社交网络影响力最大化传播模型与算法研究

摘　要

21 世纪，人们获取信息的途径不再局限于报纸、广播和电视，随着 Twitter、Facebook、Flickr 等重要社交网络的出现及迅速发展，社交网络逐渐成为了这个时代承载信息的主要媒介。由于社交网络影响力最大化问题的研究在实际应用中有着重要的指导意义，因此，该问题也成为了计算机科学研究的热点。本文围绕社交网络影响力最大化问题的模型与算法展开研究，具体包括:

1. 给出影响力最大化问题的形式化定义，针对信息传播方式(途径)，分析目前最重要的一些信息传播模型以及影响最大化问题在不同模型下的相应定义，并简要总结影响最大化问题的一些解决方法并分析其利弊，为研究影响力最大化的模型与算法打下理论基础。

2. 每个节点的不同父节点对该子节点的影响力是不同的，基于此，指出传统信息传播模型假设的不合理性，提出了一种融合节点相关性与节点重要性的 PRP 模型（PageRank-based Propagation Model，简称 PRP 模型），该模型考虑到社交网络中任何节点的不同父亲节点对该节点有不同的影响强度，实验表明，基于 PRP 模型的方法在解决影响最大化问题的效果比传统的基于线性阈值模型、加权级联模型和独立级联模型的方法更好，影响力范围更广。由于 PRP 模型考虑到了社交网络的实际情况，具有较好的实用价值。

3. 传统贪心算法及其改进算法在大规模的社交网络中解决影响最大化问题的时间复杂度很高，针对该问题，本文基于概率转移矩阵的思想提出了一种扩展的线性阈值模型，并基于该模型提出了一种新的基于概率转移矩阵的影响最大化算法（An New Algorithm Based on Probability Transfer Matrix Method，简称 PTMA）。由于 PTMA 算法节省了每个时间间隔都要统计活跃节点数目的时间，因此，该算法与其他基本贪心算法相比，节省了算法时间，降低了时间复杂度，效率更高，并适用于大规模社交网络。

关键词：社交网络；影响最大化；PageRank 算法；概率转移矩阵

Research on Propagation Model and Algorithm for Influence Maximization in Social Network

Abstract

With the emergence and rapid proliferation of social network applications in 21st century, such as Twitter, Facebook, Flickr and so on, people are acquiring information not just from newspapers, broadcasting and TV any more. Nowadays, social network is becoming a more and more important medium that conveys information. As the research of influence maximization in social network plays a very important role for guiding practical application, it has become a concerned problem in computer science field. We focus our attention on the models and algorithms of influence maximization to display our research as follows.

First of all, we formulate the problem of influence maximization in social network in details. In order to know how the information propagates in social network, some important information spread models and corresponding definitions are analyzed, and some algorithms are summarized so as to build the theoretical foundation for our research on propagation model and algorithm of influence maximization in social network.

Secondly, we point out the irrationality about the assumption of those basic models. As we know, different parent nodes have different influence on their child node. Based on this idea, we propose a new PRP model in which the correlation and authority of nodes are considered. The experimental results show that our proposed model is more effective than traditional Linear Threshold Model, Weighted Cascade Model and Independent Cascade Model in solving the influence maximization problem. As PRP model put the practical problems into consideration, it is of much greater value.

Finally, greedy algorithm and advanced greedy algorithm both have a very high time complexity for very large-scale social network as analyzed in solving the influence maximization problem. Thus, an advanced linear threshold propagation model and a new algorithm named PTMA are proposed to solve the influence maximization problem based on the idea of probability transfer matrix. As PTMA algorithm could save the time that is used to count the number of active nodes every time interval in other algorithms, it is more efficient than other algorithms and has a much lower time complexity. And it’s suitable for the very large-scale social network.

Keywords: Social network; influence maximization; PageRank Algorithm; probability transfer matrix

目　录

第 1 章 绪论

21 世纪，随着社交网络的飞速发展，人们获取信息的途径不再局限于报纸、广播和电视，社交网络正逐渐成为这个时代承载信息的主要媒介。由于人们的生活越来越依赖社交网络，信息如何在社交网络里传播以及如何影响用户已成为社交网络里一个非常活跃的研究热点，在实际生产应用中有着重要意义。本章首先分析社交网络影响最大的研究背景和意义以及国内外研究现状，然后提出本文创新点，并给出本文的组织结构。

1.1 研究背景与意义

社交网络（Social Networks）是指由个体及个体之间的关系所组成的一个复杂网络，这种复杂的网络结构对信息传播和扩散起着至关重要的作用。当一个人采纳一种新思想或接受一种产品时，他会向他的朋友或同事推荐，某些人可能会接受或采纳他的推荐，并进一步向他们自己的朋友或同事推荐，这反映了一个人的行为在很大程度上取决于周边环境的影响。这个过程称为传播或扩散(Propagation or Spreading)。

社交网络中的影响力[1]是指用户受别人影响后，感情、观点或行为发生变化的过程。在心理学和社会学研究领域里存在一个共同的研究热点：外界因素是如何影响组织或者个体的行为或者感情的，这个研究热点有着广泛的应用领域。随着 Twitter、Facebook、Flickr 等社交网络的出现及迅速发展，社交网络的影响力成为控制社交网络演化及其动态性的主要驱动力。

下面通过一个实例说明本文研究的社交网络影响力最大化问题。某创业型软件公司开发了一款文档处理软件，它希望通过在线的社交网络把该款游戏营销出去。为了让社交网络用户亲身使用并体验这款文档处理软件，该公司借助物质奖励途径给予体验软件的用户以鼓励，由于流动资金有限，它只能给予社交网络里有限用户以物质奖励。该软件公司希望此最初的用户能够喜欢这款软件并且能够影响社交网络里其它的朋友使用，然后其朋友再能够影响他们的朋友使用，依此类推，最终通过众口效应（word-of-mouth），使得社交网络里大部分用户接受并使用该款软件。因此，该公司面临的问题是如何选择最初的用户，使得社交网络

里最终影响的用户量最大以完成营销目的。

该实例描述了本文影响最大化问题。目前很多公司和个人都对该问题很感兴趣，因为他们想通过众口效应营销他们的产品，服务以及新颖想法。在线社交网络是一个很好的平台，它能提供很好的途径以达到众口效应。社交网络连接着庞大的用户群，并且它承载着巨大信息，这些信息包括社交网络的结构和动态交流信息等。然而，庞大的网络、极其复杂的组织结构和动态变化的性质，使得这个问题成为学术领域一个巨大挑战。解决这个问题的方法必须是有效的，快速的并且要具有可扩展性，这样的方法才能应用在生产实践中。

此外，随着网络的飞速发展，影响最大化问题在社会安定和舆情预警中亦有着重要指导意义，比如在我党政府倡导并努力建设的和谐稳定的社会主义社会中，一些恐怖分子、破坏集团和西方资本主义势力试图通过各种途径破坏我国稳定与和谐。那些不法分子极力尝试通过网络散播各种谣言与谬论以求蛊惑人心，这些不法组织想通过众口效应散播谣言，使被蛊惑人尽可能的多，最后达到破坏社会的稳定与团结的目的，因此政府需要找到信息传播途径并采取措施，予以制止及打击。

因此，社交网络影响最大化问题有着重要的理论及实际研究价值，它不仅对市场营销和广告发布有着重要的应用，还对舆情预警以及社会安定等方面也有着有十分重要的实际意义。

1.2 研究现状

围绕本论文主题，从影响力最大化问题的影响力、信息传播模型和算法三方面分析研究现状。

(1) 影响力最大化问题的影响力研究

在社交网络影响力分析方面，目前的一些工作是验证社会影响是否存在于社交网络里。文献[2,3]提出一些机制验证了社交网络里确实存在着社会影响力；文献[4]研究了社交网络用户影响其邻居用户的概率；此外，社会心理学研究了负面信息的影响力[5,6,7,8]，该文献中，作者指出在人们作决定时，负面信息比正面信息更具有影响力；社交网络的研究人员也在研究负面信息或者负面观点在社交网络上的传播以及这些负面信息对正面信息传播的影响[9,10]；文献[11]里提出了社交网络影响最小化问题，但在现实生活中，公司企业等关注更多的是影响力最

大的问题正面观点，这也是本文研究的出发点。

清华大学唐杰教授研究了社交网络的特定话题传播，在文献[12]里他给出不同话题下，节点影响力强度的计算公式，并且阐明了不同话题下的信息传播子网络的结构；实验表明在不同的话题下，信息传播子网的结构差异很大，并且人的个性差异决定了其兴趣话题的差异。文献[13]对两个非邻居节点关于不同话题的间接影响力建立数学模型，并给出求解算法，该文献还对某些特定节点对全局网络在不同话题下的影响力强度做了理论分析和实验。文献[14]研究了论坛里对某个特定用户影响最大的几个用户，在这篇文章里作者以回帖次数作为衡量用户之间影响强度的重要依据。

(2) 影响力最大化问题的信息传播模型研究

信息在社交网络传播过程中，用户可能被影响，然而信息的传播过程是观察不到的，其传播的规则也不明确，因此需要特定传播模型来帮助本文研究信息传播影响力。J. Leskovec 沿着信息传播路径做了一些研究[15,16,17]，并模拟了一个节点在未知网络中的传播率。文献[18,19,20]最早研究了社交网络里信息传播的两种基本传播模型：线性阈值模型（Linear Threshold Model，简称 LT）和独立级联模型（Independent Cascade Model，简称 IC）。基于以上工作，文献[21]把独立级联模型应用于研究社交网络影响力稀疏化。除了以上两种最基本传播模型，文献[22]利用物理学中的热传播过程模拟了社交网络中信息的传播。

M. Richardson 和 P. Domingos 等人[23,24]最早把社交网络影响最大化问题（即如何选择 k 个种子节点，使其在传播过程结束之后，传播的范围达到最大）转化为算法问题，且从概率角度解决了这个算法问题。进一步的研究工作出现在文献[18]中，作者 D. Kempe, J. M. Kleinberg 和 V. Tardos 最早把影响最大化问题转化为离散组合优化问题，用图理论对社交网络建模，节点表示社交网络用户，边表示它所连接节点之间的关系，信息按照某种传播模型在社交网络里传播。给定一个社交网络图，一个特定的信息传播模型和常数 k，影响最大化问题可以描述为：如何在图中找到 k 个节点（种子节点），以使得在特定信息传播模型下，该 k 个种子节点在社交网络中影响的节点数目最多。文献[25]提出了一个基于最短路径的信息传播模型。

针对影响最大化问题，目前已存在一些基本的信息传播模型，但这些模型没

有考虑网络中节点的相关性及重要性，而网络中节点的相关性和节点的重要性是衡量其影响力的重要指标，因此，本文提出了一种基于 PageRank 的信息传播模型，以提高社交网络中信息传播的影响力效果，此外，本文基于概率转移矩阵的思想建立了一种扩展的线性阈值模型，以分析大规模社交网络的信息传播问题。

(3) 影响力最大化问题的算法研究

影响最大化问题已被证明是 NP-hard 问题，文献[18]提出了适用于所用信息传播模型的近似贪心算法，该算法在每一步都选择当前最具影响力的节点作为初始传播对象进行传播（最具影响力节点是指当前能够激活最多节点的节点），然而，选择最具影响力的节点是一个非常耗时的过程，而且这种局部最优并不能保证最终的传播结果最优。该算法的缺点是效率太低，对于大型社交网络，这种贪心算法由于高耗时而不适用于实际应用。M. Kimura 和 K. Saito 提出了一个基于最短路径信息传播模型的影响力最大化算法[25]。文献[26]设计了 CELF 函数，用于选择新的节点，作者利用影响最大化目标函数的子模型性质（submodularity）去降低每一步选择当前最具影响力节点所耗用的时间，实验表明，该算法可以在每次选择当前最具影响力节点上提高 700 倍的速率，但不适用于大规模网络。文献[27,28,29,30,31]致力于研究解决影响最大化问题的算法效率问题。

目前近似求解影响最大化问题的算法有传统贪心算法及其一些改进算法，但对于规模很大的社交网络其时间复杂度很高，针对该问题，本文基于概率转移矩阵思想提出了一种扩展的线性阈值模型，并给出了一种新的影响最大化算法，即基于概率转移矩阵的影响最大化算法，该算法首先计算 t 时刻概率转移矩阵，然后利用贪心方法寻找 k 个最具影响力节点。

