分 类 号_____ 学 号 M201472720
学校代码 10487 密 级_____

华中科技大学

硕士学位论文

基于独立级联改进模型的微博影响力最大化算法研究

学 位 申 请 人 ： 来雪停

学 科 专 业 ： 计算机系统结构

指 导 教 师 ： 鲁宏伟 教授

答 辩 日 期 ： 2017 年 5 月 24 日

A Thesis Submitted in Partial Fulfillment of the Requirements
For the Degree of Master of Engineering

Research on Influence Maximization in Microblog Network Based on Improved Independent Cascade Model

Candidate : **Lai Xueting**

Major : **Computer Architecture**

Supervisor: **Professor Lu Hongwei**

Huazhong University of Science and Technology

Wuhan, Hubei 430074, P. R. China

May, 2017

独创性声明

本人声明所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除文中已经标明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本文的研究做出贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到，本声明的法律结果由本人承担。

学位论文作者签名：

日期：　　年　　月　　日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，即：学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权华中科技大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

保密□，在___年解密后适用本授权书。

本论文属于

不保密□。

(请在以上方框内打“√”)

学位论文作者签名：　　指导老师签名：

日期：　　年　　月　　日　　日期：　　年　　月　　日

摘要

社交网络由节点和边构成，节点通常指个人或者组织，边指其关联关系。信息通过社交网络进行传播，影响力最大化即是研究在网络中寻找 k 个目标节点作为信息传播的源点，使这些节点按照某种信息传播机制去影响其它节点，最终使信息达到最大化的扩散范围。研究如何使信息在社交网络中广泛的传播达到网络营销的目的就具有一定的现实意义。

随着互联网技术的不断发展，微博因具有“平民化”使用门槛在国内发展迅速，所以选择在微博网络研究影响力最大化问题。影响力最大化算法的研究主要分为两个部分：贪心算法和启发式算法。贪心算法由于其时间复杂度高而不能用于大规模网络上，启发式算法由于最终得到的目标节点影响力效果不理想也不被广泛使用。由于在线社交网络用户量庞大以及关系复杂，需要大数据集下快速准确的找到目标节点集合，现有算法的时间复杂度和准确性仍待改进，如何在大规模数据集下依据网络平台本身特性来提高算法结果的有效性和减少算法的时间消耗仍是值得研究的问题。

通过对微博数据进行分析，发现微博网络具有无标度性、异配性、以及较小的聚类系数等，同时网络中节点之间的连接存在不平等的偏斜关系，节点在网络中具有不同的可用度。传统独立级联模型没有考虑网络节点的以上特性，因此在节点间影响大小的度量上存在偏差。基于此，改进了传统独立级联模型中对节点间影响大小的计算方式，构建了基于节点偏斜关系和节点可用度的微博信息传播模型 IDSA（Information Diffusion Based on Skewed Relationship and Availability of Node），并提出了基于 IDSA 的影响力最大化算法，该算法利用节点影响力的有限性和覆盖性改进了贪心算法中对节点影响力增量的计算，将其限制在 4 层好友范围并且忽略被覆盖节点的计算。最后，分别在新浪微博和腾讯微博数据集上实施该算法并与传统的算法做比较，结果显示该影响力最大化算法有较好的表现。

关键词：病毒式营销，影响力最大化，微博网络，信息传播模型

Abstract

The social network is composed of nodes and edges. The spread of information is by means of the interaction between nodes, which is often the formation of “word of mouth”. Information can quickly reach large-scale diffusion in the network based on the principle of it named as viral marketing. The study of how to make a wide range diffusion of information on social networks to achieve the purpose of network marketing has a certain practical significance. Influence maximization is to find k target nodes in the network as the source of information dissemination, and make these nodes to affect other nodes continuously with a certain information dissemination mechanism to achieve the max diffusion of information ultimately.

With the continuous development of Internet technology, online social network plays an increasingly important role as a platform for information dissemination in people's lives, which microblog has developed rapidly. This paper studies the influence maximization in microblog network. The greedy algorithm has a high time complexity, and the heuristic algorithm does not perform well in information diffusion. In spite of some improved algorithms, it still cann’t meet the specific needs. How to improve the effectiveness of results and reduce the time consumption of the algorithm according to the characteristics of the network platform itself under the large-scale data set is still worthy of study.

By analyzing the microblog data, it is found that there are unequal skewed relationship and different availability of nodes in the network. Based on this, this paper improves the calculation of the influence between nodes and constructs the information diffusion of microblog named as IDSA. The algorithm for influence maximization based on IDSA is proposed later, which improves the time-consuming step in the greedy algorithm. The algorithm is implemented on the data set from the Sina and Tencent microblog, and compared with the traditional algorithm. The results show that it has better performance in time complexity and influence range.

Keywords: Viral Marketing, Influence Maximization, Microblog Network, Information Diffusion Model

目 录

1 绪论

1.1 研究背景与意义

网络是由节点以及节点之间的边形成，社会生活中也存在各种网络，如铁路网、电话通信网等。而社会关系网络中的节点指个人或组织，他们的关系包括朋友关系、同学关系、亲戚关系以及合作关系等。对于社会网络的分析研究起源较早，社会网络在生活的方方面面随处可见，对其的研究也显得非常重要。

互联网和 Web2.0 技术的出现带来了信息产生方式和传播模式的深刻变革，增进了人与人之间的交流。现在人们往往通过在线社交网络来接收和传播信息。在线社交网络上的信息传播形式和一般社会网络的区别在于信息传播是通过一些社交应用网站，比如微博信息的转发。而且社交网络上的信息产生往往出于广大用户群体，即自媒体形式。区别于传统媒体，自媒体限制较少且实时灵活，因此发展迅速。

微博中用户发布信息被其好友转发，并通过好友的好友继续转发，信息便通过这种方式得到扩散。通过微博我们可以关注好友、发布信息、评论以及转发信息，发布的信息可以包含图片和视频、音频等。因此微博是个灵活的社交媒体，也是相对活跃社交平台。在上面我们不仅能看到所关注好友的动态，更能实时看到一些热点信息，了解身边乃至世界所发生的大事。所以微博不仅是社交平台，更是信息传播载体。再加上其平民化的使用门槛，使微博能够得到较广泛的普及。

微博中的信息传播具有以下几个特点：1.内容的原创性，由于其操作简单，用户可以上传一张图片或者编辑一句话就可以发布一篇简短的微博，所以微博信息的内容原创性较强；2.信息传播去中心化，相对于传统媒体的信息传播，微博的每一个用户所发布的信息都有可能成为话题的焦点，所以不再是依靠某些媒体才能传播和获得信息；3.信息传播实时快速，在手机网络普及的时代，用户只需要简单的在线转发即可分享信息，因此会使信息在短时间内跨地域的大范围传播。

基于这种情况，很多人开始致力于研究如何使信息在这些社交网络中广泛的传播达到网络市场营销的目的。基于成本等因素考虑，我们不可能向网络中所有用户推荐我们的新应用，而是选择某些用户集合，使这些用户来宣传我们的产品，去影响尽可能多的人也来接受它，从而使其达到最广泛的信息传播效果。这就是社会网络影响力最大化问题又称信息扩散最大化问题。

社交网络影响力最大化问题首次由 Domingos 和 Richardson[1]提出的，节点选择是该问题的关键，节点与节点之间的影响关系是信息得到传播的依据，个人的影响力应该考虑网络的全局影响而并非单单对邻节点的影响，多节点进行信息传播需考虑节点间的影响力覆盖问题。所以影响力最大化问题是比较复杂的综合性问题，需要克服各种难题，这当然也包含效率问题。因其具有现实的社会意义（在 1.1 节开始已经提到过），所以也是值得研究并不断改进的。

在线社交网络的影响力最大化问题只是把上述问题从社会网络移到在线社交网络，其中注册用户是节点，好友关系是边，信息传播通过转发好友实现，传播平台变成这些社交网站。

1.2 国内外研究现状

1.2.1 信息传播模型

信息传播模型定义了信息是如何从一个节点传播到另一个节点的，即节点之间是如何相互影响的，以及节点在什么情况下接受某条信息等。社交网络 $G(V,E)$ 是一个有向图，V 和 E 分别是节点集合和边的集合，在信息传播模型中，每个节点有激活（接受）与非激活（未接受）两种状态。节点处于激活状态表明其参与了信息传播过程，在微博网络中说明用户转发了某条信息。被激活的节点又会通过一些途径去传播信息给自己的相邻节点。

那么在有一些初始节点情况下，这些节点是如何去传播信息给相邻节点的？如何去预测传播过程中节点是否被激活？信息传播模型解决这些问题，下面介绍信息传播模型的研究现状。

独立级联模型 IC[2][3]（Independent Cascade Model）和线性阀值模型 LT[4]（Linear Threshold Model）是两种传统的信息传播模型，在 IC 模型中，假设节点 u（$u \in V$）在某一时刻 t 变为激活状态，则 u 有且只有一次独立的机会以概率 $p_{u,v}$ 去激活其邻节点 v，如果 v 被激活则可以继续以相同的方式去激活其邻节点。而在 LT 模型中，激活节点 u 对 v 有一个影响权重 $b_{u,v}$，如果 v 的所有被激活的邻节点对其影响权重之和超过某个阀值 θ_v，则 v 也将被激活。

在 Wang[5]等人的文章中，改进了线性阀值模型中对于 $b_{u,v}$ 的计算方式，传统的计算是 $1/d(v)$（节点 v 的度数求倒）。Wang 等人认为节点间的影响权重不仅与邻节点有关系，还和邻节点之间的互相连接程度有关系。所以用

$b_{u,v} = d_{NG(v)}(u) \Big/ \sum_{w \in NG(v)} d_{NG(v)}(w)$ 定义节点 u 对 v 的影响权重，其中 $NG(v)$ 为 v 的邻居节点构成的子图。

考虑到信息本身的特征，多种信息在网络中同时传播，可能会有相互促进的作用，也可能会有相互抑制的作用，LT 的拓展模型[6]研究了相互竞争的两种信息同时传播时的相互影响。Bharathi 等人[9] 拓展了 IC 模型，建立了多种信息在网络中扩散的传播模型，该模型中节点的激活状态包含多种信息对其的激活情况。

考虑时间因素，时间限制不断激活模型[10]（Continuously activated and Time-restricted IC）是对独立级联模型的拓展，在该模型中，节点在一定时间 t 内不停的去激活周围的邻居节点。每次激活的概率满足一个递减的时间函数 $f(t)$。在实际社会网络中，如果节点 v 传播了一条信息，那他的邻居 u 可能不止一次看到 v 的这条信息，所以每次都会对 u 产生影响，只不过影响力在减小。

博弈论模型[11]中用户衡量其社会关系和信息考虑是否接受某条信息，两者采取对本身利益最大化的方式来决定。用户所处的社会关系中，好友对其的影响是不同的；信息本身存在不同，是否满足用户偏好或者增加用户利益都是用户接受信息时所要考虑的。用户会最终根据对自己最有利的方式来做出选择。

传染病模型[13]也是经典的信息传播模型，因信息传播的特点与其类似，都是从最初的几个感染节点（激活的节点）开始，不断影响周围的节点，使其或受感染，然后再去影响其它节点。传染病模型分为 SIS（Susceptible-Infected-Susceptible）和 SIR（Susceptible-Infected-Recovery）两种，SIS 模型节点有两种状态，已经被感染的和未被感染的，被感染的节点在某一时刻可能会变为未被感染的状态，然后有可能再次被感染。SIR 多了一种处于免疫状态的节点，即受过感染又恢复正常并具有免疫能力的节点。

在传统的信息传播模型中，p 往往随机产生，b 也被设定为节点度数的倒数。然而在真实的社交网络中，用户是否被激活的影响因素有很多，p 和 b 应该是具有用户特征的变量值。为了更细致的衡量用户间的这种影响力大小，Tang[16]等人提取用户感兴趣的话题，在微观上度量节点间的信息传播概率。Barbieri 等人[17]对 LT 模型进行研究，提出了基于话题信息的 TLT(Topic-aware Linear Threshold)模型。

Zaman[18]等人以 Twitter 网络提出协同过滤信息传播模型，Wang[19]等人提出了基于情感分析的独立级联信息传播模型，他们提出每种信息都含有不同的情感，不同情感的的信息对节点的影响又是不同的，从而微观上度量了信息对节点的影响大小。

1.2.2 影响力最大化算法

影响力最大化问题具有一定的现实意义，也受到越来越多的研究者的关注。Domingos P[1]等人首次提出影响力最大化问题，之后对于信息传播模型和影响力最大化算法的研究被相继提出。

首先是贪心算法及其一系列优化算法。Kemple[20]等人首次提出了用来解决影响力最大化问题的贪心算法。他们提出的贪心算法循环多次在全局网络中选取具有最大影响力增量的节点作为信息传播的初始目标节点。假设当前的目标节点集合为 A（$A \subseteq V$），下一个选取目标节点的办法是将每一个集合 A 之外的节点 v 加入 A（$\{v\} \cup A$），计算加入 v 之后的增量影响力（即增加的激活节点个数），选择具有最大增量影响力的那个节点加入集合 A。Kemple 等人的实验是基于以上介绍的两种理论模型-独立级联模型和线性阀值模型，模拟信息扩散并以信息传播效果来衡量节点的影响力。算法本身其实并不复杂，但是算法的第二步是非常耗时的，每选择一个目标节点后即需要重新计算目标集合外所有节点的增量影响力，而且每个节点的影响力增量要迭代计算上万次（一般取 20000 次），所以即使该算法选出的种子集合具有最优的传播效果也无法被使用到大规模网络计算中。

之后，Leskovec[21]等人利用影响最大化目标函数的子模性改进贪心算法提出了有效的 CELF（Cost-Effective Lazy Forward）算法。由子模型性定义 $\delta(\{v\} \cup A) - \delta(A) \geq \delta(\{v\} \cup T) - \delta(T)$，当 $A \subseteq T$ 时成立（δ 表示节点或集合的影响力大小）。所以就单个节点 v 来说随着节点集合 A 的增大，v 的影响力增量是减小的。如果记录当前一轮迭代计算的节点影响力增量，在下一轮计算时如果出现比当前记录大的值时，可以不必再计算小于该值的节点影响力增量。因为节点的下一轮计算结果肯定要小于当前值。这样，就可以省去大量节点的增量影响的重复计算。

CELF 算法在效率上比 KTT 提高大概 700 倍，并且精确度和 KTT 基本一致。虽然 CELF 算法效率上有所提高，可即便如此，对于大规模网络还是不能满足需求[22]。在 Leskovec 等人的基础上，CELF++[23]和 UBLF[24]等 CELF 的改进算法被提出，CELF++算法比 CELF 快 35-55%。

Chen[25]等人在 IC 传播模型上针对贪心算法提出改进的 NewGreedy 算法。它的主要精髓在于利用独立级联模型的特点，对于节点 v，它的每个邻节点都会尝试独立的以概率 p 来激活 v，如果在一次尝试失败后，就不再对 v 进行激活。所以该条链接便不起任何传播作用。可以据此对网络图进行优化，删除所有对传播无用的边。这

样得到子图进行每次迭代。对于它的算法复杂度由一般贪心算法的$O(knRm)$提高到$O(kRm)$（其中 k 指目标节点个数，n 指节点数，R 是每次迭代重复计算节点集合影响力的次数，m 是边数）作者通过实验并证明它是有效的。之后作者又提出该算法的改进算法 MixedGreedy，它基于 NewGreedy 的思想，计算每个节点影响力的同时删除不起传播作用的边，得到优化子图，在以后的每次迭代使用 CELF 算法。

Chen[22] 等人在之后的文章中又提出了基于 MIA（Maximum-Influence-Arborescence）模型的算法，该算法结合最短路径和影响传播概率，将节点的影响力限制在一个阀值θ之上。在此基础上减小贪心算法中计算节点影响力的时间复杂度，另一方面也避免导致仅仅使用最短路径计算影响力造成传播范围较小。

以上均是基于整体网络考虑影响力最大化算法，Galstyan[28]等人首次提出从网络结构属性出发的局部优化策略，把网络划分成两个社区，每个社区内使用简单贪心算法，这种算法的性能往往受到网络结构的影响。Tian[29]等人的工作类似，只是每个社区使用 HighDegree 算法。

