

山东大学

毕业论文(设计)

论文（设计）题目：

基于目标话题的社会网络影响力最大化研究

姓 名 张梦迪

学 号 20100301277

学 院 山东大学软件学院

专 业 软件工程

年 级 2010 级

指导教师 马军

2014 年 5 月 21 日

山东大学毕业设计（论文）成绩评定表

学院：软件学院

专业：软件工程专业

年级：2010 级

| | | | | | |
|----------|-----------------------------|-------------------|-----|----------|--|
| 学号 | 201000301277 | 姓名 | 张梦迪 | 设计（论文）成绩 | |
| 设计（论文）题目 | | 基于目标话题的社会网络影响力最大化 | | | |
| 指导教师评语 | | | | | |
| | 评定成绩：_____ 签名：_____ 年 月 日 | | | | |
| 评阅人评语 | | | | | |
| | 评定成绩：_____ 签名：_____ 年 月 日 | | | | |
| 答辩小组评语 | | | | | |
| | 答辩成绩：_____ 组长签名：_____ 年 月 日 | | | | |

注：设计（论文）成绩=指导教师评定成绩（30%）+评阅人评定成绩（30%）
+答辩成绩（40%）

摘要

病毒式营销是一种非常有效的市场营销策略，通过个人社交圈的朋友、家人或同事进行社会影响力传播的口碑效应。在这个背景下，影响力最大化问题已经成为社会网络领域的研究热点，近年来已经有大量有关社会网络影响力最大化算法各方面的研究。影响力最大化问题就是找到社会网络中一小部分结点子集（种子结点），可以最大限度地发挥影响力的传播。

基于话题的影响力最大化问题是影响力建模在内容上的扩展和完善，是加入目标市场因素后的病毒式营销策略分析，因而基于话题的影响力最大化算法可直接由经典影响力最大化方法扩展而来。影响力是通过影响力传播模型在网络中传播的，大多基于两个最基本的影响力传播模型，即线性阈值模型(Linear Threshold model,LT 模型)和独立级联模型(Independent Cascade model,IC 模型)，以及他们的扩展。影响力最大化方法主要分为贪心策略和启发式方法两种。本文通过调整经典影响力最大化算法，加入话题因素后得到了基于话题的 TGG、TNG、TDD 算法；并根据网络中影响力基于话题传播的局部性提出了基于种子选取的方法 TSRS。

最后基于 ArnetMiner 平台上的学术论文引用网络数据集和 Wikipedia 上的电影合作网络数据集进行了实验，并从效果和效率两个方面分析比较 TGG、TNG、TDD、TSRS 四种算法。实验结果表明：在基于话题的影响力最大化问题上，从传播效果上来说 Greedy 类的 TGG 算法表现最好，但时间复杂度过大，而本文提出的启发式类的 TSRS 算法仍可获得的较好传播效果，并且计算速度最快。TSRS 算法是一种有效的基于话题的影响力最大化问题算法。

关键词：目标话题、影响力最大化、社会网络、信息传播模型

ABSTRACT

Viral marketing, a very effective marketing strategy of conducting product promotions through word-of-mouth, which spread social influence among individuals' cycles of friends, families, or co-workers. Motivated by this background, influence maximization problem has become a focus in academic, the research community has recently studied the algorithmic aspects of maximizing influence in social networks. Influence maximization is the problem of finding a small subset of nodes (seed nodes) in a social network that could maximize the spread of influence. Finding out the most influential classic papers and research leader in the academic network is important to promote the scientific research cooperation and guide the scientific research work.

Topic aware influence maximization is a extended problem of influence maximization from a content perspective, a more consummate viral marketing problem with targeted consumer and the algorithm in influence maximization problem can be used in this new research. Influence spread in the network through a influence spread model. Most of these works are based on the two basic influence spread models, namely linear threshold model (LT model) and independent cascade model(IC model), and their extensions. The algorithm of influence maximization problem has two main categories: Greedy and Heuristics. This paper by adding topic factor into influence model to adjust the classic influence maximization method, have proposed three topic-aware method:TGG, TNG and TDD algorithm.What's more, based on the fact of locality of influence spreading, this paper also has proposed another new algorithm:TSRS, the topic-aware seed region separation algorithm. In the end, some carefully designed experiments are executed and have show that TSRS out perform the TDD algorithm(the state of art heuristic method in influence maximization) in spread result and far away out perform the TGG algorithm in time efficiency.

Key words: topic-aware, influence maximization, information propagation, heuristics

目录

| | |
|--|----|
| 摘要 | 3 |
| ABSTRACT..... | 4 |
| 第一章 绪论..... | 7 |
| 1.1 研究目的和意义 | 7 |
| 1.2 研究现状及主要发展趋势 | 8 |
| 1.3 本论文的主要内容..... | 9 |
| 第二章 社会网络..... | 10 |
| 2.1 概述..... | 10 |
| 2.2 图模型表示 | 11 |
| 2.3 网络元素及重要性测度..... | 12 |
| 2.3.1 结点的重要性..... | 12 |
| 2.3.2 联系的强度 | 14 |
| 2.4 社会网络统计特性..... | 15 |
| 2.4.1 幂律分布 | 15 |
| 2.4.2 六度分隔理论..... | 16 |
| 2.4.3 无尺度分布 | 16 |
| 2.4.4 小世界效应 | 17 |
| 第三章 信息扩散..... | 18 |
| 3.1 概述..... | 18 |
| 3.2 信息扩散..... | 18 |
| 3.3 扩散模型..... | 19 |
| 3.3.1 渐进模型 | 19 |
| 3.3.2 非渐进模型 | 21 |
| 第四章 社会影响力分析 | 22 |
| 4.1 概述..... | 22 |
| 4.2 影响力..... | 22 |
| 4.3 基于影响力的传播模型..... | 22 |
| 4.4 影响力最大化问题..... | 22 |
| 4.4.1 问题描述 | 23 |
| 4.4.2 问题分析 | 23 |
| 4.4.3 问题解决 | 24 |
| 第五章 基于话题的影响力最大化..... | 29 |
| 5.1 概述..... | 29 |
| 5.2 问题描述..... | 29 |
| 5.3 问题分析..... | 30 |
| 5.4 基于话题的影响力最大化方法 | 30 |
| 5.4.1 经典影响力最大化方法套用..... | 30 |
| 5.4.2 一种新方法:Topic-aware Seed Region Split..... | 32 |
| 第六章 实验..... | 33 |
| 6.1 概述..... | 33 |
| 6.2 数据集选取 | 33 |

| | |
|-------------------|----|
| 6.3 数据集处理 | 34 |
| 6.4 模型及参数设置 | 34 |
| 6.5 实验设计 | 35 |
| 6.6 结果分析 | 35 |
| 第七章 总结和展望 | 38 |
| 致谢 | 39 |
| 参考文献 | 40 |
| 附录 1 外文文献原文 | 43 |
| 附录 2 外文文献译文 | 49 |

第一章 绪论

1.1 研究目的和意义

社交应用和媒体在近十年内得到飞速发展，一大批新兴的 Web 和 Internet 信息交互形式（如 Interaction Dynamics）不断涌现：电子邮件（如 Gmail、163），即时信息（如 MSN、QQ），社交网站（如 Facebook、人人），分享网站（如 Youtube、Flickr），博客（如 WordPress、网易博客、点点），wiki（如 Wikipedia、百度百科），微博（如 Twitter、新浪微博）等等。

在飞鸽传书、烽火狼烟的年代，信息的传播相对简单：点对点或一对多的消息传播，信息生产者和消费者角色分离，传播范围和信息载量都是有限的。而当今信息时代 Internet 和 Web 的发展将人和信息更加方便并亲密的交织在一起，用户在其中同时扮演信息的生产者和消费者的角色，网络结构更加复杂、结点个数更加巨大、信息数据更加海量：根据[10]中数据显示，Facebook 用户量达 12.6 亿、月在线时间达 7000 亿分钟，人人用户量达 2.8 亿，新浪微博用户量达 5.6 亿。

各式社交媒体和网络，连接了人们的真实物理世界和虚拟信息空间，形成了形态各异的社会网络(Social networks)。内容丰富的信息在结构复杂的网络中，始发于起始用户，经传播用户扩散，最后又被终止用户接收，蕴涵着一定的传播机制。近来，基于社会网络的病毒营销、个人推荐、专家发掘等应用的出现，更是促进了对影响力驱使的社会网络中消息扩散(Influence-driven Propagations in Social Networks)机制的研究。

根据社会学理论支持和数据分析表明，“影响力”是社会网络中信息传播的主要驱动力[1]。社会影响力建模是社会网络核心研究之一，而影响力最大化问题(Influence Maximization) 是解决如何挖掘社会网络中关键用户、如何使某信息在某社会网络中实现最大传播等问题的关键。

然而，传统的影响力最大化问题往往忽略了用户结点和信息本身所具有的话题（或称属性、标签）分布。基于目标话题的社会网络信息传播建模适用于这样的场景：某商业组织希望利用 word-of-mouth 效应[3]在某社会网络中对特定用户群最大化推行自己的某类产品。显然,这是一类基于目标话题的影响力最大化问题，本文将在次研究其相关内容。

1.2 研究现状及主要发展趋势

“影响力最大化”作为算法问题最早由 Domingos and Richardson 在[2,3]中提出。其文中使用概率模型（马尔科夫随机场）对网络中的级联式信息扩散现象进行描述，提出启发式最优化求解算法。

而后 Kempe, Kleinberg, and Tardos[4]首先将其形式化描述为离散最优化问题(Discrete Optimization Problem), 证明在常用随机级联模型(Stochastic Cascade Model)类的信息扩散模型 IC、LT 模型下求解影响力最大化问题最优解是一个 NP-hard 问题，然后给出了有证明保证(Provable Guarantee)的可求得 63%最优的贪婪算法。但是 Kempe 所提出的近似算法十分耗时,在其论文的试验部分处理 15000 结点网络需要几天的时间,之后的研究主要关注于最优化近似算法的效率提高上。

