, et al."Uncovering the overlapping community structure of complex networks innature and society." arXiv preprint physics/0506133(2005).

[6] Huang,Jianbin, et al. "Density-based shrinkage for revealing hierarchical andoverlapping community structure in networks." Physica A:Statistical Mechanics and its Applications 390.11 (2011): 2160-2171.

[7] Greene, Derek, Donal Doyle, and PadraigCunningham. "Tracking the evolution of communities in dynamic socialnetworks." Advances in social networks analysis and mining (ASONAM), 2010international conference on. IEEE, 2010.

[8] The Relationship Between the Use of Virtual Social Networks with Academic Achievement and Students’ Confidence in Interpersonal Relations at Birjand University of Medical Sciences

[9] Social Network Analysis and Big Data tools applied to the Systemic Risk supervision