大规模社会网络中影响最大化高效处理技术

Daguimo Shehui Wangluozhong Yingxiang Zuidahua Gaoxiao Chuli Jishu

廖湘科 刘晓东 李姗姗 等编著

国防工业出版社
National Defense Industry Press

大规模社会网络中影响最大化高效处理技术

廖湘科　刘晓东　李姗姗　吴庆波
戴华东　彭绍亮　王　蕾　朱培栋　编著

国防工业出版社
·北京·

内容简介

本书针对社会网络最大化问题的高效处理技术提出了有效的解决方案，希望对于推进社会网络分析和影响最大化问题研究和实用化有一定的理论意义和应用价值。本书主要针对计算机学科、社会学学科等相关学科领域的广大大学师生以及科研人员提供了对社会网络基本知识、特性、发展规律以及社会网络分析各种算法的介绍。

图书在版编目(CIP)数据

大规模社会网络中影响最大化高效处理技术/廖湘科等编著. —北京：国防工业出版社，2013.8

ISBN 978-7-118-08988-2

Ⅰ.①大… Ⅱ.①廖… Ⅲ.①计算机网络—安全技术—研究 Ⅳ.①TP393.08

中国版本图书馆 CIP 数据核字(2013)第 179252 号

※

国防工业出版社出版发行

(北京市海淀区紫竹院南路 23 号 邮政编码 100048)

国防工业出版社印刷厂印刷

新华书店经售

*

开本 880×1230 1/32 **印张** 5½ **字数** 190 千字

2013 年 8 月第 1 版第 1 次印刷 **印数** 1—2500 册 **定价** 40.00 元

(本书如有印装错误，我社负责调换)

国防书店：(010)88540777 发行邮购：(010)88540776

发行传真：(010)88540755 发行业务：(010)88540717

前　言

近年来，随着互联网和 Web 2.0 技术的飞速发展，社会网络作为沟通现实人类世界的桥梁，已经成为交互沟通、知识共享和信息传播的重要媒介和平台。其中，影响最大化问题作为社会网络分析领域的一个关键问题，在许多重要场景中有着广泛的应用，例如市场营销、广告发布、舆情预警、水质监测、疫情监控等，因此具有很高的研究价值和应用价值。

许多影响最大化应用策略的制定和部署对于算法求解时间十分敏感，因此，高效的求解算法是当前学术界和工业界研究影响最大化算法的核心目标。针对已有算法求解速度慢、计算效率低的问题，已有的研究成果主要集中于一些贪心算法和启发式算法。然而，当今社会网络的数据规模海量、数据耦合度高、网络结构动态变化，当面对大规模社会网络时，已有算法暴露出许多难以克服的问题：第一，社会网络中节点影响值计算的可并行性问题。已有工作专注于降低算法的复杂度，没有充分利用已有的并行计算架构来加速问题求解。而实际社会网络中存在大量的节点影响值计算可由并行计算架构并发执行。因此，在挖掘算法的并行性方面，影响最大化算法的执行速度仍有很大的提升空间。第二，影响最大化算法效率与精度需求的折中问题。社会网络中节点的度分布服从典型的幂律分布。然而，现有贪心算法大多采用精确计算的方式来计算所有节点的影响值，导致大度数节点的计算复杂度十分高，成为算法执行的瓶颈。第三，社会网络拓扑结构的动态变化问题。现有影响最大化问题研究专注于静态网络；当网络动态变化时，大都需要针对全网进行重新计算节点的影响值，会造成大量冗余计算，导致性能无法满足大规模社会网

络的需求。

针对上述技术瓶颈，本书系统地研究了社会网络影响最大化问题的高效处理技术，从以下几个方面展开研究：

（1）针对现有方法并行度差、算法复杂度高，从而导致运行时间过长的问题，本书基于 CPU + GPU 的异构并行计算框架，设计和实现了一种具有高并行度的影响最大化算法 BUTA，并针对 GPU 体系结构做了进一步算法优化。本书通过深入分析社会网络中节点之间的层次依赖关系，发现了节点影响值计算的可并行性。在此基础上，设计了一种自底向上的逐层扫描方法 BUTA。BUTA 算法一方面可以在保证算法精度的同时大幅度降低算法复杂度，另一方面 BUTA 充分利用了节点的层次分布，以高并行度计算节点的影响值。为了使 BUTA 算法更加适配 CPU + GPU 的异构并行计算框架，本书设计了三种优化方法：K 层合并、数据重组和合并访存，分别用于降低运行时分支、减少访存次数和提高算法并行度。

