面向目标市场的信息最大覆盖算法

张伯雷[1)] 钱柱中[1)] 王钦辉[1),2)] 陆桑璐[1)]

[1)](南京大学计算机软件新技术国家重点实验室 南京 210023)

[2)](南京陆军指挥学院军队管理系 南京 210045)

摘 要 当一个企业或商家需要投放广告时,往往会先通过历史数据、个人兴趣等挖掘出可能购买自己产品的用户,即目标市场(Target Market),然后将广告信息通过电视、报纸等公共媒体的形式传递给这些目标用户,希望有更多的目标用户接受信息.然而调查显示,相比于传统大众媒体,人们更倾向于从自己认识的人那里去获取信息,因此文中考虑利用社会影响力的方式去传播广告:在社会网络中说服有限数目的初始用户,并让他们向熟识的人传播信息,期望信息可以通过级联传播覆盖尽可能多的目标用户.由于以往的信息覆盖最大化的工作集中于对全局网络的考虑,因此会忽略目标节点和全局网络之间的联系.通过数据观察可以发现,目标用户往往会由于同质性等原因而聚集在一起,因此文中提出基于聚类的 KCC 算法,算法通过对用户进行聚类分析,找出每个聚类的代表性用户,使得这些代表性节点可以影响尽可能多的目标用户,同时避免他们之间对信息覆盖的重叠.在不同的真实的数据集的实验显示 KCC 可以在大多数情况下取得优于其它常用算法的性能,尤其当种子节点数增多时,KCC 可以更多地避免节点之间信息覆盖的重叠,从而取得更好的效果;同时,KCC 只需要很短的运行时间,具有良好的可扩展性.

关键词 社会网络;目标营销;信息传播;影响最大化;社会计算

中图法分类号 TP311 **DOI 号** 10.3724/SP.J.1016.2014.00894

Maximize Information Coverage Algorithm for Target Market

ZHANG Bo-Lei[1)] QIAN Zhu-Zhong[1)] WANG Qin-Hui[1),2)] LU Sang-Lu[1)]

[1)](*State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing* 210023)

[2)](*Department of Military Training and Management, Nanjing Army Command College, Nanjing* 210045)

Abstract When an enterprise or company delivers product advertisement into the market, it might first analyze the purchase history and personal interests of users, and mine valuable customers who will potentially buy their products, which is denoted as target marketing. Compared with traditional mass media such as newspapers or TV, "word-of-mouth" diffusion in social networks is considered to be more trustful for people. So we consider delivering advertisement with social influence: by selecting a small set of people as seed nodes to spread information via social links, the information of advertisement is expected to cover as many target users as possible. Previous algorithms in influence maximization neglect the relationship between the global network and local target market, so they are not applicable in this problem. In this paper, we propose KCC algorithm to select the seed nodes. KCC clusters the nodes and finds representatives in each cluster by defining the similarities and cost of the clusters. It also reduces the overlapping of the coverage of each representative. The algorithm is conducted on different real data sets and compared with several other well-known algorithms. The results show that KCC performs better

收稿日期:2013-06-18;最终修改稿收到日期:2014-01-26.本课题得到国家自然科学基金委创新研究群体科学基金(61321491)、国家自然科学基金(61202113)、国家自然科学基金重大集成项目基金(91218302)及江苏省自然科学基金(BK2011510)资助.**张伯雷**,男,1988 年生,博士研究生,主要研究方向为复杂网络、社会网络分析.E-mail:zhangbolei@dislab.nju.edu.cn;zblhero@gmail.com.**钱柱中**(通信作者),男,1980 年生,博士,副教授,主要研究方向为云计算、社会网络分析.E-mail:qzz@nju.edu.cn.**王钦辉**,男,1985 年生,博士研究生,主要研究方向为认知无线电、网络经济学.**陆桑璐**,女,1970 年生,博士,教授,博士生导师,主要研究领域为分布式计算和复杂网络.

in most cases, especially when the seed set size grows larger. Moreover, it only needs a low running time so is scalable for large networks.

Keywords　social network; target marketing; information diffusion; influence maximization; social computing

1　引　言

近年来，互联网和在线社会网络的快速普及为企业与生产商提供了丰富的用户信息，从而促进了目标营销(Target Marketing)的发展. 商家首先分析用户的分布特征、购买历史、个人爱好等，预测出用户可能购买的产品，然后根据消费者需求的异质性进行市场细分(Market Segment)，挖掘出可能消费其产品的目标市场(Target Market)，并将营销策略集中于目标用户，希望可以获取更大的收益. 如何进行市场细分已经得到了长期关注[1-3]，然而，即使可以准确预测到可能消费的目标用户，还需要设计可信的方案让用户接受产品的广告信息，并进行消费. 调查显示，相比于传统的大众媒体用广播式的方法传播信息，人们更倾向于相信自己认识的人，并且从熟人或者朋友那里去获取信息. 因此，利用社会网络传播信息，采用病毒营销的方法使得广告信息可通过“口口相传”的模式去影响目标用户已成为一种有效的营销模式[4-6].

通过社会信息传播影响用户，在初始用户有限的前提下，使信息覆盖尽可能多的用户，也被称作影响力最大化问题(Influence Maximization)，是社会网络领域一个被长期关注的问题. Domingos 和 Richardson[7-8]首先提出了这种社会营销的模式. Kempe 和 Kleinberg 等人[9]将信息传播的过程建模为离散的随机过程，从而可以将影响力最大化问题归约为 NP 难的集合覆盖问题，并给出了与最优解比为$(1-1/e)$的近似算法. 由于该算法需要使用蒙特卡洛方法估计节点的影响力的期望值，因此需要较长的运行时间，已经有很多工作研究如何提高算法的效率与性能[10-18]，并取得了较好的效果. 事实上，在一个社会网络中，并不是所有的用户都对商家有价值，通过对市场的细分，可以区分出有价值的用户，面向这些目标用户有针对性的传播广告可以获取更大的收益.