在 Galstyan 等人的基础上，Yu Wang[30]等人提出基于社区的贪心算法 CGA，CGA 算法用社区内部用户的影响力来近似代替全局影响力。该算法分为两步，先将网络依据组合熵进行划分成多个社区，再基于 MixGreedy 算法在每个社区选择种子节点。其中组合熵定义为$\max\{\bar{R}_m(\{u\})/R_m(\{v\})\,|\,v\in C_m, u\in L[C_m], u\in C_l\}$，其中$L[C_m]$为社区$C_m$内的节点在其他社区内的激活节点个数$R(\{v\})$表示节点 v 在社区C_m中的影响度。计算节点的影响力在社区的而非整个网络，减少了计算节点影响力的复杂度。但该算法仍旧存在一定的问题，就是在网络中的社区大小规模不一，在社区划分结束后有可能存在结构失衡即有的社区规模极大，有的社区规模较小，这就使网络社区划分失去意义。

即使有不断改进，百万数据量的情况下时间问题仍是最大的问题，基于节点度和距离中心性的启发式算法[31]是具有代表性的高效算法，但该类算法牺牲了结果的准确性。他们将网络节点按照节点度或者节点中心性排序，选择 Top-K 节点作为目标节点。该类算法所达到的信息传播范围并不高。PageRank 算法[33]用来对网络节点的重要性进行排序，也被用来进行影响力最大化算法研究，还有一些依其拓展的相关研究[34][35]。

DegreeDiscount 算法[25]是基于独立级联模型的启发式算法，该算法描述为度折扣式算法。它分 k 次迭代，每次重新计算部分节点的折扣度数，一次迭代选择一个最

大度数的节点作为种子节点。前一种传统的算法会出现邻居重叠问题，从而缩小了传播的范围；而后一种正是在每次迭代过程中排除这种邻居重叠。已选目标节点的邻节点在计算新的度数时，要减去对已选目标节点的作用。该算法较之一般的度影响力算法在传播范围上有所改善。

Zhao[37]等人提出基于 k 核分解的影响力算法。田家堂等人[38]也结合贪心和启发式算法，提出了混合式算法--HPG 算法，他们在启发式阶段选择度数较大的节点作为目标节点，然后又使用了贪心算法，该算法的稳定性并不是很好。

考虑一个节点可能同时参加多个在线社交网络 OSN，Zhang[39]等人提出了多个 OSN 的影响力最大化问题。而 Du[40]等人考虑在一个网络中同时宣传多个产品的影响力最大化问题。Lei[41]等人认为应该首先度量节点间的影响概率，采用真实网络数据的反馈信息来不断改进影响力最大化目标节点的选择。

考虑到对于不同话题的信息在网络中的传播效果是不一样的[42]，Srinivasan[43]等人提出了基于话题的影响力最大化算法 Targeted Influence。每个网络用户对不同话题保持一个分布向量，代表用户的个人偏好，从而计算某种话题的信息对用户的影响大小，Chen[44]等人的工作类似。随着各种社交媒体的兴起，有针对性的进行影响力最大化算法变得越来越有意义[45]。近些年也出现一些针对具体社交平台的信息传播研究[18][47]。

影响力最大化算法的研究主要分为两个部分：贪心算法和启发式算法。贪心算法由于其时间复杂度高而不能用于大规模网络上，启发式算法由于最终得到的目标节点影响力效果不理想也不被广泛使用。尽管有针对性的进行了优化，仍不能满足具体网络平台的需求，如何在大规模数据集下依据网络平台本身特性来减少算法的时间消耗仍是值得研究的问题。本文针对微博社交网络平台，提出适用于微博网络的基于节点偏斜关系和可用度的信息传播模型 IDSA（Information Diffusion Based on Skewed Relationship and Availability of Node），以及基于该信息传播模型和传统贪心算法提出改进的影响力最大化算法。

1.3 研究内容

影响力最大化问题在网络营销等方面有着重要的意义，微博已经成为巨大的营销平台。所以就需要在微博网络上进行影响力最大化算法研究。对于真实在线社交网络，传统的影响力最大化算法并不能快速、有效的找到使信息最大程度扩散的目

标节点集合，因为没有考虑到网络的节点特性。针对此问题，本文将要研究的内容包括以下几个部分：

1. 对微博网络进行影响力最大化算法研究需要了解微博网络的特点，有针对性的实施有效的算法设计。本文先从国内流行的两大微博站点（新浪和腾讯微博）爬取数据。数据内容包含网站用户信息、用户关系以及用户发布微博，从而进一步建立微博关系网，分析微博网络结构特征和节点属性，充分了解微博用户的状态和行为。

2. 微博中信息传播是通过激活用户影响其好友用户，并使激活的好友用户其影响更多的好友用户。如何衡量这种用户之间的影响大小，提出符合微博特征的微博网络信息传播模型是本文研究的内容之一。

3. 微博用户数量庞大，形成的关系网络更是复杂多变，如何设计影响力最大化算法，使信息扩散达到较高水平并且使消耗的时间尽可能的低是本课题的另一个重要研究内容。

1.4 论文结构

全文的组织结构如下：

除了本章内容之外，第二章为微博数据获取和分析，介绍了如何从新浪微博和腾讯微博上爬取数据，以及数据的分析筛选和存储过程，并展示了最终可用的数据形式。然后根据用户关系数据建立了微博网络结构，即为有向的好友关系图；并对该有向图进行属性测量和计算，本文测量了网络的度分布、聚类系数、同配性以及 k-shell 分布等。同时利用用户微博数据测量了用户微博发布时间模式和用户交互模式。最后总结了微博网络具有偏斜关系，以及节点的具有不同的可用度，从而为下一章提出基于节点偏斜关系和可用度 SA（Skewed Relationship and Availability of Node）的信息传播模型做准备。

第三章在第二章的基础上，利用相关结论衡量用户之间的影响大小，并量化一些关键指标，结合传统 IC 信息传播模型，建立符合微博特征的基于 SA 的信息传播模型 IDSA。在此模型基础上设计了有效的基于 IDSA 的影响力最大化算法。

第四章将本文提出的影响力最大化算法在新浪微博和腾讯微博数据集上分别进行实验，并分别在信息传播范围的大小以及时间复杂度上和几种传统的影响力最大化进行比较，实验结果表明本文算法在微博网络影响力最大化问题上有较好的表现。

最后是对论文的总结和展望，对本文所做的工作和研究结果进行总结，并指明今后的研究方向。

2 微博数据获取与分析

影响力最大化问题也即信息传播最大化问题，获取足够完整的微博数据和进行全面的微博相关数据分析是进行下一步研究微博信息传播和信息传播最大化的重要基础。

2.1 微博数据获取

所需要的微博数据大概包含三个方面：微博用户关系、用户自身信息以及用户发布微博。为了得到这些数据，采用网页爬虫技术先从新浪和腾讯微博网站找到相应的信息页面，获取页面链接，从而依据链接向服务器发送页面获取请求，再对获取到的页面进行解析处理，存储所需要的数据并保存，依次递归下去，不断获得新用户数据，下面是具体的实施流程。

为了实现数据爬虫，需要解决数据的获取问题，所以需要了解数据的网络传输原理。网站的超文本网页一般通过 http 或者 https 协议进行数据传输，通过特定的地址链接，发送定向 http 请求报文，通过服务器网站验证即可接收到对应网站的应答报文。对于新浪微博和腾讯微博的 host 网站地址分别是 *http://weibo.cn* 和 *http://t.qq.com*。为了得到新浪微博某个用户 *yangmiblog* 的主页信息，可以发送表 2-1 的 http 请求报文。

表 2-1 微博主页请求报文

GET*http://weibo.cn/yangmiblog HTTP/1.1*
Host: *weibo.cn* **User-Agent:** *Mozilla/5.0 (Windows NT 6.1; rv:51.0) Gecko/20100101 Firefox/51.0* **Accept:** *text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8* **Accept-Language:** *zh-CN,zh;q=0.8,en-US;q=0.5,en;q=0.3* **Accept-Encoding:** *gzip, deflate* **Referer:** *http://weibo.cn/pub/top?cat=star&pos=65* **Cookie:***SUHB=06ZVSizU8ankY6;* *_T_WM=3063883091492a68ab11985f6d85926e* **Connection:** *keep-alive* **Upgrade-Insecure-Requests:** *1*

在这之前需要进行用户登录验证，登录成功后获取到本地的 cookie 数据，并用其来包装 http 请求，服务器验证通过之后，会发送表 2-2 的响应报文。

表 2-2 服务器响应报文

HTTP/1.1 200 OK
Connection: *close*
Content-Encoding: *gzip*
Content-Length: *4763*
Content-Type: *text/html; charset=utf-8*
Date: *Wed, 01 Mar 2017 07:56:29 GMT*
LB_NODE: *layer7-002.mweibo.se.sinanode.com*
PROC_NODE: *web354.mweibo.bx.sinanode.com*
Server: *Tengine*
Set-Cookie: *WEIBOCN_FROM=deleted; expires=Thu, 01-Jan-1970 00:00:01 GMT; path=/;* **domain=***.weibo.cn*
Vary: *Host,Accept-Encoding*

报文表示成功返回所需的数据，格式为 gzip，需要进行解压，解压后的报文的内容为 *text/html*，截取部分返回内容如表 2-3。

表 2-3 微博主页

<html xmlns="http://www.w3.org/1999/xhtml">
<img src="…" alt="M"> **女/北京**
加关注
认证演员，代表作《宫》《仙剑奇侠传三》《我是证人》等
这里有一只狐狸，幸福，感恩，知足，爱~＞＜
微博[3452]
关注[650]
粉丝[70659100]
#十里桃花妆#
听说最近很多小仙女仿了浅浅的妆啊，唇妆是关键哦，粉嫩柔润才是正宗桃花唇哦~ 你们猜我和浅浅最近涂的是
@雅诗兰黛ENVY 唇膏哪个颜色~
<img src="…" alt="图片" class="ib">
赞[158483]
转发[246929]
评论[18051]
今天 10:10 来自 OPPO R9s

从以上返回的报文信息中，可以通过网页解析提出该用户的基本信息：性别、住址、身份、关注数以及粉丝数；同时也可以提取用户发布的微博信息：微博内容、发布时间、转发数以及评论数等。

以上过程已经可以获取某个用户的部分信息，为了收集用户的好友关系数据，构建微博用户关系网络，还需要向服务器发送其它链接请求（该请求的链接也在上一个 http 的返回报文里），按照相同的方式即可获取到想要的数据，表 2-4 是部分获取的数据。

表 2-4 微博好友列表

<html xmlns="http://www.w3.org/1999/xhtml">
杨幂**关注**的人
<a href="http://weibo.cn/**esteelauder**">雅诗兰黛
<a href="http://weibo.cn/u/**1751675285**">岳云鹏
<a href="http://weibo.cn/**zhangjike1988**">张继科
<a href="http://weibo.cn/**vivawayne**">刘芮麟 Wayne
<a href="http://weibo.cn/u/**1993427351**">PAIPAI 王俐丹

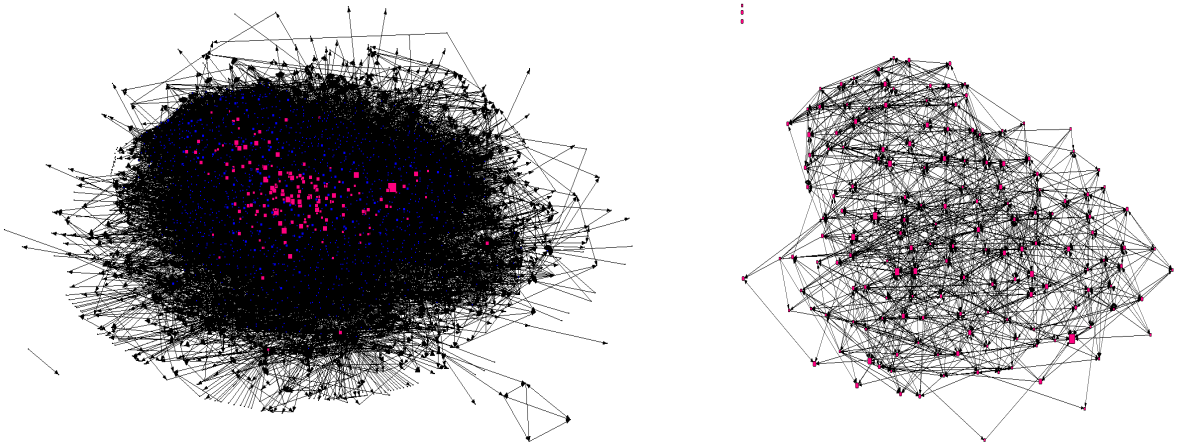
从该页面内容可以将该用户所关注的好友用户的 ID 提取出来，写入好友关系文件；同时可以提取好友主页的链接地址，这样便可以采用广度优先搜索的方式从一个用户遍历好友用户，再从好友用户遍历好友的好友，最后获取到的是一棵多路关系树。由于仅从一棵树无法描述整个微博网络结构，数据爬虫开始阶段就选择多个种子用户，这样才能抓取到更加全面的数据。多个种子用户在遍历好友用户时可能会出现重复访问的情况，这样得到的数据也会重复。为了避免这种情况，在算法中保持一个用户访问 hash 表，对将要访问的用户进行唯一性验证。从一个节点出发广度搜索，会形成一个局部较完整的网络结构，只是在网络的边缘节点可能会有部分关系的缺失。

算法为了爬取效率，一共同时开启了十个线程。将每次提取的用户页面链接都加入到一条将会被处理的就绪队列，十个线程互斥的访问该队列，共同进行上述的处理过程，并将提取到的可用数据保存到文件中，文件包含用户信息 userinfo.txt、用户关系 friendinfo.txt、微博 userweibo.txt。数据爬虫大约持续两天时间，共获得腾讯微博用户节点 2449 个（每个节点包含大约 15 条微博信息），形成的网络边为 78420 条；新浪微博数据集节点个数为 18903，边有 54172 条。

2.2 微博数据特性度量

本节将根据用户关系数据建立微博网络结构；并对该网络进行属性测量和计算，包括网络的度分布、聚类系数、同配性以及 k-shell 分布等。同时利用用户微博数据测量用户微博发布时间模式和用户交互模式。

用节点来表示微博用户，用边来表示用户之间的好友关系，则把微博用户关系网络抽象成一个有向图，从获取的数据集可以构造出微博网络图，图 2-1 通过 ucinet 展示了腾讯微博网络的整体结构。



(a) (b)

图 2-1 腾讯微博用户好友关系结构图

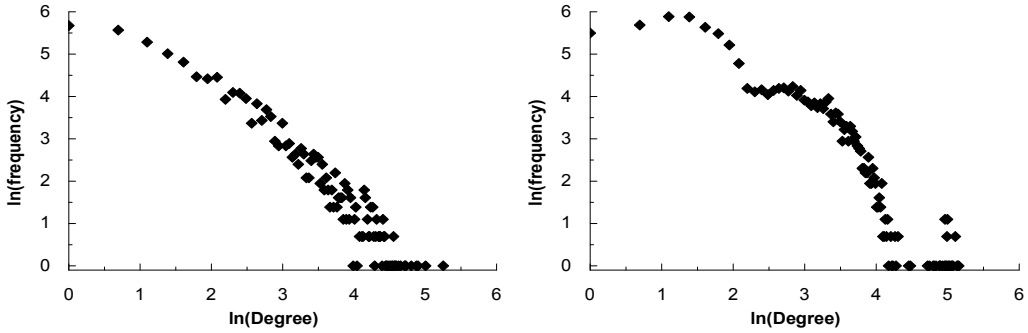
图 2-1(a)为腾讯微博网络整体结构，节点的大小表明了用户的度数大小，图 2-1 (b)是提取部分度数超过 50 的节点构成的网络图。可以看到该网络结构错综复杂，仅仅 2000 多个用户就有将近 80000 条边。

