**北京航空航天大学计算机学院**

**硕士学位论文开题报告**

**论文题目**：基于背景热点的微博转发预测研究

**专 业**：计算机技术

**研究方向**：网络信息安全

**研 究 生**： 陈 江

**学 号**： ZY1306311

**指导教师**： 王丽宏（中心）

**指导教师:** 巢文涵（校内）

**北京航空航天大学计算机学院**

2014年11月26日

**目 录**

[1 课题来源 3](#_Toc397373403)

[2 论文选题的背景和意义 3](#_Toc397373404)

[3 国内外相关研究工作分析 4](#_Toc397373405)

[4 研究内容与初步设计方案 5](#_Toc397373406)

[4.1 论文研究目标 5](#_Toc397373407)

[4.2 论文主要研究内容 5](#_Toc397373408)

[4.2.1 基于多特征融合权重计算方法的背景热点表示和提取 5](#_Toc397373409)

[4.2.2 基于LDA的用户兴趣模型构建 5](#_Toc397373410)

[4.2.3 基于背景热点及用户兴趣的微博转发模型 6](#_Toc397373411)

[4.2.4 基于背景热点的微博转发预测 6](#_Toc397373412)

[4.3 初步设计方案 6](#_Toc397373413)

[4.3.1 微博及热点数据采集 7](#_Toc397373414)

[4.3.2 基于多特征融合权重计算方法的背景热点提取 7](#_Toc397373415)

[4.3.3 基于LDA的用户兴趣提取 8](#_Toc397373416)

[4.3.4 基于背景热点及用户兴趣的微博转发预测 8](#_Toc397373417)

[5 关键技术或难点 9](#_Toc397373418)

[6 论文研究计划 10](#_Toc397373419)

[6.1 研究计划 10](#_Toc397373420)

[6.2 预期成果 10](#_Toc397373421)

[7 主要参考文献 11](#_Toc397373422)

# 课题来源

# 论文选题的背景和意义

微博即微博客（Micro Blog）的简称，是一个基于用户关系的信息分享、传播以及获取平台，用户可以通过WEB、WAP以及各种客户端组建个人社区，以140字左右的文字更新信息，并实现即时分享。时至今日，微博在中国的发展已有5年历史，微博已经成为人们获取信息的重要手段之一，同时也是社会重要的信息传播渠道，政府、企业、公众人物都使用微博进行营销或舆论引导。伴随着和诸如QQ、微信等其他社交类应用之间的竞争和发展，微博已成功转化为社会化媒体，充分发挥着社交媒体的优势。

根据2014年7月CNNIC《2014年中国社交类应用用户行为研究报告》显示，过去半年中，使用过微博的网民达到43.6%，其中使用过新浪微博的网民比例最高，为28.4%，21.7%的网民经常访问新浪微博，用户忠诚度高。在微博这样的社会化媒体平台中，信息内容的传播通过人与人之间的“关注”和“被关注”网络，一层一层传播开来，这种传播方式速度快、覆盖面广。CNNIC调查显示，80.35%的新浪微博用户通过新浪微博关注新闻/热点话题，68.1%的新浪微博用户关注感兴趣的人，60.3%的新浪微博用户主动发微博（分享/转发信息），另外50%左右的新浪微博用户在微博上发照片、看视频/听音乐。新浪微博能够满足人们的各种需求，已经成为人们生活中一个主要的沟通交流平台，亦是一个大众的舆论平台，越来越成为人们了解时下热点信息的主要渠道之一。

微博作为新兴媒体，不仅具有社交媒体的属性，还具有很大的服务价值。很多政府机关、名人、新闻媒体都通过开通微博来增加与网民之间的互动；政府方面主要利用微博征求民众意见，让民众自由发表观点建议，以期在民众心中树立亲民民主的形象，名人们通过微博发表自己的正面积极有趣的信息以获得更多支持，新闻媒体则通过微博发表精短新闻以扩大知名度。

根据香港大学新闻及传媒研究中心助理教授傅景华（Fu King-wa）最新的一项研究表示，约1000万用户创造了新浪微博约94%的消息，与此同时，还有约两亿用户只是在“转发”这些消息和帖子。傅景华通过自己编写程序，对微博上发过消息的2.087亿用户进行了调查，发现其中约93.8%的消息只是由5%的用户发布。由此可知，新浪平台中微博的产生方式主要是“转发”。

