# **基于交互行为的微博用户影响力的研究与应用**

# 绪论

## 课题背景和意义

### 研究背景

随着网络的普及和Web2.0技术的迅猛发展，不同种类的在线社交网站如雨后春夢般地涌现出来。几个著名的社交网站例如等己经成功挤身全球访问量最高的网站之列，国内的社交网站如人人网，新浪微博，开心网等也不甘示弱地正在影响着中国网民的日常生活。随着上网移动终端（如智能手机，平板电脑等）的普遍，几乎每个移动终端上都会有这些社交网站的应用，这样更是助推了这种在线社交网络的多样性和及时性。这些在线社会网络的成员数目庞大，它们的出现对传统社会网络中的影响最大化算法，包括传播模型均提出了巨大的挑战。由最初营销领域的“口碑效应”和“病毒式营销策略”的推广方式，人们发现可以在人群中选取具有代表性的节点子集，由它们可以引起更大的级联影响，逐步产生出了影响力最大化问题。影响力最大化问题被引入社会网络研究领域后，成为近年来的一大研究热点。

社交网络网络成为了人们主要的沟通手段，大量的社会媒体涌入网络，在国外，例如 Facebook，Twitter 等社交网站受到了广大网络用户的欢迎。国内代表性的有新浪微博、腾讯微博、天涯等等。2009年8月新浪微博成立，2010年国内几家知名的互联网公司，腾讯、搜狐、网易也都推出了自己的微博服务并进行大力的宣传。根据新浪微博发布的《2012 年新浪微博用户发展报告》显示新浪微博目前注册用户已达到 3.68 亿。微博也是一种包含了许多实体和关联的复杂网络，这种复杂网络在信息交流和传播中扮演着重要的角色。但是，微博具有其独特的性质，它采用一种叫做关注的动作建立起整个网络。在微博中，一个用户关注另一个用户并不需要经过其批准，同时，关注的发起者被形象化地称为“粉丝”。一旦两个用户的关注关系被建立，关注者就可以看到被关注者所发布的一切消息。微博中的消息不仅包含文本，甚至可以包含图片、超链接、视频等其他信息。

影响最大化问题的研究有着十分重要的现实意义，例如各行各业中对创新观念和新产品的推广以及政府相关政策的推广等，需要找到网络中影响力大的一些种子节点进行推广。同时，还有一些需要找出影响力较大的种子节点进行抑制影响的传播，例如抑制传染病在人群中的流行，遏制谣言在社会中的扩散等。以微博为代表的社交网络已经成为当前最重要的媒体之一．研究用户的行为以及信息的传播规律将有助于网络公司对每个用户的偏好进行更为准确地把握，并将其可能感兴趣的话题信息、其他用户或者用户社群推荐给用户。因此，非常有必要对影响力最大化问题进行研究，从而能更有效地解决实际的问题。目前研究的目标主要集中在如何扩大最终影响范围和降低算法的吋间复杂度这两个方面。

### 课题意义

近十年来，有关影响力的研究一直是社会网络领域研究的热点，无论是定性分析还是定量度量都取了一定的研究成果。社会网络上的“事物”是时刻在传播的，有时我们需要去促进事物的传播或者抑制事物的传播。如 “病毒式营销”中我们要选择最有影响力的个体来促进商品的传播，从而将商品的影响力达到最大。但是计算机病毒在计算机网络上的蔓延、传染病在人群中的流行、谣言在社会中的扩散等不良传播,我们要找出传播源头和其中影响力最大的传播种子然后将这些传播种子去除掉从而扼制他们的这种快速传播势头。实际上他们都是一种服从某种规律的网络上的传播行为,通过分析这种现象找出他们之间的传播规律,快速的找出影响力最大化的传播种子,对于影响事物的传播有重要的意义。影响最大化问题在市场营销,广告发布,病毒防治,舆情预警以及社会安定等方面有十分重要的应用。这个问题在社会科学中已有很长的研究历史。近年来,随着博客、论坛、网站等大规模的发展和应用,以及市场营销的应用需求,影响最大化已然成为社会网络研究的一个热点。

影响力最大化问题已被证明是一个NP问题,也就意味着我们只能够使用启发式的算法来近似最优结果。基于节点度数或节点中心性等是社会网络分析中常见的启发式方法,在某些领域这些简单的启发式方法效果很好。但由于影响最大化问题的特殊性,这些启发式的方法在所要解决的问题上效果并不好,甚至比较差。那么如何定义一种行之有效的启发式策略,并在此基础上如何将启发式策略发挥的更好是我们当前应该关注的焦点。

## 国内外研究现状

### 影响力度量

社交影响力的定义具有明显的因果性，而人们的思想、行为等产生变化的原因则是不胜枚举且因人而异，社交影响力只是其中之一。这就给社交影响力的建模和度量带来了很大挑战，同时也是造成社交影响力模型众多的重要原因。同样，在线社交网络中的影响力也与很多因素相关，目前大部分研究工作都是针对社交网络结构及其上的交互信息和用户行为特征进行量化和分析的，因此可以把能对信息传播过程或他人行为产生影响的个体视为具有社交影响力．

传统的社会网络利用连接结构来识别具有高中心性的节点，具体的衡量指标包括度中心性、介数中心性、紧密度中心性和特征向量中心性。通过对这些指标的计算，侧面反映了某一个节点的重要性。相对而言，该方法具有一定局限性，它的时间复杂度往往较高，不利于大规模网络的分析。

2010年Meeyoung Cha等人在对影响力的三种度量方法：节点入度、转发行为和提及行为进行了深度的比较，使用Spearmans相关系数衡量两种算法的相关性，分别比较了所有的用户、前10%和前1%的用户在三种算法上的到结果的相关性，总结得出具有高入度值，代表用户的受欢迎程度，在转发微博和提及其他用户方面，并不一定是具有重要影响力的。此外，他们还进一步得出影响力的增加并不是自发或者偶然获得的，而是需要用户保持足够的个人参与，比如将所发微博限制在同一个主题上。

2010年《Nature》物理版上，Kitsak等人首次提出了节点重要性依赖于其在整个网络中的位置的思想, 并且利用 K-核分解获得了节点重要性排序指标 (k-shell),该指标时间复杂度低, 适用于大型网络, 而且比度、介数更能准确识别在疾病传播中最有影响力的节点。

2011年Linyuan Lü提出LeaderRank 算法，在已有节点外另加一个节点(ground node)，并且将它与已有的所有节点双向连接，于是得到N+1个节点的网络，这个新的网络是一个强连通的网络，再按照原始的PageRank算法计算得到原来N个节点的“重要性”排序。于是由于加入ground node后的图是一个连通图，解决的PageRank的排序的不唯一性问题。

2012年Zeng等人考虑节点的 Ks 信息和经过Ks分解后被移除节点的信息, 提出了混合度分解方法MDD。基本思路是不仅考虑移除后剩余节点的信息，而且考虑被移除节点的度的信息，解决了在某个ks值范围内有大量相同ks值的情况。

2013年任卓明等提出了基于最小 K-核节点邻居集合中最大Ks值的深度指标H(i), 该指标依靠最小K-核节点与网络的其他层级节点的连接关系,判断最小K-核节点的重要性。

2013年Liu 等综合考虑目标节点自身 K-核的信息和与网络最大 K-核的距离, 提出了新的度量节点重要性的指标.。该指标解决了 Ks 指标赋予网络中大量节点相同的值导致其无法准确衡量其节点重要性的缺陷。

### 影响力最大化

2001年Richardson和Domingos第一次将影响力最大化问题引入到社会网络领域，他们将该问题定义为在网络中找出一个小的初始种子节点集合，使得从这些节点出发传播信息可以影响到最大范围的人群。从此，越来越多的学者致力于该问题的研究，并提出了许多算法。

Kempe等将影响力最大化问题定义为离散优化问题，并证明其是NP难解的。与此同时，他们还提出一个贪心算法，该算法能够取得最优解63%的近似度。贪心算法的基本思想是每次选择一个节点进入种子节点集合当中，该节点使影响力增量最大。尽管该算法能够取得对最优解较高的近似，但是运行时间过长使得它不适用于大规模网络。

为了提高效率，Leskovec提出了一个优化的贪心算法CELF。该算法利用影响力最大化目标函数的子模特性来减少每次需要计算的节点数，在达到与贪心算法相同准确度的前提下，大大减少了运行时间。实验验证，其运行速度比贪心算法提升了700倍。尽管如此，算法在拥有数以万计节点的网络中找出大小为的种子节点集合仍需要数小时。

随后，提出了W.Chen两种更快的贪心算法，分别是NewGreedy和MixedGreedy。NewGreedy算法是先去除原始图中对影响力传播没有作用的边，从而得到一个规模较小的图，然后在新图上进行影响力的研究。MixedGreedy算法是NewGreedy算法与算法的结合，其在第一轮使用NewGreedy算法找出种子节点，剩下轮次使用CELF算法。实验证明，MixedGreedy算法比CELF算法更高效。

2010年Y.wang提出CGA算法。该算法是基于社区的贪心算法，首先将网络化分为一系列社区，然后每一轮都利用动态规划的方法选择候选种子节点所在社区，在确定了社区之后就在该社区内利用贪心算法选择节点。以上方法均是贪心的或者基于贪心的，它们的共同特点是运行时间相对过长。为了解决这个问题，许多学者提出了启发式算法。Kimura和Saito提出了基于最短路径的影响力模型，并给出了在该模型下计算影响力传播的有效算法。然而，这些算法并不能被广泛使用，因为它们基于统一独立级联模型，在该模型中，各条边的传播概率都是相同的，这在现实中是不可能的。DegreeDiscount算法认为影响力的传播与节点度数有直接关系，其运行结果和时间效率方面都很不错，但是它也基于统一独立级联模型。

2010年，W.chen证明在独立级联模型的基础上计算影响力传播是问题，由此他们提出最大影响力路径和MIA模型。他们的实验结果证明MIA模型确实可以取得较快的速度，但是由于该模型忽略了许多可能的传播路径，因此，传播过程被大大简化，与实际过程不同。当然，计算节点的度中心性、介数中心性、紧密度中心性等也都属于启发式算法的范畴。对于启发式算法而言，它们确实能够缩短运行时间，并适用于大规模网络，但是却损失了准确度。

2011年兰如钦提出了CPWM算法，把节点的度数和节点指向的节点度数综合起来作为节点影响力的考虑因素,并且降低的贪心算法的复杂度。提出一个基于微博网络的消息影响力最大化算法构思，将微博用户按话题的主题划分用户，然后动态分配给每个话题的主题不同数量的初始节点数，最后通过算法找到每个话题的初始集合。解决了贪心算法时间复杂度高的问题、算法选择初始节点可能出现邻居重叠的问题。但是对于按照话题划分用户就不能够达到很好的消息影响力最大化效果。

2011年田家堂等利用线性阈值模型的“影响力积累”特性,提出了一个该模型下影响最大化算法的框架,并在此框架基础上给出一个新的算法HPG。HPG算法相比于KK算法影响范围和时间复杂度都比较好。但是在该启发阶段并没有考虑每个节点阈值不同的情况。会出现若一个未激活节点的阈值大于已激活节点的阈值，未激活节点的潜在价值就大于已激活节点的潜在价值。显然是不太合理的。

2012年王铁彤等利用线性阈值模型提出了一种基于节点激活阈值的启发式算法．它综合考虑了节点之间的影响力和节点的激活阈值，根据每个节点在激活过程中动态变化的阈值来计算PIN值，启发过程中，每一次都选取PIN最大的节点作为种子节点进行激活，贪心阶段中再贪心地挑选那些具有最大影响范围增量的节点作为种子节点。相比于HPG算法，该算法相对于HPG在相同的启发因子c下具有很好的激活范围以及非常低的复杂度。