为此，本文研究了面向目标市场的信息最大覆盖问题，在一个社会网络中，希望广告或者信息去覆盖特定的已知的目标人群，而不是社会网络中一些无关的用户. 如图 1 所示，在这个社会网络中，白色节点代表无关的用户，黑色节点代表目标用户，则该问题的目标是让尽可能多的黑色节点去接受信息. 假设目标市场是已知的，为全局社会网络的一个子集，本文研究如何从全局的网络中找出 k 个对目标节点有影响力的种子节点(Seed Nodes)，并从这些种子节点开始传播信息，使得信息对目标用户的覆盖达到最大. 由于这些有影响力的节点往往并不属于目标节点，因此需要对全局的网络进行搜索. 然而，虽然可以证明贪心的爬山算法具有较好的性能保证，但是由于该算法运行效率较低，并不适合于大规模的社会网络. 而一些已有改进的算法，由于忽略了目标节点与全局网络之间的联系，因此不能在该问题中取得较好的效果.

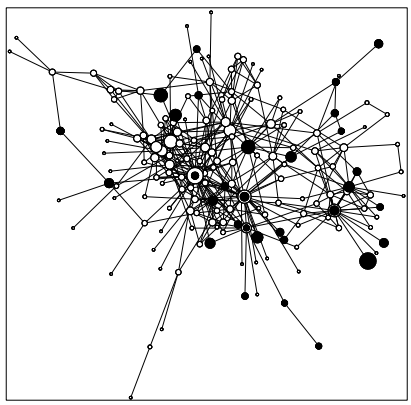

图 1　社会网络与目标市场

通过对数据的观察，发现目标用户往往存在同质性(Homophily)的现象，即相似的用户更有可能聚集起来[19-20]. 因此，本文利用这个特性而提出了 KCC(K Cluster Coverage)算法从全局网络找出这些影响力节点：首先对目标节点进行聚类，然后从每个聚类找出对其中目标市场最具影响力的节点. 可以证明所有对目标市场产生较大影响力的节点都会被限制在有限的子空间内，利用广度优先的方法获取这个子空间，可以减少算法的搜索空间. 同时，KCC 算法还定义了聚类的相似度与代价，使得每个聚类中心可以具有尽可能高的期望影响力，同时还

避免了聚类之间对信息覆盖的重叠.尤其是当种子节点数增多时,KCC 算法会具有更好的效果.本文的主要贡献有:(1)提出了用社会影响力的方式进行目标营销,并证明了该问题的复杂性;(2)证明了对目标节点有影响力的种子可以被限制在全局网络的一个有限子空间内,并提出了基于聚类的 KCC 算法,定义了聚类算法中的相似度和代价,可以高效地找出有影响力的节点;(3)在不同的真实数据中对算法进行了实验,分析并验证了所提出的算法的有效性.

本文第 2 节介绍目标营销和社会影响力的背景知识和相关工作;第 3 节介绍信息传播模型和该问题的形式化的定义以及问题的复杂性;第 4 节介绍本文的主要算法,通过减少采样空间提升算法效率,并定义聚类的相似度和代价;第 5 节介绍在不同数据集上进行的有效性和扩展性的实验和分析;最后一节是总结和未来工作.

2 相关工作

2.1 目标营销

互联网的广泛普及使得商家可以从中获取丰富的社会数据,尤其是其中关于个体用户的信息.通过分析和挖掘这些数据,可以预测出用户对哪些产品感兴趣.对一个企业或者生产商来说,当知道哪些用户更有可能消费自己的产品时,就可以通过目标营销的方式将营销资源集中到这些消费者群体(目标市场)上,从而获取更高的利润[21].随着大数据时代的来临,生产商可以获取更丰富的个体信息,并从中挖掘出更精确的对用户需求的判断[22-24],并对目标营销的预测结果做出评估.

然而,目前大多数关于目标营销的研究工作都集中于如何从已有数据中去挖掘潜在的目标人群,而很少考虑如何利用用户和用户之间的联系进行营销.调查显示,相比于传统大众媒体的广告营销,如报纸、电视等,人们更倾向于从自己所熟识的人那里去获取信息.而在目标人群中,由于同质性等原因,信息可能具有更高的传播概率[19-21].因此,本文考虑通过社会网络中的病毒营销影响目标市场进行消费,假设通过挖掘已经获得了目标人群的集合,本文研究如何利用社会影响力的方式去进行广告和信息的传播,目标是使得信息可以覆盖最多的目标人群.

2.2 社会网络信息传播

当信息通过人与人之间的接触进行传播,更容易被他人所接受[25].而在线社会网络的发展使得信息在社会网络中的传播变得更加普适和广泛[26-28],已经有很多工作研究如何从已有的行为数据估算出信息在社会网络中传播的概率与可信度[29-30].由于社会影响力在社会网络中的信息传播中扮演着非常重要的作用,因此关于如何最大化信息传播覆盖也已经得到了广泛的关注.Kempe、Kleinberg 和 Tardos 中将影响力传播的过程建模为离散的随机过程[9],并证明影响力最大化的问题是 NP 难的,同时,文献[9]还根据信息传播的次模性质提出具有性能保证的贪心算法.然而,该算法需要通过蒙特卡洛方法计算每个节点的期望影响力,因此具有较低的执行效率.

为了提高算法的执行效率,并同时保证结果的性能,很多文章研究如何设计新的算法找到影响力节点.Leskovec 等人在文献[10]中提出了一种 CELF 的算法,算法基于“懒惰向前”(lazy-forward)的思想:即在贪心算法的每一个步,之前计算过的节点的期望影响力肯定不会增加,因此可以减少对节点的期望影响力的计算,算法的执行效率可以提升大约 700 倍.在文献[11]中,Chen 等人证明了计算节点的期望影响力的是 $\sharp$P-hard 的,因此只能用近似的算法去估计节点的期望影响力.同时,文献[11]还优化了 CELF 算法,提出在算法的第一步对社会网络做多次的随机模拟,可以避免对每个节点的影响力的重复计算.Kimura 等人在文献[12]中提出了用“最短路径模型”去模拟社会影响力,即每个节点只能通过最短路径影响其它的节点,在这个假设下,算法仍然可以得到较好的结果,同时,可以大大促进计算的效率.文献[13]进一步优化了 CELF 算法,并提出了近似的 CELF++.文献[14]提出了 IRIE 算法,首先估算出每个节点的影响排名,然后用线性的方法估计出节点的增益影响力.

本文期望信息能够覆盖的节点为社会网络中所有节点的一个子集.因此,以往的影响力最大化的工作并不一定合适:如果进行全局的搜索,虽然可以取得较好的效果,但是算法的开销太大;而一些局部最优的算法,往往会忽略目标节点与全局网络之间的关系,因此也不一定能够有很好的性能,并不适用于该问题.