**（关于选题意义：并不是别人未研究，我就要研究，研究的意义在于其在实际应用中是有价值和意义的，研究热点对社会、经济、政治、舆情方面、网络营销方面的影响、网络宣传方面的影响**）

**（热点对社会的影响：**）

**（热点对转发的影响）：**

微博不止是人们获取资讯的重要渠道，其在社会各方面产生的影响也是不容忽视的。一方面，微博能在一定程度上监督司法公正性，如2013年陕西“房姐”龚爱爱案、“高铁一姐”丁书苗案、北京大兴摔童案、沈阳摊贩杀死城管安等的热议使得司法公正成为空前的网络一体；另一方面，医患纠纷、城管摊贩矛盾、微博实名举报等使得民生、吏治同样受到网民的广泛关注。此外，在这个微博自媒体时代，微博在网络营销中的作用也是不可忽视的，成功的案例包括“@致青春”、“@京东”、“@我是江小白”等。

通过国内外研究现状发现，针对微博转发以及微博转发预测的研究已有不少，但鲜有考虑背景热点的微博转发预测研究。然而，实际上，用户的转发意愿除了受用户间关系、被转发用户影响力、用户兴趣等因素影响外，事实上，用户转发微博是有既定的知识背景的，这些背景知识包括微博热点、新闻热点、搜索热点等。

基于以上考虑，加之，随着微博的不断进化，微博已经从最初的社交为主的媒体转化为如今的传播资讯为主的媒体，微博已经成为媒体发布新闻、商家进行网络营销、用户关注热点和兴趣点的重要途径，因此背景热点和用户兴趣也成为转发微博不可忽视的重要因素。所以，针对用户转发微博具有既定知识背景这一事实，并考虑到用户兴趣对微博转发的影响，本论文将研究基于背景热点的微博转发问题，构建基于背景热点及用户兴趣的微博转发模型，以期达到预测用户是否会转发某条微博的目的。

# 国内外相关研究工作分析

关于微博转发行为的研究主要集中在以下两个方面：

1. **微博中信息传播特点及效果研究。**

李英乐等人考虑用户影响力、接受用户活跃度、接收用户兴趣相似度等用户特征，结合微博消息内容的重要性和用户亲密程度，将转发预测问题转化为一个而分类问题[[[1]](#endnote-1)]。Gomez等人[[[2]](#endnote-2)]利用社交网络节点的感染时间来研究网络结构，他们基于活跃节点以一定条件概率独立地感染相邻节点这一假设，提出了NETINF算法，该算法基于模函数最优化来发现级联传播路径。Gomez等人[[[3]](#endnote-3)]通过基于两个顶点间感染时间和传播概率的概率密度函数对一定时间内一个顶点感染另一个顶点的概率建模，扩展了NETINF。文献[[[4]](#endnote-4)]提出NETRATE算法，通过解决凸最大似然问题，计算出了点对之间的传播概率以及传播图谱。考虑到社交网络的网络拓扑是随时间而变化的，Gomez等人[[[5]](#endnote-5)]于2013年扩展NETRATE算法，提出了一个随时间变化的推理算法，INFOPATH，该算法利用随机梯度算法对随时间变化的网络拓扑结构进行实时估计。

1. **微博转发预测研究**

罗知林等人提出了一种基于特征的随机森林微博转发预测算法（RFMR），该论文所研究的特征包括用户间的微网络结构、权重比率和用户个人信息等[[[6]](#endnote-6)]。张旸等人采用机器学习的分类算法，并对微博上不同特征的重要性进行分析，提出了基于特征加权的预测模型[[[7]](#endnote-7)]。论文[[[8]](#endnote-8)][[[9]](#endnote-9)]分别提出了经典的传播预测模型：Independent Cascades(IC)和Linear Threshold(LT)，这两个模型都基于有向图理论。基于这两个模型，通过对传统IC模型和LT模型的同步假设条件进行异步扩展，Saito等人[[[10]](#endnote-10)]提出了AsIC模型和AsLT模型。另外，论文[[[11]](#endnote-11)]和论文[[[12]](#endnote-12)]提出了经典的非基于图的信息传播预测模型：SIR和SIS，非基于图一类的传播预测模型忽略特定图结构的存在。