针对影响力最大化的信息传播模型和算法中存在的不足，本文利用概率转移矩阵及贪心方法思想来近似求解，将 NP-hard 问题转换为 P 问题，在降低影响力最大化问题的算法时间复杂度等方面做了相关研究。

1.3 论文创新点

本文的主要创新点如下：

1. 提出了一种基于 PageRank 的信息传播模型（PRP 模型）。该模型综合考虑了社交网络中节点的相关性及节点的重要性这两个衡量影响力的重要指标，克服了传统模型中存在的过于理想化假设的缺点（即每个

节点的不同父节点对该节点的影响力强度是相同的假设），实验表明该模型比传统的线性阈值模型、加权级联模型和独立级联模型的效果更好，影响力范围更大。

2. 提出了一种基于概率转移矩阵的影响最大化算法（PTMA 算法）。该算法通过计算 t 时刻概率转移矩阵并利用贪心方法寻找 k 个最具影响力节点，由于省去了每隔一段时间 t 就统计活跃节点的数目的时间，因此 PTMA 算法比传统的贪心算法及其改进算法的效率更高，降低了时间复杂度，且该算法的运行时间基本呈线性增长，因此，PTMA 算法适用于大规模社交网络环境。

1.4 本文组织结构

在文献查阅和调研基础上，本文主要研究了社交网络影响力最大化问题模型与算法，由五大章组成，组织结构如下：

第 1 章主要分析了社交网络影响力最大化问题的研究背景、意义、以及国内外研究现状，提出本文的创新点及论文组织结构。

第 2 章研究社交网络影响力最大化问题的相关基本理论，给出目前一些重要的信息传播模型及影响最大化问题在不同模型下的相应定义，并概括总结影响最大化问题目前解决方法。

第 3 章基于已有影响最大化信息传播模型中给定的假设过于理想化的问题，提出了一种基于 PageRank 的信息传播模型，该模型考虑了社交网络中节点的相关性及节点的重要性两个衡量影响力的重要指标。

第 4 章基于概率转移矩阵的思想提出了一种扩展的线性阈值模型，基于此模型，提出了一种新的影响力最大化算法—PTMA 算法。

第 5 章总结了全文工作，并展望了未来的研究方向。

1.5 本章小结

本章总结了社交网络影响力最大化问题的研究背景、意义和国内外重要的研究工作，提出了本文的创新点，给出本文的章节结构。

第 2 章 相关模型与理论

本章研究了社交网络影响力最大化问题的相关基本理论，给出一些重要信息传播模型和影响最大化问题在不同模型下的相关定义，并总结分析了影响最大化问题现有的解决方案。本文基于相关模型和理论，针对已有研究中存在的一些问题，在第 3 章和第 4 章给出相应的解决方法。

2.1 基本理论

基于图理论，分析社交网络结构，在图和社交网络的基础上给出社交网络中影响最大化问题的定义。

2.1.1 图理论

本文基于图论分析社交网络结构。社交网络结构一般用有向图 $G = (V, E)$来表示，在 G 中，V 和 E 分别表示网络中节点和边的集合，N 和 L 分别表示 V 和 E 中元素的个数，在图 G 上研究社交网络影响最大化问题。需要用到的相关理论如下：

有向图是指连接两个节点之间的边有方向性，边上的箭头表示方向。计算机科学家常用有向图形象地表示节点之间的因果关系与信息在节点之间的传播方向。

如果$(u_{i-1}, u_i) \in E$, (i =1,…,n) ，称$(u_0,\ldots,u_n)$为u_0到u_n的一条路径。如果从 u 到 v 有一条路径，称 u 到 v 是可达的。对于 G 的一个节点 v，定义 $F(v;G)$为 v 指向的节点的集合，$B(v;G)$为指向 v 的节点的集合，对任意的集合 A 包含于集合 V，即 V 的任何子集 A，有 $F(A;G)=\bigcup_{v\in A} F(v;G)$, $B(A;G)=\bigcup_{v\in A} B(v;G)$。

对于集合 V 的最大子集 C 构成的图称为 G 的强连通图(SCC)，因为对于所有的 $u,v \in C$，都有一条从 u 指向 v 的边。对于 G 的节点 v，称 SCC(v ; G)为包含节点 v 的强连通图。如果两个节点 u、v 之间存在一条从 u 指向 v 的有向边$(u,v) \in E$，则表示相应的两个社交网络用户之间存在联系，用户 u 可以影响用户 v。

2.1.2 社交网络影响最大化问题

社交网络中节点只有两个状态，活跃状态（active）和非活跃状态（inactive），活跃状态的用户是受影响的用户，非活跃状态的用户是当前没有受影响的用户，

活跃状态的用户可以影响非活跃状态的用户。

信息在社交网络 G 中传播，假设活跃节点的初始集合为 $A(\subseteq V)$，A 之外的所有节点都是非活跃的，$RS(A)$为社交网络中最终受影响的活跃节点的集合，则初始集合 A 的影响力范围可以定义如下：$\varphi(A)=\left|RS(A)\right|$，$\varphi(A)$ 表示最终活跃的节点的数目。因此，影响最大化问题可以表示为一个离散最优化问题：在社交网络 G 中，给定参数 k，信息根据特定的传播模型在 G 中传播，找到一个包含有 k 个节点的初始集合A，使得初始集合A在社交网络中最终影响的活跃节点数目最多，影响范围最大，即 $\varphi(A)$ 最大。

2.2 社交网络中信息传播模型

信息在社交网络中传播的最基本的模型有线性阈值模型（Linear Threshold Model，简称 LT 模型）以及独立级联模型（Independent Cascade Model，简称 IC 模型）。这两种模型成立的假定条件如下：一个节点接受信息后就变成活跃节点；节点的状态只有活跃和非活跃两种；节点能够从非活跃状态变为活跃状态，反之不可；社交网络中信息的传播表现为活跃节点的传播；在信息传播的第 0 步，本文认为活跃节点初始集合 A 中的节点最先变活跃，其他的节点都是非活跃状态；活跃节点的传播过程将会在第 $t(t\geq 0)$时刻体现出来。

在社交网络分析中，线性阈值模型和独立级联模型为目前使用最多的信息传播模型，下面将介绍这两种模型以及具有广泛应用和理论价值的其他重要模型。

2.2.1 线性阈值模型

统计学与应用数学最早研究了 LT 模型。在 LT 模型中，用户节点 $v\in V$，被其父节点 u 以正权值 $\omega(u,v)$激活，$\sum_{u\in\Gamma(v)}\omega_{u,v}\leq 1$，给定活跃节点的初始集 A，活跃节点的传播按照如下规则传播：

1. 对任意节点 $v\in V$，从 [0,1]区间上随机选取一个阈值，用θ_v表示；
2. 在传播的 t 时刻，非活跃子节点 v 被活跃父亲节点 u 以权值 $\omega(u,v)$激活；
3. 如果 v 的所有活跃父节点对其影响的权重之和大于等于 v 的阈值θ_v，

即，$\sum_{u\in\Gamma(v)}\omega_{u,v}\geq\theta_v$，那么非活跃节点$v$在$t+1$时刻被父亲节点激活，其中$\Gamma_t(v)$表示$t$时刻$v$的所有活跃父亲节点的集合；

4. 如果没有更多的节点被激活，那么该传播过程就终止。

阈值θ_v表示当父节点成为活跃节点（该节点接受某个观点或购买了某个商品）时，其子节点同样成为活跃节点的潜在倾向的不同。LT 模型是一个与 0-1 分布有关的概率模型，节点的阈值选取是随机的。对于一个初始活跃节点集合$A(\subseteq V)$，用$\varphi(A)$表示随机激活过程结束时活跃节点的个数，$\varphi(A)$是一个随机变量，用$\delta(A)$表示$\varphi(A)$的期望值，本文称$\delta(A)$为初始集合 A 的影响度。

下面本文描述在线性阈值模型和独立级联模型下的影响最大化定义[32]：在社交网络 G 中，定义 k 是一个小于 N 的正整数，对于初始集合 A，用$\varphi(A)$表示集合 A 扩散后，最终被激活的节点数目，用$\delta(A)$表示最终活跃节点数目的期望值，即$\varphi(A)$的期望，则影响最大化问题可以定义为：找到一个包含 k 个活跃节点的初始集合A_k^*，使得$\delta(A_k^*)\geq\delta(S)$，其中 S(⊆V)也是包含 k 个节点的集合。影响最大化问题可以表示为：$A_k^*=argmax_{A\in\{S\subseteq V;|S|=k\}}\delta(A)$，其中，|S|表示集合 S 的元素的个数。该定义同样适用于 2.2.2-2.2.6 小节各种模型下的影响最大化问题定义。

2.2.2 独立级联模型

独立级联模型是以发送者为中心的模型，是基于概率理论里面的交互粒子系统设计的一个信息扩散模型。在 IC 模型里，所有有向边(u,v)都存在一个实数值$p_{u,v}\in[0,1]$，$p_{u,v}$表示活跃节点u通过边(u,v)成功影响节点 v 的概率。用集合 A 表示活跃节点的初始集，则信息传播的规则如下：

1. 在 t 时刻，活跃节点 u 以概率$p_{u,v}$尝试激活非活跃节点 v，如果激活成功，则节点 v 在 $t+1$ 时刻成为活跃节点。且 u 只有一次激活 v 的机会，如果 v 没有被 u 激活，则活跃节点 u 永远不能再尝试激活 v；
2. 假设在 t 时刻 v 被多个活跃父结点尝试激活，那么他们将以任意次序

按照一定概率尝试激活 v；

3. 该传播过程的终止条件是网络中不再有新的非活跃节点被激活。

在 IC 模型中，信息在有向边间的传播成功率$p_{u,v}$是随机的，对于一个初始活跃节点集合 $A(\subseteq V)$，用 $\varphi(A)$ 表示随机激活过程结束时活跃节点的个数，$\varphi(A)$ 是一个随机变量，用 $\delta(A)$ 表示 $\varphi(A)$ 的期望值，本文称 $\delta(A)$ 为初始集合 A 的影响度。

2.2.3 递减级联模型（Decreasing Cascade Model）

在社交网络中，如果一个用户（节点）接受某个产品或信息，称该节点是活跃的，否则就称非活跃的。假定，一旦某个节点成为活跃节点，该节点永远都是活跃的。当 t 时刻，节点 u 成为活跃节点，就说 u 被感染，u 有一次机会影响非活跃的邻居节点 v，如果影响成功，则节点 v 在 t+1 时刻变成活跃节点。如果 t 时刻，有多个活跃的邻居节点影响 v，则这些邻居节点以任意顺序尝试感染 v，但是感染都在 t 时刻进行。

为了充分描述递减级联模型，需要说明节点 u 尝试激活节点 v 成功的概率，在最简单的独立级联模型中，激活成功概率用$p_v(u)$表示，跟被激活行为的历史记录无关，即$p_v(u)$跟激活行为是独立的。但是，节点 v 被激活的概率可能随着其邻居节点尝试激活它的过程而变化，即 v 已经被很多节点尝试激活很多次没有成功，新激活的邻居节点 u 对 v 的影响就会被削弱。如果 S 表示已经尝试激活节点 v 而没有成功的邻居节点的集合，那么用$p_v(u, S)$表示节点 u 成功激活节点 v 的概率。首先看一下顺序无关性级联概念：如果集合 T 中所有节点试图影响 v，那么它们尝试激活 v 的顺序不影响 v 最终被激活的概率。假设$u_1, u_2, \cdots, u_r$和$u_1', u_2', \cdots, u_{\mathrm{r}}'$是集合 T 的两个变换矩阵，可以表示为$T_i$={$u_1, u_2, \cdots, u_{i-1}$}和$T_{\mathrm{i}}'$={$u_1', u_2', \cdots, u_{\mathrm{i}-1}'$}，那么顺序无关性可以用下式表示：

$\prod_{\mathrm{i}=1}^{\mathrm{r}}\left(1-p_v\left(u_i,\ S\cup T_i\right)\right)=\prod_{\mathrm{i}=1}^{\mathrm{r}}(1-p_v(u_{\mathrm{i}}', S\cup T_{\mathrm{i}}'))$，其中$\mathrm{S}\cap \mathrm{T}=\emptyset$。