3　信息传播与面向目标最大覆盖问题

3.1　信息传播模型

一般将社会网络抽象为一个有向图 $G=(N,E)$，其中 N 代表节点集合，E 代表节点之间的有向边集. 这里用独立级联模型(Independent Cascade Model)[9]对信息在社会网络中的传播过程进行建模，在这个模型中，节点被分为活跃态和非活跃态：若一个节点已经接受了一种信息，则称之为活跃节点，否则为非活跃节点. 在一个信息级联(Information Cascade)的过程中，信息从初始的活跃节点集 S 开始，通过离散的步骤进行传播：在第 t 步时，假设刚刚被激活的节点集为 S_t，则 S_t 中的每个节点尝试去激活它的所有非活跃的邻居节点，并且以概率 p_e 成功激活，其中 p_e 为边 $e\in E$ 上的独立概率. 则由 S_t 在第 t 步激活的节点集合为 S_{t+1}. 这个过程一直持续，直至在某一步的时候，没有新的节点被激活，即 $S_t=\varnothing$. 则最终活跃的节点为每一步被激活的所有节点的并集.

3.2　问题定义

本文研究基于目标营销的信息覆盖最大化方法. 首先定义目标市场.

定义 1.　目标市场(Target Market). 目标市场 T 是一个用户节点的集合，为全局用户 N 的子集，一般是根据用户的购买历史、爱好兴趣等从全局市场中挖掘出的可能消费某种产品的用户.

本文假设目标用户已知，研究如何从全局网络中选取固定数目的影响力节点，并通过社会影响力的方式去传播信息或者广告，使得在一定的限定条件内，信息可以覆盖最多的目标人群. 定义其为面向目标的影响力最大化问题.

定义 2.　面向目标市场的信息最大覆盖问题. 假设目标人群的集合为 T，显然有 $T\subseteq N$，该问题是从全局的节点集 N 中选取有限的 k 个节点 S 作为初始的活跃节点，这 k 个节点并不一定是目标用户，通过这些节点传播信息，使得信息可以覆盖尽可能多的目标人群. 用 $\sigma(S)$ 表示初始集合为 S 时信息可以覆盖 T 中的人群的期望数目，则面向目标市场的信息最大覆盖问题的目标为

$$\max_{S\subseteq N}\sigma(S) \quad \text{s.t.}\ |S|=k \tag{1}$$

3.3　问题复杂性与贪心算法

根据定义，首先证明这个问题的复杂性：对该问题的一个特殊形式，即当目标市场为社会网络的所有节点时，面向目标用户的信息覆盖等价于面向全局的影响力最大化问题，而面向全局的影响力最大化问题可以被归约到 NP 完全的集合覆盖问题[9]，因此：

定理 1.　面向目标市场的信息最大覆盖问题是 NP 难的.

由于该问题的复杂性，本文首先探索具有性能保证的近似策略. 对于全局的影响力最大化问题，Kempe、Kleinberg 和 Tardos 证明了影响力的传播具有次模(submodular)的性质[9]：即对任意节点 v，若 S 是 Q 的子集，则 $\sigma(S\cup\{v\})-\sigma(S)\geq\sigma(Q\cup\{v\})-\sigma(Q)$. 由于次模性质，使用爬山的贪心算法依次选取具有最大增益影响力的节点可以得到 $(1-1/e)$ 的性能保证. 对于面向目标节点的信息最大覆盖问题，同样可以证明以下定理(证明过程见附录 1).

定理 2.　面向目标市场的信息传播覆盖 $\sigma(\cdot)$ 具有单调次模(submodular)的性质. 假设使用贪心算法每次选取期望的增益影响力最大的节点得到的节点集为 S，具有最优覆盖节点集为 S^*，则有 $\sigma(S)\geq(1-1/e)\cdot\sigma(S^*)$.

然而，精确计算节点的期望的影响力是一个 #P 难的问题，虽然可以通过蒙特卡洛的方法对网络进行多次的随机模拟估计影响力的近似值，但是对于大规模的社会网络，多次随机模拟的运行开销太大. 而对面向目标市场的信息覆盖，可以找出这个问题具有的特性，探索更高效的算法，并且希望新的算法可以尽可能地逼近具有性能保证的贪心算法.

4　面向目标市场的影响力最大化算法

4.1　数据观察

本文考虑利用目标市场的特性找出影响力最大的节点：在一个社会网络中，具有相似特性的用户往往存在同质性(Homophily)的现象，即相似的用户更有可能聚集在一起[30-31]. 同质性形成的原因可能是用户会选择和相同兴趣的人建立朋友关系，也可能用户会由于社会影响的关系去和朋友保持一致. 而在进行社会网络划分中，得到的目标用户往往会具有共同的兴趣或者特性，因此他们很有可能会形成较为聚集的结构. 为了验证这个现象，本文从微博

(http://weibo.com①)中获取了用户的基本信息和他们之间的关系网络观察这种现象,这个数据集包括 9341 个用户节点和 328363 条用户之间关注关系的有向边,以及这些用户的简单信息,如所在地区、标签、爱好等.假设目标用户为在相同地区、或者有相同标签的用户,并分别观察这些网络的集聚系数(Clustering Coefficient),统计结果见表 1.

表 1 不同网络的集聚系数

数据集	标签	地区	集聚系数
微博			0.427
微博	足球		0.451
微博		上海	0.649
微博	IT		0.446

从表 1 中可以看出,虽然选取的目标节点只是全局节点的一个子集,但是这些目标节点构成的子图仍然具有很高的集聚系数,即有共同爱好或者特征的人们更容易建立连接,并聚集在一起.由于微博具有较强的媒体属性,因此这个特性表现的更加明显.

为了找出最具影响力的 k 个种子节点,一方面希望这些种子节点既可以连接到很多节点,会引发较大的级联效应,同时,也希望这些节点之间尽量分散,这样可以避免每个节点对信息覆盖的重叠.基于以上考虑,本文提出了基于 K-medoids 的 KCC(K Cluster Coverage)算法找到影响力最大的节点,通过对节点进行聚类,使得每个聚类之间有相对比较独立的信息覆盖,同时每个聚类的代表性节点还具有较高的影响力.K-medoids 是数据挖掘领域一种常用的聚类方法,算法首先随机选取 k 个 medoid 对象作为聚类的中心,并将每个节点分配给最近的聚类,然后不断地更新聚类中心以获取更广泛的信息覆盖.K-medoids 以节点为聚类中心,因此对异常数据不敏感,对于一些偏离的数据,也可以取得较好的效果.