另一部分研究人员将研究工作聚焦于给定内容的流行度预测问题。例如，关于总体活动数量的研究取得了很多的研究成果——每小时新闻总量预测[[[13]](#endnote-13)]每天的标签使用总量预测[[[14]](#endnote-14)]。S. Petrovic等人[[[15]](#endnote-15)]和W. Galuba等人[[[16]](#endnote-16)]则研究用户级别的预测问题：用户是否会转发某篇特定的微博。关于串级预测的研究，一些学者将串级预测转化为回归问题[[[17]](#endnote-17)][[[18]](#endnote-18)][[[19]](#endnote-19)][[[20]](#endnote-20)]或者二分类问题[[[21]](#endnote-21)][[[22]](#endnote-22)]。

以上都是和微博转发直接或间接相关的研究和工作。关于信息传播特点，有基于统计的研究，也有基于社交网络结构的研究；关于微博转发预测，目前的研究有基于用户行为的、基于特征的、基于社交网络结构和非基于社交网络结构的。所有的这些研究，都没有考虑到微博作为人们获取信息的重要渠道之一，热点事件和用户兴趣对微博转发具有不可忽视的影响，而热点事件不仅体现在微博热点中，同时也体现在一定时间范围内的新闻热点和搜索热点中，本论文将这些热点统一归为背景热点。所以，基于前人的研究工作，本论文将研究基于背景热点的微博转发预测问题，构建基于背景热点及用户兴趣的微博转发模型。

# 研究内容与初步设计方案

## 论文研究目标

本文的**研究目标**是分析背景热点对微博转发的影响，构建基于背景热点的微博转发模型，在此基础上，对微博转发进行预测。其中，背景热点包括微博热点、新闻热点以及搜索热点等。论文的基本假设是，用户转发一条微博客观上受热点事件影响，主观上受个人兴趣影响，当热点事件相关的微博没有出现在用户可视范围之内时，转发主要受个人兴趣的影响；另一方面，当面对一个热点事件时，即使不是用户感兴趣的内容，处于社会因素的影响，用户仍然可能转发微博；同时，当发生的热点事件刚好也是用户感兴趣的领域相关的事件时，用户转发相关微博的可能性会更大。故本文的研究主要聚焦在背景热点和用户兴趣，构建基于背景热点及用户兴趣的微博转发模型，从而对微博转发进行预测。

## 研究目标形式化定义

根据论文研究目标，将微博用户转发用户的微博形式化定义为：

：原博主相关特征；

：转发用户相关特征；

w：微博内容相关特征；

h：背景热点。

## 论文主要研究内容

当前对微博转发的研究，大多或基于特征、或基于用户行为、或基于社交网络，且大多只是在微博范围内考虑的，并未将背景热点作为驱动用户转发微博的动机之一考虑进来，而且，微博作为一种新兴的社会化媒体，与其他传统媒体或者网络媒体等不是完全独立开来的。所以，基于背景热点研究微博转发问题是有研究价值的；另一方面微博是字数不超过140字的短文本，当考虑新闻热点、搜索热点、最热微博等背景热点及其相关文档时，能够有效地弥补这一点，这对用户兴趣提取等工作也是极有帮助的。

综上所述，本文将以热点作为背景，兼顾用户兴趣，研究基于背景热点的微博转发预测问题，构建基于背景热点及用户兴趣的微博转发模型。主要研究内容图示如下：



图 1 主要研究内容框架图

### 多通道背景热点表示和提取方法

热点是指一定时间、一定范围内，公众最为关心的热点话题、热点事件、热点新闻等的总称。本文考虑的热点是指包括搜索热点、新闻热点、微博热点在内的多通道热点。通过百度搜索风云榜可以获得包括娱乐、人物、小说、游戏、汽车等多领域的搜索热点，并且可以获得每个搜索热点对应的相关新闻；新闻热点则可以通过新浪新闻排行榜、网易新闻排行榜、凤凰新闻排行榜以及今日头条等获得，这些网站同样也提供热点新闻的相关新闻；微博热点则可以新浪博客中的“微博最火”频道获得，此频道同样也提供每篇热门微博的相关微博。