在递减级联模型中，函数$p_v(u, S)$是关于 S 递减的，即有$p_v(u, S)\geq p_v(u, T)$，其中 $S\subseteq T$。在该限制下，如果已经有很多节点尝试过激活节点 v，那么一个传染性的节点再去激活节点 v 成功的概率就会减小，文献[33]里对该模型作了详细的理论分析与验证。

2.2.4 加权级联模型（Weighted Cascade Model）

在独立级联模型中，激活概率$p_{u,v}$没有考虑节点的度，然而，度较高的节点影响的与被影响的概率都较高，基于此，文献[34]提出了加权级联模型，度数高的节点关联的边被赋予较低的激活概率，加权级联模型是独立级联模型一个特例模型，节点 u 对 v 的激活概率为$p_{u,v}=1/d_v$，d_v表示节点 v 的度。

2.2.5 非渐进模型（Non-progressive Model）

线性阈值模型和独立级联模型都遵循 Progressive 规则，即节点只能由非活跃状态被影响成为活跃状态，不能从活跃变成非活跃状态。非渐进模型则不受此限制，节点可以在活跃状态与非活跃状态之间相互转换，实际上可以分解成为满足 Progressive 规则的模型。文献[35]里对该模型进行了边界分析。

2.2.6 触发模型（Triggering Model）

每个节点v根据它的邻居子集的分布，独立随机选择一个触发集T_v，选定一个初始激活集合 A，如果t时刻，v有一个邻居节点在v的触发集T_v中是活跃的，那么，t+1 时刻 v 变成活跃节点。如果节点 u 在 v 的触发集T_v中，那么就称边(u,v)是活(live)边，如果 u 不在T_v中，(u,v)就是阻碍(blocked)边，在触发模型的实例中，当且仅当从初始集 A 到节点 v 有一条活边，节点 v 才能被激活。

子模性[36]是指当添加一个节点 v 到初始集合 A 时，如果集合 A 越小，v 对边际效益的增量影响就越大，它是证明算法精确度保证的必要条件。假设将有限集 W 映射到实数集 R 上，当函数 F 满足边际效益（marginal gain）递减性质时，F 称之为子模函数。子模函数形式化表示如下，F：$2^W \to R$，$A \subseteq B \subseteq W$, $v \in W/B$，如果满足 $F(A \cup \{v\}) - F(A) \geq F(B \cup \{v\}) - F(B)$，则 F 称为子模函数。

在触发模型每个实例中，影响函数 $\delta()$ 都具有子模性，该函数无输入参数。前面提到的 LT 模型和 IC 模型都是触发模型的特例，文献[37]在研究异步学习网时使用了该触发模型。

2.2.7 热传播模型（Heat Diffusion Model）

热量扩散是一种物理现象，在传播媒介中热量总是从高温部分传到低温部分。下面本文以热的传播过程模拟信息在社交网络中的传播过程，事实上，人们在影响别人时的过程跟热量的传播过程很类似，在社交网络中，热源就是某个产品的创新者或早期接受者，抑或某种创新的行为，并且具有很高的热量，这群人开始影响别人并将其影响传播给最初的某一群体，接着后期的某一群体，以致最终某一时刻，热量传播到社交网络的每个边缘，也就是产品或信息被接受。

在该模型中，用图表示社交网络，社交网络中的每个消费者或顾客用图中的节点表示，人与人之间的关系用图中的边表示。下面用三种不同的模型来描述三种不同的社交网络：无向社交网络，有向社交网络，有先验知识扩散概率的有向社交网络。

无向图 $G=(V, E)$表示无向社交网络，V 表示节点集合且$V=\{v_1, v_2, \cdots, v_n\}$，$E$ 表示所有边的集合且 $E=\{(v_i, v_j)\}$，(v_i, v_j)表示节点v_i 与v_j之间有一条边。$f_i(t)$表示 t 时刻节点v_i的热量，从最初零时刻v_i的热量$f_i(0)$开始传播。$f(t)$表示$f_i(t)$的向量。

假设在 t 时刻，每个节点v_i都在Δt时间内,从其邻居节点v_j 接受 $M(I, j,t, \Delta t)$的热量，热量 $M(I,j,t, \Delta t)$应该与时间段Δt成正比，与温差$f_j(t) - f_i(t)$成正比，并且热量是从节点v_j传到v_i。令 $M(I,j,t,\Delta t)= \alpha t \boldsymbol{H}(f_j(t) - f_i(t))\Delta t$，其中，$\alpha$ 是热传播系数。节点v_i在时刻 $t+\Delta t$时与 t 时刻的温差与它从其邻居节点接受到的热量之和相等。公式化描述如下：

$$\frac{f_i(t+\Delta t)-f_i(t)}{\Delta t}=\alpha t \boldsymbol{H} \sum_{j:(v_j,v_i)\in E} \left(f_j(t) - f_i(t)\right)$$

表示成矩阵形式如下：

$\frac{\mathbf{f}(t+\Delta t)-\mathbf{f}(t)}{\Delta t} = \alpha \boldsymbol{H}\ \mathbf{f}(t)$，其中 $H_{ij} = \begin{cases} 1, (v_i, v_j) \in E \text{ or } (v_j, v_i) \in E \\ -d(v_i), \quad i = j \\ 0, \quad \text{otherwise.} \end{cases}$，当 $\Delta t \to 0$ 时，上式变形为$\frac{d}{dt} \boldsymbol{f}(t) = \alpha \boldsymbol{H} \boldsymbol{f}(t)$, $\boldsymbol{f}(t)= e^{\alpha t \boldsymbol{H}} \boldsymbol{f}(0)$,其中 $d(v_i)$表示节点v_i的度，$e^{\alpha t \boldsymbol{H}}=\mathbf{I}+\alpha t \boldsymbol{H}+\frac{\alpha^2 t^2}{2!} \boldsymbol{H}^2+\frac{\alpha^3 t^3}{3!} \boldsymbol{H}^3+\ldots$ 矩阵 $e^{\alpha t \boldsymbol{H}}$是传播源，在此意义下热量可以无限的扩散下去。

在很多情况下，社交网络是有向的，尤其是那些在线推荐系统或知识分享网站，每个用户在知识信任网站中都有一个信任列表，信任列表中的用户会深深地影响该用户，因为用户 a 在用户 b 的信任列表里，用户 b 不在用户 a 的信任列表，所以，用户 a 与用户 b 的关系是有向的。

在有向图 $G(V,E)$中，热量从节点点v_i经过边(v_i, v_j)流到节点v_j上，假设 t 时刻，每个节点v_i都在Δt时间内从v_j接受 RH 的热量，RH=RH$(i,j,t, \Delta t)$，RH 与Δt成正比，亦与节点v_j具有的热量成正比。如果没有从v_j到v_i的边，那么 RH=0。最终，节点v_i从指向它的所有邻居节点接受到$\sum_{j:(v_j,v_i)\in E} \sigma_j \mathfrak{f}_j(t)\ \Delta t$。同时，节点$v_i$会将 DH 的热量传给其下一邻居节点，DH =DH$(I, t, \Delta t)$，热量 DH 与时间段Δt成正比，亦与节点v_i具有的热量成正比,每个节点都有同样的传播热量的能力，热量 DH$(I, t, \Delta t)$应该被统一传递给其邻居节点。真实的动态社交网络会更加复杂，但是为了模型更加简明，本文在此做了点简化。

节点v_i传递给它的每个邻居节点的热量为$\alpha \mathfrak{f}_i(t)\ \Delta t/d_i$,即每个邻居节点会接受到这么多的热量，d_i是节点 i 的出度，因此，$\sigma_j=\alpha/d_i$ 。如果节点 i 的出度是 0，那么该节点不会向其他节点传播热量。节点v_i在时刻 $t+\Delta t$与 t 时刻的温差与它从其邻居节点接受到的热量之和相等。公式化描述如下：

$\frac{\mathfrak{f}_i(t+\Delta t)-\mathfrak{f}_i(t)}{\Delta t}=\alpha\left(-\tau_i \mathfrak{f}_i(t)+\sum_{j:(v_j,v_i)\in E}\frac{1}{d_j}\mathfrak{f}_j(t)\right)$,其中$\tau_i$是判断节点 i 是否有出边的标志，如果节点 i 没有出边，那么τ_i=0，如果节点 i 有出边，则τ_i=1。可以将上一公式转换成如下矩阵形式解决:

$$\boldsymbol{f}(t)= e^{\alpha t \boldsymbol{H}}\ \boldsymbol{f}(0)，其中\ H_{ij}=\begin{cases}1/d_j, & (v_j,v_i)\in E\\ -\tau_i, & i=j\\ 0, & \text{otherwise.}\end{cases}$$

上文提到的热传播模型中，每个人都以相同的概率将信息或产品传播给其指向的邻居节点，而实际上，由于时间和精力的限制，一个人不会将信息传递给关系列表中的每个人。例如，当你发现一个Gmail或Hotmail非常有用，你会将其传给你最好的朋友或需要该信息的朋友，但不是所有的朋友，因此给出带有先验知识的分散概率有向社交网络热传播模型。有向图$G=(V, E, P)$，相关符号含义如下 ：

（1）V=$\{v_1, v_2, \cdots, v_n\}$, V表示节点集合

（2）P =$\{p_{ij}|p_{ij}$表示边(v_i, v_j)存在的概率$\}$

（3）E=$\{(v_i, v_j)|(v_i, v_j)$表示从v_i到v_j存在一条边且$p_{ij} > 0\}$

因为每个人都有其个性，有些人在社交网络中非常活跃，他们愿意跟朋友们分享自己喜欢或讨厌的东西，有些人在传播信息的行为中不活跃，因此，用ω来描述每个人的个性因素。根据以上的分析及上一小节热传播模型的分析，节点v_i在时刻t+Δ t与t 时刻的温差与它从其邻居节点接受到的热量之和相等。公式化描述如下：

$$\frac{\mathrm{f}_i(t+\Delta t)-\mathrm{f}_i(t)}{\Delta t}=\alpha\left(-\frac{\tau_i\omega_i}{d_j}\mathrm{f}_i(t)\sum_{k:(v_i,v_k)\in E}p_{ik}+\sum_{j:(v_j,v_i)\in E}\frac{\omega_j p_{ji}}{d_j}\mathrm{f}_j(t)\right)$$

其中，τ_i是判断节点 i 是否有出边的标志，参数传播概率 p 和个性因素 ω 的取值范围为[0,1]。可以将上一公式转换成如下矩阵形式解决：

$$\boldsymbol{f}(t)=\ \mathrm{e}^{\alpha t\boldsymbol{H}}\ \boldsymbol{f}(0)，\text{其中 } H_{ij}=\begin{cases}\omega_j p_{ji}/d_j, & (v_j, v_i)\in \mathrm{E}\\ -(\tau_i\omega_i/d_j)\sum_{k:(v_i,v_k)\in E}p_{ik}, & i=j\\ 0, & \text{otherwise.}\end{cases}$$

热传播模型下的影响最大化问题定义：在有N个用户的社交网络中，首先选择k个用户作为热源，用集合S_k表示，在t=0时刻，给集合S_k中每个用户一定量的热量h_0,即 $\mathrm{f}_i(0)=h_0$，其中$i\in S_k$。随着时间推移，热量最终能传遍整个社交网络。如果某一时刻t，用户j具有的热量不低于某个阈值θ，即 $\mathrm{f}_j(t)\geq\theta$,那么认为该用户被影响成功，即接受产品。用$I_{S_k}$表示$S_k$的影响集，表示在t时刻接受产品的用户的期望数值。本文的目标是找到最具有影响力的k个用户的集合S_k，使得t时刻集合I_{S_k}的元素数目最大，I_{S_k}=$\{\ j|\ \mathrm{f}_j(t)\geq\theta, j\leq \mathrm{N}\ \}$。

以上这些模型研究为本文建立信息传播模型做了基础理论铺垫，此外，本文在第4.2节基于线性阈值模型做了进一步的改进，提出了扩展的线性阈值模型，为设计PTMA算法奠定了模型基础。