2.2.1 度分布

节点的度指的是网络节点所连接的邻节点个数。对于有向网络，又分为入度和出度。度分布是对节点度的规律的一种描述，通常用 $P(k)$表示任意选择一个网络节点，其度恰好为 k 的概率。

在线社交网络 OSN 的度分布大都呈幂率分布（指数为负数）。幂率特性也称为无标度性，图表上表现为“长尾”分布。经济学中理解为二八原则，即少数人聚集了大量的财富，而大多数人的财富数量都很小。

图 2-2(a)、(b)分别描述腾讯和新浪微博的度分布。图中横坐标 *ln*(*Degree*)表示节点度数取对数的值，纵坐标 *ln*(*frequency*)表示节点度数的频率取对数的值（这里均用节点的入度）。如果节点度数分布为幂率分布，则分别取对数后应该满足斜率为负的线性分布。



(a) 腾讯微博　　(b) 新浪微博

图 2-2 度分布图

图 2-2(a)中，腾讯微博分布大致服从幂率分布，幂指数近似为 1.2(斜率为-6/5)。图 2-2(b)中新浪微博的度数分布比较特殊，如果去除某些节点，可以看到其大致可看作两段不规则的幂率分布，前半部分下降趋势平缓，幂指数约为 2。在横坐标为 2-3 之间有部分缓冲，然后急剧下降，幂指数上升约为 2.7。新浪微博有这种明显的截断分布，可能是其用户关系的两极分化现象，即小度数用户较多而大度数用户相对较少。Mislove 等人[48]统计 Flickr、YouTube 以及 LiveJounal 等社交网络的度分布幂率指数均在 1.5-2 之间。

对于腾讯微博和新浪微博，可以看出网络中存在较少的高度数节点（称为明星节点）和较多的低度数节点。明星节点的周围由于附庸者大量的粉丝节点而在网络中地位较高。这样的社交特性也决定了微博网络呈幂率分布。在这样的网络中，明星对粉丝的影响力是巨大的，如果一条用户关系是这样的，则从明星到粉丝之间的信息传播便较为容易。

度分布中由于统计的是节点入度和频率的关系，数据量小就可能使节点频率较低，从而使统计的幂律指数偏小，文章所给出的幂指数大小可能存在偏差。

2.2.2 同配性和 Kshell 分布

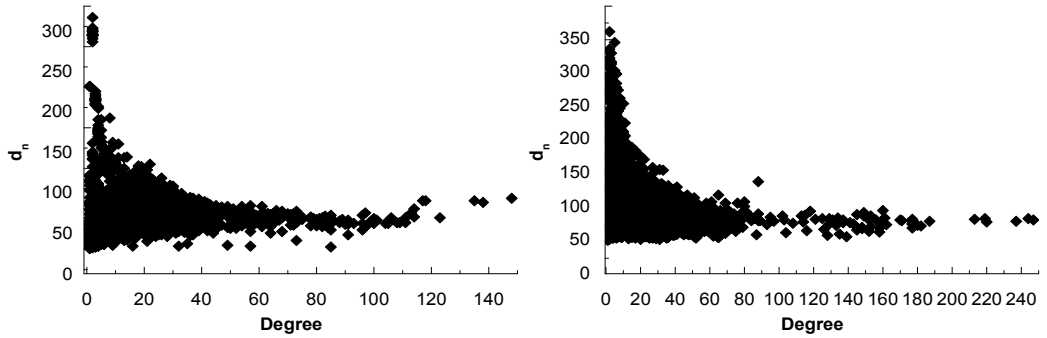
网络的同配性即节点的度相关性，指的是网络中节点与其度数相似节点的链接倾向性。若度数较大的节点趋向于和度数较大的节点链接，度数较小的节点趋向于

和小度数节点链接，则认为网络具有同配性，否则认为网络具有异配性。通常用 *Pesrson* 相关系数[49]来描述网络的度相关性：

$$r=\frac{c\sum_i u_i v_i-[c\sum_i 1/2(u_i+v_i)]^2}{c\sum_i 1/2(u_i^2+v_i^2)-[c\sum_i 1/2(u_i+v_i)]^2} \quad \text{(式 2-1)}$$

其中 c 为网络中边的倒数，u_i和v_i分别为边 i 对应的两邻节点的度数。对于在线社交网络，出于社交目的自发形成的虚拟型网络，节点用户往往会趋向于链接与其地位相似的节点，所以网络一般呈现的是同配性。如 Facebook 的同配系数为 0.17[50]，Flickr 的为 0.202[48]。而对于腾讯微博其同配系数为-0.082，新浪微博求出的结果为-0.079。

对于腾讯和新浪微博网络出现异配性，可能也是由于其自身平台特性决定的，一般对于明星用户来说更倾向于添加明星用户为其好友；对于一般度数小的粉丝节点，出于追星心理往往会附庸明星节点，而这样的节点又占网络的绝大部分。用 d_n 表示节点的平均连接度数，图 2-3 分别表示腾讯微博和新浪微博的 d_n 分布。



(a) 腾讯微博　　(b) 新浪微博

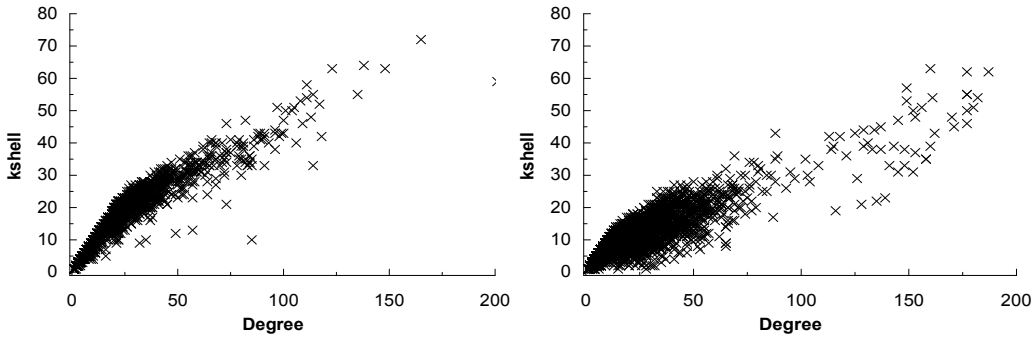
图 2-3 d_n 分布图

从图中也可以看出在节点度数比较小时，节点的 d_n 大都要比自己的度数大，说明小度数的节点趋向于连接度数大的节点；当度数逐渐增大时，节点的 d_n 值先下降后平缓（腾讯有上升趋势），与节点自身度数相差不大，说明度数大的节点更趋向于连接度数大的节点，这与之前的分析大致相同。

节点同配性涉及节点的连接好友度数和自身度数差异的比较，数据量大小影响了所连接节点的度数，因此节点的平均好友度数可能偏低，所以导致部分度数大的节点可能成异配性，从总体来看，缺失的未抓取到的其它节点的同配性没有加入统

计，可能会导致网络同配性分析的结论有所偏差。

Kshell 又称 K-核（K-core）。将一个网络中度数小于或等于 k 的节点及其连边一次去掉，剩下的节点所构成的网络即为 k 核网络，而被去除的节点的核数均为 k-1。微博网络中度数大的节点间相互连接，而度数小的单向连接于度数大的节点，导致度数大的节点称为网络的核心，所以微博网络中度数大的节点往往其核数相对较高。



(a) 腾讯微博　　(b) 新浪微博

图 2-4 Kshell 分布图

图 2-4 分别表示了腾讯和新浪微博的 Kshell 值随度数分布情况，从图中可以看出无论腾讯微博还是新浪微博的用户 Kshell 值均随度数的增加而增加，说明度数越大的节点其所处的位置越是核心位置。相比于腾讯微博网络，新浪微博的节点 K-核普遍偏低，说明腾讯网络中的节点之间连接关系更为复杂。Kshell 值可以作为节点影响力大小的衡量标准，Holthoefer 等人[51]发现核数更大的节点对影响信息传播的作用更大。

对于 kshell 值取决于网络节点的关系复杂程度，缺失的外围节点和关系会导致节点的核数较小，而缺失的节点 Kshell 值可能会导致 Kshell 和度数的相关关系结论存在偏差。

2.2.3 聚类特性和社区性

聚类系数描述了整体网络的聚集程度，节点的聚类程度表示了节点的好友相互为好友的情况，如果好友之间也为好友则节点的聚类系数相对就要越高。假设节点 i 与 k_i 个节点相连接，则他们最多会形成的边数有 $k_i(k_i-1)/2$ 条，现假设节点 i 的实际好友连接边数为 E_i，则节点 i 的聚类系数为:

$$\frac{2E_i}{k_i(k_i-1)/2} \quad \text{(式 2-2)}$$

图 2-5 绘制了腾讯和新浪微博节点聚类系数与节点度数之间的关系图，图中节点的聚类系数与度数呈负相关关系，度数越大的节点反而聚类系数要小。腾讯微博网络的聚类系数普遍高于新浪微博，腾讯网络平均聚类系数为 0.177，新浪微博为 0.037。对于明星节点由于粉丝用户过多，其维护的好友关系相对就较少；相反度数小的节点其好友关系更加聚集，相比于腾讯微博网络，新浪微博的这种现象可能更明显。

一般而言，如果相同度数节点聚类系数越大的那个，其好友之间连接关系越紧密，好友之间的影响可能也越大，这样的关系间进行信息传播的可能也越大。

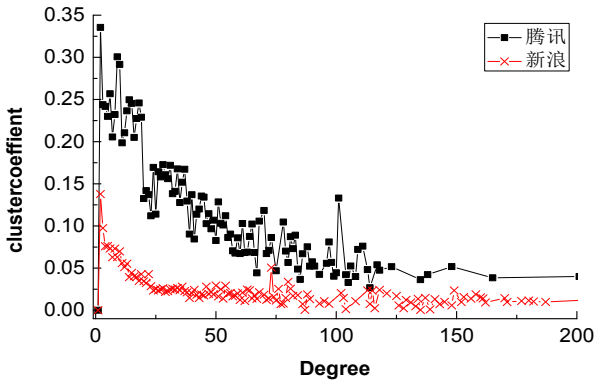


图 2-5 节点聚类系数分布

缺失的关系和节点对现有节点的聚类系数的影响是不确定的，如果随着缺失关系的增加会使节点的聚类系数增大；而增加缺失好友节点又可能会使聚类系数减小。

同节点的聚集性类似，网络的社区性指的是网络可以被分割成多个小型的区域，区域内部的节点之间联系比外部节点来说相对紧密。同样的，社区内部信息传播的可能性要比社区间要大。

网络中具有某种关系的节点之间形成特定的小团体称为社区。这种关系可以是同学关系，也可以是工作伙伴关系。在社区内部的节点之间共同分享信息，互动频繁。图 2-6 是依据 Zhao[37]等人的社区发现算法，将腾讯微博网络划分为 23 个社区，Q 为 0.35（作为度量社区划分优劣的标准，算法划分 23 个社区得到的该值最大）。可以看到腾讯微博网络具有明显的社区特性，社区的规模大小不一。

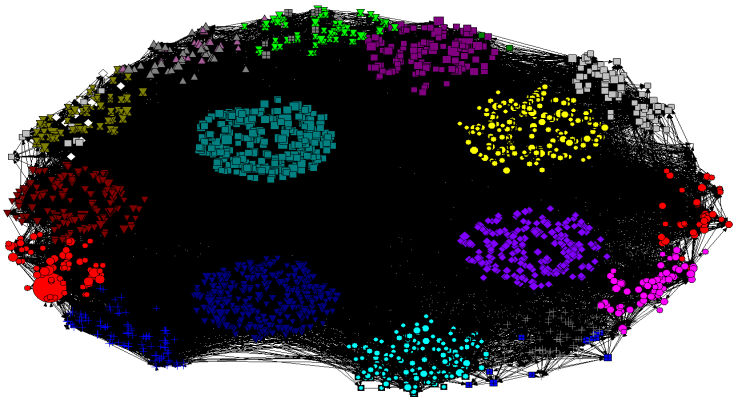


图 2-6 腾讯微博社区结构

2.2.4 交互模式和时间模式

以上利用微博好友关系数据来建立微博网络，从而进行微博网络的各种特性分析，本节对用户发布微博数据进行用户行为分析。从用户所发布的微博数据，可以看出用户的交互模式或者交互状态，即是偏向于转发还是原创；同时也可以看出用户的微博地位，是被大量追随者转发还是无人问津。

通过分析用户发布的微博数据，来对腾讯微博和新浪微博的用户的交互模式进行衡量。大致将用户所属的类型分为以下几种：原创类型但不被他人转发；原创类型并被大量转发；转发类型又被大量转发；转发类型但不被再次转发。原创类型指用户发布的微博是原创并非来自其他人，故而转发类型即是微博来自于分享。定义*model1*、*model2*、*model3*、*model4*分别代表以上所述的几种类型，并对这四种类型进行统计分析。

对于单个用户所发布的多条微博，可以计算其中的转发量和原创量（抓取的数据带有转发标签），同时可以得到每条微博的被转发次数来计算单个用户的平均被转发次数。通过这些数据便可以将用户进行模式分类，对于原创量占总发布量绝大部分的用户即属于原创类型，否则属于转发型。图2-7仅显示了新浪微博的用户交互模式。

对于新浪微博用户大部分属于*model1*和*model4*类型，*model1*可以说是自娱自乐型，对于微博网络的信息传播几乎起不到任何作用，因为他们既不会转发信息，也不会影响他人来转发信息，这样的节点在信息传播过程中的可用度可以直接归为0。所以对于这样一类又占大部分的用户，可以被直接排除在信息传播影响力节点的选

择之外。

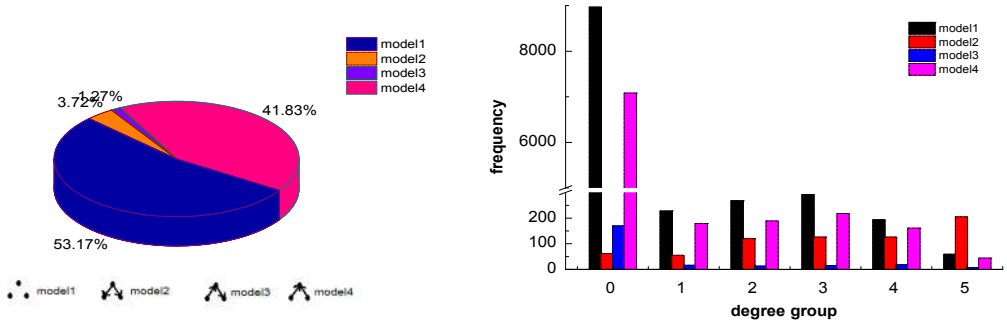
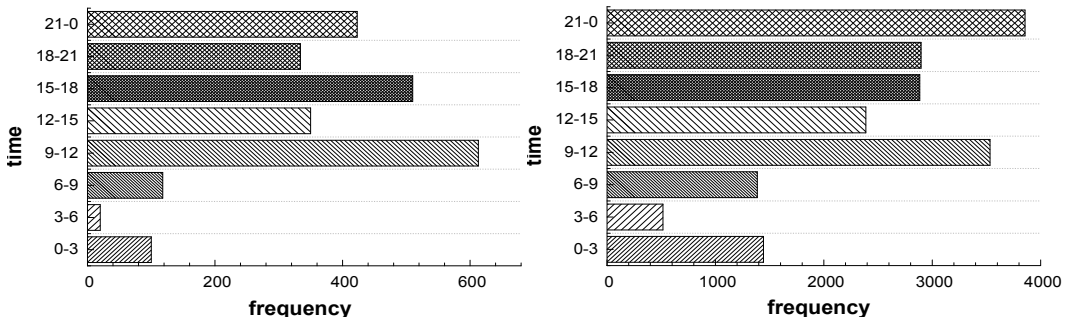


图 2-7 新浪微博用户交互模式

而对于 *model4* 类型，类似粉丝节点，由于盲目追星，大量转发明星节点的微博，而由于本身的社会地位又低并没有多少跟随者，所以又不会被转发，这类节点活跃在网络的边缘，可以称为最后的信息传播者，而他们的数量又是巨大不可忽视的。这类节点对网络信息传播还是起到重要作用的。*model2* 类型的用户稍多于 *model3*，对于这两类均属于微博中地位相对较高的角色，一类是善于原创拥有大量粉丝的明星，一类是爱好转发并有大量粉丝的明星。无论是那种类型如果进行信息传播均将会带来一定的影响力。

