分 类 号 学 号 M201372761

学校代码 10487 密 级



硕士学位论文

基于多话题的大规模社会网络影响力最大化研究

|  |  |
| --- | --- |
| 学位申请人： | 王 凯 |
| 学科专业： | 计算机应用技术 |
| 指导教师： | 李玉华 副教授 |
| 答辩日期： | 2016年5月24日 |
|  |  |

**A Thesis Submitted in Full Fulfillment of the Requirements**

**for the Degree of the Master of Engineering**

**Multi-Topic Influence Maximization in Large Scale Social Networks**

|  |  |
| --- | --- |
| **Candidate :** | Kai Wang |
| **Major :** | **Computer Application Technology** |
| **Supervisor:** | Associate Professor Yuhua Li |
|  |  |

Huazhong University of Science & Technology

Wuhan 430074, P.R.China

May, 2016

**独创性声明**

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

学位论文作者签名：

  日期：     年   月   日

**学位论文版权使用授权书**

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

保密□，在\_\_\_\_\_年解密后适用本授权书。

本论文属于

不保密□。

（请在以上方框内打“√”）

 学位论文作者签名：                   指导教师签名：

日期：    年   月   日              日期：    年   月    日

# 摘 要

近些年来，影响力最大化问题已经是数据挖掘领域炙手可热的研究方向，并且普遍应用于社会网络分析。然而，现有大部分研究在寻找最具影响力的种子节点的同时忽略了一个事实，那就是针对不同的话题，所选出的种子节点也是不同的。即使少部分现有研究考虑了话题因素，它们忽略了任何需要在网络中传播的商品或事件都是由多个话题组成的，只考虑单个话题是片面且不准确的。同时，网络中用户的兴趣往往也不是单一的，而正是用户的行为偏好直接决定了影响在社会网络中传播的结果。

针对现有传播模型和算法在传播过程中未考虑多话题因素的缺陷，本文首先利用话题模型对文本信息进行建模，再将得到的话题分布结合到传统的独立级联模型（Independent Cascade Model）中，提出了多话题敏感的独立级联模型（Multi-Topic Sensitive Independent Cascade model，MTSIC模型）。利用MTSIC模型，可以帮助选出最贴近现实生活情况的种子节点。同时利用Topical HITS算法，获得用户的权威度和从众性，并融入到模型中，使模型更加准确。由于在网络上传播的商品或事件可能存在地理位置限制，因此地理位置信息也被当作一项重要因素加入到模型中。传统影响力最大化算法并不适用于多话题场景，因此提出多话题敏感的影响力最大化算法（Activation Nodes Similarity algorithm，ANS）。考虑到在大规模网络中进行影响力最大化分析是十分耗时的，因此本文提出了基于Spark的多话题敏感影响力最大化算法（Parallelization of Multi-Topic algorithm，PMT）并行算法以提高算法效率。

由于传统评价度量未能体现多话题因素的重要性，因此本文提出了新的度量*SIS*来诠释影响力最大化算法效果。通过在数据集DBLP和Twitter上的实验结果显示，MTSIC模型可以更准确的模拟真实情况下节点的激活情况并且ANS算法可以找到在现实情况中更倾向于接受商品或事件并进行传播的种子节点。而PMT算法的高效性也被证明。从各个方面进行的实验结果证明了本文所提出的传播模型及算法是效且高效的。

**关键词：**影响力最大化，大规模社会网络，多话题敏感的传播模型，地理位置，并行化

# Abstract

Influence maximization has been widely studied in social network analysis. However, most existing works focus on global influentials while ignoring the fact that different topics may have different influentials. Even though a few works take topics into consideration, they neglect that the product or event needed to spread in the network refers to multi-topics instead of only one topic. At the same time, the users in the network also have different interests deciding the result of the influence propagation.

In this paper, we study the influence maximization of social network by incorporating the multi-topic information into Independent Cascade (IC) model and propose a multi-topic sensitive diffusion model, the Multi-Topic Sensitive Independent Cascade model (MTSIC). Based on the MTSIC model, we can find the influential nodes which are closer to the reality. The topical Hyperlink Induced Topic Search (topical HITS) algorithm is used to calculate the influence and conformity of each node in the proposed diffusion models. As the product or event needed to spread may have location limitation, so the locations of influentials nodes are also considered. Considering the traditional influence maximization algorithm is not suitable for multi-topic situation, then we propose the Activation Nodes Similarity algorithm (ANS). Given that calculating the influence maximization in large scale network is very time-consuming, a solution based on Spark, namely Parallelization of Multi-Topic algorithm (PMT), is proposed to improve the efficiency.

Considering the deficiency of traditional measurement to depict the importance of the multi-topic, we propose a novel measurement *SIS* to illustrate the effectiveness of the multi-topic influence maximization algorithms. Extensive experiments on DBLP and Twitter data show the MTSIC model can exactly simulate the activation of the nodes in reality and ANS algorithm can find out the influential nodes which are more likely to propagate in real world. The efficiency of PMT algorithm is also proved. The result of thorough experiments is the best evidence of the effectiveness and efficiency of our model and algorithms.

**Key words:** Influence maximization，large scale social network，multi-topic sensitive diffusion model，location-based，parallelization

# 目 录

[摘 要 I](#_Toc449380740)

[Abstract II](#_Toc449380741)

[1 绪 论](#_Toc449380743)

[1.1 课题研究背景](#_Toc449380744) (1)

[1.2 课题研究的目的和意义](#_Toc449380745) (1)

[1.3 国内外研究现状](#_Toc449380746) (2)

[1.4 论文的主要研究内容](#_Toc449380750) (7)

[1.5 论文组织结构](#_Toc449380751) (8)

[2 多话题敏感的影响传播模型](#_Toc449380752)

[2.1 相关问题定义](#_Toc449380753) (9)

[2.2 话题模型LDA](#_Toc449380754) (11)

[2.3 Topical HITS算法](#_Toc449380755) (12)

[2.4基于多话题的影响传播模型](#_Toc449380756) (15)

[2.5 本章小结](#_Toc449380760) (22)

[3 多话题敏感的影响力最大化算法](#_Toc449380761)

[3.1 KKT贪心算法](#_Toc449380762) (23)

[3.2 MTSIC模型的特性](#_Toc449380763) (24)

[3.3 多话题敏感的影响力最大化算法](#_Toc449380764) (26)

[3.4 本章小结](#_Toc449380767) (28)

[4 针对大规模社会网络的影响力最大化并行算法](#_Toc449380768)

[4.1 SPARK并行计算框架](#_Toc449380769) (29)

[4.2 影响力最大化并行算法PMT](#_Toc449380770) (30)

[4.3 PMT算法的流程](#_Toc449380774) (33)

[4.4 本章小结](#_Toc449380775) (34)

[5 实验与分析](#_Toc449380776)

[5.1 实验环境及数据情况](#_Toc449380777) (35)

[5.2 实验评价度量](#_Toc449380780) (36)

[5.3 实验方案设计](#_Toc449380781) (37)

[5.4 实验结果及分析](#_Toc449380782) (38)

[5.5 本章小结](#_Toc449380787) (49)

[6 总结与展望](#_Toc449380788)

[6.1 论文总结](#_Toc449380789) (50)

[6.2 论文展望](#_Toc449380790) (51)

[致 谢](#_Toc449380791) (52)

[参考文献](#_Toc449380792) (53)

[附录1 攻读硕士学位期间投稿的论文](#_Toc449380793) (59)

# [1 绪 论](#_Toc283328270)

## 1.1 课题研究背景

随着互联网突飞猛进的发展，社会网络已成为人们交际的重要场所。真实世界中存在着各种联系的用户群组在虚拟的社会网络中就被映射成为一个个有向网络，而信息将会在这些有向网络中的节点间相互传递。在传递的过程中，消息或者经由节点之间的链接关系而被一级一级地传递下去，或者由于某些原因被中断。如果想要将信息最大化的进行传播，那么就需要对动态传播机制进行研究，因此影响力最大化成为数据挖掘领域的一个重要研究方向。

在消息传播的过程中，当一个节点收到从其入度邻居传来的信息后，它会根据各种条件来判断是否接受该信息。如果接受的话，该节点会继续向其出度邻居传播信息。这种传播过程被称作口碑效应(word-of-mouth)[1]。真实情况显示，相对于网络上眼花缭乱的广告，人们更加倾向于接受来自自己关系紧密的社交群组里的“友情推荐”。随着人们对互联网的依赖的增加，像微博、Twitter、Facebook等大型社会网络平台成为了“病毒式营销”[2]策略绝佳的应用场所。

由于人与人之间偏好的差异，使得不同商品或事件在相同网络中的传播结果变得不同的。而人的偏好无法直接得到，只能通过用户在网络中遗留的文本信息分析获得，因此文本信息成为信息传播过程中的重要因素。本文所研究的论文引用网络和Twitter网络皆为文本丰富的社会网络，对论文引用网络进行影响力最大化研究可以帮助新的著作更好更快的被广大相关学者所熟知，而对Twitter网络进行影响力最大化分析可以帮助新产品或事件的推广。据此本文将多话题信息融入传统的影响力最大化问题，以期获得更有效、更贴合实际的 结果。

## 1.2 课题研究的目的和意义

现如今，利用在线社会网络来进行产品的推广与营销依然是数据挖掘方向的热门研究，而影响最大化就是其中的一项重点研究。一项产品的推广是有固定预算的，因此只能选择一个用户集合来利用“病毒式传播”使得最后受到影响的用户最多，有更多的用户愿意去购买产品。而这个过程中，如何利用已有条件选择初始用户集合来推广产品或事件，最终获得最大的影响力是问题的关键所在。

影响力最大化问题的应用场景众多，不仅在市场营销中发挥重要作用，而且在学术交流等其他方向具有很大的研究前景。目前，由于FaceBook、Tiwtter、微博等社交平台的繁荣壮大，影响力最大化问题已经有了新的起点，如何在大规模网络中准确并且高效的找出最具影响力的用户是十分具有挑战性的问题。

影响力最大化问题旨在解决如何找到网络中能够使得信息传播最大化的种子节点集合。整个选取的过程主要中涉及了传播模型和影响力最大化算法两部分内容，传播模型主要负责影响的传播与节点的激活，而影响力最大化算法负责找出符合要求的种子节点。目前已有相关研究大部分都不是话题敏感的，而不同的话题下所应对应的种子结点往往大有不同。Tang等人[3]针对话题分布会对信息传播造成的影响给出了详细的说明，据此可以体现话题因素在影响力最大化问题中的重要作用。即使少部分的传播模型考虑了话题因素，但它们都忽略了一个事实：在真实世界中，任何亟待推广的产品或事件都必定包含多个话题，只考虑单个话题是片面且不准确的。同时，网络中用户的兴趣往往也不是单一的，而正是用户的行为偏好直接决定了影响在社会网络中传播的结果。除此之外，它们同样也没有考虑地理位置、权威度等节点因素，并且无法解决大规模社会网络的问题。

本文旨在改进以往传播模型的不足，解决大规模社会网络下影响最大化问题，加入多话题、地理位置等因素，使其结果能更加贴近真实情况；构思设计了对应的影响力最大化算法和度量，让其能更好的与新模型相结合；通过采用并行计算的方式提升算法的效率，解决大规模网络的问题。

## 1.3 国内外研究现状

影响力最大化问题在数据挖掘领域及社会网络分析方面一直都是一个热门的研究方向，它主要包含两个重要的部分：传播模型以及影响力最大化算法。由于话题因素在影响力最大化问题中具有至关重要的作用，因此话题建模也成为研究影响力最大化问题的一种不可或缺的工具。本文将会对话题模型、传统影响力最大化问题及话题敏感的影响力最大化问题的国内外相关现状进行简要介绍。

### 1.3.1话题模型

对于话题建模的研究已经有十余年，最早的话题建模技术是概率潜语义分析（Probabilistic Latent Semantic Analysis， PLSA）[4]。PLSA其本质是概率生成模型，其通过在文档和单词之间引入话题层的方式对文档的生成过程进行建模。Blei 在2003年提出的潜在狄利克雷分配（Latent Dirichlet Allocation，LDA）[5]无疑是最经典的方法，它能通过对每一篇文档的分析和计算，获得文档的话题分布和每一个话题所具有的单词分布，其中，话题分布是用若干个概率值来显示用户在对应话题上的兴趣度的强弱。到目前为止出现的很多新的模型都是对LDA模型进行改进的[6-9]，Blei等[10]提出的相关话题模型(Correlated Topic Model，CTM)改进了LDA模型，可以获得话题之间的相关度，更好的诠释了真实情况下的话题的相关性。除此之外，Blei等[11]所提出的动态话题模型（Dynamic Topic Model，DTM）可以将时间分割成若干个离散的时间片段，将话题表示作多项式分布，并采用高斯分布对相邻时间片段的话题进行建模，构思了基于卡尔曼滤波法（Kalman Filtering）和无参数小波回归（Nonparametric Wavelet Regression）的变分推理算法求解DTM中的参数。

近几年有许多比较先进的话题模型，Deng等[8]把话题模型与异构信息网络(Heterogeneous Information Networks)相结合，使其得到拓展。McCallum等[12]提出作者-收取者-话题模型（Author-Recipient-Topic Model, ART），同时对文本的作者和收取者进行建模，并设计了角色-作者-收取者-话题模型（Role-Author-Recipient-Topic Model, RART），认为用户在收发信息的过程中都与一些角色相关联。还有很多优秀的话题模型，在这里就不一一列举了。