Leskovec 等人提出的 KKT 算法的改进方法 CEFL(Cost-Effective Lazy Forward)方法[5] 一种新的选使影响力最大的择初始结点优化方法，CELF 算法利用了 IC 模型的子模性(submodularity)，子模函数是一个单调递减的函数，网络中加入一个结点到集合中得到的边际收益，要大于等于加入同一个结点到集合的父集所得到的边际收益。利用子模性大大减少了计算结点影响力传播效果的次数，减少选择初始有影响力结点的计算量。实验表明，CELF 算法结果接近贪心策略近似算法，但算法运算时间要快 700 倍。在 CEFL 的基础上，Amit Goyal 等人[6]利用子模性进一步优化了 CELF 算法，又提出了更高效的 CEFL++算法，算法效率提高了 35-55%。

W. Chen 等在[7]，主张启发式(heuristics)算法在影响最大化问题研究中的更优性，提出 Degree-discount 启发式算法，是目前基于理论扩散模型影响力最大化问题最高效的方法。

以上都是传统的影响力最大化问题的研究，而关于话题相关的影响力的研究内容相对贫乏。唐杰等人[9]研究了用户之间基于话题的影响力分析，提出了 TAP 分布式学习体系来解决高效的专家发掘问题，但并没有关注影响力最大化问题。Liu 等人[8]提出了概率模型来描述话题分布和影响力之间结合，通过使用 Gibbs 样本算法来计算用户的话题分布和用户之间的影响力。Nicola Barbier 等人[]在 IC、LT 模型的基础上提出了基于话题的信息扩散模型 TIC、TLT，利用 EM 算

法计算基于话题的扩散模型中参数。这些话题相关研究主要关注的是话题建模、扩散模型及参数学习，并没有分析选使影响力最大的择初始结点这一环节的算法。如何利用基于话题的影响力特性扩展经典影响力最大化算法是本文关注重点。

1.3 本论文的主要内容

本文围绕社会网络中信息传播机制的研究内容，探讨影响力最大化问题不同解决算法的思想，最后关注点扩展到基于话题的影响力最大化问题的建模和算法解决。

第一章，介绍了当前社会网络影响力最大化的研究发展主线和现状，引出从基于目标话题的角度重新考虑影响力最大问题。

第二章，主要介绍本文所研究问题的主体——社会网络的建模和重要特性。

第三章，介绍社会网络之上的信息传播这一表象行为的建模——信息扩散模型。

第四章，介绍了信息传播中的动力学原理——社会影响力的分析，引入影响力最大化问题的探讨，分析了问题的难点及解决算法的关键技术。

第五章，研究基于话题的影响力最大化问题，通过问题抽象、扩展经典影响力最大化算法，最后提出了 TSRS 算法。

第六章，介绍了在两个真实数据集上对本论文中所提出的基于话题的影响力最大化算法的实验。

第七章，总结了本论文所做工作和分析了仍代提高和拓展方面。

第二章 社会网络

2.1 概述

作为信息传播现象的承载主体,社会网络在这里并不单指 Web 在线社交网站这一种狭义定义,而是指人与人之间形成信息交互关系网络的广义统称,可能会包含各种各样的社交媒体,如社交网站、Email、移动网络等。这些丰富的社会网络都可以用图模型来统一表示,并具有一些特殊的网络特性。

网络的研究最早要追溯到年,欧拉致力于著名的“哥尼斯堡七桥问题”的研究,图论由此萌芽。网络实际上也是一个图形结构,事物以及事物之间的某种关系构成了网络。而社会网络与上述“网络”的定义类似,“社会网络”是由多个点社会行动者和各点之间的连线行动者之间的关系组成的集合。社会网络中的点是各个社会行动者,行动者可以是个体、公司或者集体性的社会单位,也可以是学校、学院、村落、组织、社区、城市、国家等。而行动者之间的关系是多种多样的,可以是朋友关系、上下关系、国家之间贸易关系,也可以是个体之间的合作关系、互动关系等。社会网络关注的是人们之间的互动和联系,社会互动影响着人们的社会行为。

经过许多学者的研究,人们发现社会网络具有如下一些社会学角度的特征:

1.社会网络的形成是地缘、血缘、学缘、业缘等多方面的因素使然。许多网络在我们一生中是自然形成的。比如邻居网络、同乡网络的形成是因为地缘,亲戚联络是由于血缘,校友网络是学缘,同事网络是业缘。由于互联网的普遍使用,今后会有更多的人还拥有“社交,,网络。

2.社会网络反映个人和社会关系的本质。社会网络实际上是人际互动的反映,一方面它要受社会关系的规定和制约另一方面,又构成广泛的、间接的、更为复杂的个人和社会关系的基础。要认识个人和社会之间关系的本质,必须要对个人所属的社会网络及其中的社会互动加以分析。

3.社会网络是经过个人之间的社会互动所形成的。运用各种互动的媒介和符号进行交往是社会网络得以形成的前提,如果是单向的行动是无法构成社会网络的。

4.社会网络对个人来说具有效益。人们可以从自己所属的社会网络中,获得所

需要的信息、获得情感的支持、满足个人在社会生活中的多种需求、丰富个人的社会生活。

5.社会网络是相对稳定的。不同的社会网络内部,可能是紧密联系的,也可能是松散联系的,但一般而言,社会网络一旦形成,就具有相对的稳定性。

社会网络也是一个图结构,图结构中的每个结点表示一个个体,边可以表示个体之间的关系。下面,就将从数学模型和统计特性的角度来分析社会网络。

2.2 图模型表示

人与人之间进行信息交互就形成了社会网络,用图模型的方式进行网络的抽象描述,人即结点 v , 用户之间的社会联系即结点之间的边 $e=(v,u)$; V 表示网络中所有点的集合, $V=\{v_1,v_2,\dots,v_n\}$; E 表示所有边的集合, $E=\{(v,u)|v,u\in V\}$; 社会网络即结点和边组成的结构图,用 G 表示, $G=(V,E)$ 。用邻接链表或邻接邻接矩阵 A 来表示整体网络信息。其他常用符号表示如下:

| 符号 | 意义 |
|--------------|--|
| V | 网络的结点集合 |
| E | 网络的边的集合 |
| n | 结点个数 ($n= V $) |
| m | 边的个数 ($m= E $) |
| v_i | 一个结点 v_i |
| $e(v_i,v_j)$ | 一条结点 v_i 和 v_j 之间的边 |
| A | 网络的邻接矩阵, $A\in\{0,1\}^{n\times n}$ |
| A_{ij} | 若 $\forall e(v_i,v_j)\in E$,则为 1; 反之为 0 |
| N_i | 结点 v_i 的邻接结点 |

| | |
|-------|-------------------------------|
| d_i | 结点 v_i 的度 ($d_i = N_i $) |
|-------|-------------------------------|

在本论文后续出现的算法说明中都将以此表为符号表示基准。网络还可以拓展表示为加权的、有符号、有方向的网络。

2.3 网络元素及重要性测度

网络的基本元素即结点和边，下文中将介绍结点和边的统计特性如度、距离、介数、联系强度等性质，在社会网络的研究过程中,不管是对社区结构、影响力还是传播过程等方面的研究都离不开对基本性质的了解,因此了解这些统计性质对社会网络的研究有比较重要的意义。

2.3.1 结点的重要性

对事物进行排名是人类社会中一种很普遍的需求。很多时候需要给出网络中哪些结点比较重要，才能提供额外的推荐或决策信息。以下是社会网络中结点重要性，或说中心性(Centrality)度量的一些经典指标。这些中心性指标可以和影响力最大化问题结合，在选取起始结点中最为评判项（如 Degree-discount Hueristic 中就使用了度中心性为依据之一）。

1.度中心性

度的定义：结点 v_i 的所有邻接结点的数量称为它的度 d_i ；在有向图中，指向 v_i 的所有邻接结点数之和称为入度 d_i^{in} ， v_i 指向的邻接结点数之和称为出度 d_i^{out} 。结点的度分布是指的是网络图中度为的结点的概率随结点度的变化规律，大多数网络的度分布是幂律分布

结点的度中心性：很显然地，一个结点的重要性与该结点邻接的结点的数量有关，即一结点的度越大，该结点就显得越重要。用公式表示度中心性：

$$C_D(v_i) = d_i = \sum_j A_{ij}。在不同网络之间比较时规范化后的度中心性：$$

$$C'_D(v_i) = d_i / (n - 1) \quad (\text{公式 2-1})$$

2.紧密度中心性

网络中的距离：结点之间的最短路径为结点 v_i 到 v_j 经过的最少的边，称为 geodesic(测底线)；最短路径的距离即路径上边的跳数，记作 $g(v_i, v_j)$ 。 $g(v_i, v_j)$ 的最大值称为网络的直径。网络的平均路径长度定义为任意两个结点之间的距离的平均值：

$$C_{avg}(v_i) = \frac{1}{n-1} \sum_{j \neq i}^n g(v_i, v_j) \quad (\text{公式 2-2})$$

紧密度中心性：从距离的角度出发，越重要的点，到达网络中其它结点理应更快。紧密度中心性用于评价一个结点到其他所有结点的紧密程度，用结点最短路径距离的平均值的倒数表示：

$$C_c(v_i) = \left[\frac{1}{n-1} \sum_{j \neq i}^n g(v_i, v_j) \right]^{-1} = \frac{n-1}{\sum_{j \neq i}^n g(v_i, v_j)} \quad (\text{公式 2-3})$$

3.介数中心性

介数：介数分为边介数和点介数。结点的介数表示一个网络中经过该结点的最短路径的数量。

介数中心性：

$$C_B(v_i) = \sum_{v_s \neq v_i \neq v_t \in V, s < t} \frac{\sigma_{st}(v_i)}{\sigma_{st}} \quad (\text{公式 2-4})$$

上式中， σ_{st} 是结点 v_s 和 v_t 之间存在的最短路径的数量，而 $\sigma_{st}(v_i)$ 是这些最短路径中经过点 v_i 的路径数量。

两个不相邻顶点和之间的通讯,依赖于连接和路径上的其他结点。所以度量一个顶点的关联性可以通过计算经过该顶点的最短路数量来得到,定义为点的介数,边的介数同理。介数反映了相应的结点或者边在整个网络中的作用和影响力,具有很强的现实意义。例如,在社会关系网络或技术网络中,介数的分布特征反映了不同人员、资源和技术在相应生产关系中的地位,这对于在网络中发现和保护关键资源和技术具有重要意义。但是,对于但规模网络计算所有结点之间存在的最短路径是不可行的,时间复杂度为 $O(mn)$ 。