2013年于淼提出一种基于用户质量的消息影响力评估算法Bias Rank 。主要考虑参与消息传播的用户的质量，并在评估过程中加入去重机制、惩罚机制，综合地对消息进行评估。最后通过实验证明该方法能够有效的屏蔽虚假的广告消息和机器人用户为消息带来的不真实的消息影响力，更加准确的评估消息影响力。算法不仅考虑转发用户的质量还考虑交互图中边的质量，即用户的影响力是根据t 时间段内，用户对所有转发边的转发比例确定的。即被该转发次数高的用户分得的影响力比转发次数少的用户分得的多。用户活跃度越高其分配出的影响力也就越多。

2015年曹玖新提出一种基于k-核的社会网络影响最大化算法，主要思想是若节点u被选为种子节点，则与u距离小于等于d(d为自定义参数)的所有节点标识为覆盖状态，被标记为覆盖状态的节点不能被选为种子节点，每轮选择核数最大或在核数相等的情况下选择度数最大且未被覆盖的节点作为种子节点。实验结果表明：在较大传播概率下CCA算法的传播范围优于其他启发式算法，且时间复杂度非常低，并随着种子节点数的增加，影响范围保持增长势头。

## 论文的主要研究内容

我们发现针对社会网络的影响力研究已经取得了许多优秀的研究成果，但是还有一些空间值得去探讨。在影响力度量方面，虽然基于用户交互行为与网络结构的方法已经存在，但是它们往往只是从行为的角度去分析问题，而忽略了交互行为与两个用户相关，如何站在用户双方的角度去分析问题值得我们进一步探索。在影响力最大化问题上，基于贪心的算法准确性较高，但是运行时间相对过长，而启发式算法虽然所需时间较短，但是准确度却往往难以令人满意。如何结合贪心法与启发式算法的优点，找到一个方法既能给出不错的准确度，又有较高的效率，也是值得我们去研究的主题。

本文的研究工作是在前人研究的基础上,针对目前影响力最大化算法中存在的一些不足提出求解影响力最大化问题的算法,以希望能够降低求解该问题的时间复杂度和提高最终被影响的节点数目。针对该研究目标,主要做以下几个方面的工作:

1.了解该课题的研究背景和意义,并详细介绍了该课题的研究现状,针对目前研究现状存在的问题展开本文的研究工作。

2.掌握相关理论知识,尤其对求解影响力最大化问题所需要的独立级联模型和线性闭值模型这两种传播模型进行详细的研究,这些相关的基础理论研究为后续的研究工作奠定了基础。

3.研究新浪微博，使用Python爬取领域下的数据，对这些真实数据进行分析。

4.研究影响力度量算法,基于节点交互行为和基于K-核的MDD的算法。并详细分析了这些算法存在的不足，提出一种新型的影响力度量算法。

5.在真实网络网络上进行了实验,从得出的实验结果来验证提出算法的有效性,并对实验结果进行详细的分析。

6.微博用户影响力评价系统的设计与实现。

## 论文组织结构

本文针对日益发展的微博网络，围绕微博网络中的消息影响力评估以及消息影响力最大化算法展开研究，论文的组织结构如下：

第1章为绪论，详细的阐述了课题的背景与意义，并介绍了论文的研究内容和组织结构。

第2章介绍了社会网络中与用户影响力评估以及用户影响力最大化相关的研究，并分析了相关研究的特点与不足。

第3章

第4章。

第5章。

第6章

# 微博与社会网络相关理论

## 社会网络

### 社会网络的特点

社会网络分析（Social Network Analysis，SNA）是对社会网络中行动者之间的关系进行量化研究，研究其中的社会结构与社会关系等。根据研究角度的不同，社会网络分析可分为两种基本视角：关系取向(relational approach)和位置取向（positional approach)。关系取向通过社会联结本身——如强度、密集性、对称性、规模大小等来说明特定的行为和过程，这个研究角度重点关注行动者之间的社会性关联关系，。这种观点认为那些强关系、集中的以及相对孤立的社会网络可以促进集体认同；位置取向则重点关注行动者之间的社会关系模式化，这些社会关系的特点在于在结构上它们处于平等地位。它研究的是两个或者以上的个体和第三方之间的关系所折射出来的社会网络结构，注重使用网络结构等效来理解人类行为。

社会网络分析中重要的理论有“六度分割理论”和“150定律”。

### 社会网络的分析方法

自上个世界二三十年代产生以来，社会网络分析法已经发展了多半个世纪，已经成为了一套完善的理论。对网络结构的分析就是对社会网络的关系结构与属性进行分析，它研究的对象就是行动者（Action）之间的关系模型。

社会网络分析主要有两个流派，整体网模型和自我中心网模型。

为了分析网络结构以及研究社会群体中角色的关系结构，整体网络结构模型采用图论为分析工具，使用社会网络结构关系图表示社会小群体中错综复杂的人际关系，如随机网络、图论、PERT图、决策树等。后来，它为了能够研究群体数量更为庞大、网络关系更为复杂的情况以及能够全面研究一定范围内的人际关系，它引入了社会计量学、数学中的矩阵方法，能够定量的、全面的描述整个网络关系结构。目前，整体网络分析重点在于研究网络结构中个体之间间接或者直接的关系方式和网络结构时间的特性。其中主要定义有：簇（Clusters）、桥（Bridges）、紧密性、中心性、中距性等；侧重研究不同社会地位的网络成员的明星（Stars）、联络人（Liaisons）、孤立者（Isolates）、结合体（Coalitions）、小集团（Cliques）等。

自我中心网络（Ego-centric Networks，EN）分析，主要通过研究个体的人际网络是如何影响到个体行为，挖掘人际关系网络和社会团体网络形成的原因，学术渊源主要来自于英语人类学院的社区研究。目前，自我中心网络分析集中在新经济社会学的研究上，并且逐渐延伸到社会研究的各个领域，如社会阶级、社区、人口流动、社会变迁等。在此使用的主要概念有：网络的密度，网络的结构大小以及网络结构的多元性，人际关系中的强弱联系情况。从数据采集与数据处理方面来看，自我中心网络分析主要有以下几种方法：互动方法、角色关系和情感方法以及社会交换法。

### 社会网络的指标

汪小帆等人详细介绍了复杂网络的相关理论，认为社会网络分析的指标可以从节点的度、密度、中心性、凝聚子群、平均路径长度、聚类系数六个方面展开。

## 微博相关的社交网络

### 微博类社交网络

对微博研究内容的综述中，指出已经有很多学者对微博网络的结构特征进行分析，发现微博网络同样也具有无标度特性和小世界效应，即微博网络是一种复杂网络。因此，在我们获取微博用户数据时，如果采集到的数据量是足够的，微博用户网络也会呈现小世界效应和无标度特性。

### 新浪微博

新浪微博是一个由新浪网推出，提供微型博客服务类的社交网站。用户可以通过网页、WAP页面、手机客户端、手机短信、彩信发布消息或上传图片。新浪可以把微博理解为“微型博客”或者“一句话博客”。

用户可以将看到的、听到的、想到的事情写成一句话，或发一张图片，通过电脑或者手机随时随地分享给朋友，一起分享、讨论；还可以关注朋友，即时看到朋友们发布的信息。用户可以像[博客](http://baike.baidu.com/view/1509.htm" \t "_blank)、聊天工具一样发布内容。用户可以把自己喜欢的内容一键转发到自己的微博（转发功能是对twitter RT功能的改良，保留原帖，避免在传播过程中被篡改）转发时还可以加上自己的评论。用户可以对自己喜欢的用户进行关注，成为这个用户的关注者（即“粉丝”）。用户可以对任何一条微博进行评论。（这是基于中国用户习惯而设置的特殊功能，之后Yahoo Meme 和Google Buzz也都有了评论功能）。用户可以用两个#号之间，插入某一话题。像这像，#某一话题XXX#，则发出的微博，可以点击这个，自动搜索微博上所有的包含有“某一话题XXX”的相关微博。可以展开讨论，实现信息的聚合。私信功能：用户可以点击私信，给新浪微博上任意的一个开放了私信端口的用户发送私信，这条私信将只被对方看到。实现私密的交流。

微博的内容每条不能超过140个字符，仅两条中文短信的长度，可以三言两语，现场记录、也可以发发感慨，晒晒心情。用户可以通过互联网、客户端、手机短信彩信、WAP等多种手段，随时随地地发布信息和接受信息。用户发布一条信息，他的所有粉丝能同步看到，还可以一键转发给自己的粉丝，实现裂变传播。用户可以通过搜索找到其他微博用户在几秒前发布的信息，比传统搜索引擎的搜索结果更有时效性，更鲜活。“分享到新浪微博”的按钮被添加到了百度百科词条的下面，用户可以直接分享词条到新浪微博。

## 影响力评价方法

### 基于网络局部属性的最大度算法

基于网络的局部属性度量方法主要考虑节点自身信息和其邻居信息，计算简单，时间复杂度低。节点的出度和入度可以衡量社交网络中与用户影响力相关的指标（如好友数、推荐数、跟帖数等），在一定程度上可表示节点的影响力大小，而它们的方向可以表示用户影响力或者信息的传播方向．度中心度则可以用来衡量节点对其邻居的平均影响力。

最大度算法的一个基本假设是：越是拥有更多和其他节点链接的节点，越是可能让更多人看到自己发布的信息。因此，一个朴素的想法是，社交网络中度数越大的节点，其影响力也就越大。定义如下：



### 基于网络全局属性的距离中心算法

基于网络全局属性的度量方法要考虑网络的全局信息，这些指标一般准确性相对较好，但时间复杂度高。主要有紧密中心度和介数中心度等方法。紧密中心度可用来度量当前节点对其他节点的间接影响力，或者信息从该节点传播到其他节点的距离，也可间接度量该用户的社会关系强度。该值越大，表示当前用户和其他用户之间的距离越短，该用户影响其他用户的速度越快。介数中心度衡量节点在网络结构中所处位置的重要性。该度量值越大，表示网络中信息流动时经过该节点的信息量越大，即该节点在信息传播过程中的影响力越大。

距离中心算法的一个基础假设是：社交网络中的节点，如果可以更短的路径到达其他节点，那它应该有更好的机会来影响其他节点。考察网络中某个节点到其他节点所需的平均路径的长度，也是衡量影响力的一个方法。定义如下：



### 基于随机游走模型的LeaderRank算法

基于随机游走模型的度量方法主要是基于网页之间的链接关系的网页排序技术，链接关系提供了相互之间的关联和支持，从而可以判断重要程度。主要的度量方法有PageRank、LeaderRank等算法。

Google创始人Larry Page发明的PageRank算法是一个著名的排序算法，Google 的成功很大程度上决定于网页这个排序算法。它的基本思想是：从网页A指向网页B的链接被看作是页面A对页面B的支持投票，但是网页的“重要性”不能简单地看投票数(即链接数目)，而是要看是否有重要的网页链接，重要的网页链接你那么你就变得重要。有了这个想法，很容易建立网页重要性指标的数学模型，转化为一个列随机矩阵的1特征值对应的特征向量问题，特征向量的分量就是网页的“重要性”指标。

但是原始的PageRank模型仍然存在几个问题，其中一个问题是当网络有r个不连通的子网的时候，该网络的网页排序方式至少有r种，于是出现了排序不唯一的问题。PageRank的办法是作素性修正，构造一个所有元素都为1/n的列随机矩阵，与原来的矩阵按加权（一般取0.15和0.85）叠加变成所谓的素矩阵 (primitive matrix),使得网络连通（解决了排序不唯一），并且仍然保持为列随机矩阵（保证了1特征值的存在）。这显然解决了排序不唯一的问题，但是新的矩阵全连接了，但是原来的结构也遭到很大的变形。

