分类号 TP957 学号 0123456

UDC 密级 公开

工学博士学位论文

国防科大学位论文 LaTeX 模板使用手册

博士生姓名 张三

学科专业 通信与信息工程

研究方向 自动目标识别与模糊工程

指导教师 李四 教授

王五 副教授

国防科学技术大学研究生院

二〇一六 年 十一 月

How to Use the LaTeX Document Class for NUDT Dissertations

Candidate: Zhang San

Supervisor: Professor Li Si

A dissertation

Submitted in partial fulfillment of the requirements

for the degree of Doctor of Engineering

in Information and Communication Engineering

Graduate School of National University of Defense Technology

Changsha, Hunan, P. R. China

November 10, 2016

独 创 性 声 明

本人声明所呈交的学位论文是我本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，除文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表和撰写过的研究成果，也不包含为获得国防科学技术大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文题目：国防科学技术大学学位论文 LaTeX 模板

学位论文作者签名：_____ 日期：　　年　　月　　日

学位论文版权使用授权书

本人完全了解国防科学技术大学有关保留、使用学位论文的规定。本人授权国防科学技术大学可以保留并向国家有关部门或机构送交论文的复印件和电子文档，允许论文被查阅和借阅；可以将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、汇编学位论文。

（保密学位论文在解密后适用本授权书。）

学位论文题目：国防科学技术大学学位论文 LaTeX 模板

学位论文作者签名：_____ 日期：　　年　　月　　日

作者指导教师签名：_____ 日期：　　年　　月　　日

目　录

表 目 录

图 目 录

摘　要

国防科学技术大学是一所直属中央军委的综合性大学。1984 年, 学校经国务院、中央军委和教育部批准首批成立研究生院，肩负着为全军培养高级科学和工程技术人才与指挥人才，培训高级领导干部，从事先进武器装备和国防关键技术研究的重要任务。国防科技大学是全国重点大学，也是全国首批进入国家“211 工程”建设并获中央专项经费支持的全国重点院校之一。学校前身是 1953 年创建于哈尔滨的中国人民解放军军事工程学院，简称“哈军工”。

关键词: **国防科学技术大学**; 211; **哈军工**

ABSTRACT

National University of Defense Technology is a comprehensive national key university based in Changsha, Hunan Province, China. It is under the dual supervision of the Ministry of National Defense and the Ministry of Education, designated for Project 211 and Project 985, the two national plans for facilitating the development of Chinese higher education.

NUDT was originally founded in 1953 as the Military Academy of Engineering in Harbin of Heilongjiang Province. In 1970 the Academy of Engineering moved southwards to Changsha and was renamed Changsha Institute of Technology.The Institute changed its name to National University of Defense Technology in 1978.

Key Words: NUDT; MND; ME

符号使用说明

HPC	高性能计算 (High Performance Computing)
cluster	集群
Itanium	安腾
SMP	对称多处理
API	应用程序编程接口
PI	聚酰亚胺
MPI	聚酰亚胺模型化合物，N-苯基邻苯酰亚胺
PBI	聚苯并咪唑
MPBI	聚苯并咪唑模型化合物，N-苯基苯并咪唑
PY	聚吡咙
PMDA-BDA	均苯四酸二酐与联苯四胺合成的聚吡咙薄膜
ΔG	活化自由能 (Activation Free Energy)
χ	传输系数 (Transmission Coefficient)
E	能量
m	质量
c	光速
P	概率
T	时间
v	速度

第一章　第一章题目

本章的主要内容与学校提供的 Word 模板中内容一致，图片与表格均采用原始设定大小，主要是为了说明格式的统一。但是，LaTeX 的一些禁则，专业排版的能力，对公式及文献的处理都是得天独厚的，我们不必刻意去追求与 Word 的完美匹配。而且你将会发现，用 LaTeX 书写论文的美！