### 1.3.2传统影响力最大化问题

影响力最大化问题是数据挖掘领域中的一个经典问题，其主要包含影响力最大化算法和传播模型两大部分。传播模型负责抽象并模拟消息在真实网络中的传播情况及节点的激活情况，而算法负责找到满足要求的能够将影响最大化传播的中子节点集合。目前已有影响力最大化问题的相关研究已经硕果累累，具体介绍如下。

影响力最大化问题最初是由Domingos等[13]所提出的，其根本目标是解决在病毒营销的场景下找到能使影响最大化传播的种子节点集合。随后，Kempe等人[14]把该问题建模成一个离散优化问题，提出并详细介绍了两种最经典的传播模型：独立级联模型（Independent Cascade Model，ICM）和线性阈值模型（Linear Threshold Model，LTM），并且在这两种模型上提出了贪心近似KKT算法。由于KKT算法的原理是每次都选取最优解，这就注定KKT算法找到的种子节点是最优的，但与此同时也注定KKT算法的复杂性很高，效率是最差的，因此对于大规模的网络不十分适用。对于影响力算法的效率改进问题已经做了很多的工作，其改进的方向大概集中在以下两个方面：通过启发式或剪枝等方法减少计算量来提高效率，利用并行计算的方式来节约时间。

目前很多改进的影响力最大化算法都有着不错的效果，有的甚至能够很接近KKT算法的效果。Leskovec等人[15]根据函数子模性具有收益递减的性质，提出了CELF（Cost-Effective Lazy Forward）算法来缩减用于蒙特卡洛模拟的时间以提高效率，相对于KKT算法，其具有将近700倍的效率提升。Amit等人[16]改进了CELF算法而提出CELF++算法，该算法可以同时计算边际效益以更好的减少运算时间。Chen等人[17]提出了NewGreedy算法，用以对传播网络中的无用的边进行剪枝；同时又提出了MixGreedy算法，将CELF算法与NewGreedy算法相结合来获得效率提升。除此之外，Liu等人[18]利用GPU的高效计算的特点，对KKT算法进行改进并且提出了Bottom-up算法，使得节点间的边际影响统计可以相互独立，符合并行化的要求，因而又提出IMGPU算法，利用GPU实现并行计算来加速算法，来缓解在大规模网络中进行影响力最大化分析的效率问题。

以上所提到的大部分算法都是针对KKT算法的改进，虽说效率有所改进，但仍然无法达到高效的程度，因而依然很难适用于大规模网络。针对这样的情况，一些基于经典模型的启发式的算法便逐渐被设计出来。Chen等人[19]基于IC模型提出的MIA算法、 M. Kimura等人[20]提出的基于IC模型提出的shortest-path算法、Chen等人[21]基于LT模型提出的本地有向无环图算法LDAG、Goyal等人[22]提出的基于简单路径的Simpath算法等都是十分高效的启发式算法，在提高效率的同时，其效果甚至可以与KKT算法相媲美。

上面所介绍的研究都没有考虑到用户节点对于影响传播的能力会有所不同，而估计用户节点的传播能力的传统方法是采用模拟的方法，最常用的为蒙特卡洛模拟（Monte-Carlo, MC），但是这种模拟方式的效率十分低下。而Kimura等人[23]根据链接渗透和图论的相关方法，将其应用到KKT算法或者其他算法中，使其能够有效估计边际收益，相对于蒙特卡洛模拟，可以极大地减少计算量。由于种子节点的选择过程是#p-hard问题，因此 Kim等人[24]提出了并行化的算法IPA（Independent Path Algorithm），该算法利用OpenMP，从而大幅度提高了处理速度。冯小军等[25]则将每一个用户节点的潜在能力高低作为启发式算法的依据，并依此减少KKT算法阶段的无关用户。

随着影响力最大化问题的发展，关于社会网络的群组性的相关研究[26-30]也慢慢被考虑并关注。基于群组的影响力最大化问题主要是为了通过将网络划分为小一些的群组来提高运算效率。不过基于群组的影响力最大化问题只是为了提高效率而做的一些划分工作，虽然表面上是将网络进行划分，它找到的依旧还是整体的最有影响力节点。

影响力最大化问题还有一些其他相关的研究，例如其他传播模型的影响力最大化算法[31, 32]，影响传播模式因素[33-35]，多流竞争传播研究[36-39]，影响与推荐的关系[40]、影响与相关性的关系[41]等。影响力最大化问题并未设立种子节点是否愿意进行“口碑传播”这个先决条件，Yaron等人[42]针对之前已有影响力最大化问题研究并未考虑初始种子节点进行影响传播的代价的问题，提出了一种能够通过激励来获得用户进行影响传播代价的机制。传统的影响力最大化问题仅仅是针对种子节点集合的大小给予一定的限制，但并未对其他条件例如传播的时间及代价进行约束，这种情况显然在真实情况下是不合理的，因而出现了对于受限影响力最大化问题（Constrained Influence Maximization，CIM）的研究。Nguyen等人[43]对传播代价进行限制，研究在有限条件下如何最大化影响效果。Liu等人[43]对传播时间进行约束，并考虑在有限时间下如何实现影响力最大化效果。近几年仍有大量影响力最大化的相关研究，Guo等人[44]提出了一个本地影响最大化的解决方法，从而避免全局影响最大化对目标节点所产生的偏差。Cheng等人[45]提出了StaticGreedy算法，该算法在选择种子节点的中严格保证了传播算法的子模性，从而解决了可拓展性的精确度的问题。Li等人[46]提出了一致性敏感的贪心算法，从而将一致性作为考虑因素来选择种子节点。Hu等人[47]对两种经典的传播模型进行分析讨论，并揭示了在串并联图中进行影响力最大化分析的时间复杂度是线性的。Cohen等人[48]提出了基于草图的贪心算法SKIM，在效率方面，相对贪心算法可以提高两个量级。Li等人[49]研究地理位置敏感的影响力最大化问题，能够极有效率的估计用户边际影响的上限。Tao等人[50]提出的名叫TP的改进的传播模型能够在O2O环境中准确的描述商品被接受的过程。

上面介绍的相关研究都忽略了在信息传播中的一项重要因素——文本话题信息。据研究所知，用户的权威性和从众性也与话题相关，因此下面本文将介绍话题敏感的影响力分析的相关研究。

### 1.3.3话题敏感的影响力最大化问题

上一小节所介绍的影响力最大化相关研究都没有考虑话题因素，然而真实世界中，话题因素往往在影响传播的过程中扮演者重要的角色。现实生活中，用户节点所具有的影响力并不是在任何方面，而是只针对于一个或几个领域，而每个领域即可认为是一个话题，因此权威度和影响力是与话题相关的。同时，普通用户节点也并不会关注所有领域，而是只会关注某一些领域，所以话题不同，所选择的种子节点也应有所改变。Tang等人在[3]中提出了话题紧密传播(Topical Afﬁnity Propagation，TAP)模型，该模型可以利用用户的话题分布计算获得用户间基于话题的直接和间接影响。Liu等人[51]设计了一个概率衍生模型，利用该模型可以同时获得用户的话题分布和用户间基于话题的相互影响。贺人贵等人[52]通过话题将网络进行划分为若干群组，并在各个群组上进行影响力最大化分析，并将结果进行处理来找到整个网络中的种子节点，以此来提高效率。Zhang等人[53]将用户节点的行为模式作为影响传播的因素之一，并据此提出拓展独立级联模型（Extended Independent Cascade，EIC），并根据该模型特点，提出了GAUP算法，并且利用实验结果证明了用户节点的兴趣能够影响的传播起到至关重要的作用。王旭等人[54]利用TopicalHITS算法来计算用户节点的权威度和中心度并将它们融入到所设计的话题敏感传播模型TSIC中，并设计了启发式算法AHP来提高效率。Barbieri[55]等人提出的AIR模型(Authoritativeness–Interest–Relevance)是目前比较先进的话题敏感模型之一，它首先根据用户之前的消费信息进行参数学习，并且将时间动态性加入其中。Zhang等人[56]提出了一种能够在微博网络中进行影响力分析的话题敏感的解决方案。Li 等人[57]提出了基于关键词的目标影响力最大化问题（KB-TIM），其目的是在与给定的广告相关联的用户中选择能使影响力最大化的种子节点集。Chen等人[58]将MIA模型应用到话题敏感影响力最大化的场景中，并利用剪枝等方法提高效率，实现了在线影响力请求及响应。

尽管上述的已有成果考虑了话题因素，但它们忽略了一个事实：任何商品或事件都涉及多个话题，所以考虑单一话题是片面且不准确的。不仅如此，上述研究都未考虑大规模社会网络的情况，然而现实的发展却截然不同，因此研究大规模社会网络的影响最大化分析已成为当务之急。鉴于已有研究所存在的缺陷，本课题将多话题信息融合到影响力最大化问题中，同时考虑地理位置信息、用户权威度及受影响力，并针对大规模社会网络设计并行算法以提高效率。

## 1.4 论文的主要研究内容

本文以大规模社会网络为研究背景，结合多话题、地理位置等因素，找到最可能接受和传播当前需要推广的商品或事件的种子节点集合，能够应用于“病毒营销”及推荐等方向。

本文主要的研究内容包括：

1. 在经典的独立级联传播模型的基础上加入多话题、地理位置等信息，设计出一个新的多话题敏感的传播模型。

2. 改进经典的影响力最大化问题中的贪心算法，使其更加注重多话题因素，更适合所提出的新模型。

3. 基于Spark设计出一种新的并行化算法，用来缓解在大规模网络下进行影响力最大化分析的效率问题。

4. 提出了一种新的评价度量，该度量将多话题因素考虑其中，突出了多话题对种子节点选择的重要性。

5. 利用真实数据集对所提出的模型及算法进行充分的实验，并用结果来证明新的模型和算法是有效且高效的。

## 1.5 论文组织结构

第1章，绪论。对本课题的研究背景、目的及意义、国内外已有研究现状进行总结，并且概括地介绍了所要研究的内容。

第2章，多话题敏感的传播模型。针对现有模型的不足，结合多话题、位置信息等因素，提出了新的话题敏感的传播模型MTSIC。

第3章，多话题敏感的影响力最大化算法。根据所提出的多话题敏感的传播模型的特点，改进经典贪心算法以突出多话题因素在传播过程中的重要作用。

第4章，基于SPARK的并行算法。针对影响力最大化算法在大规模网络中运行的效率低下的问题，利用SPARK的基于内存计算的机制，提出了基于SPARK的并行算法。

第5章，实验与分析。对实验环境及算法评估度量进行说明，并且从多个角度对所提出的传播模型和影响力最大化算法进行了细致的分析。

第6章，总结与展望。对本课题的工作进行概括总结，针对新的影响力最大化算法及传播模型的缺点提出未来可能进行改进的地方。

# 2 多话题敏感的影响传播模型

传播模型是影响力最大化问题中极其重要的组成部分，它的主要职责是负责模拟影响在网络中的传播过程。当产品或事件在一个特定网络中进行推广时，通常假设每个用户节点将会有两种状态——“激活”（接受该商品或事件）和“未激活”（未接受该商品或事件）。当某个“未激活”用户节点的“激活”的入度邻居逐渐增加时，该节点被激活的可能性也会逐渐增大，但用户节点的状态只可能从“未激活”变成“激活”，而不可逆。若该用户节点也变成“激活”状态，那么它将会继续去尝试激活其“未激活”的出度邻居节点。如果想在网络上推广商品或事件，那么一般情况下，能接受商品或事件并继续传播的节点一定是对该商品或事件感兴趣的。判断节点是否对商品或事件感兴趣的最好的方式就是对它们的话题分布进行分析，因此话题分布作为节点传播能力的一项关键指标，可以很好刻画影响传播过程。除此之外，地理位置信息，节点的权威度和中心度也在传播过程中起到了关键作用，因此也被融入到新的模型之中。本章将对经典的传播模型进行简要介绍，并在其基础上提出融合多话题、地理位置等元素的新的传播模型。

## 2.1 相关问题定义

在介绍传播模型和影响力最大化算法之前，先要列举本文会涉及到的一些基本定义。

**定义2.1 富文本社会网络。**定义为富文本社会网络所抽象的加权有向图，其中为节点集合，表示用户节点之间的链接关系，作为节点相关的文档，*n*表示网络图中用户节点的数目。

**定义2.2 节点的集合。**定义*OutNeighbors(u)*表示为节点*u*的出度节点集合，其具体含义是指在网络中，当以节点*u*为弧头时，其弧尾所连接的节点的集合；定义*InNeighbors(v)*表示为节点*v*的入度节点集合，具体是指以节点*v*为弧尾时，其弧头所连接的节点的集合。

**定义2.3 影响力最大化问题。**经典的影响力最大化问题定义如下，在特定的网络*G*下，设定传播模型*T*以及种子节点数量*k*，需要选取*k*大小的种子节点集*S*，该集合能使获得最大值，其中表示当*S*为种子节点集时的最后被激活节点数，如公式2.1所示。其中，代表所得结果最好的种子节点集。

 (2.1)