[吕琳媛](http://bbs.sciencenet.cn/home.php?mod=space&uid=329471" \t "__blank)等人针对这个问题提出LeaderRank，在已有节点外另加一个节点(ground node)，并且将它与已有的所有节点双向连接，于是得到N+1个节点的网络，这个新的网络是一个强连通的网络，再按照原始的PageRank算法计算得到原来N个节点的“重要性”排序。于是由于加入ground node后的图是一个连通图，排序的唯一性就解决了。结果表明LeaderRank比PageRank算法排序更精准，并且对网络噪音有更好的容忍性。

### 基于网络位置的的K-核混合分解法（MDD）

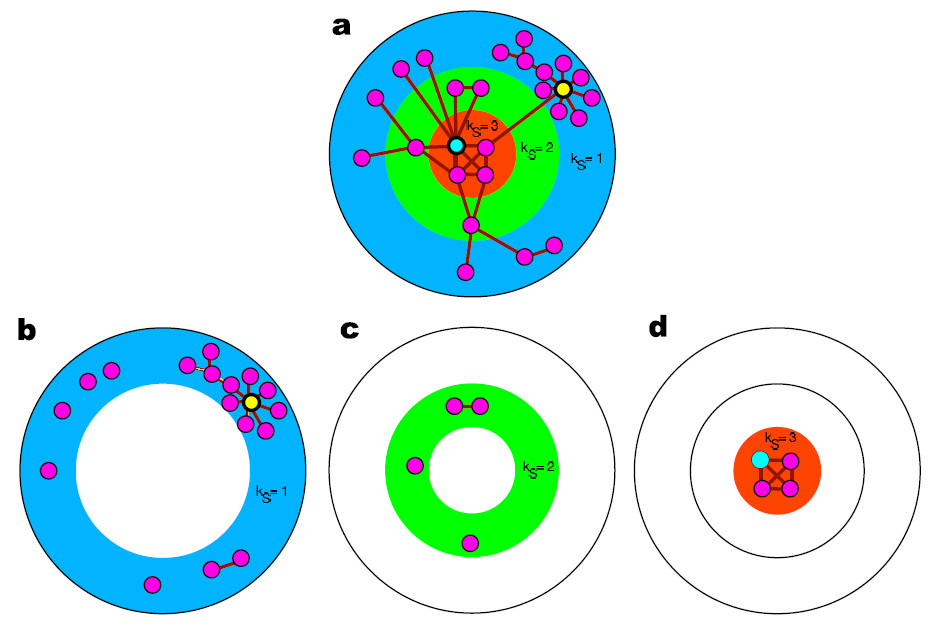
在2010年《Nature》物理版上，Kitsak等人首次提出了节点重要性依赖于其在整个网络中的位置的思想, 并且利用 K-核分解获得了节点重要性排序指标 (k-shell),该指标时间复杂度低, 适用于大型网络, 而且比度、介数更能准确识别在疾病传播中最有影响力的节点。近几年不少学者受到这种思想的启发, 对 K-核进行了扩展和改进, 使其应用范围更广, 准确性更好，并且通过实验表明在影响力传播方面，核数比度数和介数等节点属性具有更稳定的传播力，并且提出了基于覆盖的最大核算法和最大度算法。

通过SIR模型和SIS模型的建模分析指出，对于单个传播源情形，Hubs节点或者高介数的节点不一定是最有影响力的节点，而通过K-shell分解分析确定的网络核心节点（即K-shell值大的节点）才是最有影响力的节点。

K-shell是图论里的一个经典的概念，网络的外壳和边缘的K-shell为1，然后往内像剥洋葱一样进入网络的核心（K-shell值大的区域）。之所以产生Hubs节点或者高介数的节点不一定最具有影响力的原因在于，它们如果在整个网络的边缘，那么它在传播中的作用就很微弱了。而某些度数虽然比较小却位于网络核心的节点将对传播过程产生重大的影响。以CNI的网络为例说明，一个传播过程中感染的规模并不一定与传播起源节点的度k有关，即使从度数相近的不同Hub节点开始，感染结果也可能会非常不同。相反，从相同的Ks层节点开始，即使有不同的度K，传播规模却很相近，说明初始传播节点的Ks值可以准确地预测感染的规模。同时Ks值对于不同的传播率显得相当健壮。这也说明节点所处网络的位置比一个节点的局部属性度K更能决定它的影响力。同样地，高介数的节点和Hub节点一样，也远不如Ks值更能准确地预测感染的规模。

当然文章也指出，当初始存在多个传播源的时候，传播的规模很大程度依赖于初始传播源之间的距离。尽管高Ks点是最好的单一传播者，但是在多个传播源情况，度大的Hub节点往往比Ks大的节点具有更高传播效率。这是因为传播存在交叉感染现象（感染者传染给感染者），Ks大的节点往往在网络的核心，它们聚集在一起，而度大的Hub节点可以分散在网络的不同区域。因此，如果考虑多个传播源有更大的传播规模的话，应该选择处于不同shell（层）、不直接相连的较大的 k的节点作为初始传播源。这篇文章对于我们认识传播网络中到底哪些节点最有影响力，评价网络中节点的重要性，以及从事网络控制、信息传播、疫苗具有重要意义。

Ks的一个缺陷是赋予了大量节点相同的值，导致在一些较为密集的网络中无法衡量其节点的重要性。



Ks计算示意图

针对这个问题，An Zeng在[《Physics Letters A](http://xueshu.baidu.com/usercenter/data/journal?cmd=jump&wd=journaluri%3A%28fa5516ef2180f2c2%29%20%E3%80%8APhysics%20Letters%20A%E3%80%8B&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dpublish&sort=sc_cited" \t "_blank" \o "《Physics Letters A》)》上发表的一篇文章中提出了 MDD（Mixed degree decomposition method），这是一种混合K-核分解法。基本思路是不仅考虑移除后剩余节点的信息，而且考虑被移除节点的度的信息，解决了在某个ks值范围内有大量相同ks值的情况。下面是MDD的度量方式：

k(m) =k(r) + λ ∗ k(e)

λ是一个可调节的参数，值在0-1之间。详细的分解步骤如下：

1. 初始化，在网络中没有节点移除时,每个节点的k(m)等于k(r)。

2. 移除最小的k(m)值的节点（标记为M），并把他们标记为M-shell.

3. 通过公式 k(m) =k(r) +λ ∗ k(e) ，更新所有剩余节点的k(m)。然后，移除所有节点小于等于M并标记为M-shell。重复地执行直到所有节点的k(m)值大于M。

4.重复2 and 3，直到网络中的节点的值都被分配shell值。

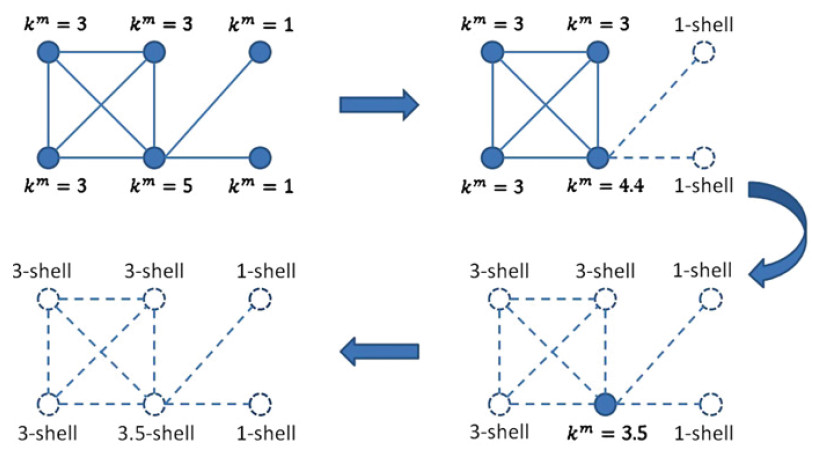


图1 MDD的一个简单的分解示例。虚线表示当前步骤需要移除的节点。MDD的参数λ设置为0*.*7. 在每步中，k(m)值被更新。

## 影响力传播模型

### 两种经典的传播动力学模型

复杂网络上传播研究的对象极广, 比如通信网络中的病毒传播、社会网络中的信息传播、经济网络中的危机扩散等. 在评价各种节点重要性挖掘方法时广泛采用的是传染病模型, 主要包括SIS 模型和SIR模型。

经典传播模型有SIR模型和SIS模型，很多研究学者在经典的传播模型的基础上，提出了多种传播模型，如基于社交网络的传播模型SEIR，谣言传播模型 HKASI 等。

在经典传染病模型中，个体状态主要分为如下几类：易染状态 S（Susceptible），感染状态 I （Infected）和移除状态 R （Removed 或 Recovered）。初始时，设定网络有一个或几个个体处于感染状态，其他个体为易染状态，处于感染状态的个体以一定概率感染处于易染状态的邻居个体。

SIR传播模型中，网络中的个体有 3 种状态，即 S ，I ，R 三种状态，与 SI 模型一样，S状态的个体接触 I 状态的个体后，以概率β变成 I 状态，不同的是，I 状态的个体以概率γ变成 R 状态，处于 R 状态不能被感染或继续传染。记 t 时刻， s(t)，i(t) 和 r(t) 为网络中 3 种状态的个体数所占比例，则 SIR 模型方程如下所示：



SIR 传播模型有两种传播模式：单点接触和全点接触。前者认为一个时间步内，状态为 I 的个体随机选择一个邻居节点中状态为 S的个体，以概率β将其感染为 I 状态，后者是选择其邻居节点中所有S状态的个体，这些个体都以概率β变为I状态。

SIS 传播模型中，网络中的个体有 2 种状态，即 S， I 两种状态，SIR 模型中， I 状态的个体以概率γ变成 R状态，而 SIS 模型中，I 状态的个体以概率γ变成 S 状态。记 t 时刻，s(t)和i(t)为网络中 2 种状态的个体数所占比例，则 SIS模型方程如下所示：



### 独立级联模型和线性阈值模型

独立级联模型，是基于相互粒子系统（Interacting Particle System）设计的一个信息扩散的模型，这是一个概率模型。在其传播机制中关键点是把一个节点对另一个节点的激活用概率p表示，激活失败的概率是1-p，概率越大表示激活成功的可能性越大。此外该模型做了这样的假设：即当节点试图激活其邻接节点时，一旦没有激活成功，那么它将永远失去再次激活节点的机会。下图是独立级联模型的传播过程：

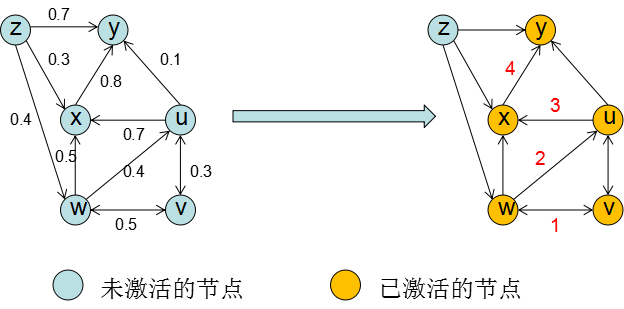


图1独立级联模型

独立级联模型的影响力传播具有次模特性。次模函数的定义如下：

当任意一个函数满足以下条件是被称为次模函数：

* 函数f将有限集合S的子集映射到一个非负实数集U
* 函数具有非严格单调递增的属性，即添加一个元素到集合里面，不

致f减小：

* 定义N为一个有限集合，
* 
* 

第三条也称为自然收益递减属性：添加一个元素到集合所得到的边际收益至少会和添加相同的元素到集合是的父集所获得的边际收益一样多。当前社会网络影响最大化算法大都基于独立级联模型的次模函数性质进行研究，这是因为对于独立级联模型的任意一个实例，影响传播函数都是次模函数，其影响传播具存次模特性，而线性阈值模型却没有良好的次模特性。

线性阈值模型是所有基于节点特异性阈值模型的核心。模型的扩散过程如下：

给定初始传播集合，所有节点的特异性阈值和节点之间的影响力，在第t步扩散时，将t-1时刻满足激活条件的节点加入到t时刻的集合中去。重复这一过程，知道不再有新的节点加入进来。