微博用户发布微博时间模式描述了用户的日常行为时间规律，即每天发布微博的平均时间段。通过爬取到的用户微博数据，单个用户的每条微博发布时间可以被提取（第 2.1 节有相关数据的介绍），然后计算平均时间段，如此得到所有用户发布微博的时间分布。经过统计，腾讯和新浪微博用户的微博发布时间模式分别如图 2-8 所示。



(a) 腾讯微博　　(b) 新浪微博

图 2-8 用户微博发布时间模式

一天中几乎每个时间段都有用户在发布或者转发微博，当然发布时间大体看来服从白昼模式，夜间的用户量还是很少的。腾讯微博发布的高峰时段是在上午的 9~12 点，下午的 3~6 点以及晚上 9~12 点人数也较多；新浪微博与腾讯微博的用户活跃时间大致相同，只是其发布微博的高峰时段是在晚上的 9~12 点。

Benevenuto 等人[52]发现在线社交网络用户的在线数量也呈现白昼模式。假设发布一条消息的时间是在晚上六点之后，则将可能会失去一大批信息传播用户，因为每天微博信息的更新量是非常巨大的，即使到了第二天的黄金时段，该条消息也许早已被湮没。所以用户的发布微博时间模式也会影响信息传播，选择合适的信息发布时间将会给信息带来更好的传播效果。

由于抓取的数据量大小的问题以及由于抓取的数据量大小以及用户发布微博的时效性问题，可能会使用户交互模式的各类型比例和用户发布微博各时间段的人数发生偏差。

2.3 微博节点偏斜关系和节点可用度

2.3.1 节点偏斜关系

在 2.2 节对微博数据进行了全面的特征分析，微博用户所建立的关系网络具有无标度性、异配性，以及具有较小的聚类系数（特别是新浪微博）。

微博网络的无标度性说明网络中度数较小的节点的个数是占大部分的，而度数较大的节点的个数要相对较少。而异配性说明一条边上连接的两个节点之间的度数具有差距，对节点的平均连接度数进行统计发现度数小的节点趋向于连接度数大的节点。这两个特性都说明微博网络存在很多节点间的不平等关系，用偏斜关系 SR（Skewed Relationship）来表示。对此提出一种符合微博特征观点，微博是一个不平等的偏斜网络，存在两种基于不同节点类型而出现的关系：

1．对等关系，它描述了一条连边所连接的两个节点在跟随者数量上是对等的，即两个节点有相似的入度。

2．偏斜关系，即节点入度出现较大差别。

用节点关系偏斜度α来衡量节点的偏斜关系，如果一条边所连接的两个节点 u、v 的度数分为 d_u 和 d_v，考虑 u 对 v 的影响时，以 u 为中心看 u、v 之间的偏斜关系。如果 d_u 远大于 d_v，则为正向的偏斜关系；若 d_u 远小于 d_v，为负向的偏斜关系；否则

为对等关系。α定义如式 2-3。

$$\alpha = \frac{d_u - d_v}{\max\{d_u, d_v\}} \quad \text{(式 2-3)}$$

α的绝对值越大说明两个节点之间的度数相差越大，偏斜程度越大，对等程度越小。如果一条关系所系的两个节点之间存在明显的偏斜，并且是单向的关注关系，可以认为该关系属于明星与粉丝之间的关系，由于明星的个人影响力作用，在这样的关系下进行信息传播的的可能性也就变得相对较大，即度数大的节点对度数小的节点的影响是较大的。因此α将作为节点间影响大小度量的参数之一，在 3.3 节具体介绍。

2.3.2 节点可用度

微博中的节点可以被分为四种类型 *model1~4*（2.2 小节对微博用户交互模式分析得出的结果），其中 *model1* 种类的节点是属于原创且不被转发的类型，可以说这类节点对微博网络的信息传播几乎没有作用，因为他们既不会接收转发任何信息，也不会被转发从而影响其它人。所以说这类用户是封闭型的，在微博网络中纯属自娱自乐。找出该类节点将其剔除，简化网络结构，对于提高算法效率有很大帮助。

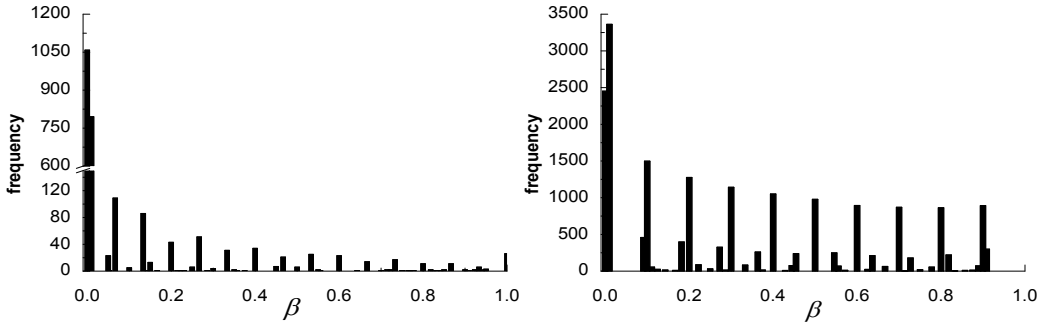
同时，通过统计新浪微博和腾讯微博中已经有 1 年以上时间没有任何发布或者转发行为的用户，发现新浪微博上有 12.97%的用户处于这种状态，腾讯微博的更多。这些用户对于信息传播也不再起到任何作用，可以称之为网络中的“死节点”。“死节点”不具有转发行为，信息传播路径不可达。

节点在信息传播过程中如果起到推动作用，那么可以说节点是有用的或者可用的，否则节点是不可用的。对于上述的两种情况，其节点的可用性是非常小的，因为其对信息传播几乎起不到什么作用。据此定义节点可用度β来描述节点对信息传播的可用程度（式 2-4），当节点为“死节点”时（即最近一次转发发生的时间如果距离现在超过一年则节点的状态标识ε为 0，否则为 1），节点可用度为 0；否则节点可用度为节点的转发率（即转发微博占总发布微博的百分比）。

$$\beta = \varepsilon \cdot \frac{n_{repost}}{n_{sum}} \quad \text{(式 2-4)}$$

计算微博所有节点的β，得到如图 2-9 的统计结果（其中纵坐标表示频率），两

个数据集都存在不少的$\beta = 0$的节点。由于数据量大小的问题和部分节点关系缺失，对于统计的节点可用度频率大小将可能存在偏差。



(a) 腾讯微博　　(b) 新浪微博

图 2-9 腾讯和新浪微博用户活跃状态

信息传播过程中，节点会受到影响被激活，作为下一个信息传播的节点，节点是否被激活的因素不仅是来自邻节点的影响，也取决于自身的状态，节点可用度为 0 或较小的情况下邻节点很难将其激活，所以节点可用度也将作为微博信息传播模型中节点影响大小度量的参数。在第三章将提出基于节点偏斜关系和节点可用度 SA 的信息传播模型 IDSA。

2.4 本章小结

本章主要介绍了如何从微博站点获取所需数据，以及数据的存储形式；同时通过获取的数据进行分析。度量了微博网络的度分布、网络的同配性、聚类性以及社区性等，也分析了微博用户的微博交互模式和发布时间模式。

发现其同一般的社交网络一样呈幂率分布，具有社区性和聚类特性，但网络呈现异配性，这可能由于微博网络本身的特点决定，微博网络是一个由明星带动，粉丝大量聚集的社交平台，很多度数很小的节点连接着度数大的节点。同时也发现微博用户大部分交互模式处于封闭型，即并不转发也不被转发，这类节点对于信息传播并没有太大的用处；微博用户发布时间模式呈白昼模式，高峰期在上午的 9~12 点，所以为了更好的进行信息传播，最好将发布时间定于该段时间。

下一章将利用本章的部分分析结果来实施微博网络的影响力最大化算法。

3 基于 IDSA 的微博网络影响力最大化算法

上一章介绍了微博数据挖掘方法和分析结果，本章将利用其相关结论对微博网络信息传播问题进行研究。研究内容包含两个方面，一是信息传播模型的研究，二是影响力最大化算法的设计。

3.1 基础概念

网络由节点和边构成，生活中有各种各样的网络，如神经网，神经元是节点，神经是边；Internet 网，网页是节点，网页之间的链接是边等。本文研究在线社交网络，在线社交平台上的注册用户是节点，他们之间通过加好友或者相互关注建立的关系是边。

随机网络中的一个节点与其它网络中的节点具有连接关系的概率相同，节点所连接的其它节点数目服从泊松分布。这样的网络最终会使节点的连接数目大致相同，不会有太大的差别。而现实世界的网络总是有“优先连接”的特点，所以随机网络无法合理的刻画真实世界网络。

无标度网络的节点连接分布较随机网络要显得参差不齐，往往呈幂率分布（第二章已给出），这种网络也称为无尺度网络。新节点 u 与网络中其它节点 w 的连接概率 P 与节点 w 的度数 d_w 成正比，即 $P(d_w)=\lambda d_w$（其中 $\lambda=1/\sum d$），它具有的这种“择优连接”使网络出现“富者更富”的现象，网络中小部分节点的度数较高，大部分节点度数低，这与微博网络的特征是一致的。

现实中的很多网络往往符合无标度网络的特征，即无标度性。所以第二章对于微博网络的属性测量可以发现其也属于无标度网络（特别是新浪微博的无标度特性造成了其网络具有更加明显节点偏斜关系）。

信息传播是发送者发送信息到接收者接收信息的过程，现实生活中的一些网络承载了信息传播的功能，如社会关系网络，个人通过口口相传向好友宣布某类信息，好友接收到之后又向好友推送，这样信息便得到传播。现在的信息传播平台已经从传统媒体逐渐转移到在线移动自媒体打开电脑或者手机，通过微信或者微博等平台即可观看好友发布的信息并且转发信息（如果在微博网络中节点转发某条信息被定义为节点被激活）。在线社交平台的信息传播具有快速、实时以及跨地域等特点。通过信息传播可以进行商品推广、广告宣传以及市场营销等。

信息传播模型指出网络中已经被激活的节点是如何影响邻节点的，以及这些未被激活的节点被激活的条件是怎样的（这里激活指节点收到某信息并参与了信息传播）。经典的信息传播模型独立级联模型 IC 和线性阀值模型 LT 描述了两种常见的信息传播机制，IC 模型提出未被激活的节点受其周围激活节点的影响是独立的，如图 3-1，节点 1 下一时刻是否被激活分别受节点 2 和节点 8 的影响；LT 模型则指出这种影响是叠加的，即节点 1 受节点 2 和节点 8 的作用之和。他们定义了这种节点间影响大小 p（也称为节点到节点的信息传播概率）以及节点的激活阀值 θ。

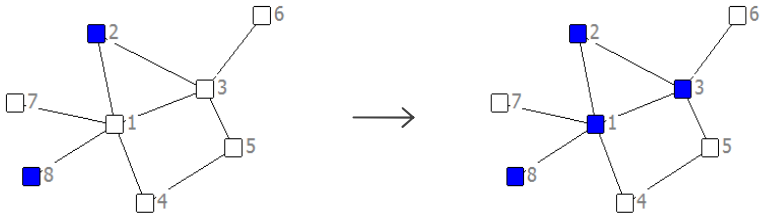


图 3-1 信息传播示意图

除了以上两类经典信息传播模型，还有一些其它的模型来描述信息传播，如传染病模型，该模型一般描述生活中传染病的扩散情况，它把人群分为三种状态：易感状态 S、感染状态 I、免疫状态 R。状态 S 代表还未染病但处于易感染的环境；免疫状态指已经染病被隔离或者恢复健康具有免疫能力不被再次传染，也不会去传染其它人。常见的传染病模型有 SIS 和 SIR 模型，还有基于这两者的多传染病模型。图 3-2 中 a、b 分别描述了这两种模型，在 SIS 模型中假设此时已经有部分人处于感染状态，他们就会去影响周围的人，周围人群可能从易感状态变为感染状态，一段时间过去由于自身或者医疗等原因恢复正常又可能变回易感状态。而在 SIR 模型中，一旦变为感染状态要么不被治愈一直处于感染状态，要么被隔离治疗恢复健康并具有免疫能力。

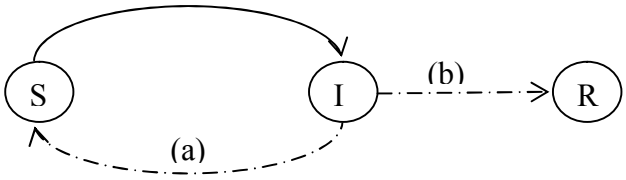
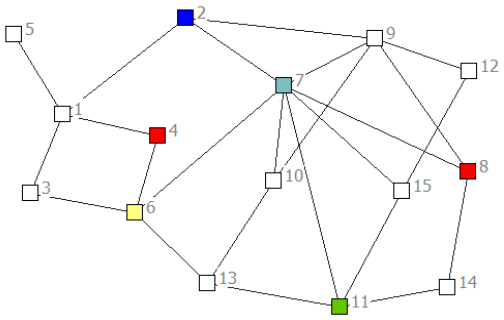


图 3-2 传染病模型状态转移图

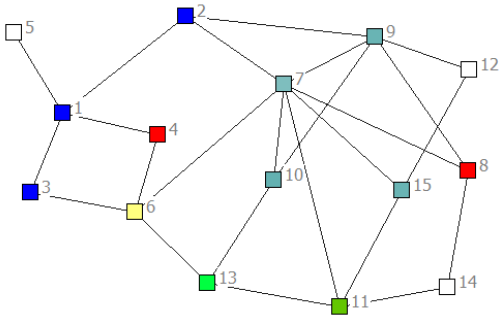
传染病模型也可被用来描述信息传播，未被激活的节点处于易感状态，当其接

收到某信息，有可能被激活变为感染状态，随着时间推移，如果对该条信息失去兴趣或者其它原因有可能删除该信息，此时节点又变为未被激活状态，这种状态可能一直保持即为 R 状态，或者成为 S 再次接收该信息。传染病模型描述的节点是多变的，状态一直随着时间发生变化。

多信息竞争传播模型同时在网络中传播多种信息，网络中的节点可能被不同的信息所激活，所以节点的激活状态是复杂的。多条信息在传播过程中存在竞争关系，节点选择被那种信息激活存在不确定性。图 3-3 中，t1 时刻网络中存在被不同信息激活的若干节点，到 t2 时刻可以看到绿色和蓝色的信息扩散程度相对较大。对于节点 1，其选择被蓝色信息激活而舍弃了红色信息；对于节点 13，它同时选择了三种邻节点发送的信息。由于多信息传播模型的传播机制比较复杂，本文中就只讨论简单的单信息传播情况。



（a）T=t1



（b）T=t2

图 3-3 多信息竞争传播

单节点影响力在本文中指单个节点依照上述某个信息传播模型的传播机制去影

响邻节点，层层递归，最后所影响的节点个数或者说激活节点个数。不同于信息传播模型中的节点间的影响大小 p，其是计算节点影响力的基础。

3.2 问题描述

影响力最大化问题 IM（Influence Maximization）可以形式化的表述如下： 在社交网络 $G(V,E)$ 中，选择 k 个初始节点集合 A（$A \subset V$），按照某种传播策略，由这 k 个信息传播的源点去影响其它节点，通过节点间的相互影响，使最终能够被这些节点影响的节点数达到最大的期望。节点 u 的 $\delta(u)$ 表示让 u 去影响其邻节点，被激活的邻节点再去影响他们的邻节点，迭代数次直到不再有节点被激活所能影响的节点总数，所以 $\delta(A)$ 表示集合 A 的影响力（即 A 中节点所能激活的节点总数），则 A 需满足：

$$\max\{\delta(A), |A| = k, A \subset V\} \quad \text{(式 3-1)}$$

不同于 3.1 节提到的单节点影响力，影响力最大化研究多节点集合的影响力，选择具有最大影响力的这样一个集合，使影响的节点个数最多。目前 Greedy 算法是解决这个问题的经典算法，假设当前的目标节点集合为 A，下一个选取目标节点的办法是将每一个集合 A 之外的节点 v 加入 A($\{v\} \cup A$)，计算加入 v 之后的增量影响力 $\Delta\delta(v|A) = \delta(A \cup \{v\}) - \delta(A)$（即增加的激活节点个数），选择具有最大增量影响力的那个节点加入集合 A，如此循环，直到集合大小满足 k 为止。