4.特征向量度中心性

特征向量中心性：通俗讲，一个人的重要性取决于其朋友的重要性。网络中，若一个结点拥有的重要的邻接结点越多，则此结点越重要。结点 v_i 的特征向量满足：

$$C_E(v_i) \propto \sum_{v_j \in V} A_{ij} C_E(v_j) \quad (\text{公式 2-5})$$

根据矩阵运算形式可知，特征向量中心性即网络邻接矩阵的主特征向量。

2.3.2 联系的强度

定义：十分明显的，人际关系中包括一些强联系(Strong tie)和弱联系(Weak tie)，区分这些连接的重要性在了解信息扩散机制十分重要。联系即网络中的边，联系的强弱反映到网络中即边权重值的大小，反映到第三章中的信息扩散模型上，即不同边的信息传播概率或阈值的设置。

意义：根据[20]中的报告显示，研究人员基于 Email 数据研究了用户之间的交互性对信息扩散的影响，发现信息从初始者到扩散者时通常会走弱关系边，而从扩散者到接收者时通常会走强关系边，说明信息在流动过程中有从弱关系到强关系的转向特征。另外发现扩散者并非中心性很强的社会化中心点(Social Hubs)，而是很普通的点，研究人员对比了所有用户的度分布和扩散者的度分布，发现非常匹配，证明了这个观点。因此可知，在信息扩散中加入联系强度因素的考虑可以帮助我们完善对信息扩散机制的了解。

研究现状：网络中的弱关系是 Granoveter[12]提出的，他们发现了弱关系可以帮助人找到新工作，过去有人研究弱关系和传统社会网络的关联。随着在线社会网络的出现，Zhao 等人[13]研究了在线社会网络中的弱关系对信息扩散的影响，关系的强弱是通过用户之间邻居的重叠数量衡量的。他们提出了 $ID(\alpha, \beta)$ 信息扩散模型，该模型可以视为独立级联模型和传染病模型的组合，它通过设置不同 α 可以选择不同强度的边扩散，而 β 是信息本身的扩散性，对于一个结点 i ，它会有 $d(i)\beta$ 个邻居被传染， $d(i)$ 是结点 i 的度。通过使用该模型在 Youtube 和 Facebook 数据集上的模拟实验发现，社会网络中的弱关系对信息扩散有着微妙的作用，弱关系相当于一个桥，它可以把孤立的团体链接起来，打破信息扩散的局部局限性，当刻意选择弱关系作为扩散路径时不能增强信息扩散的范围，但是

如果不使用弱关系则会降低信息扩散的范围。

学习：不像结点的中心性是网络结构的显性特征，可以由显性数据直接求得，联系的强度更加抽象和隐性，需要从其他信息中学习(Learning)得来。联系强度的分析方法主要有三种，从网络拓扑中学习、从用户特点和交互中学习、从用户行为序列中学习[11]。

2.4 社会网络统计特性

2.4.1 幂律分布

自然界与社会生活中存在各种各样性质迥异的幂律分布现象有时也叫长尾现象，如图 2-1 所示。



图 2-1 长尾现象

网络中的幂律分布现象：网络中度很大的结点只占网络中结点总量的少部分，而度较小的的结点数量却占大多。形式化描述为，一个网络中的度分布服从 Power Law 即有 $N(k) = k^{-r}$ 。 k 表示度为 k 的结点， N 表示度为 k 的结点的个数，则 $N(k)$ 在 k 上的 Power Law 分布如图 2-2 所示。

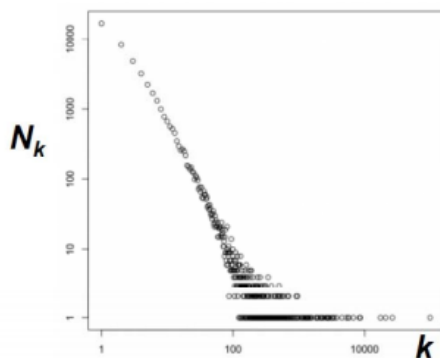


图 2-2 社会网络中的 Power Law 分布图

2.4.2 六度分隔理论

社会网络中比较著名的理论是“六度分割理论”和法则。六度分隔理论(Six degree of separation)[14]假设世界上所有互不相识的人只需要很少中间人就能建立起联系。六度分割理论是说“最多通过六个人你就能够认识任何一个陌生人”。后来1967年哈佛大学的心理学教授斯坦利·米尔格拉姆根据这概念做过一次连锁信实验，尝试证明平均只需要5个中间人就可以联系任何两个互不相识的美国人，如下图2-3所示。

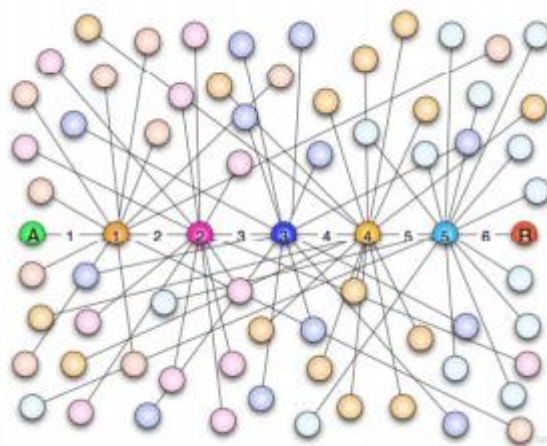


图 2-3 六度分隔现象

“六度分度”说明了社会网络中普遍存在着“弱纽带”，这些“弱纽带”使得人与人之间的距离变得非常“相近”。法则成为人们普遍公认的“一个人可以保持社交关系的人数最大值”，无论你曾经认识多少人，或者通过一种社会性网络服务与多少人建立了弱链接，那些强链接仍然在此次此刻符合法则。

2.4.3 无尺度分布

在网络理论中，无尺度网络（或称无标度网络）是带有一类特性的复杂网络，其典型特征是在网络中的大部分节点只和很少节点连接，而有极少的节点与非常多的节点连接。这种关键的节点（称为“枢纽”或“集散节点”）的存在使得无尺度网络对意外故障有强大的承受能力，但面对协同性攻击时则显得脆弱。无尺度网络的特性，在于其度分布没有一个特定的平均值指标，即大多数节点的度在此附近。在研究这个网络的度分布时，Barabási 等人发现其遵守幂律分布。

2.4.4 小世界效应

小世界现象(Small World phenomenon)指世界上的每个人之间都可以通过很短的社会关系链联系起来。小世界网络,引伸了小世界现象,不仅是社会关系网络,可以指任何网络。小世界网络就是对这种现象(也称为小世界现象)的数学描述。用数学中图论的语言来说,小世界网络就是一个由大量顶点构成的图,其中任意两点之间的平均路径长度比顶点数量小得多。

第三章 信息扩散

3.1 概述

当社会网络结构搭建起来后,信息就可以在网络上进行流动了。信息扩散在当前学术研究中有多种称谓,主要有信息扩散(Information Diffusion)、信息传播(Information Propagation, Information Spread)、信息流动(Information Flow)等。早期,信息扩散的研究人员多是一些社会学家、传染病学家和经济学家,他们主要研究创新(Innovation)、传染病(Epidemic)和产品(Product)在真实社会网络中的扩散,但是受限于真实社会中数据获取的困难,这些研究通常采用的数据集很小,而且多是一些定性的研究。伴随着社会媒体的发展,大量丰富的在线数据可以非常方便的获取,这些在线数据不仅包括大规模的在线社会网络数据,还有海量信息在网络中扩散的数据,这为信息扩散的研究带来了新的契机,从而成为研究热点。

3.2 信息扩散

在线社会网络中信息扩散研究具有极其广泛的应用价值,主要包括病毒式市场营销、在线广告投放、信息推荐和谣言控制等多个方面。这里需要说明的是,在线社会网络中扩散的信息,并不狭隘的限制于文本信息(微博、博客等),还包括多媒体信息(视频、图片等)、产品信息等各类信息。信息在网络中的扩散是用户行为引起的,例如,在微博中,用户的转发行为引发了微博信息在网络中扩散,所以信息扩散的研究本质上是用户行为的研究。

根据本论文后文研究问题的需要,将社会网络中的信息扩散现象抽象描述为以下问题:信息扩散在离散的时间步数内进行,。对每个结点 $v_i \in V$ 都有两个状态,激活态(active)和非激活态(inactive)。直观地,我们将结点的激活态表示为该结点接受了在网络中传播的新信息、新观点或新产品。相反的,非激活态表示该结点未接受传播的新信息。让 $S_t \subset V$ 表示时刻 t 激活的点的集合,即活跃点集(active set),把 S_0 叫做种子集合(seed set),即信息扩散初始化时被选中用来引发扩散级联的起始结点集合,例如那些在公司促销活动中被选中接收免费的新产

品样品的用户。

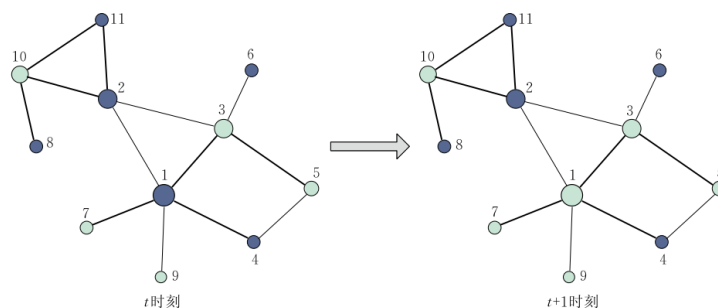


图 2-5 网络中的信息扩散示例[20]

如图 2-5 所示，在线社会网络可以使用一个有向或者无向图 $G=(V,E)$ 表示，图中结点代表用户，本文中结点和用户是同义。 V 代表结点集合，结点有激活（浅色）和未激活（深色）两类状态。图中结点的权重代表用户影响力，对应着用户影响力计算研究； E 代表边的集合，图中边的权重代表用户间关系强度，对应着信息扩散概率计算研究。信息扩散是一个时间动态性的过程，这里使用 V^t 代表时刻 t 结点的状态集合，信息扩散过程可描述为 $V^t \rightarrow V^{t+1}$ ，它是用户状态集合的跳转过程，部分未激活态的结点随着时间递进变成激活状态。