针对线性阈值模型的扩散过程进行研究，此模型具有一个很好的“影响积累”特性。在线性阈值模型中，当一个激活节点尝试去激活它的未激活邻居而没有成功时，节点对节点的影响力被“积累”下来，而不是被抛弃。这种影响力的积累对后面节点的其它邻居对的激活是有贡献的，直到节点被激活或传播过程结束。

## 影响力最大化算法

### 贪心爬山算法

Kempe等将影响力最大化问题定义为离散优化问题，并证明其是NP难解的。与此同时，他们还提出一个贪心算法，该算法能够取得最优解63%的近似度。贪心算法的基本思想是现将所有节点标记为未激活，然后在这些节点中每次选取一个能激活节点数最多的节点，标记该节点为激活，并加入到最终的集合S中去，然后在剩下的未激活节点中在进行上一步操作，知道选出所要求的k个节点为止。将蒙特卡洛模拟次数设置为1000，就可以得到相当准确的近似值。虽然原始贪心算法可以给出的近似度，但是它的运行时间过长，无法应用在大规模网络中。

在影响力最大化问题中，我们定义σ(S)为影响力传播函数，它表示信息从初始活跃节点集合S开始扩散，最终能够影响的节点数。可以看到，该函数首先是非负的，因为能够影响的节点个数不可能小于零。其次，该函数满足单调性。当加入任意一个新的节点v到集合S中时，σ(S)值不会减小。最后，Kempe证明，无论是在独立级联模型还是线性阈值模型中，影响力传播函数都具有子模特性。σ函数满足次模函数的三个特性，唯一一点区别在于次模函数的值是精确的，而σ值是未知的。为解决这个问题，我们使用蒙特卡洛模拟来对值进行近似估计。通过模拟信息传播过程，我们可以得到相当准确的近似值。正因为如此，借鉴于贪心爬山法，Kempe等人给出了原始贪心算法如下所示，该算法的时间复杂度为O(knRm)。

将社会网络图表示为G=(V,E)，种子节点数k，影响力最大化的种子节点S，蒙特卡洛模拟次数R。

|  |
| --- |
| Input:G=(V,E),k |
| Output:S |
| Initialize:S={},R=1000 |
| 1 for i=0 to k do  2 for each u∈V\S  3 for j=0 to R  4 calculate σ(S∪u) -σ(S)  5 end for  6 calculate avg(σ(S∪u) -σ(S))  7 end for  8 select u which make max(σ(S∪u) -σ(s))  9  10 end for  12 output S |

### 启发式算法

基于节点度数或节点中心性等是社会网络分析中常见的启发式方法。在社会网络和其他网络中，以度数（递减的顺序选择个最大度数节点的启发式节点选择策略，是长期以来的一个标准方法，因为度数越大说明这个节点和其他节点之间的关联就越多，且节点的度数很容易得知，在社会科学中被称为“度中心性”。

社会网络图表示为G=(V,E)，种子节点数k，影响力最大化的种子节点S。

|  |
| --- |
| Input:G=(V,E),k |
| Output:S |
| Initialize:S={} |
| 1 for each u∈V  2 calculate  3 end for  4 for i=1 to k  5 select u which make  6  7 end for  8 end for  9 output S |

在某些领域这些简单的启发式方法效果很好。但是由于每次是静态选取最大度数的节点，没有考虑影响的扩散过程，并不代表这些节点能达到最多的影响范围。因此这些启发式的方法在本文所要解决的问题上效果并不好,甚至比较差。

### 其他相关算法

基于社区的影响力最大化算法主要是利用现有的社区发现算法,将网络分成若干个社区,然后在每个独立的社区上应用已有的影响力最大化算法求解,其中OASNET算法和CGA算法是目前两种应用社区发现算法来求解影响力最大化问题的算法。

OASNET算法是等作者利用网络的社区性质把影响力最大化问题看成一种最佳的资源动态分配问题而提出的一种算法。作者利用社区发现算法把网络划分为各个独立的社区,然后利用动态规划的方法把初始的节点最佳的分配到各个社区,最后将各个社区中被影响的节点累加得到最终被影响的节点数目。

CGA算法的总体思想和OASNET算法类似,也是根据社区发现算法将网络分成各个独立的社区,然后把问题转化为将初始激活节点个数最佳的分配到各个社区,最后将各个社区影响力值求和得到最佳的影响力值。唯一不同的是CGA算法使用贪心算法作为各个社区寻找初始的影响力节点的算法,而OASNET算法是使用度数最大的算法。由于贪心算法的时间复杂度比度数最大算法的时间复杂度高很多,因此CGA算法的时间复杂度也比OASNET算法高很多。CGA算法由于将网络划分成各个小的网络,因此它相比贪心算法时间复杂度有很大数量级的降低。但是CGA算法使用贪心算法作为社区的寻找算法,时间复杂度依然比较高,而且由于网络分成各个独立社区之后会导致边的减少而影响效果。

## 本章小结

通过以上对当前国内外研究现状的总结，我们发现针对社会网络的影响力研究已经取得了许多优秀的研究成果，但是还有一些空间值得去探讨。在影响力度量方面，MDD算法已经取得非常好的效果，但是如何应用于社交网络特别是微博网络中，值得我们进一步去探索。在影响力最大化问题上，基于贪心的算法准确性较高，但是运行时间相对过长，而启发式算法虽然所需时间较短，但是准确度却往往难以令人满意。如何结合贪心法与启发式算法的优点，找到一个方法既能给出不错的准确度，又有较高的效率，也是值得我们去研究的主题。

# 微博用户影响力评价算法

本文给出了用户影响力的定义：微博用户的影响力是指用户通过传播信息或者与他人的交互行为而影响他人的思想或者激发他人产生更多交互行为的能力。针对当前微博用户影响力评价研究中存在的问题以及本文对微博用户影响力的定义，本章在对微博用户间交互行为分析和影响力的关键因素分析的基础上，构建一种基于用户间交互行为的微博影响力评价模型。

复杂网络研究学者提出了多种复杂网络中重要节点发现算法，但是没有考虑到节点对邻居节点的影响是不同的，例如在人际关系网络中，人与人之间连接紧密程度是存在差异的，基于此，本章引入影响因子概念衡量节点之间的重要度贡献，基于MDD分解算法，提出本文的微博用户影响力评价算法。

## 微博交互行为分析

用户之间的交互行为主要有关注、转发、评论、提及，除此之外，在新浪微博中，用户间的交互行为还有发私信、收藏微博、“赞”等。其中，关注、转发、评论这几种行为基本涵盖了微博用户间的所有交互行为特征，本节将对这几种主要行为进行分析。

### 转发行为

用户转发消息，可以在所转发的消息前加入文本也可直接转发，用户将消息转发后消息会被发送到用户所有的粉丝页面中，并且被转发的用户的界面会收到特定的提示。

转发关系是用户间就微博信息进行的主要交互关系，推动着信息在微博网络中的不断传播。如果微博用户发布的一条信息被粉丝转发，该条信息就会被再次传播给下一层的粉丝用户，如此下去，该微博信息就会被越来越多的人看到。转发关系是用户间的一种信息交流与反馈的关系。目前对微博用户的转发行为的研究主要集中对转发行为的预测方面，通过预测用户的微博信息能够引起的转发规模，来预测微博信息的传播效果。用户之间通过转发行为传播信息的过程，也是用户间的影响力传播的过程。用户转发他人的微博信息会显示在自己的微博页面上，并被推送给自己的粉丝，转发者成为信息的再次传播者，可以让更多的人看到该微博信息和信息的原发布者，扩散了信息原发布者的影响力。 转发关系可以体现转发者在微博中的活跃性、参与性，微博信息对转发者产生的影响效果以及被转发者的信息在微博网络中的传播广度和被转发者的影响力。

### 评论行为

用户对消息进行评论，评论消息会被发送到所评论消息的用户页面。关注关系可以视作是一种信息传播和接收的关。评论关系也是用户之间对于微博信息进行的另外一种主要行为，也可以体现出评论者在微博中的参与度。但是，与转发行为不同的是，用户在评论信息的时候，该信息不会被传播到下一级的用户，且评论的内容仅能在信息发布者的粉丝圈中传播（当然是也会被其他能看到该信息，但不是信息发布者的粉丝的用户）。

用户可以通过评论行为吸引信息发布者的目光，增加用户之间的联系强度，吸引其他粉丝用户参与对微博信息的讨论，也可以提高微博信息的受关注度，从而反应出用户的影响力。

### 赞行为

用户在阅读某一条微博时候，如果产生的共鸣或者对这条微博的观点持有赞同态度，就会产生点赞的行为。用户点赞之后，也可以在此点击取消赞，赞数代表的用户发表的微博被大众的认可程度。

在用户对用户之间有赞和没有赞行为对其产生的影响力是有所不同的。赞代表的是积极的、认同的，是高度的同意。所以赞这种行为也是统计用户影响力的一种不可或缺的因素

### 关注行为

当用户在浏览微博时，对某个人发表的微博感兴趣时，就会产生关注关系，就是成为某个人的粉丝。在以后这个人发的微博就可以非常方便的看到，并且可以作为粉丝进行转发和评论等一系列的操作。

与博客、Facebook、人人网等在线社交网络不同的是，微博中用户间的关系是一

种单向的、不对称的好友关系。一般来说，微博信息的流向是通过关注关系网络从微博用户到其粉丝，微博用户通过关注行为来获取信息，通过被关注行为传播信息。 用户间关系的不对称造成用户间影响力的不对称，关注关系能够体现出被关注者对关注者的影响。

用户之间的关注关系是构建网络模型的主要的关系。通过关注某一个用户，粉丝可以进行评论、转发和赞等相关交互行为的操作。

## 微博网络建模

微博网络是一个有向网络，这是由于微博中存在单向的关注关系。在微博网络中，节点代表微博用户，而边代表了关注关系。我们定义来表示微博网络，V表示节点的集合，E表示有向边的集合。此外，每一个用户都有自己的个人信息，比如发布的微博数、转发的微博数、评论的微博数等等。与此同时，用户之间包含了很多交互行为，比如转发、评论等。对于一条特定的有向边e，表示一个用户关注了另一个用户。

表3-4列出本文中使用的符号

|  |  |
| --- | --- |
| 符号 | 描述 |
| V | 微博网络中的节点集合 |
| E | 微博网络中的边的集合 |
| v | 网络中任意一节点 |
| e | 网络中任意一边 |
| r(u,v) | 用户u转发v的数量 |
| c(u,v) | 用户u评论v的数量 |
| r(v) | 用户v的评论数量 |
| c(v) | 用户v的转发数量 |
| A(v) | 用户v的活跃度 |
| Inf(u,v) | 用户v受到u的交互影响力 |
| IS(u,v) | 用户v受到u影响的影响因子 |
| Inf(v) | 用户v的影响力 |
| InfLv(v) | 用户v受到u影响的影响度 |

## 微博用户影响因子

### 微博用户的影响力

社交网络中的影响力只有通过人们之间的交互活动才能够体现出来，比如用户Ａ在网上的发帖吸引了用户Ｂ，使得后者成为Ａ的粉丝，即Ａ对Ｂ产生了影响力。由于社交影响力的研究工作涉及众多学科和领域，对影响力的的定义也大不相同。

目前大部分研究工作都是针对社交网络结构及其上的交互信息和用户行为特征进行量化和分析的，因此可以把能对信息传播产生影响和对他人行为产生影响视为用户的影响力。在本文所研究的工作中，在针对新浪微博的研究中发现，用户的交互行为是影响力的一个重要的影响因素，用户A发表的微博被B所看到，B产生了共鸣或者不赞同的态度，就会产生一系列的行为，比如说赞、评论、转发和关注等等。这样A便对B产生了影响。影响力不仅仅是由这些行为决定的，而且和产生影响力的用户自身属性以及该用户在网络中的重要性综合判定的。在研究了社会网络中节点的重要性，我们根据K-核理论及其改进的MDD算法，用来作为衡量节点在网络中位置的重要性的一个基本的条件，并且结合网络中的交互行为，定义出节点的影响因子，由这几个方面来综合判定微博网络中用户的影响力。