**定义2.4 函数子模性。**设为任一函数，表示将全集*U*中的全部子集映射到非负实数值。若满足公式，即边际收益呈递减趋势，则认为具有子模性。其中。

**定义2.5函数单调性。**定义一个任意函数，如果函数具有单调性，则必满足，其中。

**定义2.6 话题分布。**在一个富文本网络中，对于任意，都有与其相关联的*T*维度的向量的话题分布，该话题分布可以利用话题模型对节点的文本进行分析来获得，且满足（），每个表示节点*v*在话题*z*上的强度。

**定义2.7 相似度。**每个话题分布其本质就是一个多维的向量，而由于话题分布的所有子分量的和为1，因此可以用两个话题分布X，Y向量夹角的余弦值Cosine来表示这两个话题分布X，Y的相似度。计算公式如2.2所示，其中n表示向量维数。

**  (2.2)

**定义2.8多话题敏感的影响力最大化问题。**由于每一个商品或事件都包含不同的话题，而针对不同的商品或事件，影响力最大化问题的结果也会有所不同，因此需要将多话题因素作为影响力最大化问题的考虑的因素。本文针对多话题的特点，对多话题敏感的影响力最大化问题进行定义，其公式化的定义如公式2.3所示。

** (2.3)

*SIS*表示最后被激活的节点数与最后被激活的节点相似度总和的乘积，其定义如公式2.4所示。其具体的公式介绍请见3.2章节。

 (2.4)

根据上述定义，问题的输入输出如下所示。

**输入：**

1. 为带权网络图，*V*表示用户节点集，*E*表示用户节点链接关系集合，表示用户节点的文档，表示用户节点的话题分布。
2. 初始节点数*k*，所需推广商品或事件*c*。

**输出：**

大小为*k*的一个节点集合*S*，使得目标函数达到最大。

之后的章节会经常涉及到上述相关定义，到时将不再重述以上定义的解释及说明。

## 2.2 话题模型LDA

LDA话题模型是由David M. Blei等人[5]在2003年提出的一个具有三层变参数的贝叶斯概率模型，全称叫做潜在狄利克雷分配模型，三层变参数分别是单词、话题和文档。LDA涉及到Bayes理论、Dirichlet分布等众多理论，其本身属于[非监督的机器学习](http://en.wikipedia.org/wiki/Unsupervised_learning)技术，用于推断文档集或语料库中所包含的潜在话题。

LDA会将每一篇文档当成一个单词向量，利用这些单词向量来进行复杂的数学计算，这样就使得文本信息转变为易于进行建模的数字信息。一篇文档应包含若干个话题，而单词又是通过话题的概率分布计算获得。使用单词的多项式分布来表现一个话题分布，同样，使用话题的多项式分布来表现一篇文档。因此，LDA的建模过程如下：

1. 从每篇文档所拥有的话题分布中随机选取一个话题；

2. 从第1步所选话题的单词分布中随机选取一个单词；

3. 重复执行1、2步骤直到所有单词都已被遍历。

更详细的说法是，文档集中的每篇文档都可以认为是*T*个话题的多项分布 。而每一个话题又可以认为是单词表中*V*个单词所属的多项分布 。而单词表包含了文档集中所有的不重复的单词，但为了获得更好的效果，一般会将一些常用的停用词去掉。和分别有一个带有超参数和的Dirichlet先验分布。一篇文档中必定包含多个单词，而针对其中的每个单词，都应该从这个文档拥有的多项分布中随机选取话题，之后从话题拥有的多项分布中随机选取单词。重复上述步骤次后，便能得到所需的文档 ，该是文档中的单词数量。图2.1表示了整个生成过程：

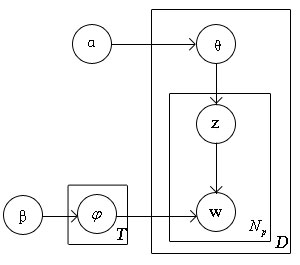


图2.1 LDA模型

上图的表示方法被称作“盘子表示法”。其中，圆圈表示可观测变量和隐含变量；方框代表多次多次抽样，其*D*是抽样次数；箭头表示变量间依赖关系。当利用该模型推断出“论文-话题”分布以及个“话题-单词”分布后，通过对与的学习，就能获得用户的话题分布和文档所包含的话题的比例等。

## Topical HITS算法

已有研究证明，文档除了文本外还包含一些可以表现节点特征的性质。Jon Kleinberg[59]认为文档具有两种潜在的属性：权威度（Authority）和中心度（Hub）。Jon Kleinberg认为，若某个页面具有很高的权威度，那么该页面一定会被许多的中心度节点所链接；同时，若某个页面具有很高的中心度，那么该页面也一定会被许多权威度高的页面所链接。据此，Jon Kleinberg提出了Hyperlink Induced Topic Search（HITS）算法，并指出权威度与中心度之间的关系，如公式2.5所示：

 (2.5)

由于在HITS刚刚被提出的时候，话题因素还并未被当作一项重要因素被大家重视，因此HITS在大多数基于文本的搜索引擎执行任务中的表现很好。但当话题因素越来越被人们所重视，在决定算法效果有着举足轻重的地位时，HITS算法就不再那么适用了。鉴于这种情况，Lan Nie等人[60]将话题因素融入HITS算法中，提出了Topical HITS算法。

在Topical HITS算法中，不再考虑单一的权威度与中心度，而是考虑权威度Authority向量和中心度Hub向量。Authority向量和Hub向量的每个维度都映射一个话题，而维度多少也就是当前文档所包含的话题量。Topical HITS算法中利用多冲浪者语义模型作为随机访问模型。根据冲浪者*A*的行为方式可以获得权威度*A*，同时根据冲浪者*H*的行为方式可以获得中心度*H*。冲浪者*A*在每次行动时有以下两种不同的决断：

1. ：从一出度链接走到其出度邻居节点，且所停留的话题与之前相同的概率为。

2. ：从一出度链接走到其出度邻居节点，且所停留的话题与之前不相同的概率为。

其概率模型如公式组2.6所示：

 (2.6)

与冲浪者*A*相比*，*冲浪者*H*具有两种相反类型的决断。

1. ：沿一入度链接走到其入度邻居，且所停留的话题与之前相同的概率为。

2. ：沿一入度链接走到其入度邻居，且所停留的话题与之前不相同的概率为。

同样，其概率模型如公式组2.7所示：

  (2.7)

根据公式2.8和2.9，可以将冲浪者*A*和*H*之间的交互关系进行描述：

  (2.8)

其中。

 (2.9)

其中。

若忽略话题因素，则以上冲浪者*A*和*H*之间的交互关系将变成一个正则化表示的HITS算法，即公式2.10所示：

 (2.10)

正则化HITS算法的节点的权威度在其入度节点之间平均分配，但经典HITS算法的入度节点便会获得其全部的权威度，对于中心度的计算也是如此，这便是两种HITS算法之间的唯一区别。

然而，并不是所有网络的链接关系都与上文所提到的网络的链接关系相同，因此在计算权威度和中心度时就要做相应的改变。举个例子来说，在DBLP网络中，若论文*x*在文中引用了论文*y*，那么影响的传播方向将会是从*x*到*y*。Twitter网络中也有同样的情况，若用户*x*关注了用户*y* ，则影响会从用户*x*会传播到用户*y*，该影响的传播方向与Twitter网络中的链接方向相反。鉴于此，在影响力最大化问题中使用Topical HITS算法时，将网络链接结构与影响传播结构区别开来就变的十分重要，并且还需根据数据集的不同做相应的调整和处理。本文将会使用上面提到的Topical HITS算法获得的权威度和中心度作为参数应用到所提出的模型中。

## 2.4基于多话题的影响传播模型

由于本文所提出的新模型是建立在经典的IC模型之上，因此需要对IC模型进行简单的介绍。

### 2.4.1 独立级联IC模型

在最早提出的影响力最大化问题中，IC模型作为最经典的传播模型也被一同提出。IC模型是一种概率模型，它将源节点与目标节点之间的概率值作为激活的概率大小。当概率值越大，则表示影响越可能由源节点传达到目标节点。在IC模型中，决定两个用户节点间的传播概率的因素有两个，一个是用户节点间的链接权重，另外一个是初始设定参数*p*，表示影响仅需通过一次节点间交互便能成功传播的概率值。传播概率的计算方式如公式2.11表示：

 (2.11)

在IC模型中，每个用户节点有且只有两种状态，分别是“激活”态和“未激活”态。每当某个用户节点被成功激活，就将它加入到初始激活节点集合，该集合是被所选的初始种子结点所激活的邻居节点集合。在整个传播过程开始前，需要先确定一个激活的种子结点集，然后传播过程按下述方式进行。在时间片t时，若用户节点u被激活，且u拥有一个未被激活的出度邻居节点v，那么u将会尝试以概率激活其出度邻居节点v，但该机会只有一次。若未激活用户节点v拥有多个已激活入度邻居节点时，这些已激活入度邻居节点会以随机顺序去尝试激活用户节点v，若任意一次激活成功，则v将在时间片时转变成“激活”态。当用户节点u尝试去激活用户节点v之后，无论v是否被激活，u将不会再去尝试激活v，即每个用户节点尝试去激活其一个未激活的出度邻居节点的机会仅有一次。当再也没有新的节点被激活时，则认为整个传播过程到此为止。

根据公式2.11能够得出，当用户节点之间权重大小相同时，其影响传播概率也是相等的。因此，当在同一个网络推广不同的产品或事件时，用户节点间的传播概率也是相等的，这显然与真实状况不符。现实情况下，即使用户节点间的链接权重相同，但由于用户节点对所推广的商品的兴趣不同，其影响传播概率也会不一样。当用户节点对所推广的商品极其感兴趣时，它被激活的概率也就会变大。若所推广的商品有所改变，则其影响传播概率也应该随之改变。举个例子来说，若有一位生产商需要利用Twitter网络推广两件商品A和B，由于预算有限额，所以只有少数的用户有机会去免费试用这两件商品。若使用IC模型进行模拟影响传播的实验，则商品A和B所得到的结果会是相同的。这是由于IC模型未考虑话题因素，所模拟出的激活用户可能对这两件商品并不感兴趣，若在真实世界中去推广，其效果可能远远不及预期。俗话说“术业有专攻”，因此针对不同商品与不同用户应该考虑商品与用户的匹配，即用户对商品的感兴趣程度。根据上述介绍，多话题因素在影响力最大化问题中的作用可见一斑，因此下文将提出多话题敏感的独立级联（Multi-topic Sensitive Independent Cascade, MTSIC）模型。

### 2.4.2多话题敏感的MTSIC模型

MTSIC模型提出是基于五个非常直观的现实观察，正是观察2.1至2.5使得MTSIC模型更加符合现实情况，获得更实际的实验结果。

**OB 2.1**如果一个用户在社会网络中大量的发布一件商品或一个事件的相关内容，那么他更容易接受这件商品或这个事件。

**OB 2.2**如果一个用户极其关心一个特定话题，他将更倾向于接受包含这个话题的商品或事件。同时，该话题在话题分布中的强度越强，用户越容易接受。

**OB 2.3**明星或名人会更容易去影响其网络中的关注者，因此这类人群也被称为权威人士，而他们所具有的能够影响他人的能力又叫权威度。若权威人士拥有越多的关注者，他的影响能力越强。

**OB 2.4**用户往往会受到其网络中的好友或所关注的明星的影响，去驱使他们接受一些商品或事件，无论该用户是否对该商品或事件感兴趣，这种会受到他人影响的能力被称作从众性。若该用户在网络中拥有越多的好友或关注越多的明星，则他越可能由于从众性而被影响。

**OB 2.5**当存在地理位置限制的情况时，用户往往只会关注与自身所在地理位置相关的商品或事件并对其进行相应的行为操作。

上面介绍的几个观察中，有几点关键的因素：首先，用户的兴趣和商品的内容基本决定了商品是否会被用户接受。其次，用户具有两个相关的能力，即权威度和从众性。基于以上几点观察，下面将给出MTSIC模型的原理及定义。

在MTSIC模型的整个激活过程包含两个大阶段——多话题激活和邻居互动激活，而多话题激活又包含两个阶段——相似度激活和最突出话题激活。事实上，相似度激活，最突出话题激活和邻居互动激活这三个激活阶段是同时进行的，只要任何一个激活阶段成功，即认为该节点激活成功。整个模型的激活过程如图2.2所示。



图2.2 MISIC模型的激活过程

#### 多话题激活

从图2.2可以看到，整个激活过程是平等的，也就是说，三个激活阶段中的任何一个阶段成功，则整个激活过程就是成功的。其中，前两个阶段与多话题因素相关，因此将其归为多话题激活阶段。

* **相似度激活**

第一个激活阶段叫做相似度激活，当用户节点*u*被激活，若它存在出度邻居*v*，那么首先*u*将以概率去尝试将影响传播给*v*。若*v*收到从*u*传来的影响，则*v*将根据自身条件来判断相似度激活阶段是否能成功。根据观察1所知，用户节点的话题分布与商品的话题分布越相似，则用户越有可能接受这个商品，因此需要计算相似度。既然话题分布其本质就是一个概率向量，而其所有维度分量的和是1，那么可以采用最直接的余弦距离来表示相似度。用余弦距离计算相似度的计算公式由定义2.7给出。得到用户与商品间的话题分布的相似度后，再与设定的阈值进行比较，一旦相似度大于阈值，则该相似度所对应的用户节点将立即被激活。该阈值是一个比1小的小数，并且在不同情况下可能有所不同。根据本文的实验数据及实验结果，这里采用0.85作为本实验的阈值。当相似度小于阈值时，程序将生成一个随机小数与相似度进行比较，如果随机小数小于相似度，则仍然认为相似度激活阶段成功，否则认为相似度激活阶段失败并准备开始最突出话题激活阶段。

