分类号 TP391.1 学号 10069068

UDC 密级 公开

工学博士学位论文

社交媒体中的主观性信息挖掘与分析

博士生姓名 谢松县

学科专业 计算机科学与技术

研究方向 自然语言处理

指导教师 王挺 教授

国防科学技术大学研究生院

二〇一四年十月

Opinion Mining and Analysis in Social Media

Candidate: Xie Songxian

Supervisor: Professor Wang Ting

A dissertation

Submitted in partial fulfillment of the requirements

for the degree of Doctor of Engineering

in Computer Science and Technology

Graduate School of National University of Defense Technology

Changsha, Hunan, P. R. China

October 10, 2014

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：社交媒体中的主观性信息挖掘与分析

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：社交媒体中的主观性信息挖掘与分析

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

现在的互联网上社交媒体随处可见，这给信息检索和传播分析工作带来了机遇与挑战。本文主要围绕在社交媒体中如何找到重要的信息以及信息是如何传播的展开。我们将 Twitter 作为研究对象，因为它是目前最著名的社交媒体之一，并且数据是公开的。这样从隐私的角度考虑，获取研究数据变得容易且能很好的为研究任务（如信息检索）服务。

信息检索的主要任务是在文档集合中，找到与给定话题相关的客观文本或主观文本。Twitter 是一个丰富的包含各种话题及其评论信息的资源库，本文将探讨如何在 Twitter 中找到相关的信息。但是 tweet 的短小化和非正式的文本特点，使得 Twitter 中的检索不同于以往的检索任务（如，网页检索）。本文将通过研究 tweet 文本特点和特有的 Twitter 社交媒体属性帮助 Twitter 检索。另外，Twitter 中信息的传播是一种普遍现象且与消息的质量相关（帮助 Twitter 中检索高质量的信息）。因此，我们从 tweet 本身和用户的角度，研究哪些因素影响了 tweet 的转发和人的转发行为。

我们的工作主要有四个部分：(1) 利用结构化信息的 Twitter 检索；(2) Twitter 观点检索；（3）Twitter 中传播观点的发现；（4）Twitter 中信息传播者的发现。四个工作具体如下：

利用结构化信息的 Twitter 检索： *Twitter 检索是在 Twitter 中找到与给定话题相关的 tweet 的任务*。绝大部分的 Twitter 检索系统在构造检索模型时一般都认为 tweet 是一个平面文本，但用户在编辑 tweet 时的一些习惯使得 tweet 文本呈现结构化的特点。这种结构化是通过一些不同的文本积木块组合而成，积木类型具体包括平面文本、主题词、链接、提及等。每一种积木都有自己独特的本质，一系列积木的排序组合又反映了一定的话语转换。以往的研究发现，通过开发文本的结构信息能够帮助结构化文本的检索（例如，网页检索）。本工作通过积木结构开发 tweet 的结构化信息，以此帮助 Twitter 检索。我们利用积木及其排列组合开发了一系列特征，并将其应用到排序学习的框架中。我们发现利用结构化 tweet 的方法进行检索能够达到目前最好的 Twitter 检索方法效果，将结构化 tweet 的方法和其他社交媒体特征一起使用能够进一步提高 Twitter 的检索效果。

Twitter 观点检索： *观点检索是在数据中找到对指定话题表达正面或反面观点的 tweet 的任务*。人们几乎在 Twitter 中表达了任何话题的观点，使其成为一个丰富的观点资源库。但是 Twitter 中也存在大量的垃圾信息和各种不同类型的文

本，使得 Twitter 中的观点检索充满挑战。我们提出了如何利用 tweet 的社交媒体信息和文本结构化信息的方法帮助 Twitter 的观点检索。特别的，基于排序学习，我们发现 tweet 的用户信息（如用户包含朋友的数目）、tweet 文本本身的结构信息和观点化程度影响着 tweet 的排序结果。实验结果表明社交媒体信息能够帮助 Twitter 的观点检索。基于无监督学习评价 tweet 观点化程度，并以此开发特征形成的检索方法能够到达手工标注 tweet 的有监督方法的检索效果，且这种方法能够帮助观点检索中话题依赖问题的解决。最后，我们在重新标注的 TREC Tweets2011 数据集上进一步验证了我们 Twitter 观点检索方法的有效性。

Twitter 中传播观点的发现： Twitter 已经变成人们收集观点做出决策的重要资源，但是数量众多且差异巨大的观点严重影响了人们使用这些资源的效果。本文我们考虑了*如何在 Twitter 中找到传播观点的任务—tweet 不仅表达了对某些话题的观点，且这个 tweet 在未来会被转发*。利用排序学习模型，我们开发了一系列特征，具体包括 tweet 的传播度特征、观点化特征和文本质量特征。实验结果证明了我们开发的特征对于 Twitter 中传播观点的发现是有效的，并且将所有特征整合的方法在发现效果上能够显著优于 BM25 方法和 Twitter 观点检索方法。最后，我们发现我们的方法在预测观点传播上可以达到人预测的水平。

Twitter 中信息传播者的发现： Twitter 和其它社交网络中一个重要的交流机制就是消息传播 —人们分享其他人创建的消息。虽然目前有许多工作研究了 Twitter 中的 tweet 是如何传播的（转发），但是一个未解决的问题是到底**谁**会转发给定的 tweet。这里我们考虑了*在 Twitter 中给定一条 tweet，发现作者的粉丝中谁会转发*。利用排序学习模型的框架，我们设计了一些特征，包括用户历史的转发信息，用户自身的社交媒体特征，用户使用 Twitter 的活跃时间，以及用户的个人兴趣。我们发现经常转发和提及作者的粉丝和与作者有相同兴趣爱好的人最有可能成为信息传播者。

通过以上四个问题的研究，我们发现 tweet 的文本信息和 Twitter 的社交媒体特征能够帮助 Twitter 信息检索和传播分析。

关键词: Twitter; **信息检索**; **观点检索**; **传播观点**; **信息传播者**

ABSTRACT

Social Media is now ubiquitous on the internet, generating both new possibilities and new challenges in information retrieval and propagation analysis. This thesis focus on finding important information and propagated information analysis in Social Media. We take Twitter as our research subject, since it is one of the most Social Media and public by default, which makes the data less problematic from a privacy standpoint, far easier to obtain and more amenable to target applications (such as information retrieval).

The main tasks in information retrieval are finding related objective or subjective documents about some topics in collection. Twitter is rich resource which contains information about various topics and opinions. Here we investigate how to find these information in Twitter. However, Twitter retrieval is different from traditional retrieval tasks (e.g, web search), since the text of tweet is short and informal. In this study we exploit textual features of tweet and the social media features to improve Twitter retrieval. Additionally, information dissemination is a prevalent phenomenon in Twitter and is related to the quality of message (which can help finding high quality information in Twitter). Therefore, from the point of view of tweets and users, we study the factors which affect tweet retweeting and users' retweeting behavior.

Our work can be divided into four parts: (1) improving Twitter retrieval by exploiting structural information, (2) opinion retrieval in Twitter, (3) finding propagated opinion in Twitter, (4) finding retweeters in Twitter. We introduce the four work in detail as follows:

Improving Twitter retrieval by Exploiting structural information. *Twitter retrieval deals with finding related tweets about some topics in Twitter.* Most Twitter search systems generally treat a tweet as a plain text when modeling relevance. However, a series of conventions allows users to tweet in structural ways using combination of different blocks of texts. These blocks include plain texts, hashtags, links, mentions, etc. Each block encodes a variety of communicative intent and sequence of these blocks captures changing discourse. Previous work shows that exploiting the structural information can improve the structured document (e.g., web pages) retrieval. In this study we utilize the structure of tweets, induced by these blocks, for Twitter retrieval. A set of features, derived from the blocks of text and their combinations, is used into a learning-to-rank scenario. We show that structuring tweets can achieve state-of-the-art performance. Our

approach does not rely upon social media features, but when we do add this additional information, performance improves significantly.

Opinion retrieval in Twitter. *Opinion retrieval deals with finding relevant documents that express either a negative or positive opinion about some topics*. Social Networks such as Twitter, where people routinely post opinions about almost any topic, are rich environments for opinions. However, spam and wildly varying documents makes opinion retrieval within Twitter challenging. Here we demonstrate how we can exploit social and structural textual information of tweets and improve Twitter-based opinion retrieval. In particular, within a learning-to-rank technique, we explore the question of whether aspects of an author (such as the number of friends they have), information derived from the body of tweets and opinionatedness ratings of tweets can improve performance. Experimental results show that social features can improve retrieval performance. Retrieval using a novel unsupervised opinionatedness feature achieves comparable performance with a supervised method using manually tagged Tweets. Topic-related specific structured Tweet sets are shown to help with query-dependent opinion retrieval. Finally, we further verify the effectiveness of our approach for opinion retrieval in re-tagged TREC Tweets2011 corpus.

Finding Propagated opinions in Twitter. Twitter has become an important source for people to collect opinions to make decisions. However the amount and the variety of opinions constitute the major challenge to using them effectively. Here we consider the problem of *finding propagated opinions – tweets that express an opinion about some topics, but will be retweeted*. Within a learning-to-rank framework, we explore a wide spectrum of features, such as retweetability, opinionatedness and textual quality of a tweet. The experimental results show the effectiveness of our features for this task. Moreover the best ranking model with all features can outperform a BM25 baseline and state-of-the-art for Twitter opinion retrieval approach. Finally, we show that our approach equals human performance on this task.

Finding retweeters in Twitter. An important aspect of communication in Twitter (and other Social Networks) is message propagation – people creating posts for others to share. Although there has been work on modelling how tweets in Twitter are propagated (retweeted), an untackled problem has been **who** will retweet a message. Here we consider the task of *finding who will retweet a message posted on Twitter*. Within a learning-to-rank framework, we explore a wide range of features, such as retweet history, followers

status, followers active time and followers interests. We find that followers who retweeted or mentioned the author's tweets frequently before and have common interests are more likely to be retweeters.

Based on the study of four work above, we find the textual information of tweet and social media features in Twitter can help Twitter retrieval and propagation analysis.

Key Words: Twitter; Information Retrieval; Opinion Retrieval; Propagated Opinion; Retweeter

第一章　绪论

1.1　研究背景

1.1.1　社交媒体

作为划时代的创新，互联网 20 年以来已深刻影响和改变着我们的生活，思维和行为方式。尤其现在，我们可以通过手机、各种穿戴式智能设备，随时随地保持与互联网不间断联系。根据中国互联网络信息中心的权威报告，截至 2014 年 7 月，我国网民规模达 6.41 亿，手机网民规模已超过 5 亿，互联网普及率为 47.4%[1]。随着互联网技术的迅猛发展，出现了形形色色吸引用户参与的社交媒体（Social Media）平台，并且已经成为人类工作、学习、生活必不可少的重要部分。图 1.1展示了在线的各种国内外的中英文社交媒体平台。

图 1.1　形形色色的中英文社交媒体

社交媒体中的互联网用户不但是单纯的信息接收者，也已经成为信息的制造者，人们通过社交媒体平台进行交流获取和产生信息。目前，中国拥有 12 亿手机用户、5 亿微博用户、5 亿微信用户，每天信息发送量超过 200 亿条，交流无处不在、无时不有。表 1.1中可以看出，根据互联网网站流量信息公司 Alexa[2]统计，流量前十的互联网网站中社交媒体平台占了绝大部分。

那么什么是社交媒体呢？作为社交媒体的维基百科是这样定义的[3]：

[1] http://www.cnnic.cn/hlwfzyj/hlwfzzx/qwfb/201408/t20140825_47878.htm
[2] www.alexa.com
[3] http://en.wikipedia.org/wiki/Socialmedia/

表 1.1　Alexa 统计访问量前十名网站

排名	网站	排名	网站
1	Google.com	6	**Wikipedia.org**
2	**Facebook.com**	7	**Amazon.com**
3	**Youtube.com**	8	**Twitter.com**
4	Yahoo.com	9	**Qq.com**
5	Baidu.com	10	**Taobao.com**

Social media are media for social interaction, using highly accessible and scalable communication techniques. It is the use of web-based and mobile technologies to turn communication into interactive dialogue.

一般来讲，社交媒体可以分为如表 1.2所示的几种类型：

表 1.2　社交媒体的类型

类型	代表性网站
Wiki	Wikipedia，Scholarpedia
Blogging	Blogger，LiveJournal，WordPress，博客
Social News	Digg, Mixx，Slashdot
Micro Blogging	Twitter，Google Buzz
Opinion & Reviews	ePinions，Yelp
Question Answering	Yahoo! Answers，百度知道
Media Sharing	Flickr，Youtube
Social Bookmarking	Delicious，CiteULike
Social Networking	Facebook，LinkedIn，MySpace

从表中可以看出，社交媒体有多中不同类型，因此会产生多种不同格式的数据，包括文本、图像以及视频等。Kaplan 和 Haenlein[1] 从数据和信息流动角度讨论了社交媒体。首先从媒体部分（media），社交媒体中最突出的方面是它区别于电视，广播和报纸等传统媒体。在传统媒体中，信息的流动是从几个内容生产者到众多用户消费者。在社交媒体中，内容产生的权利转移到了传统媒体的消费者的手中，而且信息流动的方式更加不确定。内容消费者和生产者可以多次在瞬间改变自己的角色。另外，为什么我们称这种新媒体为社会化的（social）媒体。社会化意味着信息内容不只是由个体用户产生，更是与其他用户的协作产生。因此内容（content）变得更加多样化，因为社交媒体不只是用来产生和传播的内容，也为用户互相通信交流提供了便利。社交媒体中的用户产生内容（UGC，User-Generated Content）数据具有以下特点[2]：

- **数量巨大（big）：**社交媒体中每个用户产生的数据可能是小的，但是数量巨大的用户群体以及社会化特性将用户的数据链接在一起产生了一种新形式的大数据。比如平均每天有超过 300 万条的微博（tweets）发布到 Twitter，每分钟有超过 3000 张照片上传到 Flickr，每年有超过 160 多万的博客（blogs）发表。

- **广泛链接（linked）：**社交媒体的社会化特性使得社交媒体的数据天生就是广泛链接的。比如用户产生的内容往往由于用户之间的各种社交关系链接在一起。这种链接的数据显然不是独立同分布的（IID，independent and identically distributed），与传统的数据挖掘和机器学习提出的基本假设相违背[3, 4]。

- **充满噪声（noisy）：**社交媒体中普通用户作为内容消费者和产生者常常使得社交媒体的数据质量参差不齐，充满噪声[?]。并且不仅于此，社交媒体中的网络结构也是充满噪声的，一是存在这一些传播虚假和无用信息的用户[5]，二是建立关系的方便性使得各种社会关系混杂在一起，比如好朋友和一般认识人[6]。

- **非结构化（Unstructured）：**社交媒体中用户产生的数据一般是高度非结构化的。尤其是随着移动互联方式的普及与来越多的用户使用移动设备更新 Facebook 的状态，发送微博，或者回复别人的帖子，这不但导致了文本内容短小，而且错误拼写频繁出现[7]。还有一些非自然语言的广泛使用，比如表情符（:)，:() 和缩写（h r u?）等[8]。

- **不完整性（Incomplete）：**为了用户的隐私保护，社交媒体平台一般允许用户将其一些个人数据进行隐藏不被他人看到，这些数据包括个人信息，状态更新，朋友列表，发布的视频和照片以及与他人的信息交流等。比如 Facebook 仅有很少部分用户（小于 1%）公开了他们的个人数据[9]。因此社交媒体的数据是极度不完整和稀疏的。

社交媒体的迅速普及与壮大，使得它在政治、经济、教育、社会等多方面发挥着越来越重要的作用。就如社会会随着网络而演化，无论对个人还是商业组织，社交媒体数据也逐渐变得越来越重要。Web 2.0 时代的到来，更是由于广泛的用户参与而使得用户产生内容（user-generated content (UGC)）成指数级的爆炸式增长，并且更庞大和复杂。目前常见的社交媒体数据应用有：一是基于用户个人信息、行为、位置、微博等数据而进行的个性化推荐、交叉推荐、品牌监测等营销类大数据应用，被互联网广告、电子商务、微博、视频、相亲等公司普遍采用。第二，公共服务类大数据应用，即不以盈利为目的、侧重于为社会公众提供服务的大数据

应用。典型案例如谷歌开发的流感、登革热等流行病预测应用能够比官方机构提前一周发现疫情爆发状况。国内也有搜索引擎公司提供诸如春运客流分析、失踪儿童搜寻的公益大数据服务。三是积极借助外部数据，主要是互联网数据，来实现相关应用。例如，金融机构通过收集互联网用户的微博数据、社交数据、历史交易数据来评估用户的信用等级；证券分析机构通过整合新闻、股票论坛、公司公告、行业研究报告、交易数据、行情数据、报单数据等，试图分析和挖掘各种事件和因素对股市和股票价格走向的影响；监管机构将社交数据、网络新闻数据、网页数据等与监管机构的数据库对接，通过比对结果进行风险提示，提醒监管机构及时采取行动；零售企业通过互联网用户数据分析商品销售趋势、用户偏好等等。

一些服务商拥有海量用户数据的大型互联网企业以自有社交媒体大数据资源为支撑，以 SaaS 形式为用户提供服务。典型的服务如谷歌和 Facebook 的的自助式广告下单服务系统、Twitter 基于实时搜索数据的产品满意度分析等。国内百度推出的大数据营销服务“司南”就属于此类。同时，政府也是社交媒体数据的积极使用者，2013 年曝光的棱镜门事件显示出美国国家安全部门在使用社交媒体数据应用的强大实力，其应用范围之广、水平之高、规模之大都远远超过人们的想象。2012-2013 年间美国国家安全局（NSA）、联邦调查局（FBI）及中央情报局（CIA）等联邦政府机构大量使用 Facebook，Google，Twitter 等公司的数据对全球进行监控引起全世界的关注。白宫 2014 年 5 月发布的《大数据：抓住机遇，守护价值》报告中重点提及了社交媒体的影响[4]。社交媒体大行其道的今天，自然也会成为品牌营销的手段之一。今年世界杯的主要赞助商之一可口可乐就首次尝试 iBeacon 在世界杯营销中的运用。为此，可口可乐挑选了粉丝在 Facebook 和 Twitter 上分享的照片，印制在足球场大小的旗帜上，并将在开幕式上展示。百威在圣保罗开设了社交媒体工作室。该工作室将从不同国家挑选“影响力人物”制作视频，并发布至网上。

1.1.2 主观性信息

社会化网络的出现（Social Web），为人们提供了新的内容共享服务，使百万计连接到全球资讯网（World Wide Web）的人能够在时间和代价上更高效的方式创建和共享自己的内容，思想和观点。如此巨大的信息量，主要是非结构化的（因为它是专为人类阅读消费产生的），因此不能直接使用机器处理的。文本的自动分析需要由机器对自然语言进行深入理解成，实际上我们距离这个目标还很远[10]。到目前为止，网上信息检索，汇总和处理都依据主要是依靠文本的文字表示方式。

[4]来　源: http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

这些算法非常擅长于对文本进行检索，将其拆分，检查拼写和计算词语。但是，当涉及到解释的句子，提取有用的信息，他们的能力是非常有限的。这些基于词表示的算法的很大局限就是他们只能处理字面上的信息，而对于人类来讲，我们就不会收到这样的限制，因为我们看到的每一个字激活的语义相关概念，相关的情节和感官体验的级联，所有这些都使得我们可以以快速高效方式完成一些复杂任务（如词义消歧，文字蕴涵和语义角色标注)。计算模型试图通过模仿人类大脑处理自然语言的方式来弥补这样的认知差距，比如利用在未明确表示的文本语义特征。这些计算模型是有用既为科学目的（如探索语言交流的性质)，以及用于实际用途（如能够有效地进行人机交流)。计算模型可以提供关于可再由心里语言学家（psycholinguist）进行探索的人类行为非常具体的预测。通过继续在这个过程中，我们可能最终会获得人类怎样进行语言处理深刻的理解。为了实现这样的梦想，需要具有前瞻性的思维心理语言学家，神经科学家，人类学家，哲学家，和计算机科学家的共同努力。

近年来，情感分析（或者观点挖掘，sentiment analysis，opinion mining）研究逐渐发展成为介于自然语言处理（Natural Language Processing (NLP)）与自然语言理解（natural language understanding(NLU)）之间的一个独立领域。不像其他的自然语言处理任务（文摘或文本分类)，观点挖掘主要处理与自然语言概念相关的语义信息和情感信息的推理而不需要对给定文本的深度分析[11]。

由于社交媒体的迅速发展与参与用户的数目巨大，因此社交媒体中可挖掘的信息十分丰富，应用前景也十分广泛。但是信息量的庞大规模使得手工分析其中的内容变得十分困难，因此本文从信息自动化处理的角度对社交媒体的信息进行处理与挖掘，试图为社交媒体的相关应用提供帮助。

当然社交媒体与传统媒体存在显著的差异，其自身有不同的特点，如何分析其特点为自动化信息处理服务是我们关注的问题。另外，社交媒体的平台多种多样，基于平台使用的广泛性和代表性，再加上数据的公开性，我们选择 Twitter 作为我们的研究对象[5]，对其进行深入的分析与研究。

1.1.3 Twitter

社交媒体中最具代表性的应用莫过于微博，它是一种用户可以通过及时通讯、手机、电子邮件、网页等方式发布个人状态的消息交流平台。其中 Twitter 是最具代表性的微博工具，自 2006 年 3 月 Twitter 诞生之日起，截至 2012 年 3 月，

[5]本文作者十分感谢在英国爱丁堡大学留学期间，Miles Osborne 教授提供相关 Twitter 数据进行本课题研究。

Twitter 共有 5 亿注册用户，这些用户每天会发布约 3.4 亿条 tweet（推文）[6]，并且 Twitter 每天要接收 16 亿次检索查询[7]。

有许多用户发布的 Twitter 消息创造了历史。一个早期的案例是 2008 年 4 月一位美国学生通过 Twitter 通知好友，他在埃及马哈拉被反政府抗议者扣押，后来这条消息迅速转发，随着巨大的舆论压力，他在第二天就被释放[8]。2009 年 1 月全美航空的飞机迫降哈德逊河，目击者拍摄了一张乘客们在机翼上等待救援的照片，随后该条消息被迅速转发，从此确立了 Twitter 实时新闻的传播工具地位[9]。2009 年 5 月 12 日航天员蒂莫西·克里莫（Timothy Creamer）第一次从太空发布了 tweet[10]。

另外，还有一些与 Twitter 的有关事件确立了 Twitter 在人类生活中的重要地位。例如，2010 年 4 月美国国会图书馆宣布收录所有的公共 Twitter 消息[11]。2012 年 12 月教皇本笃十六世开通 Twitter 账号“@Pontifex”[12]，成为首位使用 Twitter 的教皇，他的继任者方济各（Pope Francis）继续维护着这个账号。2013 年 1 月贾斯汀·比伯（Justin Bieber）超越 Lady Gaga，成为 Twitter 粉丝最多的用户，目前的粉丝数超过 4400 万[13]。

以上种种事件都说明 Twitter 作为社交媒体的代表具有巨大的影响力，再加上数据的公开性，因此本文以 Twitter 作为研究对象，进行详细的社交媒体分析与研究。当然 Twitter 作为社交媒体的代表也有它自身的属性，下面我们从 Twitter 的主要功能进行详细的介绍。

图 1.2给出了本文作者 Twitter 主页的示意图。我们可以看到主页中有许多不同的人发布的 tweet（消息），tweet 的话题可以从日常生活，新闻，到其他任何感兴趣的事情，每个 tweet 限制在 140 个字符以内。140 个字符的原因是由于 Twitter 在最开始设计的时候是基于短信平台，后来随着不断的发展，许多其他的客户端开始投入应用，包括网页和桌面客户端等，但是这个 tweet 的长度限制还是保留了下来，并且还被重新作为一个功能叙述。这种文本长度的限制反而促进了 Twitter 用户编辑的简便和交流的方便，Twitter 的创意总监 Zinko 曾认为“creativity comes from constraint”就是一个很好的佐证[12]。

[6]http://techcrunch.com/2012/07/30/analyst-twitter-passed-500m\
-users-in-june-2012-140m-of-them-in-us-jakarta-biggest-tweeting-city/
[7]https://blog.twitter.com/2011/engineering-behind-twitter'
s-new-search-exper
[8]https://twitter.com/jamesbuck/status/786571964
[9]http://twitpic.com/135xa
[10]https://twitter.com/Astro_TJ/status/8062317551
[11]http://www.loc.gov/today/pr/2010/10-081.html
[12]https://twitter.com/Pontifex
[13]https://twitter.com/justinbieber

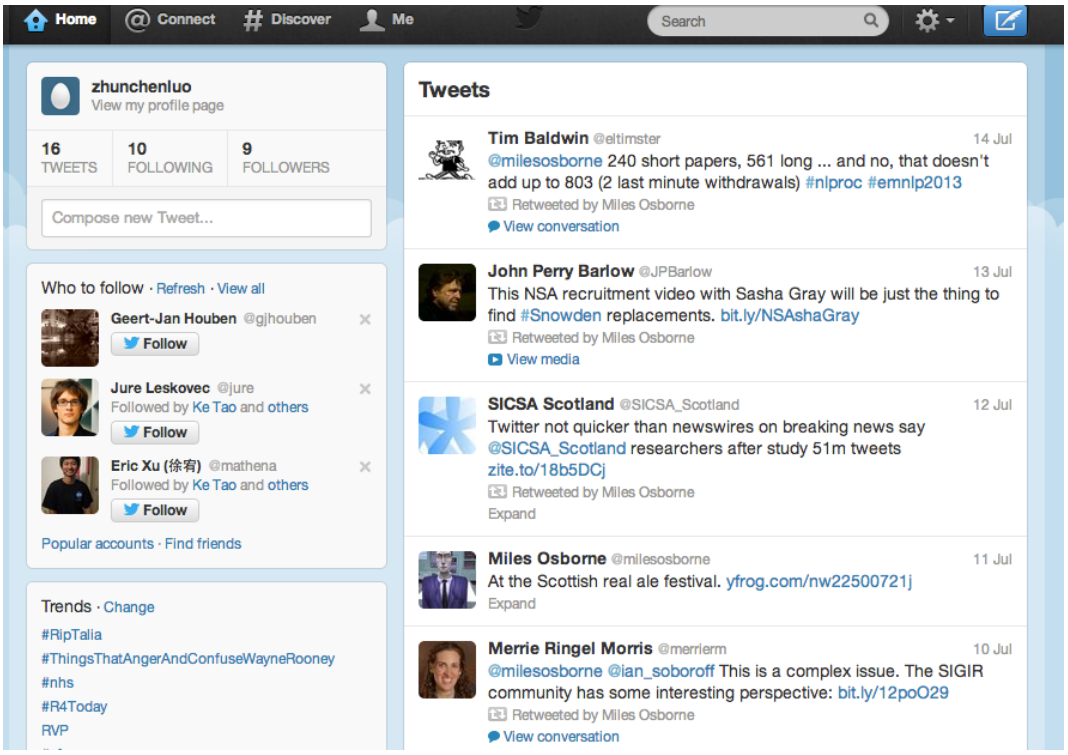

图 1.2　Twitter 示意图

随着越来越多的用户使用 Twitter，且其中许多人都投入了巨大的热情和精力，使得 Twitter 成为了一个活跃的网络社区，有着自己一套独特的交友方式和评价体系。Twitter 结合了已有的社交网络和博客的元素[13, 14]，但也存在着明显的区别：