### 影响因子

K-核分解算法对网络进行一层层剥落，最后剩下网络中最核心的节点集，考虑的只是节点自身在网络中所处的位置，并没有考虑节点所处的局部环境，而且ks指标往往赋予大量节点相同的值，导致没法区分处于同一核的节点的重要程度。K-Shell 分解算法得到的结果层次性较差，但是能够快速有效的发现网络中核心节点集，但是并不能进一步区分节点重要性差异。MDD根据引入了分解节点的度的思想，解决了K-核分解存在的大量节点ks值相同的问题，但是只是解决大量高ks值的节点区分问题，对于度为1的核数还是没有有效的方法。Liu等提出根据某一节点到最大K-核节点集合的最短距离之和来区分大量ks值相同的节点。任卓明等提出一种度量指标，根据最小K-核节点和网络中其他层节点的连接关系来区分最小K-核节点的重要性。上述方式只是考虑解决ks值相同的问题，并没有考虑到邻居节点的差异性，不同的节点对其产生的影响力是不同的。

基于这个出发点，本文引入影响因子概念，充分考虑节点的局部环境，即节点与其邻居节点的连接关系。

微博用户网络是一个有向无权网络，边代表用户之间的关注关系。在微博社交网络中，如果用户v和用户w同时关注用户u，也就是说v和w是u的粉丝，由于v和w的节点自身的属性和与u的交互各不相同，因而他们对u的影响也不相同。自身的属性由这个用户的粉丝数、发表微博数和关注数来衡量，与u的交互影响用转发、评论和关注信息来度量，用户之间的影响因子是由这两方面的影响共同决定。因此，引入影响因子来表示节点对其邻居节点的贡献度是不同的，节点v受到节点u的影响因子是由u的自身属性和u和v之间的的交互行为信息来综合度量。

节点u的自身属性的定义：



I(u)表示用户u的自身属性，mbnums表示用户u发表的微博数，attnums表示用户的关注数，fansnums表示用户u的粉丝数，zannums表示用户u的收到的赞数。α、β、γ、δ是参数。

节点u对v的交互影响：



Inf(u,v)表示用户v受到用户u的交互影响，c(u,v)表示用户u评论用户v的数量，r(u,v)表示用户u转发用户v的数量，c(v)表示用户v的评论数量，r(v)表示用户v的转发数量,表示u对v的关注关系。取值为0或1，表示存在关注关系或者不存在关注关系。

节点u对v的影响因子：



IS(u,v)表示用户u对 v的影响因子,I(u)是用户u的自身属性，Inf(u,v)是u对v的交互影响。λ和σ是参数。

## MIB算法模型

根据MDD算法，本文提出了一种基于交互行为的微博影响力度量算法MIB(Mixed degree decomposition Interaction Behavior)，该算法不仅考虑考虑了用户在网络中的位置，并且考虑了邻居节点对其的影响力，即局部属性，从全局重要性和局部重要性综合度量用户的影响力。

### 算法设计

上一步给出了u和v之间的影响因子，在本节中我们根据MDD算法得到每个节点的核数值，再结合用户之间的影响因子给出MDD-IB算法的模型。MDD-IB算法根据节点的ks值，并且考虑了邻居节点的影响力贡献，参照网络拓扑结构的邻接矩阵表示形式，构造影响力矩阵I。



其中I表示影响力矩阵。m(n)表示经过MDD分解后的数值，表示节点i对节点j的影响因子。矩阵中，,表示节点u 对节点 v 的影响力，节点u的m(u)越大，而且对其邻居节点v的影响因子越大，那么它对节点v的影响力就越大。

根据影响力矩阵I，考虑自身ks值和邻居节点的影响力贡献，给出用户的影响力的定义：



用Inf(v)表示节点的影响力，节点v 的影响力由两部分乘积构成，一个因素是在网络基于网络中节点的位置，即网络全局的重要性，用此节点自身的MDD值来度量，另一个因素是节点的局部重要性，即其相邻节点对节点v的影响力之和，该模型综合考虑了局部重要性以及全局重要性，使得在网络中能较为准确的发现重要的节点。

### 算法步骤

本小节，本文给出具体的算法步骤和算法的伪代码实现。

具体步骤：

1、由公式3.1计算出每个节点自身的影响力。

2、遍历用户集合，取得所有和当前用户产生交互行为的用户。

3、定义一个影响力计数器inf；初始值为零。

4、循环遍历与当前用户产生交互行为的用户，分别根据公式3.2计算他们之间的交互影响力，根据第1步计算出自身影响力和公式3.2计算他们之间的影响因子，计算出来的值加入inf。

5、循环步骤4，遍历完成后，得到的inf值即为当前用户的交互影响力。

6、循环步骤2，得到每个用户的交互影响力。

7、根据K-Shell的改进MDD算法，得到每个节点的MDD值。

8、将用户的交互影响力和用户的MDD值合并。首先遍历用户集合，判断用户的交互影响力是否为空，若不为空，根据公式3.5更新结果;若为空，用户最终的影响力即为自身影响力和MDD分解后的影响力乘积，

9、将最终的影响力进行排序遍历输出，即为每个节点的重要度排名。

算法的伪代码如下：

社会网络图表示为G=(V,E)，节点的重要度结果rMap，节点的自身属性I(u),节点之间的交互影响力Inf(u,v),节点之间的影响因子IS(u,v)。

|  |
| --- |
| Input:G=(V,E),uList |
| Output:rMap |
| Initialize:rMap={} |
| 1 //MDD  2 for each u∈V  3 Initial k(u) =k(r)  4 end for  5 while(size(V)!=0)  6 minSet = getMinSets();  7 update k(u) cby k(u) =k(r) + λ ∗ k(e)  8 mdd(u) = k(u)  9 remove minSets;  10 end while  11 // reaction  12 for each u∈V  13 calculate I(u)  14 for each (u,v)∈E  15 calculate Inf(u,v)  16 calculate IS(u,v)  17 end for  18 end for  19 //merge  20 for each u∈V  21 Double result=0;  22 for each v∈neib(u)  23 Result += IS(u,v)  24 end for  25 rMap.put(u,result);  26 end for |

## 算法仿真与分析

### 微博数据的爬取

我们使用python语言开发了一个微博数据爬取系统。Python语言，是一种面向对象、解释型计算机程序设计语言，由Guido van Rossum 于1989年底发明，第一个公开发行版发行于1991年。Python 语法简洁而清晰，具有丰富和强大的类库。它常被昵称为胶水语言，它能够很轻松的把用其他语言制作的各种模块（C/C++、Java等）轻松地联结在一起。根据本文研究的内容，主要爬取从指定版块出发，按广度优先策略，逐层深入，从其粉丝出发，逐个获取用户的数据。

爬取系统在爬取数据时，从指定的版块出发，爬取规定时间内的所有微博的信息，通过这些微博信息，爬去相应的粉丝与评论与转发信息，再针对评论与转发中的信息，爬取所属用户的信息。采用cookies模拟登陆，半自动爬虫，抓取相应的Html后，用正则表达式解析提取相关的信息。

爬取系统分为数据抓取层和存储层，如图3-2。数据抓取层，主要利用 cookies模拟登陆从新浪微博抓取需要的数据，并用用正则表达式执行过滤操作；异常处理层，则对在数据获取过程中发生网络错误、数据解析错误、编码错误等情况进行处理，保证系统可以无监督不间断运行；存储层，则是将抓取到的数据组织后存储到数据库中。

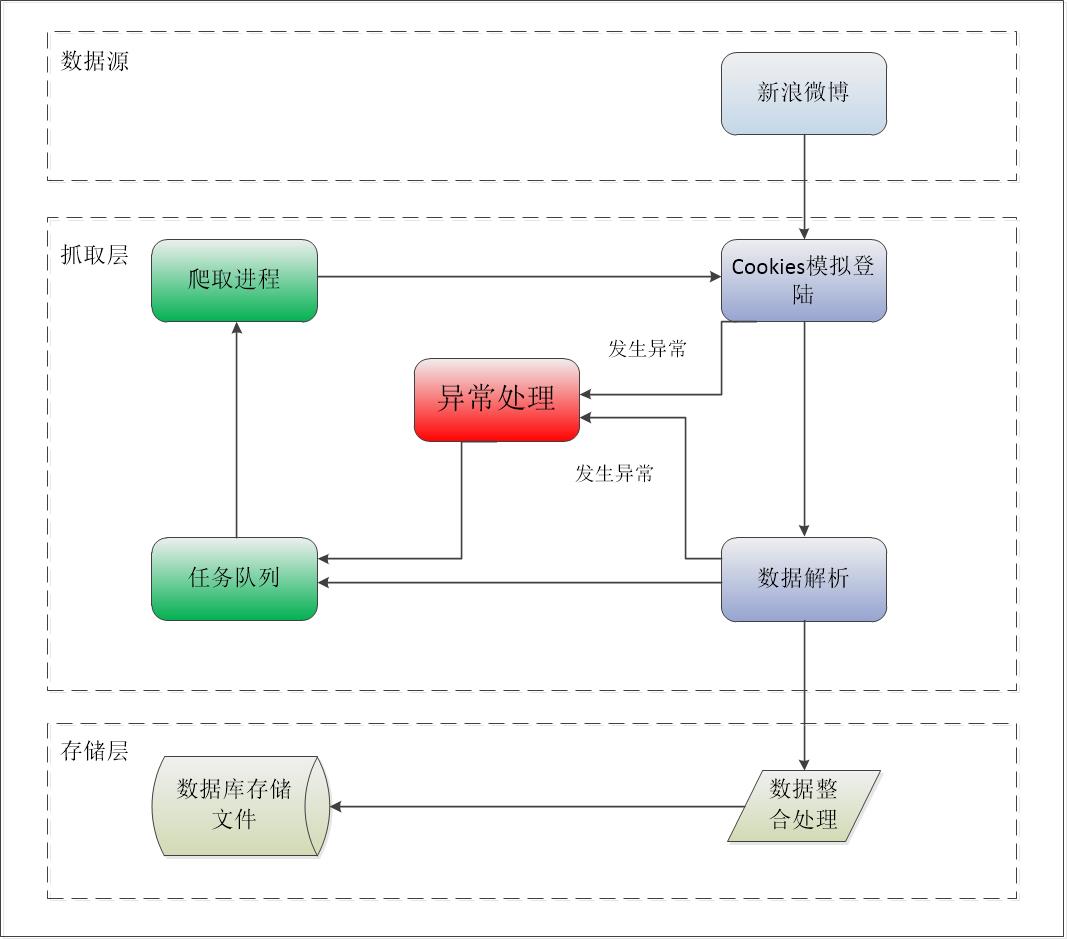
f

图 3-2 数据主抓取系统框架图

在数据抓取层，有几点需要注意。首先，新浪微博对每个账号的请求频率限制为1000 次/小时，超过1000次/每小时的请求将返回403，服务器拒绝请求。我们采取多账号的方式绕过此限制，此次系统中，共使用20个帐号，4个进程同时抓取的策略，以绕过新浪微博对请求频率的限制。同时，在实际的系统测试中，我们发现，除超限外，过快的请求频率，依然会被新浪微博封锁掉，以防止这种情况的发生，我们在系统中加入了异常处理层，对系统执行过程中出现的异常进行针对性处理，并记录异常发生的时间和发生异常的用户，同时将该用户加入二次抓取队列，以等待再次尝试抓取，二次抓取失败则放弃该用户。