2.3 基本算法

影响最大化问题是 NP 问题，即不确定是否有多项式解法，为了求解影响最大化问题，目前算法大都采用近似的方法将 NP 问题转换成多项式 P 问题求解。

2.3.1 贪心算法

D. Kempe 和 J. M. Kleinberg [18]用贪心算法来解决影响最大化问题，为了找到 2.2.1-2.2.6 模型中要求的初始扩撒集合A_i^*，一个有效的方法是每一步根据贪心算法的标准确定初始集合中的一个节点，直到找到k个节点为止。首先定义$A_0^*=\emptyset$；$I(A_i^*)$：集合A_i^*扩散后被激活节点的集合；$m(u|A_i^*)$=| I（$A_i^* \cup \{u\}$）|-| I（A_i^*）|：节点 u 相对于集合A_i^*的边际效益影响范围。然后，每一步都选择当前最具有影响力的节点，从A_0^*开始，在第 i 步，根据局部最优策略选择节点 u，则 $u=argmax_v\, m(v|A_{i-1}^*)$，且令 $A_i^* = A_{i-1}^* \bigcup \{u\}$，其中 $v \in V \setminus A_{i-1}^*$。计学与应用数学最早研究了 LT 模型。

尽管该贪心算法能在 1-1/e 的因子内近似求解最优解，但其缺点是每一步都要计算所有未被激活节点 u 的边际效益 $m(u|A_i^*)$，因此运行非常耗时，即运行效率很低，时间复杂度很高。针对这个问题，W. Chen, Y. Wang 利用独立级联模型（IC 模型）的子模性[20]，提出了一个改进的贪心算法（CELF 算法），该算法在每一步选择节点时，很多节点的增量不需要重新计算，因为它们在之前的步骤中的值已经小于其它节点在当前步骤中的值，因此降低了算法的时间复杂度，但影响范围没有扩大而且 CELF 算法不适用于 LT 模型。

2.3.2 FPTAS 算法

Even-Dar, Shaira 将影响最大化问题转换成 0-1 背包问题[38]，虽然 0-1 背包问题也是 NP 问题，但是目前已有很多方法可以将 0-1 背包 NP 问题转化成多项式 P 问题近似求解。

2.3.3 Top-K 算法

H. Ma, H. Yang, M. R. Lyu 等人[22]提出了热传播模型，最终把影响最大化问题转化成求解 $\boldsymbol{f}(t)= e^{\alpha t \boldsymbol{H}}\ \boldsymbol{f}(0)$，针对这个问题，提出了 Top-k 算法，先计算出每个用户的影响集I_i (i=1,2,⋯,N)，再找出最具影响力的 k 个节点。具体描述如下：首先确定 N 个用户的社交网络图及阈值参数 θ；然后执行计算 $\boldsymbol{f}(t)= e^{\alpha t \boldsymbol{H}}\ \boldsymbol{f}(0)$，对社交网络中每一个用户节点 i，如果 j 节点满足$f_j(t) \geq \theta$，就把用户节点 j 加到 i的影响集$I_i(t)$中，遍历社交网络中每个节点，最终得到N个节点的N个影响集I_1，

I_2，…，I_N，按影响集中元素数目由高而低排列，找到前 k 个节点的影响集对应的前 k 个节点，即最具影响力的 k 个节点。该近似算法忽略了影响集I_i中重复元素的存在，文献[22]随后给出了改进的 k-Step 贪心算法，但是随着 k 值的增大，该算法的精确度下降。在真实的社交网络中，传播过程一开始可能就有几个传播源同时开始传播，社交网络中的用户接受某一消息可能是受几个传播源的影响，于是他们在文献[22]又提出了加强的 k-Step 贪心算法，尽管该算法更贴近真实的社交网络，但是其运算比前两个算法更复杂。

以上这些算法为本文设计影响力最大化算法做了基础理论铺垫，基于本文所给出的传统的信息传播模型和传统的贪心算法及其改进算法，本文设计了PRP模型方法（见第3章）和PTMA算法（见第4章）。

2.4 本章小结

本章首先研究了与影响最大化有关的基本理论，然后分析了当前最基本的两种信息传播模型及一些重要传播模型，并在模型中给予相应的影响最大化定义，最后，给出目前解决影响最大化问题的三种基本算法并分析了算法的优缺点，为本文研究社交网络影响力最大化问题的信息传播模型与算法打下理论基础。

第 3 章 基于 PageRank 的信息传播模型

第 2 章分析了当前比较重要的一些信息传播模型，但是这些模型的假设过于理想化，没有考虑社交网络中节点的相关性和重要性这两个衡量影响力的重要指标，针对该问题，本章提出了一种基于 PageRank 的信息传播模型，该模型综合考虑了节点的相关性和节点的重要性对影响力最大化问题的影响，实验结果验证了该模型的有效性和可行性。

3.1 PRP 模型思想

社交网络影响力是指人们接受他人信息传播的过程。在该领域中，Domingos，Richardson 提出了社交网络影响最大化问题[23]，用图来表示社交网络。本文的目标是在图中找出最具有影响力的 k 个节点，使得社交网络中最终被影响的节点数目最多，信息传播范围最大。信息在社交网络传播过程中都遵循一定的规则，称之为信息传播模型。随着社交网络的出现及流行，社交网络影响力成为目前研究的热点。而在不同的传播模型基础上，研究影响最大化问题具有重要意义。

目前已存在一些基本传播模型，比如上一章阐述的线性阈值模型（LT 模型），独立级联模型（IC 模型）和加权级联模型（WC 模型）等等。这些模型都基于一个共同的假设：社交网络里所有节点的父亲节点对该节点的影响强度都是一样的。这种假设有些勉强。比如某个社交网络里用户甲有一些朋友，这些朋友对用户甲的重要性不同，有些是未谋面的网友，有些则是从小玩到大的关系非常要好的朋友，这就导致了当用户甲做某个决定或者形成某个观点时，他的朋友们对该用户的影响程度不同。所以社交网络里所有节点的父亲节点对该节点的影响强度都是不一样的。上一章描述的信息传播模型都没有考虑网络中节点的相关性和节点的重要性，本文认为网络中节点的相关性和节点的重要性是衡量其影响力的重要指标。只有把社交网络里所有节点的父亲节点对该节点的影响强度按照不同父亲节点对该节点的重要性的不同区别开，才能使得信息传播模型更具有实际意义和基于该模型的影响最大化问题更有应用价值。

本文用 PageRank 算法计算网络中节点的 PageRank 值，而该数值恰好可以衡量节点的相关性和节点的重要性。于是，本文试图通过 PageRank 算法计算网络中所有节点的 Pagerank 值。本文认为如果该节点的 Pagerank 值比较高，那么

该节点的影响强度就比较大。基于此，本文提出了一种基于 PageRank 算法的信息传播模型 (PRP 模型)，并在该模型下利用贪心算法来近似求解影响最大化问题。

3.2 PageRank 算法

PageRank [39,40]也就是网页排名，又称网页级别、Google 左侧排名或佩奇排名。它是一种由搜索引擎根据网页之间相互的超链接计算的技术，该技术由 Google 创始人拉里•佩奇和谢尔盖•布林于 1998 年在斯坦福大学研发。该算法在商业界取得巨大的成功，因此，也成为其他搜索引擎与学术界最关注的算法之一。目前很多分析网络的计算模型都是基于 PageRank 算法基础之上的。Google 根据网页的 PageRank 值来判断该网页的等级或者重要性。在融合了很多诸如关键词标识和题目标识等其他影响因子之后，Google 利用改进的 PageRank 算法获得了更相关的质量更高的搜索结果。PageRank 值的区间为[0，10]，PageRank 值越高，就表明该网页越重要。比如，一个网页的 PageRank 值为 2 ，那么关注该网页的人很少，该网页不重要。一个网页的 PageRank 值为 7 到 10 之间，该网页很重要，很容易影响到其他人，也容易引起其他人的关注。

在 PageRank 算法提出之前，研究人员利用入链的数量来进行链接分析。他们假设网页的入链越多，该网页就越重要。在入链数量的基础之上，PageRank 算法还考虑了网页的质量，这使得 PageRank 算法获得更好的搜索结果。PageRank 算法有两大基本假设。

- 质量假设：链入页面甲的页面质量不同，页面甲的质量也不同。质量高的页面会通过链接把更高的权重指向其他页面，也就是说指向页面甲的页面质量越高，那么页面甲的质量就越高。
- 数量假设：在网络模型中，如果一个页面的入链数目越多，那么该页面就越重要。

基于以上的两大基本假设基础之上，初始时，PageRank 算法为每个网页赋予相同的初始值，然后迭代计算每个网页的 PageRank 值，直到算法收敛为止。网页的 PageRank 值用以表示该网页的重要性。它和用户用以搜索的关键词没有关系。如果一个搜索引擎完全采用 PageRank 值来排序，那么，无论用户输入什么查询词，搜索引擎总是返回 PageRank 值最高的页面。为了解决这个问题，斯

坦福大学的学者提出了基于主题的 PageRank 算法[41]：不再采用平均传递网页的 PageRank 值，而是根据链接到网页的主题相关性来传递。本小节重点阐述 PageRank 的原理，以便于本章提出的 PRP 模型计算节点之间的权重。

PageRank 算法通过两大步骤来计算各个网页的 PageRank 值的。

1. 初始阶段：初始时，PageRank 算法为每个网页赋予相同的初始值，随着每一轮迭代计算，各个网页的 PageRank 不断更新，一直到算法收敛为止。
2. 迭代阶段：在迭代更新网页的 PageRank 值时，每个页面将其当前的 PageRank 值按照出链平均分配给它链出的页面。然后，每个页面对所有分配给它的权重加和就可以得到更新的 PageRank 值。当每个 PageRank 值更新后，就完成这一轮的迭代计算。

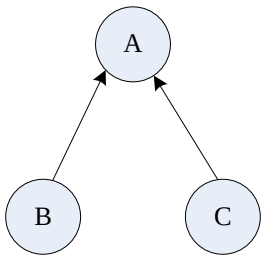

图 3.2.1 包含三个节点的社交网络

本文用 PR(W)表示网页 W 的 PageRank 值，L(W)表示网页 W 的出链数目。下面通过一个例子来看如何计算每个网页的 *PR* 值的。如图 3.2.1 所示，图里包含三个节点 A，B，C。B 和 C 都有一条指向 A 的链接。初始时假定 $PR(A)=PR(B)=PR(C)=0.5$。因为 B 和 C 都有一条指向 A 的链接，所以 B 和 C 都为 A 贡献了 0.5 的 *PR* 值。A 的 *PR* 值更新为：

$PR(A) = PR(B) + PR(C) = 1$。

由于没有链接指向 B 和 C，那么在这次计算中，$PR(B)$和 $PR(C)$的值将被赋为 0。

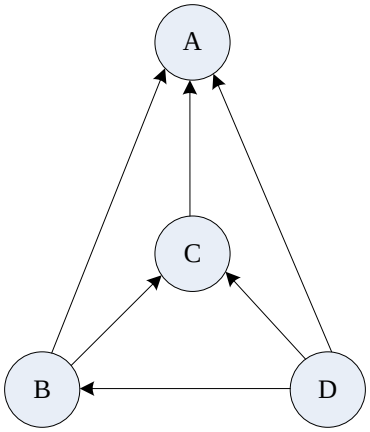

图 3.2.2 包含四个节点的社交网络

如图 3.2.2 所示，图里包含四个节点 A，B，C 和 D，其中 B 有两条链分别指向 A 与 C；D 有三条链分别指向 A，B 和 C；而 C 有一条链指向 A。对于 B 来说，它把自己的 PR 值平均分给了 A 和 C，A 与 C 分别得到来自于 B 的 0.25 的 *PR* 值。对于 C 和 D 同理而言。一个网页传递给其他网页的 *PR* 值等于该网页的 *PR* 值除以该网页的出链总数，那么在这次计算中，各个网页的 *PR* 值更新为：

$PR(A)= PR(B)/2 + PR(C)/1 + PR(D)/3$

$PR(B)= PR(D)/3$

$PR(C) = PR(B)/2 + PR(D)/3$

$PR(D) =0$

对于 A 来说，本文形式化 A 的 *PR* 值计算公式如下：

$$PR(A)=\frac{PR(B)}{L(B)}+\frac{PR(C)}{L(C)}+\frac{PR(D)}{L(D)}$$

由于社交网络里存在一些孤立网页，这些网页的出链数目为 0，因此要对 PageRank 算法改进，也就是引入 q（阻尼系数 damping factor）。一般设定阻尼系数 q=0.85。q 的意义是用户沿着链接向后浏览的概率，用户以 1-q 的概率浏览新页面的概率，那么 A 的 PR 值按照引入阻尼系数的计算公式为：