为了得到背景热点简洁且信息量大的表示，本论文将提取背景热点相关新闻的热点词，并以此热点词来表示对应背景热点。本论文的热点词从背景热点相关新闻中提取，能够体现当前的热点事件、重要资讯和流行趋势，这样的热点词具有简洁、时效、信息量大、相关词语之间的互信息大等特点。

### 时序化微博用户兴趣提取

用户兴趣是决定用户是否转发一条微博的重要因素，本论文所讨论的用户兴趣是指通过用户所发布历史微博体现出来的用户偏好。在新浪平台，对于一个新注册用户，用户兴趣可以通过用户标签来反映，但随着用户转发微博量的增加，用户所转发或者发布的微博体现出来的用户兴趣虽然能在一定程度上体现标签所反映的用户兴趣，但却不可能是完全吻合的，因为一个人的兴趣虽然具有一定的稳定性，但并不是一直不变的。本论文称这种具有一定稳定性亦具有随时间的波动性的微博用户兴趣为时序化微博用户兴趣。

本论文的用户兴趣模型同时考虑了用户标签和用户历史微博。本论文假设用户标签是用户兴趣的初始状态，随着用户发博数量的增加，本论文将从历史微博中提取用户兴趣，然后结合用户兴趣的初始状态作为用户的兴趣度量。本论文假设用户兴趣可以通过用户标签、用户兴趣空间来标识，在这个用户兴趣标识模型中，随着时间的推进，用户标签和用户兴趣词在标识模型中所占的权重是变化的。用户标签是指用户固有属性里的用户标签，用户兴趣空间是指从用户历史微博中提取的用户兴趣表示。通过用户兴趣模型提取的用户兴趣，将结合其他特征构建微博转发模型。

### 融合背景热点和微博用户兴趣的微博转发预测模型

微博转发作为信息传播的一种具体形式，既具有信息传播所体现出来的网络特征，同时也具有信息传播所体现出来的社会学特征。一方面，用户之间的关注和被关注体现了用户之间的关系，亿级数量级的微博用户构成一张庞大的用户关系网络；另一方面，用户之间的微博转发，是用户获取资讯的方式，也是用户之间沟通交流的社会化行为之一。用户转发微博是具有既定的背景知识的，本文假设背景热点就是用户转发微博的背景知识，背景知识刺激用户转发微博时客观因素；主观上，用户转发微博很大程度上受用户兴趣的支配。所以，本论文将融合背景热点及微博用户兴趣构建微博转发模型。

### 基于背景热点的微博转发预测系统设计与实现方案

对于一个给定的微博转发数据集，微博是否被转发是既定的事实，一个微博转发预测模型或算法效果的好坏可以通过和数据集的比较得出，转发数据集中被转微博的转发概率为1，预测模型或算法预测的转发概率约接近于1，则预测效果越好。

有了基于背景热点及用户兴趣的微博转发模型之后，本论文将研究该模型对微博转发的预测效果，并对加入背景热点和不加入背景热点的模型预测效果进行比较，研究背景热点对微博转发预测的影响，设计基于背景热点的微博转发预测系统，并实现其功能。

## 初步设计方案

根据主要研究内容和研究目标，给出如下的初步设计思路：

### 数据采集

数据抓取模块主要包括两部分工作：微博数据的采集和热点数据采集。本论文将结合两种方式采集微博及热点数据：新浪API采集新浪微博数据、网页爬虫采集新闻热点数据和新浪微博数据。

1. **微博数据采集**

新浪API抓取新浪微博数据的优点是操作简单，采集的数据噪音少；缺点是开放程度有限，普通开发者能调用的API接口有限，且在调用频率上也有限制。本论文将采取新浪微博开放平台API和网页爬虫两种方式采集微博数据。

1. **热点数据采集**

网页爬虫采集数据的优点能够抓取整个网页内容，可以从中选择自己需要的数据提取出来；缺点是所采集的数据噪音大，同时也会受到反爬虫策略的限制。本论文将利用网页爬虫采集热点数据。