### 数据集描述

本文重点研究的是微博中用户间的交互行为。交互主要包括发布微博、转发微博、评论微博等。实验选取新浪微博作为数据来源，通过模拟cookies登陆得到所需的数据，将收集到的数据按照以下几个方面进行存储统计：

（1）用户信息表：用户ID、昵称、粉丝数、关注数、微博数；

（2）微博信息表：微博ID、微博内容、发布时间、转发数、评论数、赞数；

（3）用户转发关系表：转发用户ID、被转发用户ID、转发内容；

（4）用户评论关系表：评论用户ID、被评论用户ID、评论内容。

（4）用户评论关系表：关注用户ID、被关注用户ID。

本人爬取了2016年新浪微博7月1日到8月2日#西安身边事#主题下的一个月的微博相关的数据。总共4472条微博，爬取了相关的评论数、转发数、赞数和内容。共有45531个用户、20383条转发关系和59404条评论关系。统计数据如下表3-3所示：

|  |  |
| --- | --- |
| 数据项 | 数据量 |
| 用户 | 45531 |
| 微博 | 4472 |
| 转发关系 | 20383 |
| 评论关系 | 59404 |
| 关注关系 | 52437 |

表3-3数据统计资料

### 数据集分析

对微博研究内容的综述中，指出已经有很多学者对微博网络的结构特征进行分析，发现微博网络同样也具有无标度特性和小世界效应，即微博网络是一种复杂网络。因此，在我们获取微博用户数据时，如果采集到的数据量是足够的，微博用户网络也会呈现小世界效应和无标度特性。

为了验证本文数据的有效性，下面将通过对关注关系网络的出度入度、聚类系数和平均路径长度等参数的分析，以验证我们所抓取的微博用户网络是否具有复杂网络特性。

图3-3是整个数据集的网络拓扑图。

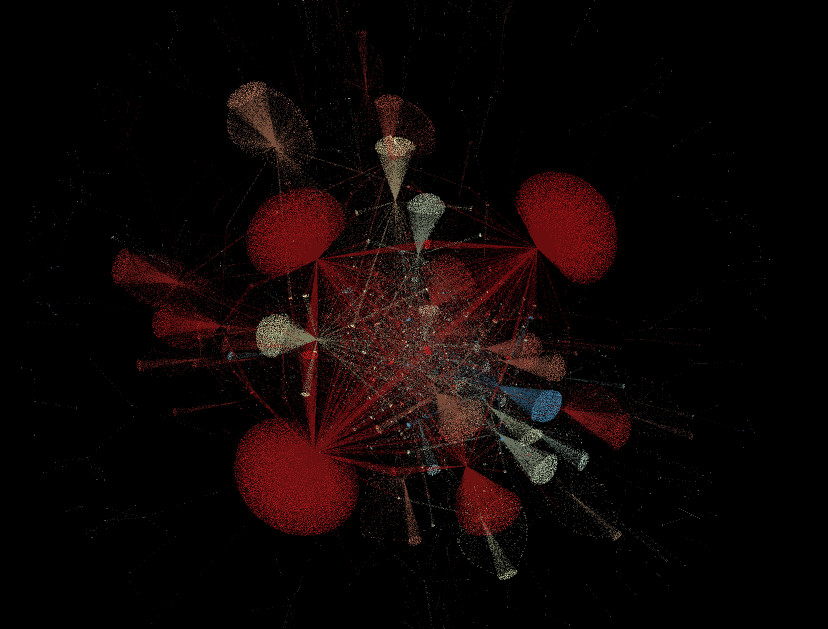


图3-3

经过计算给出一些该微博网路的一些相关的数据，如表3-4

|  |  |  |
| --- | --- | --- |
| 平均聚类系数 | 平均路径长度 | 网络直径 |
| 0.017 | 3.149 | 9 |

表3-4

微博网络都具有较低的平均距离与较高的集聚系数，都具有明显的小世界性质。虽然微博网络是基于微博用户构建的，但互联网技术使得虚拟社交变得更加容易。与现实的社会网络不同，用户可以跳过朋友的朋友关系而直接关注自己感兴趣的对象，使得整个网络的平均距离大大缩短，大部分网络的平均距离都在 3-4之间，这意味着信息的传播效率非常高，只需要经过 2-3个中心节点，信息就可以在任意 2 个节点之间传播。所以此网络具有”小世界效应”。

社会网络中节点的度会遵从幂律分布，因此，为了进一步验证子数据集的有效性，我们给出用户数分别和粉丝人数、关注人数分布图，分别如图3-5和图3-6所示。由图得知，它们都服从幂律分布。因此，有足够的理由证明将该子数据集用做实验数据集是合理的。

### 实验项目

本节为了突出算法的优劣性，分别从以下的几个方面进行了实验：

1）本节算法与其他算法在微博数据下的节点评价效果

3）本节算法与其他算法之间的相关性对比

4）本节算法与其他算法之间的影响范围对比

### 实验结果

为了评估算法的有效性，我们选择MDD算法来做比较。

1. 算法结果

下表分别给出了MDD算法和MDD-IB算法排出的网络中最具有影响力的前二十个用户。

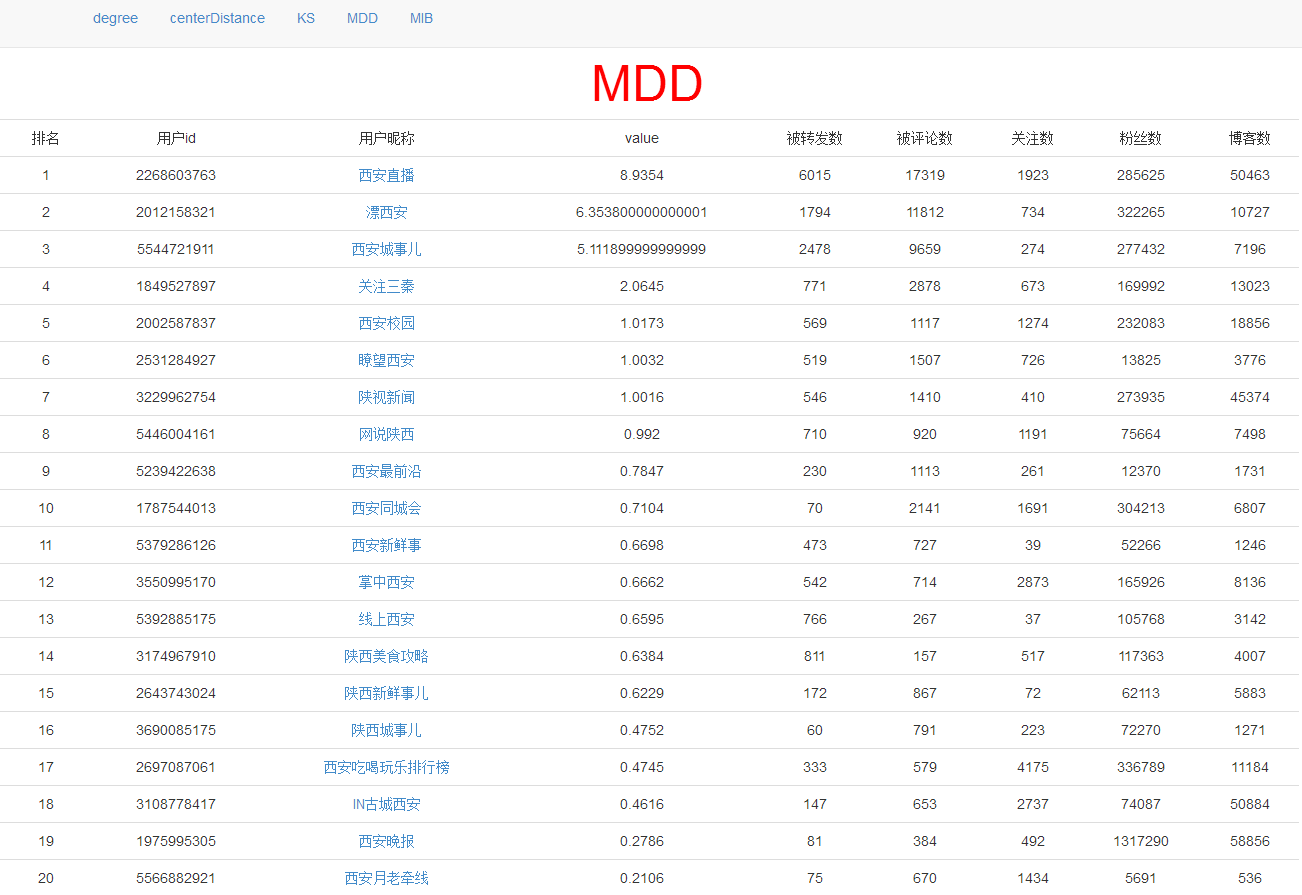


图- MDD算法

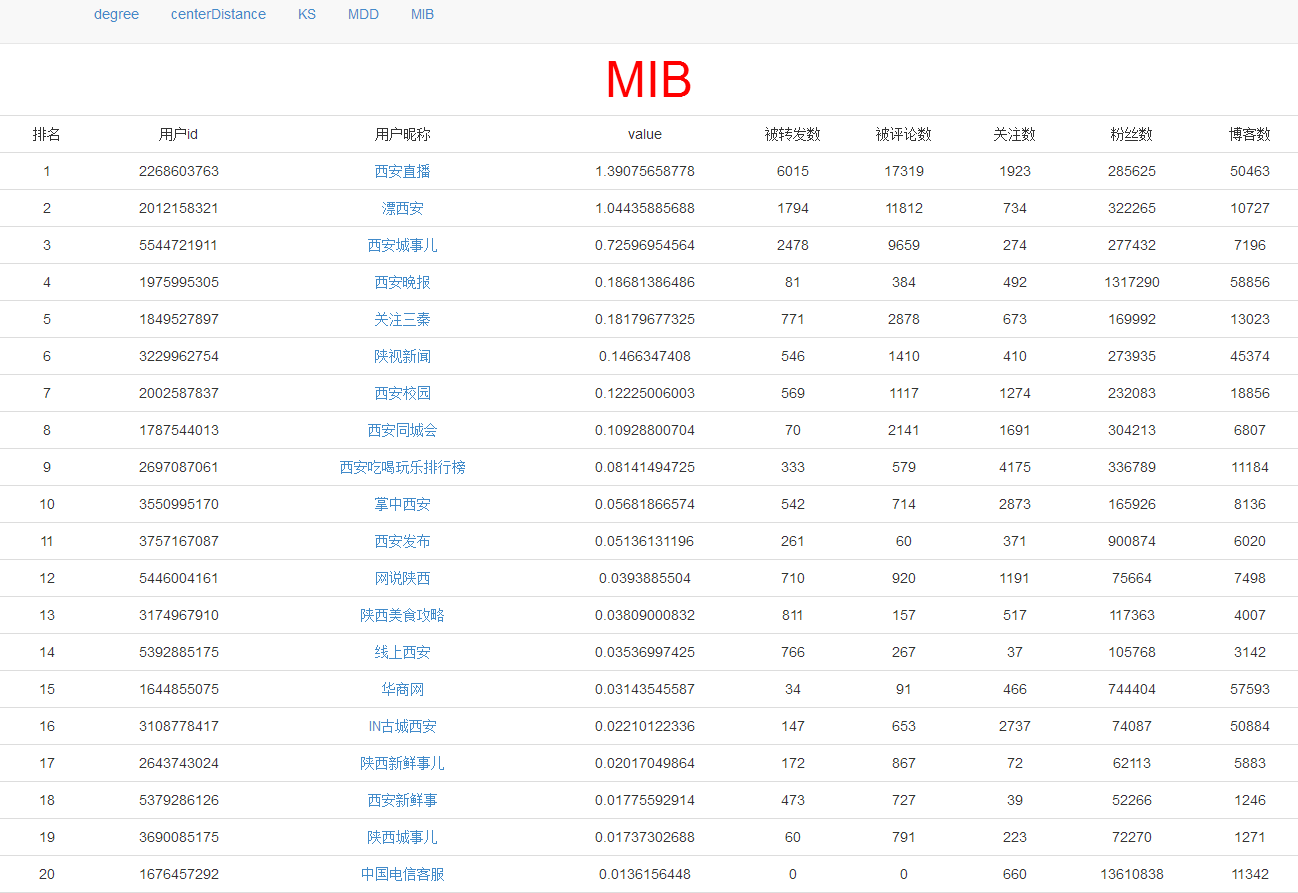


图- MIB算法

图2和图3分别给出了MDD和MIB排出的网络中最具有影响力的二十个用户，在这二十个用户中，有十七个用户是重叠的。通过MIB算法，用户西安发布、华商网和中国电信客服成为涌进前十的新用户，但是这三个用户在MDD算法下的排名分别是第26、第32和第62。虽然这些用户的排名都得到了提升，但是原因是不同的。对于用户西安发布来说，尽管发表的博客数相对较少，但是其微博被别人转发的次数非常多，也就是说，作为用户交互行为的一种，转发行为的密集使得用户排进了前二十。对于华商网和中国电信客服而言，尽管用户交互行为也不密集，但是却粉丝的数量众多，而且用户中国电信客服被排名第二和排名第六的用户的用户关注。粉丝的质量高，用户的影响力也会得以提升。此外，在表中，用户关注三秦的排名比用户西安晚报要高，这是因为用户关注三秦与其他用户做了更多的交互。总之，实验结果告诉我们，MDD算法依赖于网络局部的位置，即是用户的粉丝的数量和质量，而我们的算法不仅依赖于粉丝的数量和质量，还充分考虑了用户之间的交互行为。