4.2 基于广度优先搜索的取样空间选取

在 K-medoids 聚类算法中,需要从全局网络随机地选取 k 个节点作为聚类的中心,并不断地优化,由于这 k 个节点并不一定是目标节点之一,而搜索全局节点则效率太低.考虑到只需要影响目标节点,因此可以在希望保证算法性能的前提下,用减少取样空间(sample space)的办法提升算法效率.本文首先证明,所有对目标市场有较大覆盖的节点与目标市场的距离都在一定的范围内.

定理 3. 假设活跃节点仅能通过最短路径激活非活跃节点,则当每条边上的激活概率为相等的 p 时,只要从目标节点集 T 进行深度为 $t>\log_p\left(\frac{1}{T}\right)-1$ 的广度优先搜索得到节点集合 U,则对目标市场影响力最大的节点都在集合 U 中.

证明. 当活跃节点的集合为 S 时,定义一个新的节点 v 的期望影响力为 $\delta(v)=\sigma(S\bigcup\{v\})-\sigma(S)$.则对任意节点 w,若 $w\in V\backslash U$,它的期望影响力为

$$\delta_S(w)=\sum_{v\in T}(1-pp(S,v))pp(w,v) \qquad (2)$$

其中 $pp(S,v)$ 是活跃的集合 S 可以激活节点 v 的概率.假设每个节点只能通过和另一个节点之间的最短路径之一去激活,由于 w 不在集合 U 中,所以 w 到 T 中节点的最短路径的长度小于 $t+1$.因此有

$$\delta_S(w)\leq\sum_{v\in T}(1-pp(S,v))p^{t+1}.$$

用 $u_{\max}$ 表示当前在集合 U 中具有最大影响力的节点,即 $u_{\max}=\arg\max_{u\in U}(\sigma(S\bigcup\{u\})-\sigma(S))$.由于 T 中的节点也属于集合 U,所以对 T 中的单个节点具有最大影响力的节点就是其本身,即 $v_i=\arg\max_{v\in T}pp(v,v_i)$,且 $pp(v_i,v_i)=1$.因此,对 U 中影响力最大的节点 $u_{\max}$,有

$$\begin{aligned}\delta_S(u_{\max})&=\sum_{v\in T}(1-pp(S,v))pp(u_{\max},v)\\&=\frac{1}{|T|}\sum_{v_i\in V}\sum_{v\in V}(1-pp(S,v))pp(u_{\max},v)\\&\geq\frac{1}{|T|}\sum_{v\in V}(1-pp(S,v)).\end{aligned}$$

因此,当 $t>\log_p\left(\frac{1}{T}\right)-1$ 时,有 $\delta_S(u_{\max})\geq\delta_S(w)$,即在有活跃节点 S 时,U 中影响力最大的节点是全局最优的. 证毕.

由定理 3 可以看出,只要从目标节点出发,做限定步骤内的广度优先搜索,则必然会覆盖所有的有影响力的节点,可以保证算法的性能.同时,由于社会网络往往比较稀疏,而且信息传播的概率也较小,因此只需要通过搜索很小的步骤 t,就可以得到合适的采样空间,保证覆盖所有可能有影响力的节点.

4.3 面向目标节点的 KCC 信息覆盖算法

当利用 4.2 节的广度优先方法获取一个较小的搜索子空间之后,再以这些节点为采样空间进行聚类,由于聚类之间有相对独立的信息覆盖,因此在每个聚类中找出代表节点,作为信息覆盖的种子节点

① Weibo. http://weibo.com

集. 在聚类算法中，首先从采样空间随机地选取 k 个节点作为每个聚类的中心，然后将其它每个节点分配到与之相似度最高的聚类. 由于希望信息可以覆盖尽可能多的目标节点，因此只需要考虑目标节点的分配，而不需要对无关节点进行聚类.

在聚类的每一步，需要计算目标节点与聚类之间的距离，本文定义节点与聚类的相似度为聚类中心对节点的影响力的期望，由于计算影响力的期望是 #P 难的问题，因此采用近似的模型估算该影响力. 定义任意一个节点 v 到它所属的聚类 C_i 的相似度为

$$sim(v,C_i)=\sum_{P\in SIP(v,C_i)}\prod_{e\in P}p_e,$$

其中，$SIP(v,C_i)$ 为节点 v 与聚类 C_i 的中心节点 m_i 之间的所有点不相交的最短路径(Shortest Independent Paths)的集合. 为了得到这个集合，可以从节点 v 出发进行广度优先搜索，并从具有传播概率最大边开始进行选取，同时记录到达每个节点时这条路径上的传播概率，若在第 r 步搜索得到聚类中心 m_i，则停止搜索，并找出所有在 $r-1$ 步搜索到的节点也和该中心节点相连的节点，从 v 经过这些节点到 m_i 的所有不相交的路径即为 $SIP(v,C_i)$. 由于每条边上的传播概率是独立的，因此，一条路径上的信息传播概率为每条边上的概率 p_e 的乘积. 由于一般边上的概率较小，因此可以设定阈值 θ，即只保留传播概率大于 θ 的路径.

若一个节点到多个聚类的最短路径的长度相同，则在其它聚类中也分别加入对这个节点的期望影响力，则节点 v 和一个没有所属关系的聚类 C_j 的相似度为

$$sim(v,C_j)=(1-sim(v,C_i))\sum_{P\in SDP(v,C_j)}\prod_{e\in P}p_e,$$

其中 C_i 为节点 v 被分配到的聚类，而 C_j 为到节点 v 的距离与 C_i 相同的另一个聚类. 例如，在图 2 中，节点 v 到聚类 C_1 和聚类 C_2 的距离都为 1，虽然 v 已经被分配到聚类 C_1 中，但是在计算总的代价时，也需要包括 C_2 对 v 的影响力.