3.3 扩散模型

信息在社会网络传播过程中都遵循一定的规则，称之为信息传播模型 (diffusion model)。信息传播建模是对信息扩散机制的探索。为了能准确地分析信息的传播，通常构建符合该网络传播特征的信息传播模型。信息传播模型不仅能够可视化网络的信息传播过程，并且能够预测信息未来的传播路径和传播趋势。构建模型的前提需要清楚信息的传播过程。

根据信息扩散过程中结点状态改变是否保持，将扩散模型分为渐进模型 (Progressive model) 和非渐进 (Non-progressive model) 模型[15]。其中渐进的扩散模型即经典基于影响力的扩散模型。由于本文关注的是影响力驱使下的信息扩散问题，所以下文中将主要介绍两种渐进模型。

3.3.1 渐进模型

1.IC 模型

独立级联模型[16](Independent Cascade Model, ICM)，借鉴了交互粒子系统和概率论的理念。在该模型中，每个初始激活结点会产生自己独立的扩散级联(信息级联的例子如图所示)，级联之间是互相独立、互不干扰的。该模型关注的是信息的发送者(sender)胜过接收者(receiver)。在独立级联模型中，一个结点 w 一旦在第 t 步被激活，它只有一次机会激活它的邻接结点。对于其邻接结点 v ，其被 w 激活的概率为 $p_{w,v} \in [0,1]$ 。如果 v 被成功激活，那么 v 就是第 $t+1$ 步被激活的结点。在往后的信息扩散过程中， w 将不再试图激活其余的邻接结点，该进程直到不再有激活行为发生而终。

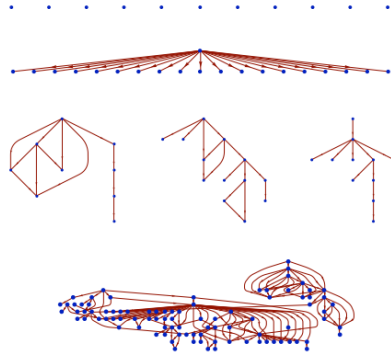


图 3-2 Twitter 中信息级联 (Information cascade) 的例子

2.LT 模型

线性阈值模型[17](Linear Threshold Model, LTM)。该模型可简单表述为：如果一个用户的采取行动的朋友的数量超出某一阈值，那么该用户才会采取行动。该模型关注的是信息的接收者(receiver)胜过发送者(sender)。在该模型中，对于网络中每个结点 v 在 $0 \sim 1$ 内均匀分布随机抽取一个阈值 θ_v 。阈值 θ_v 表示为了激活结点 v ，结点 v 的朋友需要被激活的比例。假设邻接结点 w 能够影响结点 v 的强度为 $b_{w,v}$ ，不是一般性，假设： $\sum_{w \in N_v} b_{w,v} \leq 1$ ，随机给网络中的所有结点分配一个阈值，同时给定初始活动结点集合 S_0 ，那么网络中信息的扩散过程是唯一确定的。在信息扩散的每一步，所有前一步被激活的结点仍保持活动。满足以下条件的结点将被激活： $\sum_{w \in N_v, w \text{ 是活动的}} b_{w,v} \geq \theta_v$ ，这个过程不断执行下去直到没有结点可以激活。

3.IC 和 LT 比较

IC 和 LT 在信息扩散问题上有如下共同点：

1. 社会网络表示成一个有向图，每一个用户是一个结点。
2. 每个结点的初始状态或为活动的或为不活动的。
3. 一个结点被激活后将激活它的邻接结点。
4. 一旦一个结点被激活，这个结点就不能被吊销。

不同点：

IC 是从激活用户的角度解释影响力传播过程，即激活过程。而 LT 则是从被激活用户的角度解释影响力传播过程。

3.3.2 非渐进模型

在社会网络中常使用的非渐进信息传播模型有 SIR 模型（传染病传播模型），巴斯模型,投票模型等。由于本文关注的是影响力驱使下的信息扩散问题，将不展开探讨非渐进模型，详细分析可参考[15]。

第四章 社会影响力分析

4.1 概述

在社会学中对影响力已经有很长的研究，向著名的六度分隔定理(Six Degree of Separation)和三阶影响(Three Degree of Influence)[18]。根据 Dunbar's number[19]，你能对世界上 1,000,000 多人产生影响。

4.2 影响力

根据维基百科中的定义，当个人的意见、情感或行为因他人而，有意或无意地，产生改变时，社会影响力就产生了。在社会网络中，社会影响力表示人们倾向于遵循朋友的行为。

在商业上能说明影响力扩散导致商业成功的比较有名的案例是 Hotmail 现象[19]。在 90 年代初期，Hotmail 还属于不太知名的邮件服务提供商，然后他们想出一个很简单的想法：在他们用户所发的每封邮件最后附上“Join the world's largest e-mail service with MSN Hotmail.<http://www.hotmail.com>。”事实证明这个策略创建并宣传了一个品牌。在仅仅 18 个月后 Hotmail 成为第一大邮件服务提供商，拥有了 8 百万用户群。这个案例背后的道理是首先一小部分用户受邮件附加信息诱使而接受并试用 Hotmail，然后当他们发送给其他人邮件时，又会有一小部分人受到相似诱惑。这种现象被迅速的扩散，在人群中形成病毒式感染。

病毒营销是由欧莱礼媒体公司(O'Reilly Media)总裁兼 CEO 提姆奥莱理(Tim O'Reilly)提出，其讯息传递策略是通过公众将信息廉价复制，告诉给其它受众，从而迅速扩大自己的影响。

4.3 基于影响力的传播模型

典型的基于影响力的模型就是在本论文第三章中提到的两个渐进模型：IC 模型和 LT 模型。模型定义请参考前文。

4.4 影响力最大化问题

信息扩散最大化问题[2,3]是应用性很强的研究型问题，通常也称作影响力最

大化问题，下面通过一个事例来说明什么是影响力最大化问题：一个小公司推出了一款新产品，但是由于资金等问题，它只能选择一部分用户来试用这款产品（通过赠送礼品或者提供优惠的方法），这个公司希望这些用户会喜欢这种产品，并且影响他们在社会网络中的朋友们去使用这款产品，接着，他们的朋友再影响他们朋友的朋友，类推下去，这样新产品信息就能在社会网络中逐渐的扩散开来。

为了赢得病毒营销活动，需要完成两个重要任务：（1）详尽描述影响力在网络中的扩散过程，包括学习扩散模型的参数（2）在学习来的扩散模型下，设计有效的方法来识别网络中用来推销的起始点。本论文在重点关注第二个问题，对第一个问题不展开详细讨论。

影响力最大化问题就是如何选择初始结点集合，使得这些结点可以最大程度地影响社会网络中的其它结点，使信息在社会网络上可以获得最大程度的扩散。此类研究可以广泛应用于病毒式市场营销、广告投放等，具有重要的商业价值。

通常的，影响力最大化问题可以定义为：给定一个社会网络和一个扩散模型，任务就是如何在社会网络上获取一个指定大小的结点集合，使得这个结点集合可以达到影响力最大化的效果。

4.4.1 问题描述

给定一个社会网络图 $G=(V, E)$ 、在 G 上的扩散模型 m (如 ICM 或 LTM)、预算 k ，寻找 $S \subseteq V$ 且 $|S|=k$ 作为初始点集，使在 S 的影响下最终激活的结点的数量的期望， $\sigma_m(S)$ ，最大化。即，寻找

$$S^* = \arg \max_{S \subseteq V, |S|=k} \sigma_m(S) \quad (\text{公式 4-1})$$

4.4.2 问题分析

1.影响力最大化问题的定位区别

一个在问题描述中需要注意的地方是“求一个大小为 k 的使影响力最大的结点集合”不等于“求 Top- k 个影响力最大的结点”。直观地，如果两个影响力很大的结点都能影响相同的结点集，那么这两个结点都会被视为影响力最大的结点，但对于影响力最大化问题，只有这两个结点的一个会被选做种子结点。

2.影响力最大化问题 NP 难的来源[15]

- a) 影响力最大化问题本身含有的组合问题的难度--贪婪算法
- b) 影响力计算的难度--Monte Carlo simulations

3.目标函数的子模性

Kempe 证明在 ICM 和 LTM 下求解影响力最大化问题是 NP 难的问题，近似算法贪心算法求得的的影响力扩散解可以保证达到最优影响力扩散解的 $(1-1/e)$ ，约 63%[4]。对于大规模的社会网络，只能采用一些优化算法获取近似的较优解。在[Kempe]中 Kempe 给出如下定理：“定理 2.1 对于一个非负单调次模函数 f ，设 S 是大小为 k 的由每次添加使 f 获得最大边缘收益的元素而得来的集合。设 S^* 是所有大小为 k 的集合中使 f 的值最大的集合。那么有 $f(S) \geq (1-1/e) \cdot f(S^*)$ ；换句话说， S 可以提供 $(1-1/e)$ 的近似。”

利用影响力最大化问题的目标函数的子模性(Submodularity)是优化寻找最优种子集合算法的关键技术。

4.4.3 问题解决

4.4.3.1 求解过程

在给定 G 和 K 后，影响力最大化问题的求解过程：

1.确定影响力最大化模型：即选择使用的扩散模型 m ，定义目标函数——影响力扩散函数 $\sigma_m(\cdot)$ 。

2.模型学习：学习扩散模型的参数，如从该网络过去的扩散数据(past propagation)中学习来 IC 模型中使用的每个边上的影响概率。或直接简单指定每条边的传播概率为 0.01 等定值[Kempe]。

3.设计寻找种子集的算法：目前解决影响力最大化问题的方法主要有贪心算法和启发式算法有两类。经典算法将在下一节中描述。

4.4.3.2 贪心算法

贪心算法在对问题求解时，总是做出在当前情况下最好的选择。使用贪心算法来解决影响力最大化问题的思路是，每次选择的结点都可以达到当前影响力最大化的效果。假设，目前已经选出的结点集合为 S ，使用的贪心算法在选取某个结点时，会把集合 S 同每个集合 S 之外的每个结点 v 结合，计算每个 $S \cup \{v\}$ 的影响力，选出产生最大影响力的那个 v ，加入到集合 S 中，集合 S 初始值为空。