1. 类似于社交网络，Twitter 用户也通过一定的社会关系或兴趣爱好建立联系形成社交网络，但是这种用户之间的联系不是双向的而是单向的。用户可以关注其他用户（“follow”），订阅对方发布的 tweet，而对方并不需要关注该用户。

2. 类似于博客，Twitter 用户的主页中按时间顺序（默认设置是由近至远）展示他所发布的 tweet，但并不允许其他人直接对每个 tweet 进行评论[14]。

3. Twitter 用户的隐私设置一般是公开的，但用户也可以设置成保护状态，使得其他用户无法看到。

Twitter 最主要的功能就是当用户注册以后，一个其关注对象发布的 tweet 流将按时间倒序展示出来（见图 1.2）。每个用户都有自己不同的原则关注其他用户（见图 1.2中“FOLLOWING”），例如有些用户关注了成千上万的用户，有些用户

[14]这个明显区别于新浪微博的评论功能。

只关注了几个个人熟悉的用户，有些关注名人，有些关注了日常生活不认识但发布内容感兴趣的用户。

至于用户在 Twitter 中可以交流的工具，不仅有 Twitter 网站本身，还有许多第三方应用，包括手机应用，桌面客服端等，这些不仅帮助用户发布 tweet 和交流，有时还提供其他一些功能，如帮助用户发现热门话题，提供感兴趣的其他用户进行关注等。这些应用的大量设计与开发主要归功于 Twitter 开放自身的 API[15]，使得研发人员可以利用 Twitter 数据，设计相关应用。

由于 tweet 140 个字符的限制，使得 Twitter 用户渐渐采用了一些独特方式表示一些功能。例如当用户发布 tweet 想涉及某人时，他们将“@”符号与其涉及对象的用户名结合在一起，如见图 1.2中的“@milesosborne”。这是一种特殊的约定，起源于以前的即时聊天系统，这个方式目前在 Twitter 中有两个功能：

1. 直接发消息到某人作为回复（reply），Honeycutt 和 Herring 将其成为“地址性（addressivity）”[16]。

2. 在发布 tweet 时涉及某人，例如，tweet：*I saw @oprah's show today*，以引起对方的注意。

Twitter 中另外一个重要的约定方式就是“标签（hashtag）”，主要表示 tweet 的话题，它由符号“#”与关键词组合而成。这种方式类似与网页中“社会标签(tag)”，方便标注主题以此进行分类或检索[17]。另外，hashtag 可能起源于以前程序员之间使用特定的词汇与标点符合，例如“$”和“*”表示各种变量与指针或使用“#”符合识别 HTML 锚点的历史。

最后 Twitter 中最特别的一个功能就是转发（Retweet），即重新发布其他人发布过的 tweet。它不同于“@username”和 hashtag 功能表示单一，转发的表示有许多不同的形式，最常用的形式就是拷贝其他人的 tweet，然后在这个 tweet 的前面添加“RT”和来源地址（@username），例如：

A: *Hello world!*

B: *RT @A: Hello world!*

B 就是转发来自 A 的 tweet。当然实际的转发方式可能远比上面的例子复杂。目前除了 Twitter 提供了一个自动的转发按钮便于用户转发，其他的人工添加符号表示转发的方式还有很多，例如，“RT”、“rt”，“via@username”等等。由于 tweet 140 个字符的限制，用户转发时可能只截取原始 tweet 部分内容，还有些用户在转发的 tweet 的文本前面加上一段自己的评论，甚至有些用户并不是完全复制原始 tweet

的内容，有些进行重新编辑然后再发布。以上种种使得 tweet 在转发时可能在内容上有一定的变化，转发方式的不一致都使得转发的识别变得困难，影响了 Twitter 中信息传播研究的信息跟踪。

但是 Twitter 的转发功能也为相关研究提供了机遇，如 tweet 被转发的次数可以反映 tweet 所涉及话题的热度，另外，转发的内容一般都是高质量的消息，最后转发还反映了转发者对被转发者发布消息的肯定态度。

毫无疑问，Twitter 使用的广泛性，以及用户通过这个平台发布大量的消息，使得从 Twitter 中获取相关信息变得十分有意义。但是数据量的巨大也造成手工方式的获取变得困难，因此本文通过自动化的方式从海量的 Twitter 数据中获取相关信息为其它 Twitter 应用服务。另外，Twitter 是一定的关系建立的社交网络，通过“@username”进行交流，利用 hashtag 标注 tweet 主题，使用转发传播信息都使得 Twitter 成为一个研究者分析社会热点，理解人类交流方式，解析人物关系网等情况的重要数据。这些功能与方式都使得 Twitter 具有其自身独特的特点，本文将对其特点进行深入的探讨。

1.2 研究问题

随着以 Twitter，Facebook、新浪微博为代表的社交媒体迅速发展，如何帮助人们利用平台更好地交流，并且在这些社交媒体中发现有意义的信息变得越来越重要。而信息检索技术是满足以上需求的重要技术手段。信息检索是在文档集合中，找到与给定话题相关的客观文本或主观文本。它能够帮助人们在海量的社交媒体信息中，快速找到相关内容，帮助有意义的信息发现，以此满足人们的需求，方便人际之间的交流。另外，以 Twitter 为代表的社交媒体的一个显著特点，就是信息的传播性。人们通过转发分享新闻与观点，加速信息的流动、扩大信息传播的范围。另外，Twitter 中已有的研究发现，转发的信息往往意味着高质量的信息[18]，这是基于人们在 Twitter 传播行为上的一个基本假设：当人们认为一个 tweet 非常重要且值得和大家分享此信息时，他们将通过转发传播这个 tweet。因此研究社交媒体中信息传播的内在规律变得十分重要，且可以帮助在 Twitter 中检索高质量的信息。

但是 Twitter 中的信息检索与传播分析任务也存在着挑战。由于 Twitter 客户端使用的多样性，如大量使用移动平台，以及 tweet 文本本身 140 个字符的限制，造成 tweet 文本与其他文本（如新闻）编辑质量与风格的巨大差异。同时移动平台的广泛使用，使得 Twitter 中信息传播速度更快，范围更广。再加上 Twitter 用户参与的低门槛性，使得信息在 Twitter 中的传播不像以往媒体（如报纸）的新闻传播，会对信息的正确性进行层层验证，这就造成了 Twitter 中信息传播的随意性，

使得信息的质量难以保证。因此本文主要从两个科学问题来思考与研究 Twitter，以此帮助 Twitter 中的信息检索与传播分析：

1. **人们在 Twitter 中如何用自然语言描述话题和表达观点？**
2. **以 Twitter 为代表的社交媒体有何新特点？如何利用这些特点帮助获取信息和对信息进行传播分析？**

第一个问题的研究主要是从自然语言处理的角度分析人们在社交媒体上如何组织语言来描述客观话题和表达主观观点。显然 Twitter 参与的低门槛特点使得大量的用户参与其中，由于参与者编辑文本的水平参差不齐，编辑的内容与目的也多种多样，另外，再加上 tweet 本身的字符限制都使得 Twitter 中的文本呈现低质量、噪音大的短文本特点，这给传统的以正式文本（如新闻）为处理对象的自然语言处理技术带来了挑战。因此深入地研究 Twitter 文本的特点，对于解决 Twitter 中的信息检索与传播分析任务变得十分重要，并且分析 Twitter 中文本的特点也能够帮助其他以语言为基础的 Twitter 应用，如 Twitter 中的事件发现、观点挖掘等等。

第二个问题的研究主要是分析 Twitter 作为新型媒体的特点，从中发现一些规律和有价值的信息帮助 Twitter 中信息检索与传播分析问题的解决。显然 Twitter 中无论是 tweet 本身还是 tweet 的用户都呈现了一些新的特点。比如，tweet 中包含 hashtag，以此可以作为 tweet 的内容主题。tweet 中也包含大量的链接，这些链接与 tweet 中描述链接的文本存在什么样的关系？另外，每个 tweet 都有作者，Twitter 的一个显著特点就是用户信息的公开化，包括用户的朋友关系，发布信息的历史，个人的属性信息等，如何发现这些用户信息的普遍规律与内在价值，以此帮助 Twitter 中的信息检索与传播分析任务是本文研究的主要问题，当然社交媒体的新特点研究同样也能帮助 Twitter 中其他任务的解决，如用户推荐等。

1.3 相关研究

无论是 Twitter 中的信息检索还是传播分析，对 tweet 文本的理解都是其中一个重要的环节。但是 tweet 文本的短小与大量的噪音文本（存在着许多缩写词、错别字等等）都造成自然语言处理技术在 Twitter 上的应用存在着新的挑战，我们在本节将介绍自然语言处理技术在 tweet 文本处理上的相关工作。另外，Twitter 上信息检索的研究也离不开以往传统信息检索技术的借鉴与应用，我们主要利用 tweet 的文本特征与社交媒体的特征帮助 Twitter 中的信息检索，而这些特征如何有效地整合到检索模型中是需要考虑的问题，因此本节将详细介绍基于机器学习的信息检索技术。最后本节还将介绍已有的 Twitter 中的传播分析工作，以此为本文具体的传播分析任务的解决提供帮助。

这里要强调的是，本文的相关工作分析主要从整体相关工作和局部相关工作进行阐述。本章的相关研究主要介绍的是整体的相关工作，因为这些研究成果可以为本文所研究的具体任务提供思想借鉴和技术支持。以后各个章节中的相关工作则会具体地分析已有的类似工作，以及研究成果。

我们从整体上介绍了三个相关研究，包括 Twitter 与自然语言处理（见 1.3.1），主要介绍目前已有的自然语言处理技术在 Twitter 中的研究成果，以此帮助本文的关于 tweet 的文本分析；信息检索与机器学习（见 1.3.2），主要介绍目前机器学习技术在信息检索中的应用，以此为解决本文的具体 Twitter 信息检索和传播分析任务提供问题解决框架；Twitter 中的传播分析（见 1.3.3），主要介绍目前已有的关于 Twitter 转发的研究成果，为本文 Twitter 传播分析的具体任务提供借鉴。

1.3.1 Twitter 与自然语言处理

随着以 Twitter 为代表的社交媒体的广泛使用，自然语言处理技术在 Twitter 的文本处理中得到广泛应用，但是研究者发现 Twitter 的文本明显区别于以往的很多文本类型。Eisenstein 将这种文本类型称为*坏语言*（*bad language*）：文本“无视”我们以前期望的词汇、拼写和语法[2]。

研究人员发现最先进的自然语言处理技术在 Twitter 的文本应用中都显著差于其他文本。在自动地词性标注测试中，Stanford tagger 在 Wall Street Journal 语料上的正确率可以达到 97%[19]，而 tweet 的文本处理仅仅只有 85%[20, 21]。利用 CoNLL 数据训练的 Stanford 命名实体识别器，对 CoNLL 测试语料进行实体识别，F1 值可以达到 86%[22]，而在 Twitter 的文本中仅仅只有 44% 的 F1 值[23]。另外，Foster 等人也对语法分析效果进行分析，发现最先进的语法分析器在 Twitter 的文本应用中，正确率下降约 15%[24]。

为了解决自然语言处理技术在 Twitter 中所遇到的挑战，研究人员主要从两个方面进行了相关研究：

1. **文本的正常化（Normalization）**，即把坏语言变成好的语言，以其适合于传统文本的自然语言处理技术。Han 和 Baldwin 开发了一个分类器，能够识别“非正常（ill-formed）”的词，然后利用基于形态音位（morphophonemic）相似的方法将其转换为正确的词[25, 26]，Han 等人还提出了一种构造词典的方法，简单替换词的变形（例如 tomorrow 替换 tmrw），这种方法结合词语的上下文评估各种变换的可能性[27]，但 Liu 等人提出一种没有明确分类的方法，进行词的正常化[28]。另外，Liu 等人提出了一种基于图模型的方法同时解决 tweet 中命名实体识别和 tweet 文本正常化的方法[29] 。Liu 等人设计一个正常化认知驱动系统解决 Twitter 中文本的正常化问题[30]，该系统整合人

们对于“非正常（ill-formed）”词的各种认知角度，包括字符转换、视觉感知、字符串和语音相似等等。最近 Hany 和 Menezes 提出了一种无监督学习的方法对 Twitter 中的文本进行正常化，他们在大量 tweet 文本中构造 n 元词串，以此构造语境相似的二部图，然后利用 Random Walks 算法发现“非正常（ill-formed）”词与正常词的对应关系[31]。以上所有的方法都在一定程度上解决了 Twitter 中文本正常化的问题。

2. **领域化（Domain adaptation）** 与其让 Twitter 的文本适应以前的自然语言处理技术，不如改变这些技术适应 Twitter 文本。一系列的工作从领域化的角度出发进行了相关研究。这些工作包括适合 Twitter 文本的自动词性标注器[20, 21]，自动命名实体识别的方法[23, 29, 32–36]，语法分析器[24]，对话模型[37]，自动摘要[38–46] 等等。这些工作采用各种方法，使其能够很好地适应 Twitter 文本的特点，具体为:

   预处理（preprocessing） 减少词语中某些重复的字符（经常有些词用重复的字符表达感情[47]）, 去掉 hashtag、链接、提交（@username）等等;

   标注新数据（new labeled data） 根据任务在 Twitter 中标注部分数据[23, 32]，以此进行有监督学习;

   自定义标注标准（new annotation schemes） 定义适合 Twitter 的标注标注，如词性标注中对 hashtag、链接、提交（@username）等定制特定的标注类型[20, 21];

   “远端”监督（distant supervision） 通过一定的规则，构造大量粗糙的训练数据帮助 Twitter 的文本机器学习模型训练，然后应用到具体的任务中[23]。

毫无疑问，tweet 文本的特殊性使得传统自然语言技术在 Twitter 上的应用充满挑战，我们将利用以上所涉及到的方法、思想或已有的开发工具，按照信息检索任务和传播分析任务的具体需求，设计对应的 tweet 文本自然语言处理方法，开发有效的文本特征，提高 Twitter 中的信息检索效果与传播分析的准确性。

1.3.2 信息检索与机器学习

Twitter 中的信息检索是本文的一个重要研究任务，由于 tweet 文本的特点和丰富的社交媒体属性使得 Twitter 中的信息检索不同于以往的信息检索任务（如图书馆文档检索)。在 Twitter 检索任务中需要考虑因素很多，如 tweet 用户的信息等。传统的检索模型在构造排序函数的时候往往只需要考虑不多的因素，如查询词在文档的频率、位置等，因此可以手工构造这些函数对文档排序，但是 Twitter

中的检索需要考虑的因素相当多，造成手工构造排序函数变得复杂，但是基于机器学习的排序模型可以通过训练数据自动构造排序函数，因此这里我们详细介绍基于机器学习的信息检索模型。

信息检索与机器学习领域有很多研究的重叠，上世纪 60 年代提出的相关反馈就是一个简单的机器学习算法，它构建一个分类器区分相关文档和非相关文档，以此作为用户关于初始排序中文档重要性的反馈[48]。到了 80、90 年代，研究人员开始使用机器学习方法来基于用户反馈学习排序算法。但是，许多机器学习算法在信息检索上的应用都受到训练数据规模较小的影响，如果系统要对每个查询构造分类器，基本上是不现实的。

但进入 21 世纪，随着网络搜索引擎的出现，从用户交互中积累了海量的查询日志，潜在的训练数据的规模非常庞大。借此基于点击流数据的排序学习算法被提出[49–56]。由于对于每个查询中文档的相关性判断非常稀疏，但是有一定数量文档在网络搜索引擎的检索返回结果中被用户点击浏览，这些行为可以隐性地认为是用户对文档相关性的判定。例如，如果一个用户在一个查询的排序中点击了第三个文档而不是前面两个，那么可以假设第三个文档应该在下次排序中获得较高的排序位置。

在排序学习模型中，最著名的排序函数莫过于基于支持向量机（SVM）的方法，通常被称为 Ranking SVM。它的输入是针对一组查询的偏序排序信息的训练集合：

$$(q_1, r_1), (q_2, r_2), ..., (q_n, r_n)$$

其中 q_i 是一个查询，r_i 是所需排序的文档关于查询的部分排序信息或相关性级别。这意味着如果文档 d_a 应该比 d_b 排序更高，那么 $(d_a, d_b) \in r_i$。这些排序信息可以通过点击流数据获得，然后训练排序模型。相关的研究从排序学习的数据[57–60]、排序学习模型[61–64] 和评估学习效果[65–69] 三个方面展开。

本文将主要采取基于排序学习的机器学习算法，针对 Twitter 信息检索的具体问题，将 tweet 文本分析的结果和 Twitter 社交媒体的新特点转换成特征，整合到机器学习的模型框架中，帮助 Twitter 中各项检索任务的解决。

1.3.3 Twitter 中的传播分析

Twitter 中一个重要的机制就是转发，即重新发布其他人发布过的 tweet。这种简单的机制可以使得作者的全部粉丝看到转发的信息，使得信息迅速、广泛的传播。我们本节将介绍 Twitter 中已有的对于转发行为的研究，以此分析涉及影响转发行为的因素，包括 tweet 的文本内容与转发的关系，用户的属性如何决定其他人

的转发；Twitter 中信息的一般传播路径与规律。这些研究成果可以帮助本文具体的传播分析任务。

boyd 等人[15]研究了 Twitter 中转发的各种类型以及转发的原因，他们分析了不同用户，用户属性，用户交流方式对于转发的影响，同时也分析了人们在 Twitter 中喜欢转发的内容[70]。他们发现 18% 的转发 tweet 包含 hashtag，52% 的转发 tweet 包含链接，11% 的转发 tweet 包含连续的转发符号串（如，“RT @user1 RT @user2 ”），另外，9% 的转发 tweet 都包含回复原 tweet 作者的回复字符串（“@reply”）。这说明 tweet 文本中的 hashtag，链接、回复、提交和转发符号都与 tweet 的转发存在着一定的对应关系。

Yang 和 Counts 通过 Twitter 中的提及（“@username”）抽取了用户之间的关系，并在此基础上构造了用户关系的复杂网络。他们研究了信息在这个复杂网络上是如何传播的，包括信息传播的速度，规模，以及范围[71]。他们发现大约只有 25% 的 tweet 是被信息作者的朋友转发，大部分是被粉丝但非朋友转发。这说明 Twitter 中用户形成的复杂网络，影响着人们的转发行为，因此信息在传播路径上具有一定的规律可循。

Macskassy 和 Michelson 分析了一个月用户的 Twitter 数据，他们解释了各种信息传播的方式，尤其是转发的行为模式，他们发现 tweet 的内容是 tweet 被转发的决定因素，因此他们构建了基于内容的转发模型[72]。

Starbird 等人对具体事件在 Twitter 上的传播进行了深入研究，他们分析了 2011 年埃及的政治事件，演示了这个事件的相关信息在 Twitter 上是如何生成，发展，传播的[73]。

Comarela 等人研究了影响用户回复或转发的因素，他们发现以前是否回复，发布信息的频率，信息的时效性，tweet 的长度决定用户是否回复[74]。

除了以上的工作，最新的研究还从不同角度对 Twitter 中的转发行为进行了深入的研究[75–78]。

综上所述，我们发现影响人们转发行为的因素主要包括 tweet 文本的内容、tweet 文本的社交媒体属性（如，是否包含链接、hashtag、提及等）、tweet 作者的用户属性，tweet 作者的朋友圈子，当然以上的研究都是从宏观上大规模分析 Twitter 转发数据得出的研究结论。从微观的角度则可以考虑给定一个 tweet，未来这个 tweet 是否会被转发，我们将在 4.2.1介绍相关工作。

虽然已有的 Twitter 转发研究从许多不同的角度进行了考虑，但是仍然有许多问题与因素被忽视，例如 tweet 的转发预测针对的一般是普遍 tweet，并未细粒度的划分类型，本文我们将针对特定类型 tweet 进行转发研究。另外，目前的转发大

[15]Danah Boyd 因为家庭的原因一般使用小写拼写姓名，这里并不是拼写错误。

多都是从 tweet 本身进行考虑，并未从受众的角度进行分析，本文将对 tweet，作者、受众三个方面在转发过程中的相互关系进行探讨。

1.4 研究内容与方法

1.4.1 本文研究内容

本文的研究内容主要是围绕在给定话题的情况下，如何在 Twitter 中找到与话题相关的主客观 tweet。主要利用 tweet 文本特点和社交媒体属性帮助 Twitter 中的信息检索。本文中 Twitter 的传播分析主要是从内容和受众两个方面进行考虑，如何在 Twitter 中发现会传播的观点和如何在 Twitter 中发现信息的传播者。同样也通过分析 tweet 文本特点和社交媒体特点帮助这两个问题的解决。参见图1.3本文研究框架。

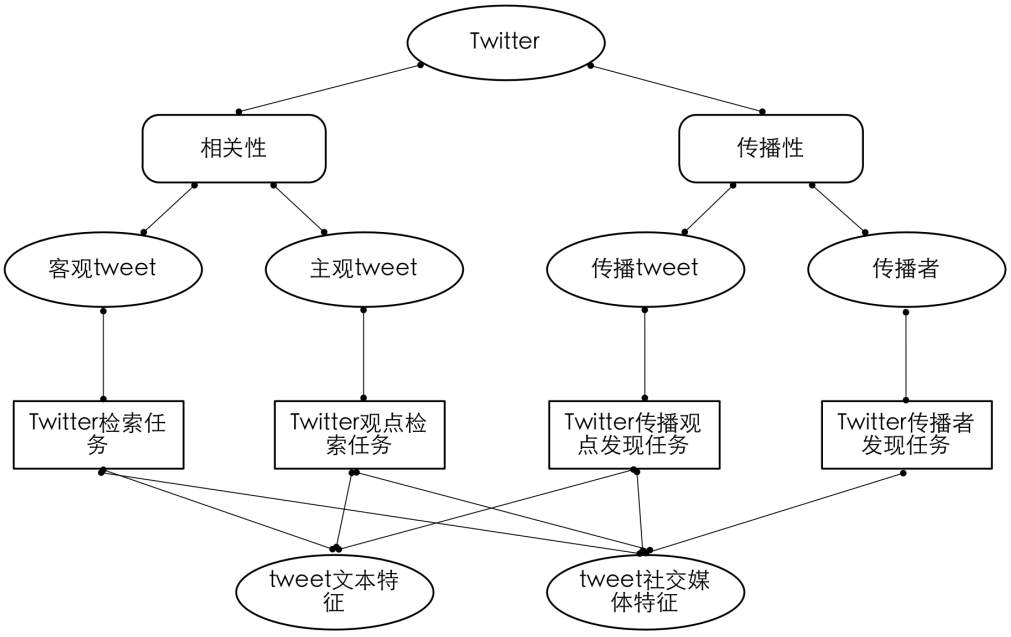

图 1.3 本文研究框架

具体四个研究内容的定义如下:

1. **Twitter 检索**: 给定关键词，在 Twitter 中找到话题相关的 tweet。通过文本的结构化信息（我们发现 tweet 文本呈现结构化特点）和社交媒体信息帮助提高 Twitter 检索，以此分析了人们在 Twitter 中是如何描述话题以及社交媒体特征与相关话题 tweet 的内在联系。

2. **Twitter 观点检索**: 给定关键词, 在 Twitter 中找到话题相关且带观点的 tweet。通过 tweet 观点化信息（基于结构化信息）和社交媒体信息帮助提高 Twitter

观点检索，以此分析了人们在 Twitter 中是如何表达观点以及社交媒体特征与 Twitter 中观点的内在联系。

3. **Twitter 传播观点发现**：给定关键词，在 Twitter 中找到话题相关且带观点的 tweet，并且这个 tweet 在未来会被转发。通过 tweet 观点化信息（基于结构化信息）和社交媒体信息帮助发现 Twitter 中传播观点，以此分析了人们在 Twitter 中是如何表达高质量的观点以及社交媒体特征与 Twitter 中传播观点的内在联系。

4. **Twitter 传播观者发现**：给定 tweet，发现 tweet 的粉丝中，谁会在未来传播这个 tweet。通过社交媒体信息帮助发现 Twitter 中的信息传播者，以此分析了社交媒体特征与 Twitter 信息传播者的内在联系。

总的来说，针对 1.2的第一个问题，本文通过分析 tweet 文本的结构化信息和观点表达方式，找出规律与特点，将其利用到 Twitter 检索、观点检索、传播观点发现任务。针对 1.2的第二个问题，本文通过分析社交媒体的用户属性、社会网络结构、文本属性，发现用户之间、用户与 tweet 文本（主客观信息)、用户与传播行为、tweet 文本与传播行为之间的内在联系，通过开发社交媒体特征，帮助解决 Twitter 检索、观点检索、传播观点发现和信息传播者发现任务。

1.4.2 本文研究方法

针对以上研究内容，本文基于自然语言处理技术和机器学习技术，深入分析 Twitter 中 tweet 的文本特点和社交媒体属性，解决 Twitter 中若干检索与传播分析问题。我们希望通过对 Twitter 中若干检索与传播分析问题的研究达到如下几个主要目标:

1. 认识以 Twitter 为代表的社交媒体的新特点，包括文本表现形式，用户属性，Twitter 中信息的传播行为等等。

2. 传统的信息检索技术如何在新型的社交媒体中使用，重点研究基于机器学习的信息检索技术在 Twitter 中的应用。

3. 深入研究 Twitter 中观点检索问题，寻找人们在 Twitter 中表达观点的方式，以及其它相关因素。

4. 针对 Twitter 中 tweet 文本质量较低，以及质量评价问题，帮助人们进一步理解 Twitter 中高质量文本的评价问题。

5. 通过研究特定用户查询问题，找到 Twitter 中 tweet、作者和粉丝之间的关系，帮助 Twitter 中传播分析的研究。

根据各个具体的研究内容，我们的具体研究方法为:

1. 针对 Twitter 中的信息检索问题，我们深入分析 tweet 中的文本特点，找到文本特定结构与社会属性之间的关系，开发文本结构特征，然后结合 tweet 的社交媒体特征（用户属性等)，将其整合到机器学习的框架中，通过排序学习，提高 Twitter 中信息检索的效果。

2. 针对 Twitter 中的观点检索问题，首先对 Twitter 中的观点检索问题进行定义，构造测试数据集，然后分析 Twitter 中用户表达观点的文本特点以及 Twitter 中观点所对应的潜在用户属性，开发特征，利用排序学习，解决 Twitter 中如何找到观点的问题。

3. 针对 Twitter 观点检索中大量返回低质量观点的问题，从发现传播观点的角度提出了 Twitter 中高质量观点的客观评价指标，这个指标利用 Twitter 中高质量信息大量传播的特点，分析了 Twitter 传播观点发现的问题，利用 tweet 中如何判定是否转发的方法，tweet 中文本是否包含观点以及 tweet 文本本身的语言质量帮助相关任务的解决。

4. 针对 Twitter 中信息传播者发现的问题，我们首先进行问题定义，构造数据集，分析信息传播者的特点，找到信息传播者与转发 tweet、转发 tweet 作者之间的联系，设计相关特征，将其利用到机器学习框架中，解决信息传播者发现的问题。

1.5 本文主要贡献

本文主要围绕分析 Twitter 文本的特点与 Twitter 社交媒体属性展开，通过 Twitter 中的信息检索和传播分析任务，发现哪些因素能够帮助或影响检索效果的提高与传播分析的准确性。

在 tweet 文本分析方面，我们发现，虽然 tweet 是短文本，但是它具有结构化的特点。不同的 tweet 文本结构对应不同的属性和文本质量，通过挖掘 tweet 的文本结构信息能够帮助 Twitter 的信息检索。另外由于某些特定结构的 tweet 具有某种属性（如主观化)，因此可以利用结构化的 tweet 收集大量的相关文本，构造情感词典帮助 tweet 主观化判定，提高观点检索的效果。