贪心算法的最大缺陷就是时间消耗太大，尽管有改进的相关算法，仍旧不能解决在节点数量巨大的复杂网络下快速寻找目标节点集合的问题，特别是对于真实的在线社交网络，用户群体以及相互关系又是复杂多变的。3.4 节将会提出贪心算法的改进算法，之前先对微博网络信息传播模型进行研究。

3.3 微博网络信息传播模型 IDSA

在 3.1 节信息传播模型中提到节点间影响大小 p，其是衡量一个节点对邻节点的影响概率，是能否激活邻节点的依据。而经典的信息传播模型（如 IC 和 LT 模型）在相邻节点 u、v 的 $p_{u,v}$ 的取值上往往是度数的倒数 $1/d(v)$。当然也有其它改进计算方式如 *Wang* 等人[5]提出节点间的影响大小不仅与邻节点有关系，还和邻节点之间的互相连接程度有关系，所以用 $p_{u,v} = d_{NG(v)}(u) \Big/ \sum_{w \in NG(v)} d_{NG(v)}(w)$ 定义节点 u 对 v 的影响大小，其中 $NG(v)$ 为 v 的邻居节点构成的子图。

然而对于微博网络，这样对于 $p_{u,v}$ 的定义就显得不太合理。由第二章对微博网络以及用户的分析，发现微博网络是一个由明星带动，粉丝聚集的特殊型网络；很多网络节点的状态为封闭的（即节点可用度较小），对网络信息传播的作用并不太大。所以微博网络是个特殊的关系不平等的偏斜社交网络。这种关系的不平等性在影响信息传播方面是不可忽视的，所以不能仅仅依靠网络连接状态来衡量节点间的影响大小。

图 3-4 显示了具有代表性的局部微博网络图，图中用符号 1、2、3、4、5 分别表示 han_qiaosheng、duanxuan、fengyumaijia、vibole 和 xuyirenweb，可以看出 1、2、3 是拥有大量粉丝的明星节点，其它节点是环绕的粉丝节点。

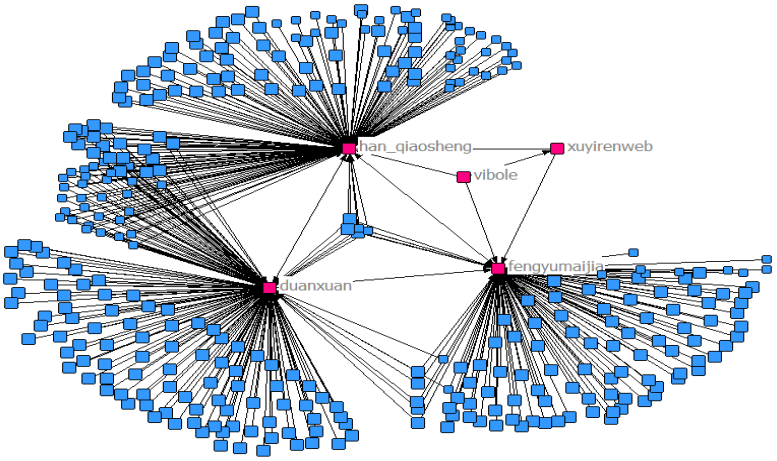


图 3-4 微博网络局部图

如果依据上述的 $p_{u,v}$ 的计算方式，那么在节点 4 的子图结构中，明星节点 1 和节点 5 对节点 4 的影响大小是一样的；在明星节点 1 的子图中，节点 2 和节点 3 对节点 1 的影响大小是一样的。后一种是正常的，而前一种在微博网络中就是不正常的，微博网络中明星对粉丝的影响是巨大的，往往其一条信息由粉丝转发可能达到上万尽管他们之间的关系并不是多么紧密，这处于一种追星的狂热心理和盲目表现。所以 $p_{u,v}$ 的这种度量方式已经不再适用于微博网络，因为其没有考虑到可能存在的偏斜关系。

微博网络存在偏斜关系，而且节点具有不同的节点可用度（在 2.3 节介绍），所以衡量一个节点对与其相邻的节点的影响大小 $p_{u,v}$ 变得相对复杂。对于微博网络信息传播，这种影响是指影响其它节点转发某条信息的能力大小。对于节点 u 和 v，若他们是偏斜关系，节点 u 的个人影响力 $\gamma_{u\_\inf}$ 作为 $p_{u,v}$ 的主要衡量依据；若是对等关系，

节点 u 和 v 之间的关系紧密度 $\gamma_{u,v}$ 则作为主要依据。前提是对于 v 的节点可用度 β 不为 0 时才有意义，若节点可用度为 0，则相应的 $p_{u,v}$ 也应该为 0，所以基于节点偏斜关系和节点可用度 SA，定义节点 u 对 v 的影响函数为：

$$f_{(u,v)} = \beta(\alpha\gamma_{u\_\inf} + (1-\alpha)\gamma_{u,v}) \quad \text{(式 3-2)}$$

节点间的影响函数值说明了用户转发某条信息的行为发生的可能性，其值越高则转发的可能性越大，所以节点 u 对 v 的影响概率大小 $p_{u,v}$ 可以定义为（为了标准化取 *0-1* 之间的值）：

$$p_{(u,v)} = \frac{1}{1+\exp\{-f_{(u,v)}\}} \quad \text{(式 3-3)}$$

为计算概率 $p_{u,v}$，需要量化节点 u 的个人影响力 $\gamma_{u\_\inf}$ 和关系紧密度 $\gamma_{u,v}$，本文节点影响力通过节点发布信息被转发的均值来衡量（n 表示数量，下面通过归一化处理，使 $\gamma_{u\_\inf}$ 值在 *0-1* 之间）：

$$\gamma_{u\_\inf} = \frac{n_{u\_avg} - n_{\min\_avg}}{n_{\max\_avg} - n_{\min\_avg}} \quad \text{(式 3-4)}$$

节点之间的关系紧密度 $\gamma_{u,v}$ 采用 *Wang* 等人[5]的计算方式，NG_v 代表节点 v 的邻居集合形成的局部图，如果节点相互关注则 τ 值为 1，否则为 0，则定义如下：

$$\gamma_{u,v} = \frac{d_{NG(v)}(u) \Big/ \sum_{w \in NG(v)} d_{NG(v)}(w) + \tau}{2} \quad \text{(式 3-5)}$$

独立级联模型中，邻节点的影响大小是独立互不干涉的，每个邻节点以概率 p 单独的产生影响，而在信息传播过程中，如果一个节点的邻居节点不断被激活，则对该节点是否接收信息的心态将会产生影响[53]，Jin 等人[54]提出了一种完全级联传播模型，在该模型中激活节点 u 对 v 的影响概率随着节点 v 的邻居节点被激活的个数的增加而发生变化，这种变化具有不确定性。因为随着邻居节点被激活的个数的增加，节点 v 可以选择随波逐流，也可以选择坚持自我，亦或是选择消极对待（失去新鲜感），如图 3-5，这取决于每个人对待事务不一样的心态。图中节点 1 的激活状态随着其周围激活节点个数增加而可能发生三种变化：

1） 更易激活，这说明随着邻节点被激活个数的增加，对节点 1 产生正向作用。

2） 没有改变，尽管不断有邻节点被激活，也不影响节点 1 的状态，回归到独立

级联模型。

3） 更难激活，随着时间推移，有较多的邻节点被激活，对节点 1 产生负向作用。

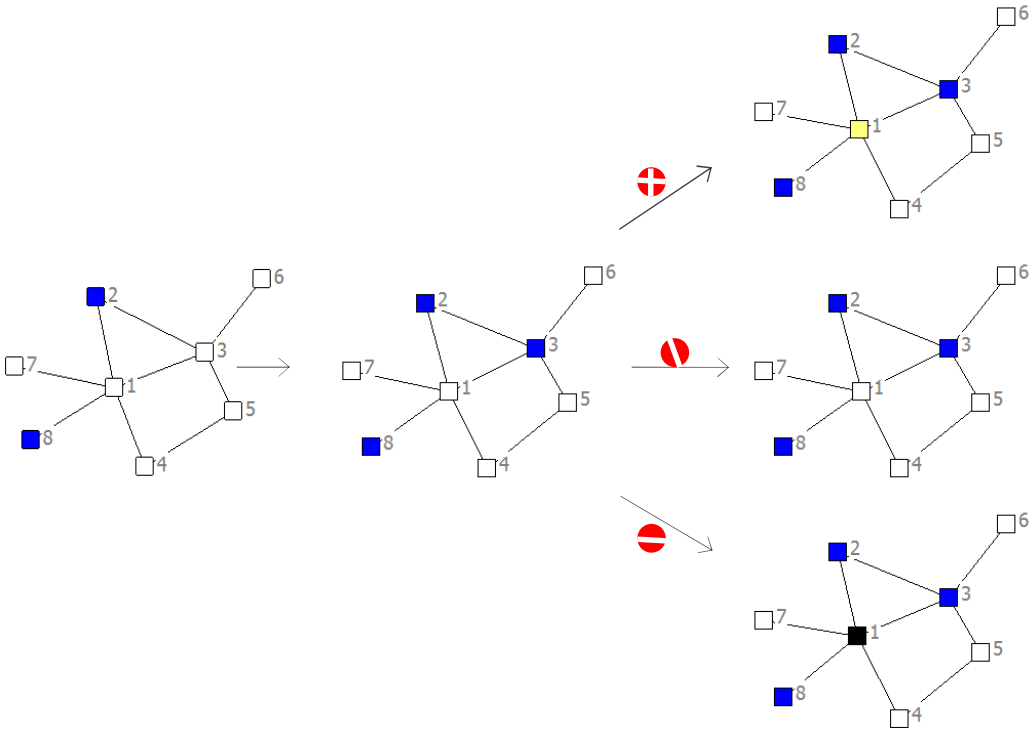


图 3-5 节点的激活状态

基于此，我们改进对节点 v 的激活阀值为式 3-6。

$$\vartheta_v = \theta_v - c * (|A_v|/|N_v|) * \theta_v \quad \text{(式 3-6)}$$

其中 N_v 是节点 v 的邻居节点集合，A_v 表示尝试激活 v 但未成功的节点集，c 的值则是从 $\{-1,0,1\}$ 中选取的随机值（用蒙特卡洛模拟计算）。至此，在独立级联模型基础上，改进节点间影响大小和节点激活阀值的计算方式，提出基于节点偏斜关系和节点可用度 SA 的微博网络信息传播模型 IDSA，该模型将被应用到影响力最大化问题。

3.4 基于 IDSA 的影响力最大化算法

3.4.1 微博数据预处理

解决信息传播中 $p_{u,v}$ 的问题，构建基于 SA 的信息传播模型 IDSA 便可以来实施影响力最大化算法。在实施算法之前，先做一部分预处理工作。需要预处理的数据

都在 3.3 小节的 $p_{u,v}$ 问题中提到，如计算节点 u 的个人影响力 $\gamma_{u\_\inf}$ 需要对用户 u 所发布的微博被转发的次数进行统计，得到均值再进行标准化处理；计算节点的关系情况 $\gamma_{u,v}$ 需要得到节点 v 的好友关系信息等。在第二章的数据挖掘中描述了所爬取到的微博数据有用户信息 userinfo.txt、用户关系 friendinfo.txt 以及微博 userweibo.txt。预处理过程是对这些文件的读取和计算过程。在此过程中将计算 $p_{u,v}$ 并将所有不可用的“死节点”统计出来放入集合 B，算法 3-1 是数据处理的伪代码。

算法 3-1. *Preprocessing with weibo data*

Input: *friendinfo.txt*、*userweibo.txt*

1: $B=\phi$ //B 表示不可用的节点集合
2: $P=\phi$ //P 是每对节点的 $p_{u,v}$ 的结果集
3: **read** *friendinfo.txt* **do**
4: **construct** $G(V,E)$ //构建微博网络结构
5: **end**
6: **for** $u \in V$ **do**
7: **read** *userweibo.txt* **and**
8: **compute** $\gamma_{u\_\inf}$、$\gamma_{u,v}$、ε、α、β //分别计算中间参数
9: **if** $\varepsilon = 0$ **then**
10: $B = B \cup \{u\}$ //节点 v 的可用度为 0
11: **end if**
12: **end for**
13: **for** $u \in V$ **do**
14: **for** $v \in N_u$ && $v \notin B$ **do**
15: $p_{u,v} = 1/(1+\exp\{\beta(\alpha\gamma_{u\_\inf} + (1-\alpha)\gamma_{u,v})\})$
16: $P = P \cup \{p_{u,v}\}$
17: **end for**
18: **end for**

output: $G(V,E,P)$, B

预处理输出了带有 P（P 是每对节点的 $p_{u,v}$ 的结果集）微博网络的结构图和网络中所有节点可用度为 0 的节点集合 B，为下一步实施影响力最大化算法奠定基础。在信息传播过程中，相邻节点 u 和 v，v 被激活的条件是 $p_{u,v}$ 足够大，超出 v 的阀值 ϑ，而 v 的节点可用度表示了 v 当前在微博网络的状态，为 0 的表示不再参与信息传播。

为近似对信息传播模型进行评估，在微博网络上模拟信息传播，度量节点的影响力（节点以概率 p 去影响周围节点，依次递归循环所能影响的节点总数），并与节

点发布微博实际被转发次数比较。两者越接近说明模型的效果越好，实验结果将在第四章给出。

3.4.2 影响力最大化算法

如同 3.2 节描述，影响力最大化算法即是在给定的信息传播模型下选择具有最大影响力的节点集合 A 的办法。信息传播模型描述网络中节点间的影响大小 p 以及节点被激活的阀值 ϑ（$p >= \vartheta$ 时节点被激活）。给定一个节点，通过信息传播模型便可以求出其所能激活的节点个数。那么在网络中寻找一个大小为 k 的集合，集合中所有节点遵循信息传播模型能够激活的其它节点个数最多（即影响力最大），这是影响力最大化算法所要解决的问题。

微博网络信息传播模型在上述两个小节已经讨论，现在本节将设计算法尽可能快速并且准确的找出目标节点集合 A。传统的影响力最大化算法 Greedy 虽然是求解该问题的最优化算法，但其时间复杂度太高，往往难以满足大数据量的社交网络。其算法最费时的一步是每次选择一个目标节点都需要计算所有节点的影响力增量 $\Delta\delta(u \mid A)$（即目标节点中新增加一个节点时，增加的激活节点个数），取其中最大的节点加入目标节点集合。算法 3-2 为其求解过程，其中对于多节点集合影响力 δ 的具体计算就不再详细描述（本章的前两个小节做过介绍）。

算法 3-2. *Greedy*

Input: G,k //G 是网络图，k 为目标节点个数

1: $A = \phi$ //初始化 A

2: while $|A| < k$ do

3: select $u = \arg\max_{u \in V-A}(\delta(A \cup \{u\}) - \delta(A))$ //选择影响力增量最大的节点