1.1　（1.1 题目）

正文内容

1.1.1　（1.1.1 题目）

正文内容
正文内容

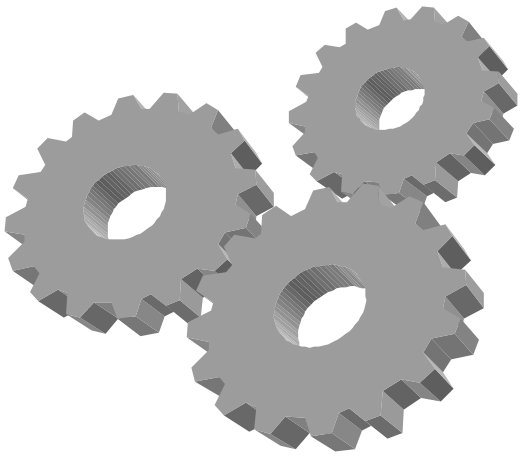

图 1.1　图 1.1 名称

1.1.1.1　（1.1.1.1 题目）

正文内容
正文内容
正文内容

1.1.1.2　（1.1.1.2 题目）

正文内容
正文内容
正文内容

1.1.2　（1.1.2 题目）

正文内容

正文内容

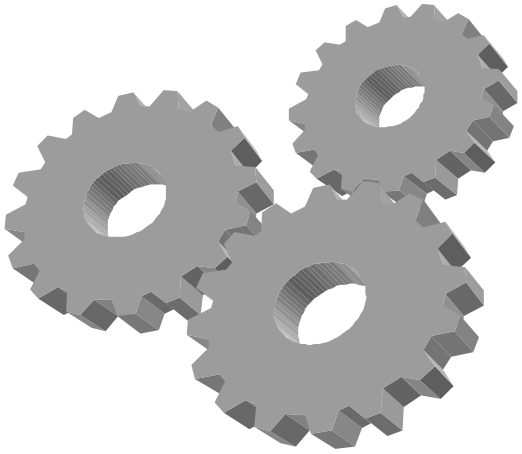

图 1.2　图 1.2 名称

1.2　（1.2 题目）

正文内容
正文内容

表 1.1　表 1.2 名称

正文内容
正文内容
正文内容
正文内容

1.3　（1.3 题目）

正文内容
正文内容
正文内容
正文内容
正文内容
正文内容

1.3.1 （1.3.1 题目）

正文内容

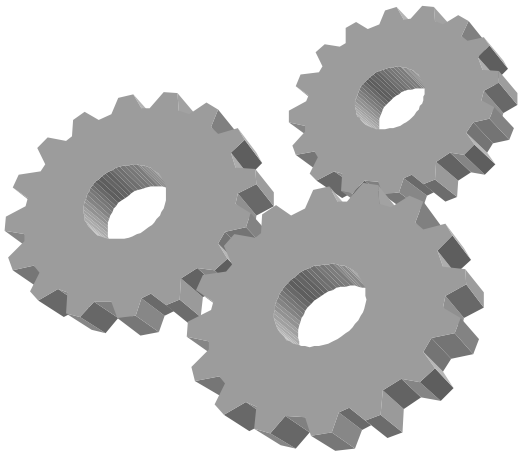

图 1.3　图 1.3 名称

1.3.2 （1.3.2 题目）

正文内容
正文内容

表 1.2　表 1.2 名称

第二章　基于语义扩展和文本质量的实时个性化搜索

随着社交网络中信息爆炸式的增长，用户越来越难以从海量的信息中获取到自身所需的信息，用户所需的信息往往淹没在其中，这种现象可称之为**信息洪流**（*information flood*）现象。在社交网络、电子商务网站、即时通信等应用中，信息产生的速率以及数量都是过去无法比拟的，如在推特（Twitter）、脸书（Facebook）、新浪微博等社交媒体中，每天都会产生海量的文本信息。根据用户输入的查询，在海量的信息流中实时地检索出高质量的、相关的信息是一个极具挑战性的问题。社交网络中的信息流实时个性化搜索相对于传统的信息检索提出了以下挑战：（1）社交网络中充斥着各种各样话题的信息，而且信息大多数都是以短文本表示，相对于传统的新闻等长文本信息，难以进行语义理解；（2）社交网络中的信息质量参差不齐，难以从中遴选出高质量的信息；（3）社交网络中的信息产生速率快，如何能够实时地检索出用户所需要的信息，将其推送给用户也是一大难点。因此，由于社交网络中数据的信息海量性、主题多样性、数据稀疏性以及社交互动性等特性，传统的个性化检索方法不足以解决社交网络中的信息实时个性化搜索问题。本章针对以上的问题，提出了一个面向推特信息流的实时个性化搜索框架来实现用户的信息实时推荐。本章针对社交网络中信息的特性，提出了一种基于语义扩展和文本质量的实时个性化搜索算法，并在 TREC 2015 Microblog Track[1] 测评中验证了算法的性能。首先，我们构造了一个逻辑规则过滤器来选择核心关键词，提高检索的准确率。其次，我们对文本质量进行建模，利用的标注的数据进行训练，以此来对文本的质量的进行打分。训练好的文本质量模型提高了检索的排序性能。然后，我们使用外部语料库来实现语义扩展，例如搜索引擎，知识库等。语义扩展能够使得我们更好地理解用户的偏好和兴趣。最后，我们采用了一个动态的推送策略来自动地推送高质量且相关的信息给特定的用户，这能够避免信息过载。本章中的算法结合了社交网络文本的语义特征和社交属性，针对不同的用户搜索，做了综合性地排序。我们使用 TREC 2015 Microblog Track 中的真实数据流进行了实验，实验结果显示了本章提出的算法在不同测评指标下，与其他算法相比的优越性。

本章的内容组织如下：第2.1 节介绍了研究动机，讨论了在社交网络环境下进行实时个性化搜索的必要性。第2.2节介绍了相关定义，对本章中涉及的相关概念和知识进行了符号化的定义。第2.3 节介绍了方法描述，详细地阐述了本章提出的系统框架和算法。第2.4节进行了实验分析，验证了本章提出的方法，并且分析了实验结果。最后，第2.5对本章的内容进行了总结。

2.1 研究动机

随着大数据时代的到来，诸如推特（Twitter）、脸书（Facebook）、新浪微博等的社交网络平台逐渐取代传统的媒体平台，成为新时代的实时信息交互平台。以推特平外为例，据统计，每天平均约有 58,000,000 条推文（tweet）发布，每天平均处理约 2,100,000,000 次搜索查询，每月约有 115,000,000 个活跃的用户[1]。在大数据时代，人人都可以是信息的生产者、传播者和接收者，这在一定程度上加速了信息的传播速率。然而，如此庞大的数据量使得用户在社交网络平台中搜索查询时面临了信息过载的问题，用户很难检索到自己需要的信息，亦或用户所需的信息淹没在了众多的信息之中。尤其是社交网络平台中的信息内容囊括了众多领域，其中的话题种类繁多，这使得用户难以搜索到相关性强而且高质量的文本信息。在社交网络平台中，传统的信息检索方法变得耗时长而且信息检索方式难以适用。因此，在大数据时代，为了满足用户实时获取相关信息的需求，需要一种面向社交网络的新的信息检索方法。

在传统的信息检索流程中，往往是用户根据自己所需，输入关键字进行查询，系统根据用户的查询，搜索到相关的结果，并进行排序，返回给用户。而在社交网络平台中，信息产生速率快，信息以数据流的形式给出，同时用户希望系统能够自动地推送相关的信息，而不是用户通过查询来获取信息。因此，在社交网络中的信息实时个性化搜索的流程与传统的信息检索流程不尽相同。在社交网络中，信息的实时个性化搜索流程一般是用户将查询搜索以关键词的形式给出，然后系统根据用户的查询在信息流中实时地处理文本信息，将相关的信息自动地推送给用户。

由美国国家标准与技术研究院（NIST）主办的文本检索会议（TREC）是由多个测评项目组成的一个致力于解决新时代信息检索问题的测评大会，涉及智能问答、医疗诊断、实体识别、信息推荐等领域。TREC 自 2011 年起设立了微博实时推荐（Microblog Track）[1–5] 这一个子任务，目标是解决在社交网络中信息的实时个性化搜索问题。从设立后，TREC Microblog Track 便吸引了全世界的参赛者参加，与智能实时问答（TREC LiveQA Track）成为了 TREC 中最火热的两个子任务。Microblog Track 的任务是针对不同用户，智能地分析用户的兴趣爱好，自动地、实时地为用户推送相关的、高质量的信息，涉及到众多科学领域，包括机器学习、自然语言处理、信息检索、人工智能等。如图2.1所示，在社交网络平台中，用户是信息的生产者，用户将源源不断地发布信息，其中包括有趣的信息和一些无用的信息。同时，用户又是信息的接收者，用户通过定制自身感兴趣的话题和