贪心算法的近似最优保证依赖于影响力扩散函数（Influence spread function） $\sigma_m(\cdot)$ 的子模性和单调性。只要影响力扩散函数满足子模和单调性，贪心算法本身可以在任何扩散模型下工作。为了简化，我们只关注在 IC 和 LT 模型下的贪心算法。

贪心近似最大化方法基本的流程：（1）使用扩散模型在社会网络上使用 Monte Carlo 模拟模仿扩散进程——解决影响力计算的难度（2）根据扩散效果来衡量候选种子结点，使用贪心策略选择本轮最优结点——解决影响力最大化问题本身含有的组合问题的难度。

4.4.3.2.1 GeneralGreedy

| Algorithm 1: GeneralGreedy(G, k) |
|---|
| <pre> 1: initialize $S = \emptyset$; and $R = 20000$ 2: for $i = 1$ to k do 3: for each vertex $v \in V \setminus S$ do 4: $sv = 0$; 5: for $i = 1$ to R do 6: $sv += RanCas(S \cup \{v\})$ 7: end for 8: $sv = sv / R$ 9: end for 10: $S = S \cup \arg \max_{v \in V \setminus S} \{sv\}$ </pre> |

```

11: end for
12: output S
    
```

4.4.3.2.2 NewGreedy

Algorithm 2: NewGreedyIC(G,k)

```

1: initialize  $S = \emptyset$ ; and  $R = 20000$ 
2: for  $i = 1$  to  $k$  do
3:   set  $sv = 0$  for all  $v \in V \setminus S$  for each vertex  $v \in V \setminus S$  do
4:     for  $i = 1$  to  $R$  do
5:       compute  $G'$  by removing each edge from  $G$  with probability  $1-p$ 
6:       compute  $R_{G'}(S)$ 
7:       compute  $|R_{G'}(\{v\})|$  for all  $v \in V$ 
8:       for each vertex  $v \in V \setminus S$  do
9:         if  $v \notin R_{G'}(S)$  then
10:            $sv += |R_{G'}(\{v\})|$ 
11:         end if
12:       end for
13:     end for
14:   set  $sv = sv / R$  for all  $v \in V \setminus S$ 
15:    $S = S \cup \arg \max_{v \in V \setminus S} \{sv\}$ 
16: end for
17: output S
    
```

4.4.3.2.3 CELF

算法思想：利用次模性质为每个结点维护一个上界，以减少计算量。

关键技术: Lazy evaluation

Input: k : size of returned set; f : monotone and submodular set function

Output: selected subset

```

1: initialize  $S \leftarrow \emptyset$ ; priority queue  $Q \leftarrow \emptyset$ ;  $iteration \leftarrow 1$ 
2: for  $i = 1$  to  $n$  do
3:    $u.mg \leftarrow f(u | \emptyset)$ ;  $u.i \leftarrow 1$ 
4:   insert element  $u$  into  $Q$  with  $u.mg$  as the key
5: end for
6: while  $iteration \leq k$  do
7:   extract top (max) element  $u$  of  $Q$ 
8:   if  $u.i = iteration$  then
9:      $S \leftarrow S \cup \{u\}$ ;  $iteration \leftarrow iteration + 1$ ;
10:  else
11:     $u.mg \leftarrow f(u | S)$ ;  $u.i \leftarrow iteration$ 
12:    re-insert  $u$  into  $Q$ 
13:  end if
14: end while
15: return  $S$ 

```

4.4.3.3 启发式算法

贪心近似算法解决影响力最大化问题的优点是可以保证 63% 最优，缺点是需要大量蒙特卡罗模拟来估计每个候选种子集合的影响力传播，速度慢。但如果仅仅减少蒙特卡罗模拟次数，即算法描述中 ROUND 的值，不仅会显著地减小影响力扩散，且会非显著地减少运行时间[15]。

为了解决这个问题，一些启发式策略被提出[Chen et al., 2009, 2010, Goyal et al., 2011b, Jung et al., 2012, Wang et al., 2012]。这些策略背后的一个普遍思想是通过探究网络图结构和扩散模型的特质，来提高影响力计算的速度，而避免使用蒙特卡罗模拟。

下面将介绍两个经典的结点和距离中心性启发式算法和一个由 Wei Chen 提出的 DegreeDiscount 启发式算法。

4.4.3.3.1 Degree Discount

算法思想：照度降序选择结点，每选中一个结点，将其邻居结点的度减小一定值。

Algorithm 3: DegreeDiscountIC(G,k)

```

1: initialize  $S = \emptyset$ 
2: for each vertex  $v \in V$  do
3:   compute its degree  $d_v$ 
4:    $dd_v = d_v$ 
5:   initialize  $tv$  to 0
6: end for
7: for  $i = 1$  to  $k$  do
8:   select  $u = \arg \max \{dd_v \mid v \in V \setminus S\}$ 
9:    $S = S \cup \{u\}$ 
10:  for each neighbor  $v$  of  $u$  and  $v \in V \setminus S$  do
11:     $tv = tv + 1$ 
12:     $dd_v = d_v - 2tv - (d_v - tv)tv \cdot p$ 
13:  end for
14: end for
15: output  $S$ 

```

第五章 基于话题的影响力最大化

5.1 概述

论文的第四章中已引入影响力最大化问题并对现有的解决算法进行了介绍和分析，这些内容也是近来社会网络影响最大化的主要研究关注点，但是以上工作都忽略了结点用户本身具有不同个话题分布，或者说有不同的兴趣关注。病毒营销活动中存在目标市场这一实际情况，实际中为了使产品得到最广泛的推销应该把目标锁定在对产品感兴趣的用户群上。基于目标话题的影响力最大化问题其实是影响力最大化问题的扩展。在本章中将描述这一类，调整普通影响力最大化问题中的扩散模型和算法以解决基于目标话题的影响力最大化问题，并提出了自己的新方法基于种子选区的方法。

基于话题的影响力最大化问题的解决过程主要有两步（1）话题建模：学习结点和消息话题分布（2）影响力最大化方法：遵循普通影响力最大化问题的解决思路，选择扩散模型和设计使基于话题的影响力最大化的初始集 S 的选择策略。本论文将重点关注第二步。即假定已知社会网络中的话题分布。

5.2 问题描述

给定(1)一个已经标注好的社会网络图 $G=(V, E, T_v)$ ，其中 $T_v = \{T_{v_1}, T_{v_2}, \dots, T_{v_n}\}$ 是每个结点的话题兴趣分布， T_{v_i} 是结点 v_i 感兴趣的话题的集合。(2)一个目标话题集合 $T = \{t_1, t_2, \dots, t_q\}$ (3)一个基于目标话题 T 的结点收益评估函数 pf_T ，这是一个二值函数，如果结点 v_i 的话题集合 T_{v_i} 中含有目标话题集合 T 中的任一话题则 $pf_T(v_i)=1$ ；反之，如果结点 v_i 不含有任何目标话题集合 T 中的任何话题则 $pf_T(v_i)=0$ (4)在 G 上的扩散模型 m 。

那么，基于话题的影响力最大化问题即寻找寻找 $S \subseteq V$ 且 $|S|=k$ 的初始点集，使在 S 的影响下最终激活的含有目标话题的结点的数量的期望 $R_T(S)$ 最大化，

$R_T(S) = \sum_{v_i \in S} pf_T(v_i)$ 。即，寻找

$$S^* = \arg \max_{S \subseteq V, |S|=k} R_T(S) \quad (\text{公式 5-1})$$

5.3 问题分析

比较第三章中普通影响力最大化问题中 S^* 的定义，可发现在基于话题的影响力最大化目标函数 $R_T(S)$ 只是对 $\sigma_m(S)$ 进行了限定，且有 $R_T(S) = \{v \mid v \in \sigma_m(S), \mathcal{T}_v \cap T \neq \emptyset\}$ ，显然 $R_T(S)$ 仍保持子模性、非负、单调，因而贪心策略仍适用于基于话题的影响力最大化问题且能保证约 63% 的最优。

5.4 基于话题的影响力最大化方法

5.4.1 经典影响力最大化方法套用

5.4.1.1 Topic-aware General Greedy

| Algorithm 5: TGG(G,k) |
|--|
| <pre> 1: initialize $S = \emptyset$; and $R = 20000$ 2: for $i = 1$ to k do 3: for each vertex $v \in V \setminus S$ do 4: $sv = 0$; 5: for $i = 1$ to R do 6: $sv += R_T(S \cup \{v\})$ 7: end for 8: $sv = sv / R$ 9: end for 10: $S = S \cup \arg \max_{v \in V \setminus S} \{sv\}$ 11: end for 12: output S </pre> |

5.3.2 Topic-aware New Greedy

| Algorithm 6: TNG-IC(G, k) |
|---|
| <pre> 1: initialize $S = \emptyset$; and $R = 20000$ 2: for $i = 1$ to k do 3: set $sv = 0$ for all $v \in V \setminus S$ for each vertex $v \in V \setminus S$ do 4: for $i = 1$ to R do 5: compute G' by removing each edge from G with probability $1-p$ 6: compute $\Gamma_{G'}(S)$ 7: compute $\Gamma_{G'}(\{v\})$ for all $v \in V$ 8: for each vertex $v \in V \setminus S$ do 9: if $v \notin \Gamma_{G'}(S)$ 10: $sv = \Gamma_{G'}(\{v\})$ 11: end if 12: end for 13: end for 14: set $sv = sv / R$ for all $v \in V \setminus S$ 15: $S = S \cup \arg \max_{v \in V \setminus S} \{sv\}$ 16: end for 17: output S </pre> |

5.3.3 Topic-aware Degree Discount

| Algorithm 7: TDD-IC (G, k) |
|--|
| <pre> 1: initialize $S = \emptyset$ 2: for each vertex $v \in V$ do </pre> |

```

3:   compute its degree dv
4:   ddv=dv
5:   initialize tv to 0
6: end for
7: for i = 1 to k do
8:   select u = arg max{ddv | v ∈ V \ S}
9:   S = S ∪ {u}
10:  for each neighbor v of u and v ∈ V \ S do
11:    if  $T_{v_i} \cap T \neq \emptyset$  then
12:      tv = tv + 1
12:    else do
13:      ddv =  $(1-p)^{tv+rv} \cdot (\sum_{u \in dv-tv} (pf_T(u) \cdot p))$ 
13:    end for
14: end for
15: output S
    
```