$PR(A) = (PR(B)/L(\mathrm{B}) + PR(\mathrm{C})/L(\mathrm{C}) + PR(\mathrm{D})/L(\mathrm{D}))*q+1-q$

这个公式就是文献[42]定义的公式，而文献[43]更精确的表示了每个网页的 PR 值的计算公式。

$$PageRank(p_i)=\frac{1-q}{N}+q\sum_{p_j}\frac{PageRank(p_j)}{L(p_j)}$$

$p_1, p_2, \cdots p_N$ 表示页面，N 是页面的总量。下面可以用一个向量来表示所有

网页的 PR 值：

$$R=\begin{bmatrix} PageRank(p_1) \\ PageRank(p_2) \\ \vdots \\ PageRank(p_N) \end{bmatrix}$$

R 则是如下方程的解：

$$R=\begin{bmatrix} (1-q)/N \\ (1-q)/N \\ \vdots \\ (1-q)/N \end{bmatrix}+q\begin{bmatrix} \ell(p_1,p_1) & \ell(p_1,p_2) & \cdots & \ell(p_1,p_N) \\ \ell(p_2,p_1) & \ddots & & \\ \vdots & & \ell(p_i,p_j) & \\ \ell(p_N,p_1) & & & \ell(p_N,p_N) \end{bmatrix}R$$

其中，$\ell(p_i,p_j)=\dfrac{1}{L(p_j)}$，而且$\sum_{i=1}^{N}\ell(p_i,p_j)=1$。如果网页 i 不指向网页 j，则 $\ell(p_i,p_j)$ 为 0。

任何页面的 PR 值都是按照上述的公式计算得到的。PageRank 算法会一直迭代更新所有页面的 PageRank 值，一直到所有的 PageRank 值趋于稳定为止。

3.3 PRP 模型

由于传统传播模型没有考虑节点的相关性和节点的重要性，而节点的相关性与重要性则是衡量影响力的重要因素，因此，本文提出了一种基于 PageRank 的传播模型（PRP 模型）。

本文用 G 代表网络图结构，E_G 代表链接的集合，PageRank 值越高的父节点对其子节点的影响力越大。节点 v_i 的 PageRank 值以 p_i 来标识。节点 v_i 的第 j 个父亲节点的 PageRank 值以 p_i^j 来标识。v_i 父亲节点的集合以 $parent_i$ 标识。链接 (v_j,v_i) 上的权值以 ω_{ji} 标识。

假设不存在链接 (v_j,v_i)，则 ω_{ji} 为 0。如果存在，用节点 j 的 PageRank 值除以全部链入节点 i 的节点的 PageRank 值的总值就得到了 ω_{ji}。ω_{ji} 的数学表达式如下：

$$\omega_{ji}=\begin{cases}\dfrac{p_i^j}{\sum\limits_{k\in parent_i}p_k}, (v_j,v_i)\in E_G\\ 0 \qquad\quad , (v_j,v_i)\notin E_G\end{cases}$$

假设每个节点都存在一个 0-1 函数，即，如果节点处于非激活状态，则该节点的状态函数值为 0，如果节点处于激活状态，则该节点的状态函数值为 1。首先，所有节点处于非激活状态，即状态函数值都为 0。随机选取一些节点（传播源），把这些节点激活，其状态函数值都是 1。接下来，这些激活的节点将尝试去激活那些处于非激活状态的邻居节点。如果那些邻居节点被激活，则这些节点的状态函数值从 0 变为 1，且不可逆。否则，这些节点的状态函数值就是 0。在图 G 里存在两大类永远处于非激活状态的节点：一类是不存在从传播源到达这些节点传播路径的节点；一类就是没有入链的节点。下面给出三种状态稳定的节点的含义。

(1) 激活节点：最初选择的激活节点，它们始终处于激活状态，其状态函数值永远是1；

(2) 非激活节点：两类永远处于非激活状态的节点，其状态函数值总是0；

(3) 状态未知的节点：那些被激活状态节点尝试激活过的节点，该节点激活与否都已明确，那么其状态也是稳定的。

为了避免死锁，先随机设定其中一个节点的状态函数值，等待另一个节点的所有邻居节点尝试激活该节点后，该节点的状态稳定后，解决了死锁问题。

在 t 时刻，v_i 的状态函数值是 1 的父亲节点以 $parent_{i_t}^1$ 来标识；v_i 的状态函数值是 0 的父亲节点以 $parent_{i_t}^0$ 来标识。在 t+1 时刻，v_i 成为活跃节点的概率是 $\sum\limits_{v_j\in parent_{i_t}^1}\omega_{ji}$，用 p_1 表示；而 v_i 成为非活跃节点的概率是 $\sum\limits_{v_j\in parent_{i_t}^0}\omega_{ji}$，用 p_0 表示。在该时刻，v_i 的状态函数 $f_{t+1}(v_i)$ 可以用以下公式阐释：

$$f_{t+1}(v_i)=\begin{cases}1 \quad , p_1=\sum\limits_{v_j\in parent_{i_t}^1}\omega_{ji}\\ 0 \quad , p_0=\sum\limits_{v_j\in parent_{i_t}^0}\omega_{ji}\end{cases}$$

3.4 基于 PRP 模型的贪心算法

在PRP模型中，为寻找Top-k节点以解决影响最大化问题，需要找到种子集，一种有效的方法是每一步都根据贪心算法确定初始集合中的一个节点，直到找到 k 个节点为止。基于 PRP 的贪心算法将从空的初始集合开始，每次将使得影响范围函数获得最大边际效益的节点加入初始集合，而每次信息在节点之间的传播都是按照 PRP 模型进行，具体过程如算法 1 所示。

算法 1：基于 PRP 模型的贪心算法

输入：有向图 G，最终扩散集合大小 k

输出：集合大小为 k 的种子集 S

1：初始化：$S=\varnothing$，$R=10000$
2：**for** $i=1$ *to* k **do**
3： **for each vertex** $v \in V \setminus S$ **do**
4： $S_v = 0$
5： **for** $t=1$ *to* R **do**
6： $S_v += |RS(S \cup \{v\})| - |RS(S)|$
7： **end for**
8： $S_v = S_v / R$
9： **end for**
10： $S = S \cup \{\arg\max_{v \in V \setminus S}\{S_v\}\}$
11： **end for**

算法 1 中，首先定义 $S=\varnothing$；$RS(S)$ 表示集合 S 按照 PRP 模型激活节点，每一时刻处于活跃状态的节点集合；并定义 $S_v += |RS(S \cup \{v\})| - |RS(S)|$ 表示节点 v 的边际效益影响范围函数。从初始集合 $S=\varnothing$ 开始，每一步都选择使得当前影响范围函数获得最大边际效益的节点，选择策略如下：根据局部最优策略，对集合 $V \setminus S$ 中所有的节点，依次计算 $t=1, t=2, \cdots, t=R$ 时刻节点的边际效益，并对这些时刻的边际效益求均值，最后选择使边际效益均值最大的节点 u，形式化表示为 $u = \arg\max_{v \in V \setminus S}\{S_v\}$，将节点 u 并入集合 S 中，即 $S = S \cup \{u\}$。经过 k 步后，算法 1 选择出影响范围最大的 k 个节点。

3.5 实验评估与分析

本实验的目的是验证基于本文提出的 PRP 模型的效果比基于传统的线性阈

值模型、加权级联模型和独立级联模型的效果更好，影响力范围更大。

本小节通过实验来分析比较，PRP 模型与基于传统的线性阈值模型、加权级联模型和独立级联模型的影响效果与最终的影响范围。

3.5.1 实验数据集

本文基于 4 个真实的数据集进行实验，4 个数据集的具体信息如下：

数据集 1 是一个物理领域的合作者网络[32]，节点表示研究者，边表示研究者之间的合作关系，该数据集有 10748 个节点，53000 条边。

数据集 2 来自社群服务平台 Flickr[1]，数据集的实体有用户以及他们的关系，包含用户文件和关系文件，从该数据集抽取 11328 个节点，54870 条边。

数据集 3 来源于 Meme Tracker[44]，是一个在线新闻网络，节点表示新闻门户或新闻博客，边表示网站之间的影响关系，该数据集有 339936 个节点，有 1574596 条边。

数据集 4 是一个社会新闻分享和投票的网站[2]，数据集里包含有不同实体和这些实体之间的联系，从该数据集抽取 10536 个节点，52400 条边。

3.5.2 实验设计

本实验的基准比较模型是第 2 章相关工作中提到的线性阈值模型、独立级联模型和加权级联模型。在 4 个真实数据集上进行实验，设定目标集合大小 k 分别为 0，5，10，15，20，25，30，通过实验分析本文提出的 PageRank 传播模型（在下图中用网页排名传播模型表示）以及基准比较模型的影响范围。

3.5.3 实验结果分析

基于数据集 1 的实验结果见图 3.1。

[1] http://socialnetworks.mpi-sws.org/data-imc2007.html
[2] http://arnetminer.org/heterinf

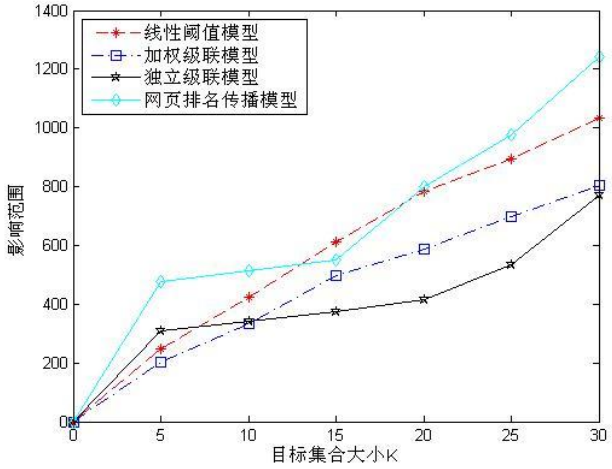

图 3.1　数据集 1 上不同 k 影响范围曲线

从图 3.1 可知，当 k 值为 15 时，PageRank 传播模型影响范围只比线性阈值模型影响范围略低，而比另外两个基准模型的影响范围高；当 k 取其它值时，PageRank 传播模型影响范围比所有基准比较模型的影响范围都高。

基于数据集 2 的实验结果见图 3.2。

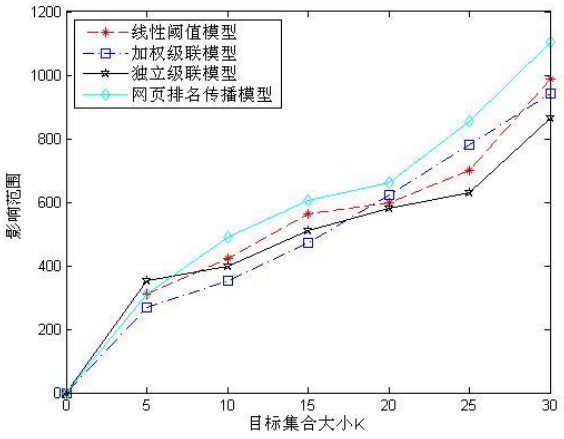

图 3. 2　数据集 2 上不同 k 影响范围曲线

从图 3.2 中可以观察出当 k 值等于 5 时，PageRank 传播模型影响范围比独立级联模型影响范围低，比加权级联模型影响范围高，与线性阈值模型影响范围不差上下，而当 k 取其它值时，PageRank 传播模型影响范围都高于基准比较模型的影响范围。

数据集 3 的实验结果见图 3.3。

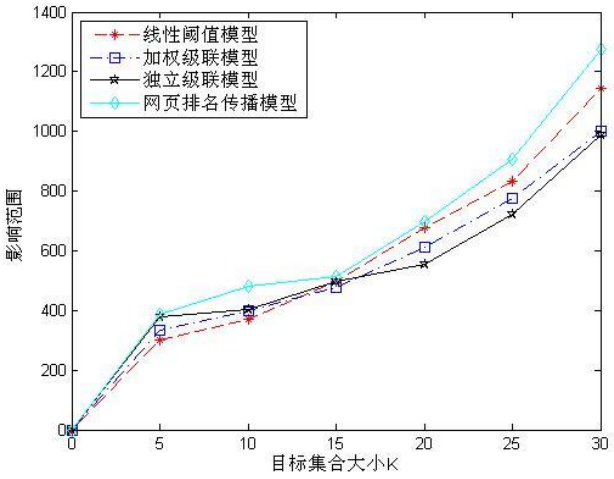

图 3.3　数据集 3 上不同 k 影响范围曲线

从图 3.3 中可以观察出当 k 值等于 5 时，PageRank 传播模型影响范围与独立级联模型影响范围相近，比其它两个基准模型影响范围高；当 k 值等于 15 时，PageRank 传播模型影响范围与所有基准模型的影响范围都相近；当 k 取其它值时，PageRank 传播模型影响范围都高于基准比较模型的影响范围。