[1] http://www.statisticbrain.com/twitter-statistics/

内容，系统智能地分析其兴趣爱好，针对不同的用户，在推特信息流中自动地、实时地为用户推荐相关的、高质量的信息。例如将饮食信息推荐给美食家、将财经信息推送给金融家、将商旅信息推送给旅行者等，同时将一些低质量无意义或者无人关注的信息丢弃。

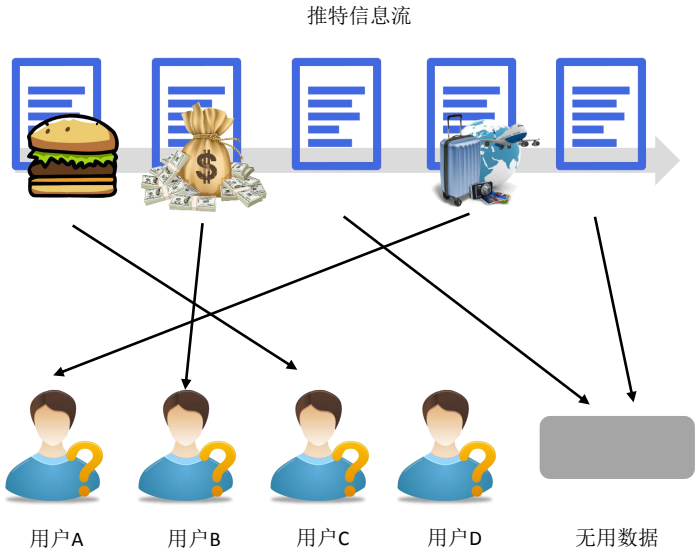

图 2.1　社交网络平台中信息的实时个性化搜索示意图

为了解决针对不同用户，为其实时地搜索出相关性强、高质量的信息，许多已有的个性化推荐[6–9] 以及协同搜索[10–13] 的工作对这个问题进行了一定的研究。但是很少有面向社交网络个性化搜索的研究，特别是对于实时搜索推荐技术的研究。目前已有的机器学习、自然语言处理、信息检索、人工智能等研究对于社交网络中的实时个性化搜索问题提供了许多帮助。文本分类以及排序的研究[14–17] 能够为了社交网络中的信息检索提供支持。同时，查询分析以及优化技术[18–20] 能够使得信息检索有着显著的提升。

然而，社交网络的环境与传统的新闻、论坛等 Web 环境有着较大的区别。因此，为了满足社交网络中用户获取信息的需求，这将需要建立新的模型、研究新的方法来解决这一问题。在社交网络中，实现信息的个性化实时推荐面临的主要挑战可以总结如下：

- **信息海量性**，在社交网络中，信息产生的速率快，网络中的每一个用户同时扮演着信息生产者、信息传播者和信息接收者的角色。如此高容量的信息流需要一个新的模型来适应持续不断变化的语义特征。

- **主题多样性**，社交网络中的信息内容包罗万象，覆盖了许许多多的领域和话题。如果主题模型不能区分众多的话题，这将导致噪声的引入以及不准确的话题模型以及用户模型。

- **数据稀疏性**，在社交网络中，信息在不同的主题上的分布是不均匀的，在某些主题上信息量大，而在某些主题上信息是稀疏的。有效的主题模型需要解决数据稀疏性所带来了的影响。

- **社交互动性**，社交网络中的用户之间有着丰富的互动信息，与传统的文本信息不同，社交网络中的信息包含了许多有价值的结构化的社交属性。适当地利用这些社交属性能够提高搜索的性能，但是这需要对这些社交属性进行相关性的选择。

为了解决上述面临的挑战，本章提出了一种基于语义扩展和文本质量的实时个性化搜索框架，该框架综合考虑了用户的偏好、语义特征和社交属性。首先，本章基于语义扩展提出了一种布尔逻辑关键词过滤（Boolean Logic Keyword Filter）的用户模型。该模型依靠外部搜索引擎提供的知识进行建立，建立的用户模型充分利用了查询扩展以及检索结果的重排序来提高推荐结果的相关性。最终的实验评估证明该模型显著地提高了检索的召回率。此外，本章还基于逻辑回归提出了一种文本质量模型，该模型利用推文的社交属性来评估其文本的质量。该模型能够对推文的文本内容，是否受到大众的认可等进行评估，因此它能帮助系统返回高质量的推文。最终的实验评估证明该模型显著地提高了检索结果的排序性能。

2.2 相关定义

本节将对社交网络中的实时个性化搜索的任务进行定义，并将其形式化。

2.3 方法描述

2.4 实验分析

2.5 本章小结

第三章　社交网络中传播效率最大化

影响力最大化（*Influence Maximization*）在许多的社交网络应用中扮演着举足轻重的角色，例如市场营销、商业活动以及竞选活动等等。研究信息传播模型以及机理的一个典型目的便是市场营销。在现实社会中，人或者事物都是通过某种关系来连接，形成了一个巨大的网络。因此，信息传播可以选择一小部分节点激活做为种子节点集合，然后通过信息的传递，在整个网络中产生一个大范围的影响。从技术方面来说，影响力最大化问题是在给定网络和传播模型的条件下，研究如何选择种子节点集合使得传播的影响范围最大化。影响力最大化问题由于其的问题真实、应用性广的特点，吸引了许许多多的研究，包括信息传播模型的研究和计算影响力最大化的方法。但是，仍然有着一些需求在实际应用中得不到满足。