5.4.2 一种新方法:Topic-aware Seed Region Split

算法思想：利用影响扩散的局部性，在选取种子节点时保证候选节点非当前 S 集中任意节点的距离为 3 的邻接节点。

算法描述：

Algorithm 7: TSRS-IC (G,k)

```

1: initialize S = ∅
2: for each vertex v ∈ V do
3:   compute its degree dv
4:   ddv=dv
5:   initialize tv to 0
6: end for
    
```



```

7: for  $i = 1$  to  $k$  do
8:   select  $u = \arg \max\{dd_v \mid v \in V \setminus S\}$ 
9:    $S = S \cup \{u\}$ 
10:  for each neighbor  $v$  of  $u$  and  $v \in V \setminus S$  do
11:    if  $T_{v_i} \cap T \neq \emptyset$  then
12:       $tv = tv + 1$ 
12:    else do
13:       $dd_v = (1-p)^{tv+rv} \cdot (\sum_{u \in dv-tv} (pf_T(u) \cdot p))$ 
13:    end for
14:  end for
15: output  $S$ 
    
```

第六章 实验

6.1 概述

本章将在两个真实社会网络数据集上对四个基于话题的影响力最大化方法进行实验验证,分别是第五章中的 Topic-aware General Greedy(TGG)、Topic-aware New Greedy(TNG)、Topic-aware Degree Discount(TDD)和 Topic-aware Seed Region Split(TSRS)方法。通过不同模型在相同网络上的对比实验,比较分析每种算法的影响效果和运算效率,并区分多目标话题和单目标话题任务两种基于话题的最大化问题。

6.2 数据集选取

数据集 1: Citation 网络,来自 ArnetMiner 网站,是一个来自 10 个研究话题领域的论文引用关系数据集,包含 2555 个结点,5967 条边。结点话题分布由 <http://arnetminer.org/topicBrowser.do> 的 author-conference-topic 模型处理得来。

数据集 2: Movie-actor-director-writer 网络,简称“Movie”,来自 Wikipedia 下的“English-language films”类条目,是一个异构网络,话题由 wiki 条目中的分类标签得来,包含 10 个话题。

数据集分析：

| 指标\数据集 | Citation | Movie |
|--------|----------|---------|
| 网络图 | 有向图，无权重 | 无向图，无权重 |
| 结点数 | 2555 | 34283 |
| 边数 | 5976 | 142427 |
| 结点平均度数 | 4.7 | 7.6 |
| 最大结点度数 | 98 | 1501 |

6.3 数据集处理

Citation 数据集中包含 10 个话题分别是：1.Data Mining ,2.Web Services, 3.Bayesian Networks, 4.Web Mining, 5.Semantic Web, 6.Machine Learning, 7.Database Systems, 8.Information Retrieval, 9.Pattern recognition, 10.Natural Language System。在 Citation 数据集构成的是有向图，结点是论文，边是引用的反向，由论文指向它的所有引用者。

Movie 数据集的 10 个话题分类是：1.American film actors,2. American television actors,3.Black and white films, 4.Drama films, 5.Comedy films, 6.British films, 7.American film directors, 8.Independent films, 9.American screenwriters, 10.American stage actors。Movie 中包含 movies, actors, directors, writers 四种类型的结点。结点的类型（movie、star）和话题分布都来自 wiki 上的标签。若两个结点出现在 wiki 的同一个页面中，则这两个结点之间就有一条相应的边。在本实验中结点类型不区别对待。

每个网络的数据抽象为 $G=(V,E,T_v,A)$, V 是{结点 id}的列表， E 是{边的 id, 结点 1, 结点 2}的列表， T 是{结点 id, {话题数组}}的列表， A 是网络的邻接矩阵。

每个结点的话题分布 T_v 用一个长度为 10 的 0-1 数组存储，来表示该结点是否包含对应位置的话题。

6.4 模型及参数设置

1. 扩散模型：独立级联模型。IC 模型中的传播概率： $p=0.01$ 和 0.05 两个

2. Greedy 类算法的蒙特卡罗模拟次数: Round=100

3. 初始种子集 K: 根据数据集规模不同, 对于 Citation 数据集 K 的实验取值范围为[1,50],对于 Movie 数据集是[1,100]。

4. 目标话题 T:

(1) 单目标话题, 如 $T=\{t\}$: 分别以每个数据集的每个话题为目标话题进行影响力最大化计算。本实验中将分别以每个数据集的 10 个话题作为目标话题进行一次实验。

(2) 多目标话题, 如 $T=\{t_1, t_2, t_3\}$: 任意选取三个数据集中的话题设为目标话题集, 以最大化含有这三个目标话题组合的结点为目标。本实验中对于 Citation 数据集选择 $T=\{\text{Data Mining, Machine Learning, Pattern recognition}\}$, 对于 Movie 数据集选择 $T=\{\text{American film actors, Comedy films, Independent films}\}$ 。

6.5 实验设计

1.实验环境: Dell 笔记本, Intel(R) core i3 CPU, 2.4Ghz, 内存 2G。

2.程序设计: 使用 Python 编程实现算法和实验, 使用了 NetworkX 可视化软件包和 Numpy 科学计算软件包。实现了交互式实验流程设置、数据的可视化显示。

3.实验思路: 在不同的目标话题、K、数据集运行各最大化方法, 纪录每次的 S 集合、S 激活的点个数、运行时间。最后绘制实验结果分析图表。算法比较的基线是任意选取 k 个点的 Random 方法。

6.6 结果分析

对于影响力最大化算法的评估主要有两个方面: 1.所得 S 集的影响效果 2.计算 S 的所需时间。下面将分别进行统计分析:

1.影响效果比较

a) Citations

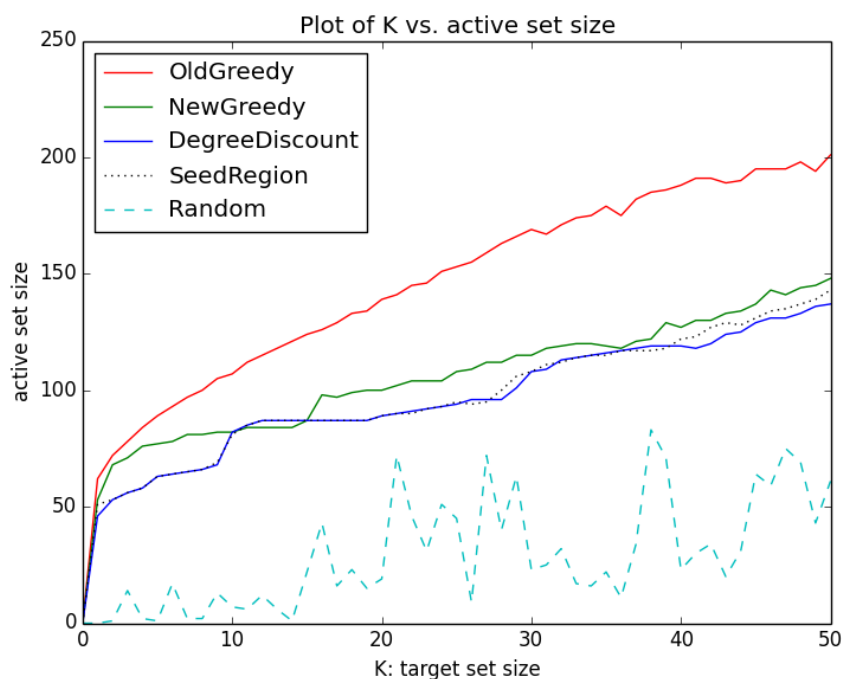


图 6-1 Citation 数据集下目标话题={Data Mining}时 TGG、TNG、TDD、TSRS 和 Random 算法的影响效果比较

b) Movie

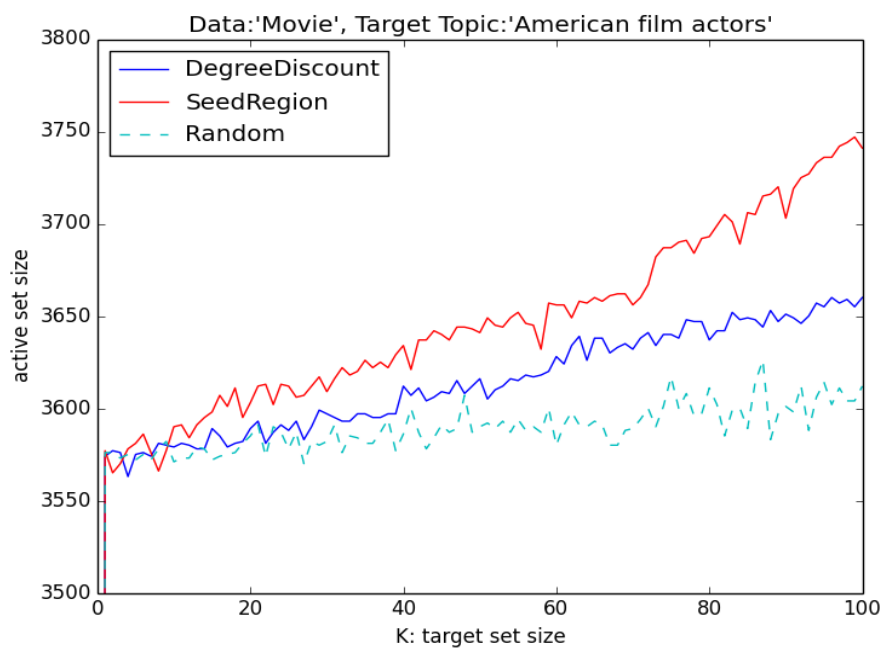


图 6-2 Movie 数据集下目标话题={America film actor}时 TDD 和 TSR 影响效果比

较

2.运算时间比较

| 数据集 | 目标话题个数 | 指标 | TGG | TNG | TDD | TSRS |
|----------|--------|---------------|---------|-------|------|-------|
| Citation | 单目标话题 | 运行时间（秒 级别） | 10000 s | 100 s | 0.1s | 0.01s |
| Movie | 单目标话题 | 运行时间 | -- | -- | 1s | 0.1s |