基于数据集 4 的实验结果见图 3.4。

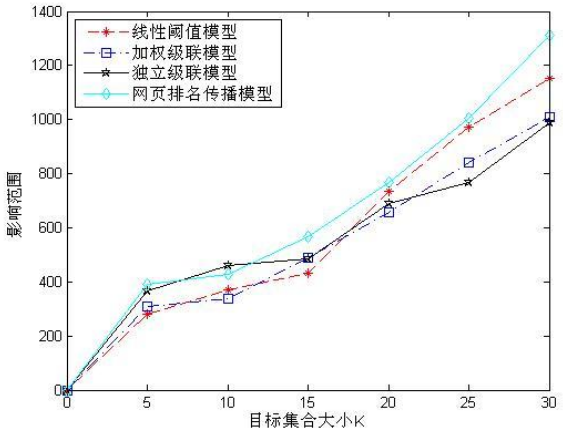

图 3.4　数据集 4 上不同 k 影响范围曲线

从图 3.4 中可以观察出当 k 值等于 10 时，PageRank 传播模型影响范围比独立级联模型影响范围略低，比其它两个基准模型影响范围都高，而当 k 取其它值时，PageRank 传播模型影响范围高于所有基准模型的影响范围。

通过在 4 个真实的数据集上进行实验可知，PageRank 传播模型解决影响最大化问题的效果比传统的线性阈值模型、加权级联模型和独立级联模型下的效果更好，影响力范围更大。

本章 PRP 模型摈弃了传统的线性阈值、加权级联和独立级联方法的共同假设，该假设认为社交网络里所有节点的不同父亲节点对该节点的影响强度相同，因此过于理想化。本章 PRP 模型按照不同父亲节点对该节点的重要性不同，区分出社交网络中节点的所有父亲节点对该节点的影响强度，使得该信息传播模型更具有实际意义，基于该模型的影响最大化问题更有应用价值。

3.6 本章小结

本章首先指出了目前的信息传播模型的前提假设的问题，由此分析了 PRP 模型的思想，详细阐述了 PageRank 算法的原理以及实际意义；其次基于 PageRank 算法，提出了融合社交网络中节点的相关性和节点的重要性的信息传播模型—PRP 模型，且分析了社交网络两大类永远处于非激活状态的节点以及节点的稳定状态的概念；最后设计了基于 PRP 模型的贪心算法，给出了实验评估与分析。

第 4 章 基于概率转移矩阵的影响最大化算法

解决社交网络的影响最大化问题往往采用贪心算法及其改进算法，但这些算法对于规模很大的社交网络其时间复杂度很高。针对该问题，本文基于概率转移矩阵的思想提出了一种扩展的线性阈值模型，在扩展的线性阈值模型基础上，设计了一种新的影响最大化算法：基于概率转移矩阵的影响最大化算法（PTMA 算法），实验结果验证了该方法比其他基本算法更有效，时间复杂度更低。

4.1 PTMA 算法思想

随着社交网络的出现及流行，社交网络影响力成为目前研究的热点。社交网络影响力[1]是指人们接受他人信息传播的过程。在该领域中，Domingos and Richardson 提出了社交网络影响最大化问题[23]，用有向图来表示社交网络。本文的目标是在图中找出最具有影响力的 k 个节点，使得最终社交网络中被影响的节点数目最多，信息传播范围最大。而信息在社交网络传播过程中都遵循一定的规则，称之为信息传播模型。

对于研究社交网络影响最大化问题来说，信息传播模型尤为重要与特殊，因为所有解决社交网络影响最大化问题的算法都是基于一个特定的信息传播模型基础之上的，所以信息传播模型能够决定解决社交网络影响最大化问题的算法的时间复杂度以及实用价值。两类最基本的信息传播模型是线性阈值模型和独立级联模型，本节重点分析线性阈值模型。

由线性阈值模型传播特性可以知道，t 时刻父亲节点的状态直接影响 t+1 时刻孩子节点的状态，而孩子节点是否被成功激活，取决于 t 时刻父亲节点是否活跃且父亲节点对其影响权重是否大于等于子节点阈值，该模型下影响力的传播具有很强的“时刻依赖性”，也就说每隔一段时间 t，那么就要观察社交网络里哪些是活跃节点，哪些是非活跃节点。这就导致了所有基于此模型的算法就要每隔一段时间 t，统计活跃节点的数目，这种计算模型浪费了大量的不必要的时间，特别对于大规模或特大规模社交网络，基于此模型的算法的效率都是不可接受的。目前存在很多基于线性阈值模型传播特性的改进的贪心算法和其他算法，这些算法的时间复杂度相当高。

因此，本文提出了一个更有效的扩展的线性阈值模型。该模型省去了每隔一段时间 t 就要统计活跃节点的数目的时间。传统线性阈值模型是该模型中时间间隔为 1 时的特殊情况。并且在扩展的线性阈值模型下，本文提出了基于概率转移矩阵的影响最大化算法（PTMA 算法）。PTMA 算法首先计算 t 时刻概率转移矩阵，然后利用贪心方法寻找 k 个最具影响力节点。下文将详细阐述扩展的线性阈值模型以及 PTMA 算法。

4.2 扩展的线性阈值模型

用有向图$G=(V,E)$表示社交网络，V表示节点集合，E表示边集合，$|V|=n$。

由传统的线性阈值模型传播特性可以知道，t时刻父节点的状态直接影响$t+1$时刻子节点的状态，而子节点是否被影响成功，取决于t时刻父节点是否活跃且父节点对其影响权重是否大于等于子节点阈值，因此该模型具有很强的“时刻依赖性”。针对该问题，本文提出了扩展的线性阈值模型，只需知道初始时刻节点状态及图中节点之间的权重，本文就能确定任意时刻某节点能否被成功影响。

在扩展的线性阈值模型中，初始时刻有向边(u,v)的权重ω_{uv}等于节点u的出度d_u之倒数，即$\omega_{uv}=1/d_u$，由此可得到初始时刻的概率矩阵$P_{n\times n}$，其中$P_{ij}=\omega_{ij}$。给定初始活跃节点集合S，传播过程按如下规则进行：

1. 对节点$v\in V$，从[0,1]区间随机产生一个阈值θ_v。
2. 在初始$t=0$时刻，初始概率矩阵P的每个元素$\omega_{uv}=1/d_u$，其中ω_{uv}表示节点u到v的权重。这里本文规定$\omega_{uu}=1$；若节点u到v没有边，则$\omega_{uv}=0$。
3. 传播T时刻，即$t=t+T$，初始活跃节点$u(u\in S)$以影响力$I^T(u\to v)$影响非活跃节点v，$I^T(u\to v)=\sum_{\beta=1}^{r}\omega_{\beta 1}^{u}\omega_{\beta 2}^{\beta 1}\cdots\omega_{\beta j}^{\beta i}\omega_{v}^{\beta j}$，其中$r$表示从节点$u$到节点$v$有$r$条路径，$\beta$表示第$\beta$条路径，$\omega_{\beta j}^{\beta i}$表示第$\beta$条路径上$i$节点到$j$节点之间的权重，用$P^T$表示初始矩阵$P$经过$T$次乘积变换后

的矩阵，那么 $I^T(u \to v)$ 就是矩阵 P^T 的 u 行 v 列元素。

4. 如果 T 时刻，非活跃节点 v 的所有影响力之和 $I^T(v)$ 大于等于 v 的阈值 θ_v，那么非活跃节点 v 就被成功影响，变成活跃节点，其中 $I^T(v) = \sum_{q \in S} I^T(q \to v)$。

5. 节点的状态只能由非活跃变成活跃，反之不能。

该模型是对线性阈值模型的扩展，当 T=1 时，该模型就是线性阈值模型。

4.3 PTMA 算法

传统贪心算法近似求解影响最大化问题的时间复杂度很高，尤其对于规模很大的社交网络。本文利用线性阈值模型的传播特性，并基于概率转移矩阵提出了扩展的线性阈值模型，并给出了解决影响最大化问题的框架及 PTMA 算法。

传统贪心算法如果不观察 t 时刻父节点状态，就无法确定 t+1 时刻子节点被影响的权重，而在扩展的线性阈值模型基础上，本文提出的 PTMA 算法将初始时刻节点之间的权重记录在一个初始概率矩阵中，通过矩阵乘积的方法直接得到 T 时刻节点之间的影响权重，不需要每时刻都计算所有非活跃节点的边际效益，大大提高了运行效率和时间复杂度。整个算法框架分为两个阶段：第一阶段：计算最终时刻的概率矩阵；第二阶段：寻找 k 个目标节点。下面给出算法框架伪代码，详见算法 2。

算法2：PTMA算法

输入：有向图$G=(V, E)$, $|V|=n$，初始传播集合大小k，初始概率矩阵P

输出：集合大小为k的种子集S

1：初始化：$S=\varnothing, T=100000$

2：$W = P$

3：$i=1, j=\log_2 T, u=1$

4: **while** $i \le j$ **do** /*把T个初始矩阵连乘的形式转换成一棵二叉树形式*/

5: $W = W \cdot W$ /*计算两个矩阵的乘积*/

6: $i = i+1$

7: **end while**

8: $m = 2^{\lfloor j \rfloor}$ /*已作连乘运算的矩阵个数*/

9: $W = W \cdot P^{T-m}$ /*计算得到最终的概率矩阵*/

10: **while** $u \le k$ **do** /*采用贪心方法选择节点*/

11: **for each vertex** $v \in V \setminus S$ **do** /* 计算$V \setminus S$ 中每个节点在T时刻的影响范围*/

12： **for each vertex** $v' \in V \setminus S \bigcup \{v\}$ **do** /* 判断$V \setminus S \bigcup \{v\}$ 中每个节点在T时刻是否活跃*/

13： $I^T(v') = \sum_{q \in S \bigcup \{v\}} I^T(q \to v')$ /*初始活跃节点集合对v' 的影响力之和*/

14： 随机产生节点v'阈值$\theta_{v'}, 0 \le \theta_{v'} \le 1$ /*线性阈值模型机制*/

15： **if** $I^T(v') \cdot \frac{1}{|S \cup \{v\}|} \ge \theta_{v'}$ **then** $f(v') = 1$

/* 如果v'受到的影响力不小于其阈值$\theta_{v'}$，那么该节点变为活跃状态*/

16： **else** $f(v') = 0$ /*否则保持非活跃状态*/

17： **end for**

18： $|RS(S \cup \{v\})| = \sum_{v' \in V \setminus S \cup \{v\}} f(v')$ /*v节点在T时刻的影响范围*/

19： **end for**

20： $S = S \cup \{\arg\max_{v \in V \setminus S} \{|RS(S \cup \{v\})|\}\}$ /*选择影响范围最大的结点并入初始活跃节点集合S中*/

21： $u = u + 1$ /*选择下一个节点*/

22：**end while**

23：输出集合 S

执行步骤 4～9 得到最终 T 时刻的概率矩阵，其中 4～7 步时间复杂度为$O(n^3 \cdot \log_2 T)$，第 8、9 两步计算当$\log_2 T$不能取整的情况，其时间复杂度为$O(Cn^3)$，常量C小于$\log_2 T$，所以步骤 4～9 的时间复杂度为$O(n^3 \cdot \log_2 T)$。步骤 12～17 通过比较概率矩阵中节点影响力权重与其阈值大小确定节点状态，该步时间复杂度为$O(n)$，然后执行步骤 18 计算每个非活跃节点的影响范围，因此循环 11～19 的时间复杂度为$O(n^2)$，最后执行步骤 22 可得到了一个目标节点，k 步就得到 k 个目标节点，因此循环 10～22 的时间复杂度为$O(k \cdot n^2)$，于是，PTMA 算法复杂度为$O(n^3 \cdot \log_2 T)$。

4.4 TGA 与 PTMA 算法时间复杂度分析

为了方便表示，本文把传统贪心算法称之为 TGA 算法，为对 TGA 与 PTMA 算法的时间复杂度进行分析与对比，首先需要对 TGA 算法进行分析，TGA 算法如算法 3 所示。