4: $A = A \cup \{u\}$

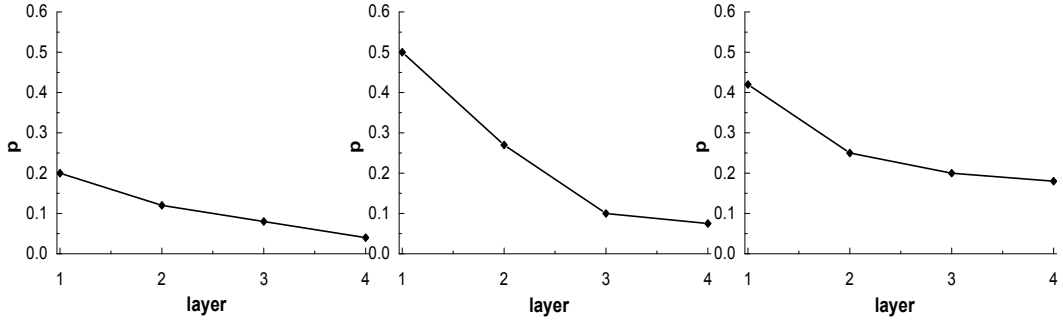
5: end while

output: A

所以针对这一问题，本小节采用以下几种优化方式：

1．图 3-6 显示了三种不同网络的节点 *n-layers* 影响力模式[55]，即节点能够影响的好友节点层数，如 *layer* 为 2 表示节点最多能够影响好友节点，好友的好友节点。由图中可以看出节点的影响力范围大多在 4-*layers* 之内，所以在计算节点的影响力增量时，将计算范围限制在路径长度为 4 之内，即让节点去激活其邻节点，被激活的邻节点再去影响他们的邻节点，这样迭代 4 次便不再继续，因为后面的影响相对于

前面几次迭代非常小可以忽略。



(a) Twitter (b) Digg (c) Citation

图 3-6 *n-layers influence*

2．依据 Zhao[37]等人提出的思路，在选择第 i 个目标节点时（此时的目标几点集合为 A_i），如果非目标节点 u 激活了非目标节点 w，即节点 w 在节点 u 可影响的范围之内或者说节点 u 覆盖了节点 w 可激活的节点，则会有 $\Delta\delta(w \mid A_i) \subseteq \Delta\delta(u \mid A_i)$。这样的话如果 u 被选为该步的目标节点，w 就成为目标节点结合 A_i 的可影响范围内的节点；如果未被选择，下一步也会优先选择 u 而非 w。所以两种情况下 w 的影响力增量 $\Delta\delta(w \mid A_i)$ 均无需再被计算，这样把一个节点可以覆盖的其它节点放在一个集合内，每次选择目标节点时排除该集合内节点的影响力增量计算。

3．对于上一小节输出的节点可用度为 0 的集合 B，可以直接排除这些节点，不必计算，从而简化网络以及计算方式。

基于此，设计改进的影响力最大化算法，算法开始先将 V 中去除集合 B 中的可用度为 0 的节点。然后在每一步选择目标节点时，计算所有非目标节点的 4-*layers* 增量影响力，即进行 4 层迭代计算能够激活的全局节点个数，并标记非目标节点可覆盖的影响节点，将他们放入集合 MI（集合中的节点不再被考虑到下一步循环的目标节点选择范围）。最后求出当前循环拥有最大影响力增量的非目标节点加入目标节点集合 A。至此完成单步循环所要做的工作，循环 k 次，求出集合 A。Influence 代表节点的影响力增量，算法的伪代码描述如 3-3。

算法尽可能减少无用的节点和多余的计算，同时缩小节点影响力增量的计算范围等来减少贪心算法的时间复杂度。每次计算 *MI* 集合都会被更新，添加一些被覆盖影响力的节点，之所以能够这样做是因为计算节点影响力增量的这个过程存在相同的重复计算。

算法 3-3. *Improved Greedy*

Input: $G(V,E,P)$, B, k //k 表示目标节点的个数

1: $A=\phi, MI=\phi$

2: $Influence_{max}$=0,u_{max}=0 //*Influence* 表示影响力变量

3: **for** $u \in V$ & & $u \in B$ **do** //去除网络中节点可用度为 0 的节点

4: $V=V-\{u\}$

5: **end for**

6: **while** $|A_i|<k$ && $i \le k$ **do**

7: **for** $u \in V$ && $u \notin A_i$ **do** //求每个非目标节点的影响力增量

8: $\Delta\delta(u \mid A_i)=\phi$ //初始化为空

9: **if** $u \notin MI$ && $u \notin A_i$ **then**

10: $Influence=\Delta\delta(u \mid A_i)$

11: $MI = MI \cup \arg\delta(u \mid A_i)$ //被覆盖的节点集合

12: **if** $Influence_{\max} < Influence$ **then**

13: $Influence_{\max} = Influence$

14: $u_{\max} = u$

15: **end if**

16: **end if**

17: **end for**

18: $A = A \cup \{u_{\max}\}$

19: **end while**

output: A

假设网络存在 N 个节点和 M 条边，传统贪心算法的时间复杂度为 O(kNMR)，其中 R 为蒙特卡洛模拟随机事件的迭代次数。因为将节点的影响力范围限制在 4 层好友，所以节点所能激活的其它节点个数是个可见常量 d，又由于节点的覆盖现象以及节点的可用度简化了网络结构，每次循环计算的节点个数由 N 会缩小到 n，则算法的时间复杂度为 O(kndR)。

3.5 本章小节

在本章中，提出微博是一个存在不平等关系的偏斜网络，一种是出于交友的好友对等关系，一种是出于追星心理的明星和粉丝之间的偏斜关系。之所以称之为偏斜，是因为对于第一种关系邻节点双方在微博中的地位是平等的，可能都是明星也

可能都是粉丝（即节点的度数之差不大）。对于这样的关系来说，好友之间的关系很大一部分决定了信息是否能从一个节点传播到另一个节点，即影响另一个节点转发某条信息；而对于第二种关系，明星的度数较大，粉丝出于盲目追星，随波逐流状态，所以明星的影响力变成了信息传播的关键因素。

除此之外，本章还对用户节点的可用度进行衡量，节点的状态决定了是否转发某条信息，处于节点可用度为 0 的节点几乎对信息传播无任何作用。结合以上两点本章提出了节点间影响力大小 p 的度量方式来适应微博网络的信息传播模型。之后再次基础上提出了改进的影响力最大化算法，下章将通过实验证明该算法的有效性。

4 实验与结果分析

以新浪微博和腾讯微博数据作为实验数据集，比较传统影响力最大化算法和本文提出的改进贪心算法 IG，分别从信息传播范围的大小（影响力大小）以及时间复杂度两个方面进行对比，结果发现本文 IG 在微博网络影响力最大化问题上有较好的表现。代码的开发平台是 eclipse，JDK1.8，所有的实验都是在 32 位的 Windows 7 平台下运行，内存是 4GB。

4.1 数据集

实验将在国内两个微博网络上进行（新浪微博和腾讯微博），这两个是国内流行的微博网络。微博数据集通过网络爬虫取得，数据集中主要包含：用户微博数据、用户信息、用户之间的关注关系形成的数据。微博数据和用户信息主要用来挖掘用户的行为和状态，用户之间的关注关系形成微博信息传播网络，在该网络上实施影响力最大化算法，找到使信息扩散范围最大的目标节点集合。表 4-1 是部分截取数据（省略了微博数据）。

表 4-1 微博数据集

friendinfo.txt	*userinfo.txt*
Weiboshijiao 5704834946	hanhan 男/上海微博[456]关注[1090]粉丝[42455971]
matianyu caofang	1093897112 男/北京微博[935]关注[386]粉丝[4732316]
hanhan ps3auto	matianyu 男/北京微博[3331]关注[360]粉丝[16056947]
hanhan 2391195543	hu_ge 男/上海微博[3461]关注[719]粉丝[48192392]
1093897112 hu_ge	5054721980 女/海南微博[21]关注[44]粉丝[8]
1093897112 tennischannel	tianyazaixian 男/北京微博[14273]关注[333]粉丝[1384415]
hu_ge 2850809427	weiboshijiao 女/北京微博[1761]关注[199]粉丝[5052073]

腾讯和新浪微博网络的基本参数表如表 4-2，表中统计了网络的平均度、聚类系数和同配系数等，在第二章已经详细介绍过。可以看出尽管新浪微博的数据量较大，然而腾讯微博的聚类系数较高，相互连接关系更复杂，而且节点的 Kshell 值明显偏高；尽管腾讯微博的节点入度较大，但是节点的可用度较少（活跃节点仅占总节点的一半），而新浪微博的用户活跃度相比较高。

表 4-2 腾讯和新浪微博网络参数表

统计项 \ 统计值	腾讯	新浪
节点数(n)	2449	18903
边数(m)	78420	54172
平均入度($Degree_{in}$)	32	3
最大入度($Degree_{max}$)	2253	498
聚类系数(C)	0.177	0.037
同配系数(r)	-0.082	-0.079
平均 K-核	14	3
节点活跃度(%)	43.2	12.97

4.2 对比算法

分别在新浪和腾讯微博网络上实施贪心算法 *Greedy*（*CELF++*）[21]、度中心性算法（*Highest Degree*）[20]、随机算法（*Random*）、*NewGreedy*[25]、*MixedGreedy*[25]、*KDA*[37]和 *IG*（*Improved Greedy*，本文提出的算法）。实验主要在 IC 模型上实施，节点的激活阀值采用所有 p 的均值，由于要在时间复杂度和影响力大小两个方面进行比较，所以选择了常见的启发式算法和以贪心算法为基础的一些改进算法。

1. *Highest Degree*，从算法名字就可以知道，选择目标节点的办法是直接统计网络节点的度数，然后选择 Top-K 节点它是一个基于度的启发式方法。

2. *Greedy*（*CELF++*），改进的贪心算法，利用最大影响力目标函数的子模性，时间复杂度比一般贪心算法有所降低。

3. *Random*，该算法的目标节点为随机游走选择的，采用随机函数直接得到 k 个节点，这种选择方式的时间复杂度最低，然而影响力也是最小的，因为所选的节点不具有标志意义。

4. *NewGreedy*，在贪心算法计算节点影响力增量时，如果节点在当前迭代无法激活其邻节点，就直接删除与邻节点的连接关系，下次迭代不再使用。

5. *MixedGreedy*，算法结合了 *NewGreedy* 和 *CELF* 的思想。

6. *KDA*，基于 k-shell 分解的影响力最大化算法。

4.3 结果分析

4.3.1 信息传播模型

如第三章所述，用真实网络数据集计算用户的影响力 n_1（即用户发布微博被转发的次数）与利用模型计算用户的影响力 n_2（用户所能激活其它用户总数）进行比较来近似的说明所建立模型的可用性。因为用户的被转发量是基于用户的所有跟随者 f_1 的转发行为决定的，而现有的爬取到的用户关系数据有可能是不完整的（仅有部分跟随者 f_2），所以需利用相关比例来进行比较。表 4-3 为选取部分节点的比较结果。

表 4-3 模拟和真实节点影响力对比

节点 ID	f_2/f_1	n_2/n_1
3819125669	9/110	2/15
CaliGurl	17/216	8/69
1504489613	20/445	3/60
3267347020	45/888	1/12
2393930942	13/1028	1/28
2793038397	201/1510	23/126
2808300230	54/2775	5/163
newbang	98/3462	9/205

尽管个别结果有差距，总体来看实际与预测的影响力大小相差并不是很大，因为网络情况复杂，网络节点特征多样，要想建立足够准确可靠的信息传播模型是非常困难的。为得到更加明确的评估结果，对微博网络中 50%节点进行测量，这些节点中大部分节点的预测水平与实际是相近的。

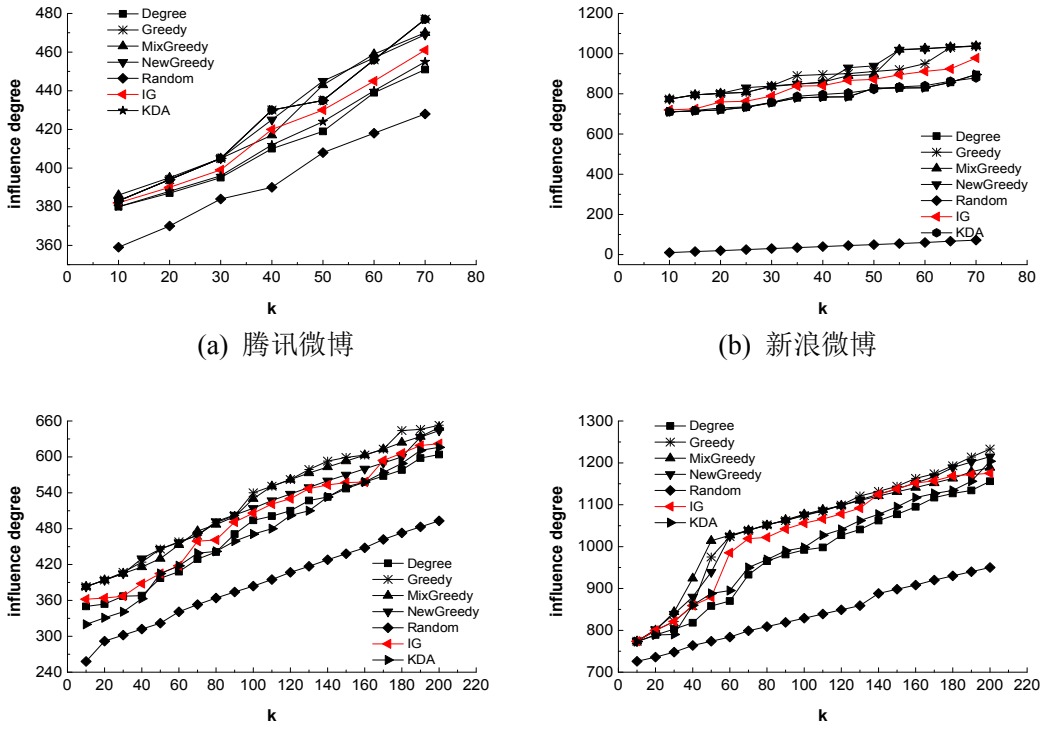
4.3.2 影响力最大化算法

影响力最大化算法的好坏一般从两个方面来比较，一是最终得到的目标节点的影响力大小（反映在能够激活的其它节点个数）；一是算法的时间复杂度，能够满足大规模数据的需求。以下分别在腾讯和新浪微博上比较上述的六种算法，先比较目标节点的影响力大小，实验取目标节点个数 $k=70$ 和 $k=200$ 两种不同大小。结果如图

4-1 所示，其中横坐标为目标节点集合大小 k，纵坐标 Influence degree 代表影响力大小或者激活节点个数。

对于新浪微博，可以看出当目标节点 $k=70$和 $k=200$时，各种算法的影响力大小变化趋势大致相同，$k=200$的前半部分正是 $k=70$的缩小图，所以无论 k 取值多少，Top-K 节点几乎不变。新浪微博目标节点取 70 的图中，可以看出本文提出的算法在影响力大小上相对 Degree 算法是具有优势的，KDA 算法比 Degree 稍好，而 Random 算法是结果最差的，Greedy 及其优化算法 NewGreedy、MixedGreedy 表现几乎一样，只是在 30-65 这段个数范围内稍有波动，各个算法的影响力大小交叉递增，最后趋于一致。当节点个数升至 200 时，对于新浪微博来说这几种算法的表现与 $k=70$时并无太大的差异。

对于腾讯微博，单从影响力大小上来看，无论是 $k=70$或者 $k=200$都是新浪微博网络得到的目标节点具有更大的影响力。腾讯微博在 $k=70$时各个算法的表现和新浪微博一样，而在 $k \geq 100$ 时 MixedGreedy 能够好于其它几种算法，Random 算法效果一直处于最后，这跟其目标节点的选择有关，具有一定的随机性，而没有考虑到节