在信息传播过程中，一个被激活的节点将会尝试在下一个时刻激活它的邻居节点。所以，除去种子节点外，每一个被激活的节点在被激活之前都会有一个时间延迟，我们称之为**传播时延**（*propagation time delay*）。如果一个节点在信息传播结束时，仍然未被激活，则该节点的传播时延可以看作无穷大。在传统的影响力最大化问题中，传播时延没有被考虑在内，而研究整个网络的传播时延是非常有意义的，它可以度量选择的种子节点集合的传播效率。出于这种需求，本章提出了一个新的问题，传播效率最大化问题。该问题将传播时延考虑在内，在给定网络和传播模型的条件下，研究如何选择种子节点集合使得传播效率最大化。传播效率最大化问题与影响力最大化问题虽然相似，但是两个问题的侧重点不尽相同。传统的影响力最大化问题研究的是如何使得传播的范围最广，不考虑节点的传播时延。而本章提出的传播效率最大化问题将传播时延考虑在内，研究如何使得传播效率最大化。

本章主要的工作可以总结如下。首先，基于传统的影响力最大化问题，我们将传播时延考虑在内了来探究传播效率，提出了传播效率最大化问题，研究给定网络和传播模型的条件下，研究如何选择种子节点集合使得传播效率最大化。其次，本章证明了传播效率最大化问题是一个 NP-hard 问题，而且在独立级联模型下计算传播效率的过程是一个 #P-hard 的问题。然后，我们证明了传播效率函数在独立级联模型下是**子模**（*submodular*）的。最后本章设计了三个算法来解决提出的传播效率最大化问题，并且在真实数据集上验证了这些算法。实验结果展示了所提出的算法的性能。

本章的内容组织如下：第3.1节介绍了研究动机，讨论了传统的影响力最大化问题的不足和考虑了传播时延的传播效率最大化问题的意义。第3.2节介绍了相关定义，对本章中相关概念和所提出的问题进行了符号化的定义。第3.3介绍了方法

描述，详细地阐述了本章解决问题的方法以及相关证明过程。第3.4节进行了实验分析，设计了一系列的实验，验证了本章所提出的方法，并对实验结果进行了分析。最后，第3.5对本章的内容进行了总结。

3.1 研究动机

随着社交网络的兴盛发展，信息传播的速率变得越来越快。在传统的媒体环境下，一条消息需要通过较长时间才能传播到一定的范围。在社交网络中，个人通过关系（例如关注、好友等）连接形成网络，信息在网络中通过转发、复制等行为进行传播，信息传播的速率极大的加快了。在传统的媒体环境下，个人都是通过单一信息源（例如新闻、论坛等）来接收信息。而在社交网络环境下，个人既可以是信息的接收者也可以是信息的发布者，个人可以接收到在网络中与之相连的个人推送的信息。与此同时，信息通过个人构成的网络进行传播。影响力最大化问题是信息传播中的一个典型问题，是在给定网络和传播模型的条件下，研究如何选择种子节点集合使得传播的影响范围最大化。市场营销是研究信息传播模型以及机理的一个典型驱动。市场营销的过程可以简述如下，一个公司希望在社交网络中通过“口碑效应”推动一款新产品或者一种新理念。一种有成本效益的方法是寻找到整个网络中有影响力的个人，然后投入资源来使他们接受这款产品，例如赠送样品、免费试用、优惠折扣等。这样的举措是希望这些有影响力的个人在接受新产品或者新理念后，能够驱使社交网络中的其他个人也来接受新产品或者新理念，然后在社交网络中产生一个大的级联效应，从而使得更多的个体接受新产品或者新理念。为了达到市场营销这一目的，我们需要对两个重要的问题进行研究：（1）如何对网络中的信息传播过程进行建模，包括模型参数的学习等；（2）如何在给定的传播模型的条件下，设计一个有效的方法来寻找能够最大化影响力的节点集合。本章着重对第二个问题进行研究。

影响力最大化问题作为市场营销的一种算法技术首先被 Domingos 和 Richardson[21] 所提出，该问题基于马尔科夫随机场的概率框架。而后，Kempe 等人[22] 首次将影响力最大化问题形式化成为一个离散的随机优化问题。信息传播的过程可以简述如下，在一个给定的网络中，选择部分节点做为种子集合，这些种子节点将会按照一定的规则去激活它们的邻居节点。被激活的节点在下一时刻将拥有能力去激活它们的邻居节点，这个过程将一直持续到没有新的节点可以被激活，整个信息传播的过程才会停止。影响力最大化问题是在给定网络和传播模型的条件下，研究如何选择种子节点集合使得传播的影响范围最大化，即激活的节点数目最大化。从上述信息传播过程的描述中，我们可以得知，一个被激活的节点会在被激活的下一时刻尝试激活它的邻居节点。因此，除去种子节点外，网络中的节

点在被激活前都会有一个时间延迟，我们称之为传播时延。如果一个节点在信息传播结束时，仍然未被激活，则该节点的传播时延可以看作无穷大。而影响力最大化问题仅仅考虑了传播范围，忽略了节点的传播时延。在真实的场景中，例如市场营销、商业活动、竞选活动等，传播时延在信息传播中是一个非常重要的因素。为了让其他人接受自己的新产品或者新理念，人们总是希望能够尽快地将消息传播到群体中。我们以**传播效率**（*influence efficiency*）来表示传播时延的倒数，传播时延小，则传播效率大。如果节点在信息传播结束时仍然未被激活，那么该节点的传播效率则为 0。以上是针对单个节点的传播时延和传播效率的分析，下面我们对整个网络的传播时延和传播效率进行分析。在给定一个传播网络和初始的种子节点集合的情况下，如果整个网络中所有节点的传播效率高，这就意味着在信息传播过程中，网络中的节点将被迅速地激活。我们设想如下，针对传统的影响力最大化问题有两种选择种子节点集合的策略，它们有着相同的传播范围，即能够在信息传播过程结束时能够激活相同的节点数目。但是，这两种策略中，网络中的传播时延可能是不同的，即不同策略在同一网络中的传播效率是不同的，而这一问题在传统的影响力最大化问题中是没有讨论的。

3.2 相关定义

在本节中，第3.2.1节首先对**独立级联模型**（*Independent Cascade Model*）以及在独立级联模型下的影响力最大化问题进行回顾，然后介绍了几种解决影响力最大化问题的方法，并且对这些算法进行分析，包括影响力函数期望的单调性和子模性等。其次，第3.2.2节提出了**传播效率最大化**（*Influence Efficiency Maximization*）问题，该问题是基于传统的影响力最大化问题，将传播时延考虑在内，研究如何使得整个网络的传播效率最大化。同时，第3.2.2节对传播效率最大化问题进行了形式化的描述，分析了传播效率最大化和影响力最大化问题的区别。