首先来看 TGA 算法，下面具体分析 TGA 算法的时间复杂度。

算法3：TGA算法

输入：有向图 G，最终扩散集合大小 k

输出：集合大小为 k 的种子集 S

1：初始化：$S=\varnothing$，R=10000

2：**for** $i=1$ *to* k **do**

3： **for each vertex** $v \in V \setminus S$ **do**

4： $s_v = 0$

5： **for** $t=1$ *to* R **do**

6： 按边搜索整个网络，观察并记录活跃节点个数

7： $s_v += |RS(S \cup \{v\})| - |RS(S)|$

8： **end for**

9： $s_v = s_v / R$

10： **end for**

11： $S = S \cup \{\arg\max_{v \in V \setminus S}\{S_v\}\}$

12： **end for**

对于算法 3，每选一个目标节点，需要执行算法 1 中的步骤 3～10，其中步骤 5～8 为计算每一时刻所有非活跃节点的边际效益，而每次计算需要沿边搜索整个网络节点并观察记录活跃节点个数，其运算复杂度为 $O(T \cdot n^2)$ ，因此步骤 3～10 运算复杂度为 $O(T \cdot n^3)$ ，选取 k 个最具影响力节点需执行步骤 2～12，于是 TGA 算法的运算复杂度为 $O(n^3 \cdot k \cdot T)$。

接下来分析比较 TGA 算法与 PTMA 算法的时间复杂度。

TGA 算法考虑了整个传播过程，每一时刻都要计算所有未被激活节点的边际效益，而计算时，几乎都要遍历整个网络观察被影响的节点，时间复杂度非常高。

PTMA 算法是基于概率转移矩阵的一种影响最大化算法，该算法分为两个阶段，第一阶段计算 T 时刻网络节点的概率矩阵，第二阶段贪心法选取 k 个目标节点。第一阶段，通过初始概率矩阵的 T 次乘积得到 T 时刻概率矩阵，而本文把 T 个初始矩阵连乘的形式转换成一棵二叉树，每步只需计算两个矩阵的乘积，其运算复杂度为 $O(n^3)$ ，总步骤为树高 $\log_2 T$ ，于是第一阶段时间复杂度为 $O(n^3 \cdot \log_2 T)$ 。第一阶段结束后，就得到了 T 时刻网络所有节点之间相互影响的权重，以便于下一阶段选取目标节点。第二阶段贪心法选取目标节点都是选取影响力最大的节点，但已计算出 T 时刻的概率矩阵，就不需要每个时刻都遍历整个网络，这样就比 TGA 算法大大降低了时间复杂度。

以下用表 4.4.1 来对比 TGA 与 PTMA 算法的时间复杂度。

表 **4.4.1　PTMA** 和 **TGA** 算法时间复杂比较

算法	PTMA 算法	TGA 算法
时间复杂度	$O(n^3 \cdot \log_2 T)$	$O(n^3 \cdot k \cdot T)$

4.5 实验评估与分析

本实验的目的是为了证明在大规模社交网络中，本文提出的基于扩展的线性阈值模型的 PTMA 算法效果更好，时间复杂度更低。

本小节以 TGA 算法与 HPG 算法作为基准比对算法，从两个角度分析对比本文提出的 PTMA 算法与这些基准算法的效果与效率。本小节的实验数据使用 3.5.1 描述的 4 个数据集，首先在数据集 1 和数据集 2 上分析对比 PTMA 算法与这些基准算法的影响范围；然后在另外两个真实数据集上分析对比 PTMA 算法与这些基准算法的运行时间。

4.5.1 实验设计

本文在扩展线性阈值模型基础上提出了 PTMA 算法，与第 4.4 小节中描述的 TGA 算法和文献[45]中提出的 HPG 算法比较影响范围和运行时间。本文在 3.5.1 描述的 4 个真实数据集上进行实验。设定目标集合大小 k 分别为 0，5，10，15，20，25，30，在数据集 1，2 上，通过实验分析 PTMA，TGA 和 HPG 的影响范围，在数据集 3，4 上分析 PTMA 算法与对比算法的运行时间。

4.5.2 影响效果分析

数据集 1 的实验结果见图 4.1。

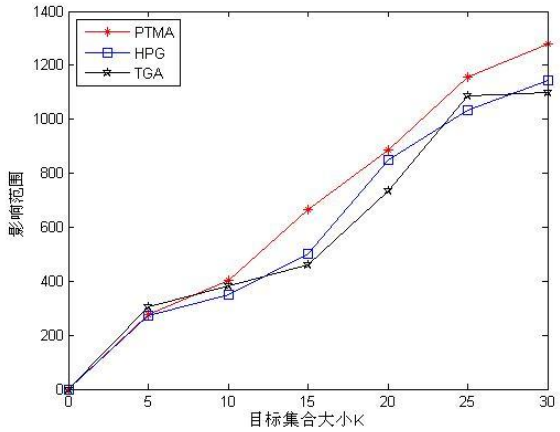

图 4.1　数据集 1 上不同 *k* 值影响范围曲线

从图 4.1 中可以观察出当 k 值为 0，5，10 时，这三个算法的影响范围大致相当；当 k 逐渐增大时，PTMA 算法在影响范围上比所有基准比较算法的影响范围都高并且呈线性增长。

数据集 2 的实验结果见图 4.2。

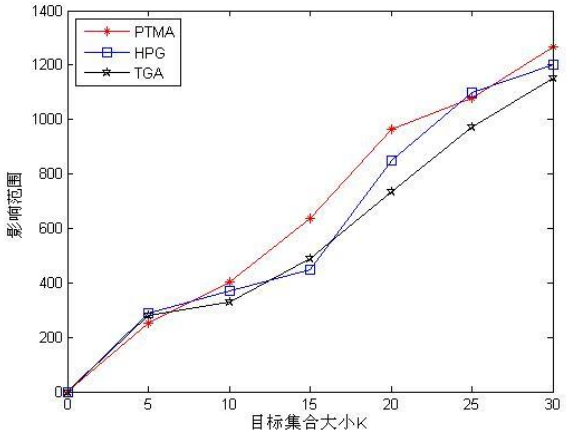

图 4.2　数据集 2 上不同 *k* 值影响范围曲线

从图 4.2 中可以观察出 k 为 5 时，三个算法的影响范围大致相同；k 为 25 时，PTMA 算法与 HPG 算法影响范围大致相同，都比 TGA 算法影响范围略高；当 k 取其他数值，PTMA 算法比其他算法影响范围都高。

4.5.3 时间代价分析

数据集 3 的实验结果见图 4.3。

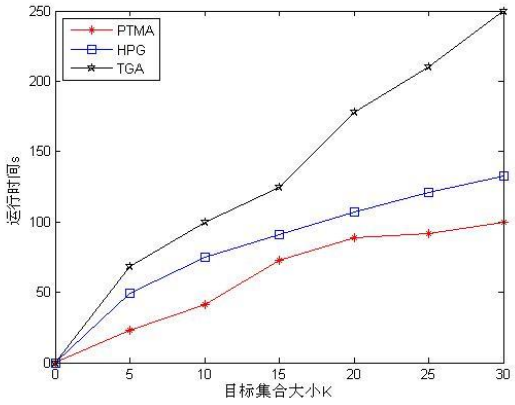

图 4.3 数据集 3 上不同 *k* 值运行时间曲线

从图 4.3 中可以观察出 PTMA 算法运行时间最少，依次是 HPG 算法，和 TGA 算法。当 k 为 30 时，三个算法运行时间分别是 80s，120s 和 250s。

数据集 4 的实验结果见图 4.4。

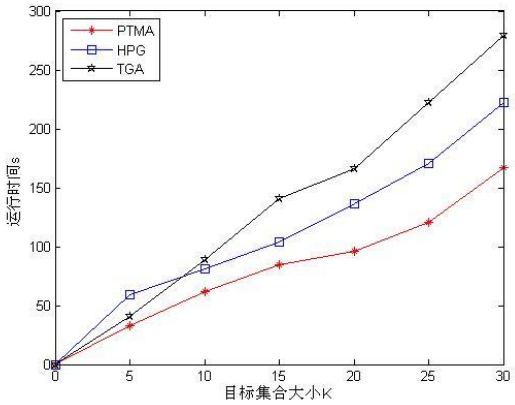

图 4.4 数据集 4 上不同 *k* 值运行时间曲线

从图 4.4 中可以观察出 PTMA 算法运行时间最少。当 k 为 5 时，TGA 算法比 HPG 算法运行时间少，当 k 取其他数值时，HPG 算法比 TGA 算法运行时间少。当 k 为 30 时，三个算法运行时间分别是 160s，220s 和 280s。

由以上实验分析得出，本章提出的 PTMA 算法比 HPG 算法和 TGA 算法影响力范围更大。由于传统的模型每隔一段时间都需要花费大量的时间去统计活跃节点的数目，而 PTMA 算法省去了每隔一段时间 t 统计活跃节点的数目的时间，因此，PTMA 算法的运行时间更少，且随着数据集和 k 值的逐渐增大，PTMA 的运行时间基本呈现线性增长，说明本文提出的算法可以适用于大规模社交网络。

4.6 本章小结

本章首先阐述了信息传播模型的重要性，然后分析了基于传统的线性阈值模型的贪心算法以及改进算法的缺陷，这些算法在大规模社交网络中的时间复杂度很高。针对该问题，本章基于概率转移矩阵思想提出了扩展的线性阈值模型，在扩展的线性阈值模型基础上提出了基于概率转移矩阵的 PTMA 算法，该算法的时间复杂度比已有的相关算法降低了一个数量级。

第 5 章 总结与展望

本章对本文的全部工作进行总结，并分析本文取得的成果。然后指出本文工作的不足之处，并对未来的工作进行了展望。

5.1 总结

社交网络是指由个体及个体之间的关系所组成的一个复杂网络，这种复杂的网络结构对信息的传播和扩散起着至关重要的作用。社交网络中的影响力是指用户受别人影响后，感情、观点或行为发生变化的过程，该研究最早出现在心理学和社会学研究领域。随着 Twitter、Facebook、Flickr 等社交网络的出现及迅速发展，社交网络的影响力也逐渐体现在这些主流网络里，并成为计算机领域中一个重要的研究热点问题。当前国内外主要研究工作大都集中在社交网络影响力最大化问题上，本文针对这些研究中存在的问题，对社交网络影响力最大化的模型与算法作了一些进一步的研究。本文主要工作总结如下：

（1）信息传播模型对研究社交网络影响最大化问题尤其重要，涉及到解决社交网络影响最大化问题的算法往往都是基于某个信息传播模型。信息传播模型决定着社交网络影响最大化问题的算法效率以及实用价值，所以本文首先着重分析了目前最重要的一些信息传播模型，以及影响最大化问题在不同模型下的相应定义，并简要总结影响最大化问题的一些解决方法并分析其利弊，为本文的后续工作打下理论基础。

（2）基本的线性阈值模型，独立级联模型和加权级联模型都基于一个共同假设，该假设认为社交网络里所有节点的父亲节点对该节点的影响强度都是相同的，但实际的社交网络并非如此，可见该假设过于理想化。只有信息传播模型更适合实际情况，才能使得基于该模型的算法具有应用价值。因此，本文考虑到社交网络中节点的相关性和节点的重要性两个衡量其影响力的重要指标，提出了一种融合节点重要性与相关性的 PRP 模型，解决了假设过于理想化的问题。

（3）在规模很大的社交网络中，解决影响最大化问题时，往往采用传统的贪心算法以及其改进算法，这些算法的时间复杂度都很高。针对该问题，本文基于概率转移矩阵的思想提出了一种扩展的线性阈值模型，在扩展的线性阈值模型基础上，给出了一种新的影响最大化算法（PTMA 算法），该算法节省了每个时

间间隔都要统计活跃节点数目的时间，因此，比其他基本算法更有效。

（4）本文在四个真实数据集上进行实验，由实验分析可知，本文提出的 PRP 模型比传统的线性阈值模型、加权级联模型和独立级联模型下的影响力范围更大。通过实验也验证了本文提出的 PTMA 算法比 HPG 算法和 TGA 算法的影响力范围广且运行时间更少。随着数据集和 k 值的逐渐增大，PTMA 的运行时间基本呈现线性增长，说明本文提出的 PTMA 算法适用于大规模社交网络。

5.2 未来工作展望

虽然本文提出的 PRP 模型在影响力范围上有提升，但依然还存在值得改进的地方，例如 PRP 模型只适用于有向网络中信息的传播，将来进一步的工作将研究适用于有向网络以及无向网络的通用信息传播模型。

本文提出的 PTMA 算法比 HPG 算法和 TGA 算法影响力范围更大，运行时间更少，且本文提出的算法可以适用于大规模数据，但是该算法只是适用于分析静态的网络结构，将来进一步的工作将研究适用于分析动态网络结构的信息传播模型和算法。

参考文献

[1] D. A. Easley and J. M. Kleinberg. Networks, Crowds, and Markets - Reasoning About a Highly Connected World[M]. Cambridge University Press, 2010.