在 Twitter 社交媒体属性的分析方面，我们通过对 tweet 中是否包含链接、hashtag、提及等的研究，确定这些符号串或功能与 Twitter 信息检索的对应关系，

因此帮助该任务的解决。我们还分析了 tweet 作者的属性，包括作者的粉丝数目、朋友数目、分组数目、兴趣爱好、圈子、活跃时间等等，试图发现这些因素与 Twitter 信息传播之间的内在联系。

在具体的 Twitter 信息检索任务中，我们从给定关键词找到主客观相关 tweet 的两个方面进行研究。获取客观 tweet 方面，我们开发了 tweet 文本的结构化特征和社交媒体特征，将其整合到基于排序学习的模型中，实验结果验证了我们的方法是有效的。获取主观 tweet 即 Twitter 中的观点检索是一个全新的工作，我们定义了这个问题，发布了关于研究这个问题的实验数据[16]，截止到 2013 年 9 月这个实验数据已经有 40 多个国家和地区超过 200 多个研究单位和个人下载，并且这个数据还被 ICWSM 会议推荐为官方的社交媒体研究数据[17]，针对 Twitter 中的观点检索我们也提出了我们的方法，主要是开发 tweet 的文本特征与社交媒体特征结合排序学习框架进行解决，同时我们也提出了一种基于无监督学习的 tweet 观点化评价的方法，目前有许多其它研究单位的工作围绕我们的 Twitter 观点检索工作展开[79–82]。

在具体的 Twitter 传播分析任务中，我们从传播的内容和受众两个方面进行考虑，提出了 Twitter 中传播观点发现的新任务和信息传播者发现的新任务。Twitter 中传播观点发现可以帮助我们解决 tweet 质量评价主观化的问题，由于以往的研究主要是围绕给定 tweet，预测该 tweet 在未来是否会被转发，我们对这个问题进行了细粒度的研究，从观点化的 tweet 能否被转发的角度进行了探讨，通过开发 tweet 的文本特征和社交媒体特征解决传播观点发现的问题。Twitter 信息传播者发现的问题，针对的是以往研究忽视“谁”会转发的问题，我们定义了这个问题，提出了解决这个问题的方法，发现兴趣与转发的历史信息是决定信息传播者的重要因素，同样我们也公布了研究这个问题的数据集[18]，供以后科研人员继续使用。

以上所有的工作都通过论文的形式公开发表[83–86]。

1.6 本文结构

本文的研究工作主要围绕社交媒体中检索与传播分析任务展开，我们可以将这两方面的工作分为以下几个主要部分：在 Twitter 检索方面，我们首先分析了 tweet 的文本信息与社交媒体信息，以此帮助 Twitter 中传统的信息检索任务；然后以此为基础，进一步探讨 Twitter 中观点检索问题，给定关键词，检索到话题相关且带观点的 tweet；在 Twitter 的观点检索任务中，我们发现检索结果存在大量的低质量观点，结合 Twitter 中的传播分析，我们从传播的内容角度考虑，转发

[16]下载地址：http://sourceforge.net/p/ortwitter/wiki/Home/

[17]http://www.icwsm.org/2013/datasets/datasets/

[18]下载地址：https://sourceforge.net/projects/retweeter/

的 tweet 一般是高质量的信息，因此我们再进一步研究了在 Twitter 中如何发现传播观点的问题；Twitter 信息的传播分析不仅可以从 tweet 的本身进行研究，也可以从受众的角度进行分析，因此最后我们讨论了在 Twiter 中如何需找信息传播者的问题。上述工作共分为六个章节, 论文主体结构以及章节之间的关系如图 1.4所示。，每个章节内容具体安排如下:

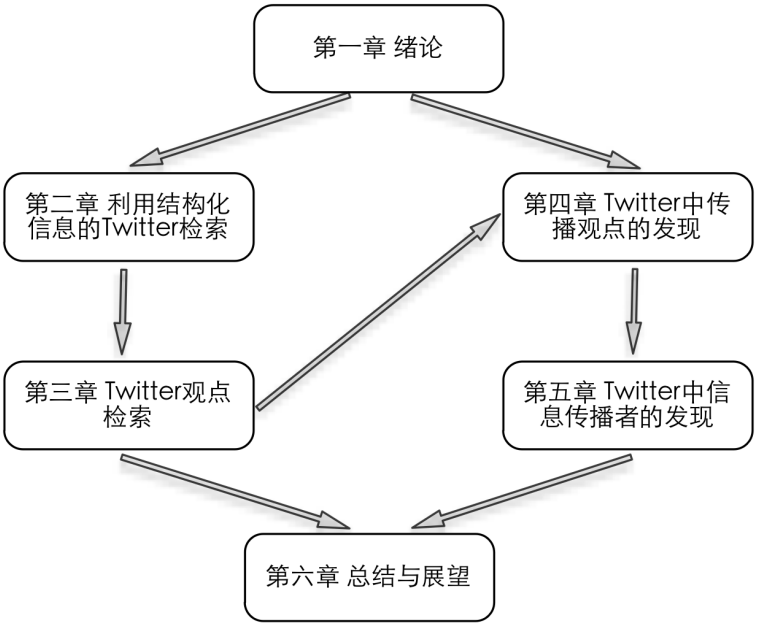

图 1.4　论文整体结构图

第一章是绪论，首先介绍了本文研究的背景，介绍了社交媒体和 Twitter 的一些基础知识，接着提出研究动机，阐明了本文所涉及的科学问题、研究内容，并给出了研究方法，然后分析了研究问题，确立了依托自然语言处理技术与机器学习方法解决这些问题的基本思路，最后介绍了本文的主要工作和文章的结构。

第二章是 Twitter 中的信息检索，首先介绍了 Twitter 信息检索的研究背景，然后提出了以往 Twitter 信息检索方法忽视 tweet 文本结构化特点以及存在大量社交媒体信息的问题，以此设计了一种标注 tweet 文本结构的自动标注器，最后利用自动标注器标注 tweet 文本开发结构特征，结合社交媒体特征帮助 Twitter 信息检索，实验验证了方法的有效性。这一章回答了 Twitter 中人们是如何用自然语言描述客观话题，并且社交媒体特征与 tweet 话题相关性存在怎样的联系。

第三章是 Twitter 中的观点检索，本章开头定义了 Twitter 中的观点检索问题，分析了 Twitter 观点检索与以往观点检索的不同特点，接着提出了一种自动获取主观 tweet 与客观 tweet 的方法自动生成情感词典，依靠词典对 tweet 的文本进行主观化判定，结合 tweet 的用户属性信息和文本信息，利用排序学习算法，实现观点

检索。实验部分，我们构造了自己的 Twitter 观点检索语料，并发布了语料供以后的研究者使用，实验结果证明了我们的观点检索方法有效。这一章回答了 Twitter 中人们是如何用自然语言表达主观观点，并且社交媒体特征与 tweet 观点相关性存在怎样的联系。

第四章中我们针对 Twitter 观点检索存在大量低质量观点的问题，依据转发 tweet 一般是高质量文本的既有研究成果，提出了在 Twitter 中发现传播观点的问题。我们首先定义了问题，然后构造了数据集，接着开发了 tweet 传播度特征、观点化特征和文本质量特征，将其整合到排序学习的机器学习模型框架中。实验结果说明了这些特征对于 Twitter 中发现传播观点是有帮助的，另外我们的方法可以达到人判定传播观点的效果。这一章回答了 Twitter 中人们是如何用自然语言表达传播性的观点，并且社交媒体特征与传播性的观点存在怎样的联系。

第五章中我们探讨了 Twitter 中发现信息传播者的问题，给定一个 tweet，发现 tweet 的作者粉丝中谁会转发该消息。我们开发了转发历史特征、用户特征、用户活跃时间特征和用户兴趣特征，并依然将其应用到排序学习的框架中构造模型进行排序。实验部分，我们构造了自己的测试数据与基准系统，我们公布了数据，实验结果显示了我们的方法能够成功找到 tweet 转发者。这一章回答了社交媒体特征与信息传播者存在怎样的联系。

最后一章是总结部分，我们阐明了本文工作的贡献点，并且指出了工作的一些不足，并对未来社交媒体检索与传播分析的一些问题和方法进行了尝试性地思考。

第二章　利用结构化信息的 Twitter 检索

2.1　引言

Twitter 中的信息检索任务主要是给出关键词，然后在大量的 tweet 数据中找到与关键词话题相关的 tweet。图2.1是 Twitter 官方提供的检索主页，用户输入任何关键词，Twitter 搜索引擎返回如图2.2的检索结果（Twitter 中检索“sigir13”）。

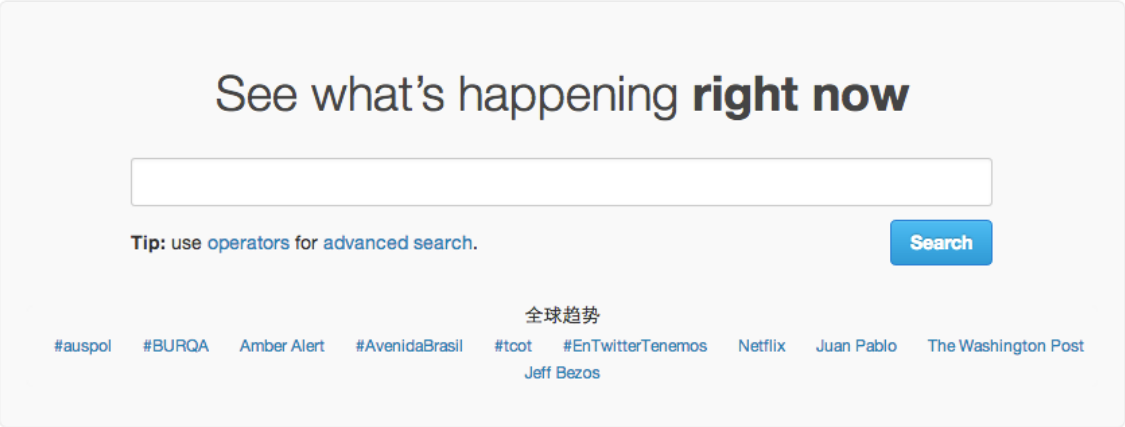

图 2.1　Twitter 检索主页

目前传统的信息检索在 Twitter 中的应用主要还是度量关键词与 tweet 的话题相关性进行排序，另外辅助一些时间信息，如图2.2的返回结果就是在 tweet 与关键词在考虑话题相关程度的同时，按时间由近及远进行排序。但是 Twitter 是典型的社交媒体，每一个 tweet 都包含了丰富的社交媒体信息（如 tweet 作者的属性信息），这些信息并未得到足够的利用与开发，帮助 Twitter 中的信息检索。另外，传统 Twitter 中的信息检索都认为 tweet 的文本是短小非正式的噪音文本，不利于传统信息检索技术在 Twitter 中的应用。因此本章将研究如何利用 tweet 的文本信息和社交媒体信息帮助传统的 Twitter 信息检索。

现有 Twitter 上的信息检索方法简单的将每个 tweet 看成一个平面文本 [87–90]，以前的工作表明，网页和普通的文本可以基于内容或结构划分成不重叠的小块，这些块和它们的组合信息可以被用来改进信息检索的效果[91–93]。虽然 tweet 是一个短文本，但是它也可以被看成是一个由多个块组合而成的文本。

图 2.3给出了一些由 Yao Ming、BBC News 和 Lady Gaga 发布的 tweet 样本[1]。我们可以看到三者发布的 tweet 在文本结构上存在明显的差异：

[1]姚明是一位在 NBA 退役的中国职业篮球运动员, BBC News 是英国广播公司的 Twitter 账户，Lady Gaga 的是美国流行歌手。

图 2.2　Twitter 中检索 sigir13 返回结果

1. Yao Ming 发布的 tweet 仅仅是平面文本。

2. BBC News 发布的 tweet 都以链接结尾。

3. Lady Gaga 发布的 tweet 中包含大量的 hashtag，链接和“@”符号。

图 2.3　Yao Ming、BBC News 和 Lady Gaga 的 tweet 样本示意图

虽然平面文本、链接、hashtag 和“@”符号的长度有差异，但我们都将其视为组成 tweet 的小块。利用这些小块，本章中我们将介绍一种从 tweet 文本中获取结构信息的方法，并以此来提高 Twitter 信息检索的效果。这种方法的动机主要是基于一个词在不同块中的出现往往存在着不同的排序权重。因为每个块都有自己本身的话题、功能、长度、位置、文本质量和语境等等。另外，tweet 所对应块的序列组合结构也反映了 tweet 的话语转换信息以及质量信息。

我们将这种 tweet 中的块称为“Twitter 积木（Twitter Building Blocks 或 TBB）”，块的序列组合结构（TBB structures）则反映了 tweet 的结构信息。这些结构信息可以用来对 tweet 进行聚类，并且每个类别都有自身的属性。例如，与 BBC News 发布 tweet 结构相同的 tweet（见图 2.3）很有可能是新闻信息。另外，这种结构也与 tweet 文本的质量相关。因此，我们将开发这些结构信息并将其用到基于排序学习模型的 Twitter 信息检索方法中。

本章中，为了提高 Twitter 信息检索的效果，我们将设计一系列基于 Twitter 积木及其结构的特征。这些特征的主要优点在于它们不仅反映了 tweet 的结构信息而且信息的获取仅仅来自于 tweet 文本本身，并不依赖于其他社交媒体特征。我们将基于结构化信息的 Twitter 排序学习模型与目前最好的 Twitter 排序学习模型[88]进行了比较，实验结果表明仅仅只用我们设计的结构化特征开发的 Twitter 排序学习模型在 tweet 排序上可以达到最好的 Twitter 排序学习模型效果，而将社交媒体相关特征与结构化特征结合使用，tweet 排序效果更好。

本章的主要工作如下：

1. 我们提出了 Twitter 积木的概念，Twitter 积木是一种词串，它反映了各种交流信息，而 Twitter 积木各种序列组合则反映了 tweet 本身的话语转换。

2. 我们设计了基于结构化信息的 Twitter 排序模型，并验证了该模型在对 tweet 排序上达到最好的 Twitter 排序模型效果，并且这种模型并未利用其他社交媒体特征。

3. 我们将社交媒体特征与结构化特征相结合使用，获得了更好的 Twitter 信息检索效果。

2.2 相关工作

我们从两个方面讨论相关工作：基于结构化的信息检索和 Twitter 信息检索。

2.2.1 基于结构化的信息检索

基于结构化的信息检索主要是利用结构化信息和文本的内容信息提高信息检索效果。这类方法的动机是基于同一个词在不同的文本块中拥有不同的排序权重，另外，文本块的特定组合结构也反映了特定的信息，这些信息可以被用于提高信息检索效果。Ahnizer 等人利用手工设定块的权重来提高文本检索质量[92]，它们利用不同块的不同权重综合计算不同词在排序当中的权重，以此改善排序效果。它们实验验证了这种方法有利于提高结构化文档的检索，特别是数据更新快的结构文档，例如，数字图书、论坛、新闻网站等。Fernandes 和 Moura 等人提出了一种自动设定块权重的方法并以此提高网页检索的效果[93, 94]。Cai 等人也提出了一种利用切分网页获取结构化信息提高网页检索效果的方法[95]。以上方法都是基于同一词在不同块中存在不同的排序权重的思想。

2.2.2 Twitter 信息检索

O'Connor 等人开发了 TweetMotif 系统，主要用于 Twitter 中的话题发现及其摘要[96]。Efron 提出了一种基于语言模型 hashtag 检索方法[87]，它利用检索到的 hashtag 对感兴趣的话题进行查询扩展，以此提高 Twitter 检索效果。Massoudi 等人提出了一种基于 tweet 文本质量的检索模型并取得了不错的检索效果[89]。Naveed 等人提出了一种 Twitter 检索方法，解决了 tweet 文本长度归一化与数据稀疏的问题[90]。但是以上的所有方法都并没有将 tweet 的结构化信息引入 Twitter 检索中。Duan 等人提出了基于排序学习的 Twitter 检索方法[88]，该方法不仅考虑了 tweet 内容的话题相关性，还考虑了 tweet 作者的权威性和 tweet 的其他特征。我们将此方法作为我们其中的一个基准系统进行比较，但是他们的方法依然没有考虑 tweet 文本的结构化信息。除了以上工作，最新的 Twitter 信息检索都没有考虑 tweet 文本的结构化信息[97–105]。

绝大部分的 Twitter 检索系统在构造检索模型时一般都认为 tweet 是一个平面文本，但是用户在编辑 tweet 时的一些习惯使得 tweet 文本呈现结构化的特点，以往通过开发文本的结构信息能够帮助结构化文本的检索（例如，网页检索），因此我们希望利用 tweet 的文本结构信息帮助 Twitter 检索。我们首先定义 Twitter 积木以此捕获 tweet 文本结构，然后构造了 Twitter 积木自动识别器，通过对 tweet 积木的自动识别开发特征，设计了基于结构化信息的 Twitter 排序模型，最后实验验证该模型对于 Twitter 信息检索的有效性。

2.3 Twitter 积木（TBB）

这一节我们主要介绍 Twitter 积木以及积木的自动识别。通过 Twitter 积木可以定义 tweet 的文本结构，这种结构反映了 tweet 的文本属性（如话语转换和文本质量）和潜在的社会属性。我们希望通过 tweet 文本的结构信息开发特征，帮助 Twitter 信息检索任务。

一个 tweet 可以看成由一些文本块组合而成的文本，而文本块则是一些词的排列组合而成。我们将这些文本块称为“Twitter 积木（Twitter Building Blocks 或 TBB）”。各种 Twitter 积木的排列组合则构成了“Twitter 积木结构（TBB structure）”。

2.3.1 Twitter 积木的定义

在 Twitter 中人们经常使用三种方式表现其行为，包括标注（为 tweet 添加标签指明 tweet 的主题）、提及（在 tweet 中提及某人）、转发（再次重新发布其他人发布的 tweet）。另外，tweet 本身的内容也可以分成三类，包括链接、观点、普通文本。基于此我们定义了六种不同的 Twitter 积木：

1. **标签积木（TAG）**：“#”和关键词的绑定使用，主要用来指明 tweet 的主题，如“*#iphone*”。
2. **提及积木（MET）**：在 tweet 中提及某人的符合串，可以让对方看到此 tweet，如“*@ladygaga*”。
3. **转发积木（RWT）**：指明复制或再次传播其他人发布 tweet 的符合串，如“*RT @ladygaga*”。
4. **链接积木（URL）**：链接到 tweet 本身内容以外的网址，如“*http://www.facebook.com*”。
5. **观点积木（COM）**：用来表达作者对其他 Twitter 积木的态度、评估、情绪。
6. **普通文本积木（MSG）**：其他文本。

图 2.4给出了手工标注 Twitter 积木的两个 tweet 样例。每个由下划线标注的词序列是一个 Twitter 积木。从图 2.4中我们看到 Tweet (a) 是一个由“COM RWT MET MSG”积木结构组成的 tweet，Tweet (b) 则是由“MSG URL TAG”积木结构组成的 tweet。积木本身的内容和组合反映了 tweet 的话语转换信息。例如，Tweet (a) 是作者转发（RWT）用户 @miiisha_x 的普通文本积木（MSG），这个普通文本积木（MSG）提及了（MET）用户 @XPerkins，同时作者还给出了对于普通文本积

木（MSG）的观点 (COM)。Tweet (b) 中作者则先后给出了普通文本积木（MSG）、一个链接积木（URL）和两个 hashtag 组成的标签积木（TAG），作者在后面两个积木提供了其他资源且标注了 tweet 的主题，使得其他读者能更好地理解原来普通文本积木（MSG）的内容。

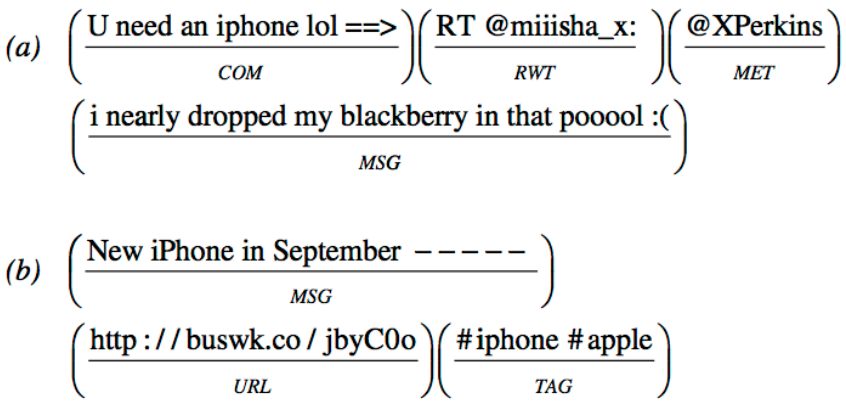

图 2.4　手工标注 Twitter 积木的 tweet 示意图

为了理解人们是如何使用这些 Twitter 积木，我们随机的选择了 2000 个英文 tweet[2]，并对其进行了 Twitter 积木及其结构的手工标注。在标注之前我们首先使用了 O'Connor 等人开发的工具[3]自动地对每个 tweet 进行单词化（tokenized）[96]。表 2.1给出了不同 Twitter 积木结构的比例分布。这里列出 14 个最常用的 Twitter 积木结构，其他非常用的 Twitter 积木结构则统称为“OTHERS”。我们可以看出“MSG”结构比例最高，这说明这种最简单的结构是最常用的 Twitter 积木结构。其他比例较高的 Twitter 积木结构是比较简单的，一般不超过三块积木。而由其他比例较低 Twitter 积木结构归类而成的“OTHERS”仅仅存在 13%，以上说明人们大多使用简单和固定的结构发布 tweet。

2.3.2　Twitter 积木自动标注

手工标注每个 tweet 的 Twitter 积木及其结构显然是不可行的，因此我们设计了自动标注器来完成 Twitter 积木自动标注的任务。这个任务可以看成由两个子任务组成：Twitter 积木的分类和 Twitter 积木的边界识别。这个任务十分类似于自然语言处理中的命名体识别问题（Named Entity Recognition）[106]，因此我们采用序列标注的方法（Sequential Labeling Approach）一同解决 Twitter 积木自动标注的两个子任务，另外，我们还采用了 IOB-类型标注模式[23, 33]。

[2]我们使用语言识别工具对非英语 tweet 进行过滤 http://code.google.com/p/language-detection/

[3]下载地址为：https://github.com/brendano/tweetmotif

表 2.1　Twitter 积木结构比例分布

Twitter 积木类型	比例 (%)	Twitter 积木类型	比例 (%)
MSG	30.25	TAG MSG	1.55
MET MSG	20.70	TAG MSG URL	1.20
MSG URL	18.40	RWT MSG URL	0.95
OTHERS	13.20	COM RWT MSG	0.85
COM URL	4.10	MET MSG URL	0.85
MSG TAG	2.65	MSG MET MSG	0.70
MSG URL TAG	2.10	RWT MSG TAG	0.70
RWT MSG	1.75		

输入一个 tweet，我们的目的是输出一个序列 Twitter 积木块 $B_1B_2...B_m$，其中 B_i 是一个词串 $t_{i1}t_{i2}...t_{in}$。每一个在 tweet 中的词 t_{ij} 将被标注一个标签“*X_Y*”*(X = TAG*，*MET*，*RWT*，*URL*，*COM*，*MSG*；*Y = B*，*I)*，这个标签说明了词的类型和是否是边界词。每个在同一个积木块中的词具有相同的 *X* 值，“*Y = B*”表示词 t_{ij} *(j=1)* 为边界词，而 “*Y = I*”表示为非边界词。例如，图 2.4中 Tweet (b) 的词“iPhone”和“#iphone”，它们的标签分别是“MSG_I”和“TAG_B”。

我们利用条件随机场（*Conditional Random Field*）进行序列标注[107]，这种方法能够将设计的局部特征很好的纳入学习模型中。我们利用 Twitter 不同积木经常出现的词语组成样式、长度范围、词性特点、所在 tweet 位置的前后环境来设计特征，帮助 Twitter 积木识别，具体特征包括：

1. **词类型（Token Type）**：长度为 7 个词的文本窗口，其中待标注的词在窗口中间。

2. **词性（Pos）**：每个词在 tweet 中的词性[4]。

3. **长度（Length）**：每个词中有几个字符。

4. **前后缀（Pre_Suf_fix）**：词中前面和后面三个或三个以下的字符串。

5. **Twitter 启发式（Twitter orthography）**：一些简单的规则识别 Twitter 积木：

   a)：待标注词是否以“#”开头，如果是则该词一般是标签积木（TAG）的组成部分；

   b)：待标注词是否以“www.”，“http:”开头或以“.com”结尾，如果是则该词一般是链接积木（URL）的组成部分；

[4]我们使用了 CMU 的 tweet 词性标注器进行词性标注[21, 108]，下载地址为`http://www.ark.cs.cmu.edu/TweetNLP`

c)：如果某段词串是“@usename :”或“@usename”，则该词串一般是提及积木（MET）；

d)：如果某段词串是“RT @usename :”、“RT @usename”、“RT”或“via @usename”，则该词串一般是转发积木（RWT）；

e)：词串“RT @username”的前面和“via @username”或“«”的后面一般是观点积木（COM）。

表 2.2　自动标注 Twitter 积木结构

标签类型	数目	准确率 (%)	召回率 (%)	F1 (%)
TAG_B	72	88.00	91.67	89.80
TAG_I	34	93.94	91.18	92.54
URL_B	164	95.62	93.29	94.44
URL_I	24	55.56	41.67	47.62
MET_B	145	91.45	95.86	93.60
MET_I	63	94.34	79.37	86.21
RWT_B	72	93.06	93.06	93.06
RWT_I	129	90.51	96.12	93.23
COM_B	70	67.27	52.86	59.20
COM_I	550	64.48	46.55	54.07
MSG_B	482	90.50	90.87	90.68
MSG_I	5708	94.27	97.06	95.64
AVG		84.92	80.79	82.80

我们利用前面手工标注的 2000 个 tweet 进行训练和测试。其中 1000 个 tweet 作为训练集，500 个 tweet 作为开发集，500 个 tweet 作为测试集。另外，我们使用工具 FlexCRFs[5] 对手工标注数据进行训练。表 2.2 是自动标注 Twitter 积木的结果。其中 tweet 中词标签正确识别的平均 F1 值达到 82.80%，而识别词为“COM_B”和“COM_I”的 F1 值相对较低，主要原因可能是观点积木（COM）在训练集中很少被标注造成模型泛化能力不足，另外，观点挖掘也一直是自然语言处理研究中一个难点问题[109]。识别词为“URL_I”的 F1 值也较低，主要原因是一些链接词被 Twitter 单词化器错误切分[96]，不过我们从表 2.2 中可以看到标签为“URL_I”的词的数目很少，因此从整体上对 tweet 进行 Twitter 积木标注影响很小。通过对 tweet 中的词进行标注我们可以获得 tweet 中的积木及其结构，最后 tweet 积木结构的整体识别率可以达到 82.60%。

[5]下载地址为：http://flexcrfs.sourceforge.net/

2.3.3 Twitter 积木分析

我们可以对 tweet 按照不同的 Twitter 积木结构进行聚类，并以此分析了各种 Twitter 积木结构的不同属性和文本质量。我们发现了以下特点：

1. **公共广播式**：一些由类似于 BBC News 发布的 tweet 往往具有如下 Twitter 积木结构：“MSG URL”、“MSG URL TAG”、“TAG MSG URL”，这些 tweet 一般首先给出一段介绍文本然后紧跟着一个相关链接。