* **最突出话题激活**

与相似度激活相同，当*v*收到从*u*传来的影响后，根据自身条件判断最突出话题激活阶段是否能成功。从观察2可以得出，仅仅是单独的一个话题也有可能决定整个激活过程是否是成功的，而这就是最突出话题激活阶段存在的原因。在该阶段的激活过程中，首先要做的就是分别计算用户节点话题分布中的每个子话题与对应商品话题分布中子话题的权重乘积，然后选出所得乘积最大的子话题作为最突出话题。由于该阶段的激活结果直接取决于用户节点话题分布权重与商品话题分布中对应话题的权重，并且话题分布中每个话题分量的值都不会超过1，因此本文简单的将两个话题权重相乘作为最突出话题激活概率。正如相似度激活阶段所做的一样，程序将生成一个随机小数并将之与最突出话题概率相比较，如果小于，则认为激活成功，否则，则认为该阶段激活失败并开始准备邻居互动激活阶段。极端情况下，如果用户只关心唯一的一个话题时，其节点话题分布中对应的话题分量的权重将会接近1，当推荐的商品同样只涉及这一个话题时，所得到的最突出话题激活概率也将会接近1，那么该用户节点将非常有可能被激活，这种情况与真实世界是一致的。

值得注意的是，现实情况下，当且仅当用户第一次收到某一特定商品或事件的讯息后，才会根据的自身的兴趣与商品或事件的内容进行是否接受该商品或事件的判断。若用户再次收到该商品或事件的讯息，由于已经对该商品或事件的喜好做出过判断，因此不会再去重新评判，无论这条讯息是由谁发来的，所以多话题激活对于同一商品和同一用户只做一次判断。但上述情况又分为两种，即第一次评判时，用户是否已经接受该商品或事件。倘若用户第一次已经接受此商品，那么可以认为，该用户的任何入度邻居只要能将讯息传达给该用户，则该用户必定接受并激活。否则，则只能靠邻居互动激活来尝试激活该用户。

#### 邻居互动激活

最后的邻居互动激活阶段包含三个小阶段：推荐阶段、传播阶段和接受阶段。从观察3和4可以得知，即使用户暂时并不对所推广的商品太感兴趣，但该用户仍然有可能受到其邻居的影响并接受该商品。如图2.3所示，一次成功的激活与三个参数相关，因此需要获得这三个参数的激活概率。对于传播阶段，针对节点u和v来说，其传播概率的计算方式与经典的IC模型相同，公式如2.11所示。对于推荐阶段，如果仅仅将话题分布的相似度作为推荐概率的话，将与现实情况有出入。举例来说，如果一位名人u与一位普通人v具有相同的话题分布，如果设相似度作为推荐概率的话，那么这位名人和这位普通人对他们的出度邻居将会有相同的影响力，而实际情况明显是这位名人更具影响力，更能影响关注他的人。因此相对于普通人v来说，名人u在推荐阶段更容易成功。鉴于以上情况，仅仅将相似度作为推荐概率是不合理的。同样地，仅仅将相似度作为接受概率也是不妥的。事实上，每个用户的推荐概率和接受概率是不同的，所以需要找到另外一种方式来同时计算推荐概率和接受概率。



图2.3 邻居互动激活过程

在一个给定的具有影响关系的网络G中，如果节点u有条有向边指向节点v，那么节点u有机会去影响节点v。当节点u有非常强力的意愿去向节点v推荐某样商品，那么节点v将很容易被影响并接受该商品。同时，当节点u的推荐概率很高时，节点v的接受概率也随之提高，反之亦然。上述情况表明，节点的推荐概率与其出度邻居节点的接受概率之间存在一种相互增强的关系。这种关系与Topical HITS算法得到的权威度（Authority）和中心度（Hub）十分相似。因此本文利用Topical HITS算法去为每一个节点计算推荐概率向量*R*和接受概率向量*A*。

在MTSIC模型中，Topical HITS算法所得到的权威度是无法直接被直接用作推荐概率的，这是因为Topical HITS算法所得到的权威度是分话题的。但注意到商品的话题分布的本质其实是描述商品每个话题的强度的概率向量，因此本文利用商品的话题分布和Topical HITS算法获得的权威度去计算用户节点的推荐概率。具体的计算公式如公式2.12所示，其中n表示话题数，代表节点u在话题i下的权威度，表示商品c的话题分布。

** （2.12）

同样地，本文利用商品的话题分布和Topical HITS算法获得的中心度去计算用户节点接受概率，其计算公式如2.13所示。其中表示用户u在话题i下的中心度。

** （2.13）

同经典的IC模型相比，MTSIC模型的激活过程更贴近真实的影响传播方式。与IC模型类似，新模型中的从节点u到v的集成的传播概率如公式2.14所示。

** （2.14）

### 2.4.3基于地理位置的用户节点选择

在绝大多数影响力最大化的情景中，地理位置经常是一项至关重要的因素。举例来说，如果一位生产商想要在某个特定的城市推广一款商品，那么所被影响的用户节点所在的地理位置只能是在该城市的地理范围内。但事实上存在一种特殊的情况，由于选出的种子节点是自动被激活的，因此即使一位用户其本身不属于该城市，但只要其拥有属于该城市的出度邻居节点，这位用户依然有可能被选作种子节点去影响属于该城市的其他用户。再举个例子，如果美国的某个市要进行市长选举，而其中一个候选人想要通过网络来扩大他的影响力来使更多市民为他投票，那么在网络中受到影响的用户节点应该都属于该城市。但是如果一位明星有许多的粉丝属于这座城市，即使该明星本人不属于这座城市，他依然能影响许多人为该候选人投票。鉴于以上情况，当影响力最大化问题有地理位置的限制时，就需要重构网络图。像图2.4所示的那样，新构建的图将包含属于该地理位置范围的所有用户，并且包括那些其本身不属于该地理位置范围但拥有属于该地理位置范围的出度邻居节点的用户。

网络图的重构过程十分简单，所有属于地理位置范围内的节点间的相互关系全部保留，但只保留不属于地理位置范围内的节点的出度关系。这么做的原因是由于，不属于该地理位置范围的节点只可能被选作种子节点而自动激活而不可能被其他节点激活，因此需要去掉这些节点的入度关系。至此，将得到新的网络图供影响力最大化传播使用。



图2.4 网络图重构

## 2.5 本章小结

本章介绍了论文涉及到的一些关键定义，并且详细介绍了对本文非常重要的LDA模型以及Topical HITS算法的原理和推理过程，给出了经典的IC传播模型的概念，并且在经典传播模型的基础上引入多话题、地理位置、权威度和中心度等因素，提出了多话题敏感的MTSIC模型以及模型中涉及到的关键参数的计算方法。下一章将根据本章所提出的MTSIC模型的特点，设计适用的多话题敏感影响力最大化算法。

# 3 多话题敏感的影响力最大化算法

影响力最大化问题的另一重要部分就是影响力最大化算法。当传播模型完成激活过程之后，影响力最大化算法将会从激活的节点中选出初始种子节点集合。已有的影响力最大化算法忽略了多话题等因素，不适用于多话题敏感影响力最大化场景，因此本章在原有的经典影响力最大化算法的基础上融合多话题因素，构思了具有更好效果的影响力最大化算法。

## 3.1 KKT贪心算法

KKT[14]算法是最早被提出用于解决影响力最大化问题的算法，可以得到一个最优解的近似解。其原理十分简单易懂，即让每一个未被激活的非种子节点依次作为种子节点在传播模型上进行蒙特卡洛（Monte-Carlo）模拟传播，模拟传播完成后，统计每个“暂时”的种子节点所能激活的节点数，再将其除以蒙特卡洛模拟次数，来评估每个候选种子节点的影响传播能力。拥有最大影响传播能力的节点将作为该轮的种子节点，并将该节点所激活的节点加入到最后激活的节点集。经过*k*轮种子节点的选取，就能得到所求的种子节点集与最后被激活的节点集。KKT贪心算法的伪代码如算法3.1所示。

虽然KKT算法可以直接与多话题敏感的MTSIC模型相结合来求解影响力最大化问题，但KKT算法并未考虑话题因素，因此得到的效果并不理想。同样，传统的目标函数仅仅考虑了最后激活的节点数，并未考虑多话题因素。鉴于此，本文将提出一个新的目标函数，用以强调话题因素的重要性。由于目标函数所应满足的几个限制，即非负性、单调性及子模性，因此需对新的目标函数进行证明。非负性的意思是，目标函数所得值将永远不小于0；单调性的意思是，目标函数所得值将会随集合中节点数的增多而增大；子模性的意思是，目标函数拥有收益递减的特性，即边际收益会随集合中节点数目的越多而减小。只有满足以上三个条件，目标函数才成立。因此，下一节将详细说明在MTSIC模型下新的目标函数的特性。

|  |
| --- |
| 算法**3.1** KKT算法 |
| **输入：**网络图*G*，种子节点数*k*，模拟次数*M*，节点数*n*  **输出：**大小为*k*的种子节点集合*S* |
| 1： 初始化 ;  2： , *V*是网络*G*的节点集合，*n*为节点数, *Active[n]*是当前种子节点能够激活的节点数  3： **For** i=1 to k{  4： Active[n] 🡨0;  5： **For** r=1 to M{//*M*为蒙特卡洛模拟的次数  6： **For** j=1 to n {//*n*为节点数  7： //*Influence[j]*是在当前的传播图中节点*j*能够激活的节点数。  8： }//**End\_For**  9： }//**End\_For**  10： //找到*Active*中激活节点数总和最大的节点  11： //增加节点*u*到集合*S* 12： }//**End\_For**  13： **return** *S* |
|  |

## 3.2 MTSIC模型的特性

在传统的影响力最大化问题中，最后被激活的节点数往往是最终的目标函数。然而在考虑多话题的情景下，该目标函数就变得并不十分适用。若是仅仅考虑最后被激活节点数，那么将忽略用户节点对商品的兴趣度，使得现实结果与实验结果相差甚远。因此只有将多话题因素融入目标函数中才能使得结果更加贴合实际。刻画用户节点对商品或事件的兴趣的最好的度量就是话题分布的相似度，这也是本文所一直使用的度量。鉴于以上情况，本文设计了一个全新的且适用于多话题情景的目标函数，即当*S*为初始种子节点的时，最后被激活的节点集的节点数与其相对于商品c的相似度累积总和的乘积，其公式已在公式2.4中给出。

由于和恒大于等于0且单调递增，因此的单调性与非负性是极易证明的，或者说是显而易见的，但子模性就需要进行一些推理。证明具有子模性即证明在S为任何集合时，是否成立。但是由于MTSIC模型的传播过程是随机不确定的，因此非常难以确定，所以很难证明上述条件成立。在影响力最大化问题提出的时候，Kempe等人[14]也遇到这个问题，借鉴他们的做法，可以利用一个等价的过程来模拟MTSIC模型的传播过程，该过程如下。

在MTSIC模型中，当用户节点*u*被激活后，若*u*有出度邻居*v*，则*u*会尝试以概率去将影响传播给*v*。若影响传播成功，*v*收到到传播的影响后，会根据自身条件来判断是否接受该影响，此时三阶段激活过程就开始进行了。若三阶段激活过程中任一阶段激活成功，则*v*便激活成功。由于三个阶段的激活过程在真实情况下看作是同时发生且等价的，因此概率是相加的，即，其中表示用户节点*u*与商品c的话题分布相似度，表示用户节点*u*最突出话题与商品c对应话题的乘积。在这种情况下，*u*到*v*的总的激活概率为。倘若已知*v*在多话题激活阶段能够成功激活，则由*u*到*v*的激活概率仅仅取决于。但无论是何种情况，对于任意节点对(*u,v*)，它们的激活概率只取决于其自身，与其他节点相独立。因此以上激活过程就能够看作是抛硬币的过程，该过程跟*v* 被*u*所激活的时间是毫无关系的，无论在任何时间点去投掷硬币，其结果是相同的。当所有“硬币”都抛完，便可开始进行激活统计，如果激活成功，则在新的激活图中为这对节点加一条激活边。若将抛硬币的结果和初始种子节点集*S*保持固定，则即可确定。

抛硬币得到的所有结果所构成的一个可能集合被称作结果概率空间，不妨设*X*为结果概率空间中的任一结果集，则表示未是在初始节点是*S*的情况下，结果集*X*中所包含的最终被激活节点的*SIS*值。由于结果集*X*已经确定，则所获得的*SIS*值也是一个确定的值。现在令表示为，在结果集为*X*的情况下，将u选作种子节点所获得的*SIS*值，则。若用*S*和*T*表示两个集合，并且满足条件，定义为当将节点*u*加入到种子节点集后，所增加的边际收益，该值至少与的值一样大，即不会比小。以上所述满足公式3.1，即目标函数具有子模性。

 (3.1)

而公式3.2给出了 在所有概率空间下的期望值：

  (3.2)

## 3.3 多话题敏感的影响力最大化算法

### 3.3.1 *激活节点数相似度乘积ANSM*度量

由于多话题因素是本文研究的关键所在，而已有的影响力最大化算法都没有考虑多话题因素，因此即使有些已有算法具有很高的效率却依然无法满足需求。鉴于以上情况，本文将改进经典的KKT算法，融入多话题因素，使其与本文所提出的MTSIC模型相匹配。经典的KKT算法仅仅将最后被激活的节点数作为度量来选择种子节点，而最后被激活的节点数并不能刻画用户节点对商品的兴趣度，因此本文提出新的激活节点数相似度乘积度量（Activation Nodes Similarity Product，ANSM）作为选择种子节点的标准，如公式3.3所示。ANSM度量理解起来非常简单，它由两部分组成：最后被激活节点数和相似度的累积度。其中的为一个不大于1的可调节参数，在不同场景下可进行相应修改。例如，如果商品的生产商更加重视受影响用户对商品的感兴趣程度，那么应设置更加接近1；但如果商品的生产商同时关心最后被激活节点数和受影响用户对商品的感兴趣程度，那么就应设为0.5。若将置为0，则该算法将退化成KKT算法。