通过相似度计算，可以将每个节点分配至与其距离最近的聚类. 当对所有节点进行分配之后，需要计算生成的聚类的总代价，并找出最高的聚类方案. 聚类的总代价为每个聚类的可以产生的期望影响力之和. 形式化地定义选择形成聚类的代价为

$$cost(S)=\sum_i\sum_{v\in S}sim(v,C_i),\ i=1,2,\cdots,k.$$

通过用随机的节点替换聚类的中心，并计算新的代价，直至取得最高的代价为止，所找到的 k 个聚类中心即为种子节点. 具体的算法如算法 1.

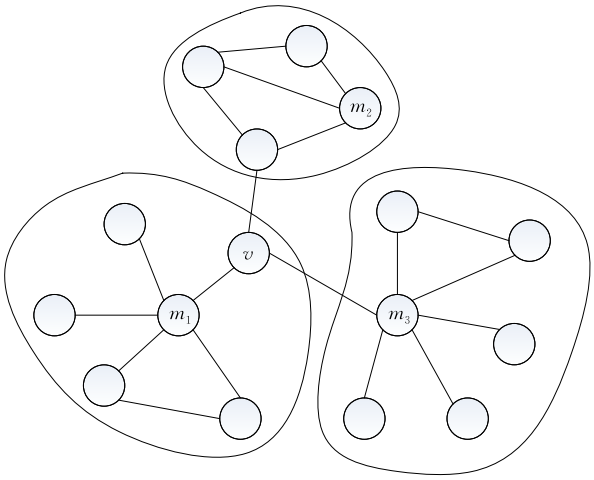

图 2　聚类中的边缘节点

算法 1.　KCC 信息最大覆盖算法.

输入：有向图 $G=(N,E)$，目标节点集 T，每条边 e 的信息传播概率 p_e

输出：种子节点集 S

```
1.  U←T, Q=∅        /* BFS 获取采样空间 U */
2.  FOR v in T do
3.    Q←{v}, v.traversed←True;
4.    FOR i←1 to t do
5.      FOR u in G.adjacent(Q) do
6.        Q←Q∪u, U←U∪u;
7.        u.traversed←True;
8.  S←U 中随机选取 k 个节点;   /* 聚类算法 */
9.  WHILE S 还会发生变化
10.   FOR S_i in S do
11.     FOR o∈U
12.       IF not o∈U
13.         IF cost(o,S_{-i})>cost(S)
14.           替换 o 与 S_i;
15. RETURN S
```

其中 $\boldsymbol{S}_{-i}$ 为 S 中除了 S_i 以外其它所有节点构成的向量.

复杂性分析：从算法 1 中可以看出，运行时间主要和 8～13 行的选取聚类中心有关，假设循环的次数为 r 次，则算法主要部分的复杂性为 $O(rk|U|)$. 对于大多数情况，r 和 k 都是较小的常数，而 $|U|$ 是通过较小的步 t 得到的采样空间，主要和目标节点的个数有关，因此，该算法可以在多项式内高效地执行.

5 实验和评估

5.1 实验设置

本文在 3 个真实的数据集上进行了实验. 数据集 1 为从微博(http://weibo.com)上爬虫得到的社会网络数据集,其中节点为微博中的用户,有向的边代表用户之间的关注关系,该数据集包含 9341 个节点和 328363 条有向边;数据集 2 为 ca-GrQc[①],是学术合作网络,节点代表单个的作者,边代表两个作者的合作关系,该数据集共包含 5242 个节点和 28980 条无向边(可以看成两条有向边);数据集 3 为 com-Youtube,是 Youtube 上用户关系网络,节点代表单个的用户,边代表两个用户之间的好友关系,该数据集共包含 1134890 个节点和 2987624 条无向边. 这 3 个数据集代表的社会网络都具有小世界、无尺度等复杂网络的特征.

在挖掘目标节点时,对于微博数据集,实验在爬虫的过程中获取一些用户的基本信息,如地理信息、标签信息等,并选取在某个特定区域或者有共同标签的用户,这些目标信息往往会使得广告的投放更加有效. 本实验分别选取了在地区"上海"的目标用户和具有"足球"标签的用户作为目标用户;对于 com-Youtube 和 ca-GrQc 数据集,由于其包含的信息有限,因此实验随机选取了一部分节点作为目标节点,从实验结果可以看出,算法在任意选取目标节点的数据集上都有较好的效果.

KCC 算法分别和以下的算法进行比较.

(1) 全局贪心的 CELF 算法(Global CELF). 由定理 2 可知,目标用户的信息最大覆盖具有次模的性质,因此该算法所得到的信息覆盖至少为最优解的$(1-1/e)$;同时使用懒惰向前(Lazy Forward)的方法,避免了对节点影响力的重复计算.

(2) 局部贪心的 CELF 算法(Local CELF). 该算法只搜索目标节点和目标节点之间的连接. 由于社会网络中目标节点以外的结构往往也对信息的传播扮演很重要的作用,甚至可能对目标市场产生最大信息覆盖的节点并不是目标节点之一,因此该算法并不一定能够取得较好的效果.

(3) Degree Discount 算法. 该算法基于文献[12],计算在每个节点对邻居节点的节点度. 当某个节点被计入种子集合,则从其邻居节点的影响力中去除该节点所占的比例.

(4) IRIE 算法. 基于文献[14],算法首先计算每个节点的影响力排名,并采用线性的方法估算出节点的增益影响力,从而选出种子节点集.

(5) 随机(Random). 这个算法随机地从目标节点中选取一些节点.

5.2 信息传播覆盖

首先观察不同算法所得到的种子节点集对目标市场的信息覆盖,即每个节点的影响力. 由于这个值的计算是 #P 难的,因此采用蒙特卡洛的方法估计节点的影响力:对网络中的每条边 e 进行概率为 p_e 的抛硬币,并计算节点集的连通分量的节点数目,将这个过程重复 $R=1000$ 次,然后计算平均值,即为估计的影响力. 不失一般性,假设信息在每条边上的传播的概率为相同的 0.05 或 0.01. 算法分别选取影响力节点的个数为 1～15. 分别在不同的数据集上进行模拟实验,并记录各个算法产生的种子节点集的平均覆盖大小. 图 3～图 6 为各种算法在不同数据集上的模拟实验,其中 x 轴为所选取的种子节点的数目,y 轴为对图进行蒙特卡洛模拟之后计算的平均信息覆盖.