2. **私人广播式**：而一些由普通用户（粉丝数目不多）发布的 tweet 则往往具有如下 Twitter 积木结构：“COM URL”、“MET MSG URL”。例如 tweet “*I like it and the soundtrack* `http://www.imdb.com/title/tt1414382/`” 的 Twitter 积木结构是“”COM URL”。人们关心这些结构 tweet 的人数一般远远少于关心公共广播式 tweet 的人数。

3. **高质量新闻**：高质量的新闻 tweet 用得最多 Twitter 积木结构则是“RWT MSG URL”。例如，tweet “*RT @CBCNews Tony Curtis dies at 85* `http://bit.ly/dlSUzP`” 不仅仅是一个新闻，而且还是一个热点事件。

4. **杂乱信息**：那些不经常使用且较复杂的结构一般是“OTHERS”Twitter 积木结构。例如，tweet “*RT @preciousjwl8: Forreal doeee? (Wanda voic) #Icant cut it out #Newark* `http://twipic.com/2u15xa`*...lmao!!WOW ...* `http://tmi.me/1UwsA`”。这种结构的 tweet 一般话语转换复杂不容易被人们马上理解。

另外，Twitter 实时地在线发布大量的消息，其文本质量差异性很大，高质量的 tweet 如新闻媒体发布的信息，低质量的 tweet 则可能仅仅包含一些毫无意义的词[2, 25, 27, 30, 110]。以前的研究发现将 tweet 的文本质量因素考虑到 Twitter 信息检索中有利于性能的提高[88–90, 100, 111]，基于此我们分析了不同 Twitter 积木结构与其所对应 tweet 的文本质量关系。

通过我们的自动 Twitter 积木标注器标注 tweet，我们对每种 Twitter 积木结构随机的选取了各 10000 个 tweet，然后计算每种 Twitter 积木结构所对应 tweet 中 OOV 值（Out of Vocabulary Value）。OOV 值是计算 tweet 的普通文本积木（MSG）和观点积木 (COM) 中包含未登录词的个数，然后除以 tweet 的这两个积木中总的词个数。这个值可以粗略地估计 tweet 的文本质量。为了适用 tweet 中包含许多 Twitter 所独有的词汇，我们自己构造了词典。构造的方法是从 1 百万个 tweet 中选取频率最高的 50 万个词组成词典。通过这种方法我们发现大多数未登录词是

一些拼写错误的词或缩写词。表 2.3给出了不同 Twitter 积木结构所对应的 OOV 值。我们发现不同结构的 OOV 值存在明显的差异。例如，“RWT MSG TAG”和“RWT MSG”Twitter 积木结构具有较小的 OOV 值，这说明人们经常转发其他人高质量的信息。而“OTHERS”Twitter 积木结构则 OOV 值最高，可能的原因是每个 tweet 最多包含 140 个字符，而“OTHERS”Twitter 积木结构中一般具有较多的标签积木（TAG)、提及积木（MET)、转发积木（RWT）和链接积木（URL)，这些积木的大量存在造成人们在普通文本积木（MSG）和观点积木 (COM) 中大量使用缩写词来压缩这两个积木块的长度。

表 2.3　不同 Twitter 积木结构的 OOV 值

Twiter 积木结构类型	**比例 (%)**	**Twiter 积木结构类型**	**比例 (%)**
OTHERS	4.30	MET MSG URL	1.42
TAG MSG URL	3.42	MSG	1.32
MSG URL	1.93	MSG TAG	1.31
MSG URL TAG	1.91	RWT MSG URL	1.30
COM URL	1.80	MET MSG	1.15
COM RWT MSG	1.78	RWT MSG	0.82
MSG MET MSG	1.64	RWT MSG TAG	0.58
TAG MSG	1.63		

2.4　基于 Twitter 积木的 tweet 排序学习

我们通过 tweet 的 Twitter 积木及其结构开发特征，并将其加入到排序学习的框架中（Learning to Rank)，以此评估 Twitter 积木对于 tweet 信息检索效果的影响。

2.4.1　Twitter 检索排序学习框架

排序学习是一种将特征有效地整合到排序模型的机器学习算法[112]。以往的检索模型，主要是利用词频，倒转文档频率和文档长度等几个因素来人工拟合一些排序公式，以此计算查询词与文档的相关性，然后对文档进行排序。因为所要考虑的因素不多，人工进行的公式拟合是可以实现的，但随着搜索引擎技术的发展，网页在排序过程中需要考虑的因素越来越多，比如网页的 PageRank 值[113]，网页的 URL 信息等等都会影响网页的排名，这些影响排序的因素有时可以达到几十甚至上百种，所以以前通过手工拟合排序公式的方法变得越来越复杂。而机器学习对这种类型的工作却十分合适，它可以将影响排序的因素转换为特征进行训练与测试，按照不同特征的组合，通过排序效果，验证哪些特征影响排序，以此说明哪些因素会影响查询词与文档的相关性。

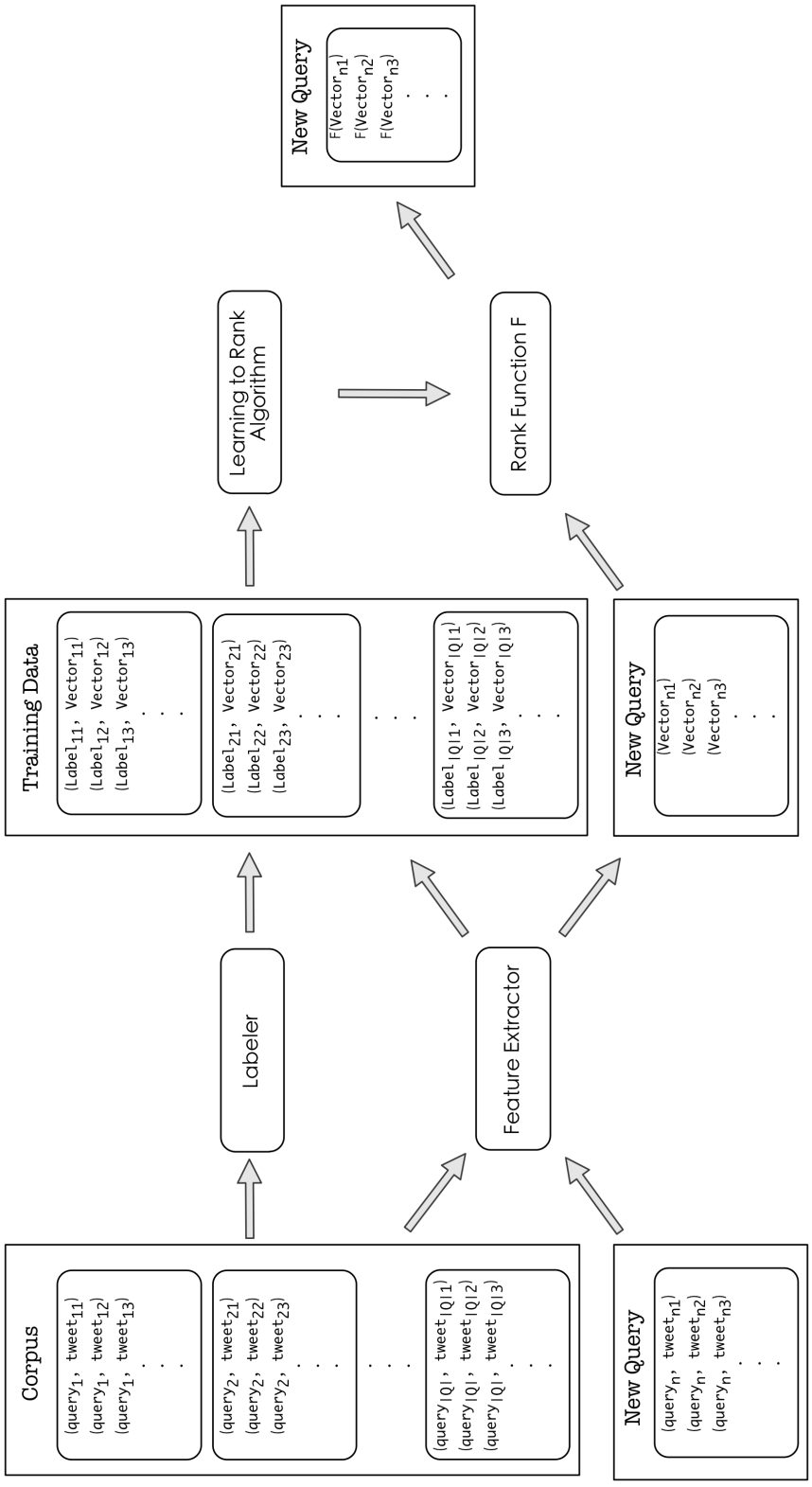

图 2.5 Twitter 检索排序学习框架

另外，对于有监督学习来说，一个重要的条件就是大量的训练数据，但是在排序学习过程中，大量人工标注训练数据是不现实的，但是搜索引擎可以通过用户的点击动作模拟点击网页与用户查询的相关性，以此获得大量的训练数据。

对于 Twitter 中的信息检索，我们可以通过排序学习模型对给定关键词的 tweet 集合进行排序。我们的目的是验证我们开发的特征是否对于 Twitter 的排序有效，因此使用不同特征集合排序模型在 Twitter 中 tweet 排序结果可以反映不同特征的有效性。图 2.5给出了基于排序学习的 Twitter 检索框架。

首先给定一个查询集合 Q 及其对应的 tweet 作为训练集，每个 $tweet_{ij}$ 被手工标注了是否与对应的查询词 $query_i$ 相关。一系列与 tweet 查询相关的特征被设计和开发，形成特征向量 $Vector_{ij}$。接着，应用排序学习算法对手工标注的数据进行训练生成排序学习模型。对于一个新的查询和其所对应的 tweet，抽取相同的特征形成特征向量 $Vector$，然后利用生成好的排序学习模型对 tweet 进行相关排序。对特定一组特征生成的排序模型在测试数据集上的排序表现能够反映特征对于排序任务的有效性。

这里要强调的是，本文所涉及的 Twitter 检索、Twitter 观点检索、Twitter 传播观点发现和 Twitter 传播者发现任务都采用了排序学习框架，通过不同特征组合形成的模型，在排序效果上的表现，说明 Twitter 中哪些 tweet 文本特征和 Twitter 的社交媒体特征会影响对应的任务。

2.4.2 Twitter 积木特征

对于每一个 tweet，我们从 Twitter 积木及其结构中设计了一系列特征，并称之为 Twitter 积木特征（TBB features）。这些 Twitter 积木特征仅仅利用了 tweet 的文本信息并未利用其他社交媒体属性。我们将这些特征分为如下几个类别：

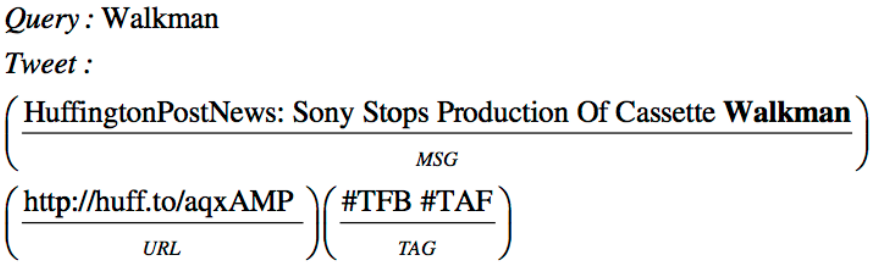

图 2.6 一个查询词和一个 tweet

1. **Twitter 积木结构类型（TBB Structure Type）**：每个 tweet 都有自身不同的结构。我们设计了一个 15 维的特征向量，其中特征向量中的每个维度表示当前 tweet 是否是特定的某种 Twitter 积木结构。我们选取 14 种经常使用的

Twitter 积木结构作为特征维度，其他很少使用的 Twitter 积木结构作为剩下的特征维度（见表 2.1）。如果 tweet 是某个 Twitter 积木结构，则特征向量中所对应的特征维为 1，其他特征维为 0。例如，图 2.6中 tweet 的 Twitter 积木结构是“MSG URL TAG”，则特征向量所对应的该维特征取值为 1，其他维取值为 0。

2. **查询词 Twitter 积木位置（TBB Query Position）**：我们利用六个维度的特征向量表示查询词所在 Twitter 积木的位置。因为一个查询词经常是一个短语或是一个 hashtag，所以特征表示为查询词是否在普通文本积木（MSG）、观点积木 (COM)、标签积木（TAG）的开头或中间。例如，图 2.6中查询词“Walkman”就在普通文本积木（MSG）中间。如果查询词在对应的积木位置，则特征值为 1，否则为 0。

3. **邻近 Twitter 积木类型（Neighbour TBB Type）**：查询词所在 Twitter 积木的前后积木类型同样具有重要的信息，我们将其利用与开发。特征包括查询词所在 Twitter 积木的前后积木是否是标签积木（TAG）、提及积木（MET）、转发积木（RWT）、链接积木（URL）、普通文本积木（MSG）、观点积木 (COM)。例如，图 2.6中查询词“Walkman”所在积木的后个积木是链接积木（URL）。

4. **Twitter 积木数目（TBBs Count）**：直觉上如果一个 tweet 存在较多的包含查询词的 Twitter 积木，则该 tweet 更有可能是人们需要的相关 tweet。因此，我们设计了一个特征，这个特征表示 tweet 中存在几个包含查询词的 Twitter 积木。例如，图 2.6中 tweet 中仅仅存在一个包含“Walkman”查询词的 Twitter 积木。

5. **Twitter 积木长度（TBB Length）**：直觉上如果 tweet 中包含查询词的 Twitter 积木长度越长，则该 tweet 中与查询词相关的信息则越多。因此，我们设计了最长包含查询词 Twitter 积木词个数的特征。

6. **Twitter 积木 OOV 值（TBBs OOV）**：这个特征计算 tweet 中存在未登录词的比例，以此评估 Twitter 积木的质量。

7. **Twitter 积木语言（TBB Language）**：这是一个布尔特征用来表示包含查询词的 Twitter 积木是否是英语，因为人们一般会选择语言是母语的 tweet 作为相关 tweet。

2.5 Twitter 信息检索实验

2.5.1 Twitter 信息检索实验数据

我们利用 Twitter streaming API 每天获取 80 万个 tweet 并建立索引，然后构造了一个搜索引擎。三个用户参与实验并标注数据，这三个用户都是计算机科学研究人员，它们从 2010 年 10 月 4 日至 2010 年 10 月 28 日使用我们的搜索引擎。它们任意输入自己喜欢的查询词，然后搜索引擎根据 BM25 算法[6]对收集的 tweet 进行排序并返回结果，搜索引擎默认返回 10 个 tweet，另外，返回的 tweet 中同时显示时间信息和 tweet 的作者信息。然后这三个用户根据返回 tweet 是否与查询词相关对 tweet 进行标注，如果相关则标为 1，否则为 0。最后我们总共收集了 100 个查询词和其返回的 tweet。我们对这 100 个查询词及其 tweet 进行分析与统计，发现这 100 个 tweet 大致可以分成六类（见表 2.4)。这六类中数目最多的是“热点”类，如：查询词“Chilean miner”，另外，还有很大一部分查询词与科技相关（归为“科技”类)，如查询词“java flaw”，“新闻”类是一些重要的新闻信息，如“wikileaks”，有意思的是三个用户还输入了一些“地理”查询词，以此期望在 Twitter 中找到一些所在地的信息，“娱乐”类中包含了一些电影和明星的查询词，剩下的查询词都归为“其他”类。

表 2.4 Twitter 信息检索查询词类别及其数目

类别	数目
热点	34
科技	27
新闻	17
地理	8
娱乐	8
其他	6

表 2.5则给出了实验数据的一些统计信息。

[6]BM25 具体算法将在 3.6.3介绍。

表 2.5　Twitter 信息检索实验数据统计信息

查询词长度（词数）	1.48
查询词对应 tweet 平均数目	9.36
相关 tweet 总数目	184
非相关 tweet 总数目	752

2.5.2　信息检索评价指标

平均准确率（Mean Average Precision-MAP）是信息检索中经常使用的评价指标[114, 115]。它对相关文档与不相关文档的排序非常敏感。我们利用平均准确率来评价排序系统的有效性。计算平均准确率的公式如下:

$$MAP(Q) = \frac{1}{|Q|}\sum_{q=1}^{|Q|} AveP(q)$$

其中 Q 是查询词集合，$|Q|$ 是查询词集合中查询词数目，$AveP(q)$ 是查询词 q 的平均准确率。

$$AveP(q) = \frac{1}{|rel|}\sum_{n:p_n=1} Prec@n$$

其中 $|rel|$ 是相关文档数目，$Prec@n$ 是检索到 top n 个文档中相关文档的比例。

$$Prec@n = \frac{|i : 1 \le i \le n, p_i = 1|}{n}$$

其中 $p_i = 1$ 表示排在第 i 个位置的文档是相关文档。

2.5.3　Twitter 信息检索实验设置和基准系统（Baseline）

我们进行了 10 次交叉验证，并将 Duan 等人的方法作为我们的基准系统（Baseline）[88]，这个系统目前是最好的基于排序学习的 Twitter 信息检索系统。基准系统在我们构造的数据集上 tweet 的排序平均准确率为 0.344。我们没有采用 Duan 等人使用的以 BM25 分值作为排序基础的基准系统, 因为我们的系统和 Duan 等人的系统在排序效果上都显著好于 BM25 系统。另外，我们还开发了一个利用社交媒体特征的 Twitter 信息检索系统，称之为 SM_Rank。表2.6列出了 Baseline 和 Social Media 两个系统所使用的特征。我们将利用 Twitter 积木标注器自动标注积木然后抽取结构化特征开发的 Twitter 信息检索系统称为 TBB_Rank。最后基于三组不同特征集合组合的系统分别称为 Baseline+SM_Rank, Baseline+TBB_Rank, SM+TBB_Rank。

表 2.6　基准系统特征（Baseline Features）和社交媒体特征（Social Media Features）

基准系统特征（Baseline Features）	**描述**
链接（Link）	tweet 中是否包含链接
长度（Length）	tweet 中包含的词数目
重要粉丝（Important_follower）	tweet 的作者或转发该 tweet 的用户中 follower score[1]最高的用户粉丝数目
提及总和（Sum_mention）	tweet 的作者和转发该 tweet 的用户 mention scores[2]的总和
分组数目（First_list）	tweet 的作者 List score[3]
社交媒体特征（Social Media Features）	**描述**
粉丝数目（Followers）	tweet 的作者的粉丝数目
朋友数目（Friends）	tweet 的作者的朋友数目
分组数目（Listed）	tweet 的作者的分组数目
提及（Mentions）	tweet 是否包含提及
Hashtag 数目（Hashtags）	tweet 中包含 hashtag 的数目
回复（Reply）	tweet 是否是回复
转发（Retweeted）	tweet 是否是转发
来源（Source Web）	tweet 的消息是否来源于 web
发布数目（Statuses ）	tweet 的作者以往发布 tweet 的数目
转发数目（Retweet Count）	tweet 被转发的次数
作者被转发数目（Author Retweet Count ）	tweet 的作者在收集的 tweet 中被转发的次数
公共词数目（Overlap Words）	查询词与 tweet 公共词数目 (Jaccard score)
Tweet 发布时间（Recency）	tweet 发布时间到用户输入查询词的时间差 (以秒计)

[1] Follower Score：用户粉丝数目；

[2] Mention Score：在收集的 tweet 中用户被提及的个数；

[3] List Score: 用户的分组数目。

2.5.4　Twitter 信息检索实验结果及分析

表 2.7给出了各个 Twitter 信息检索系统的 tweet 排序实验结果。我们可以看到仅仅从 tweet 文本获取结构化信息构造的 TBB_Rank 系统，在排序效果上可以达到与 Baseline 和 SM_Rank 系统相当的效果。我们进行了显著性测试（paired t-test），发现三个系统在实验结果上没有显著性差异（$p = 0.05$）。我们进一步发现将 Twitter 积木特征和社交媒体特征结合使用的 SM+TBB_Rank 系统能够显著提高 tweet 排序效果。最后将三组特征集合一起使用的 Baseline+SM+TBB_Rank 系统获得最高平均准确率 0.4712。所有这些都说明结构化信息能够帮助 Twitter 信息检索。

表 2.7　基于排序学习的 Twitter 信息检索实验结果

	MAP
Baseline	0.4197
SM_Rank	0.4338
TBB_Rank	0.4235
Baseline+SM_Rank	0.4546
Baseline+TBB_Rank	0.4326
SM+TBB_Rank	0.4710*†
Baseline+SM+TBB_Rank	0.4712*†

* 和 † 分别表示排序结果显著高于 Baseline 和 SM_RankTwitter 信息检索系统。

Duan 等人发现一个 tweet 中是否包含一个链接是基于排序学习的 Twitter 信息检索系统中最重要的特征[88]。我们在此基础上进一步研究分析了那种包含链接积木（URL）的 Twitter 积木结构对于 tweet 排序更加重要。我们将问题转化为表 2.8中的特征那个对于 tweet 排序更加重要。我们将这些特征一个一个替换基准系统中的链接（Link）特征，然后观察平均准确率的变化。

表 2.8　基于链接积木（URL）的 Twitter 积木结构特征

链接积木（URL）特征	**描述**
MSG URL	tweet 的 Twitter 积木结构是否是“MSG URL”
RWT MSG URL	tweet 的 Twitter 积木结构是否是“RWT MSG URL”
COM URL	tweet 的 Twitter 积木结构是否是“COM URL”
TAG MSG URL	tweet 的 Twitter 积木结构是否是“TAG MSG URL”
RWT MSG URL	tweet 的 Twitter 积木结构是否是“RWT MSG URL”
MSG URL TAG	tweet 的 Twitter 积木结构是否是“MSG URL TAG”
OTHER URL	tweet 的 Twitter 积木结构是否是其他包含链接积木（URL）的非常用积木结构

表 2.9给出了实验结果。我们发现特征“MSG URL”能够在基准系统中（Baseline）替换链接（Link）特征时，平均准确率没有显著下降（$p = 0.05$），而其他特征的替换都显著地影响了系统的 tweet 排序效果，这说明这个特征能够取代基准系统中（Baseline）的链接（Link）特征，同时也说明了一个 tweet 的 Twitter 积木结构是否是“MSG URL”对于 tweet 排序具有重要的影响。这个原因可能是因为大多数包含链接 tweet 的 Twitter 积木结构是“MSG URL”（见表 2.1），而如果一个 tweet 它的结构是“MSG URL”则它很有可能是相关 tweet。例如在我们的实验数据中存在如下例子，查询词为“wikileaks”，它有两个对应的 tweet：

(a) *Obama administration braces for WikiLeaks release of thousands of secret documents on Iraq war (Star Tribune)* `http://bit.ly/9lnBGB`

(b) *BBCWorld: Wikileaks files 'threaten troops'* `http://bbc.in/c4Sznk`: *BBC-World: Wikileaks files 'threaten troops'...* `http://dlvr.it/7P7zM`

标注者将 Tweet (a) 标注为相关 tweet 而 Tweet (b) 标注为不相关 tweet。Tweet (a) 的 Twitter 积木结构是“MSG URL”，Tweet (b) 的 Twitter 积木结构是“MSG URL MSG URL”。标注 Tweet (b) 为不相关 tweet 的原因可能是这个 tweet 包含两个链接积木（URL），造成读者理解这个 tweet 比较混乱。在我们的实验中，我们的基准系统（Baseline）和 SM_Rank 系统都将 Tweet (b) 排在 Tweet (a) 前面，而 TBB_Rank 系统则将 Tweet (a) 排在 Tweet (b) 前面。这说明我们的 Twitter 积木能够获取更多的包含链接 tweet 的信息，并以此改善 tweet 的排序效果。

表 2.9　基于链接积木（URL）的 Twitter 积木结构特征排序实验结果

	MAP
Baseline	0.4197
MSG URL	0.4019
MSG URL TAG	0.3327
RWT MSG URL	0.3289
TAG MSG URL	0.3245
COM URL	0.3191
OTHER URL	0.1984
MET MSG URL	0.1932

2.6　小结

本章我们介绍了 Twitter 积木及其结构的定义与识别，并以此获取 tweet 的结构化信息。我们分析了不同 Twitter 积木结构具有不同属性，例如，不同 Twitter 积木结构所对应 tweet 的 OOV 值明显不同。我们基于 Twitter 积木及其结构设计了一系列特征，并将其利用到基于排序学习的 Twitter 信息检索应用中，实验结果表明，在 tweet 排序上仅仅使用我们设计的结构化特征，能够达到当今最好的基于排序学习的 Twitter 信息检索系统的效果。如果将社交媒体特征和结构化特征结合使用能够得到更好的 tweet 排序效果。以上工作说明虽然 tweet 文本十分简短，但是其文本依然具有结构化信息，并能够帮助提高 Twitter 信息检索的效果。

第三章　Twitter 观点检索

3.1　引言

Twitter 是一种流行的社交媒体，拥有超过 5 亿的用户并且每天产生多于 3.4 亿个 tweet[1]。人们喜欢在 Twitter 中分享消息并对人物、政治、商品、公司、事件等进行评论，使得 Twitter 变成一个拥有丰富观点信息的资源，不仅可以帮助人们做出决策，还可以辅助政府与公司收集有价值的反馈信息。例如，Jansen 等人利用 tweet 作为某些商品的口碑信息进行分析[116]；O'Connor 等人分析了 tweet 的文本情感因素，以此作为公共的舆情分析[117]；Bollen 等人利用 Twitter 中人们通过文本反映的情绪变化来预测股票市场[118]。但是大多数已有的工作都是在 tweet 话题已经明确（涉及特定话题）的基础上的相关观点进行分析，并没有在给定某些人、商品或事件时，如何在 Twitter 中找到相关评论的工作。

本章中我们将讨论如何在 Twitter 中进行观点检索的工作，相对于上一章的 Twitter 信息检索工作，相关的观点 tweet 需要满足如下两个标准：

1. 与查询词话题相关。
2. tweet 包含涉及查询词的主观化观点，但不考虑观点的态度是正面的还是负面的。

以下是两个关于查询词“UK strike”的 tweet：

a) *“Perhaps if the public sector workers on #strike today go Christmas shopping then at least it will give the high street / UK economy a boost!”*

b) *“UK: BBC - Up to TWO Million Set to Strike http://t.co/wBrsgrKh #tcot #gop #ows”*

这两个 tweet 中，Tweet(a) 是与查询词相关的观点 tweet（本章以后我们简称为相关 tweet）；Tweet(b) 不是与查询词相关的 tweet，因为它没有作者对于话题“UK strike”的观点。

博客和 Web 网页中的观点检索已经被大量的研究[119–123]。最近文本检索会议（Text Retrieval Conference-TREC）也组织了对于博客的观点检索评测（博客 Track）[124–127]。传统的观点检索面对的主要问题是：（1）文档的观点性判定；（2）文档的观点是否是对给定话题的评价。除了以上普遍存在的问题，Twitter 中的观点检索难度更大，主要原因是：

[1] http://www.mediabistro.com/alltwitter/500-million-registered-users_b18842

1. **文本简短**：tweet 文本十分简短且最多只有 140 个字符。这种短文本属性造成人们经常使用缩写词或短语等压缩 tweet 的文本长度[110]，Luo 等人也发现如果 tweet 中存在很多链接、hashtag、固定字符串（例如，“@username”和“RT @username”），也会造成人们经常使用缩写词或短语[83]，这样 tweet 中就会大量存在未登录词，结果造成许多词无法识别，使得查询词与 tweet 中的文本无法在检索时进行很好的词匹配[97]。