1. 算法之间的相关性

相关性分析是指研究两个变量之间的相关紧密程度，衡量这两个变量之间是否存在某种关联。社交网络用户影响力研究领域中，学者们通常采用统计学中相关系数指标来衡量算法之间的相关性。

统计学中的斯皮尔曼等级相关系数（Spearman’s correlation coefficient）是用来衡量两个变量之间的相关性。它对两个变量的分布没有要求与不管样本容量的大小，都可以使用斯皮尔曼等级相关来研究这两个变量之间的相关性。它对两个变量的分布没有要求与不管样本容量的大小，都可以使用斯皮尔曼等级相关来研究这两个变量之间的相关性。对于样本容量为n的样本，按照升序或者降序原则把n个原始数据和转化为等级数据和，则原始变量 X和Y之间的相关系数为：



式子中, 分别是原始数据在数据集 X，Y 中的平均排名位置。而在实际应用中，可以采用如下公式来近似计算斯皮尔曼等级相关系数：



其中，斯皮尔曼等级相关系数的取值范围[-1,1],当越接近 1，则表明两种排序算法之间相关性越大，当越接近-1，则表明两种排序算法之间相关性越小，

甚至相反的。在本文中提到的相关性系数均指按照降序原则，对两个变量的初始值进行排序，然后计算这两个变量的斯皮尔曼等级相关系数。

研究社交网络用户影响力的领域中，我们采用以下两种算法与新算法做个关联性分析。

MDD：基于K-Shell算法，根据节点所处的位置，结合节点的度移除节点，再根据剩下的节点和移除节点的度的信息更显剩下的节点的影响力值。节点的位置是衡量用户影响力的一个非常直观的因素，所以在研究复杂网络问题时，学者们把它作为一个重要衡量节点影响力的算法，如 Zeng 等人，他们研究表明MDD划分算法更好的解决的K-Shell无法区分大量节点影响力相同的情况。

MBUserRank：基于不同的用户交互，即不同的用户对用户影响力，一个用户的影响力被不等划分，并传递给其关注者，对用户交互行为的充分考虑使得我们的方法更符合微博的特性，计算得到的影响力值也更准确。与PageRank类似，阻尼系数被引入，这是因为微博行为也会存在各种随机偶然性，比如转发微博的偶然性等。该算法基于PageRank，但是不同之处在于在该算法中，任意一个用户对其关注者的影响力贡献是不均等的，这依赖于用户之间的交互程度，也就是不同的用户对用户影响力。MBUserRank算法认为不仅依赖于粉丝的数量，并且依赖于粉丝的质量。

|  |  |  |
| --- | --- | --- |
|  | MIB和MDD | MIB和MBUserRank |
| 斯皮尔曼等级相关系数 | 0.62 | 0.34 |

表 2 算法之间的相关性

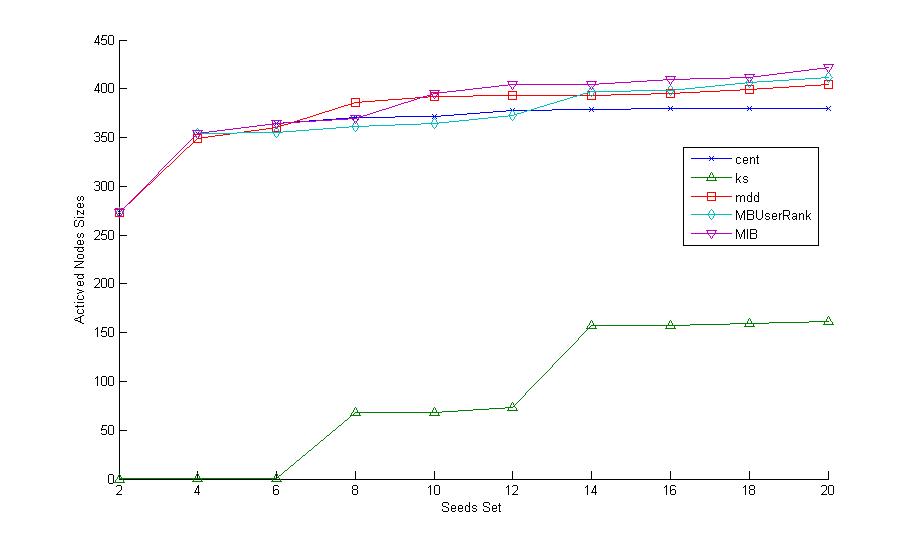
由表可看出 MIB算法的运行结果与 MBUserRank的运行结果之间关联性不强，即MIB算法的运行结果与用户粉丝数量没有必然的联系，这与国外研究 Twitter 用户影响力的结论一致。MIB与MDD之间的相关系数为0.62，表明 MURank 算法虽然是考虑到用户的交互行为，但是其运行结果与 MDD结果之间的关联性还是较强。

3）验证

在排序出前二十的用户后，下一步就是要验证这二十个用户到底是不是最具有影响力的。我们选择了其他的一些算法作比较，包括距离中心算法、K-Shell算法、MBUserRank算法和MDD算法。

将种子集合的大小标注为。对于每个算法，每次挑选出排名最靠前的k个节点（对于ks算法是随机选出核数最高的k个节点），从它们出发传播信息，并获得最终激活的节点数。图-展示了不同方法选出的种子集合最终激活的用户数量，这里主要统计了从到的情况。将节点u对节点v的交互影响作为激活概率，反应了用户之间的交互行为，交互影响的取值是在0到1之间。此外，激活概率不是我们重点考虑的，主要的是模拟信息的传播。采取独立级联模型进行实验的验证。

可以看到，由从MIB选出的种子节点集合每次都会比其他方法激活更多的用户，这足以证明我们的影响力计算方法确实能够给出相对较优的结果。



### 实验结论

通过以上实验，我们观察到（1）一般来说，用户的粉丝数量在一定程度上影响了该用户的影响力，但并不是唯一的因素；（2）—个用户的影响力与该用户和其他用户的交互行为密切相关；（3）—些用户在粉丝数量和交互行为上并没有优势，但是只要他们能获得影响力大的用户的关注，他们依然可以成为具有影响力的用户。综上实验结果表明，改进的MDD算法的比传统的算法要明显。并且改进的算法的效率比传统的算法效率要高。综合以上几个方面，本章提出的算法是比较具有实用性的。

## 本章小结

本章先介绍了微博中主要的几种交互行为，并且给出了微博网络构建的模型。MDD算法是较为理想的粗粒度的度量节点重要性的一种方法，但是由于划分结果层次性较差，得到的结果并不能进一步区分节点的重要性差异。因此本文引入“影响因子”概念，构造节点之间的影响力矩阵来表明邻居节点之间的影响力的贡献度。最后提出MIB算法模型，即基于用户交互行为的影响力度量模型，该算法克服以往传统算法片面性，综合考虑了节点的全局重要性与局部重要性，并且考虑了节点对其邻居节点的影响力贡献不同，用节点的MDD值和相邻节点之间的影响力两方面综合度量，更加符合实际情况，能够得到更准确的结果。

# 微博用户影响力最大化算法

尽管针对影响力最大化问题已经有许多算法被提出，但是它们不是太耗时，就是缺乏准确度，或者是前提条件太多。本章对该问题进行了进一步的探讨。首先，对算法做了修改，得到新的影响力度量算法。在此基础上引入启发，提出了新的影响力最大化算法。最后通过实验对该算法进行了分析与验证。

## 影响力度量

上一章重点讨论了微博中影响力的度量方式。首先引入用户对用户交互影响力用来度量一个用户对另一个用户的影响程度，并分别从直接和间接的角度出发，讨论了体现用户交互程度的几个个因素。由于交互是两个人的行为，因此我们从关注者与被关注者的角度分析双向的交互行为，并给出了用户对用户影响力计算公式。其中，转发、评论是用户直接交互行为相关的，而赞和关注是用户和用户之间的间接的交互行为。接着，基于MDD算法，结合用户对用户之间的交互关系，提出了用户之间的的影响因子，最后根据MDD和用户间的影响因子，构造用户影响力矩阵，提出了计算微博中用户影响力的算法。该算法与MDD算法的本质区别在于充分考虑到用户之间的影响不同，这取决于用户自身的属性和用户之间的交互。

在本小节中，各个节点的影响力值用的是上一章计算出来的数值。采取了将用户行为和网络位置结合其来的方法，给出一个较为合理的影响力度量方式，在接下来的最大化问题的研究中，各个节点的影响力值就以这个为前提。

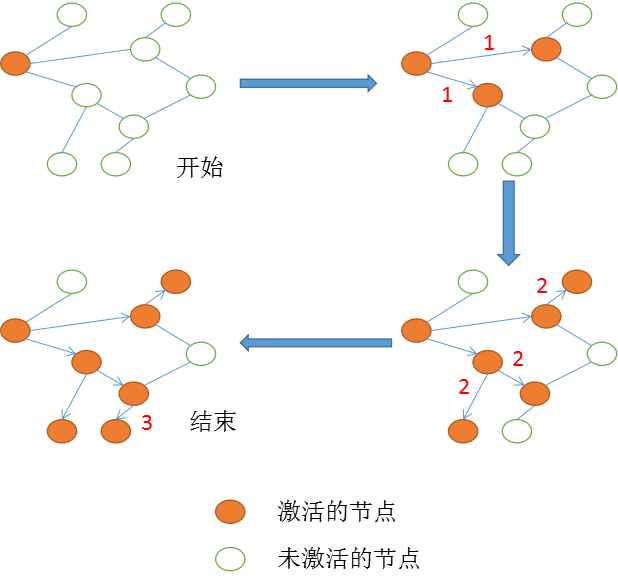
## 影响力传播模型3-4

### 传统的IC模型

关于影响最大化问题的模型选择，目前研究最广泛的、最基本的两个传播模型是：线性阈值模型(LT)和独立级联模型(IC)，所有的研究工作都是基于这两个模型或者其拓展模型。