 (3.3)

### 3.3.2 基于多话题的影响力最大化算法

由于已有影响力最大化算法并没有考虑多话题等因素，因此无法契合MTSIC模型的特点，因此无法应用到基于多话题的影响力最大化研究中。鉴于这种情况，本文提出了名叫ANS（Activation Nodes Similarity）的多话题敏感的影响力最大化算法。由于前面已经给出了ANSM度量，只需要将ANSM度量与经典算法KKT相结合，就能得到ANS算法。所以ANS算法的复杂度与KKT算法相同，即*O(kMn)*。ANS算法的伪代码描述如算法3.2所示。

|  |
| --- |
| 算法**3.2** ANS算法 |
| **输入：**网络图*G*，种子节点数*k*，模拟次数*M*  **输出：**大小为*k*的种子节点集合*S* |
| 1： 利用初始数据构造MTSIC模型;  2： 初始化 ,;  3： **For** i=1 to k{  4： **For** j=1 to M{  5： **For** each {  6： compute its *ANSM(v)*;  7： *Cur\_ANSM(v) += ANSM(v);*  8： }**End For**  9： }**End For**  10： select ;//选择*ANSM*值最大的节点  11： //增加节点*u*到集合*S*  12： //增加节点*u*到集合*S* 13： }**End\_For**  14： return *S* |
|  |

在算法3.2中，*Cur\_ANSM(v)*表示节点v当前的ANSM值。利用堆结构来存储*Cur\_ANSM(v)*值可以更快的找到拥有最大*Cur\_ANSM(v)*值的节点。每轮拥有最大*Cur\_ANSM(v)*的节点将成为种子节点。由于该算法是根据KKT算法改进而来，因此KKT算法所存在的效率问题对ANS算法来说依然存在。为了缓解这个问题，本文将基于SPARK设计多话题敏感影响力最大化并行算法来提高效率。

## 3.4 本章小结

本章首先对影响力最大化问题中经典的贪心算法KKT进行了详细阐述，且根据MTSIC模型的多话题等特点提出了新的度量并对其子模性进行证明。针对已有影响力最大化算法无法满足多话题因素的问题，提出了ANSM值并将其与贪心算法相结合，给出了能够与MTSIC模型完美匹配的改进的影响力最大化算法ANS，且描述了详尽的算法流程，及时间复杂度分析。下一章将针对在大规模网络上进行影响力最大化分析的效率问题设计基于SPARK的多话题敏感影响力最大化并行算法。

# 4 针对大规模社会网络的影响力最大化并行算法

众所周知，运行影响力最大化算法程序是一件非常耗时的事，而在大规模网络上进行影响力最大化分析的运行效率更是低下，因此许多已有研究针对影响力最大化算法的低效率问题提出了不同的解决方法，例如启发式学习算法、剪枝算法、并行算法等等。本章将针对新模型与算法的特点，给出并行化的影响力最大化算法。目前，市面上有着各种各样的并行计算框架，而最夺人眼球的莫过于基于内存计算的SPARK框架，其适用于图迭代的特性非常有利于影响力最大化问题效率的提高。

## 4.1 SPARK并行计算框架

SPARK[61]是基于内存计算的开源的大数据并行计算框架，其研究的原因是为了更快速的分析数据。SPARK具有高可伸缩性和高容错性，这使得用户能够将SPARK部署在大量廉价的硬件设备上来搭建集群。SPARK是基于SCALA 语言实现，并将SCALA作为它应用程序框架，这使得SCALA可以对分布式数据集进行操作时如同操作本地集合对象那样轻松。

SPARK之所以能够有很高的效率，一个很重要的原因是其引入了一个被称为弹性分布式数据集（RDD，Resilient Distributed Dataset）的概念。RDD具有不可变、高容错、分布式的性质，这使得用户可以轻松高效的并行操作它。不仅如此，任何类型的对象都能够转变成RDD，并支持转换和行动两种操作。转换（例如过滤、联接、映射等）用来在一个RDD上执行操作，再生成新的RDD来保存结果；行动（例如计数、归并等）用来在一个RDD上进行某种计算，之后将结果返回。

SPARK与另一个开源项目HADOOP非常相似，但两者之间还是存在这一些不同。首先，基于MapReduce计算引擎的HADOOP往往需要将中间数据重新写回到HDFS，而SPARK会将中间数据保留在内存中，避免了读写硬盘造成的效率低的情况。其次，RDD作为SPARK的分布式数据结构，可以作为分布式索引，并且具有很高的操作效率。最后，SPARK在Shuffle中的排序策略与HADOOP不同，即并不是任何场景都需要进行排序，并且SPARK可以通过线程池复用线程来避免线程的启动和切换开销。

目前SPARK已成为一种非常成熟的技术，许多知名企业都运用SPARK来进行大数据处理。因此本文将基于SPARK提出多话题敏感的影响力最大化并行算法PMT。

## 4.2 多话题敏感影响力最大化并行算法PMT

由于影响力最大化问题步骤繁杂，因此并不是整个影响力最大化模型及算法都可以进行并行化处理。考虑到每一阶段的特性与并行化的要求，本文提出了多话题敏感影响力最大化并行算法（Parallelization of Multi-Topic algorithm，PMT），该并行算法对整个程序流程中的三部分进行并行化处理。

### 4.2.1 相似度计算的并行化处理

由于考虑多话题因素，因此在整个影响力最大化的过程中，用户节点与商品的话题分布的相似度被应用于各个地方。但考虑到大规模网络中的节点数可能极其庞大且相似度的浮点数计算较为复杂，因此若要计算每个用户节点与商品的话题分布的相似度也将会十分耗时。鉴于每个用户节点的话题分布是与其他用户节点相独立的，符合并行化的要求，所以本文将相似度的计算工作进行并行化，以提高效率，节约时间。

### 4.2.2 用户节点激活的并行化处理

用户节点激活过程是影响力最大化算法中重要的一环，同时也是算法中较为耗时的一个阶段。在激活过程开始前，激活所需要的所有条件都已经准备好了。在激活时，即使信息是否能够传达到一个特定的用户节点取决于该用户节点的入度邻居，但该节点是否能被激活则完全取决于其本身。因此，当获得了和后，整个激活过程可以完全地并行化。

### 4.2.3 边际影响统计计算的并行化处理

当激活过程结束之后，将迎来影响力最大化流程中另一重要的阶段，即边际影响统计计算阶段。在传统的流程中，需要从网络中的每个用户节点开始，逐一遍历整个激活节点网络图，即便是在普通网络中进行此操作也将十分耗时，更不用说是在大规模网络中。在研究[18]中，一个基于GPU的名叫*Bottom-Up Traversal Algorithm*的算法被提出，用并行化的方式来解决这个问题。由于该算法十分巧妙且同样可以移植到SPARK，因此本文对其进行稍许改进并应用于PMT算法中。根据影响力最大化问题的特点，可以看出每个用户节点的边际影响统计只取决于其出度邻居。在本文所提出的算法中，网络图中的每个用户节点都有一个入度邻居节点集合一个出度邻居节点集。入度邻居节点集包含了所有能够访问当前用户节点的邻居的编号而出度邻居节点集包含了当前用户节点能够直接访问到的所有邻居的编号。利用这两个集合，就可以将所有的用户节点进行等级划分。首先，当且仅当某一用户节点没有任何出度邻居时，该用户节点被标记为等级0。找到所有等级0的用户节点后，遍历每个0级节点用户节点的入度邻居节点集，并将遍历的每个用户节点的等级设为1。然后重复上述操作直到网络中所有的用户都被标记。在这个过程中，由于一个用户节点可能是多个不同等级的节点的入度邻居，因此该用户节点可能会被标记多次，而该节点的最终等级取决于其最后一次的等级，该等级同样也是该用户节点获得的最大的等级。但标记过程是否成功取决于当前网络图是否是有向无环图（DAG）。若不是，则标记过程将会永远执行下去，因此如何确定网络图为有向无环图就变得至关重要。幸运地是，已有研究已经能够找到一个网络图中的所有强连通分量（SCC）。从研究[18]可知，在强连通分量中的用户节点的边际影响与该强连通分量的中的节点数相等，所以对于同一强连通分量，只需计算其中一个用户节点的边际影响，而无需全部计算。因此需要将属于同一强连通分量的全部用户节点整合成一个具有强连通分量节点数权重的节点。通过这种做法，可以避免大量的非必要的计算。图4.1是一个等级划分的有向无环图。

但是仍有一个问题没有解决，那就是重叠问题。图4.2中有4个用户节点，*D*为等级0，*B*和*C*为等级1，*A*为等级2。*D*是*B*和*C*的共同的出度邻居，这就造成了重叠问题。当计算*A*的边际影响时，不能仅仅将*A*和*A*的出度邻居的边际影响的和作为*A*的边际影响，还需用所得的和减去重叠的*D*的边际影响。那么，这个问题就转换成了如何确定重叠的用户节点。在PMT算法中，每个用户节点都有一个可访问集合用来存储该节点能够访问到的所有的不重复的节点的编号，只需要利用该集合即可找到重叠的用户节点。通过以上分析可以看出，属于同一级别的用户节点的边际影响的计算过程是相互独立的，这是因为一个节点的边际影响只取决于其可访问的低等级的节点及其本身，因此可以将同等级节点的边际影响计算进行并行处理。



图4.1 等级划分的有向无环图



图4.2 重叠问题

## 4.3 PMT算法的流程

通过上述介绍与分析，这里将给出PMT算法整体的算法流程。

|  |
| --- |
| 算法**4.1** PMT算法 |
| **输入：**网络图*G*，种子节点数*k*，模拟次数*M*，商品c  **输出：**大小为*k*的种子节点集合*S* |
| 1： 利用地理位置信息重构网络图G;  2： 初始化 ,;  3： **For** each {  4： 计算相似度 **In Parallel**  5： }**End For**  6： **For** i=1 to k{  7： **For** j=1 to M{  8： **For** each {  9： 激活计算 **In Parallel**  10： }**End For**  11： 利用激活节点创建激活节点图  12： 将激活节点图重构成DAG  13： **For** each {  14： 标记*v*的等级  15： }**End For**  16： **For** each level{  17： 计算当前等级中每个节点的边际影响 **In Parallel**  18： }**End For**  19： 对结果进行统计  20： }**End For**  21： select ;//选择*ANSM*值最大的节点  22： //增加节点*u*到集合*S*  23： //增加节点*u*到集合*S* 24： }**End\_For**  25： return *S* |
|  |

通过上述的详细的PMT算法流程可以看出，核心部分为三层迭代，外层迭代次数为种子节点的数目，中层迭代次数为模拟次数，内层迭代次数为节点的数目。算法具体的加速效果将在下一章介绍。

## 4.4 本章小结

本章提出了基于SPARK的多话题敏感影响力最大化并行算法，以解决多话题场景下的影响力最大化分析的运行效率的问题。首先给出了SPARK计算框架的简介，原理等，然后根据SPARK的原理和本文所提出的基于多话题的传播模型和影响力最大化算法的特点给出了基于SPARK的并行的PMT算法，并详细介绍了该算法中可并行部分的原理与过程。之后给出了PMT算法的整体流程与时间复杂度。下一章将给出在真实数据集下MTSIC模型、ANS算法及PMT算法的实验结果，并给出详尽的实验分析。

# 5 [实验与分析](#_Toc283328287)

本章将简要地对实验环境及数据情况进行介绍，并根据现有模型和算法的特点提出新的评价度量，并在真实数据集上设计并实施实验来验证MTSIC模型及ANS、PMT算法的效果和效率。

## 5.1 实验环境及数据情况

### 5.1.1 实验环境

由于实验环境的不同可能导致实验结果的不一致，因此首先要介绍一下本文实验所处的环境，如表5.1所示：

表5.1 运行环境

|  |  |  |
| --- | --- | --- |
| 单机 | 硬件平台 | 联系Y400，i5 3230，8G内存，1T外存 |
| 操作系统 | Windows 8 家庭版 |
| 开发工具 | Eclipse，Matlab2013b |
| 分布式 | 硬件平台 | DELL服务器，4核，16G内存，1TB外存，4台 |
| 操作系统 | CentOS |
| 开发工具 | Hadoop2.2.0，Spark0.9.0 |

### 5.1.2 实验数据集

本文实验所采用的数据集为Arnetminer[[1]](#footnote-1)和Twitter[[2]](#footnote-2)平台数据集，其中Arnetminer数据集又被称为DBLP，该数据集包含了从2008年到2014年间发表过论文的作者间的引用关系，并消除了部分孤立的作者节点。而Twitter数据集是佛罗里达州2014年1月29号到3月4号的部分用户的Tweet博文数据，并同样进行了优化处理。两个数据集的详细情况如表5.2所示：

表5.2 实验所用数据集

|  |  |  |  |
| --- | --- | --- | --- |
| 属性  数据集 | 节点数 | 边数 | 边语义 |
| DBLP | 254495 | 1641978 | 作者间引用关系 |
| Twitter | 23882 | 24527 | 关注关系 |