为了便于参照，表3.1中列出了频繁使用的符号。

3.2.1 传播模型以及影响力最大化问题

本章采用一种广泛采用的信息传播模型，独立级联模型，进行传播影响的研究。在该模型下，一个社交网络可被建模表示为一个有向图 $\mathcal{G} = (\mathcal{V}, \mathcal{E})$，其中 $\mathcal{V}$ 代表网络中的个体，$\mathcal{E}$ 表示个体之间的社会关系。此外，图中的每一条边 $(u, v) \in \mathcal{E}$ 上都关联着一个传播概率 $p^{\mathcal{G}}_{u,v}$，表示着节点 u 到节点 v 的影响力度。如果传播概率 $p^{\mathcal{G}}_{u,v}$ 越大，则节点 u 更加可能激活节点 v。如果图 $\mathcal{G}$ 与上下文无关，本章则用 $p_{u,v}$ 来表示节点 u 到节点 v 的传播概率。

独立级联模型描述了一个直观的信息传播过程，其过程如下。在独立级联模型下，网络中的个体会被其邻居所影响，这些影响之间是独立的。给定一个种子

表 3.1　常用符号列表

符号	描述
$\mathcal{G}=(\mathcal{V},\mathcal{E})$	$\mathcal{G}$ 是社交网络构成的图，$\mathcal{V}$ 是节点集合，$\mathcal{E}$ 是边的集合
$\mathcal{H}=(\mathcal{V},\mathcal{Z})$	$\mathcal{H}$ 是基于图 $\mathcal{G}$ 生成的超图（参见**??**），$\mathcal{V}$ 是节点集合，$\mathcal{Z}$ 是超边集合
n	$\mathcal{G}$ 或者 $\mathcal{H}$ 的节点的数目
m	$\mathcal{G}$ 的边的数目
k	种子节点集合的大小
$p^{\mathcal{G}}_{u,v}$	节点 u 激活节点 v 的概率
$I(S)$	种子节点集合 S 的影响力
$RR(v)$	节点 v 的反向连通节点集合 (see Definition 3.1)
$e_{u,v}$	节点 u 到节点 v 的传播效率（参见**??**）
$T(S)$	种子节点集合 S 在图 $\mathcal{G}$ 中的传播效率（参见**??**）
$T'(S)$	种子节点集合 S 在超图 $\mathcal{H}$ 中的传播效率（参见**??**）

节点集合 $S\subseteq\mathcal{V}$，信息传播在独立级联模型下是如下运作的。定义 S_t 为在 $t\geq 0$ 的时刻激活的节点集合。显然，在 $t=0$ 时刻时，满足 $S_0=S$。在 $t+1$ 时刻，每一个在 t 时刻被激活的节点 $u\in S_t$ 会独立地去尝试激活它的出度边指向的未被激活的邻居节点 $v\in\mathcal{V}\setminus\cup_{0\leq i\leq t}S_i$，激活节点 v 的概率等于 $p_{u,v}$。当节点 u 尝试了去激活所有它的出度边指向的节点后，它在信息传播的之后过程中将不再会有机会去激活。即在 t 时刻被激活的节点 u 只会在 $t+1$ 时刻去尝试激活它的出度边指向的邻居节点。当 t 满足 $S_t=\emptyset$ 时，整个信息传播过程停止。

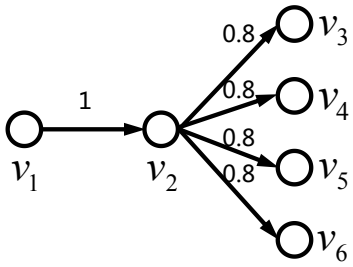

图 3.1　社交网络中的信息传播概率图 $\mathcal{G}$

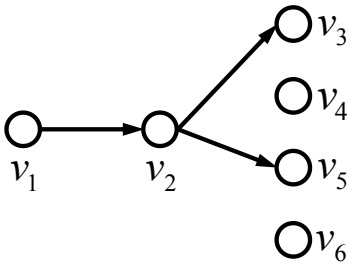

图 3.2　随机实例图 g

以图3.1中的社交网络 $\mathcal{G}$ 为例，考虑种子节点集合 $S=\{v_1\}$ 的信息传播过程。图3.1中边上的数值代表着节点之间的传播概率 $p^{\mathcal{G}}_{u,v}$。整个信息传播的过程可以描述如下。在 $t=0$ 时刻，由于节点 v_1 是种子节点集合中唯一的节点，因此节点 v_1 被激活。在 $t=1$ 时刻，因为节点 v_1 在 $t=0$ 时刻被激活，而且图 $\mathcal{G}$ 中存在一条从 v_1 到 v_2 概率为 1 的边，节点 v_1 将依概率 1 去激活节点 v_2。因此，节点 v_2 将在

$t=1$ 时刻被激活，$S_1=\{v_2\}$。此后，在 $t=2$ 时刻，节点 v_2 将尝试去激活节点 v_3、v_4、v_5 以及 v_6。假设这一次信息传播过程如图3.2所示，节点 v_3 和 v_5 被激活，则 $S_2=\{v_3,v_5\}$。然后，在 $t=3$ 时刻，由于被激活的节点没有后继节点可被激活，整个信息传播过程在此时刻停止。定义 $I(S)$ 为在种子节点集合是 S 的条件下，整个信息传播过程中激活的节点数目，代表信息传播过程中的影响力。在上述的图 $\mathcal{G}$ 的一次信息传播过程中，影响力 $I(S)=4$。