[2] A. Anagnostopoulos, R. Kumar and M. Mahdian. Influence and correlation in social networks[C]. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008:7–15.

[3] D. J. Crandall, D. Cosley, D. P. Huttenlocher, J. M. Kleinberg and S. Suri. Feedback effects between similarity and social influence in online communities[C]. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008: 160–168.

[4] A. Goyal, F. Bonchi and L. V. S. Lakshmanan. Learning influence probabilities in social networks[C]. In Proceedings of the Third ACM international Conference on Web Search and Data Mining, 2010:241–250.

[5] G. Peeters and J. Czapinski. Positive-negative asymmetry in evaluations: The distinction between affective and informational negativity effects[J]. European Review of Social Psychology, 1990, 1:33–60.

[6] S.E.Taylor. Asymmetrical effects of positive and negative events: The mobilization-minimization hypothesis[J]. Psychological Bulletin, 1991, 110(1):67–85.

[7] P. Rozin and E.B. Royzman. Negativity bias, negativity dominance, and contagion[J]. Personality and Social Psychology Review, 2001, 5(4):296–320.

[8] R. F. Baumeister and E. Bratslavsky. Bad is stronger than good[J]. Review of General Psychology, 2001, 5(4):323– 370.

[9] W. Chen, A. Collins and R. Cummings. Influence Maximization in Social Networks When Negative Opinions May Emerge and Propagate[C]. In Proceedings of the Eleventh SIAM International Conference on Data Mining, 2011:379– 390.

[10] K.K. Cai, S.H. Bao, Z. Yang and J. Tang. OOLAM: an opinion oriented link

analysis model for influence persona discovery[C]. In Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, 2011:645–654.

[11] M. Kimura, K. Saito, and H. Motoda. Blocking links to minimize contamination spread in a social network[J]. ACM Transactions on Knowledge Discovery from Data, 2009, 3(2): 1–23.

[12] J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks[C]. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009:807–816.

[13] L. Liu, J. Tang, J. Han, M. Jiang, and S. Yang. Mining topic-level influence in heterogeneous networks[C]. In Proceedings of ACM 19th Conference on Information and Knowledge Management, 2010:199–208.

[14] G. Huillier, H. Alvarez, S. A. Rios and F. Aguilera. Topic-based social network analysis for virtual communities of interests in the dark web[J]. ACM Special Interest Group on Knowledge Discovery and Data Mining Explorations, 2010, 12(2): 66–73.

[15] S.A. Myers and J. Leskovec. On the Convexity of Latent Social Network Inference[C]. In Advances in Neural Information Processing Systems, 2010:1741–1749.

[16] J. Leskovec and A. Krause. Inferring networks of diffusion and influence[C]. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2010:1019–1028.

[17] J. Yang and J. Leskovec. Modeling information diffusion in implicit networks[C]. In Proceedings of the 2010 IEEE International Conference on Data Mining, 2010:599–608.

[18] D. Kempe, J. M. Kleinberg and V. Tardos. Maximizing the spread of influence through a social network[C]. In Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining， 2003: 137–146.

[19] D. Kempe , J. Kleinberg and É. Tardos. Extracting influential nodes on a social network for information diffusion[J]. Data Mining and Knowledge Discovery,

2010, 20 (1):70–97.

[20] W. Chen, Y. Wang and S. Yang. Efficient influence maximization in social networks[C]. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining，2009:199–208.

[21] M. Mathioudakis, F. Bonchi, C. Castillo, A. Gionis, and A. Ukkonen. Sparsification of influence networks[C]. Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2011:529–537.

[22] H. Ma, H. Yang, M. R. Lyu, and I. King. Mining social networks using heat diffusion processes for marketing candidates selection[C]. In Proceedings of ACM 17th Conference on Information and Knowledge Management. 2008: 233–242.

[23] P. Domingos and M. Richardson. Mining the network value of customers[C]. In Proceedings of the 7th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2001:57–66.

[24] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing[C]. In Proceedings of the 8th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2002:61–70.

[25] M. Kimura and K. Saito. Tractable models for information diffusion in social networks[C]. In Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases, 2006:259–271.

[26] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. S. Glance. Cost-effective outbreak detection in networks[C]. In Proceedings of the 13th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2007: 420–429.

[27] R. Narayanam and Y. Narahari. Determining the top-k nodes in social networks using the shapley value[C]. In Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems, 2008:1509–1512.

[28] W. Chen, C. Wang and Y. Wang. Scalable influence maximization for prevalent viral marketing in large scale social networks[C]. In Proceedings of the 16th

ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2010:1029–1038.

[29] W. Chen, Y. Wang and S. Yang. Efficient influence maximization in social networks[C]. In Proceedings of the 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2009:199–208.

[30] Y. Wang, G. Cong, G. Song and K. Xie. Community-based greedy algorithm for mining top-k influential nodes in mobile social networks[C]. In Proceedings of the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2010: 1039-1048.

[31] W. Chen, Y. Yuan and L. Zhang. Scalable influence maximization in social networks under the linear threshold model[C]. In Proceedings of the 10th IEEE International Conference on Data Mining, 2010: 88–97.

[32] M. Kimura, K. Saito, R. Nakano and H. Motoda. Extracting Influential Nodes on a Social Network for Information Diffusion[J]. Data Mining. Knowledge Discovery, 2010, 20(1):70–97.

[33] D. Kempe, J. M. Kleinberg and E. Tardos. Influential nodes in a diffusion model for social networks[C]. In Proceedings of the 32Nd International Conference on Automata, Languages and Programming, 2005: 1127–1138.

[34] R. Shultz and F. Rivest. Using knowledge to speed learning: A comparison of knowledge-based cascade-correlation and multi-task learning[C]. In Proceedings of the Seventeenth International Conference on Machine Learning, 2000:871–878.

[35] M. Fazli, M. Ghodsi, J. Habibi, P. J. Khalilabadi and V. Mirrokni. On the Non-progressive Spread of Influence through Social Networks[J]. Lecture Notes in Computer Science, 2012, 7256:315–326.

[36] E. Mossel and S. Roch. On the submodularity of influence in social networks[C]. In 39th Annual ACM Symposium on Theory of Computing, 2007:128–134.

[37] R.Aviv, Z.Erlich and G.Ravid. Network Analasis Of Knowledge Construction in Asynchronous Learning Networks[J]. JALN , 2003,7(3):1–23.

[38] E. Even-Dar and A. Shapira. A note on maximizing the spread of influence in

social networks[C]. In Proceedings of WINE, 2007:281–286.

[39] L. Page, S. Brin, R. Motwani and T. Winograd. The pagerank citation ranking: Bringing order to the web[EB/OL]. http://ilpubs.stanford.edu:8090/422/, 1999.

[40] 金迪, 马衍民. PageRank 算法的分析及实现[J]. 计算机应用, 2009, 18(1001): 118–118.

[41] T. H. Haveliwala. Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search[J]. IEEE Trans. on Knowl. and Data Eng, 2003, 15 (4): 784–796.

[42] L. Page and S. Brin. The Anatomy of a Large-scale Hypertextual Web Search Engine[C]. In Proceedings of the Seventh International Conference on World Wide Web, 1998: 107–117.

[43] A. Arasu and J. Cho and H. Garcia-Molina. Searching the Web[J]. ACM Trans. Internet Technol, 2001, 1(1): 2–43.

[44] J. Leskovec, L. Backstrom and J. M. Kleinberg. Meme-tracking and the dynamics of the news cycle[C]. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009:497-506.

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

致 谢

回首将近三年的硕士生涯，我在学业、科研及生活各个方面得到了很多老师、同学、家人、朋友的关心帮助与支持。在此毕业之际，向他们致以最诚挚的感谢。

首先，我要感谢我的导师张佩云副教授对我的悉心栽培与谆谆教导，张老师在科研上给了我很多有益的建议以及具体而耐心的指导，才得以让我在研究生阶段顺利完成小论文发表等任务，并顺利完成毕业论文的撰写等工作。此外，张老师关心照顾我的生活，并给了我很多帮助与指导，让我对自己的人生有了新的认识与反思，并激励我努力抒写属于自己的人生篇章。在此毕业季，我怀着深深的敬意与感动，再次向张老师表示我万分感谢与感激。

其次，感谢曾经教过我的各位老师，让我在知识与智慧中成长；感谢许勇老师、罗永龙老师、王杨老师、章昭辉老师、李汪根老师、左开中老师、方群老师、陈付龙老师、孙道清老师、齐学梅老师等在我学业上的指导与帮助，正是因为有了他们的严谨、高尚、追求学术理想的治学态度，我才能在这几年的学习过程中不断汲取知识、提升学术水平。

再次，我要感谢我的室友、研究生同学，在生活学习工作上给了我很多有益的建议，在我落寞无助时给予我真诚的关心与温暖，让我感觉她们就像家人一样亲切。这份永远不能忘怀的情谊，让我对校园生活以深深的眷恋，在此我要向她们表示最真诚的谢意。

再次，我要感谢我的爱人刘威威同学，在学习上给予我的宝贵建议，在生活中对我的宽容与理解，让我的研究生生活变得充实而有意义。

最后，我要感谢我的父母和亲人，二十几年对我无微不至的照顾以及精神上物质上的支持，让我一想到他们就不再惧怕这个世界。从呱呱坠地弱小的我到今天长大成人，父母对我倾注了无数心血，培养我上学至硕士毕业，让我有机会走出小村庄见识外面的世界，一切都是父母给我的，我要用整个人生来感谢家人对我的爱，对我的付出。

附录：攻读学位期间参与的科研项目与公开发表的论文

参与的主要科研项目如下：

1. 安徽师范大学“千人培养计划”联合培养研究生项目：社交网络中影响最大化传播模型与算法研究。(2012-2014)，主持。
2. 国家自然科学基金青年基金(61201252)：社会网络环境下可信服务组合动态协同模型与算法研究。(2013-2015)，参与。
3. 安徽省自然科学基金(1308085MF100)：个性化需求驱动下基于社会网络的服务组合动态协同模型与算法研究。(2013-2015) ，参与。
4. 安徽省高校省级自然科学研究重点项目(KJ2011A118)：基于个性化需求的可信服务组合研究。(2011-2012) ，参与。

已发表和接收的论文：

1. **宫秀文**, 张佩云. 基于 PageRank 的社交网络影响最大化传播模型与算法研究[J]. 计算机科学, 2013, 40(Z6):136-140.
2. 张佩云, **宫秀文**. 基于概率转移矩阵的社会网络影响最大化算法[J], 计算机工程, 2013, 39(11): 41-45 转 51.
3. 张佩云, **宫秀文**, 谢荣见. 农业信息资源共享与信息服务系统构建研究[J], 计算机技术与发展, 2013, 23(11): 157-160.
4. 张佩云, 黄波, **宫秀文**. 一种基于社会网络的可信服务最大覆盖方法[J], 计算机工程与科学, 2013, 35(10): 36-43.
5. 张佩云, 陈恩红, 谢荣见, **宫秀文**, 黄波. 基于元数据与领域概念树的文本相似度计算[J], 系统工程与电子技术, 2014, 37 (3): 585-591.

已完成和再投的论文：

1. Weiwei Liu, Zhi-Hong Deng, **Xiuwen Gong**, Xiaoran Xu and He Liu. Mining Top K Spread Sources for A Specific Topic and A Given Node[C], submitted to CIKM’14.
2. Weiwei Liu, Zhi-Hong Deng, **Xiuwen Gong**. Whether and When A Topic Will Be Prevalent in Near Future[J], submitted to IEEE Transactions on Parallel and Distributed Systems.