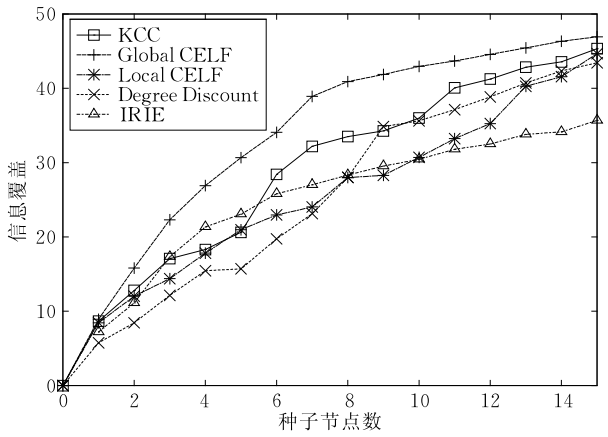

图 3 微博数据集,目标用户为具有"足球"标签的用户,目标节点个数为 207,传播概率为 0.01,$\theta=0.005$

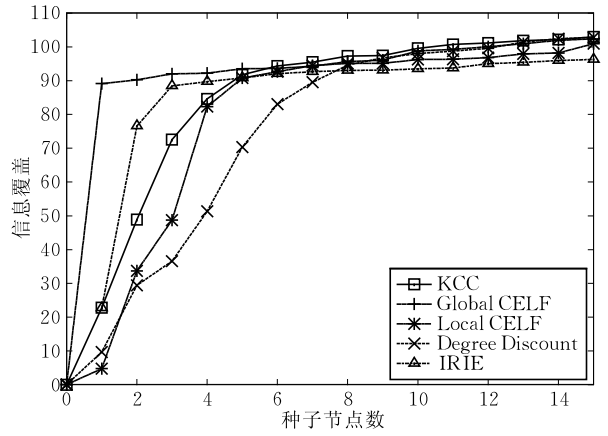

图 4 微博数据集,目标用户为所在地区为"上海"的用户,目标节点个数为 350,传播概率为 0.01,$\theta=0.005$

① SNAP. http://snap.stanford.edu/data/index.html

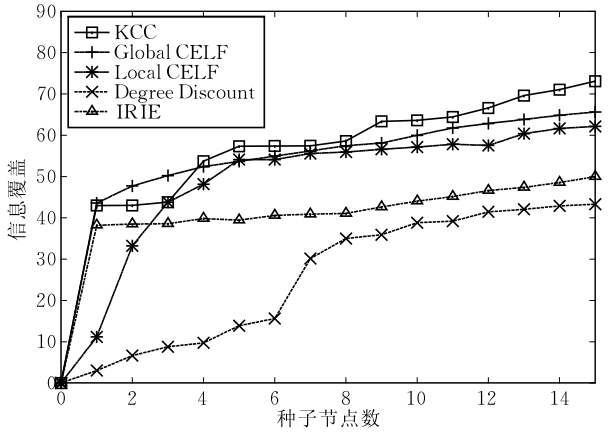

图 5　ca-GrQc 数据集，随机的从网络中选取了 1021 个节点，传播概率 0.05，$\theta=0.005$

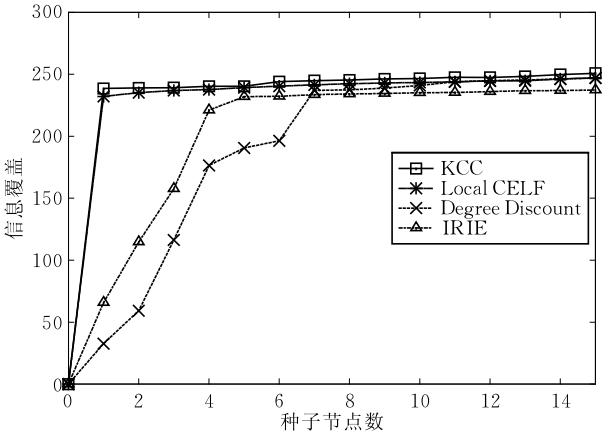

图 6　com-Youtube 数据集，随机的从网络中选取了 8022 个节点，传播概率 0.01，$\theta=0.005$

图 3、图 4 为在微博数据的模拟实验. 在图 4 中，由于目标节点的集聚系数较高，即他们之间的连接更多，在这种情况下，随机的复杂网络中很有可能存在一个较大的连通分量，因此很容易通过少量的活跃节点就引起一个较大的信息级联，在种子节点较少的时候就激活很多的节点. 从图中可以看出，在大部分时间里，Global CELF 算法都可以取得最好的效果，这是由于该算法具有性能保证，可以取得不低于最优解$(1-1/e)$的效果. 当选取的种子节点较少时，由于 KCC 算法采用了对节点影响力的估算，因此并不一定会取得很好的效果，但是当种子节点集增大时，KCC 算法可以尽量避免信息覆盖之间的重复，同时，KCC 选取的每个节点都有较大的影响力，因此会取得较好的效果，甚至会在某些时候优于 Global CELF 算法；而 Local CELF 算法并没有考虑到目标节点以外网络的结构对信息传播的影响，因此取得的效果和目标节点所处的位置和它们之间形成的结构有关，在图 3 中，目标节点之间的结构较为稀疏，因此 Local CELF 并不能取得很好的效果. 在图 4 中，节点之间的连接较为密集，目标节点之外的结构对信息传播的影响较小，Local CELF 也可以取得较好的效果. Degree Discount 算法会计算每个节点在一步之内的影响力期望，并会去除已加入种子节点集的节点的影响；尤其是当目标节点之间的连接较少时，该算法由于只考虑距离为 1 的影响力，会忽略很多的信息传播路径，因此在图 4 中 Degree Discount 并不能取得较好的效果. IRIE 算法会用线性的方法估算当一个节点加入种子节点集之后对其它节点的影响力的干扰，然而在面向目标节点的信息覆盖时，由于有空闲节点的干扰，因此在计算增益影响力时并不十分准确，在节点增多时，可以获取的增益影响力比较有限.