给定一个种子节点集合 S，定义 $\mathbb{E}_{\mathcal{G}}[I(S)]$ 表示种子节点集合在图 $\mathcal{G}$ 中影响力的期望，它等于以种子节点集合为传播源，在图 $\mathcal{G}$ 中信息传播结束时激活节点数目的期望值。在独立级联模型下，影响力最大化问题的目标是寻找一个大小至多为 k 的种子节点集合，使得影响力函数的期望值 $\mathbb{E}_{\mathcal{G}}[I(S)]$ 最大化。给定一个输入 k，影响力最大化问题可以形式化为如下，

$$\begin{aligned} &S^*=\arg\max \mathbb{E}_{\mathcal{G}}[I(S)] \\ &s.t.\ \ S\subseteq\mathcal{V},|S|=k \end{aligned} \tag{3.1}$$

以图3.1的社交网络 $\mathcal{G}$ 为例，给定输入 $k=1$ 来考虑影响力最大化问题。为了解决例子中的影响力最大化问题，根据公式3.1所示，我们需要计算所有 $k=1$ 的种子节点集合的影响力期望值，即每个节点的影响力期望值。对于节点 v_1 组成的种子节点集合，其影响力期望值 $\mathbb{E}_{\mathcal{G}}[I(\{v_1\})]=1+1+4\times0.8=5.2$。对于种子节点集合 $\{v_2\}$，影响力期望值 $\mathbb{E}_{\mathcal{G}}[I(\{v_2\})]=1+4\times0.8=4.2$。对于其他的种子节点集合，$\mathbb{E}_{\mathcal{G}}[I(\{v_3\})]=\mathbb{E}_{\mathcal{G}}[I(\{v_4\})]=\mathbb{E}_{\mathcal{G}}[I(\{v_5\})]=\mathbb{E}_{\mathcal{G}}[I(\{v_6\})]=1$。由此，我们可以得出，在给定图 $\mathcal{G}$ 以及 $k=1$ 的条件下，影响力最大化问题的最优解为 $S^*=\{v_1\}$。

在上述的例子中，因为图 $\mathcal{G}$ 的结构简单，并且种子节点集合的大小 $k=1$，所以能够直接计算得出种子节点集合的影响力期望值，从而选择得出最优解。而在实际情况下，图 $\mathcal{G}$ 的结构往往会复杂得多，而且 $k>1$，很难直接计算种子节点集合的影响力期望值。Kempe 等人[22] 首先证明了在独立级联模型下，影响力最大化问题是一个 NP-hard 问题。因此，直接计算出有影响力的节点是十分困难的。为了解决此问题，Kempe 等人又证明了在独立级联模型下，影响力函数的期望 $\mathbb{E}_{\mathcal{G}}[I(S)]$ 是单调的以及子模的。这两个性质为求解影响力最大化问题的近似算法提供了理论保证。形式上，一个单调的函数对于任意的节点 u 和任意集合 S，都满足 $f(S\cup\{u\})\geq f(S)$。而一个子模的函数对于任意的节点 u 以及任意的两个集合 $S\subseteq W$，都满足 $f(S\cup\{u\})-f(S)\geq f(W\cup\{u\})-f(W)$。针对具有单调性以及子模性的函数，Nemhauser 等人[23] 提出了一种朴素的贪心算法来解决此类问题。算法的核心思想是首先从一个空的种子节点集合 $S=\emptyset$ 开始，重复地选择当前边际

收益（即 $f(S \cup \{u\}) - f(S)$）最大的节点 u 加入到种子节点集合 S 中，直到满足种子节点集合的大小为 k 时结束迭代。选择节点 u 的准则可以形式化如下，

$$u = \arg\max_{w \in V \setminus S} \left(\mathbb{E}_{\mathcal{G}}[I(S \cup w)] - \mathbb{E}_{\mathcal{G}}[I(S)]\right) \tag{3.2}$$

Nemhauser 等人证明了朴素的贪心算法得出解以 $1 - 1/\mathrm{e}$ 的因子近似于最优解，即任意一个由贪心算法的出来的解 S 都满足 $I(S) \geq (1 - 1/\mathrm{e})I(S^*)$，其中 S^* 代表最优解，e 为自然常数。虽然贪心算法的核心思想比较简单，但是由于计算影响力期望值的过程是 #P-hard[24] 的问题，因此实现该算法并不是简单的。为了解决该问题，Kempe 等人[22] 提出了使用**蒙特卡罗方法**（*Monte Carlo method*）来对 $\mathbb{E}_{\mathcal{G}}[I(S)]$ 在一定精度内进行估计。蒙特卡罗方法的步骤如下。假设我们对图 $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ 中的所有的边 $e \in \mathcal{E}$ 都进行抛硬币实验，图 $\mathcal{G}$ 中边的连接的概率为 $p(e)$，我们以 $1 - p(e)$ 的概率移除掉边 e。定义 g 为得到的结果图，$R_g(S)$ 为在图 g 中从种子节点集合 S 出发可达的节点集合。我们需要注意的是，图 g 中的边不再是概率性连接的边，而是确定性的边。对于任意节点 $v \in g$，如果在图 g 中存在一条路径从节点集合 S 出发到达节点 v，那么称节点 v 是从节点集合 S 可达的。Kempe 等人[22] 证明了 $R_g(S)$ 的期望值与 $\mathbb{E}_{\mathcal{G}}[I(S)]$ 是相等的，即可表示为如下，

$$\mathbb{E}_{\mathcal{G}}[I(S)] = \mathbb{E}_{g \sim \mathcal{G}}\left[I_g(S)\right] \tag{3.3}$$

其中 $I_g(S) = \left|R_g(S)\right|$，即在图 g 中的影响力等于从种子节点集合出发可达的节点数目。因此，我们可以通过估计 $R_g(S)$ 的期望值来估计 $\mathbb{E}_{\mathcal{G}}[I(S)]$，即通过估计图 g 中可达的节点数目的期望来估计原图 $\mathcal{G}$ 中的影响力期望值。在实际操作中，我们首先根据原来的社交网络生成多个实例 $g \sim \mathcal{G}$，然后对每一个实例进行计算其影响力 $I_g(S)$，最终计算其平均值作为 $\mathbb{E}_{\mathcal{G}}[I(S)]$ 的一个估计。假设我们在估计 $\mathbb{E}_{\mathcal{G}}[I(S)]$ 的过程中生成了 r 个实例图 g，并且 r 足够大，那么在独立级联模型下，贪心算法能够得到一个 $(1 - 1/\mathrm{e} - \varepsilon)$ 的近似最优解，其中 ε 是一个与图 $\mathcal{G}$ 和 r 相关的常数[25, 26]。一般来说，Kempe 等人建议设置 $r = 10,000$，许多其他的工作都采用了相似的设置参数。