由于需要获得每个用户节点的兴趣情况，所以需要借助话题模型LDA来计算话题分布。若用户节点的文本信息越丰富，则获得的话题分布就越准确，所以需要对用户文本信息进行汇总和处理。对于DBLP数据集，需要汇总每位被选中的作者在08-14年间发表论文的情况；而针对Twitter数据集，由于每条博文最多只含140个字，因此需要汇总每位被选中的用户在14年1月29号到3月4号间发表的博文的内容。汇总的文本信息中存在大量的无用信息，例如a，the，of等，因此需要对文本信息进行停用词处理。之后，分别利用话题模型去计算DBLP数据集和Twitter数据集话题数为5、10、20、30的用户话题分布，为之后的实验做准备。

由于MTSIC模型中还需要获取用户的权威度和中心度，因此在通过话题模型获得用户节点的话题分布后，再将该话题分布与用户间的链接关系作为参数，运行Topical HITS算法，来获得每个用户节点的权威度与中心度的值。

## 5.2 实验评价度量

影响力最大化问题最早被提出时，用来评价传播模型和影响力最大化算法的优劣的评价度量为最后被激活的节点数。然而随着话题因素在影响力最大化问题中扮演越来越重要的角色，最后被激活节点数就变得并不能完全体现结果的优劣。举个例子来说，如果一位生产商想要通过“病毒式营销”在网络中推广一款产品，那么他最终的目的是希望网络中的用户对他的商品感兴趣甚至来购买该款商品，而不是只是过眼云烟，转瞬即逝。因此，该生产商找到的群体应该既有足够多的受众，同时也对该商品有足够的兴趣。据此，根据多话题因素的特点，新的评价度量应该同时关注两部分内容：最后被激活节点数与最后被激活节点累积相似度。因此，本文提出新的评价度量激活节点数与累积相似度乘积（Similarity Influence Sum，*SIS*）作为实验的评价度量来判断传播模型和影响力最大化算法的好坏，公式如2.4所示。

表示当种子节点集为*S*时，最后被激活节点数与最后被激活节点与商品*c*的累积相似度的乘积，其中表示节点*v*与商品*c*的话题分布相似度。当越大，表明最后被激活节点集在节点数量较大的同时对商品较为感兴趣。通过该度量标准，可以准确评价各个传播模型和影响力最大化算法在多话题因素下的优劣情况。

## 5.3 实验方案设计

实验进行前首先要对实验中所需的参数进行初始参数设置。对于传播概率*p*而言，在经典的IC模型中，该值为0.01[14]，因此本文实验也沿用IC模型的设置。由于实验需要进行模特卡洛模拟（Monte-Carlo, MC），并对模拟次数进行设定。而根据已有经验[14]，当模拟次数超过1万次以上时，实验结果变化极小，因此本文实验设定模特卡洛模拟次数为1万。在MTSIC模型中，在相似度激活阶段需要设定一个阈值来作为节点是否能在该阶段被激活的判断。该阈值的设定需要仔细的考量，倘若设的过大，则很难有用户与商品的话题分布有如此高的契合度，激活效果将变差；倘若设置的过小，则大量的用户将在该阶段被激活，可能会与现实情况有所出入。因此，MTSIC模型将采用PageRank算法中经过大量实验考量的阻尼系数0.85作为阈值。由于已有研究都未考虑多话题因素，很难进行实验与它们进行比较，因此本文将仅与最经典且效果不会被超越的KKT算法以及研究[54]所提出的AHP算法进行对比。

本文的实验将会以以下几个方向展开：首先，需要综合的进行考量，分别对MTSIC模型结合ANS算法、IC模型结合KKT算法以及TSIC模型结合AHP算法进行传统度量与新度量下的结果比较。其次，分别统计MTSIC模型中三个激活阶段中激活的节点数目，据此来体现MTSIC算法的有效性。再次，将MTSIC模型分别于ANS算法和KKT算法结合，进行实验来比较KKT算法与ANS算法所选择的最后激活的节点集，判断是否ANS算法能够更有针对性的找到更适合所推广商品的用户集合，判断ANS算法的有效性。最后，将并行算法PMT与单机算法运行结果进行比较，以此证明PMT算法的效率。

## 5.4 实验结果及分析

### 5.4.1 MTSIC模型结合ANS算法的有效性

表5.3展示了MTSIC模型结合ANS算法在DBLP数据集上进行话题数分别为10、20、30，种子节点数为5时的结果。

表5.3 MTSIC结合ANS在DBLP上不同话题数下的结果

|  |  |  |  |
| --- | --- | --- | --- |
| 话题数 | 最后激活节点数 | 累积相似度 | 平均相似度 |
| 10 | 9685.7 | 6339.975 | 0.6546 |
| 20 | 5434.85 | 2807.267 | 0.5165 |
| 30 | 3149.27 | 1397.646 | 0.4438 |

表5.4展示了IC模型结合KKT算法在DBLP数据集上进行话题数分别为10、20、30，种子节点数为5时的结果。

表5.4 IC结合KKT在DBLP上不同话题数下的结果

|  |  |  |  |
| --- | --- | --- | --- |
| 话题数 | 最后激活节点数 | 累积相似度 | 平均相似度 |
| 10 | 13021.4 | 1858.308 | 0.1427 |
| 20 | 12451.6 | 889.7557 | 0.0715 |
| 30 | 12679.4 | 267.6208 | 0.0211 |

根据表5.3与5.4所示，考虑到KKT算法在最后激活节点数指标上是最好的且无法被超越的算法，MTSIC模型结合ANS算法在该指标上与IC模型结合KKT算法的差距并不大；在累积相似度指标上，MTSIC模型结合ANS算法的效果远超IC模型结合KKT算法，在平均相似度上也有相同的结果，这说明MTSIC结合ANS算法所找到的激活节点是对所推广的商品或事件非常感兴趣的，而在现实情况下，这些用户节点会更容易接受该商品或事件。图5.1给出了MTSIC模型结合ANS算法与IC模型结合KKT算法在DBLP数据集上*SIS*度量下的实验结果。可以看出，MTSIC模型结合ANS算法在*SIS*度量下的结果是要明显优于IC模型结合KKT算法的，尤其是当话题数较少时。当话题数增加时，由于话题数的增加使得用户节点与商品或事件的话题分布相似度减小，使得MTSIC模型的激活效果变差。但从另外的方面考虑，话题数越多，所找到的最后激活节点在现实情况下就越对该商品或事件感兴趣，更有针对性。



图5.1 MTSIC结合ANS与IC结合KKT在DBLP上*k*=5时*SIS*值的比较

同样地，表5.5和5.6显示了在Twitter数据集上不同话题数下MTSIC模型结合ANS算法与IC模型结合KKT算法的结果。可以看出，当话题数为5、10时，实验结果与在DBLP数据集上的实验结果相似。但当话题数为20和30时，无论是MTSIC模型结合ANS算法还是IC模型结合KKT算法，它们的结果都有所反常，即话题数增多反而效果变好。这是因为每次实验的结果不仅取决于网络中用户节点的链接及属性，还取决于所推广的商品或事件的话题分布。造成话题数为20和30的情况下，结果变好的原因是，相对于话题5和10，话题数为20和30时所选的推广商品或事件在网络中被更多的用户所感兴趣，因此无论是最后激活节点数还是累积相似度都有所提高，因此出现反常。

表5.5 MTSIC结合ANS在Twitter上不同话题数下的结果

|  |  |  |  |
| --- | --- | --- | --- |
| 话题数 | 最后激活节点数 | 累积相似度 | 平均相似度 |
| 5 | 259.14 | 235.4608 | 0.9086 |
| 10 | 167.20 | 78.9872 | 0.4724 |
| 20 | 177.68 | 88.7597 | 0.4995 |
| 30 | 192.95 | 121.4363 | 0.6294 |

表5.6 IC结合KKT在Twitter上不同话题数下的结果

|  |  |  |  |
| --- | --- | --- | --- |
| 话题数 | 最后激活节点数 | 累积相似度 | 平均相似度 |
| 5 | 269.29 | 64.8799 | 0.2409 |
| 10 | 259.03 | 48.0997 | 0.1853 |
| 20 | 261.18 | 32.0462 | 0.1227 |
| 30 | 267.37 | 50.9103 | 0.1904 |

图5.2给出了MTSIC模型结合ANS算法与IC模型结合KKT算法在Twitter数据集上*SIS*度量下的实验结果。可以看出，在*SIS*度量下，MTSIC模型结合ANS算法具有更好的效果，但结果依然受到所推广的商品或事件的话题分布的影响。

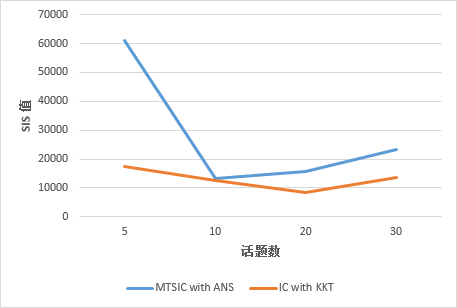


图5.2 MTSIC结合ANS与IC结合KKT在Twitter上*k*=5时*SIS*值的比较

表5.7展示了TSIC模型结合AHP算法在Twitter数据集上进行话题数为10，种子节点数为5时的不同话题的结果。

表5.7 TSIC结合AHP在Twitter上话题数为10时不同话题下的结果

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 话题 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 最后被激活节点数 | 65.55 | 93.4 | 56.5 | 61.93 | 71.22 | 73.63 | 83.12 | 57.52 | 92.08 | 76.79 |
| 累积相似度 | 15.74 | 23.43 | 7.351 | 11.31 | 12.52 | 17.47 | 13.47 | 10.76 | 15.43 | 14.48 |

图5.3显示了MTSIC模型结合ANS算法与TSIC模型结合AHP算法在Twitter数据集上话题数为10时，不同话题下*SIS*度量的结果。

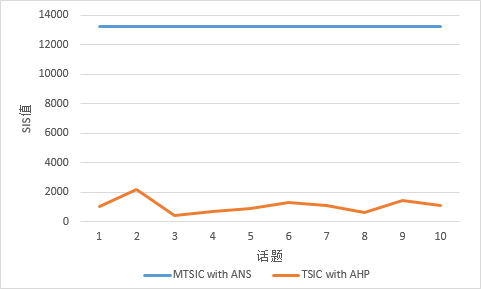


图5.3 MTSIC结合ANS与TSIC结合AHP在Twitter上话题数为10、*k*=5时*SIS*值比较

由于MTSIC模型结合ANS算法是考虑的多话题因素，而并不是考虑单一话题，因此将其与考虑单话题的TSIC模型结合AHP算法进行比较是有一定困难的。这里，本文选择Twitter数据集并选择话题数为10，分析TSIC模型结合AHP算法在每个分话题下得到的结果与MTSIC模型结合ANS算法在全部10个话题下得到的结果的优劣。如表5.7显示了TSIC模型结合AHP算法任一分话题下所得到的最后被激活节点数与其累积相似度都远远不如MTSIC模型结合ANS算法在考虑多个话题下所得结果。图5.3显示了TSIC模型结合AHP算法在每个分话题上所得*SIS*值，由于MTSIC模型结合ANS算法是考虑多话题，因此在图中表示为一条水平线，以作对比。很明显，MTSIC模型结合ANS算法在多话题下所得结果要明显好于TSIC模型结合AHP算法在分话题下所得结果。

### 5.4.2 MTSIC模型的有效性

5.4.1节单单从整体来体现MTSIC模型结合ANS算法的有效性，但无法细致的体现出MTSIC模型各个阶段的激活情况，也就无法体现MTSIC模型的每个激活阶段是否都有不可或缺的作用，因此需要对MTSIC模型的三个阶段分别统计其激活的情况。表5.8展示了MTSIC模型在DBLP数据集上不同话题数下，三个激活阶段所分别激活的情况。图5.4给出了MTSIC模型在DBLP数据集上不同话题数下三个激活过程所激活情况的趋势变化。

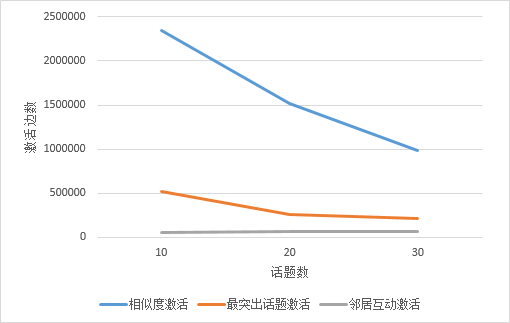


图5.4 MTSIC在DBLP上不同话题数下三个激活过程所激活情况的变化趋势

表5.8 MTSIC模型在DBLP上不同话题数下激活情况

|  |  |  |  |
| --- | --- | --- | --- |
| 话题数 | 相似度激活 | 最突出话题 | 邻居互动 |
| 10 | 2346496.7 | 525065.4 | 56749.8 |
| 20 | 1521344.4 | 260998.1 | 69459.7 |
| 30 | 987411.3 | 214026.5 | 62298.7 |

表5.9展示了MTSIC模型在Twitter数据集上不同话题数下，三个激活阶段所分别激活的情况。

表5.9 MTSIC模型在Twitter上不同话题数下激活情况

|  |  |  |  |
| --- | --- | --- | --- |
| 话题数 | 相似度激活 | 最突出话题激活 | 邻居互动激活 |
| 5 | 70663.9 | 11969.2 | 1647.83 |
| 10 | 31354.22 | 6189.14 | 10077.71 |
| 20 | 34573.56 | 4405.99 | 10060.35 |
| 30 | 42984.98 | 5509.61 | 8471.54 |