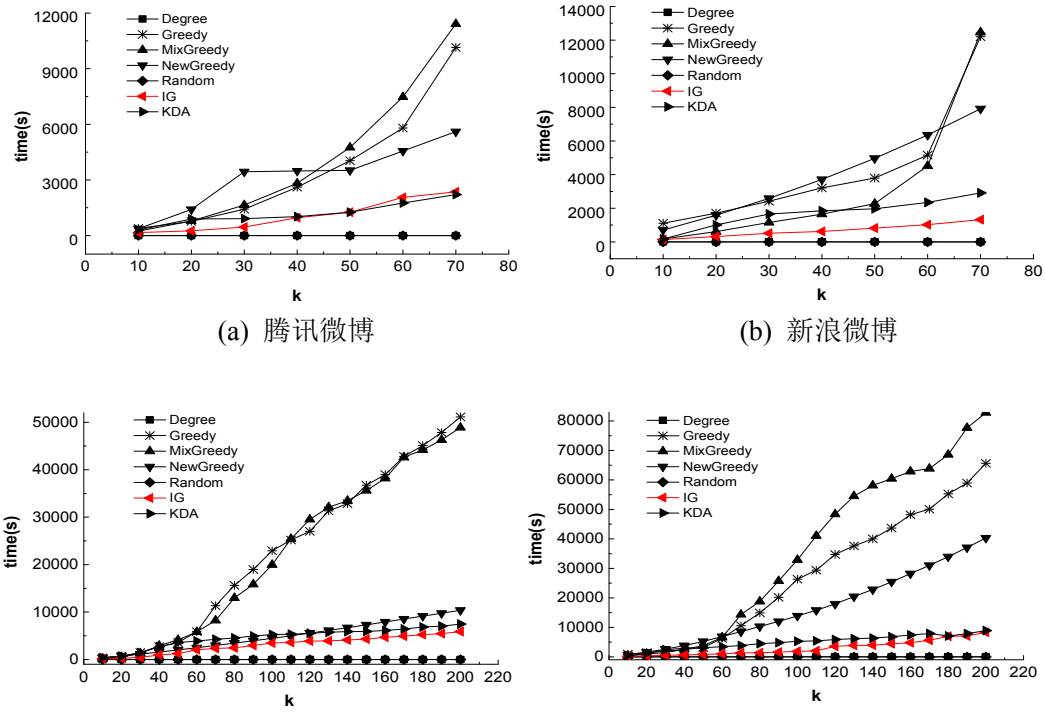
(a) 腾讯微博　　(b) 新浪微博

(c) 腾讯微博　　(d) 新浪微博

图 4-1 影响力大小对比

点的网络属性。而 KDA 算法在腾讯微博$k=200$时出现不稳定现象，在$k \geq 150$ 时才表现比 Degree 好。

总体来看，贪心算法及其优化算法得到的目标节点在影响力大小方面是具有相对优势的，而 Random 算法明显处于劣势，本文的算法要好于 Degree 算法，影响的节点个数要多。这是由于本文不仅仅从节点所处的网络结构出发，还考虑了节点的角色属性和状态。即使度数较大的节点，如果周围节点的可用度均比较低的话，节点的影响力便会被削减，所以从某些角度考虑，Degree 算法的缺点也是不可忽视的。KDA 表现有些不稳定，总体来说要比 Degree 好。



(a) 腾讯微博　　(b) 新浪微博

(c) 腾讯微博　　(d) 新浪微博

图 4-2 两个数据集上的运行时间对比

图 4-2 为各种影响力最大化算法的时间复杂度对比，同样实验分别在新浪和腾讯微博网络上进行，目标节点$k=70$和$k=200$。横坐标不变，纵坐标为时间（单位：s）。在$k=70$的新浪微博数据集上，可以看出 Degree 和 Random 算法耗费的时间是非常小的，接着是本文提出的算法，之后是 KDA 算法，而 Greedy、NewGreedy 和 MixedGreedy 三种算法是非常耗时的，达到 3 个小时及以上更多。在$k \leq 60$时，NewGreedy 耗时大于 Greedy，Greedy 大于 MixedGreedy；而当目标节点增致 70 时，Greedy 和 MixedGreedy

突然上升，超过 NewGreedy。

在图 4-2(d)中，MixedGreedy 耗时就明显超出 Greedy 很多，而 NewGreedy 仍在相对低水平稳定上升，但是这三种算法的时间复杂度高出其它很多，几乎超过 10 个小时。另外四种算法的时间复杂度保持极低水平，完全可以满足大规模数据的需求。从新浪微博数据集的实验结果来看，Greedy 及其另外两种优化算法非常耗时，并且 MixedGreedy 作为优化算法表现的并不稳定。从腾讯微博数据集上也可以看出。腾讯微博上，当$k=70$时，NewGreedy 表现的异常耗时，而$k=200$时又处于正常水平，这可能是实验过程中存在了其它干扰性的因素。

通过目标节点影响力大小和运行时间两方面对各种算法做对比，发现 Greedy 系列算法非常耗时，MixedGreedy 表现不稳定，可以看到虽然 MixedGreedy 和 NewGreedy 同样作为 Greedy 的优化算法，MixedGreedy 并不能很好的体现出优化效果，无论从时间上还是影响力上。NewGreedy 算法作为 Greedy 的优化算法具有较好的稳定性和优化效果。虽然时间上他们比较耗时，但是算法所求出的目标节点具有较好的影响力。同样 KDA 作为优化算法，在时间复杂度上要较低于其它贪心算法。但在影响力大小方面稍显不足。

而 Degree 和 Random 算法作为时间复杂度极低的算法，其目标节点的影响力大小方面稍显逊色，当然 Degree 还是远好于 Random。至于本文提出的算法，虽然在时间复杂度上高于 Degree 和 Random，但是影响力方面要高于他们，而且耗时小于传统贪心算法，可以作为一个折中的优化算法。而本算法的使用范围仅限于类似于新浪微博这种具有明显偏斜关系的网络，虽然信息传播模型具有普遍适用性，但是本文的节点间影响力p的度量是依据微博网络来设计的，观察其计算公式，偏斜度α完全是依据邻节点度数决定，不同的α匹配不同的关系。

表 4-4 是对不同算法所获取到的目标节点进行比较。仅截取了腾讯微博数据集上的前 15 个目标节点，其中表格中的数据是用户的 ID（唯一标识微博用户身份）。从表格的部分数据显示，Greedy、NewGreedy 以及 MixedGreedy 这三种算法在目标节点的选择上有很大一部分是重合的，NewGreedy 与 MixedGreedy 因为是贪心算法的优化算法，其本质与贪心算法的选择是相似的，所以会出现极大的节点相似性，而 Degree 算法也有部分节点与其重合，这是因为度数高的节点在微博网络中往往是明星节点，其被选为目标节点的可能性很高。而本文的算法尽管在节点选择顺序比较上与贪心算法有着差异，但就其相似节点个数来说差别并不是太大。

表 4-4 不同算法的目标节点集比较

Greedy	NewGreedy	MixedGreedy	Degree	Random	IG
gdga110	gdga110	gdga110	luchuan	bbc020	Hejiong
liyuntao	liyuntao	liyuntao	gdga110	cdga-110	gdga110
qdangdang	qdangdang	qdangdang	liyuntao	liai	liyuntao
liulishg2008	liulishg2008	liulishg2008	pony	guagdgfabu	mengfei
cyb6206477	cyb6206477	cyb6206477	cyb6206477	lfga110	mayili007
chenggang	chenggang	chenggang	bojuegolf	jq52163	fengyumaijia
bojuegolf	bojuegolf	bojuegolf	imjerry	pzhqz2010	jjlin
liuyikun	cherryhe	jingtime	chenggang	jiangjiuzhen	bojuegolf
daixueneng	iivf8888	gzga110	liuyikun	kkkj_7285	liuyikun
MGgaoyan	guazhoxiaofg	cherryhe	cherryhe	mikhh_205	fegxiogg088
qhdga110	qhdga110	guazhoxiaofg	dj77600624	lovqg13140	qhdga110
cherryhe	liuyikun	qhdga110	jingtime	jgt759	li_nian
gzga110	dangqing	dangqing	qhdga110	cuizixuan19	aliaszhang
ZHANG66166	ZHANG66166	liuyikun	gzga110	duahonghao	jiajingwen
aliaszhang	aliaszhang	aliaszhang	daixueneng	Trends-Men	chenggang

相对于这些算法，Random 的目标节点选择上就有很大差距了，因为是随机选择并不能保证目标节点的质量问题。为了更好的对这些算法所选择的目标节点的相似性进行比较，以 Greedy 为参照，绘制了目标节点相似度比较示意图 4-3。

在$k=70$的目标节点集合中每$d=10$个节点计算一次相似度，如式 4-1 为 Greedy 和 Degree 结果集相似度衡量。

$$Similarity=|A_{Greedy}^{d}\cap A_{\deg ree}^{d}|/d \quad \text{(式 4-1)}$$

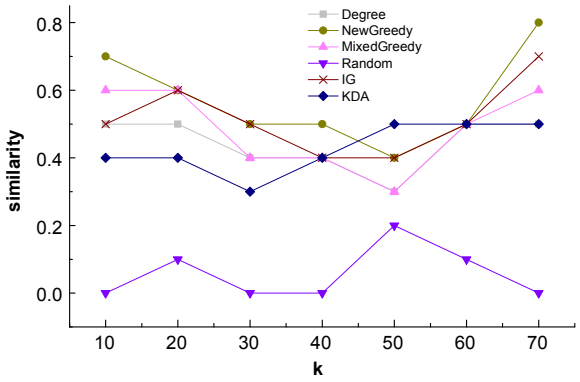


图 4-3 目标节点相似度比较

其中 A 表示算法得到的目标节点集合，图 4-3 展示了 Greedy 算法的结果和其它算法的相似度比较。可以看出图中除了 Random 算法和 Greedy 的结果集相似度比较差距非常大之外，其它几种算法所求出的目标节点均和 Greedy 的目标节点较相似，平均水平均在 0.3 以上。细分的话，其中 Degree 要低于其它几种，KDA 在前期表现稍显逊色后期较好，而 MixedGreedy 相似度曲线波动比较大，IG 和 NewGreedy 相对稳定，并且 NewGreedy 与 Greedy 的相似度一直处于较高水平。

这与目标节点的影响力比较结果大致相同，IG 与 Greedy 的相似度曲线在随目标节点个数增加而呈现上升趋势，说明其得到的目标结果集中的节点质量是较优的（质量好坏指的是所选择的节点与应该选择的最优节点之间的差距大小，差距越小的质量就越好）。

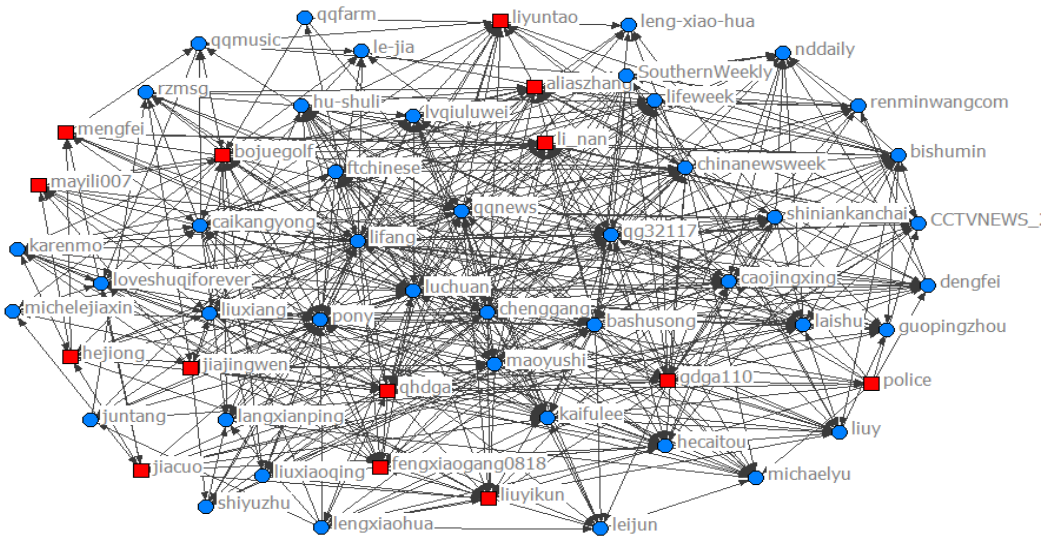


图 4-4 腾讯网络

图 4-4 显示了腾讯网络的节点链接有向图，因为抓取到的腾讯网络节点较多以及相互链接关系复杂，所以提取了部分节点，去除了网络中入度小于 200 的节点及其连边。显示在图中的节点都是入度超过 200 的，不过有些入边被删除了。其中红色方块标记出的 14 个节点是通过实施本章的算法得到的部分目标节点，其它的蓝色圆点是未被选择的。图中的箭头表示关注关系，箭头指向了关注对象，而另一头指向了追随者（如粉丝），入度越大的节点说明其追随者越多。有些连接是双向的，表明节点间相互关注。

尽管已经将节点个数缩小到 57 个，但从图上仍可以看到错综复杂的连接关系。目标节点分散占据在网络的各个区域，并不是聚在一起，而且目标节点并不都是核心的度数较大的明星节点（如 luchuan、pony）。这正是体现了选择节点时的策略，每次选择的是具有最大影响力增量的节点，聚在一起的多个节点之间可能相互影响，造成影响力覆盖。

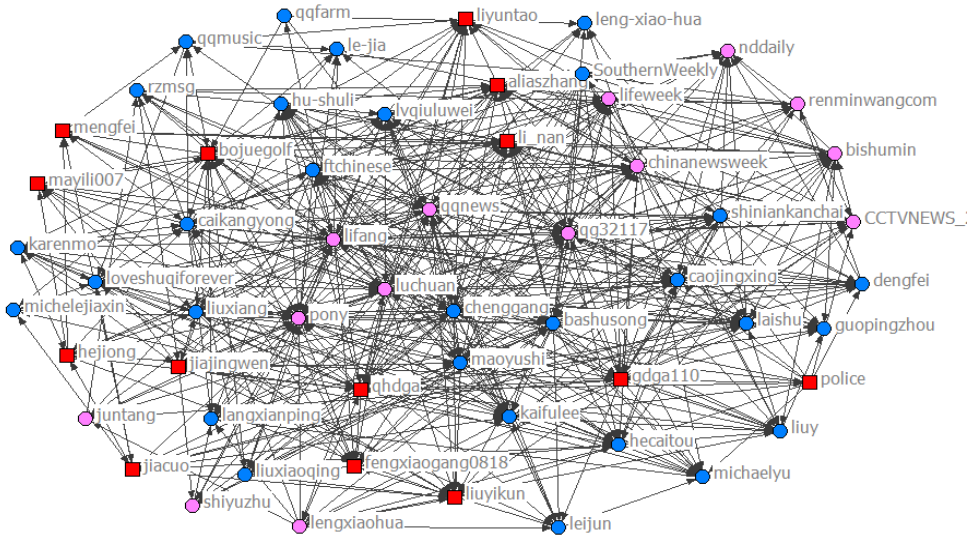


图 4-5 目标节点激活邻节点示意图

图 4-5 中模拟信息传播，粉色节点为目标节点进行信息传播时激活的节点（目标节点尝试以概率 p 去影响邻节点，超过激活阀值的邻节点被激活，这里的是第三章建立的微博信息传播模型所计算的），图上仅显示了部分激活节点。从图中也可以看出，有些目标节点的邻节点被激活，而有些并没有；同时节点的影响力是存在覆盖现象的，被激活节点的影响力被目标节点的影响力覆盖，这些被激活的节点将不再作为目标节点的候选对象，而是从其它未被激活的节点中选择，这也解释了目标节点的非聚集分布。

4.4 本章小结

在上一章中，提出微博是一个存在不平等关系的偏斜网络，这种偏斜关系出于追星心理的粉丝和明星之间，利用这种关系提出改进的适用于微博网络的独立级联模型。并依据节点可用度、节点间影响力覆盖、节点的影响力有限性等改进了传统贪心算法。

本章主要对上一章介绍的影响力最大化算法进行结果评估，首先介绍了两种数据集；因为要做对比实验，然后介绍了几种对比算法：*Highest Degree*、*Random*、*Greedy*、*NewGreedy* 以及 *MixedGreedy*。首先从影响力大小的角度来评估算法的优劣性，发现本文的算法处于中间水平，好于 *Degree* 而比 *Greedy* 系列稍低；之后从时间复杂度来比较各个算法，可以看到贪心算法的时间复杂度是相对来说比较高的，在 k=200 时将近要用 10 个小时以上的时间才能完成计算。而本文的算法时间复杂度要低于 *Greedy* 系列，在可容忍的范围内求出目标节点。所以综上所述，结果表明本文提出算法优劣并存，但不失为一种可以满足大规模数据集的这种算法。