尽管朴素的贪心算法是有效的，但是在复杂网络的应用中，算法的效率是极其低的。算法的时间复杂度为 $O(knmr)$。确切来说，算法进行了 k 次迭代来选择种子节点集合，每一次迭代需要对 $O(n)$ 个节点进行影响力的期望值的估计。每一次估计需要对生成的 r 个实例进行计算，而每一次计算需要消耗 $O(m)$ 的时间。因此，整个计算过程的时间复杂度为 $O(knmr)$。

对于具有子模性的函数，**惰性计算**（*lazy evaluations*）技术是一种比较知名的优化方法，它能够大大的降低计算的次数，而不改变贪心算法的输出。这个技术首先由 Minoux[27] 作为一种加速的贪心算法提出，Leskovec[28] 等人通过实验验证了惰性计算针对影响力最大化问题能够加速近 700 倍。

即使惰性计算能够提高计算的性能，但是贪心算法仍然在效率上是不足的。究其原因，贪心算法的弊端主要是在于计算影响力期望值的过程中，它需要对 $O(kn)$ 个节点进行估计。然而，其中大多数估计都是无用的，因为我们只关心影响力期望值最大的节点。在朴素贪心算法的框架下，这些无用的计算又是不可避免的。

为了解决朴素贪心算法的这一弊端，Borgs 等人[25] 提出了一种新的方法，突破了朴素贪心算法的限制。Tang 等人[29] 将这种方法称之为**反向传播采样**（*Reverse Influence Sampling*），并且阐述了其工作原理。为了解释反向传播采样算法的工作原理，我们首先引入如下的概念。

定义 3.1 (**反向可达集合**): 给定一个图 $\mathcal{G} = (\mathcal{V}, \mathcal{E})$，我们对图中的 $\mathcal{G}$ 中的每一条边 $e \in \mathcal{E}$ 进行抛硬币实验，依概率 $1 - p(e)$ 移除掉边 e。定义 g 为得到的图，对于任意的节点 $v \in \mathcal{V}$，节点 v 在图 g 中的反向可达集合 $RR(v)$ 定义为图 g 中可达节点 v 的节点集合。这就是说，如果节点 $u \in RR(v)$，则至少在图 g 中存在一条路经从节点 u 到达节点 v。

根据定义3.1可知，如果节点 u 在节点 v 的反向可达集合 $RR(v)$ 中，那么节点 u 在图 $\mathcal{G}$ 中能够依一定概率通过一条路径到达节点 v。这也就表示，如果采用节点 u 作为种子节点集合 $S = \{u\}$ 在图 $\mathcal{G}$ 中进行信息传播，那么节点 u 是有一定概率激活节点 v 的。

3.2.2 传播效率最大化问题

3.3 方法描述

3.4 实验分析

3.5 本章小结

致　谢

衷心感谢导师 xxx 教授和 xxx 副教授对本人的精心指导。他们的言传身教将使我终生受益。

感谢 NudtPaper，它的存在让我的论文写作轻松自在了许多，让我的论文格式规整漂亮了许多。

参考文献

[1] Lin J, Efron M, Wang Y, et al. Overview of the TREC-2015 Microblog Track [C]. In Proceeddings of the 24th Text REtrieval Conference (TREC 2015). 2015.

[2] Ounis I, Macdonald C, Lin J, et al. Overview of the TREC-2011 Microblog Track [C]. In Proceeddings of the 20th Text REtrieval Conference (TREC 2011). 2011.

[3] Soboroff I, Ounis I, Macdonald C, et al. Overview of the TREC-2012 Microblog Track. [C]. In Proceeddings of the 21th Text REtrieval Conference (TREC 2012). 2012.

[4] Lin J, Efron M. Overview of the TREC-2013 Microblog Track [C]. In Proceeddings of the 22th Text REtrieval Conference (TREC 2013). 2013.

[5] Lin J, Efron M, Wang Y, et al. Overview of the TREC-2014 Microblog Track [C]. In Proceeddings of the 23th Text REtrieval Conference (TREC 2014). 2014.

[6] Xie H, Li X, Wang T, et al. Personalized Search for Social Media via Dominating Verbal Context [J]. Neurocomputing. 2015, 172 (C): 27–37.

[7] Sontag D, Collins-Thompson K, Bennett P N, et al. Probabilistic models for personalizing web search [C]. In Proceedings of the fifth ACM international conference on Web search and data mining. 2012: 433–442.

[8] Wang H, He X, Chang M-W, et al. Personalized ranking model adaptation for web search [C]. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. 2013: 323–332.

[9] Tang L, Jiang Y, Li L, et al. Personalized recommendation via parameter-free contextual bandits [C]. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2015: 323–332.

[10] Liang C, Leng Y. Collaborative filtering based on information-theoretic co-clustering [J]. International Journal of Systems Science. 2014, 45 (3): 589–597.

[11] Vosecky J, Leung K W-T, Ng W. Collaborative personalized twitter search with topic-language models [C]. In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. 2014: 53–62.

[12] Xue G-R, Han J, Yu Y, et al. User language model for collaborative personalized search [J]. ACM Transactions on Information Systems (TOIS). 2009, 27 (2): 11.

[13] Yang X, Guo Y, Liu Y, et al. A survey of collaborative filtering based social recommender systems [J]. Computer Communications. 2014, 41: 1–10.

[14] Canuto S, Gonçalves M, Santos W, et al. An Efficient and Scalable MetaFeature-based Document Classification Approach based on Massively Parallel Computing [C]. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2015: 333–342.

[15] Severyn A, Moschitti A. Learning to rank short text pairs with convolutional deep neural networks [C]. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2015: 373–382.

[16] Ren Z, Peetz M-H, Liang S, et al. Hierarchical multi-label classification of social text streams [C]. In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. 2014: 213–222.

[17] Paik J H. A novel TF-IDF weighting scheme for effective ranking [C]. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. 2013: 343–352.

[18] Gao J, Xu G, Xu J. Query expansion using path-constrained random walks [C]. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. 2013: 563–572.