独立级联模型，是基于相互粒子系统（Interacting Particle System）设计的一个信息扩散的模型，这是一个概率模型。在其传播机制中关键点是把一个节点对另一个节点的激活用概率p表示，激活失败的概率是1-p，概率越大表示激活成功的可能性越大。此外该模型做了这样的假设：即当节点试图激活其邻接节点时，一旦没有激活成功，那么它将永远失去再次激活节点的机会。

给定初始传播节点集合，以及所有节点之间相互激活成功的概率。当传播至第t步时，利用在t-1步中被激活的节点，根据激活概率试图去激活它们的邻居节点，并将在这一步中被激活的节点加入到激活节点的集合中。重复这一过程，直至不再有新的节点被激活。下图是独立级联模型的激活过程。这个激活概率是一个系统变量，与其它尝试激活节点而未成功的节点无关，这也是该模型命名的来历。



/\* 图需要重画\*/

传统IC模型中认为处于激活状态的节点就会自发地去尝试激活其邻居节点，即一旦节点v处于激活状态，节点v就以100％的概率去对其邻居节点执行尝试激活行为，而邻居节点是否被激活则由节点间的激活概率决定，这个激活概率是是一个系统变量，通过多次蒙特卡洛模拟得到。显然，在微博网络中激活概率并不是随机的，与节点之间的影响以及节点的活跃度有必然的联系。而且，并不是出于激活状态的节点必然会激活其他节点，也存在着激活状态的节点对其他节点没有产生任何影响的情况。

### 活跃度

在独立级联模型模拟信息传播时，每一条边都会对应一个0到1之间的数值，该值表示信息有多大可能从一个节点扩散到另一个节点，称为激活概率。由于在微博中的激活概率并不是像在IC模型中随机的，而是根据节点的一些行为因素来进行决定的。因此针对上述问题，我们提出了活跃度的概念。

基于这个出发点，本文引入活跃度概念，充分考虑节点的活跃程度。活跃度用来度量发布者在此网络中的活跃程度，通过该用户在网络中的发微博数和参与的回复数以及参与的转发数来衡量。

用户u的在网络中的参与程度为；



AC(u)表示用户u的参与程度，表blog(u)示节点u发布的微博数，comment(u)表示用户u参与的回复数，repost(u)表示用户u参与度转发数。

用户u的活跃度为：



A(u)表示用户的活跃度，max(AC)表示最大活跃值的用户。

### 基于活跃度的改进的IC模型

在研究影响力最大化问题时，我们的方法基于独立级联模型，但是在激活概率的设置上进行了调整。众所周知，统一独立级联模型将图中的每条边都赋予一个相等的概率值，而在加权独立级联模型中，每条边对应的激活概率等于起始节点出度的倒数。无论是统一独立级联模型还是加权独立级联模型，都不适合直接用于微博网络，这是因为微博网络中每条边的激活概率与这条边连接的两个用户之间的交互行为息息相关。用户对用户影响度表示一个用户对另一个用户的影响程度，它基于用户交互行为。

IC模型是对实际社交网络的一种抽象，模型中节点的主要属性是节点度与节点间的激活概率，节点影响力的定义也都是基于这两个属性。本节对IC模型进行了扩展，引入节点之间的影响度属性，使节点影响力的定义更科学。

节点u对v影响度：



InfLv(u,v)表示u对v 的影响度，Inf(u,v)表示u对v的影响因子，A(u)为用户u的活跃度。

从用户交互行为计算得到的用户对用户影响力代表了过去一段时间内信息从一个用户到另一个用户的传播情况，在乘以用户的活跃度，构成了用户之间的影响度。我们将此作为我们模型的激活概率。不同的用户对用户影响力体现了不同的信息传播可能性。此外，它的值位于0到1之间，符合激活概率的数学要求。考虑到我们的目标是对影响力最大化问题进行研究，而不是激活概率的精确计算，因此将用户对用户的影响度贡献近似为激活概率是合理的。

。。。

。。。

改进的IC模型图

。。。

。。。

。。。

## 影响力最大化SMC算法4-6

有关影响力最大化问题的算法可以归纳为两种类型，一种是基于贪心的算法，它的特点是结果准确性较高，但是运行时间过长；另一种是启发式算法，它在运行时间上具有巨大的优势，但是准确性较低。为了结合这两种算法的优点，本节提出了一个新的算法SMC（Sina Microblog CELF），该算法可以取得近似最优解，而且运行时间相对较短。

### 次模函数

次模函数的定义如下：

当任意一个函数满足以下条件是被称为次模函数：

* 函数f将有限集合S的子集映射到一个非负实数集U
* 函数具有非严格单调递增的属性，即添加一个元素到集合里面，不

会导致f减小：

* 定义N为一个有限集合，

若

则

第三条也称为自然收益递减属性：添加一个元素到集合所得到的边际收益至少会和添加相同的元素到集合是的父集所获得的边际收益一样多。即就是当我们添加一个节点到种子集合时，如果集合越小，节点对影响范闱的增量影响就越大。当前社会网络影响最大化算法大都基于独立级联模型的次模函数性质进行研究，这是因为对于独立级联模型的任意一个实例，影响传播函数都是次模函数，其影响传播具存次模特性，而线性阈值模型却没有良好的次模特性。

### 算法设计

上一节介绍了MIB算法，该算法在计算出用户影响力后，会对用户进行排序。一个用户的影响力代表了该用户在网络中的权威性以及影响其他用户的可能性，相比排名较低的用户，排名靠前的用户更具有话语权，能够影响到范围更大的人群。在影响力最大化问题中，我们的目的是找到大小为的种子节点集合，从该集合出发传递信息能够激活最多的节点。因此，在选择种子节点时，不需要考虑普通节点，因为这些节点对于信息传播的意义不大。相反，要考虑具有一定影响力，也就是排名靠前的用户，因为这些用户往往能够使信息传播的更远。在选择大小为的种子节点集合时，首先根据活跃度去除活跃度值是零的一些用户，接着通过MIB算法对全网络用户进行影响力排序，然后选出排名前ck的用户组成候选种子节点集合，c是启发因子，取值在[0,1]。再从剩下的集合中利用CELF算法挑选出剩下的k-ck个种子节点。当c==0时候，退化成了CELF算法；当c==1时，即是启发式的MIB算法。我们使用贪心方式从剩下未激活的集合挑选种子节点，但是候选种子数却从全网络用户减少到了只有k-ck个用户，这是与贪心算法的区别。

我们选取CELF算法从剩下的未激活节点中找寻剩余的节点，是因为在独立线性模型具有次模特性。CELF算法利用次模特性，在每一步选择初始种子节点时，大量节点的增量影响不需要被重新计算，这是因为它们的增量影响在之前步骤中的值已经小于其它节点在当前步骤中的值。如在第i步时得到的增量影响为420、400、380、320、280、200、70，在第i+1步时计算得到有一个节点的增量影响为320，则在第i+1步中计算出来增量影响为280、200、70的这些节点在第步就不需要被重新计算了，因为根据次模函数的性质，这些节点的增量影响肯定不会大于节点的增量影响。因此它缩短了Greddy算法的时间。

CELF比Greedy时间复杂度提高了700多倍，可以实现对原始贪心算法高达倍的加速。也正因为如此，可以用它代替原始贪心算法完成挑选个种子节点的工作。改进后的算法称为尽管它的时间复杂度与办算法一样，但是由于从第二开始减少了每一轮需要计算的节点数，该算法的运行时间得以大大降低。能够在较短时间内取得近似最优解。

与原始CELF算法相比，SMC算法考虑了影响旳整个传播过程，每一步都需要计算所有未激活节点的边际影响。刚开始时，所有的节点均处于未激活状态，计算每个未激活节点的边际影响均要遍历整张网络图来进行影响的传播，非常耗时。SMG算法由于找出ck个节点后激活了大量的节点，接着是从未激活中节点中寻找k-ck节点，但是由于己经经历了启发阶段，己经有大量的节点被激活，此时未激活的节点比原始数据集会少很多，相应的会比算法少遍历很多遍网络图，因此的时间复杂度大大降低。

### 算法步骤

在影响力最大化问题上，基于贪心的算法准确性较高，但是运行时间相对过长，而启发式算法虽然所需时间较短，但是准确度却往往难以令人满意。

基于这个出发点，在下一步工作中，以前面的影响力度量算法为基础，提出一种基于两阶段的算法。由于贪心爬山算法的时间复杂度主要集中在前部分的节点选择，所以准备将影响力最大化算法的分为三个阶段：

第一阶段是根据上述公式计算用户的活跃度，去除活跃度为零的用户。

第二阶段是根据第三章计算出的用户的影响其，在所有节点中启发式的选出ck个影响力最大的节点作为激活节点，进行传播。c是启发因子，取值在[0,1]

第三阶段再在剩下未激活节点中选出的中用CELF算法求出剩下的k-ck个节点。

算法的伪代码如下：

社会网络图表示为G=(V,E)，种子节点数k，影响力最大化的种子节点S。

|  |
| --- |
| Input:G=(V,E),k,c |
| Output:S |
| Initialize:S={} |
| 1 for each u∈V  2 calc A(u)  3 if(A(u)==0)  4 Remove(u)  5 end if  6 end for  7 for i=1 to ck  8 select u which make max(inf(u))  9  10 end for  11 for i=1 to k-ck  12 select u which make  13  14 update state  15 end for  16 output S |

## 算法仿真与分析4-6

### 实验数据集

在上一章中，我们使用爬取的新浪微博的数据来验证MIB算法。本节依旧采用该数据做实验。为了进一步保证准确度，将粉丝数少于10的用户从原始数据集中删除。这样，就得到了一个新的子数据集，数据集包含个38426个节点和52473条边。众所周知，社会网络中节点的度会遵从幂律分布，因此，为了进一步验证子数据集的有效性，给出用户粉丝人数分布和关注人数分布，分别如图和图所示。由图得知，它们都服从幂律分布。因此，有足够的理由证明将该子数据集做实验数据集是合理的。

图1

图2

### 实验结果

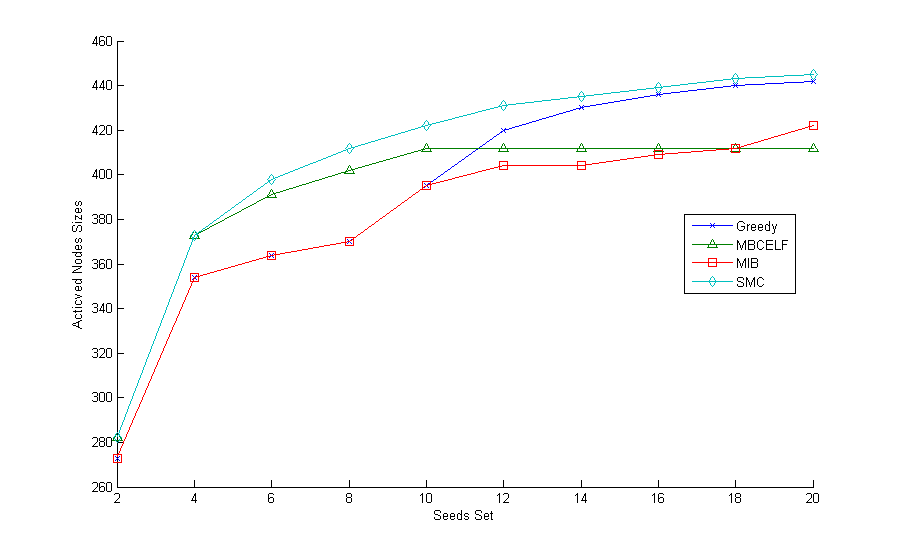
为了验证算法的有效性，在此选择了著名的Greedy算法、MaxDegree算法和MBCELF算法来作比较。

Greedy算法：原始的贪心算法。

MBCELF算法：中科大李翔提出的一种影响力最大化算法。

MIB算法:结合MDD和用户的交互关系的一种算法。

下面是各个算法的实验对比图：



### 实验结论

## 本章小结

# 微博用户影响力评价系统的设计与实现

传

F