### 基于多特征融合权重计算方法的背景热点提取

借鉴于程肖等人基于多特征融合权重计算方法的热点提取方法，本论文拟采用多特征融合权重计算的方法提取背景热点。本文把噪音词提取所用到的语料, 按时间的顺序进行分割、预处理; 通过ICTCLAS进行分词, 对分词后的语料进行噪音词提取; 通过联合方差和在时间标签的基础上来提取噪音词。热点词提取所用到的新闻语料经过多级过滤, 包括: 停用词过滤、规则过滤和噪音词过滤, 根据词频、文档频率和词语权重多特征融合的热点词判断公式, 进行综合权重计算, 并降序排序, 实现热点词的提取。

1. **基于多级过滤的候选热点词提取**

本文拟采用中科院自动分词系统（ICTCLAS）先对语料进行分词然后对切分词进行多级过滤：

1. **停用词过滤**

载入一个常用停用词表，用该词表过滤常见的无意义词。该词表中包括130多个英文停用词，500多个中文停用词，以及一些标点符号、数字等特殊符号，攻击800余条。

1. **词性过滤**

对于特殊词性的词, 如数词、量词、代词、方位词、叹词等, 它们对于热点词的提取贡献不大, 如“一”、“条”、 “中间”、 “吗”等, 通过编写相应的规则, 判断其词性并进行过滤。

1. **噪音词过滤**

噪音词是语料中普遍存在的与主题无关的无意义词。为了找出噪音词，需要一些标准用语估计词的有效性。本文拟采用联合方差的概念，基于在语料库中出现的频数与包含该词的文档频数分别计算其在语料中发生的概率。联合方差的计算公式如下：

其中， , ,tf为词频，Ztf为全部词条的词频总和，为Ttf的均值，sum表示总的词条数。 , ,df为文档数，Zdf为全部文档数总和，为Tdf的均值，count表示总的文档数。

根据联合方差计算公式进行单个词的联合方差计算，依据联合方差的稳定性选取噪音词。在时间标签的基础上，联合方差越小，变化越小，说明该词的稳定性越好，这个词无意义的概率大，越容易被认为是噪音词。

1. **热点词提取**

用来进行热点词提取的背景热点相关新闻经过多级过滤后构成热点词集，在多级过滤的时候，把停用词表和实验提取的噪音词表加载到热点词提取算法中，可以把该类词表称为非热点词表，当发现该词的时候将其过滤掉，最后形成热点词集。热点词的特征包括词频（TF）、文档频数（DF）、反文档频率（IDF）。对于热点词的权值计算采用如下公式：

其中，为词的频率；表示在整体样本数据中出现词i的文档数目；M表示所有文档数目；|d|表示文档向量的长度；用来对命名实体赋予不同权重。

根据热点词提取的结果进行排序，确定其是否为热点词。热点词提取的结果进行排序，确定其是否为热点词。按测试语料中的词频、文档频数和权重总结出热点词的加权综合判断公式，更具HK的值进行排序。

综合文献[[[23]](#endnote-23)]的排序方法，得到如下公式：

其中， ，设定为经验值。

### 基于LDA的用户兴趣提取

LDA是一种文档主题生成模型，也成为一个三层贝叶斯概率模型，包含词、主题、文档三层结构。LDA是一种非监督学习技术，可以用来识别大规模文档集或语料库中潜藏的主题信息。LDA可以简明易懂地将每个文档表示出来，而且无论文档长度如何，都缩减为固定长度的向量。所以，本论文将LDA应用于用户历史微博数据集上，得到历史微博数据集上的多个主题，将其作为用户的兴趣表示。

### 兴趣与热点之间的关系

热点由热点词组成，热点词以一定的包含于热点；兴趣由兴趣词组成，兴趣词以一定概率包含于兴趣。

热点和兴趣相似是一个概率事件，可以表示成如下关系：



这样，热点和兴趣之间的相似概率便可以通过热点词和兴趣词之间的相似性得到。

### 基于背景热点及用户兴趣的微博转发预测

根据调研，一些研究人员从概率模型的角度去研究微博转发，构建模型，而另外一些人则将其转化为二分类问题，加以研究。本论文将考虑将微博转发妆花为二分类问题来研究。

鉴于微博的特点，可以用有向无权图G=(V,E)来表示，节点表示网络中的地个用户，边表示用户和的关注关系，其方向表示信息传播的方向，指向粉丝用户一侧。

假设用户发布了一条与热点h相关的微博消息，表示粉丝在看到后采取的行为，当表示转发，表示不转发。因此，转发行为预测问题可以描述为：在已知，，和的条件下，寻找一个目标函数，，将，，和映射到-1和+1两个类别中，这是一个典型的二分类问题。对转发行为的预测是针对一个特定的关注关系而言，是网络G中的一个局部预测问题。