（2）针对影响最大化算法效率与精度需求的折中问题，本书提出了一种基于蒙特卡洛理论的采样估计算法 ESMCE，大幅度提升了计算效率。本书对社会网络中节点的分布特性进行了建模和挖掘；针对大度数节点计算时间长的问题，本书引入蒙特卡洛理论，设计了一种节点影响值估计方法 ESMCE。在采样过程中，ESMCE 算法设计了一种由幂律指数指导的采样节点个数计算方法。之后，根据估计误差同精度要求之间的差距，本书提出了一种基于灰度预测模型的后续采样节点个数预测方法，以通过多次迭代采样来提高算法精度直至采样误差满足设定的精度要求。

（3）针对社会网络拓扑结构的动态变化问题，本书设计了一种增量式的影响最大化算法 IncInf。本书深入分析了社会网络拓扑结构的演化特征，发现社会网络的拓扑变化满足优先连接原则，同时最有影响力节点的度数要明显大于普通节点。基于上述发现，本书设计了一种基于局部化理论的影响变化量高效计算方法。基于节点的影响变化量和原有网络对应的最有影响力节点信息，设计了一种剪

此为试读，需要完整PDF请访问：www.ertongbook.com

枝策略，将候选节点范围有效缩小到影响值增长迅速、度排序靠前的节点集合，从而大幅度降低了算法复杂度，减少了程序运行时间。

（4）针对当前内容分发方法忽略了社会网络中的用户关联关系、地理位置等社会信息的问题，本书设计了一种基于影响最大化的内容分发方法 SCORE。同已有的内容分发方法不同，SCORE 方法充分利用了社会网络中的用户信息，提出了一种基于影响最大化算法的缓存内容选择策略，以快速准确地定位未来访问频率较高的关键内容。为了最小化访问延迟，SCORE 方法通过挖掘用户之间的关联关系和地理位置信息，设计了一种基于 K - MEANS 聚类算法和加权球面平均计算方法的边缘服务器选择策略，从而将关键内容预先分发到离潜在访问用户最近的边缘服务器，以便于就近响应用户请求。实验结果表明，SCORE 方法可以大幅度降低用户访问延迟，提升用户体验质量，同时减轻网络流量压力。

综上所述，本书针对社会网络影响最大化问题的高效处理技术提出了有效的解决方案，并通过在真实数据集上进行实验验证了所提算法的有效性，对于推进社会网络影响最大化问题的研究和实用化具有一定的理论意义和应用价值。

PREFACE

In recent years, with the in - depth research of Internet and Web 2.0 techniques, social network, serving as an important medium for communication, knowledge sharing and information spreading, has been widely used for bridging the human world. Influence maximization, as one of the key issues in the filed of social network analysis, has been extensively applied to many crucial scenarios, such as viral marketing, advertisement publishing, public sentiment warning, water quality and epidemic monitoring, which shows substantial research and application importance.

Currently, many researchers in academia and industry work on the influence maximization problem, and propose lots of greedy algorithms and heuristics, which fairly improve the efficiency of influence maximization algorithms. Nevertheless, modern social networks are mostly large - scale, highly complicated and essentially dynamic, which pose serious challenges to the high efficient processing of influential user identification. Most existing algorithms suffer the following problems:

First, existing algorithms only focus on designing algorithms with low computation complexity, while ignoring the parallelism of influence spread computation. They also take no advantage of existing heterogeneous parallel computing frameworks, such as CPU + GPU, for acceleration. In fact, the influence spread computation of many nodes in social network can be performed in parallel to overlap the running time and thus the efficiency can be dramatically improved. Second, they take no con-

sideration of the node distribution characteristics in social network. Existing works mainly focus on computing the exact influence spread, leading to low computational efficiency and limiting their application to real – world social networks. Third, previous studies overlook the dynamic characteristics exhibited during the evolution of real – world social networks. Most of them are proposed to deal with static social network. While, as a matter of fact, real – world social networks keep evolving over time. When confronting dynamic social networks, existing works will suffer from computing from scratch.