2. **文本质量差异大**：对于观点检索来说，Twitter 是一个全新的领域，存在大量的垃圾信息和文本变化大的 tweet，但是大多数网页（如新闻网页）都是一些文本质量相对较高的正式文档，而 tweet 大多数是个人发布的非正式文本。这就造成 tweet 的文本质量明显不同于其他文本，对于观点检索来说是个巨大的挑战。

但是 Twitter 自身的特点可以帮助 Twitter 中的观点检索。因为 Twitter 中包含丰富的社交媒体特征，例如，以前发布过多少 tweet 的用户信息。这些特征可以弥补词方面检索不匹配的缺点，间接提高观点检索的效果。另外，一些人们发布 tweet 的习惯使得 tweet 可以被看成一个结构化文本，这种结构化信息也可以帮助 Twitter 中的观点检索。

例如，人们经常在字符串“RT @usename”前面加评论信息，这种类型的 tweet 一般都是主观化的 tweet。而大多数新闻媒体用户（例如，BBC News）发布的 tweet 一般都是给出一段介绍文本，紧接着给出一个相关链接，直观上这种类型的 tweet 一般是客观的信息。更重要的是，这些结构化信息不依赖于特定话题。因此这就促成我们利用这些社交媒体特征与 tweet 的结构化信息帮助 Twitter 的观点检索。

本章中我们将介绍如何利用机器学习构造检索模型的方法实现 Twitter 中的观点检索，这种方法利用了社交媒体特征，观点化特征和话题相关特征（例如，Okapi BM25[128, 129] 和向量空间模型[130]）。实验结果表明当检索模型考虑 tweet 中是否存在链接、提及、作者曾发布过多少 tweet、粉丝数目和观点化特征时，Twitter 中的观点检索效果明显提高。

另外，我们利用基于语料的方法构造观点化词典评价 tweet 中文本的观点化程度，但是这种方法依赖手工标注的语料，手工标注语料耗费人力与时间，而且对 tweet 的观点化评价是一个与话题相关的问题[131–133]，在对 tweet 进行观点化评价中不可能针对每个话题手工标注语料进行词典构造，为此我们提出了一种新的方法来解决以上问题。这种方法利用社交媒体特征与 tweet 文本的结构化特征自动收集一些近似主观化 tweet（‘pseudo’ subjective tweet-PST）和近似客观化 tweet（‘pseudo’ objective tweet-POT），收集好的这两种数据集合能够自动生成观点化词

典来评价 tweet 的观点化程度。我们将这种方法与传统的手工标注语料进行了实验比较，发现在 Twitter 观点检索中，该方法取得的效果与手工方法效果相当，当自动收集与话题相关的近似主观化 tweet 和近似客观化 tweet 时，Twitter 观点检索效果获得进一步提高。

本章的主要工作如下:

1. 根据我们的调研这是第一次对 Twitter 进行观点检索的研究。我们发布了研究过程中所使用的检索数据[2]，这个数据包括 50 个查询词和其对应的 5000 个手工标注相关与不相关的 tweet。

2. 我们提出了一种 Twitter 观点检索方法，这种方法利用社交媒体特征与 tweet 观点化特征，并在实验结果上显著优于优化的 BM25 基准系统和 VSM 系统(分别提高 56.82% 和 33.75%)。

3. 另外，我们还提出了一种基于社交媒体特征与 tweet 文本结构化信息收集近似主观化 tweet（‘pseudo’ subjective tweet-PST）和近似客观化 tweet（‘pseudo’ objective tweet-POT）构造主观化词典的方法，并以此评价 tweet 的观点化程度，实验结果表明这种方法对于 tweet 的主观化评价能够取代传统手工标注语料构造词典的方法，并在 Twitter 观点检索中取得效果相当的结果，当考虑针对话题收集 PST 和 POT 时，使用该方法的 Twitter 观点检索效果更好。

4. 最后我们重新标注了 TREC Tweets2011 数据[134]，使其能够对 Twitter 观点检索方法能够进行评价，实验结果进一步验证了我们方法的有效性。

3.2 相关工作

我们从两个方面讨论相关工作：Twitter 中的观点挖掘和 TREC 观点检索。Twitter 中的观点挖掘主要是通过自然语言处理技术在 Twitter 中发现和跟踪人们对事件、产品等的观点和态度；TREC 观点检索是在博客数据集合中，给定查询词，找到与查询词话题相关且带观点的博客文档。前者能够帮助我们在观点检索任务中对 tweet 进行观点判定，后者可以为我们在 Twitter 中进行观点检索提供技术思路。

3.2.1 Twitter 观点挖掘

Twitter 吸引了成千上万的用户在这个平台上发布各种观点，这也使得其成为一个研究的热点对象。

[2]下载地址：https://sourceforge.net/projects/ortwitter/

Jansen 等人研究了关于特定品牌的 tweet[116]，他们设计了一个舆论感知器，量化品牌的公司与消费者之间的紧密关系，他们发现 19% 的 tweet 包含或提及与商品有关的信息，这其中大约 20% 包含人们对商品的观点。这说明 Twitter 可以成为公司收集消费者对于商品文本评价的资源库。

O'Connor 等人通过分析 tweet 的态度度量了社会舆情[117]，这种方法可以替代传统的问卷调查。他们分析了消费者对于商品态度的 tweet 和对美国总统施政评价的 tweet，结论是这种简单的方法可以替代费用高昂且实时性要求强的社会问卷调查方式。

Bollen 等人从大规模 Twitter 数据中分析了道琼斯工业平均指数（Dow Jones Industrial Average[3]）与 tweet 的情绪倾向性存在很强的联系[118]。他们发现公众的情绪状态的变化的确可以在大规模 Twitter 数据中被跟踪，这种不断变化的信息，可以成功地帮助预测股市。这表明 Twitter 是一个收集公众意见的重要来源。

以上的研究主要集中在给定话题的情况下挖掘公众的观点，而不是如何在 Twitter 中找到特定话题的观点 tweet。我们本章考虑的问题是在给定查询词的情况下，在 Twitter 中检索相关观点。

另外的相关工作是 Twitter 的情感倾向性分析，主要分析观点 tweet 是正面的肯定态度，还是负面的否定态度。Davidov 等人利用机器学习的方法对给定 tweet 进行情感倾向性判定[135]，他们利用 hashtag 和表情符号自动构造了大量的训练样本，以此提高分类效果。Barbosa 和 Feng 从网站 *Twendz*[4], *Twitter Sentiment*[5] 和 *TweetFeel*[6]中收集大量的训练数据进行情感倾向性判定[136]，Jiang 等人通过结合话题相关的特征和 tweet 所在上下文信息，改善了话题相关的 tweet 的情感倾向性分析[131]。其他还有大量的相关工作从不同的角度分析和提高 tweet 情感倾向性分析的效果[47, 137–155]。

情感分析的目的是识别文本态度的倾向性，我们的工作考虑的是如何在 Twitter 中找到观点性的 tweet，不论观点是正面的还是负面的。

3.2.2 TREC 观点检索

博客数据中的观点检索第一次在 TREC 2006 中组织评测，然后 TREC2007 与 TREC2008 也分别进行了相同任务的评测[124, 125, 125]。大多数参与评测的机构都采用两阶段的方法，首先在数据中检索到与查询词话题相关的文档集合，然后对文档集合中的文档根据各种不同的观点化度量方法进行文档的重排序，以此获得相关文档。

[3] http://en.wikipedia.org/wiki/Dow_Jones_Industrial_Average
[4] http://twendz.waggeneredstrom.com/
[5] http://twittersentiment.appspot.com/
[6] http://www.Tweetfeel.com/

另外还有一些观点检索的工作。Eguchi 和 Lavrenko 第一次提出了观点检索的问题[156]，他们结合了话题相关模型与情感倾向性相关模型，以此组合成一个整体模型对文档进行排序，这种方法的有效性在 MPQA 语料中得到验证。Zhang 和 Ye，Huang 和 Croft 也提出了他们自己的整合话题相关与情感倾向性相关的模型，以此检索观点[121, 157]。Gerani 等人第一次将排序学习的方法引入观点检索中[158]。进一步的工作还从不同的角度进行相关研究[122, 158–176]。以上所有的工作针对的数据不是博客就是 Web 网页，我们考虑的问题是 Twitter 中的观点检索，这种新的社交媒体存在大量的社交媒体特征，需要将此考虑到检索模型当中。

观点检索任务的难点在于如何度量文档的观点化程度。He 等人基于主观词的出现来评价文档的观点化程度[120]。这些主观词自动地由标注语料产生，他们将观点化的相关文档视为主观文档，话题相关的文档视为客观文档，然后词的主观化程度可以通过度量词在两个文档集合中的分布差异性进行度量。Amati 等人采用类似的方法自动构造观点词来帮助检索[177]。Seki 和 Uehara 利用结合邻近词依赖关系的统计语言模型来度量文档的观点化程度进行排序[172]。Jijkoun 等人提出了一种依赖语法规则自动标注数据，并以此自动构造话题相关主观词的方法[178]。Li 等人提出了一种基于话题词与情感词配对评价文档观点化程度的观点检索方法[167]。以上所有的方法都依赖人们手工标注的语料构造情感词典。

不同于上面的方法，Zhang 等人利用网站 RateitAll.com 的评论信息和其他网页作为近似主观句（‘pseudo’ subjective sentence-PSS）的资源，然后利用维基百科的页面作为近似客观句（‘pseudo’ objective sentence-POS）的资源[119]。他们假设主观性句子应该在评论信息中占据主要部分，客观性句子可以被忽略，而维基百科的客观性句子占据比例恰好相反。接着他们利用近似主观句集合和近似客观句集合构造 SVM 句子主观性分类器，这个分类器可以对新句子进行观点化度量，最后他们利用这个观点化度量与句子话题相关程度综合对文档进行排序。本章中，我们采用类似的方法，利用社交媒体特征与 tweet 文本的结构化信息自动生成近似主观化 tweet（‘pseudo’ subjective tweet-PST）和近似客观化 tweet（‘pseudo’ objective tweet-POT），以此帮助 Twitter 中的观点检索。

3.3 基于排序学习的 Twitter 观点检索框架

为了构造一个能够帮助 Twitter 观点检索的模型，我们将研究社交媒体特征与 tweet 观点化特征对于这个任务的有效性。我们将设计一系列特征，并将其应用到排序学习模型的场景中（Learning to Rank）[112]。

3.3.1 Twitter 观点检索排序学习框架

正如 2.4.1所介绍的，排序学习是一种将特征有效地整合到排序模型的机器学习算法。我们将一系列与 Twitter 观点检索相关的特征引入此框架，话题对应的 tweet 分别人工标记是否为相关 tweet。我们使用 RankSVM 进行排序模型的学习[179]。

3.3.2 Twitter 观点检索相关特征

第一章介绍了本文主要的两个研究问题：（1）人们在 Twitter 中如何用自然语言描述话题和表达观点？（2）以 Twitter 为代表的社交媒体有何新特点？如何利用这些特点帮助获取信息？因此，我们从两个方面设计 Twitter 观点检索相关特征：

1. **社交媒体特征（Social Feature）**：主要涉及一些 Twitter 特有的属性特征和与 tweet 作者有关的用户信息。

2. **观点化特征（Opinionatedness Feature）**：主要涉及如何评估 tweet 的观点化程度。

社交媒体特征主要是想通过分析 Twitter 新媒体的特点，发现哪些社交媒体新特征与观点化的 tweet 存在关系，以此帮助 Twitter 观点检索。比如，用户的粉丝数目、认证用户等，是否会影响 tweet 的观点化偏置。Tweet 的观点化评价（即观点化特征）是观点检索的重要部分，我们的目的是通过分析人们在 Twitter 中如何表达观点及其特定的文本表现形式，以此提出方法来帮助 tweet 观点化的判定，最终提高 Twitter 观点检索的效果。

下两节我们将详细介绍这些特征是如何设计并应用到 Twitter 观点检索当中的。

3.4 社交媒体特征

为了分析哪些 Twitter 新媒体特点与 tweet 观点化存在潜在的关系，我们从 Twitte 社交媒体特征的两个方面进行考虑：Twitter 特定的属性和 tweet 的用户信息。

3.4.1 Twitter 特定特征

人们在发布 tweet 时经常有一些习惯，这些习惯可能与 tweet 的观点化程度相关：

1. **链接（URL）**：Twitter 中分享链接是非常普遍的。大多数包含链接的 tweet 都是首先给出一个客观描述性的文本，然后紧接着给出一个与该文本相关的链接（例如，BBC News 发布的 tweet）。另外，Twitter 中大多数的垃圾 tweet 都包含链接。因此，我们设计了链接特征，用来描述给定 tweet 中是否存在链接。如果 tweet 中存在链接，则该特征值为 1，否则为 0。

2. **提及（Mention）**：在一个 tweet 中，人们经常使用符合“@username”来提及某人或回复某人。这种类型的 tweet 往往都是一些个人消息。以前的研究发现个人消息比官方消息更有可能包含观点[169]，所以我们用一个布尔特征来描述一个 tweet 中是否包含“@username”，并以此帮助 Twitter 中的观点检索。

3. **主题词（Hashtag）**：hashtag 是由“#”符合与一个词组合而成，它主要用来表示 tweet 的主题。我们设计布尔特征来描述 tweet 中是否包含 hashtag。

4. **Tweet 发布时间（Recency）**：Twitter 实时地产生大量的 tweet 流，直觉上如果发布的 tweet 离用户查询时间越接近，则越有可能是相关 tweet，因此我们设计了 tweet 发布时间到用户输入查询词的时间差 (以秒计) 的特征。

3.4.2 用户特征

Twitter 的用户和用户之间的关系是一个典型的社交网络，这些丰富的用户信息可以用来帮助 Twitter 的观点检索：

1. **发布数目（Statuses）**：Twitter 用户以往发布 tweet 的多与少反映了用户的活跃程度。直觉上，如果一个用户发布大量的 tweet，他很有可能是垃圾用户。因此我们设计了特征 tweet 的作者发布 tweet 数目帮助 Twitter 观点检索。

2. **粉丝数目和朋友数目（Followers and Friends）**：在 Twitter 中用户可以关注任意数量他感兴趣的其他用户。如果 *userA* 关注 *userB*，则 *userB* 发布的所有 tweet 用户 *userA* 都能在自己的页面中自动显示。我们将 *userA* 称为粉丝 (Follower)，将 *userB* 称为朋友（Friend)。用户的粉丝数目决定了用户受欢迎的程度，例如，新闻媒体用户往往比普通用户拥有更多的粉丝。同样，用户的朋友数目也反映了用户的特点，例如，大多数垃圾用户的朋友数目都远远大于其粉丝的数目。我们设计了这两个特征帮助 Twitter 观点检索。

3. **分组数目（Listed）**：一个用户可以将其朋友按照一定的标准（例如，兴趣和社会关系）划分到几个组中。如果一个用户被分到多个组中，则说明很多

人都对这个用户的 tweet 感兴趣。我们用一个特征来记录给定 tweet 的作者被分到不同组的数目，以此帮助 Twitter 观点检索。

3.5 观点化特征

对于观点检索任务来说，评估一个 tweet 的观点化程度显然是不可或缺的。以前对于评估文档观点化程度的方法大致可以分成两类[160]：

1. 基于机器学习的分类方法。

2. 基于情感词典的方法。

我们采用基于情感词典的方法，因为这种方法简单且不需要利用机器学习技术。但是，常用的情感词典，如 MPQA Subjectivity Lexicon[7]，可能对于 Twitter 来说不适用，因为 tweet 的文本很短且非正式，存在大量的情感表达方式不出现在情感词典中。因此，我们利用语料来构造适应 Twitter 的情感词典。通过计算 tweet 中特定的词，帮助评估 tweet 文本的观点化程度。我们利用手工标注的主观 tweet 集合和客观 tweet 集合，计算词的 chi-square 分值，此分值可以评估词的观点化权重。chi-square 分值是一种词对于主观 tweet 集合或客观 tweet 集合关联程度的度量方法。对于一个 tweet，我们只保留那些 chi-square 分值不低于 m 的词。对于 tweet d 的观点化程度，评价函数是：

$$Opinion_{avg}(d) = \sum_{t \in d, \chi^2(t) \geq m} p(t|d) \cdot Opinion(t)$$

其中 $p(t|d) = c(t,d)/|d|$ 是词 t 在 tweet d 中的相对词频，$c(t,d)$ 是该词在 tweet d 中的词频，$|d|$ 表示 tweet d 所包含的词数。

$$Opinion(t) = sgn(\frac{O_{11}}{O_{1*}} - \frac{O_{21}}{O_{2*}}) \cdot \chi^2(t)$$

这里，$sgn(*)$ 是一个符号函数，$\chi^2(t)$ 是计算词 chi-square 值的函数。

$$\chi^2(t) = \frac{(O_{11}O_{22} - O_{12}O_{21})^2 \cdot O}{O_{1*} \cdot O_{2*} \cdot O_{*1} \cdot O_{*2}}$$

表 3.1中，O_{ij} 是用来表示在主观 tweet 集合或客观 tweet 集合中存在多少条 tweet 包含词 t 或不包含词 t，例如，O_{12} 表示在主观 tweet 集合中不包含词 t 的 tweet 数目。

[7]下载地址：http://www.cs.pitt.edu/mpqa/

表 3.1　pearson's chi-square 表

	t	¬*t*	Row total
Sub. set	O_{11}	O_{12}	O_{1*}
Obj. set	O_{21}	O_{22}	O_{2*}
Col. total	O_{*1}	O_{*2}	O

$O_{1*} = O_{11} + O_{12}$；$O_{2*} = O_{21} + O_{22}$；$O_{*1} = O_{11} + O_{21}$；$O_{*2} = O_{12} + O_{22}$；$O = O_{11} + O_{12} + O_{21} + O_{22}$

大规模手工标注主观 tweet 和客观 tweet 非常耗时耗力，且 tweet 是否存在观点还与领域有关[131–133]。例如，与话题“android”相关且存在观点的 tweet 可以用词“*open*”、“*fast*”、“*excellent*”进行描述，但是这些词不可能是对某些特定事件的观点描述（例如，“UK strike”）。另外，我们也不可能对于每个话题人工标注大量的主观 tweet 和客观 tweet。基于以上我们提出了一种自动收集近似主观化 tweet（'pseudo' subjective tweet-PST）和近似客观化 tweet（'pseudo' objective tweet-POT）的方法。

在 Twitter 中一些简单的 tweet 文本结构化信息和用户信息可以帮助自动收集大量的 PST 和 POT。例如，人们经常在转发其他人发布的 tweet 前面加上自己的评论信息，这种结构的 tweet 大部分是主观化的 tweet。另外，许多新闻媒体用户喜欢发布一段客观描述的文本加上链接的 tweet，这些新闻媒体用户一般曾经发布过大量的 tweet 且粉丝众多。我们定义近似主观化 tweet（'pseudo' subjective tweet-PST）和近似客观化 tweet（'pseudo' objective tweet-POT）如下：

1. **近似主观化 tweet（'pseudo' subjective tweet-PST）**：一种 tweet，它在转发 tweet 的字符串“RT @username”前面存在其他文本。例如，tweet“*I thought we were isolated and no one would want to invest here! RT @BBCNews: Honda announces 500 new jobs in Swindon* `bbc.in/vT12YY`”是一个 PST。

2. **近似客观化 tweet（'pseudo' objective tweet-POT）**：如果一个 tweet 满足如下条件：

   a：包含链接；

   b：tweet 的作者曾经发布过大量的 tweet 且粉丝众多。

   这种 tweed 称为 POT。例如，tweet“*#NorthKorea:#KimJongil died after suffering massive heart attack on train on Saturday, official news agency reports* `bbc.in/vzPGY5`”。

利用上面的定义，我们可以简单地构造一些规则并从 Twitter 中收集大量的 PST 和 POT。我们假设在收集的 PST 集合中，所有的 tweet 都是主观性的文本，

而收集的 POT 集合中，所有的 tweet 都是客观性文本。虽然这并不是 100% 准确，但是，PST 集合中主观性的 tweet 必定占据绝大部分，使得客观性的 tweet 可以被忽略，而 POT 集合情况则恰好相反。另外，文本的结构化信息和用户信息与 tweet 的话题无关，因此如果能有大量话题关联的 tweet，则可以从中收集大量话题相关的 PST 和 POT。

3.6 Twitter 观点检索实验

本章我们将上面讨论的特征对于 Twitter 观点检索的效果进行实验验证。

3.6.1 Twitter 观点检索实验数据

据我们所知，目前还没有关于 Twitter 观点检索的评测数据。为了从 Twitter 中找到给定话题的话题相关 tweet 且其包含对此的观点，我们自己构造了 Twitter 观点检索数据。

我们在 2011 年 11 月利用 Twitter API 爬取了 3000 万个 tweet，所有的 tweet 都是英文的[8]。我们使用 Lucene[9]对所有 tweet 建立索引并构造搜索引擎。7 个用户使用了该搜索引擎，7 人中包括 6 名男性和 1 名女性，所有的用户都不是母语为英语的人，但是他们的英语都有很好的基础。这些用户可以在搜索引擎中提交他们感兴趣的任意查询词，给定查询词，搜索引擎将根据 BM25 返回 100 个得分最高的 tweet。我们使用 Lucene-BM25[10] 计算每个 tweet 对于查询词的 BM25 分值。在计算 BM25 分数时，相关参数都使用默认设置（$k_1 = 2$；$b = 0.75$)。最后所有的查询词都是在 2011 年 12 月 1 日提交的。

根据 tweet 是否与查询词话题相关且包含观点的定义，提交查询词的用户对返回的 tweet 进行标注，如果 tweet 与查询词话题相关且包含观点，则这个 tweet 的类别分值标为 1，否则标为 0。我们在这里只考虑 tweet 是否包含观点，并不考虑观点的情感倾向性，另外，tweet 仅仅话题相关或只包含话题无关的观点，我们都将其视为不相关 tweet。

最终我们搜集了 50 个查询词和 5000 个标注数据，其中每个查询词对应 100 个 tweet。我们将这 50 个查询词分成 6 个类别，表 3.2给出了这 6 个类别的查询词及其相关 tweet 数量。另外，所有查询词的平均长度为 1.94 个词，每个查询词平均相关 tweet 为 16.62 个。

[8]我们利用语言检测工具过滤非英语 tweet，下载地址：http://code.google.com/p/language-detection/

[9]下载地址：http://lucene.apache.org

[10]下载地址：http://nlp.uned.es/~jperezi/Lucene-BM25/

我们进一步考虑了这些标注数据的可靠性。对于每一个查询词，我们随机选取了 10 个对应的 tweet 叫另外两个标注者进行相关标注。这两个标注者标注数据的 kappa 分数为 0.54，这说明我们的 Twitter 观点检索数据具有“好”的可靠性。

3.6.2 Twitter 观点检索实验设置

我们将验证上两节介绍的特征对于 Twitter 观点检索的作用。对于排序学习，我们使用了 SVM Rank 工具[11]。在排序学习模型的训练与测试过程中，我们使用了线性核函数。另外，所有结果都是将模型参数调到最优。为了解决测试数据规模较小造成结果偏置的问题，我们使用了 5 次交叉验证的方法。每次折叠过程，我们都将数据分成训练数据、测试数据、验证数据，其中训练数据包含 30 个查询词，而测试数据和验证数据各包含 10 个查询词。对于评价指标，我们依然使用平均准确率（*Mean Average Precision-MAP*），因为这个评价指标在 TREC 的检索任务中被经常使用且具有很好的排序算法效果区分性与稳定性[180]。

3.6.3 基准系统（Baseline）

目前据我们所知，Twitter 中的观点检索还没有相关的工作，因此我们只能通过选择相近的工作（比如上一章的 Twitter 信息检索的方法）作为比较。但是由于我们的研究目的是分析哪些因素能够影响 Twitter 中观点检索的效果，因此在基准系统的选取上，我们选择了简单的基于话题相关的模型（如 BM25）。一般地，与多个基准系统比较检索效果能够更好地验证检索方法的有效性，因此在我们的实验中，我们采用了两个基准系统：

1. **BM25**：利用优化的 Okapi BM25 对 tweet 进行打分，然后排序。
2. **VSM**：利用向量空间模型（Vector Space Model）对 tweet 与查询词进行话题相关性打分，然后排序[130]。

在信息检索的排序算法中，Okapi BM25 是其中最著名的基于概率的评估查询词与文档内容相关性的算法。给定查询词 q，t 为查询词中的词，一个文档 d 的 BM25 分数为：

$$BM25(q,d) = \sum_{t \in q} log(\frac{N - df_i + 0.5}{df_i + 0.5} * \frac{(k_1 + 1)(tf_i)}{k_1((1-b) + b\frac{dl}{avdl}) + tf_i})$$

其中 tf_i 表示词 t 的词频，df_i 表示词的文档频率，N 表示数据集合中的文档数量，dl 是文档长度（词数），$avdl$ 是文档集合中所有文档的平均长度。两个参数 k_1

[11]下载地址：http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

表 3.2　Twitter 观点检索查询词及其相关 tweet 数目

组织	产品	人物
pixar, 23	Mac book pro, 37	Jennifer Aniston, 40
manchester city, 21	iphone4s, 32	chris paul, 39
htc, 19	kinect, 30	Obama, 35
Syria, 17	itouch, 25	bill gates, 16
iran, 15	kindle fire, 16	Maggie Q, 13
Manchester United, 15	iOS5 Jailbreak, 12	owl city, 12
disney, 12	galaxy note, 11	paul graham, 10
Lenovo, 10	Xbox 360, 7	steve jobs, 9
microsoft, 6	google venture, 7	Kai-Fu Lee, 1
Calvin Klein, 5	new Google Bar, 5	
fossil, 5	EA Daily Deals, 0	
intel, 3	SIEMENS fridge freezer, 0	
channel, 0		

其他	事件	电影、电视剧
job hunting, 79	iran nuclear, 35	Breaking Dawn, 49
speech recognition, 15	American Music Awards, 25	big bang, 41
machine learning, 4	UK embassy, 21	Two And A Half Men, 35
new start-ups, 2	UK strike, 7	inside job, 2
immigrate to canada, 1	ARTIST OF THE YEAR, 7	
systems biology, 0		
text mining, 0		

和 b 分别决定了词频和文档长度的归一化对于分数的影响。以下的实验中，我们将利用验证数据集对参数 k_1 和 b 进行优化，以此得到最优的基于 Okapi BM25 的 Twitter 观点检索排序算法。

因为 tweet 是典型短文本，而 BM25 算法对于文本长度的变化十分敏感，因此我们引入了另一种简单的信息检索排序算法 VSM。VSM 最早由 Salton 等人提出[130]。VSM 算法中，文档 d_j 和查询词 q 可以分别表示成两个 t 维的词向量 $d_j = (w_{1,j}, w_{2,j}, ..., , w_{t,j})$ 和 $q = (w_{1,q}, w_{2,q}, ..., , w_{t,q})$，其中 t 表示文档集合中不同词数目，每一维表示一个词，有许多方法可以决定词的权重 w，我们使用经典的 tf-idf 度量词的权重：

$$w_{t,d} = tf_{t,d} * log(\frac{|D|}{|d' \in D|t \in d'|})$$

$tf_{t,d}$ 是词 t 在文档 d 中的词频，$log(\frac{|D|}{|d' \in D|t \in d'|})$ 是倒转文档频率，$|D|$ 是文档集合中文档数量，$|d' \in D|t \in d'|$ 文档集合中包含词 t 的数量。文档 d_j 和查询词 q 的相似度可以通过计算两个向量的余弦夹角得出：

$$sim(d_j, q) = \frac{\sum_{i=1}^{N} w_{i,j} * w_{i,q}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2} * \sqrt{\sum_{i=1}^{N} w_{i,q}^2}}$$