基于以上分析，本论文拟选取合适的特征，本论文拟考虑的特征有：

基于SVM算法构建微博转发预测模型。本文将提取尽可能多的特征，分析特征对微博转发的影响，最后选取合适的特征，作为SVM算法的特征，构建微博转发预测模型。

本文拟考虑的特征有：

1. 基于内容的特征

* 微博用户兴趣：由用户历史微博体现出来的用户特征；
* 背景热点：来自搜索、新闻、微博的多通道热点；
* 微博是否含有链接

1. 基于被转发用户特征

* 粉丝数、关注数、发博量
* 活跃度：一定时间范围内发微博数量
* 用户性别

1. 基于转发用户特征、

* 粉丝数、关注数、发博量
* 活跃度
* 用户性别

本论文将在以上特征中选取合适的特征作为机器学习算法的输入，构建微博转发预测模型。本论文将考虑的机器学习算法还包括线性回归、k-近邻分类的算法，最后比较不同模型的预测效果，选择最合适的模型。

为了比较不同模型的效果，本论文还将考虑朴素贝叶斯、多元线性回归、k-近邻分类等模型，比较不同模型应用于微博转发预测的效果。最后，选取效果最好的预测模型。

# 关键技术或难点

本论文在研究过程中存在的关键技术或难点主要有以下几个方面：

1. 影响微博转发的因素有很多，如何选取合适的特征构建微博转发模型将直接影响到转发预测的准确程度；
2. 多通道背景热点提取和表示：来自于微博热点、新闻热点和搜索热点等多通道的热点存在交叉融合的部分，也存在差异性。单个通道热点简洁、大信息量的表示和提取，以及多通道融合热点的表示和提取，都是本文不可回避的重点和难点；
3. 如何将背景热点作为影响用户转发微博的影响因素之一构建微博转模型，并取得良好的预测效果，是本文的主要工作也是重点和难点；
4. 如何选择适用于本文所选特征的微博转发模型，使得其能够取得好的预测效果，是本文的最终目标，同时也是难点。

# 论文研究计划

## 研究计划

表1列出了论文的研究计划，论文研究分为六个阶段，每个阶段都有相应的目标和任务。

表1 研究计划表

|  |  |
| --- | --- |
| 2014.11 — 2014.12 | 学习相关资料和技术 |
| 2015.01 — 2015.03 | 系统概要设计，技术尝试 |
| 2015.03 — 2015.05 | 系统详细设计 |
| 2015.05 — 2015.09 | 编码实现，撰写小论文 |
| 2015.09 — 2015.11 | 系统完善，文档整理 |
| 2015.11 — 2015.12 | 撰写毕业论文，答辩 |