To well address the above challenges, this book systematically investigates some key issues of influence maximization in large – scale social networks, especially on the high efficient processing techniques. The research mainly focuses on the following aspects.

For parallel computation of influence spread, this book in – depth investigates the dependency relationship among nodes in social networks. To improve the parallelism and reduce the complexity of existing algorithms, we propose a bottom – up traversal algorithm with inherent parallelism. On the one hand, BUTA can greatly reduce the time complexity through DAG conversion and bottom – up traversal. On the other hand, BUTA is designed with sufficient parallelism and can be mapped to modern heterogeneous parallel computing frameworks. For this reason, we map BUTA to GPU to exploit the parallel processing capability of GPU, thus further reducing the execution time. To best fit BUTA with the GPU architecture, we further develop an adaptive K – Level combination method to maximize the parallelism and reorganize the influence graph to minimize the potential divergence.

In this book, we also exploit Monte – Carlo estimation to significantly improve the efficiency at only negligible cost of precision. We first analyze the node distribution characteristics in social network. To address

the key bottleneck of influence spread computation for nodes with large degree, we design a power - law exponent supervised Monte - Carlo estimation method, named ESMCE. ESMCE exploits the power - law exponent of the social network to guide the initial sampling. Then, based on the disparity of estimation error and precision requirement, ESMCE utilizes the grey forecasting method to forecast the number of child nodes needed in further iteration. Multiple iterative steps run until the precision requirement is finally achieved.

To deal with the influence maximization problem in dynamic social networks, we investigate the dynamic characteristics of social networks and observe from real - world traces that the evolution of social network follows the preferential attachment rule and the influential nodes are mainly selected from high - degree nodes. Such observations shed light on the design of IncInf, an incremental approach that can efficiently locate the top - K influential individuals based on previous information instead of calculation from scratch. In particular, IncInf quantitatively analyzes the influence spread changes of nodes by localizing the impact of topology evolution to only local regions, and a pruning strategy is proposed to effectively narrow the search space into nodes experiencing major increases or with high degrees.

None of previous content distribution methods comprehensively exploits the valuable social information of social network, such as social relationship and user geographic information. In this book, we propose SCORE, a social - aware content distribution method based on influence maximization problem. SCORE fast locates the contents that are potential to trigger large cascades. Then SCORE leverages the social relationship and geographic information, and selects target servers by K - MEANS clustering algorithm and weighted spherical mean calculation. Through this, SCORE effectively pushes selected content to geo - located servers

before potential users actually request the content so as to reduce user latency, improve quality of experience, and alleviate network traffic pressure.

To summarize, our works present solutions to several essential issues of influence maximization which are key requirements in social networks. Comprehensive experiments demonstrate that proposed algorithms can properly achieve their design goals.

目　录

此为试读，需要完整PDF请访问：www.ertongbook.com

第1章 绪 论

互联网和 Web 2.0 技术带来了信息产生方式和传播模式的深刻变革，推动了社会媒体的蓬勃发展。各种社会网络服务（Social Networking Service，SNS）层出不穷，已经渗透到大众生活的方方面面，成为人类交互沟通、知识共享和信息传播的重要媒介和平台[1]，人类社会已经进入了 Web 2.0 的社会网络时代。当前知名的大型社交网站包括国外的 Facebook[2]、Twitter[3]、Google +[4] 以及国内的新浪微博[5]、人人网[6]等。社会网络在信息、观点、创新的传播扩散过程中发挥了基础性的作用；而且随着用户数目的持续增长，社会网络的规模快速扩大，其核心作用越来越显著，因此社会网络技术引起了学术界和工业界的高度重视，被认为是对 21 世纪产生重大影响的技术之一。

1.1 社会网络研究概述

1.1.1 基本概念和特点

社会网络是由社会个体成员以及个体成员之间的社会关系所组成的一种复杂网络结构[7]。其中社会个体成员可以是个人、组织等不同含义的实体或虚拟个体；而个体成员之间的社会关系可以是合作、朋友、血缘、敌对等各种类型的关系。社会网络通过一种或者多种特定类型的社会关系将多个社会个体成员组织成某种社会结构。