图5.5给出了MTSIC模型在Twitter数据集上不同话题数下三个激活过程所激活情况的趋势变化。

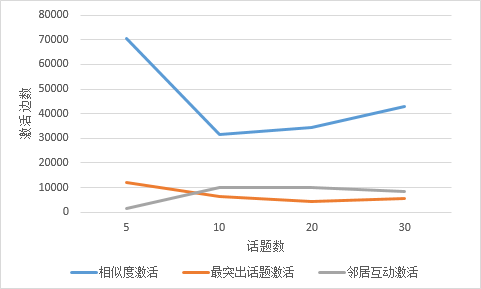


图5.5 MTSIC在Twitter上不同话题数下三个激活过程所激活情况的变化趋势

需要说明的是，表格中的激活数并不是指的激活节点数，而是激活的边数。根据表5.8和表5.9所展示的结果，可以看出，无论是在DBLP数据集上还是在Twitter数据集上，相似度激活过程所激活的边数占据总激活的边数的绝大部分，超过80%的比例；而邻居互动激活过程占据总激活的边数的小部分，大约在5%-15%之间；而最突出话题激活过程所占总激活的边数比例最少，大概在5%-10%之间。上面情况产生可能是以下原因的结果：首先，在所选网络中有大多数用户节点对所选择进行推广的商品或事件感兴趣，因此相似度激活过程就变得非常顺利；其次，仅有很小部分用户节点所感兴趣的某一单独话题刚好在所推广的商品或事件的话题分布中具有较高的权重并且并未在相似度激活过程中被激活，因此该小部分用户节点最终通过最突出话题激活过程所激活；最后，由于已有部分用户节点通过相似度激活过程或最突出话题激活过程被激活，而这部分节点将不会被纳入邻居互动激活过程的激活结果，而且所选网络中可能绝大多数的用户的权威度和中心度并不高，因此造成邻居互动激活过程所激活的边数所占比例并不十分显著，但仍占有一定分量。倘若所选网络中的用户节点的平均权威度与中心度都相对较高，即网络中的用户间都相互信任且用户间的链接丰富，则邻居互动激活过程将具有更好的效果。

以上实验结果证明，整个多话题激活过程在MTSIC模型中起到了至关重要的作用，尤其是相似度激活过程，这也从侧面证明了MTSIC模型的有效性。

表5.10 MTSIC与IC在Twitter上不同话题数下最后被激活节点符合地理位置限制比例

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 话题数  模型 | 5 | 10 | 20 | 30 |
| MTSIC | 99.58% | 100% | 100% | 99.51% |
| IC | 20.29% | 21.93% | 18.02% | 14.87% |

而地理位置因素同样是MTSIC模型中一项重要的因素，因此也需要验证其作用。由于DBLP数据集中并不包含地理位置信息，因此这里只使用Twitter数据集来进行实验。通过对整个Twitter数据集进行统计，数据集中所有用户的最高经度为36.67798，最低经度为24.578373，最高维度为-75.65408，最低维度为-90.21815。可以选择其中的一块区域作为地理位置限制，本实验选择的经度区间为28到33，维度区间为-87到-80。表5.10给出了利用MTSIC模型及IC模型所得最后被激活节点中属于所限定区域的节点比例。可以看出由于MTSIC模型会根据地理位置限制对网络进行重新构建，因此无论话题数为多少，最后被激活节点都基本属于所限定区域。之所以当话题数为5和30时，有极少节点并不属于所限定区域，这是因为这些不属于所限定区域的节点是被选作种子节点而主动激活的。这也验证了即使节点其本身不属于该限定区域，但只要其拥有属于该限定区域的出度邻居节点，这位节点依然有可能被选作种子节点去影响属于该限定区域的其他节点这一观点。反观IC模型的结果，其大部分最后被激活节点都不属于所限定区域，由于现在真实情况中，若影响到达限定区域外的节点，则该影响传播很有可能被中断，因此实际结果将远远不如实验结果。综上所述，MTSIC模型更加贴近实际，具有更高的有效性。

### 5.4.3 ANS算法的有效性

由于已有影响力最大化算法并未考虑多话题因素，并不适用于融合了多话题因素的MTSIC模型，因此本文提出了基于多话题的影响力最大化算法ANS。但ANS算法是否真的能够根据需求找到所期望的种子节点，这就需要对ANS算法的有效性进行证明。通过对*ANSM*度量中的参数值的改变，来记录ANS算法在不同累积相似度比例下所找到的最后被激活节点集，并将其与KKT算法所找到的最后被激活节点集进行比较，看是否ANS算法所找到的最后被激活节点与KKT算法所找到的有所不同，以此来证明ANS算法相对于KKT算法能够更好的选择所需的种子节点，即ANS算法的有效性。

表5.11给出了ANS算法与KKT算法在DBLP数据集与Twitter数据集上不同话题数下最后被激活节点集的重合程度。

表5.11 ANS与KKT在不同数据集上不同话题数下最后被激活节点重合度

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 话题数  数据集 | 5 | 10 | 20 | 30 |
| DBLP |  | 74.97% | 71.32% | 65.86% |
| Twitter | 89.43% | 71.03% | 64.17% | 58.29% |

表5.12给出了ANS算法与KKT算法在DBLP数据集上不同话题数下未相互包含的最后被激活节点的平均相似度情况。

表5.12 ANS与KKT在DBLP、不同话题数、未互相包含的最后被激活节点平均相似度

|  |  |  |  |
| --- | --- | --- | --- |
| 话题数  算法 | 10 | 20 | 30 |
| ANS | 0.1846 | 0.1415 | 0.1286 |
| KKT | 0.1749 | 0.1324 | 0.1041 |

表5.13给出了ANS算法与KKT算法在Twitter数据集上不同话题数下未相互包含的最后被激活节点的平均相似度情况。

表5.13 ANS与KKT在Twitter、不同话题数、未互相包含的最后被激活节点平均相似度

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 话题数  算法 | 5 | 10 | 20 | 30 |
| ANS | 0.3086 | 0.2069 | 0.1471 | 0.1641 |
| KKT | 0.2945 | 0.1541 | 0.1048 | 0.1240 |

图5. 6给出了ANS算法与KKT算法在DBLP数据集与Twitter数据集上不同话题数下最后被激活节点集的重合度趋势。

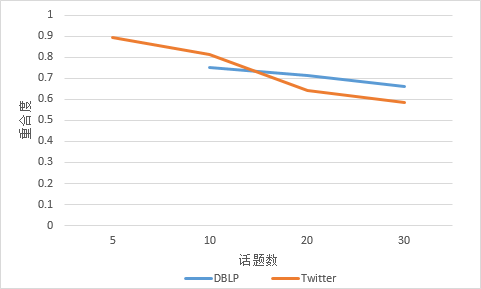


图5.6 ANS与KKT在不同数据集、不同话题数下最后被激活节点集的重合度趋势

以上都是在值为0.5时所得结果。从表5.10可以看出，虽然ANS算法与KKT算法所挑选的最后被激活节点有部分是相同的，但仍有一部分ANS算法所挑选的节点是KKT算法的结果未包含的。从表5.12和表5.13可知，无论话题数是多少，ANS算法所选择的那些KKT算法未选择的节点的平均相似度都高于KKT算法所选择的那些ANS算法未选择的节点。这从侧面证明ANS算法所选择的用户节点对所推广的商品更加感兴趣，若在真实世界中也就越容易接受所推广的商品或事件。而从图5. 6可以看出，当话题数逐渐增多时，ANS算法与KKT算法所选择的最后被激活节点的重合度越来越小，说明所选择的节点的差异也变得越来越大。这是因为当话题数增多时，用户节点的兴趣被划分的更加细致，在KKT算法只关注最终被激活节点数的同时，ANS算法还需关注被激活节点与所推广商品或事件的话题分布的相似度，因此造成ANS算法与KKT算法所选最后被激活节点集的差异。

表5.14给出了选择DBLP数据集和Twitter数据集，并设置话题数为10时，当改变ANS算法中*ANSM*度量的参数值的时候，ANS算法与KKT算法所选最后被激活节点集的差异变化情况。

表5.14 ANS在不同数据集、话题数为10、不同值时与KKT的最后被激活节点集重合度

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 值  数据集 | 0.2 | 0.5 | 0.7 | 0.9 | 1 |
| DBLP | 78.76% | 76.82% | 74.97% | 73.41% | 71.73% |
| Twitter | 82.98% | 81.50% | 78.92% | 77.90% | 78.22% |

图5.7给出了ANS在不同数据集上话题数为10下不同值时与KKT的最后被激活节点集的重合度变化趋势。

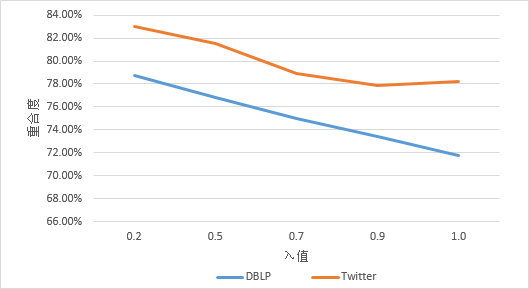


图5.7 ANS在不同数据集、话题数为10、不同值时与KKT的最后被激活节点集的重合度变化趋势

从表5.14和图5.7可以看出，当逐渐增大时，ANS算法与KKT算法所选择的最后被激活节点集的差异也在逐渐变大。这是由于，值表示的是在*ANSM*度量中累计相似度所占的比重。当逐渐增大时，ANS算法将逐渐倾向于去寻找具有更高累计相似度的用户节点作为种子节点，此时所选的种子节点就很有可能与KKT算法所选种子节点不同，因此造成ANS算法与KKT算法所选择的最后被激活节点集的差异逐渐增大的结果。这也证明了ANS算法更适合去寻找对所推广商品或事件感兴趣的用户节点，而不是只关注最后被激活节点数，同时也证明了ANS算法的有效性。

### 5.4.4 PMT算法的有效性

在大规模网络上研究影响力最大化问题是十分效率低下的，需要针对效率问题提出合理的解决方案。相对于串行算法，本文提出的并行的多话题敏感的影响力最大化算法PMT能够提高有效提高加速比。所使用的分布式集群配置如5.1.1所示。实验证明，在DBLP数据集上，当话题数为10时，PMT算法具有2.89倍的加速比；而当话题数为20时，PMT算法具有3.07倍的加速比。之所以话题数越多时加速比也有所提升，是因为并行计算的部分中的大量工作是与话题分布相关的，因此话题数越多，平分给集群的计算量就越大，加速效果也就越显著。

## 5.5 本章小结

本章首先对实验环境及数据情况进行了简要的介绍，然后根据多话题因素提出了新的度量*SIS*并对实验方案进行设计。整体的实验总共分为四个部分，分别是对MTSIC模型结合ANS算法从不同话题数、不同数据集上证明其组合的有效性；根据MTSIC模型中三个激活过程的激活结果统计来证明MTSIC模型的有效性；在MTSIC模型下，通过对比ANS算法与KKT算法各自激活的节点的重合度来证明ANS算法的有效性；通过比较单机程序与PMT算法的加速比来证明PMT算法的有效性。实验结果充分证明了本文所提出的模型和算法具有优良的效果及效率。

# [6 总结与展望](#_Toc283328293)

## 6.1 论文总结

影响力最大化问题一直以来都是数据挖掘领域中的热门的研究方向，对于影响力最大化问题的相关研究也是从出不穷。然而随着像FaceBook、Twitter等社会网络的快速壮大，已有的影响力最大化的模型与算法已经无法满足当前的需求，更不能达到预期的效果。然而，随着社会网络的盛行，利用“病毒式营销”、“口碑传播”等方式推广商品或事件就有着巨大的商业前景，这就使得影响力最大化问题面对着前所未有的挑战。

通过对已有工作的研究发现，之所以已有研究并不太适用于当前的社会网络的原因主要有两点：第一，绝大多数的已有研究忽略了用户话题的信息，即使有少部分研究考虑了话题元素，但却忽略了任何的商品或事件都不是话题单一的且用户的兴趣也具有多样性；第二，目前的社会网络都是大规模网络，已有研究在处理大规模网络时的效率问题尤为突出。上述出现的问题也就是本文所需解决的问题。因此，本文的主要工作大致如下：

1. 对于影响力最大化问题的国内外现状做深入的了解，根据已有研究分析它们的不足与缺陷，发现已有研究都忽略了任何商品或事件都包含多个话题，因此没有考虑多话题因素。针对这种情况，展开多话题敏感的影响力最大化问题的研究。

2. 基于对于现实世界的5个观察，提出MTSIC模型。该模型融入了多话题因素，同时考虑用户节点的权威度与中心度，使得结果更加符合真实情况。考虑到现实应用往往存在地理位置的限制，将地理位置因素加入其中，使得结果更加准确，更具有针对性。

3. 由于已有影响力最大化算法无法适用于基于多话题的MTSIC模型，因此将多话题因素与经典的贪心算法相结合，提出了ANSM度量，并根据该度量给出了ANS算法。

4. 考虑到当前的社会网络皆为大规模网络，已有的算法在大规模网络下的运行效率较差，而基于内存计算的SPARK计算框架适用于图迭代的特性，因此给出了基于SPARK的并行算法，以解决影响力最大化问题的效率问题。