我们利用 Lucene 实现 VSM 的计算。

3.6.4 Twitter 观点检索实验结果及分析

我们将实验分成以下几个方面进行讨论：

1. 社交媒体特征能否帮助 Twitter 中的观点检索？（见 3.6.4.1）
2. 观点化特征能否帮助 Twitter 中的观点检索？（见 3.6.4.2）
3. 基于话题相关的 PST 和 POT 能否比话题不相关的 PST 和 POT 更好地帮助 Twitter 中的观点检索？（见 3.6.4.3）
4. PST 和 POT 能否帮助观点化 tweet 的识别？（见 3.6.4.4）
5. 那种 Twitter 观点检索模型达到最佳效果？（见 3.6.4.5）

另外，表 3.3 简要概述了 Twitter 观点检索所使用的相关特征。

3.6.4.1 社交媒体特征实验结果

我们首先验证社交媒体特征是否能够帮助 Twitter 中的观点检索。我们将各种社交媒体特征与基准系统中的 BM25 或 VSM 结合使用形成排序模型，表 3.4和表 3.5显示了各种排序模型在 Twitter 观点检索任务中的排序效果。我们发现链接特征（URL）和粉丝数目特征（Followers）与 BM25 特征结合使用时能够比 BM25 基准系统显著提高[12]Twitter 观点检索效果（见表 3.4）。虽然提及特征（Mention）和发布数目特征（Statuses）能够提高 MAP 值，但是它们并不是显著性地提高。而表 3.5显示，链接特征（URL）、提及特征（Mention）、发布数目特征（Statuses）能够显著提高检索任务的效果。

以上实验结果表明一些社交媒体属性能够帮助 Twitter 中的观点检索。特别是链接特征（URL）对于该任务特别有效，这可能是大多数包含链接的 tweet 都是客观性的描述文本，同时大多数垃圾信息也包含链接，使用该特征能够有利于减少

[12]我们使用 t-test 进行显著性测试。

表 3.3　Twitter 观点检索特征概况

基准系统特征（Baseline Features）	取值范围	描述
BM25	$(0, +\infty)$	tweet 的 BM25 分值
VSM	$(0, +\infty)$	tweet 的 VSM 分值
社交媒体特征（Social Media Features）	**取值范围**	**Description**
链接（URL）	0 or 1	tweet 中是否包含链接
提及（Mentions）	0 or 1	tweet 是否包含提及
主题词（Hashtag）	0 or 1	tweet 是否包含 hashtags
tweet 发布时间（Recency）	$N^+ = \{1, 2, 3, ...\}$	tweet 发布时间到用户输入查询词的时间差 (以秒计)
发布数目（Statuses ）	$N^+ = \{1, 2, 3, ...\}$	tweet 的作者以往发布 tweet 的数目
粉丝数目（Followers）	$N = \{0, 1, 2, ...\}$	tweet 的作者的粉丝数目
朋友数目（Friends）	$N = \{0, 1, 2, ...\}$	tweet 的作者的朋友数目
分组数目（Listed）	$N = \{0, 1, 2, ...\}$	tweet 的作者的分组数目
观点化特征（Opinionatedness Features）	**取值范围**	**Description**
MPQA 词汇（MPQA_Lexicon）	0 or 1	tweet 是否包含主观词典 MPQA 中的词或短语
TwitterSenti	0 or 1	由 Twitter Sentiment API 对 tweet 的观点性打分
Gold	$(-\infty, +\infty)$	认为标注 tweet 是否是观点
Q_I	$(-\infty, +\infty)$	由话题不相关的 PSTs and POTs 数据对 tweet 的观点性打分
Q_D	$(-\infty, +\infty)$	由话题相关的 PSTs and POTs 数据对 tweet 的观点性打分

垃圾信息。链接特征（URL）、发布数目特征（Statuses）、粉丝数目特征（Followers）对于 Twitter 中观点检索的有效性也再次证明了我们利用社交媒体信息与 tweet 文本结构化信息帮助收集近似主观化 tweet（PST）和近似客观化 tweet（POT）的合理性。而提及特征（Mention）在观点排序上的有效性同样也验证了普通个人发布的 tweet 比官方发布的 tweet 更有可能包含观点[169]。令我们惊讶地是 tweet 发布时间（特征 Recency）对于 Twitter 中的观点检索没有帮助。我们认为可能的原因是所有的查询词都是在一个月的 tweet 收集完以后提交的，所以造成大多数的查询词都不涉及新事件（见表 3.2），因此 tweet 的实时性对于参与数据标注的用户来说并不是最重要的。

表 3.4　基于社交媒体特征的 Twitter 观点检索系统实验结果（BM25）

	MAP
BM25	0.2831
BM25+URL	0.3305▲
BM25+Mention	0.2920
BM25+Hashtag	0.2734
BM25+Recency	0.2576
BM25+Statuses	0.2931
BM25+Followers	0.2946△
BM25+Friends	0.2799
BM25+Listed	0.2822

△ 和 ▲ 分别表示排序结果显著高于 BM25 观点检索系统（$p < 0.05$ 和 $p < 0.01$）。

表 3.5　基于社交媒体特征的 Twitter 观点检索系统实验结果（VSM）

	MAP
VSM	0.2812
VSM+URL	0.3171▲
VSM+Mention	0.2932△
VSM+Hashtag	0.2803
VSM+Recency	0.2757
VSM+Statuses	0.2928△
VSM+Followers	0.2829
VSM+Friends	0.2808
VSM+Listed	0.2801

△ 和 ▲ 分别表示排序结果显著高于 VSM 观点检索系统（$p < 0.05$ 和 $p < 0.01$）。

3.6.4.2　基于话题不相关的观点化特征实验结果

接下来我们验证观点化特征对于 Twitter 观点检索效果的影响。这里为了自动产生近似主观化 tweet（PST）和近似客观化 tweet（POT），我们设计了一些简单的规则帮助识别这些 tweet:

1. 对于 PST，我们选择那种 tweet 中包含字符串“RT @username”且第一次出现这种字符串的位置[13]前面还存在长度不小于 10 个字符的文本。

2. 对于 POT，我们选择那种 tweet 中包含链接且其作者以前发布不少于 10000 个 tweet 还要粉丝数目不小于 1000。

[13]一个 tweet 中可能包含多个字符串“RT @username”。

在我们收集的 2011 年 11 月 3000 万个 tweet 中，总共有 4.64% 的 tweet 是 PST，1.35% 的 tweet 是 POT。

我们请了一位标注者对上面方法自动生成的 tweet 进行主观化 tweet 的质量检测。检测中随机抽取了自动生成的 100 个 PST 和 100 个 POT，然后请标注者标注那个是主观化 tweet 那个是客观化 tweet。结果发现 95% 的 PST 是主观化 tweet，85% 的 POT 是客观化 tweet，实验结果证明了我们的方法可以自动的大规模生成高质量的主观化 tweet 和客观化 tweet。因此我们随机选择了 3000 个英文的 PST 和 POT 作为话题不相关的数据集，帮助情感词典构造。

在基于情感词典的方法计算 tweet 的观点化程度中，我们首先使用 Porter English Stemmer 和停用词表[14]预处理 tweet 文本。为了达到最佳效果，我们设定 chi-square 分值的阈值 m 为 5.02，这使得词在两个集合中显著差异水平达到 0.025，这个设定也与 Zhang 等人的工作一致[119]。我们将利用话题不相关的数据集计算 tweet 的观点化程度的特征称为 Q_I 。以前的工作中采用标注好的主观博客数据与客观博客数据对新博客进行主观性打分[120, 158]，我们的实验中也同样采取类似的方法，利用训练集中标注好的主观性 tweet 与客观性 tweet 对测试集中的新 tweet 进行主观性打分，我们称之为 Gold 特征。另外，我们还利用词典 MPQA Subjectivity Lexicon 对 tweet 进行主观化判定，称为 MPQA_Lexicon 特征，如果 tweet 中包含 MPQA Subjectivity Lexicon 中的词或短语，则 tweet 的观点化得分为 1，否则为 0。最后我们还设计了 TwitterSenti 特征，该特征利用 Twitter Sentiment API[15]对 tweet 的情感倾向性进行 tweet 的主观性评价[181]，如果 Twitter Sentiment API 判断 tweet 为无情感倾向的 tweet，则该特征的值为 0，否则为 1[16]。

表 3.6和表 3.7给出了各种观点化特征对于 Twitter 观点检索的影响。我们发现所有的观点化特征加入到两个基准系统中都能将检索效果显著提高，这说明评价 tweet 的观点化程度，对于 Twitter 中观点检索是必不可少的。我们也发现虽然利用 MPQA_Lexicon 特征可以提高 Twitter 观点检索的效果，但是提高的幅度远不如利用其他观点化特征的方法，可能的原因是 tweet 文本是一种明显区别于网页评论或博客的文本，这就使得从网页评论或博客文本中构造的情感词表并不适合 tweet 的主观化程度评估。最后我们发现利用 Q_I 特征在排序效果上可以达到与利用特征 Gold 相当的效果，且显著性测试上发现没有明显差别，这说明基于社交媒体信息与 tweet 文本结构化信息生成的 PST 和 POT 可以帮助 Twitter 中的观点检索，更重要的是这种方法不需要人为标注数据。

[14]我们使用标准的停用词词表，并且再加上“RT”字符串。

[15]网址：http://www.sentiment140.com

[16]这里强调的是不带情感倾向的 tweet 有时也可能是带观点的 tweet，这里我们将其忽略，以后的工作在进一步讨论。

表 3.6　基于观点化特征的 Twitter 观点检索系统实验结果（BM25）

	MAP
BM25	0.2831
BM25+MPQA_Lexicon	0.2895
BM25+TwitterSenti	0.3279▲
BM25+Gold	0.3739▲
BM25+Q_I	0.3792▲

△ 和 ▲ 分别表示排序结果显著高于 BM25 观点检索系统（$p < 0.05$ 和 $p < 0.01$）。

表 3.7　基于观点化特征的 Twitter 观点检索系统实验结果（VSM）

	MAP
VSM	0.2812
VSM+MPQA_Lexicon	0.2876
VSM+TwitterSenti	0.3244▲
VSM+Gold	0.3485▲
VSM+Q_I	0.3566▲

△ 和 ▲ 分别表示排序结果显著高于 VSM 观点检索系统（$p < 0.05$ 和 $p < 0.01$）。

3.6.4.3　基于话题相关的观点化特征实验结果

我们方法的另一个优点就在于它能够很容易的生成大量话题相关的 PST 和 POT。我们利用我们搜集的所有 PST 和 POT 建立索引构造搜索引擎。给定查询词，搜索引擎能够返回任意数量的话题相关的 PST 和 POT。因此，利用这个搜索引擎，我们针对 50 个查询词，构造了 50 个话题相关的 PST 和 POT 集合，每个集合都包含 3000 个 tweet。我们将利用话题相关 PST 和 POT 计算 tweet 观点化程度的特征称为 Q_D。表 3.8和表 3.9给出了这个特征与 Q_I 特征在 Twitter 观点检索中的实验比较结果，我们发现 BM25+Q_D 系统和 VSM+Q_D 的排序效果都优于 BM25+Q_I 系统和 VSM+Q_I，这说明我们的方法可以帮助 Twitter 观点检索中 tweet 的观点化程度评估与话题相关的问题。

表 3.10给出了由不同话题相关 PST 和 POT 与话题不相关 PST 和 POT 集合生成的高 $\chi^2(t)$ 得分情感词列表。我们可以看出我们的方法发现了一些人称代词，例

表 3.8　基于话题相关观点化特征的 Twitter 观点检索系统实验结果（BM25）

	MAP
BM25+Q_I	0.3792
BM25+Q_D	0.3907△

△ 和 ▲ 分别表示排序结果显著高于 M25+Q_I 观点检索系统（$p < 0.05$ 和 $p < 0.01$）。

表 3.9　基于话题相关观点化特征的 Twitter 观点检索系统实验结果（VSM）

	MAP
VSM+Q_I	0.3566
VSM+Q_D	0.3599

△ 和 ▲ 分别表示排序结果显著高于 VSM+Q_I 观点检索系统（$p < 0.05$ 和 $p < 0.01$）。

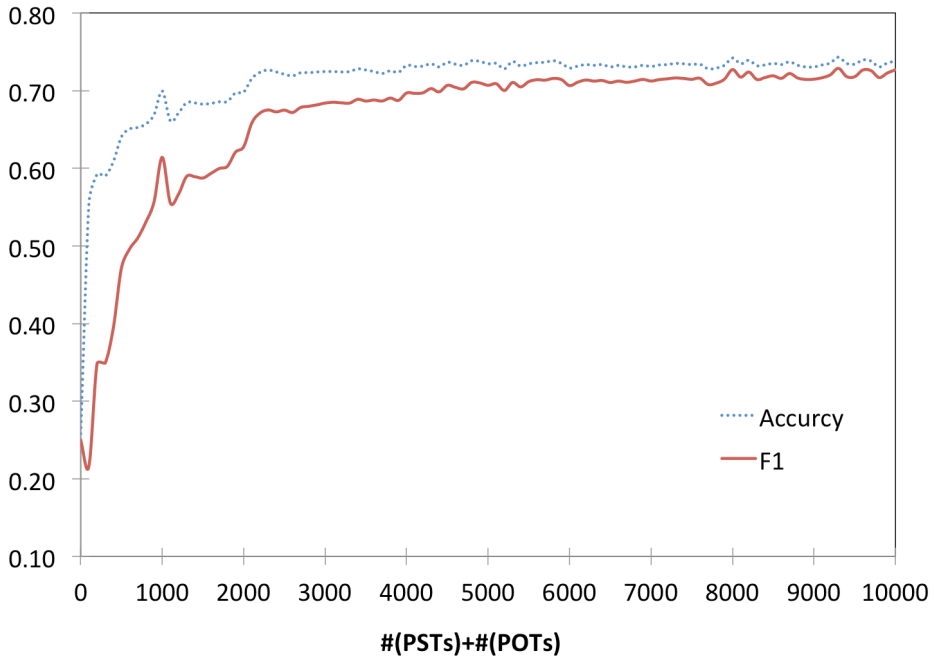

图 3.1　基于观点化特征的 Twitter 观点分类实验结果

如，“i”、“u”、“my”，也发现了一些表情符合，例如，“:)”、“:(”、“:d”。可能的原因是个人发布的 tweet 很有可能是主观性文本。对于话题相关的 PST 和 POT 集合，我们的方法则成功的发现了情感词“excit”（$Opinion(t) > 0$），该情感词与电影“Breaking Dawn”相关，且这个情感词不可能描述与其他话题（例如，“UK strike”）有关的观点。另外，我们还发现了一些 $Opinion(t) < 0$ 的词，例如，词“bbc”很有可能出现在一些由 BBC News 发布的关于话题“UK strike”的 tweet 当中。

3.6.4.4　基于观点化特征的 Twitter 观点分类实验结果

我们也对 tweet 观点化评价函数 Opinion_avg(d) 能否作为主观化 tweet 的分类器感兴趣（见 3.5）。我们利用 1000 个手工标注的主观化 tweet 和 1000 个客观化 tweet 作为测试集。我们将 Opinion_avg(d) 得分大于 0 的 tweet 视为主观化 tweet，小于或等于 0 的 tweet 视为客观化 tweet。正确率（Accuracy）和 F1 值作为分类器的评价指标。图 3.1给出了实验结果。我们发现当选取大约 2200 个 PST 和 POT tweet 时，正确率与 F1 值趋于稳定，分别是 0.72 和 0.67，再选取更多的 PST 和 POT 对

表 3.10　基于话题独立与话题相关观点化特征打分的高 $\chi^2(t)$ 分数的情感词

序号	Breaking Dawn	HTC	Obama	UK strike	Q_I
1	i +	i +	i +	... -	i +
2	video -	lol +	you +	i +	lol +
3	go +	.. +	#obama -	followfridai -	:) +
4	.. +	u +	my +	rank -	.. +
5	me +	my +	lol +	you +	u +
6	lol +	new -	u +	my +	* +
7	new -	:) +	!! +	lol +	new -
8	via -	me +	me +	week -	my +
9	!!! +	* +	barack -	last -	morn +
10	wait +	rezound -	#tcot -	:) +	me +
11	pattinson -	you +	... -	u +	!!! +
12	robert -	phone -	cont +	me +	good +
13	... -	like +	.. +	thi +	:d +
14	so +	:d +	presid -	so +	via -
15	too +	!!! +	* +	!! +	!! +
16	:) +	morn +	i'm +	#ows -	cont +
17	see +	i'm +	:) +	#jobs -	haha +
18	can't +	good +	we +	x +	ya +
19	:d +	!! +	do +	come +	too +
20	premier -	... -	he +	3 +	... -
21	kristen -	too +	obama' -	gener -	i'm +
22	excit +	cream -	!!! +	onli +	:(+
23	again +	cont +	#news -	good +	thank +
24	i'm +	so +	know +	bbc -	,) +
25	im +	thank +	lmao +	here +	@damnitstrue +

+ 为词的 *Opinion*(*t*) 分数不小于 5.02，– 为词的 *Opinion*(*t*) 分数不大于 -5.02。

于分类结果没有显著提高。这说明评价函数 Opinion_avg(d) 对于 tweet 的主观化分类是有效的。

3.6.4.5　最佳 Twitter 观点检索系统实验结果

最后我们将所有在前面的实验中能够帮助 Twitter 观点检索的特征都加入到基准系统中形成新的检索模型。这些特征包括：BM25（或 VSM）、链接（URL）、提及（Mention）、发布数目（Statuses）、粉丝数目（Followers）、Q_D。表 3.11和表 3.12给出了 Twitter 观点检索最好的模型 BM25_Best 和 VSM_Best 在 Twitter 观

表 3.11　基于全部与最佳特征的 Twitter 观点检索系统实验结果（BM25）

	MAP
BM25	0.2831
BM25_Best	0.4181▲
BM25_All	0.4128▲

△ 和 ▲ 分别表示排序结果显著高于 BM25 观点检索系统（$p < 0.05$ 和 $p < 0.01$）。

表 3.12　基于全部与最佳特征的 Twitter 观点检索系统实验结果（VSM）

	MAP
VSM	0.2812
VSM_Best	0.3761▲
VSM_All	0.3721▲

△ 和 ▲ 分别表示排序结果显著高于 VSM 观点检索系统（$p < 0.05$ 和 $p < 0.01$）。

点检索数据集上最好的实验结果，MAP 值分别比系统 BM25 和 VSM 提高 56.82% 和 33.75%。

另外，我们还将表 3.3中的所有特征整合到一个排序学习模型当中[17]。这里的两个排序模型分别称为 BM25_All 和 VSM_All。表 3.11和表 3.12给出了这两个模型的排序结果，它们的排序能力稍微弱于 BM25_Best 和 VSM_Best 系统，但没有显著性差异。

以上结果表明，对于 Twitter 中的观点检索来说，仅仅使用特征 BM25（或 VSM)、链接 (URL)、提及 (Mention)、发布数目 (Statuses)、粉丝数目 (Followers)、Q_D 就能达到最佳效果。

3.6.5　Twitter 观点检索实验数据偏置分析

因为目前还没有关于 Twitter 观点检索的数据，所以我们构造并标注了自己的数据集。但是仅仅使用 BM25 算法对每个查询词所对应的 tweet 进行排序，然后收集前 100 个 tweet 进行相关判定可能造成其它研究算法无法在这个数据集上实验验证。另外，用 BM25 算法收集的数据集可能由于该算法本身的问题造成选取数据的偏置。基本上由 BM25 算法选取的部分标记是否相关的文档去验证其它算法，有效性往往都无法很好的验证，这是因为所有排名 100 以外的未进行相关性判定的 tweet 都当做了不相关 tweet，这也是 TREC 使用 2 个以上的检索算法返回结果形成“池”(Pool)，且每个算法返回 1000 个或更多排序靠前文档进行数据相关性

[17]这里我们只使用 Q_D 评价一个 tweet 的观点化程度，因为它在前面的实验中对 tweet 的观点化评估效果最好。

判定的原因。基于以上分析，我们利用了 TREC2011 年的 Twitter 信息检索任务的数据和其相关判定的结果来评价我们 Twitter 观点检索系统的有效性[18]。

TREC2011 年的 Twitter 数据由 1600 万个 tweet 构成，这个数据从 2011 年两个星期的所有 tweet 数据中采样得到[134]。总共有 59 个来自世界各地的研究组织参加了 Twitter 信息检索任务的评测，并且提交了他们自身排序算法所得到的 tweet 排序结果。在 TREC2011 的 Twitter 信息检索数据集中总共包含 49 个话题（查询词）。每个研究组织被要求针对话题从 1600 万个 tweet 中检索 30 个 tweet 进行评价。最后返回结果所形成的 tweet“池”有 50324 个 tweet，其中 2965 个 tweet 被判定成话题相关 tweet[19]。但是，这些被判定为话题相关的 tweet 仅仅适合评价 Twitter 信息检索算法的有效性，而我们的目的是在 Twitter 中进行观点检索。因此，我们重新再手工标注了部分 TREC2011 年的 Twitter 数据，判定哪个是话题相关且包含观点的 tweet。我们利用 Metzler 等人提交的 Twitter 信息检索返回的结果重新标记 (他们提交数据的 ID 是 isiFDL)[97]，选择 Metzler 等人提交的 tweet 是因为他们所提交的 tweet 返回结果是目前 TREC2011 年的 Twitter 信息检索任务最好的结果[134]，它包含了最多的话题相关的 tweet。最后，我们重新标注的数据集中有 98 个话题相关且包含观点的 tweet。

我们的方法是对 Twitter 中的数据进行观点检索，我们利用重新标注的 TREC2011 年的 Twitter 数据进行 Twitter 观点检索的实验验证，验证的方法是利用我们的系统对新标注的 TREC 数据进行重排序。我们将 Metzler 等人提出的 Twitter 信息检索系统作为基准系统[97]，称为 isiFDL。而我们利用链接（URL），提及（Mention），发布数目（Statuses），粉丝数目（Followers），Q_D 特征构造的系统称为 isiFDL_Best，这里强调的是我们用算法 isiFDL 对 tweet 的话题相关性得分替换我们原先使用的 BM25 算法。我们同样使用 5 次交叉验证实验验证算法的有效性。表 3.13给出了实验结果。我们可以看到 isiFDL_Best 系统对与 Twitter 观点检索的排序效果显著优于 isiFDL 系统。这说明我们的观点检索系统在重新标注的 TREC2011Twitter 数据上依然有效。

显然，虽然重新标注的数据最大限度地减小了话题相关的偏置问题，但是依然没有解决在观点化相关上的偏置问题。我们也不可能像 TREC 的方式一样组织多个研究机构提交 Twitter 观点检索的数据构造 Twitter 观点检索的标准数据。未来如果有像 TREC 一样的方式构造的 Twitter 观点检索数据，我们将再进一步验证我们方法的有效性。

[18]下载地址：http://trec.nist.gov/data/tweets/
[19]下载地址：http://trec.nist.gov/data/microblog/11/microblog11-qrels

表 3.13　观点检索系统在 TREC Tweets201 数据上的实验结果

	MAP
isiFDL	0.1639
isiFDL_Best	0.2181△

△ 和 ▲ 分别表示排序结果显著高于 isiFDL 观点检索系统（$p < 0.05$ 和 $p < 0.01$）。

3.7　小结

据我们所知，我们是第一个提出如何在 Twitter 中进行观点检索的研究机构。我们的方法是利用社交媒体信息与 tweet 的观点化信息帮助检索。实验结果表明当检索模型中考虑链接（URL）、提及（Mention）、发布数目（statues）、粉丝数目（statues）、tweet 文本的观点化程度等特征时，可以帮助提高 Twitter 观点检索效果。

另外，我们还提出了一种利用社交媒体信息和 tweet 文本结构化信息帮助生成近似主观化 tweet（PST）和近似客观化 tweet（POT）的方法，以此提高评估 tweet 观点化的程度，实验发现将其引入观点检索中可以替代传统手工标注数据的方法，且考虑话题相关因素构造 PST 和 POT 时，效果更好。

第四章　Twitter 中传播观点的发现

4.1　引言

以往的传播分析中有一个假说：“如果人们觉得自己的观点是公众中的少数派，他们将不愿意传播自己的看法；而如果他们觉得自己的看法与大多数人一致，他们就会勇敢地说出来”[182]。以前的大众媒体通过关注多数派的观点来营造一种意见环境，因此影响和制约舆论，舆论的形成不是社会公众理性讨论的结果，而是舆论环境的压力作用于人们惧怕孤立的心理，强制人们对优势意见采取趋同行为[183]。而以 Twitter 为代表的社交媒体则部分打破了这种环境，人们在这些平台上面实现相对平等的信息传播、意见的交流与交锋、讨论协商公共事务等。社交媒体强大的互动性、进入的低门槛、以及人人享有表达观点等的优点，彻底打破了以往传统媒体对话语权。随着以 Twitter 为代表的社交媒体的迅速发展，越来越多的人在这个社交媒体上分享消息，发表自己对人，事件，产品等的看法。实际上 Twitter 已经变成了一个巨大的观点库，不仅可以帮助人们做出决策，还可以帮助政府与公司收集人们对政策或产品的评价，以此做出相应的调整。

但是，这种观点库因为 tweet 数据量的庞大和质量的显著差异使其很难被有效利用。首先，由于 Twitter 拥有超过 5 亿的用户并且每天产生多于 3.4 亿个 tweet[1]，造成用户面临数据过载的问题；其次，观点的重要性往往差别较大，这就造成一些应用中用户可能忽略某些观点。例如，以下是两个关于话题“Obama”的 tweet，并且这两个 tweet 都包含观点：

(a) *“RT@KG_NYK: The fact that Obama “lost” the debate b/c he didnt call Romney’s lies out well enough is pretty harrowing commentary on surf ”*

(b) *“MyNameisGurley AND I HATE OBAMA*

用户一般会认为 Tweet（a）比 Tweet（b）重要，因为 Tweet（a）介绍了 Obama 竞选美国总统的第一次电视辩论的基本情况，并且给出了相关的观点，而 tweet（b）只是对 Obama 的一个一般性观点，对大多数用户来说，没有多少价值。另外，Tweet（a）是对用户 @KG_NYK 观点 tweet 的转发，这也表明了 Tweet（a）同意用户 @KG_NYK 的观点。

评估观点的重要性是非常主观的，但是在 Twitter 中人们一般通过转发行为间接说明 tweet 的重要性[18]。这是基于人们在微博传播行为上的一个基本假设：当

[1] http://www.mediabistro.com/alltwitter/500-million-registered-users_b18842

人们认为一个 tweet 非常重要且值得和大家分享此消息时，他们将通过转发传播这个 tweet。基于此本章将研究如何在 Twitter 中进行传播观点检索（发现）[2]。不同于第二章的 Twitter 信息检索，本章的传播观点检索试图查找的 tweet 是包含观点的 tweet。另外，本章的问题也不同于第三章的 Twitter 观点检索，因为我们的目的不仅仅是找到话题相关的观点，更重要的是发现能够传播的观点。总的来说，相关的 tweet 必须满足三个条件：