为了验证 KCC 算法在任意选取的目标节点集上都具有较好的效果，实验还从 ca-GrQc 数据集和 com-Youtube 数据集中分别随机地选取了一部分节点作为目标节点，并比较了各个算法在这两个数据集上的性能. 由于 Global CELF 需要的计算开销太大，因此并没有在 com-Youtube 数据集上运行该算法. 从图 5、图 6 中可以看出，即使对于任意选取的目标节点，KCC 也可以取得较好的效果，并且会在种子节点数增多时取得优于 Global CELF 的信息覆盖. 对于 ca-GrQc 数据集，由于节点之间的传播概率较大，因此目标节点之间的结构较为紧密，而随机选取的节点在整体上较为分散，因此更有利于 KCC 算法进行聚类，可以取得非常好的效果. 在图 5 中，目标节点比较分散，因此 Local CELF 算法取得的效果较为一般；在图 6 中，目标节点集存在一个较大的连通分量，因此 Local CELF 可以取得较好的性能. 对于 Degree Discount 算法，由于这两个数据集都具有相对比较稀疏的结构，因此估算单个节点的信息传播期望和实际差距较大，在有些时候只能取得 KCC 算法的 50%.

5.3　可扩展性

本节统计不同算法在选取 15 个种子节点时的运行时间，结果如图 7 所示.

从图 7 可以看出，虽然 Global CELF 算法选取的种子节点可以取得很好的信息覆盖，但是该算法需要通过蒙特卡洛的方法估算每个节点的期望影响力，对于规模较大的社会网络，运行时间非常长，因此不适宜应用于面向目标节点的信息覆盖；com-Youtube 数据集包含网络节点较多，因此并没有运行该算法. Local CELF 算法的运行时间和目标节点的数目有关，同样会由于目标节点的增多而产生明

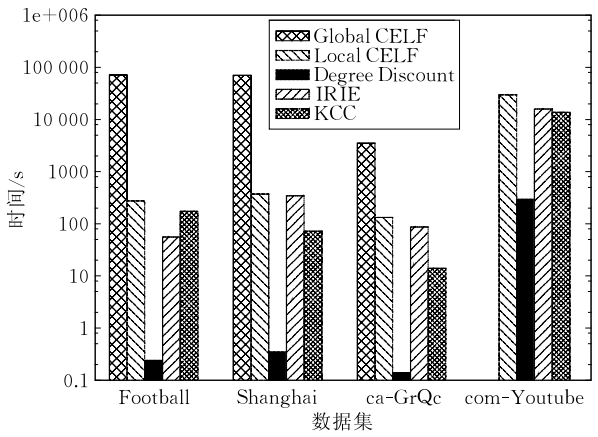

图 7 运行时间

显的计算开销的增大. Degree Discount 算法运行时间较快,但是计算的结果并不稳定,在目标节点之间较密集的时候并不能取得很好的效果. IRIE 算法也是一个较为高效的算法. 从图 7 中还可以看出,KCC 算法在取得较好的实验结果的同时,其运行时间也在一个可以接受的范围内;对于更大规模的网络如 com-Youtube,该算法的运行时间也不会有过快的增长,因此具有较好的可扩展性.

6 总 结

本文采用社会影响力传播的方式进行社会营销,在目标用户已知的情况下,研究如何从社会网络中选取小部分的初始用户传播信息,使得信息可以覆盖最多的目标用户. 通过对真实数据的分析,可以看出,目标用户往往会形成聚集的结构,因此本文提出了 KCC 算法找出种子节点,首先通过减少采样空间提升执行效率,并证明了最具影响力的节点都在该空间内;进一步定义了聚类中的相似性和聚类代价,从而准确找出具有影响力的种子节点,同时可以尽量避免节点之间对信息覆盖的重叠. 在真实数据集上的实验验证了该算法既可以找到有影响力的节点,同时具有较好的效率.

在未来的工作中,我们将研究社会网络的聚类结构(社区)对社会信息传播的影响;并通过不同的数据集研究不同目标用户在社会网络中所占据的位置以及它们对网络结构的影响;同时还会分析当多种相关信息同时传播时如何采取合理的传播方案.

参 考 文 献

[1] Rossi P E, McCulloch R E, Allenby G M. The value of purchase history data in target marketing. Marketing Science, 1996, 15(4): 321-340

[2] Allenbya G M, Rossi P E. Marketing models of consumer heterogeneity. Journal of Econometrics, 1998, 89(1): 57-78

[3] Farahat A, Bailey M C. How effective is targeted advertising? //Proceedings of the 21st International Conference on World Wide Web. Lyon, France, 2012: 111-120

[4] Granovetter M. Threshold models for collective behavior. The American Journal of Sociology, 1978, 83(6): 1420-1443

[5] Datta S, Majumder A, Shrivastava N. Viral marketing for multiple products//Proceedings of the 10th IEEE International Conference on Data Mining. Sydney, Australia, 2010: 118-127

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

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

[8] Richardson M, Domingos P, Glance N. Mining knowledge-sharing sites for viral marketing//Proceedings of the 8th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Edmonton, Canada, 2002: 61-70

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

[10] Leskovec J, Krause A, Guestrin C, et al. Cost-effective outbreak detection in networks//Proceedings of the 13th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. San Jose, USA, 2007: 420-429

[11] Chen W, Wang Y, Yang S. Efficient influence maximization in social networks//Proceedings of the 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Paris, France, 2009: 199-208

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

[13] Goyal A, Lu W, Lakshmanan S L V. CELF++: Optimizing the greedy algorithm for influence maximization in social networks//Proceedings of the 20th International Conference Companion on World Wide Web. Hyderabad, India, 2011: 47-48

[14] Jung K, Heo W, Chen W. IRIE: Scalable and robust influence maximization in social networks//Proceedings of the 12th International Conference on IEEE Data Mining (ICDM). Brussels, Belgium, 2012: 918-923

[15] Cui P, Jin S, Yu L, et al. Cascading outbreak prediction in networks: A data-driven approach//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China, 2013: 901-909

[16] Cheng S, Shen H, Huang J, et al. StaticGreedy: Solving the scalability-accuracy dilemma in influence maximization//Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management. San Francisco, USA, 2013: 509-518

[17] Alon N, Gamzu I, Tennenholtz M. Optimizing budget allocation among channels and influencers//Proceedings of the 21st International Conference on World Wide Web. Lyon, France, 2012: 381-388

[18] Chen W, Yuan Y, Zhang L. Scalable influence maximization for prevalent viral marketing in large scale social networks//Proceedings of the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Washington, USA, 2010: 1029-1038

[19] Anagnostopoulos A, Kumar R, Mahidian M. Influence and correlation in social networks//Proceedings of the 14th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. San Francisco, USA, 2008: 57-66