[19] Letelier A, Pérez J, Pichler R, et al. Static analysis and optimization of semantic web queries [J]. In: Proc. of the 31st Symposium on Principles of Database Systems (PODS. 2012, 38 (4): 84–87.

[20] Si J, Li Q, Qian T, et al. Users' interest grouping from online reviews based on topic frequency and order [J]. World Wide Web. 2014, 17 (6): 1321–1342.

[21] Domingos P, Richardson M. Mining the network value of customers [C]. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. 2001: 57–66.

[22] Kempe D, Kleinberg J, Tardos É. Maximizing the spread of influence through a social network [C]. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. 2003: 137–146.

[23] Nemhauser G L, Wolsey L A, Fisher M L. An analysis of approximations for maximizing submodular set functions—I [J]. Mathematical Programming. 1978, 14 (1): 265–294.

[24] Chen W, Wang C, Wang Y. Scalable influence maximization for prevalent viral marketing in large-scale social networks [C]. In in KDD 2010. 2010: 1029–1038.

[25] Borgs C, Brautbar M, Chayes J, et al. Maximizing social influence in nearly optimal time [C]. In Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms. 2014: 946–957.

[26] Kempe D, Kleinberg J, Tardos É. Influential nodes in a diffusion model for social networks [M] // Kempe D, Kleinberg J, Tardos É. Automata, languages and programming. Springer, 2005: 2005: 1127–1138.

[27] Minoux M. Accelerated greedy algorithms for maximizing submodular set functions [M] // Minoux M. Optimization Techniques. Springer, 1978: 1978: 234–243.

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

[29] Tang Y, Xiao X, Shi Y. Influence maximization: Near-optimal time complexity meets practical efficiency [C]. In Proceedings of the 2014 ACM SIGMOD international conference on Management of data. 2014: 75–86.

作者在学期间取得的学术成果

发表的学术论文

[1] Yang Y, Ren T L, Zhang L T, et al. Miniature microphone with silicon- based ferroelectric thin films. Integrated Ferroelectrics, 2003, 52:229-235. (SCI 收录, 检索号:758FZ.)

[2] 杨轶, 张宁欣, 任天令, 等. 硅基铁电微声学器件中薄膜残余应力的研究. 中国机械工程, 2005, 16(14):1289-1291. (EI 收录, 检索号:0534931 2907.)

[3] 杨轶, 张宁欣, 任天令, 等. 集成铁电器件中的关键工艺研究. 仪器仪表学报, 2003, 24(S4):192-193. (EI 源刊.)

[4] Yang Y, Ren T L, Zhu Y P, et al. PMUTs for handwriting recognition. In press. (已被 Integrated Ferroelectrics 录用. SCI 源刊.)

[5] Wu X M, Yang Y, Cai J, et al. Measurements of ferroelectric MEMS microphones. Integrated Ferroelectrics, 2005, 69:417-429. (SCI 收录, 检索号:896KM.)

[6] 贾泽, 杨轶, 陈兢, 等. 用于压电和电容微麦克风的体硅腐蚀相关研究. 压电与声光, 2006, 28(1):117-119. (EI 收录, 检索号:06129773469.)

[7] 伍晓明, 杨轶, 张宁欣, 等. 基于 MEMS 技术的集成铁电硅微麦克风. 中国集成电路, 2003, 53:59-61.

研究成果

[1] 任天令, 杨轶, 朱一平, 等. 硅基铁电微声学传感器畴极化区域控制和电极连接的方法: 中国, CN1602118A. (中国专利公开号.)

[2] Ren T L, Yang Y, Zhu Y P, et al. Piezoelectric micro acoustic sensor based on ferroelectric materials: USA, No.11/215, 102. (美国发明专利申请号.)

附录 A　模板提供的希腊字母命令列表

大写希腊字母：

Γ \Gamma	Λ \Lambda	Σ \Sigma	Ψ \Psi
Δ \Delta	Ξ \Xi	Υ \Upsilon	Ω \Omega
Θ \Theta	Π \Pi	Φ \Phi	
$\varGamma$ \varGamma	$\varLambda$ \varLambda	$\varSigma$ \varSigma	$\varPsi$ \varPsi
$\varDelta$ \varDelta	$\varXi$ \varXi	$\varUpsilon$ \varUpsilon	$\varOmega$ \varOmega
$\varTheta$ \varTheta	$\varPi$ \varPi	$\varPhi$ \varPhi	

小写希腊字母：

α \alpha	θ \theta	o o	τ \tau
β \beta	ϑ \vartheta	π \pi	υ \upsilon
γ \gamma	ι \iota	ϖ \varpi	ϕ \phi
δ \delta	κ \kappa	ρ \rho	φ \varphi
ϵ \epsilon	λ \lambda	ϱ \varrho	χ \chi
ε \varepsilon	μ \mu	σ \sigma	ψ \psi
ζ \zeta	ν \nu	ς \varsigma	ω \omega
η \eta	ξ \xi	$\varkappa$ \varkappa	$\digamma$ \digamma
$\upalpha$ \upalpha	$\uptheta$ \uptheta	o \mathrm{o}	$\uptau$ \uptau
$\upbeta$ \upbeta	$\upvartheta$ \upvartheta	$\uppi$ \uppi	$\upupsilon$ \upupsilon
$\upgamma$ \upgamma	$\upiota$ \upiota	$\upvarpi$ \upvarpi	$\upphi$ \upphi
$\updelta$ \updelta	$\upkappa$ \upkappa	$\uprho$ \uprho	$\upvarphi$ \upvarphi
$\upepsilon$ \upepsilon	$\uplambda$ \uplambda	$\upvarrho$ \upvarrho	$\upchi$ \upchi
$\upvarepsilon$ \upvarepsilon	$\upmu$ \upmu	$\upsigma$ \upsigma	$\uppsi$ \uppsi
$\upzeta$ \upzeta	$\upnu$ \upnu	$\upvarsigma$ \upvarsigma	$\upomega$ \upomega
$\upeta$ \upeta	$\upxi$ \upxi		

希腊字母属于数学符号类别，请用\bm 命令加粗，其余向量、矩阵可用\mathbf。