5. 在DBLP和Twitter真实数据集上进行不同传播模型与影响力最大化算法的对比实验，使用最后激活节点数与本文所提出的新的度量分别对不同模型与算法进行分析，给出新模型MTSIC与算法ANS的性能优势。通过分析在分布式平台的与单机运行所花费的时间，给出所提出的并行算法PMT的效率优势。

## 6.2 论文展望

鉴于本人学术水平和时间所限，论文中还存在着一些需要进一步完善和探讨的地方，以下是几点未来可以着手改进的地方：

1. 尽管本文所提出的模型将多话题、地理位置、权威度和中心度等因素融入其中，但该模型依然是静态的。事实上，随着时间的推移，用户节点的兴趣和网络结构也在随之改变，因此考虑动态因素就成为一个需要研究的方向。

2. 本文是利用Topical HITS算法来计算用户节点的权威度和中心度，但已有研究有很多其他可选方法来进行模型的参数化学习。以后可以研究其他更有效地方式来估计这些参数值，使得模型更准确且更贴近现实。

3. 由于本文仅仅将多话题因素融入到经典的贪心算法中，而并未考虑其他高效的先进的算法，也并没有利用启发式、剪枝等策略，因此算法的效率依然很低，仅仅单纯满足了模型的需求。未来可以研究更加高效的算法，来提升运行效率。

4. 本文所提出的模型和算法只是作为研究，并未上升到实际应用。因此可以设计一个平台，去响应商品或事件的影响力最大化的请求，并记录结果。对于相似商品或事件的请求，可以将已经计算得到的之前相似的商品或事件的结果返回，避免重复计算，节约运行成本。

最后，由于本人能力有限，文中未免会出现一些错误或纰漏，本人愿意接受老师和同学们的悉心指导和批评指正。

# 致 谢

三年的时间转瞬即逝，三年前刚刚踏入校园的我转眼间已经将要离开。美好的时光总是短暂的，三年间我不仅收获了知识、能力，也收获了亲情、友情和美好的回忆。在智能与分布式计算实验室学习生活的这段时间，我充分感受到了实验室老师们对于学术的严谨的精神，也体会到了老师们对学生们的关爱。在学习上，是老师们的谆谆教导使我在知识的海洋能畅快遨游；在生活上，是同学朋友们的热心帮助使我克服各种各样的困难。因此，借此机会，希望向在我研究生生涯中指导过我，帮助过我的老师和同学们表达我衷心的感谢。

我在华中科技大学的学生生涯能够顺利的结束，最应该感谢的是我的导师李玉华老师。李老师在学术上给了我正确的指导；在研究方向上给了我良好的建议；在研究过程中给了我宝贵的意见；在日常生活上给了我很大的帮助。在论文的写作过程中，李老师帮我规划进度并监督，认真仔细的帮我进行整理和修改，终使我能够按时地，保质保量的完成论文的撰写。在这里，由衷的感谢李玉华老师为我所做的一切。

感谢实验室主任李瑞轩老师对全体学生严格要求，李瑞轩老师对学术严谨的态度和对以身作则的精神是我们学习的榜样。同时也感谢文坤梅老师、辜希武老师和章衡老师，是实验室所有老师的共同努力才使得我们拥有了良好的学习环境。

感谢我的师姐王旭和师兄牛进保在学术上给我的帮助和建议；感谢我的同学王葵、吴浩、吴文哲、潘杰、周倩婷、付永刚在学习和生活中给予的帮助；感谢师弟张军和袁清亮给予的各方面帮助。

特别感谢父母和亲戚朋友对我的支持，没有你们的支持和帮助，我不可能取得今天的成就，你们的关心和鼓励是我最大的动力。希望我可以不负众望，做一个对社会对国家有贡献的高学历人才。

最后，由衷地感谢各位答辩评委们对我提出的宝贵意见，真心感谢老师们对我论文的悉心指正。

参考文献

[1] Brown J J, Reingen P H. Social ties and word-of-mouth referral behavior. Journal of Consumer research, 1987, 14(3): 350~362

[2] Goldenberg J, Libai B, Muller E. Using complex systems analysis to advance marketing theory development: Modeling heterogeneity effects on new product growth through stochastic cellular automata. Academy of Marketing Science Review, 2001, 1(3): 10~14

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

[4] Hofmann T. Probabilistic Latent Semantic Analysis. In: Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence, 1999. 289~296

[5] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation. Journal of Machine Learning Research, 2003, 3(2): 993~1022

[6] Wang C, Blei D M, Heckerman D. Continuous Time Dynamic Topic Models. In: Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence, 2008: 579~586

[7] Steyvers M, Griffiths T. Probabilistic topic models. Handbook of latent semantic analysis, 2007, 427(7): 424~440

[8] Deng H, Han J, Zhao B, et al. Probabilistic topic models with biased propagation on heterogeneous information networks. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2011. 1271~1279

[9] Blei D M, Mcauliffe J D. Supervised Topic Models. ArXiv Preprint, 2010, 4(2): 7~16

[10] Blei D M, Lafferty J D. Correlated Topic Models. Advances in Neural Information Processing Systems, 2005, 18(2): 147~156

[11] Blei D M, Lafferty J D. Dynamic topic models. In: Proceedings of the 23th International Conference on Machine Learning, 2006. 113~120

[12] Mccallum A, Wang X, Emmanuel A E S C. Topic and Role Discovery in Social Networks with Experiments on Enron and Academic Email. Journal of artificial intelligence research, 2007, 30(1): 249~272

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

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

[15] Leskovec J, Krause A, Guestrin C, et al. Cost-effective outbreak detection in networks. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining 2007. 420~429

[16] Goyal A, Lu W, Lakshmanan L V. Celf++: optimizing the greedy algorithm for influence maximization in social networks. In: Proceedings of the 20th international conference companion on World Wide Web, 2011. 47~48

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

[18] Liu X, Li M, Li S, et al. IMGPU: GPU-Accelerated Influence Maximization in Large-Scale Social Networks. IEEE Transactions on Parallel and Distributed Systems, 2014, 25(1): 136~145

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

[20] Kimura M, Saito K. Tractable Models for Information Diffusion in Social Networks. In: Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases, 2006. 259~271

[21] Chen W, Yuan Y, Zhang L. Scalable Influence Maximization in Social Networks under the Linear Threshold Model. In: Proceedings of the 10th International Conference on Data Mining, 2010. 88~97

[22] Goyal A, Lu W, Lakshmanan L. SIMPATH: An Efficient Algorithm for Influence Maximization under the Linear Threshold Model. In: Proceedings of the 11th International Conference on Data Mining, 2011. 211~220

[23] Kimura M, Saito K, Nakano R. Extracting Influential Nodes for Information Diffusion on a Social Network. In: Proceedings of the 22th National Conference On Artificial Intelligence, 2007. 1371~1376

[24] Kim J, Kim S K, Yu H. Scalable and parallelizable processing of influence maximization for large-scale social networks?. In: Proceedings of the 29th International Conference on Data Engineering, 2013. 266~277

[25] 冯小军. 社会网络环境下一种基于潜力的影响最大化算法. [硕士学位论文]. 上海: 复旦大学, 2010.

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

[27] Yan Q, Guo S, Yang D. Influence Maximizing and Local Influenced Community Detection Based on Multiple Spread Model. Advanced Data Mining and Applications, 2011, 3(2): 82~95

[28] Galstyan A, Musoyan V L, Cohen P R. Maximizing Influence Propagation in Networks with Community Structure. Physical Review, 2009, 79(5): 56~102

[29] 兰如钦. 社会网络上的影响力最大化算法研究. [硕士学位论文]. 北京: 北京交通大学, 2011.

[30] 黎雷. 社会网络影响力模型及其算法研究. [硕士学位论文]. 北京: 北京交通大学, 2010.

[31] Dar E E, Shapira A. A Note on Maximizing the Spread of Influence in Social Networks. Internet and Network Economics, 2007, 2(5):281~286

[32] Vez P A E E, Vera P A, Saito K. Selecting the Most Influential Nodes in Social Networks. In: Proceedings of the 2007 International Joint Conference on Neural Networks, 2007. 2397~2402

[33] Bonchi F. Influence Propagation in Social Networks: A Data Mining Perspective. Web Intelligence. 2011, 12(1): 8~16

[34] Dhanjal C, Blanchemanche S, Clemen S, et al. Information Diffusion within Social Networks. 2011, 5(3): 481~486

[35] Wang D, Wen Z, Tong H, et al. Information spreading in context. In: Proceedings of the 20th International Conference on World Wide Web, 2011. 735~744

[36] Li H, Bhowmick S S, Sun A. Casino: towards conformity-aware social influence analysis in online social networks. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, 2011. 1007~1012

[37] He X, Song G, Chen W, et al. Influence Blocking Maximization in Social Networks under the Competitive Linear Threshold Model Technical Report. Arxiv Preprint, 2011, 1(5): 81~86

[38] Chen W, Collins A, Cummings R, et al. Influence Maximization in Social Networks When Negative Opinions May Emerge and Propagate. In: Proceedings of the 11th SIAM International Conference on Data Mining, 2011. 181~186

[39] Budak C, Agrawal D, Abbadi A E. Limiting the spread of misinformation in social networks. In: Proceedings of the 20th international conference on World Wide Web, 2011. 665~674

[40] Huang J, Cheng X, Shen H, et al. Exploring social influence via posterior effect of word-of-mouth recommendations. In: Proceedings of the 15th ACM international conference on Web search and data mining, 2012. 573~582

[41] Anagnostopoulos A, Kumar R, Mahdian M. Influence and correlation in social networks. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, 2008. 7~15

[42] Singer Y. How to win friends and influence people, truthfully: influence maximization mechanisms for social networks. In: Proceedings of the 5th ACM International Conference on Web Search and Data Mining, 2012. 733~742

[43] Nguyen H, Zheng R. On Budgeted Influence Maximization in Social Networks. IEEE Journal on Selected Areas in Communications, 2013, 31(6): 1084~1094

[44] Guo J, Zhang P, Zhou C, et al. Personalized influence maximization on social networks. In: Proceedings of the 22nd ACM international conference on Conference on information & knowledge management, 2013. 199~208

[45] Cheng S, Shen H, Huang J, et al. StaticGreedy: solving the scalability-accuracy dilemma in influence maximization. In: Proceedings of the 22nd ACM international conference on Conference on information & knowledge management, 2013. 509~518

[46] Li H, Bhowmick S S, Sun A. CINEMA: conformity-aware greedy algorithm for influence maximization in online social networks. In: Proceedings of the 16th International Conference on Extending Database Technology, 2013. 323~334

[47] Hu J, Meng K, Chen X, et al. Analysis of influence maximization in large-scale social networks. ACM SIGMETRICS Performance Evaluation Review. 2014, 41(4): 78~81

[48] Cohen E, Delling D, Pajor T, et al. Sketch-based influence maximization and computation: Scaling up with guarantees. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, 2014. 629~638

[49] Li G, Chen S, Feng J, et al. Efficient location-aware influence maximization. In: Proceedings of the 2014 ACM SIGMOD international conference on Management of data, 2014. 87~98

[50] Zhou T, Cao J, Liu B, et al. Location-Based Influence Maximization in Social Networks. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, 2015. 1211~1220

[51] Liu L, Tang J, Han J, et al. Mining topic-level influence in heterogeneous networks. In: Proceedings of the 19th ACM Conference on Information and Knowledge Management, 2010. 199~208

[52] 贺人贵. 基于话题的学术网络影响力最大化研究. [硕士学位论文]. 武汉: 华中科技大学, 2012.

[53] Zhang Y, Zhou J, Cheng J. Preference-Based Top-K Influential Nodes Mining in Social Networks. In: Proceedings of the 10th IEEE International Conference on Trust, Security and Privacy in Computing and Communication, 2011. 1512~1518

[54] 王旭. 富文本社会网络中话题敏感的影响力最大化问题研究. [硕士学位论文]. 武汉: 华中科技大学, 2014.

[55] Barbieri N, Bonchi F, Manco G. Topic-Aware Social Influence Propagation Models. Knowledge and information systems, 2013, 37(3): 555~584

[56] Zhang C, Sun J, Wang K. Information propagation in microblog networks. In: Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining, 2013. 190~196

[57] Li Y, Zhang D, Tan K L. Real-time Targeted Influence Maximization for Online Advertisements. Proceedings of the VLDB Endowment, 2015, 8(10): 1070~1081

[58] Chen S, Fan J, Li G, et al. Online Topic-Aware Influence Maximization. Proceedings of the VLDB Endowment, 2015, 8(6): 666~677

[59] Kleinberg J M. Authoritative sources in a hyperlinked environment. Journal of the ACM, 1999, 46(5): 604~632

[60] Nie L, Davison B D, Qi X. Topical link analysis for web search. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, 2006. 91~98

[61] Zaharia M A. An Architecture for and Fast and General Data Processing on Large Clusters. Dissertations & Theses, 2013

# 附录1 攻读硕士学位期间投稿的论文

[1] Li Y, Liu C, Zhao M, et al. Multi-Topic Tracking Model for dynamic social network. Physica A: Statistical Mechanics and its Applications, 2016, 454(1): 51-65.

[2] Kai Wang, Yuhua Li, Ruixuan Li, Xu Wang. Multi-Topic Influence Maximization in Large Scale Networks. Proceedings of the 22th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.（已投稿）

1. http://arnetminer.org/ [↑](#footnote-ref-1)
2. https://twitter.com/ [↑](#footnote-ref-2)