3.总结

从传播效果上来说 Greedy 类的 TGG 算法表现最好，但时间复杂度过大，而本文提出的启发式类的 TSRS 算法仍可获得的较好传播效果，由图 6-1 和 6-2 可看出 TSRS 的传播效果优于基于 Degree Discount 算法的 TDD，并且计算速度最快，是 TDD 速度的 10 倍。可见本文提出的 TSRS 新算法是一种有效的基于话题的影响力最大化问题算法。

第七章 总结和展望

成果:

本文通过调整经典影响力最大化算法，加入话题因素后得到了基于话题的 TGG、TNG、TDD 算法；并根据网络中影响力基于话题传播的局部性提出了基于种子选取的方法 TSRS。最后基于 ArnetMiner 平台上的学术论文引用网络数据集和 Wikipedia 上的电影合作网络数据集进行了实验，并从效果和效率两个方面分析比较 TGG、TNG、TDD、TSRS 四种算法。实验结果表明：在基于话题的影响力最大化问题上，从传播效果上来说 Greedy 类的 TGG 算法表现最好，但时间复杂度过大，而本文提出的启发式类的 TSRS 算法仍可获得的较好传播效果，并且计算速度最快。本论文的实验 TSRS 算法是一种有效的基于话题的影响力最大化问题算法。

展望:

完整的基于话题的影响力最大问题解决流程应该包含 1.话题分布建模（本论文并没有探讨话题建模部分，实验部分使用以标记好话题分布的数据集）、2.用户之间基于话题的影响力计算（传播概率本论文实验部分人为设置 IC 模型传播概率为经验值 0.01，参照了 Kempe 论文[4]中的设置）、4.传播模型选择（本论文只关注了 IC 模型）、3.影响力最大化计算。本论文主要关注于第四步中的初始结点集 S 的高效选择算法，前三步工作仍待完善。

致谢

时光荏苒，岁月如梭，四年的大学时光转瞬即逝，四年中收获了知识享受过青春。随着毕业论文的完成，大学生活也即将结束。在此仅对我身边的亲友老师表达我最真挚的感谢。

首先，感谢我的父母，慈母手中线，游子身上衣。临行密密缝，意恐迟迟归。谁言寸草心，报得三春晖。感谢我的父母对我养育之恩和谆谆教诲。

其次，感谢我的导师马军老师，在我的论文写作中给予我及时的指导和帮助，对于我的论文提出了多次实用的修改意见，不仅使我顺利的完成此次论文写作，同样使我的论文写作水平获得提升。

第三，感谢大学四年中所有身边的老师与同学，他们陪伴我度过最美好的青春年华，让我成熟成长。

最后，感谢各位评审老师感谢他们抽出宝贵的时间来阅读本文，并提出宝贵的意见和建议

参考文献

- [1] H.C.Kelman. Compliance, identification, and internalization: Three processes of attitude change. *Journal of Conflict Resolution*,2(1):51–60, 1958.
- [2] P. Domingos and M. Richardson. Mining the network value of customers. In *Proceedings of the 7th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 57–66, 2001.
- [3] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In *Proceedings of the 8th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 61–70, 2002.
- [4] D. Kempe, J. Kleinberg, and E.Tardos, “Maximizing the Spread of Influence through a Social Network,”*Proc. of ACM International Conference on Knowledge Discovery and Data Mining KDD*, 2003.
- [5]Leskovec J, Krause A, Guestrin C, Faloutsos C, VanBriesen J, Glance NS (2007) Cost-effective outbreak detection in networks. In: *Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining (KDD’07)*
- [6]A. Goyal, W. Lu, and L. V. S. Lakshmanan, "CELF++: Optimizing the greedy algorithm for influence maximization in social networks," in *WWW (Companion Volume)* 2011.
- [7]Chen YC, Peng WC, Lee SY (2012) Efficient algorithms for influence maximization in social networks.*Knowl Inf Syst* 33(3):577–601

- [8]Lin X, Mei Q, Han J, Jiang Y, Danilevsky M (2011) The joint inference of topic diffusion and evolution in social communities. In: Proceedings of IEEE international conference on data mining (ICDM' 11)
- [9]Tang J, Sun J, Wang C, Yang Z (2009) Social influence analysis in large-scale networks. In: Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining (KDD' 09)
- [10] Tang J, (2014)Models and Algorithms for Social Influence Analysis. In WWW14 international conference
- [11]L. Tang and H. Liu, "Community Detection and Mining in Social Media," Morgan & Claypool Publishers, Synthesis Lectures on Data Mining and Knowledge Discovery, 2010.
- [12]Granoveter MS. The Strength of Weak Ties. Chicago: University Chicago Press, 1974
- [13]Zhao J, Wu J. Xu K. Weak ties: Subtle role of information diffusion in online social networks, Physical Review E, 2010, 82(1):016105
- [14]S. Milgram. The Small World Problem. Psychology Today, 1967, Vol. 2, 60–67
- [15] Wei Chen, Laks V.S. Lakshmanan, Carlos Castillo. Information and propagation in social network,Synthesis Lectures on Data Management 2013 5:4, 1-177
- [16]Goldenberg J,Libai B,Muler E. Talk of the network:A complex systems look at the underlying process of word of mouth. Marketing Letters, 2001,12(3):211-223

[17]Granoveter M. Threshold Models of collective behaviour. American Journal of Sociology,1987,83(6):1420-1443

[18]J.H. Fowler and N.A.Christakis. The Dynamic Spread of Happiness in a Large Social Network: Longitudinal Analysis Over 20 Years in the Framingham Heart Study. British Medical Journal 2008; 337: a2338

[19]R. Dunbar. Neocortex size as a constraint on group size in primates. Human Evolution, 1992, 20: 469–493.

[20]李栋，徐志明等，在线社会网络中信息扩.计算机学报第 37 卷第 1 期. 2014 年 1 月

附录 1 外文文献原文

Influence and Propagation in Social Networks

Introduction

In this chapter we motivate the study of influence and information propagation by providing numerous examples. In addition, we provide some basic definitions.

1.1 SOCIAL NETWORKS AND SOCIAL INFLUENCE

Social networks have been studied extensively by social scientists for decades (e.g., see [Barnes,1954, Radcliffe-Brown, 1940, Wasserman and Faust, 1994]). Earlier studies have had to confine themselves to extremely small datasets. Enabled by the Internet and sparked by the recent advent of online social networking sites such as Facebook, LinkedIn, and Tumblr, research on social networks is witnessing an unprecedented growth due to the ready availability of large scale social network data. This has at once led to the development of many exciting applications of online social networks and to the formulation and the subsequent study of many research questions.

A rich body of such studies has come to be classified as the analysis of influence and information propagation in social networks. It is our aim in this book to outline some of the key concepts, developments, and achievements in this area, as well as studying the driving applications that underlie this research and highlight important challenges that remain open. For convenience and consistency of terminology, we will use the term social influence analysis or just influence analysis to indicate the analysis of the diffusion of information or influence through a social network.

1.1.1 EXAMPLES OF SOCIAL NETWORKS

By a social network, ¹we mean a possibly directed graph. A social network may be homogeneous, where all nodes are of the same type, or heterogeneous, in which case the nodes fall into more than one type.

Examples of homogeneous networks include the underlying graphs representing friendships in basically all of the social networking platforms (e.g., the list of “friends” in Facebook), as well as the graphs representing co-authorship or co-worker

relationships in collaboration networks.

Examples of heterogeneous networks include rating networks consisting of people and objects such as songs, movies, books, etc. Such networks can be found in media appraisal and consumption platforms such as Last.fm and Flixster. Here, people may be connected to one another via friendship or acquaintance, whereas objects (songs, movies, etc.) may be linked with one another by means of similarity of their meta data. For example, two songs may be linked since both are in the same genre or by the same artist. Similarly, two movies may be directed by the same director. In addition, links may be present between people and objects owing to their rating relationship: e.g., user Sam rating a specific model of Nikon SLR Camera produces a link between the corresponding nodes. Another example of a heterogeneous network is a scientific collaboration

network between authors, augmented with articles (that are the result of collaboration), and the venues they are published in. This network consists of three types of nodes—authors, articles, and venues—and links between nodes of the same type as well as between nodes of different types.

In the bulk of this book, our focus will be on information propagation in homogeneous networks. We briefly return to heterogeneous networks in Chapter 7.

1.1.2 EXAMPLES OF INFORMATION PROPAGATION

We begin with some concrete examples of propagations or information cascades in current online social networking sites. Consider Facebook, where a user Sally updates her status or writes on a friend’s wall about a new show in town that she enjoyed. Information about this action is typically communicated to her friends. When some of Sally’s friends comment on her update, that information is passed on to their friends and so on. In this way, information about the action taken by Sally has the potential to propagate transitively through the network. Sam posts (“tweets”) in Twitter about a nifty camera he bought which he is happy about. Some of his followers on Twitter reply to his tweet while others retweet it. In a similar

fashion, viewing of movies by users tends to propagate on Flixster and MovieLens, information about users joining groups or communities tends to spread through Flickr, adoption of songs and artists by listeners spreads through last.fm, and interest in research topics propagates through scientific collaboration networks. Is there a pattern to these propagation phenomena? What can we learn from analyzing them and how can we benefit from the results of such analysis? In this chapter, we will address these questions.

1.2 SOCIAL INFLUENCE EXAMPLES

We begin with a brief overview of several real-life stories that motivate the study of information propagation in social networks.

In a famous study published in the *New England Journal of Medicine*, Christakis and Fowler [2007] analyzed the medical records of about 12,000 patients. They extracted a real offline (as opposed to online) social network from these records, based on the relationships between the patients, including friendship, sibling, spouse, immediate neighbor, etc. Their goal was to study the relationship between non-infectious health conditions, including obesity, and one's social neighbors and understand the correlation between having obese social network neighbors and being obese oneself. Among other things, they found that having an obese friend makes an individual 171% more likely to be obese compared to a randomly chosen person. In cases of obese spouse and obese sibling, the corresponding numbers were 37% and 40%, respectively. It is to be noted that their study did not focus on causation but instead on correlation. Still, their study shows having obese social contacts is a good predictor of obesity.

The same authors, in an influential book [Christakis and Fowler, 2011] “present compelling evidence for our profound influence on one another’s tastes, health, wealth, happiness, beliefs, even weight, as they explain how social networks form and how they operate.” As specific examples, they argue that back pain spread from West Germany to East Germany once the Berlin wall came down, that suicide spreads

through communities, that specific sexual practices spread through friendship networks among teenagers, and political beliefs and convictions propagate through networks, the conviction being more intense the denser one's connections.