1. 与查询词话题相关。

2. tweet 包含涉及查询词的主观化观点，但不考虑观点的态度是正面的还是负面的。

3. 这个 tweet 将会被转发。

Twitter 中传播观点的发现可以帮助政府、企业、用户发现高质量的观点，从应用的角度来说，传播性观点的发现，能够帮助政府部门迅速掌握舆论情况，发现问题，引导公共舆论或及时调整政策，维护社会的和谐与稳定。企业通过对传播性观点的分析，能够从中挖掘出消费者对本公司产品哪些地方不满意，对其他竞争公司的相同产品哪些地方比较送欢迎，以此改进本公司的产品质量与服务。从用户的角度来说，传播性的观点往往更加影响他们的决策与判定，因为传播性的观点往往会影响用户周围圈子的观点，而用户做出决策往往具有协同性。

但是 Twitter 中传播观点的发现充满了挑战，以前的研究发现预测一个 tweet 是否被转发与 tweet 的话题紧密相关，而话题可以通过一元的词向量来有效表示[18, 184]。而在这个任务中，因为初步检索返回的结果都是话题相关的 tweet，tweet 之间在话题上差别不大，所以同一话题下的 tweet 是否将被转发依赖于其他非话题因素，例如，tweet 是否可信，tweet 的文本质量是否较高，tweet 的写法是否规范等。以前的研究发现这些文本分析比基于话题的文本分析难度更大[185, 186]，另外，tweet 的文本还存在着文本短小的特点。但是 tweet 丰富的社交媒体信息，包括 tweet 特定的特点和用户信息，可以作为额外的辅助信息潜在地帮助 tweet 的文本分析，最终提高 Twitter 中传播观点检索（发现）的效果。

本章中我们依然用机器学习的方法学习排序模型，此模型利用一系列特征捕获 Twitter 中传播观点相关的信息。这些特征包括：tweet 的传播度特征、tweet 的观点化特征、tweet 的文本质量特征。其中 tweet 的传播度特征是一个 tweet 能够被转发的置信度；另外，我们继续利用前一章介绍的 tweet 观点化特征帮助传播观点

[2]本章将传播观点发现问题用检索的方法解决，需要强调的是分类方法同样可以解决此问题，但将其视为排序问题并不失一般性。

检索，这个特征利用社交媒体信息和文本的结构化信息；最后，我们还设计了一些与文本质量有关的特征，例如，文本长度、语言属性、文本的流畅程度等。

实验结果表明这三大类特征对于 Twitter 中的传播观点检索都是有效的，我们整合三大类特征形成的系统明显优于 BM25 基准系统和最好的 Twitter 观点检索系统（第三章中介绍的最好的观点检索系统[84]）。而且将我们最好的系统与人进行观点传播预测比较发现，我们的方法能达到人判定的程度。

本章的主要工作如下：

1. 我们定义了 Twitter 中传播观点发现的新任务，意在 Twitter 中发现高质量的观点。

2. 我们设计了基于 tweet 的传播度特征、观点化特征、文本质量特征的 Twitter 传播观点检索方法，并在真实的数据集上验证了方法的有效性。

3. 实验结果验证了我们设计的特征对于传播观点检索是有效的，并且系统显著优于 BM25 基准系统和最好的 Twitter 观点检索系统。

4. 另外，我们的方法在识别 Twitter 中的观点是否会被转发上可以达到人预测的水平。

4.2 相关工作

Twitter 中传播观点的发现主要是在 Twitter 中找到会被传播的观点，因此一个 tweet 首先必须是观点化的文本，这就与 Twitter 中的观点检索任务相关。Twitter 中一个观点能否传播需要考虑哪些影响人们转发行为的因素，因此可以借鉴预测一般 tweet 转发（不考虑是否为观点化的 tweet）的相关工作。正如前面强调的，高质量的文本往往与信息的传播性相关，因此如何分析观点的质量也是必须考虑的问题。基于以上分析，我们从三个方面讨论相关工作：tweet 转发预测，Twitter 观点检索，观点质量评价。

4.2.1 Tweet 转发预测

在 Twitter 中，消息的重要性可以通过是否被转发来体现。因此，有许多工作是直接预测一个发布的 tweet，能否在将来某个时候被其他人转发。

Suh 等人是第一个提出研究预测 tweet 转发的研究单位[187]，他们设计了一系列特征，并在大规模数据上进行验证，最后发现 tweet 内容，tweet 是否包含链接或 hashtag，与 tweet 是否转发十分相关。

Zaman 等人也在较早的时候利用协同过滤的方法预测一对用户，其中一人发布 tweet，另外的用户是否会转发[188]。他们发现用户信息是最有效的特征。

Naveed 等人则基于 tweet 的内容构建了一个计算 tweet 被转发可能性的模型[189]，他们定义了 tweet 的“有趣”性。实验结果表面他们定义的“有趣”性能够帮助 tweet 的转发预测。

Petrovic 等人利用基于 passive-aggressive 算法的机器学习方法对一个 tweet 能否被转发进行预测[184]。他们发现 tweet 的文本内容，用户的分组数目，粉丝数目和作者是否是验证用户是预测 tweet 转发任务最重要的特征[18]。

Hong 等人采用与 Petrovic 等人类似的机器学习方法，不仅预测了 tweet 是否将被转发，还预测了 tweet 将被转发的次数。

Artzi 等人则将转发与回复行为看做对一个 tweet 的响应[190]，他们依然基于机器学习的方法对 tweet 能否引起响应进行预测。另外，许多其它工作从不同角度预测 tweet 的转发行为或热点事件在 Twitter 中的传播[191–194]。

Stieglitz 等人分析了 tweet 文本的情感倾向性是否对 tweet 转发有影响[195]。他们选取与政治话题有关的 tweet 作为研究对象，发现某些与政党和政治人物有关的情感词的确对相关 tweet 能否被转发有影响。与他们的工作不同，我们的工作不是检测情感因素能否影响 tweet 的转发，而是分析哪些因素影响 Twitter 中观点的传播。

4.2.2 Twitter 观点检索

上一章中，我们第一次研究了 Twitter 中的观点检索[84]，发现利用社交媒体信息与 tweet 的观点化信息可以帮助检索。实验结果表明当检索模型中考虑链接(URL)、提及（Mention)、发布数目（statues)、粉丝数目（statues)、tweet 文本的观点化程度等特征时，可以提高 Twitter 观点检索的效果。另外，我们还提出了一种利用社交媒体信息和 tweet 文本结构化信息帮助生成近似主观化 tweet（PST）和近似客观化 tweet（POT）的方法，以此提高评估 tweet 观点化的程度，实验发现将其引入观点检索中可以替代传统手工标注数据的方法，且考虑话题相关因素构造 PST 和 POT 时，效果更好。本章中我们将在这个工作的基础上进一步研究如何在 Twitter 中找到传播观点，并将观点检索方法作为其中一个基准系统。

Paltoglou 等人利用 Twitter TREC2011 数据重新标注了 tweet 的主观性[196]，他们的工作是我们 3.6.5节工作的进一步扩展，主要是标注了 TREC 数据中更多的 tweet，并请多人交叉标注，但他们的工作并未提出任何观点检索的方法，也没有涉及 Twitter 中观点是如何传播的问题。

4.2.3 观点质量评价

根据观点的质量进行商品评论的排序一直是购物网站十分关注的问题，例如，Amazon.com 和 Ebay.com。但是，大多数的网站都是根据用户的投票（如：thumbs up 和 thumbs down）进行评论的排序，自动化地对评论进行排序一直是个热点和难点。

Kim 等人与 Zhang 和 Varadarajan[197, 198] 最先讨论了如何自动化地评估商品评论的质量，他们利用回归模型对每个评论进行有用性打分，以此对评论进行排序。他们发现利用评论文本的句法和语法信息能够帮助评论文本质量。实验结果表明，浅层语法分析特征，如：评论中名词、动词、形容词不同的比例能够影响评论的质量。

Liu 等人详细分析了 Amazon 中发布的评论[199]，发现用户对评论的投票存在三类明显的偏置:

1. **不平衡投票偏置（Imbalanced Vote Bias）**：有些评论存在大量的用户投票，有些评论很少有人投票，甚至没有人投票。

2. **优胜者投票偏置（Winner Circle Bias）**：一旦某些评论被其他用户投票的次数超过一定的数量后，这个评论将会被更多人投票。

3. **新旧评论投票偏置（Early Birds Bias）**：发布较早的评论所吸引的投票热度明显高于发布较晚的评论。

基于以上三种偏置，Kim 等人与 Zhang 和 Varadarajan 研究工作中所采用的训练数据是不可信的，因此 Liu 等人选取了专家数据训练模型对新评论进行质量评估。

Ghost 等人研究了评论对于商品销售的影响[200]，例如，商品的销售量与评论的关系。他们发现评论的主观性程度，信息量，可读性，语言用词的正确性影响商品的销售量。

Liu 等人研究了电影评论的质量[201]，他们发现除了文本信息的因素，评论者的领域知识程度，评论发布时间也与评论的质量有关。Danescu-Niculescu-Mizil 等人和 Lu 等人还发现评论的有用性不仅与评论的内容有关[202, 203]，还与其社会属性相关，例如评价对象的质量，评价者的权威性等[204–214]。

另外，随着评论信息变得越来越重要，很多商家由于恶性竞争，在相关评论网站上发布垃圾评论甚至虚假评论，因此有一些工作围绕垃圾评论或虚假评论识别展开[215–222]。

以上工作大部分处理的对象都是传统的网站评论。但是 Twitter 作为一种新型社交媒体，它简短的文本和丰富的社交媒体属性应该被用于考虑评论的质量。

4.3 Twitter 传播观点检索实验数据

为了研究哪些因素与 Twiter 中的传播观点有关，我们利用了第三章介绍的 Twitter 观点检索数据[3]。这个数据包括 50 个查询词和 5000 个已经被判定是否为相关观点的 tweet。每一个查询词，平均有 16.62 个话题相关且带观点的 tweet。这个数据是在 2011 年 11 月通过 Twitter streaming API 收集的。

我们本章的目的是在 Twitter 中找到话题相关的观点并且在未来会被转发。因此，我们在 2012 年 4 月利用 Twitter statuses API[4]重新抓取了原数据中的 tweet。根据我们在 4.1对相关 tweet 的定义，我们将以前相关观点 tweet 进行再标注，如果在这 6 个月中，相关观点 tweet 被转发了，则该 tweet 是相关 tweet，否则为不相关 tweet，而其他在原始数据中剩余的 tweet 都为不相关 tweet。我们认为这些 tweet 是否相关的状态已经稳定，它们不太可能被再次转发。另外，当我们重新抓取这些 tweet 时，我们发现某些 tweet 因为各种原因已经被删除，我们将这些被删除的 tweet 统统认定为不相关 tweet，因为它们已经无法通过转发进行消息的传播。最终，对于每一个查询词，平均有 3.4 个相关 tweet，这说明了 Twitter 中只有很少一部分的观点会被转发，大部分的观点是不会被传播的。有趣的是，这个数据中观点被转发的比例是 20.5%，而普通 tweet 被转发的比例是 16.6%，这说明 Twitter 中，观点比普通 tweet 更容易被传播。

4.4 基于排序学习的 Twitter 传播观点检索框架

我们的目的是构造一个排序系统，这个系统能够在 Twitter 中找到未来会被传播的观点。为此我们设计了 tweet 传播度特征，tweet 观点化特征和 tweet 文本质量特征。我们将这些特征引入到排序学习算法中，通过标注的训练数据学习模型，以此为新的 tweet 进行排序获得相关文档。

4.4.1 Twitter 传播观点检索排序学习框架

我们依然采用排序学习作为我们系统的基本框架。通过一些查询词和一些被标注是否与查询词相关的 tweet 构成训练集合，然后设计开发针对相关定义的特征进行模型训练，学习到的模型能够对新的查询词和对应的 tweet 进行排序，排序位置高的为相关 tweet。特征的有效性可以通过模型中引入不同的特征在测试集合中的排序效果进行评估。

[3]下载地址：https://sourceforge.net/projects/ortwitter/

[4]https://dev.twitter.com/docs/api/1/get/statuses/show/%3Aid

4.4.2 Twitter 传播观点检索相关特征

正如 4.2介绍的相关工作，Twitter 的传播观点检索首先需要考虑的是，tweet 能否在将来被传播，因此如何评估 tweet 被转发的可能性是需要解决的问题。其次，Twitter 传播观点检索中对 tweet 观点化的评价是非常重要的一部分，tweet 是否是观点，是决定 tweet 能否成为相关 tweet 的必要条件。最后，在 Twitter 中文本的高质量与 tweet 的传播性息息相关，因此需要分析哪些因素影响文本的质量。因此为了在 Twitter 中检索到传播的观点，我们设计了传播度特征，观点化特征，文本质量特征：

1. **传播度特征（Retweetability Feature）**：主要评估 tweet 未来会被转发的可能性。
2. **观点化特征（Opinionatedness Feature）**：主要涉及如何评估 tweet 的观点化程度。
3. **文本质量特征（Textural Quality Features）**：主要涉及 tweet 文本有关的属性特征。

接下来，我们将具体介绍有关的特征。

4.5 传播度特征

Twitter 中 tweet 的转发被认为是最重要的信息传播手段，并且有大量的工作是关于给定一条 tweet，预测该 tweet 在未来能否被转发[18, 184]。因此，我们设计了一个特征来预测 tweet 是否会被转发。我们利用 Petrovic 等人的工作来设定该特征[184]。该工作通过机器学习利用 passive-aggressive 算法来计算 tweet 会被转发的置信度。算法中利用了一系列与预测转发有关的特征，包括：

1. **内容（Content）**：tweet 中的文本内容。这个特征主要表示 tweet 的话题，因为有关特定话题的 tweet 比有关其他话题的 tweet 更容易被转发。例如，人们可能更加关注与话题“ran nuclear”有关的 tweet，而对话题“systems biology”相关的 tweet 兴趣较小。
2. **粉丝数目（Followers）**：tweet 作者的粉丝数目。这个特征反映了用户受欢迎的程度。一个受欢迎的用户发布的 tweet 往往比较容易被转发。
3. **分组数目（Listed）**：tweet 作者被分组的数目。这个特征同样反映了用户受欢迎的程度。

4. **验证用户（Verified）**：tweet 作者是否被官方验证。这个功能主要是 Twitter 官方来验证某些用户的权威性（如，明星）。在 Petrovic 等人的工作中，他们发现 91% 由验证用户发布的 tweet 都会被转发，而只有 6% 的非验证用户发布的 tweet 被其他人转发[184]。

对于 tweet 是否会被转发，时间因素也非常重要，例如，人们可能在 2011 年 11 月对“American Music Awards”的关注度远远高于 2012 年 4 月。因此，我们在训练预测 tweet 是否会被转发的模型时，使用了由 Twitter streaming API[5]抓取的 2011 年 11 月的 tweet。这个训练数据中总共包含 3000 万条 tweet，我们将那些由“retweet”按钮转发的 tweet 视为正例，其他 tweet 视为反例。最后，我们测试了我们模型的预测性能，我们随机选取了 10 万条 tweet 作为测试集，结果显示我们预测模型的正确率到达 95.99%。

利用这个模型，我们可以对我们传播观点检索中的数据进行转发预测，我们利用 passive-aggressive 算法对 tweet 转发预测的置信度作为特征的分值，并称该特征为 Retweetability 特征。

4.6 观点化特征

显然传播观点检索任务中，对 tweet 的观点化程度进行评估依然是不可或缺的。这里我们采用 3.5节介绍的方法对 tweet 进行观点化程度的评估，具体采用表 3.3中介绍的 Q_D 特征：由话题相关的 PSTs 和 POTs 数据对 tweet 的观点性打分，因为这个特征在 Twitter 观点检索任务中，取得最佳的检索效果。另外，我们将这个特征称为 Opinionatedness 特征。

4.7 文本质量特征

Twitter 作为一种非常流行的社交媒体吸引了许多用户在其发布各种消息，例如，个人信息，垃圾信息，对话信息等等。不同于传统新闻网站发布的正式消息，这些信息大部分都没有进行认真地编辑，因此存在大量的错别字，缩写语和口语化的用词。Tweet 文本的质量往往参差不齐，我们设计了一些特征来帮助判定 twee 文本的质量，以此帮助 Twitter 中传播观点的检索：

1. **长度（Length）**：tweet 文本中词的个数。Kim 等人发现文本的长度能够有效评估评论的质量[197]。直觉上，更长的 tweet 一般包含更多的信息，我们用此特征间接表示 tweet 信息的丰富程度。

[5]http://stream.twitter.com/

2. **词性（PosTag）**：第三章的研究已经发现个人发布的信息更有可能在 tweet 中包含观点，这些 tweet 经常包含人称代词，例如，“i”，“u”，“my”，另外还包含一些表情符合，例如，“:)”，“:(”，“d”，但是一些很少包含“开放类（open-class）”词的观点往往是一些低质量的信息，例如，这个 tweet “@fayemckeever Jennifer Aniston :)”几乎对读者来说没有任何价值。因此我们设计了一些特征，意在捕获 tweet 文本的语言属性，具体包括 tweet 中“开放类（open-class）”词的比例，名词的比例，动词的比例，形容词的比例。这里我们使用 Twitter Part-of-Speech Tagging[6] 对 tweet 进行词性标注[21, 108]。

3. **流畅度（Fluency）**：tweet 的流畅程度反映了 tweet 的可读性，我们利用语言模型（language model）来解决评估 tweet 文本的流畅程度[223]。我们将 tweet t 在特定语言模型下的概率作为 tweet 文本流畅度的分值 $F(t)$，具体公式如下：

$$F(t) = \frac{1}{m}P(w^m) = \frac{1}{m}\prod_{i}^{m} P(w_i|w_{i-N+1}, w_{i-N+2}, ..., w_{i-1})$$

其中 tweet t 可以表示成一个词向量 $w^m = (w_1, w_2, ..., w_m)$。为了解决语言模型中长度的偏置问题，我们通过 tweet 文本的长度来对其值进行归一化。我们利用 2011 年 11 月的 3000 万条 tweet 计算得到 N 元语言模型（N=4）。

4.8 Twitter 传播观点检索实验

4.8.1 传播观点人工预测实验

在评估我们的 Twitter 中传播观点检索方法之前，我们首先测试人是否可以正确判断哪些是传播的观点。我们请了两个用户参与测试，总共向他们给出了 100 对 tweet，然后要用户判断那个是未来会被转发的观点。每一对 tweet 都只与一个话题相关，并且只有一个是转发的观点。给出每一对 tweet 请用户判断时，tweet 的顺序被随机打乱，为的是避免 tweet 排列顺序的偏置。

我们以正确率衡量人的判断能力，正确率是 100 对 tweet 中用户准确判断那个是传播观点的 tweet 对数。实验结果显示，两个用户都能够击败随机的选取系统(正确率为 50%)，其中一个用户判断的正确率为 75% ，另一个为 69%。实验结果说明人能够有效地在 Twitter 中判断那个是传播的观点，那个不是。

[6]下载地址：http://www.ark.cs.cmu.edu/TweetNLP/

4.8.2 Twitter 传播观点检索实验设置和基准系统（Baseline）

本节我们介绍一下我们的实验设置和基准系统。对于排序学习的实现，我们依然使用了 SVM Rank 工具[7]。我们选取线性核函数，实验结果都是选取最优的参数设定。为了避免过拟合，我们使用了 10 次交叉验证，因此每次验证中包含 45 个查询词和对应的 tweet 作为训练集合，5 个查询词和对应的 tweet 作为测试集。我们还是使用平均准确率（Mean Average Precision-MAP）作为我们测试的验证指标。

Twitter 传播观点检索是一项新的工作，直接研究该问题的的工作还没有，因此我们选取另外两个方法作为我们的基准系统。一个是使用 Okapi BM25 计算的查询词与 tweet 的话题相关性进行排序得到的 BM25 基准系统，这个系统被广泛应用于各种 Twitter 检索任务中[83, 84, 88]。另一个基准系统采用第三章介绍的 Twitter 观点检索的方法。这个方法利用了社交媒体特征与 tweet 的观点化特征，我们将这个基准系统称为 TOR（Twitter Opinion Retrieval）。TOR 具体的特征在表 4.1中给出。

表 4.1　TOR 基准系统的相关特征

TOR 特征（TOR Features）	**描述**
BM25	Okapi BM25 分数
链接（URL）	tweet 中是否包含链接
提及（Mentions）	tweet 是否包含提及
发布数目（Statuses ）	tweet 的作者以往发布 tweet 的数目
粉丝数目（Followers）	tweet 的作者的粉丝数目
观点化程度（Opinionatedness）	tweet 的观点化程度

4.8.3 Twitter 传播观点检索实验结果及分析

本节我们验证上面设计的特征是否对于 Twitter 中的传播观点检索有效。我们将设计的特征一个一个加到两个基准系统中进行验证。表 4.2 简要概述了 Twitter 传播观点发现所使用的相关特征。表 4.3和表 4.4给出了实验结果。

我们看到当传播度特征（Retweetability），词性特征（PosTag），流畅度特征（Fluency）整合到 TOR 基准系统时，检索传播观点的 MAP 值能够显著提高。这说明考虑 tweet 被传播的可能性，tweet 文本的语言属性，tweet 文本的可读性可以帮助发现 Twitter 中的传播观点。另外，我们发现 TOR 基准系统的 MAP 值显著高于 BM25 基准系统的 MAP 值（$p<0.01$），这说明 tweet 的观点化程度和 tweet 的一些社交媒体信息同样能够帮助发现 Twitter 中的传播观点。虽然当 BM25 基准系统

[7]下载地址：http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

表 4.2　Twitter 传播观点检索特征概况

基准系统特征（Baseline Feature）	**取值范围**	**描述**
BM25	$(0, +\infty)$	tweet 的 BM25 分值
Twitter 观点检索系统特征（TOR Features）	**取值范围**	**描述**
BM25	$(0, +\infty)$	tweet 的 BM25 分值
链接（URL）	0 or 1	tweet 中是否包含链接
提及（Mentions）	0 or 1	tweet 是否包含提及
发布数目（Statuses ）	$N^+ = \{1, 2, 3, ...\}$	tweet 的作者以往发布 tweet 的数目
粉丝数目（Followers）	$N = \{0, 1, 2, ...\}$	tweet 的作者的粉丝数目
观点化程度（Opinionatedness）	$(-\infty, +\infty)$	tweet 的观点化程度
传播度特征（Retweetability Feature）	**取值范围**	**Description**
转发置信度（Retweetability）	[0,1]	tweet 转发的置信度
观点化特征（Opinionatedness Feature）	**取值范围**	**Description**
观点化程度（Opinionatedness）	$(-\infty, +\infty)$	tweet 的观点化程度
文本质量特征（Textural Quality Features）	**取值范围**	**Description**
长度（Length）	[1,140]	tweet 的词数
词性（PosTag）	[0,1]	不同词性的词分布比例
流畅度（Fluency）	[0,1]	语言模型计算的 tweet 文本流畅程度

表 4.3　Twitter 传播观点检索系统实验结果（TOR）

	MAP
BM25	0.0997
BM25+Retweetability	0.1077
BM25+Opinionatedness	0.1146
BM25+Length	0.0881
BM25+PosTag	0.1157
BM25+Fluency	0.1046
BM25+Textural Quality	0.1277
BM25+All	0.1317

$^{\triangle}$ 和 $^{\blacktriangle}$ 分别表示排序结果显著高于 BM25 传播观点检索系统（$p < 0.05$ 和 $p < 0.01$），BM25+All 使用了 BM25，Retweetability，Opinionatedness 和 Textural Quality features 特征。

整合传播度特征，词性特征，流畅度特征时，MAP 值都有提高，但并不显著，可能的原因是仅仅使用这些特征与 BM25 特征组合不足以发现 Twitter 中的传播观点。有趣的是我们发现长度特征（Length）整合到 TOR 基准系统时，能够帮助传

表 4.4　Twitter 传播观点检索系统实验结果（BM25）

	MAP
TOR	0.1521
TOR+Retweetability	0.1806▲
TOR+Length	0.1580
TOR+PosTag	0.1917▲
TOR+Fluency	0.1875△
TOR+Textural Quality	0.1930▲
TOR+Retweetability+Textural Quality (Best)	0.1992▲

△ 和 ▲ 分别表示排序结果显著高于 TOR 传播观点检索系统（$p < 0.05$ 和 $p < 0.01$）。

播观点检索，但整合到 BM25 基准系统时，MAP 值反而下降，这说明 tweet 的长度信息不像其他网站的评论信息一样[197]，对观点的质量判断有效。可能的原因是 tweet 的长度限制在 140 个字符，因此 tweet 的长度差异性不像其他网站的评论一样那么明显。最后我们将整合的文本质量特征（Textual Quality）加到两个基准系统中，文本质量特征包括长度特征，词性特征，流畅度特征，实验结果显示文本质量特征能够帮助 Twitter 中的传播观点检索。以上种种说明我们设计的传播度特征，观点化特征，文本质量特征对于 Twitter 中的传播观点检索是有效的。

最后我们将所有的特征都整合到 TOR 基准系统中（Best），表 4.4给出了 Twitter 传播观点检索最好的结果，MAP 值达到 0.1992，这个 MAP 值超过 TOR 基准系统 MAP 值 30.97%，超过 BM25 基准系统 MAP 值 99.80%。以上说明我们最好的系统不仅能够在 Twitter 中找到话题相关的观点，而且观点能在未来被转发传播。例如，以下一个例子是我们数据中与话题“American Music Awards”相关的三个 tweet：

(a) *Watch Olnine Free| The 39th Annual American Music Awards (TV 2011): The 39th Annual American Music Awards (TV 20... http://t.co/SxrjVVmx*

(b) *We’re so excited for the American Music Awards this weekend*

(c) *That awkward moment when the American Music Awards is really the American Minaj Awards*

在我们的各个系统中，BM25 基准系统将 Tweet（a）排在 Tweet（b）和 Tweet（c）前面，但是 Tweet（a）只是与查询词“American Music Awards”话题相关，并不包含观点。基准系统 TOR 将 Tweet（b）排在其他两个 tweet 前面，虽然 Tweet（b）包含了对话题“American Music Awards”的观点，但是该 tweet 并没有被转发。

而我们最好的传播观点检索系统（Best）将 Tweet（c）排在最前面，我们可以看到这个 tweet 不仅与话题“American Music Awards”相关，而且还包含有意思的观点，另外，六个月后这个 tweet 还被转发了 143 次。

4.8.4 Twitter 观点传播预测 VS Twitter 普通 tweet 传播预测

已经有许多工作是关于预测普通的 tweet 能否被转发[18, 184]。我们想研究分析一下在 Twitter 中普通 tweet 转发与观点转发之间的关系。因此，我们构造了一个系统只使用传播度特征（Retweetability）进行 Twitter 中传播观点的检索。表 4.5 显示了实验结果。我们发现只使用传播度特征的检索系统明显不如 Best 系统效果好，这说明 Twitter 中的传播观点转发预测不同于普通 tweet 的转发预测。因此，正如我们前面的实验结果所得，对于 Twitter 中的传播观点检索应该考虑更多的因素，如 tweet 的观点化程度和文本的质量。