随着国内外社交网络服务的兴起和流行，各种社会网络的用户

数目飞速增长。如图 1 - 1 所示，Facebook 自从 2004 年上线以来，已经成长为目前世界上最大的社交网站；截至 2012 年 10 月，Facebook 活跃用户数已经超过 10 亿，拥有 1250 亿的用户连接。Google + 自从 2011 年 6 月面世之后用户数目飞速增长，在 2012 年 12 月已经拥有超过 5 亿的用户。另外，Twitter 平均每天都有 100 万新用户注册加入[8]，国内新浪微博注册用户数截至 2012 年 12 月底也已经超过 5 亿[9]。如此迅猛的发展势头，使得社会网络广受工业界和学术界的关注，社会网络分析已经成为了当前重要的研究领域。

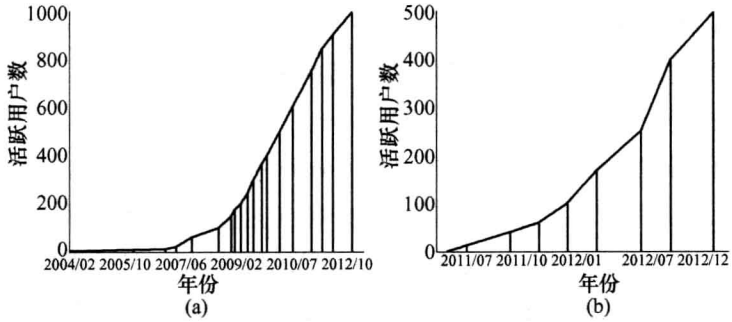

图 1 - 1　Facebook 和 Google + 用户增长趋势

(a) Facebook；(b) Google + 。

尽管社会网络中节点个数众多、结构复杂，然而社会网络中节点之间的连接关系并非完全随机，而是呈现出一些社会网络所独有的规律，例如“小世界特性”(Small World)、“无标度特性”(Scale Free)以及“高聚类系数”(High Clustering Coefficient)等。

1. 小世界特性

1967 年哈佛大学的社会心理学家 Stanley Milgram 在美国内布拉斯加州的一次实验发现，邮件平均通过六次转发就可以到达指定的陌生人，这就是著名的六度分隔理论(Six Degrees of Separation)[10]。美国哥伦比亚大学的 Duncan Watts 等于 1998 年提出小世界网络模型[11]，并于 2001 年在互联网上发动六万余名志愿者通过转发电子邮件的方式验证了六度分隔理论[12]。2011 年，Facebook 同米兰大学

联合研究发现 Facebook 上两个用户之间的平均距离仅为4.74[13]，同样满足六度分割理论。

2. 无标度特性

无标度特性由 Albert - Laszlo Barabasi 和 Reka Albert 等于 1999 年首次提出[14]，他们经过研究发现万维网的出度和入度、电力网络以及神经网络都可以用幂律分布很好地描述。另外，Alan Mislove 等研究发现 YouTube[15]、Flickr[16]、LiveJournal[17] 和 Orkut[18] 四个在线社交网站的节点度分布也满足幂律特性[19]。幂律分布的形式化描述如下所示：

$$P(k) \sim k^{-\gamma} \tag{1-1}$$

式中：$\gamma > 0$ 为幂指数，通常取值在 2 ~ 3 之间；$P(k)$ 是网络中一个随机选择的节点的度为 k 的概率。

3. 高聚类系数

聚类系数是一个体现网络图中节点之间聚集成团的系数。具体来说，它反映了一个点的邻接点之间相互连接的程度。节点 i 的聚类系数 C_i 的形式化描述如下所示：

$$C_i = \frac{E_i}{(k_i(k_i - 1))/2} \tag{1-2}$$

式中：k_i 是节点 i 的邻接点个数；E_i 是节点 i 的 k_i 个邻接点之间实际存在的边数。由于社会网络中的用户关系在一定程度上反映了真实世界的网络结构，因此各个节点之间倾向于形成密度相对较高的群体，所以，相对于随机网络，社会网络的集聚系数更高。Holger Ebel 等[20]以及 Mark Newman 等[21]的研究证实了社会网络的高聚类特性。

1.1.2 研究现状

社会网络是目前人工智能和数据挖掘领域的研究热点。越来越多的知名大学和研究院所意识到社会网络的潜力，纷纷投入其中并

此为试读，需要完整PDF请访问：www.ertongbook.com