In the business area, a famous case demonstrating information propagation leading to commercial success, is the Hotmail phenomenon [Hugo and Garnsey, 2002]. In the early 1990s, Hotmail was a relatively unknown e-mail service provider. They had a simple idea, which was appending to the end of each mail message sent by their users the text “Join the world’s largest e-mail service with MSN Hotmail. <http://www.hotmail.com>.” This had the effect of building and boosting a brand. In a mere 18 months, Hotmail became the number one e-mail provider, with 8 million users [Hugo and Garnsey, 2002]. The underlying phenomenon was that a fraction of the recipients of Hotmail messages were inspired by the appended message to try it for themselves. When they sent mail to others, a fraction of them felt a similar temptation. This phenomenon propagated transitively and soon, adoption of Hotmail became viral.

Viral phenomena of the sort discussed above have sometimes changed lives, as in the ragsto-riches story of Ted Williams [Zafar, 2012]. He was a homeless person in Columbus, Ohio, USA, and had had many a brush with the law. He was found at a street corner in January 2011 when he was interviewed by a journalist. The interview was posted on YouTube, including details that Williams was a former voice-over artist. Within months, the video attracted 11 million views, and triggered numerous messages of support including job offers, changing his life for ever.

On November 16, 2011, a song from the soundtrack of a then upcoming Indian (Tamil) movie, called “Why this kolaveri di?” was released. By November 21, it was a top trend in Twitter. Within a week of its release, it had attracted 1.3 million views on YouTube and more than a million “shares” on Facebook, reaching and propagating through many non-Tamil speakers. It eventually went on to win the Gold Award from YouTube for most views (e.g., 58 million as of June 2012) and was featured in mainstream media such as Time, BBC, and CNN.

“Gangnam Style,” a South Korean song released in July 2012, became the first video to reach 1 billion views on YouTube as of December 21, 2012. Within one year of its first release, it has been viewed more than 1.745 billion times, even surpassing Justin Bieber’s “Baby!”

The power of online information diffusion has also been utilized by citizens responding to natural or man-made disasters. When there was a coordinated terror attack in Mumbai in November 2008, as the events were unfolding, tweets were being sent via SMS at the rate of about 16 per second, including in them such information as eye witness accounts, pleas for blood donors, location of blood banks and hospitals, etc. A Wikipedia page was up in minutes, providing a staggering amount of detail and extremely fast “live” updates. A newswire service Metroblog was set up in short order, containing 112 Flickr photos by a journalist giving a firsthand account of the aftermath. A Google map with main buildings involved in the attacks, with links to background and news stories was immediately set up. In Vancouver, Canada, in the summer of 2011, there were riots following the Stanley Cup final. Rioters, many of them teens, looted and destroyed properties in downtown. Many of them were bragging about it in social media, e.g., posing with Gucci bags in front of burning cars. This triggered a widespread reaction of disgust and was leveraged in mobilizing a cleanup effort. The amount of data made available for forensics was staggering: contrasted with 100 h of VHS footage from 1994 riots, there now was 5000 h worth of 100 types of digital video available for forensic analysis. This along with cooperation from the public enabled the police to apprehend most of the rioters.

1.3 SOCIAL INFLUENCE ANALYSIS APPLICATIONS

The study of information and influence propagation has found applications in several fields, including viral marketing, social media analytics, the spread of rumors, stories, interest, trust, referrals, the adoption of innovations in organizations, the study of human and non-human animal epidemics, expert finding, behavioral targeting, feed ranking, “friends”

recommendation, social search, etc.

Among these, viral marketing or word of mouth marketing as it is otherwise called, is a “poster” application of influence analysis. The vision behind this is to activate a small number of “influential” individuals in a social network through which a large number of other individuals can be influenced by a viral propagation. Formally, consider a social network represented as a directed graph $G = (V, E)$ with nodes V corresponding to individuals and links $E \subseteq V \times V$ representing social ties. Furthermore, suppose there is a function $p: E \rightarrow [0, 1]$ that associates a weight or probability $p(u, v)$ with every link (u, v) , representing the influence exerted by user u on v . This informally captures the intuition that whenever u performs an action, then v also performs the action after u , with probability $p(u, v)$. The idea behind viral marketing is that by getting a small set of users in V (a seed set) to use a product, for instance by giving it to them for free or at a discounted price, we can reach a much larger set of users through transitive propagation of influence.

附录 2 外文文献译文

社会网络中的影响力传播

介绍

在本章中我们将通过举例来初识影响力和信息传播,另外我们想给出一些基本的定义。

1.1 社会网络和社会影响力

近几十年来, 社会网络在社会学家中被广泛的研究(例如[Barnes,1954, Radcliffe-Brown, 1940, Wasserman and Faust, 1994])。早期的研究受限于极小的数据集, 近来, 像 Facebook, LinkedIn, Tumblr 这样的在线社会网络的进步, 使大数据研究成为可能, 社会网络的相关研究正在见证一场不可预见的突飞猛进。许多令人激动的在线社会网络应用蓬勃而出, 并伴随而来一系列新的研究问题。

一个研究热点是社会网络中的影响力传播。本书意在描述这个研究领域的关键的内容、发展和成果, 并研究营运而出的相关应用以及仍待解决的问题和挑战。

为了方便性和术语的一致性, 我们使用社会影响力分析或影响力分析来表示社会网络中的信息扩散或影响力扩散的分析。

1.1.1 社交网络的范例

我们普遍认为社交网络是一个有向图。一个社交网络可能是每个点都是同类型的同质网络, 也可能是具有多种类型结点的异质网络。

同质网络的例子包括所有社交网络平台中的基本的关系, 如 Facebook 中的朋友列表; 或者表示在合作网络中的同事, 同级作者。

异质网络的例子有同时包含人和诸如歌曲, 电影, 书籍等的对象的等级网络。这种网络在媒体评价和消费平台上有所体现, 如 Last.fm 和 Flixster。人们或许可以通过朋友或熟人认识其他人, 而对象(歌曲, 电影等)或许因为通过具有相似媒体与其他结点联系。例如, 两首歌可能因为相同的题材或者歌手而联系起来, 同样地, 两个电影也会因为同一个导演联系起来。此外, 这些联系也代表了人们和对象在网络中所属的等级关系, 例如, 用户 Sam 与尼康相机产生了一个一致的结点因而产生了一个特殊的模型。关于异质网络的另一个范例则是一个在作家之间的合作网络, 这使得他们的文章以及发表的场合都在不断增加。这个网络

包含三种类型的结点，作者，文章，以及发表地点，都有相同类型结点的联系，也有不同类型结点的联系。

在本书中，我们的重点是同质网络中的信息传播，我们将在第七章简单地介绍同质网络。

1.1.2 信息传播的例子

我们先来看一些当前在线社交网络中存在的信息传播和倾泻的例子。考虑 Facebook 中，一名用户 Sally 更新了她的状态或者在朋友的留言板记录了关于她在城镇里欣赏的一场秀。有关这个举动的一般信息会传播到她的朋友那里。当 Sally 的一些朋友评论她的更新时，信息会传播到朋友的朋友们等。通过这种方式，相关 Sally 举动的信息就具备在网络中过渡传播的潜能。Sam 在 Twitter 上贴出他买的一个很好的相机，为此他很高兴。一些他的粉丝回复了他的微博并且转发。相似的方式，Flixster 和 MovieLens 的用户倾向于在观看电影后在 Flixster 和 MovieLens 上传播信息，有关用户加入组或者社区的信息会在 Flickr 上传播，听众对于歌曲和艺术家的接受程度会在 last.fm 上传播，对于研究课题的兴趣会在科学合作网络中传播。那是否存在一个模式表示所有的传播现象？我们可以从分析它们中获得什么，我们如何从分析结果中受益？在这一章中，我们将会解决这些问题。


1.2 社交影响的例子

我们首先来简短回顾几个真实生活中的故事，这些故事促进了社交网络中对信息传播的研究。

在一篇发表在新英格兰医学期刊的著名研究中，克里斯塔基斯和福勒分析了约 12000 个病人的病历。他们从这些病历中抽取了一份真实的离线（相对于在线）社交网络，根据的是病人之间的关系，包括友情、兄妹、配偶、当前邻里等关系。他们的目标是研究非传染性的健康状态，比如肥胖和一个人的社交邻里之间的关系，并且理解社交邻里肥胖和自身肥胖之间的关联性。此外，他们发现相比于一个随机抽取的对象，一个有肥胖朋友的人会有 171% 的可能性变胖。对于有肥胖配偶或者肥胖兄弟姐妹的情况，关联指数分别是 37% 和 40%。需要指出的是，他们研究的重点并不是因果关系而是关联性。还有，他们的研究表明具有

肥胖社交关系是一个很好的肥胖的前兆。

上述两位作者，在一本很有影响力的书中展示了有力的证据，在解释社交网络是如何形成的以及如何工作的时候，证明我们对他人的品味、健康、财富、幸福、信仰，甚至体重方面的巨大影响。至于具体的例子，他们提到了柏林墙推倒后从西德传到东德的背痛，社区中传播的自杀，青少年中通过朋友网络传播的特定的性练习，网络中德政治信仰和信念的传播，一个人的联系越紧密他的信念越强烈。

在商业场景中，有一个非常著名的例子可以证明信息传播能使得的商业成功：即是 Hotmail 现象[Hugo and Garnsey, 2002]。在 1990 年早期，Hotmail 是一个相对默默无闻的 Email 服务提供商，他们有一个简单的想法，即在每一封 Email 邮件后追加一句话“加入世界最大 Email 服务 MSN Hotmail,  <http://www.hotmail.com>。”这的确对建立和发展一个品牌有巨大的影响。仅仅 18 个月后，Hotmail 变成了一个拥有八百万用户的 Email 提供商[Hugo and Garnsey, 2002]。潜在的现象则是参与使用 Hotmail 的一部分人受到追加信息的鼓励并且愿意尝试它。当他们发给别人邮件的时候，一部分人感受到了相同的激励。这种现象传播的及时快速，使得 Hotmail 像“病毒”一样风靡。