## 预期成果

1. 基于LDA的微博用户兴趣模型的构建算法
2. 基于背景热点及用户兴趣的微博转发模型
3. 基于背景热点的微博转发预测算法及系统

# 主要参考文献

1. [] 李英乐, 于洪涛, 刘力雄. 基于SVM的微博转发规模预测方法[J]. 计算机应用研究, 2013, 30(9). DOI:10.3969/j.issn.1001-3695.2013.09.008. [↑](#endnote-ref-1)
2. [] Gomez Rodriguez M, Leskovec J, Krause A. Inferring networks of diffusion and influence[C]//Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010: 1019-1028. [↑](#endnote-ref-2)
3. [] Rodriguez M G, Balduzzi D, Schölkopf B. Uncovering the temporal dynamics of diffusion networks[J]. arXiv preprint arXiv:1105.0697, 2011. [↑](#endnote-ref-3)
4. [] I. CVX Research. CVX: Matlab software for disciplined convex programming,version 2.0 beta. <http://cvxr.com/cvx>, sep 2012. [↑](#endnote-ref-4)
5. [] Gomez Rodriguez M, Leskovec J, Schölkopf B. Structure and dynamics of information pathways in online media[C]//Proceedings of the sixth ACM international conference on Web search and data mining. ACM, 2013: 23-32. [↑](#endnote-ref-5)
6. [] 罗知林, 陈挺, 蔡皖东. 一个基于随机森林的微博转发预测算法[J]. 计算机科学, 2014, (4):62-64. [↑](#endnote-ref-6)
7. [] 张旸, 路荣, 杨青. 微博客中转发行为的预测研究[J]. 中文信息学报, 2012, 26(4). DOI:10.3969/j.issn.1003-0077.2012.04.015. [↑](#endnote-ref-7)
8. [] Goldenberg J, Libai B, Muller E. Talk of the network: A complex systems look at the underlying process of word-of-mouth[J]. Marketing letters, 2001, 12(3): 211-223. [↑](#endnote-ref-8)
9. [] Granovetter M. Threshold models of collective behavior[J]. American journal of sociology, 1978: 1420-1443. [↑](#endnote-ref-9)
10. [] Saito K, Ohara K, Yamagishi Y, et al. Learning diffusion probability based on node attributes in social networks[M]//Foundations of Intelligent Systems. Springer Berlin Heidelberg, 2011: 153-162. [↑](#endnote-ref-10)
11. [] Hethcote H W. The mathematics of infectious diseases[J]. SIAM review, 2000, 42(4): 599-653. [↑](#endnote-ref-11)
12. [] Newman M E J. The structure and function of complex networks[J]. SIAM review, 2003, 45(2): 167-256. [↑](#endnote-ref-12)
13. [] Yang J, Counts S. Predicting the Speed, Scale, and Range of Information Diffusion in Twitter[J]. ICWSM, 2010, 10: 355-358. [↑](#endnote-ref-13)
14. [] Ma Z, Sun A, Cong G. On predicting the popularity of newly emerging hashtags in twitter[J]. Journal of the American Society for Information Science and Technology, 2013, 64(7): 1399-1410. [↑](#endnote-ref-14)
15. []Petrovic S, Osborne M, Lavrenko V. RT to Win! Predicting Message Propagation in Twitter[C]//ICWSM. 2011. [↑](#endnote-ref-15)
16. [] Galuba W, Aberer K, Chakraborty D, et al. Outtweeting the twitterers-predicting information cascades in microblogs[C]//Proceedings of the 3rd conference on Online social networks. USENIX Association, 2010: 3-3. [↑](#endnote-ref-16)
17. [] Bakshy E, Karrer B, Adamic L A. Social influence and the diffusion of user-created content[C]//Proceedings of the 10th ACM conference on Electronic commerce. ACM, 2009: 325-334. [↑](#endnote-ref-17)
18. [] Kupavskii A, Ostroumova L, Umnov A, et al. Prediction of retweet cascade size over time[C]//Proceedings of the 21st ACM international conference on Information and knowledge management. ACM, 2012: 2335-2338. [↑](#endnote-ref-18)
19. [] Szabo G, Huberman B A. Predicting the popularity of online content[J]. Communications of the ACM, 2010, 53(8): 80-88. [↑](#endnote-ref-19)
20. [] Tsur O, Rappoport A. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities[C]//Proceedings of the fifth ACM international conference on Web search and data mining. ACM, 2012: 643-652. [↑](#endnote-ref-20)
21. [] Hong L, Dan O, Davison B D. Predicting popular messages in twitter[C]//Proceedings of the 20th international conference companion on World wide web. ACM, 2011: 57-58. [↑](#endnote-ref-21)
22. [] Jenders M, Kasneci G, Naumann F. Analyzing and predicting viral tweets[C]//Proceedings of the 22nd international conference on World Wide Web companion. International World Wide Web Conferences Steering Committee, 2013: 657-664. [↑](#endnote-ref-22)
23. []丁伟莉, 赵华, 郑德权等. 中文Blog热门话题检测与排序技术研究[C]. //中文信息处理前沿进展-中国中文信息学会二十五周年学术会议. 2006. [↑](#endnote-ref-23)