[20] McPherson M, Smith-Lovin L, Cook M J. Birds of a feather: Homophily in social networks. Annual Review of Sociology, 2001, 27(1): 415-444

[21] Cui P, Wang F, Liu S, et al. Who should we share what? Item-level social influence prediction for users and posts ranking//Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. Beijing, China, 2011: 185-194

[22] Yan J, Liu N, Wang G, et al. How much can behavioral targeting help online advertising//Proceedings of the 18th International Conference on World Wide Web. Madrid, Spain, 2009: 261-270

[23] Zhang Y, Wu Z, Chen H, Sheng H. Mining target marketing groups from users' Web of trust on epinions//Proceedings of the AAAI Spring Symposium. Chicago, USA, 2008: 116-122

[24] Chen Y, Pavlov D, Canny J F. Large-scale behavioral targeting //Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France, 2009: 209-218

[25] Tang J, Sun S, Wang C, Yang Z. Social Influence analysis in large-scale networks//Proceedings of the 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Paris, France, 2009: 807-816

[26] Jo Y, Hopcroft J E, Lagoze C. The Web of topics: Discovering the topology of topic evolution in a corpus//Proceedings of the 20th International Conference on World Wide Web. Hyderabad, India, 2011: 257-266

[27] Budak C, Agrawal S, Abbadi A. Limiting the spread of misinformation in social networks//Proceedings of the 20th International Conference on World Wide Web. Hyderabad, India, 2011: 665-674

[28] Zhang B, Qian Z, Wang X, Lu S. Tracing influential nodes in a social network with competing information//Proceedings of the 17th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. Part II, Gold Coast, Australia, 2013: 37-48

[29] Gomez-Rodriguez M, Leskovec J, Krause A. Inferring networks of diffusion and influence//Proceedings of the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Washington, USA, 2010: 1019-1028

[30] Choudhury M, Sundaram H, John A, et al. "Birds of a Feather": Does user homophily impact information diffusion in social media?. arXiv Preprint arXiv: 2010, 1006(1702)

[31] Goyal A, Bonchi F, Lakshmannan L V S. Learning influence probabilities in social networks//Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. New York, USA, 2010: 241-250

附录 1.

定理 2 证明. 由于 $\sigma(\cdot)$ 为影响力的期望值，而每条边上的传播概率是独立的，因此可以采用蒙特卡洛的方法，对概率图中的每条边 e 进行概率为 p_e 的抛硬币(coin flip)，则可以得到一个确定的图的实例 X，列举所有这样的情况，即有

$$\sigma(S)=\sum_{outcom\ X} Prob[X]\cdot\sigma_X(S),$$

其中 $Prob[X]$ 为得到实例 X 的概率，$\sigma_X(S)$ 为每个确定子图中 S 的连通分量与 T 的交集的节点数目. 如果可以证明 $\sigma_X(S)$ 具有子模的性质，那么显然它的线性组合 $\sigma(S)$ 也具有同样的性质. 假设 $R(S,X)$ 为确定的实例 X 下 S 的连通分量；$R_T(S,X)$ 表示在确定的实例 X 下，S 的连通分量所覆盖的目标节点的集合，则 $R_T(S,X)=R(S,X)\cap T=\sigma_X(S)$. 显然，对于一个确定的图，$R_T(S,X)$ 对于集合 S 单调递增，因此若 S 是 Q 的子集，因此 $R_T(v,X)-R_T(S,X)\geq R_T(v,X)-R_T(Q,X)$，即 $\sigma_X(S\cup\{v\})-\sigma_X(S)\geq\sigma_X(Q\cup\{v\})-\sigma_X(Q)$，因此 $\sigma_X(S)$ 具有次模的性质. 同时，$\sigma_X(S)$ 还是单调递增的，而对于单调的次模函数，通过贪心算法选取的 k 个节点集 S，可以取得 $(1-1/e)$ 的近似保证. 证毕.

ZHANG Bo-Lei, born in 1988, Ph.D. candidate. His research interests include complex networks and social networks, with an emphasis on social influence and information diffusion analysis.

QIAN Zhu-Zhong, born in 1980, Ph. D. His research interests include cloud computing and social network analysis.

WANG Qin-Hui, born in 1985, Ph. D. candidate. His research interests include network economics, cognitive radio.

LU Sang-Lu, born in 1970, Ph. D. Her research interests include distributed computing and complex networks.

Background

With the great evolution of online social networks such as Facebook, Twitter, Weibo, etc., it provides rich data for analysis and new opportunities for leveraging applications using social network platforms. Substantial attention has been gained to investigate the information spreading in these networks. One problem that has been studied extensively is influence maximization in social networks. Companies can convince a small set of people as initial adopters and spread the products information via social links. The goal is to generate a large cascade of adoption. Kempe et al. first modeled the information diffusion as a discrete step process and show the problem is NP-hard. They further propose greedy algorithm using hill-climbing strategy with performance guarantee. Since then, a lot of research has been done to improve the performance and efficiency of the algorithm.

However, most of these works ignore the value of customers. Motivated by the progress in target marketing, we can first identify those who will potentially buy the products using their purchase history, personal interests etc. And the company would like deliver their advertisement to these valuable customers. As the social links are considered trustful relationships, we intend to use influence maximization to cover as many valuable customers as possible. Previous algorithms are not applicable in this problem because they ignore the relationship between the target market and the global structure.

In this paper, the authors propose a cluster based algorithm KCC to find influentials for the target market. The independent cascade model (ICM) is adopted to simulate the spreading process of information. They analyze real data set and find the targeted users are usually clustered together due to homophily. With this observation, KCC tries to cluster the targeted users and find a representative in each cluster. They first reduce the sample space using diffusion theory and show that the influential nodes are within a limited step from the target nodes. They then define the similarity and cluster cost for the cluster algorithm so that the algorithm can separate the nodes and find the most influential nodes in each cluster. The experiments on real data set show that our algorithm is effective in finding influential nodes and has low running time compared to the inefficient greedy algorithm. The authors have been working on information diffusion and social influence analysis. They published papers in international conferences like PAKDD 13 etc.

This work is partially supported by the National Natural Science Foundation of China under Grant Nos. 61321491, 61202113, 91218302; Jiangsu Natural Science Foundation under Grant No. BK2011510.