通过对比各种算法所得到的目标节点，以及计算他们和 *Greedy* 之间的相似度可以看出引起算法影响力大小的原因，*Degree* 所得到的目标节点和 *Greedy* 还是有一定的差距的，之后通过显示目标节点在网络图中的位置，也可以看出 *Degree* 算法的缺点。

5 总结与展望

5.1 研究总结

微博的快速发展带动着自媒体信息传播的快速发展，信息的多样化和来源的广泛化改变着生活的方方面面。现在人们了解信息的方式更多的是在线的信息发布平台，而微博作为信息传播的在线社交平台，正发挥着巨大的作用。通过微博，可以进行信息推广、商品营销以及舆论控制等。在进行商品营销时，不可能选择网络中的所有节点来传播商品信息，只能选择部分节点，而如何选择这些节点又要使信息传播的很广泛，传统影响力最大化问题即是研究该类问题。

在微博中进行影响力最大化问题不同于一般的社交网络，微博网络拥有庞大的用户群体，所以形成以个体为特色的信息传播，一个人的行为习惯以及网络地位（往往代表着社会地位）影响着信息传播。研究成果主要有以下几点：

1. 对微博数据进行分析，发现微博关系网络的一些特征。利用抓取到的微博好友数据集建立微博用户关系网络，通过对网络的测量发现网络该网络具有度分布的无标度特性、社区性、聚类性等，同时发现网络是异配的，即网络中存在大量度相异的节点链接，这是由于微博网络的明星粉丝效应，所以发现微博网络是一个关系偏斜的网络。同时利用用户发布微博数据，分析了微博用户的交互模式和时间模式，发现微博中存在大量封闭的用户（原创型又不被转发）。

2. 基于微博数据的分析，构建了适用于微博网络的改进的独立级联信息传播模型。对于微博网络，明星对粉丝的影响力是巨大的，而明星对明星，粉丝对粉丝相对而言并没有那么突出，所以利用这种偏斜和对等关系衡量用户对用户的影响大小；同时用户的节点可用度决定了用户转发某条信息的可能性，结合两者提出了符合微博特征的基于 SA 的信息传播模型 IDSA。

3. 基于 IDSA 提出了改进的贪心算法 IG。算法依据以下三点提出改进方案：

（1）节点可用度为 0 或者较低的用户往往是封闭型的，或者是不活跃的，这类节点对网络中信息传播并没有用处，删除这些节点以及他们的连边可以简化网络结构；

（2）一个节点的影响范围（激活节点范围）是有限的，统计大部分节点发现一般在 4 层邻节点之内；

（3）一个节点可以影响其它节点，被影响的节点的影响力于是被覆盖，所以不用选择该类被覆盖的节点。

同时在新浪和腾讯微博上实施该算法并与传统的算法做比较发现该算法有较好的表现。时间上比传统贪心算法消耗低，但比启发式算法稍高；得到的目标节点质量好于启发式算法，比贪心算法稍低，可以作为一种折中的解决办法。

5.2 下一步工作展望

尽管对微博数据进行了多方面的测量，提出符合微博特征的影响力最大化解决办法，仍然存在以下问题有待研究：

1. 文中提到微博网络具有偏斜关系和节点可用度，没有给出是否两个特性具有普遍性，可对此进行进一步的研究。

2. 为了更准确的计算节点间的影响大小和节点激活阀值，可从用户微博内容进行分析，与用户发布微博相似的更容易受到用户的关注，也更有可能进行转发，所以基于内容用户行为分析对微博信息传播建模将会有一定的作用。

3. 微博网络的用户群体是在不断变化着的，用户之间的连接关系也会发生变化，如何在动态的网络上设计有效影响力最大化算法还有待研究。微博网络的数据量不断，每次计算的消耗时间会随之增大，并不能对每次变化的网络都进行重新计算，必须设计可扩展的可靠算法。

4. 微博中可能存在多种信息源，他们之间可能存在相互作用，如竞争性关系或者促进性关系，这都影响了信息的传播，所以信息传播的影响因素并不是单一的，在信息传播建模中需要考虑的现实情况还有很多。

致谢

三年的研究生生活转瞬即逝，对于将要走向工作岗位的我来说，这三年时光既珍贵又难忘。三年里充满同学间的欢声笑语，同时也充满师生间的谆谆教诲。这三年让我经历了很多同时也成长了很多，毕业之际，想感谢那些给了我无限感动和温暖的老师、同学、朋友以及家人。

首先，要感谢我最尊敬的鲁宏伟导师。跟着鲁老师三年的时间里，让我受益良多。记得刚来到华科的那段日子，由于没能适应研究生生活和学习，让我困顿许久，也变得比较沉默。鲁老师经常开会告诉我们应该如何利用研究生三年做一些有意义的事情，并且开导我要多跟师兄师姐交流，这样才能进步。慢慢的，我开始不再那么忧虑和浮躁，跟着鲁老师一步步做项目以及到后来的论文写作，我的生活开始变得充实和有目标。鲁老师的谆谆教诲和细心教导让我从迷茫中走出来，渐渐领悟到求知的重要以及时间的珍贵。同时鲁老师豁达的生活观教会了我在生活上要积极向上，乐观面对每一天。

其次我要感谢甘早斌副教授。进入华科我所读的第一本书就是甘老师推荐的，这本书对我之后的研究方向起了指导作用。甘老师对待工作中出现的问题，会废寝忘食的去解决，这种执着认真的态度也让我学习很多。他时常教导我们要开拓思维，多动脑筋去解决生活和学习中所遇到的各种问题，这对我以后的工作将会起到很大的帮助。

同时感谢参加我的论文开题与评审的计算机学院的专家与老师们，感谢他们能抽出宝贵的时间来阅读我的论文，并给出有价值的意见。

依然要感谢陪我度过难忘的研究生生活的师姐马霄、赵倩、朱晓恒，12 级师兄魏涛、吴新庭、张垂友和郭文鹏以及 13 级师兄许雷永、戴振民、熊乐、赵将、项海峰、肖雄志和马会新，还要感谢同届的张传超、张建雄、何雷和刘菲，以及 15、16 级的师弟师妹们，还有我的室友们，是你们让我的研究生生活充满色彩。

最后感谢支持和鼓励我的家人。这三年虽然短暂，但是留给我的记忆却将影响我今后的人生，感谢的话语还有很多，就让它藏在心底。我将带着这些勇敢积极的面对今后的工作和生活。

参考文献

[1] Domingos P, Richardson M. Mining the network value of customers. In: Proceedings of the 7th ACM International Conference on Knowledge Discovery and Data Mining,San Francisco, USA, 2001:57-66

[2] Goldenberg J, Libai B, Muller E. Talk of the network: A complex systems look at the underlying process of Word-of-Mouth. Marketing Letters, 2001, 12(3):211-223

[3] Goldenberg J, Libai B, Muller E. Using complex systems analysis to advance marketing theory development. Academy of Marketing ScienceReview, 2001, 9(3):1-18

[4] Granovetter M. Threshold models of collective behavior.American journal of sociology, 1978, 83(6): 1420-1443

[5] Wang Y, Feng X. A potential-based node selection strategy for influence maximization in a social network. In: Proceedings of 5th Internatinal Conference, ADMA, Beijing, China, 2009: 350-361

[6] Borodin A, Filmus Y, Oren J. Threshold models for competitive influence in social networks. In: Proceedings of the 6th International Conference on Internet and Networks Economics, Stanford, USA, 2010:539-550

[7] He X, Song G, Chen W, et al. Influence blocking maximization in social networksunder the competitive linear threshold model.In: Proceedings of the International Conference on Sustainable Design and Manufacturing, California, USA, 2012: 463-474

[8] Chen W, Zhang H. Complete submodularity characterization in the comparative independent cascade model. Computing Research Repository, 2017: abs/1702.05218

[9] Bharathi S, Kempe D, Salek M. Competitive influence maximization in social networks. In: Proceedings of the 3rd International Workshop on Internet and Network Economics, San Diego, USA, 2007: 306-311

[10] Jinha K, Wonyeol L, Hwanjo Y. CT-IC: Continuously activated and Time-restricted Independent Cascade model for viral marketing. In:Proceedings of the 2012 IEEE 12th International Conference on Data Mining, Washington, USA, 2012:57-68

[11] Morris S, Contagion. The review of economic studies. Review of Economic Studies, Oxford University Press, 1999, 66(4): 825-52

[12] Young H P. The Diffusion of Innovations in Social Networks. General Information, 2000, 413(1): 2329-2334

[13] Hethcote, Herbert W. The mathematics of infectious diseases. SIAM Review-Society for Industrial and Applied Mathematics, 2000, 42(2): 599-653

[14] May R M, Lolyd A L. Infection dynamics on scale-free network. Physical Review E, 2001, 64(2): 066112

[15] Satorras R P, Vespignani A. Epidemic spreading in scale-free networks. Physical Review Letters, 2001,86(14):3200-3203

[16] Tang J, Sun J, Wang C, et al. Social influence analysis in large-scale networks. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, USA, 2009: 807-816

[17] Barbieri N, Bonchi F, Manco G. Topic-aware social influence propagation models. Knowledge and information systems, 2013, 37(3): 555-584

[18] Zaman TR, Herbrich R, VanG J, et al. Predicting information spreading in twitter. Computational Social Science, 2010, 104(45): 599-601

[19] Wang Z, Zhao J, Xu K. Emotion-based Independent Cascade model for information propagation in online social media. International Conference on Service Systems and Service Management, 2016:1-6

[20] Kempe D, Kleinberg J, Tardos E. Maximizing the spread of influence through a social network. In: Proceedings of the 9th ACM International Conference on Knowledge Discovery and Data Mining, Washington, USA, 2003:137-146

[21] Leskovec J, Krause A, Guestrin C, et al. Cost-effective outbreak detection in networks. In: Proceedings of the 13th ACM International Conference on Knowledge Discovery and Data Mining, New York, USA, 2007:420-429

[22] Chen W, Wang C, Wang Y. Scalable influence maximization for prevalent viral marketing in large-scale social networks.In: Proceedings of the 16th ACM International Conference on Knowledge Discovery and Data Mining, Washington, USA, 2010:1029-1038

[23] Goyal A, Lu W, Lakshmanan L. V. S. CELF++: Optimizing the greedy algorithm for influence maximization in social networks. In: Proceedings of the 20th International Conference Companion

on World Wide Web, Hyderabad, India, 2011: 47-48

[24] Zhou C, Zhang P, Guo J, et al. UBLF: An upper bound based approach to discover influential nodes in social networks. In: Proceedings of the 13th International Conference on Data Mining, Dallas, TX, 2013: 907-916

[25] Chen W, Wang Y., Yang S. Efficient influence maximization in social networks. In: Proceedings of the 15th ACM International Conference on Knowledge Discovery and Data Mining, New York, USA, 2009: 199-208

[26] Borgs C, Brautbar M, Chayes J, et al. Maximizing social influence in nearly optimal time. In: Proceedings of the Twenty-fifth Annual ACM-SIAM Symposium on Discrete Algorithms, Philadelphia, USA, 2014: 946-957

[27] Tang Y, Xiao X, Shi Y. Influence maximization: Near-optimal time complexity meets practical efficiency. In: Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, New York, USA, 2014:75-86

[28] Galstyan A, Musoyan V, Cohen P. Maimizing influence propagation in networks with company structure. Physical Review E, 2009, 79(5):056-102

[29] Cao T, Wu X, Wang S, et al. OASNET: An optimal allocation approach to influence maximization in modular social networks. In: Proceedings of the 2010 ACM Symposium on Applied Computing, New York, USA, 2010:1088-1094

[30] Wang Y, Cong G, Song G, et al. Commuity-based greedy algorrithm for mining top-k influential nodes in mobile social networks. In Proceedings of the 16th ACM SIGKDD Conference on Konwledge Discovery and Data Mining, New York, USA, 2010:1039-1048

[31] Wasserman S, Faust K. Social Network Analysis. Cambridge University Press, New York, 1994

[32] Kimura M, Saito K. Tractable models for information diffusion in socal networks. In: Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases, Berlin, Germany, 2006:259-271

[33] Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. Computer Networks, 1998, 30(1):107-117

[34] Richardson M, Domingos P. The intelligent surfer: Pmbabilistic combination of link and content information in PageRank.Cambridge, MA: Mrr Press, 2002

[35] Haveliwala T H. Topic-sensitive PageRank. In: Proceedings of the 11th International World Wide Web Conference, 2002

[36] Chao J, Han X, Wang Z. Community influence maximizing based on comprehensive cascade deffuse model. Journal of Jilin Uniersity, 2009, 47(5):1032-1034

[37] Zhao Q, Lu H, Gan Z, et al. A K-shell Decomposition Based Algorithm forInfluence Maximization. International Conference on Engineering the Web in the Big Data Era, 2015: 269-283

[38] 田家堂, 王铁彤, 冯小军. 一种新型的社会网络影响最大化算法. 计算机学报, 2011, 34(10): 1956-1965

[39] Zhang H, Nguyen D T, Zhang H, et al. Least cost influence maximization across multiple social networks. IEEE/ACM Transactions on Networking, Piscataway, USA, 2016, 24(2): 929-939

[40] Du N, Liang Y, Balcan M F, et al.Scalable influence maximization for multiple products in continuous-time diffusion networks. Journal of Machine Learning Research, 2017,18(2):1-45

[41] Lei S, Maniu S, Mo L, et al. Online influence maximization. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, USA, 2015:645-654

[42] Saito K, Kimura M, Ohara K, et al. Learning continuous-time information diffusion model for social behavioral data analysis. Asian Conference on Machine Learning: Advances in Machine Learning, Nanjing, China, 2009:322-337

[43] Srinivasan B V, Anandhavelu N, Dalal A, et al. Topic-based targeted influence maximization. Sixth International Conference on Communication Systems and Networks, 2014:1-6

[44] Chen S, Fan J, Li G, et al.Online topic-aware influence maximization. In: Proceedings of the Vldb Endowment, 2015, 8(6):666-677

[45] Cha M, Haddadi H, Benevenuto F, et al. Measuring userinfluence in twitter: The million followerfallacy. In 4th International AAAI Conference on Weblogs and SocialMedia, 2010:10-17

[46] Weng J, Lim E P, Jiang J, et al. Twitterrank: Finding topic-sensitive influential twitterers. In Proceedings of the third ACM international conference on Web search anddata mining, New York, USA, 2010:261-270

[47] Yang Z, Guo J, Cai K, et al. Understanding retweeting behaviors in social networks. In

Proceedings of the 19th ACM international conference on Information and knowledge management, New York, USA, 2010:1633-1636

[48] Mislove A, Marcon M, Gummadi K P, et al. Measurement and analysis of online social networks. ACM SIGCOMM Conference on Internet Measurement, San Diego, California, USA, 2007:29-42

[49] Newman M E. Assortative mixing in networks. Physical Review Letters, 2002, 89(20):208701

[50] Wilson C, Boe B, Sala A, et al. User interactions in social networks and their implications. EUROSYS Conference, Nuremberg, Germany, 2009:205-218

[51] Borge-Holthoefer J, Moreno Y. Absence of influential spreaders in rumor dynamics. Physical Review E, 2011, 85(2 Pt 2):026116

[52] Benevenuto F, Rodrigues T, Cha M, et al. Characterizing user behavior in online social networks. ACM SIGCOMM Conference on Internet Measurement, Chicago, USA, 2009:49-62

[53] Narayanam R，Narahari Y. A shapley value-based approach todiscover influential nodes in social networks. IEEE Trans on Automation Science and Engineering, 2011, 20(1):130-147

[54] Chao J, Han X, Wang Z. Community influence maximizing based on comprehensive cascade deffuse model. Journal of Jilin Uniersity(Science Edition), 2009, 47(5):1032-1034

[55] Liu L, Tang J, Han J, et al. Mining topic-level influence in heterogeneous networks. ACM Conference on Information and Knowledge Management, Toronto, Canada, 2010:199-208

附录 I 攻读硕士学位期间参与的科研工作

[1] 企业横向项目：武汉东浦信息公司供应链管理信息系统，2015.9-2016.8

[2] 企业横向项目：上海艾络格工业物联网数据管理云系统软件开发，2016.12-2019.11