表 4.5 Twitter 传播观点检索系统实验结果（Retweetability）

	MAP
Retweetability	0.0936
TOR+Retweetability+Textural Quality (Best)	0.1992▲

△ 和 ▲ 分别表示排序结果显著高于 Retweetability 传播观点检索系统（$p < 0.05$ 和 $p < 0.01$）。

4.8.5 Twitter 传播观点预测 VS Twitter 传播观点人工预测

最后，我们将 Best 系统与人进行 Twitter 中传播观点预测的实验比较，我们利用 4.8.1节介绍的 100 对 tweet，然后利用 Best 系统对每一对 tweet 进行排序，排序高的判定为 Twitter 中会被传播的观点。最后实验结果显示我们的 Best 系统的正确率达到 71%，这个结果略为小于人预测的平均正确率（72%），但没有显著性差别（$p=0.05$）。这个结果说明了，在 Twitter 中的传播观点检索的任务上，我们的方法能够达到人预测的效果。

4.9 小结

本章中我们详细介绍了如何在 Twitter 中发现传播观点。一系列特征被整合到基于排序学习的框架中解决传播观点发现的任务，这些特征包括：传播度特征（Retweetability），观点化特征（Opinionatedness），文本质量特征（Textural Quality）。

实验结果表明这些特征对于 Twitter 中的传播观点的发现是有效的，而且结合所有特征构造的最佳检索系统性能显著优于 BM25 系统和目前最好的 Twitter 观点检索系统。最后，我们还发现我们的系统在预测 Twitter 中观点是否会被转发的能力可以达到人判断的效果。

第五章　Twitter 中信息传播者的发现

5.1　引言

Twitter 的快速发展使得信息的交流变得越来越方便和快捷。人们每天在 Twitter 上不仅接受资讯，同时也发表自己的观点。在 Twitter 中一个很重要的机制就是转发（retweeting）：重复发送其他用户发布的 tweet。这种机制是 Twitter 中最常用的信息传播手段，当一个 tweet 被某人转发时，该用户的所有粉丝都将看到此信息，同时这种转发行为也反映了转发者对 tweet 原作者所持观点的一种肯定的态度。

现有对转发的研究主要集中在给定一条 tweet，预测未来是否会被转发上[18, 184, 190]；另外一部分研究则把重点放在研究转发的行为模式上[70, 71, 224]。但是以上的研究都忽略了一个重要的问题：到底“谁”会转发给定的 tweet。我们将这个“谁”称为信息传播者。上一章，我们讨论了 Twitter 中传播观点发现的问题，这是从 Twitter 传播分析的 tweet 角度进行考虑，本章中我们将从传播的受众角度对 Twitter 进行传播分析的研究，即如何在 Twitter 中找到信息传播者。

由于用户参与 Twitter 的低门槛特点，因此许多人在 Twitter 上发布低质量甚至虚假的信息，其中对社会危害最大的就是谣言。这种信息加上转发机制，可以在 Twitter 上迅速传播，造成大众的恐慌，危害极大。目前在 Twitter 上减低谣言的危害，除了自动识别谣言信息以外，我们认为对谣言传播路径的监督与控制也是有效的手段之一。如果当一条 tweet 确定为一条谣言以后，我们通过信息传播者分析，能够预测哪些用户会传播这些谣言，对于维护社会和谐稳定是具有积极的意义。另外，研究信息传播者可以帮助商业或娱乐公司减少广告成本，因为这个研究可以找到性价比高的用户发布相关用户感兴趣的信息，并且迅速传播，扩大商业公司或娱乐公司所推广商品的影响力。

研究在 Twitter 中如何找到信息传播者十分有意义，它能够帮助我们加深理解信息是如何在这个社交媒体中传播的。这些影响信息传播的因素可能与消息的本身，消息的作者，以及消息的接受者有关。三种因素之间相互影响，决定信息的传播。我们将 Twitter 中找到信息传播者的问题看做一个排序问题，即给定一个 tweet，在作者的粉丝中发现“谁”将转发该消息。我们利用机器学习的方法，为用户设计一系列特征构建排序学习模型。这些特征包括了用户历史的转发信息，用户自身的社交媒体特征，用户使用 Twitter 的活跃时间，以及用户的个人兴趣。

我们构造了自己的实验数据来验证 Twitter 信息传播者发现方法的有效性，实验结果表明我们的方法显著优于随机排序的方法和基于用户以往转发原作者 tweet

数量的排序方法。另外，我们发现用户历史的转发信息、用户的兴趣、以及用户的活跃时间是决定转发者的重要因素。

本章的主要工作如下:

1. 我们定义了 Twitter 中信息传播者发现的新任务，帮助理解信息在 Twitter 中是如何传播的。

2. 我们设计了基于用户转发历史信息、用户属性、用户活跃时间和用户兴趣的 Twitter 信息传播者发现排序方法，并在真实的数据集上验证了方法的有效性。

3. 实验结果验证了我们的方法对于信息传播者发现是有效的，并且系统显著优于随机系统和基于用户历史转发记录的排序系统。

4. 另外，我们还发现用户历史转发信息，兴趣和活跃时间是决定信息传播者的重要因素。

5.2 相关工作

由于 Twitter 的流行性，数据的公开性，以及独特的转发属性使得 Twitter 的研究变得异常活跃。我们将相关工作分成两个部分：tweet 转发预测和 Twitter 转发行为分析。具体参见 4.2.1和1.3.3。但是要强调的是不同于 tweet 的转发预测，本章我们的工作集中在预测给定的 tweet 到底“谁”会转发，而 Twitter 转发行为的分析可以帮助提高发现信息传播者的效果。这里需要强调的是，据我们所知目前 Twitter 的研究中还没有对信息传播者发现的相关研究，因此我们自己构造了基准系统进行对比实验。

5.3 基于排序学习的 Twitter 信息传播者发现框架

给定一个 tweet t，它的作者为 user u，user u 的粉丝集合为 $Followers(u)$，我们的目的就是学习一个排序模型来评估每个粉丝 f_i ($f_i \in Follower(u)$) 未来转发 t 的可能性。不失一般性，我们将这个问题当做排序问题，而没有看成分类问题。另外，由于粉丝集合 $Followers(u)$ 中可能存在任意多个转发者，因此我们从 $Followers(u)$ 中选取 top-k 个评估分数最高的粉丝作为信息传播者。

5.3.1 Twitter 信息传播者发现排序学习框架

为了生成一个排序函数 F 能够根据粉丝是否转发 tweet 对其进行排序，我们将设计一些特征，并将其引入到基于排序学习的模型中。正如前几章介绍的，排

序学习是一种整合特征以数据驱动的机器学习模型。这里在训练数据中，每个粉丝 f_i 都被标注是否转发 t，一系列与 Twitter 信息传播者相关的特征都从 u 与 f_i 的关系和各自的属性中抽取，并且特征的有效性，可以通过特定的特征组合，在测试数据的排序表现中体现出来。

5.3.2 Twitter 信息传播者相关特征

一个 tweet 能否被转发，是 tweet 本身、tweet 的作者、作者的粉丝相互作用的结果。因此在 Twitter 信息传播者发现任务中，我们需要考虑三者的属性以及相互的关系。我们首先考虑 tweet 的作者与作者的粉丝以往历史的转发信息，以此捕获粉丝在转发 tweet 上是否具有偏向性。其次，粉丝本身的特征也会影响其转发行为，比如，直觉上对于同一个 tweet 转发的概率，非名人就比名人的转发概率高，因为名人更加注重声誉，对信息质量要求更高。再次，时间因素也是需要考虑的，如果转发者与信息发布者具有相同的使用 Twitter 时间，那么信息在两人之间传播的可能性就比较大。最后，tweet 的内容会影响人的转发行为，感兴趣的信息比不感兴趣的信息对人的转发行为影响更大。因此，为了在 Twitter 中找到信息传播者，我们设计了转发历史特征，用户特征，用户活跃时间特征，用户兴趣特征：

1. **转发历史特征（Retweet History Feature-RH）**：主要涉及用户以前转发 tweet 的信息，同时也包括自身 tweet 被转发的情况。

2. **用户特征（Follower Status Feature-FS）**：主要涉及用户的社交媒体属性。

3. **用户活跃时间特征（Follower Active Time Feature-FAT）**：主要涉及用户发布 tweet 的活跃时间。

4. **用户兴趣特征（Follower Interests Feature-FI）**：主要通过 tweet 的内容发现用户的兴趣爱好。

接下来，我们将具体介绍有关的特征。

5.4 转发历史特征

直觉上，如果粉丝 f_i 以前经常转发或提及 user u 的 tweet，那么粉丝 f_i 很有可能再次转发 user u 的 tweet。因此我们设计了两个特征来获取转发信息：

1. **用户转发数目（Num_fRu）**：以往历史记录中，粉丝 f_i 转发 user u 的 tweet 数目。

2. **用户提及数目（Num_fMu）**：以往历史记录中，粉丝 f_i 提及 user u 的 tweet 数目。

用户的交流是相互的，如果一方经常转发或提及对方，那么对方也很有可能转发或提及对方，因此我们设计了另外两个特征：

1. **用户被转发数目（Num_uRf）**：以往历史记录中，粉丝 f_i 的 tweet 中被 user u 转发的 tweet 数目。

2. **用户被提及数目（Num_uMf）**：以往历史记录中，粉丝 f_i 的 tweet 中被 user u 提及的 tweet 数目。

最后，我们发现有些用户（例如，垃圾用户）只转发其他人的信息，不撰写原始的 tweet，我们设计了两个特征来对这些情况进行建模：

1. **用户转发比例（Ratio_retweet）**：以往历史记录中，粉丝 f_i 的 tweet 中转发 tweet 的比例。

2. **用户提及比例（Ratio_mention）**：以往历史记录中，粉丝 f_i 的 tweet 中提及 tweet 的比例。

5.5 用户特征

信息的传播一般是从社会地位高的用户（例如，名人）流向地位低的用户，这主要跟用户的社会属性相关[225]。

我们随机抽取了 10 万条转发 tweet 进行了详细的研究分析，发现只有 38.8% 的转发是从发布 tweet 较少的用户到发布 tweet 较多的用户；仅仅 23.8% 的转发是从粉丝较少的用户到粉丝较多的用户；0.04% 的转发是从非官方验证的用户到官方验证的用户。这些统计结果充分说明了在 Twitter 中不同社会地位的用户存在不同的转发行为。

因此，我们设计了一些特征来对用户的社会属性进行描述：

1. **发布 tweet 数目（Posts）**：用户发布 tweet 的数目。

2. **粉丝数目（Followers）**：用户的粉丝数目。

3. **朋友数目（Friends）**：用户的朋友数目。

4. **分组数目（Listed）**：用户被分组的数目。

5. **验证用户（Verified）**：用户是否被官方验证。

5.6 用户活跃时间特征

Twitter 用户一般不太可能在很晚的时间与其他人进行交流，而且如果一条 tweet 在很晚发布，那么该 tweet 作者的粉丝很可能第二天忽视了这条消息，因为它很有可能淹没在大量的新消息中，而大部分 Twitter 用户浏览 tweet 的模式是从最新的信息开始。

我们随机抽取了 1 万条回复 tweet，并对其发布时间进行了分析，发现只有 12.4% 回复 tweet 发生在 00:00 到 06:00 之间，这说明用户的活跃时间有一定的规律。

因此，为了捕获这些用户交流的时间信息，我们设计了两个特征：

1. **时区时间（Timezone）**：粉丝 f_i 是否与 user u 在同一个时区。
2. **用户活跃时间（PostTimeConsis）**：以往历史记录中，粉丝 f_i 发布 tweet 不同时间的数目比例与 tweet t 发布时间的关系，选取比例值为特征值。

我们用以上两个特征来反映用户与粉丝在 Twitter 上活跃时间的一致性。

5.7 用户兴趣特征

当一个粉丝转发一条 tweet 时，通常意味着粉丝与 tweet 作者存在一些共性，例如，他们有共同的兴趣（“we both love cats”），这是一种不太明显的关系，不像 Twitter 中的朋友或粉丝关系。

这种潜在的共性可能可以帮助发现信息传播者，我们利用相似性模型（Similarity Model）来计算 tweet t 与粉丝 f_i 之前发布 tweet 的兴趣相似程度。我们将 tweet t 和粉丝 f_i 之前发布的 tweet 都表示成词向量，然后用 $tf\text{-}idf$ 表示词的权重，用余弦夹角计算两者的相似性，我们将这个特征称为“**相似兴趣（SimInterest）**”。

在计算这个特征的时候，对于每一对 tweet t 和粉丝 f_i 以往发布的 tweet，我们首先过滤到在 6 百万 tweet 中词频最高的 100 个词和词频小于 5 的词来表示词向量，具体的余弦夹角的计算利用了向量空间模型（Vector Space Model）[130]。

5.8 Twitter 信息传播者发现实验

5.8.1 Twitter 信息传播者发现实验数据

到目前为止还没有关于在 Twitter 中发现信息传播者的数据，因此我们自己构造了相关数据[1]。

[1]下载地址：https://sourceforge.net/projects/retweeter/

我们从 Twitter Streaming API 中随机选取了 500 条 tweet，每个 tweet 至少被其作者的粉丝转发过一次。这些数据时间分布在 2012 年 9 月 14 日到 2012 年 10 月 1 日之间。Kwak 等人发现一半以上的转发行为发生在原始 tweet 发布一个小时以内，75% 发生在一天以内[226]，因此我们在一天以后重新抓取了 500 个 tweet，检测最后到底有哪些粉丝转发了对应的 tweet。另外，由于 Twitter API 的限制和有些受欢迎的用户粉丝很多，我们不太可能全部抓取所有的粉丝，所以对每一个 tweet，我们只研究了 tweet 作者最新的 100 个粉丝，并获取了每个粉丝最新的 200 个 tweet。

像前面介绍的，我们对每个粉丝进行了标注，转发者标注为 1，非转发者为 0。这里要强调的是，有些用户由于隐私问题未使其数据公开，更重要的是我们未对每个 tweet 抓取所有的粉丝，造成有些 tweet 并没有转发者，这就使得最后的评测结果数值偏低。

表 5.1给出了数据的一些基本情况：

表 5.1　Twitte 信息传播者发现实验数据统计信息

被转发 tweet 总数目	500
平均每个 tweet 粉丝数目	81.15
转发者总数目	257
非转发者总数目	40317

5.8.2　Twitter 信息传播者发现实验设置

我们使用 SVM Rank 来实现排序学习算法并构造排序模型[2]。排序学习依然使用线性核函数，所有的模型参数都调到最优。为了避免数据的过拟合，我们使用了 10 次交叉验证。评测指标继续使用平均准确率（Mean Average Precision-MAP）。

5.8.3　Twitter 信息传播者发现基准系统（Baseline）

Twitter 信息传播者发现是一个新的工作，目前还没有其他方法能够进行直接的比较。因此，我们自己选择了两个基准系统：

1. **Random**：对所有粉丝随机排序。

2. **RPT**：如果一个粉丝在历史上经常转发某用户的 tweet，那么在未来他很有可能继续转发，因此我们根据粉丝历史数据中（每个粉丝 200 个最新的 tweet）转发用户 tweet 的数目进行排序。

[2]下载地址：http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

表 5.2　Twitter 信息传播者发现特征概况

转发历史特征（RH）	**取值范围**	**Description**
用户转发数目（Num_fRu）	$N = \{0, 1, 2, ...\}$	粉丝转发作者 tweet 的数目
用户提及数目（Num_fMu）	$N = \{0, 1, 2, ...\}$	粉丝提及作者 tweet 的数目
用户被转发数目（Num_uRf)	$N = \{0, 1, 2, ...\}$	作者转发粉丝 tweet 的数目
用户被提及数目（Num_uMf)	$N = \{0, 1, 2, ...\}$	作者提及粉丝 tweet 的数目
用户转发比例（Ratio_retweet）	$[0, 1]$	粉丝的 tweet 中转发 tweet 的比例
用户提及比例（Ratio_mention）	$[0, 1]$	粉丝的 tweet 中提及 tweet 的比例
用户特征（FS）	**取值范围**	**Description**
发布 tweet 数目（Posts）	$N^+ = \{1, 2, 3, ...\}$	作者以往发布 tweet 的数目
粉丝数目（Followers）	$N = \{0, 1, 2, ...\}$	作者的粉丝数目
朋友数目（Friends）	$N = \{0, 1, 2, ...\}$	作者的朋友数目
分组数目（Listed）	$N = \{0, 1, 2, ...\}$	作者的分组数目
验证用户（Verified）	0 or 1	作者是否被官方验证
用户活跃时间特征（FAT）	**取值范围**	**Description**
时区时间（Timezone）	0 or 1	粉丝是否与作者在同一个时区
用户户活跃时间（PostTimeConsis）	$[0, 1]$	粉丝发布 tweet 不同时间的数目比例
用户兴趣特征（FI）	**取值范围**	**Description**
相似兴趣（SimInterest）	$(-1, 1)$	tweet 与粉丝以往发布 tweet 的相似度

Random 系统是一个“弱”的基准系统，与它的比较能够反映哪些因素能够帮助 Twitter 中信息传播者发现。**RPT** 系统是一个“强”的基准系统，与它的比较能够说明我们最好的 Twitter 信息传播者发现方法的能力。

5.8.4　Twitter 信息传播者发现实验结果及分析

我们利用不同特征集合：转发历史特征、用户特征、用户活跃时间特征、用户兴趣特征，分别构造了排序系统 RH，FS，FAT 和 FI。表 5.2 简要概述了 Twitter 信息传播者发现的相关特征。

表 5.3给出了不同系统发现信息传播者的结果。我们可以看到基于用户兴趣特征的 FI 排序系统效果最好，而基于转发历史特征的 RH 系统和基于用户特征的 FS 其次，基于用户活跃时间特征的 FAT 系统没有明显效果。这说明粉丝的兴趣爱

好，粉丝的转发历史，以及粉丝的社会地位可以帮助我们发现 Twitter 中的信息传播者。

我们将所有特征整合到一个排序系统中，称为 All。这个系统达到最高的 MAP 值，高出 Random 301.4%，高出 PRT 25.6%。

表 5.3　基于不同特征组的 Twitter 信息传播者发现系统实验结果

	MAP
Random	2.17
PRT	6.93
RH	6.27*
FS	3.66*
FAT	2.91
FI	8.12*
All	8.71*†

* 和 † 分别表示排序结果显著高于 Random 信息传播者发现系统和 PRT 信息传播者发现系统（$p < 0.05$）。

接着我们分析了具体的特征对于信息传播者发现的影响。我们根据不同的特征进行单独的数据训练与测试。表 5.4给出了各个特征的排序表现，这里 PRT 的 MAP 值与 Num_fRu 的 MAP 值不同的原因是因为前者根据数值直接排序，后者作为特征基于排序学习算法进行排序。

我们可以看到转发历史特征集合中的各个特征都能显著提高检索信息传播者的效果，另外，我们还发现用户活跃时间（PostTimeConsis）对于发现信息传播者也是有效的，最后利用特征相似兴趣（SimInterest）取得的最好的排序结果充分说明粉丝转发 tweet 是根据自己的兴趣与 tweet 的内容是否匹配来进行的。

这里是我们数据中的一个例子，反映粉丝转发的历史对于检索信息传播者的有效性：

We are having a bake sale today in the Student Union from 11-2! Come buy a midday snack from the Pretty Poodles!

有个该 tweet 作者的粉丝在这条信息发布之前已经转发此作者的信息 30 多次，而且该粉丝继续转发了这条 tweet。我们的检索模型 RH 成功地将该粉丝排在了第一位。

这是另一个验证粉丝兴趣对于检索信息传播者有效的例子：

Excited to announce our debut London show. Full details here - http://t.co/P60Wc3Lj

表 5.4　基于不同特征的 Twitter 信息传播者检索系统实验结果

	MAP
Random	2.17
PRT	6.93
Num_fRu	6.83*
Num_fMu	7.08*
Num_uRf	6.20*
Num_uMf	7.62*
Retweet_Ratio	4.45*
Mention_Ratio	3.05*
Posts	3.79*
Followers	2.37
Friends	2.03
Listed	2.17
Verified	2.34
Timezone	2.37
PostTimeConsis	2.86*
SimInterest	8.12*

* 和 † 分别表示排序结果显著高于 Random 信息传播者检索和 PRT 信息传播者检索（$p < 0.05$）。

有一个该作者的粉丝转发了这个 tweet，并且该粉丝在以前的 tweet 中经常发布一些与音乐和演唱会有关的信息。我们的 FI 检索系统也成功地将其找到。

5.9　小结

在 Twitter 中寻找信息传播者可以帮助我们更有效地向其他用户传送信息，我们对于信息传播者检索的工作能够帮助其他研究者更好地了解社交媒体中信息是如何传播的。本章中我们发现粉丝转发的历史记录，粉丝的社会地位，粉丝个人的兴趣爱好对于信息传播者检索是有效的。

未来我们将设计更多的特征帮助发现信息传播者。例如，是否亲密的朋友会经常转发用户的 tweet，地理位置信息是否有所帮助等等。

第六章　总结与展望

社交媒体是一个新兴领域，本文主要围绕 Twitter 中文本特点和社交媒体特征展开研究。通过 Twitter 中的信息检索和传播分析任务，我们发现 Twitter 中的文本结构化信息和 tweet 的社交媒体信息可以帮助这些问题的解决。

Twitter 中的检索研究能够从 Twitter 的海量数据中快速找到有意义的信息，对于 Twitter 中的其他研究具有重要的意义。以往的信息检索研究主要是对图书馆文档或网页进行处理，我们针对 Twitter 数据，具体涉及了 Twitter 中的传统信息检索问题研究和 Twitter 中观点检索研究，以此解决如何在 Twitter 中找到主客观 tweet 的问题。

Twitter 中的传播分析问题，我们主要从 tweet 本身的传播和传播的受众角度进行分析，提出了 Twitter 中传播观点发现与传播者发现的问题。通过任务的定义与方法的研究，最后通过实验验证，找到了一些 tweet 文本特征和社交媒体特征与 Twitter 中信息传播的内在联系。

6.1　工作总结

本文的主要工作可以从以下四个方面来总结：

首先，针对现有 Twitter 信息检索工作忽视 tweet 文本结构信息对 tweet 排序重要性的问题，我们对 tweet 文本进行了结构化研究，以此帮助 Twitter 中的信息检索。这个工作的动机是基于普通文本和网页结构信息能够帮助传统信息检索的已有研究结论。虽然 tweet 文本短小，但是也存在结构化属性的特点。我们定义了 tweet 文本中的几个结构化模块，称之为 Twitter 积木。然后构造自动标注器，对 tweet 文本进行积木的标注。任何一个 tweet 文本都是由若干积木块排列组合而成，而 tweet 文本特定的积木组合又对应了文本特殊的属性。我们通过这种积木结构开发特征，然后结合 tweet 的社交媒体特征，将其应用到基于排序学习的 Twitter 信息检索任务中。实验结果发现我们的 tweet 文本结构化信息能够帮助 Twitter 信息检索。

其次，针对目前政府、企业、个人都通过 Twitter 来收集大量的观点帮助决策，但是并未对观点收集的基础工作观点检索展开系统研究，我们第一次提出了 Twitter 中观点检索的新问题。我们发布了 Twitter 观点检索的新语料，该语料已经作为 ICWSM 会议的常用语料供后续研究者研究使用[1]。另外，我们根据 Twitter 中观点检索与博客观点检索的不同点，利用社交媒体特征与 tweet 观点化特征，提

[1] http://www.icwsm.org/2013/datasets/datasets/

出了 Twitter 观点检索的方法，该方法在实验结果上显著优于优化的 BM25 基准系统和基于向量空间模型的基准系统。再者，我们还提出了一种基于社交媒体特征与 tweet 文本结构化信息收集近似主观化 tweet 和近似客观化 tweet 构造主观化词典的方法，该方法能够有效构造适合 Twitter 的情感词典并以此评价 tweet 的观点化程度。最后我们重新标注了 TREC Tweets2011 数据，证明了我们的 Twitter 观点检索方法在 TREC 数据上依然有效。

再次，针对 Twitter 观点检索中时常包含大量的低质量观点，而以往的研究认为转发的 tweet 通常是高质量文本，我们提出了 Twitter 中传播观点发现的新任务。我们根据新任务的特点开发了一系列特征以此提高传播观点发现的效果，这些特征包括了 tweet 的传播度特征、tweet 的观点化特征、tweet 的文本质量特征。我们在真实的数据集上进行了测试，结果验证了我们设计的特征对于传播观点发现是有效的，并且我们的方法显著优于 BM25 方法和我们的观点检索方法。另外，令我们鼓舞的是我们的方法能够在 Twitter 中预测观点是否会被转发到达人预测的水平。

最后，针对以往 tweet 转发预测研究中忽视“谁”转发的问题的重要性，我们研究了在 Twitter 中发现信息传播者的问题。我们定义了 Twitter 中信息传播者发现的新任务，以此帮助理解 Twitter 中信息是如何传播的。我们同样开发了一系列特征，并将其应用到排序学习的机器学习框架中，具体的特征包括用户历史的转发信息，用户自身的社交媒体特征，用户使用 Twitter 的活跃时间，以及用户的个人兴趣。由于以往没有相同的工作，因此我们自己构造了数据，并发布了数据供以后的研究者继续使用。实验结果证明了我们方法对于 Twitter 中信息传播者发现是有效的，方法优于随机系统和基于用户历史转发记录的排序系统。最终我们发现用户历史转发信息，兴趣和活跃时间是决定信息传播者的重要因素。

6.2 工作展望

展望未来，社交媒体中的信息检索和传播分析研究及其相关方向还有很多工作需要完成。这里总结以下亟待探索的研究方向和路线：

1. 以 Twitter 为代表的社交媒体一个重要特点就是消息的实时性，许多研究工作都围绕在 Twitter 中发现实时信息展开，包括新事件发现[227–233]、实时灾害报道（如地震、疾病、火灾等）[234–238]，另外，TREC 的 Twitter 检索[99, 102, 105, 239–242] 也将实时性作为一个重要指标。本文的研究中，我们并未对话题检索和观点检索深入讨论实时性对检索效果的影响。这个问题的关键是找到与话题相关的时间点，如何找到这个相关时间点是我们未来研究的一个重点。

2. 本文的社交媒体研究仅仅以 Twitter 为代表展开，实际上流行的社交媒体还有很多，如 Facebook[2]、YouTube[3]、Flickr[4]等等。这些社交媒体肯定有自己独特的特点，在其数据上进行检索任务和传播分析需要研究其特殊性；另外，未来一个可能的需求就是多种社交媒体综合检索和跨媒体的信息传播，这就需要研究者在充分理解各种社交媒体的特点和人们对各种社交媒体不同的需求上，提出方法解决问题。

3. 目前跨媒体之间的研究是一个新的研究方向，主要是基于不同媒体之间的差异，利用各自的优点，解决其他媒体存在的问题。例如，有的研究利用维基百科的知识，帮助扩展 tweet 文本的语义，以此克服 tweet 文本短小，信息缺失的缺点[243, 244]；有些研究利用维基百科的访问信息帮助 Twitter 中的事件发现[245]；还有些研究各种媒体之间的联系以此帮助其他任务的解决[246–251]等等。未来我们将利用其他社交媒体的优点，帮助 Twitter 中已有的信息检索和传播分析任务进一步提高效果。

总之，社交媒体的研究还有许多问题等待着去解决，我们将继续深入研究相关问题。

[2]https://www.facebook.com/
[3]https://www.youtube.com/
[4]https://www.flickr.com/

致　谢

时间应该“浪费”在美好的事物上，我很庆幸自己有机会从事这些课题的研究。

本课题承蒙国家自然科学基金 -面上项目“融合网络特征的文本观点挖掘”(项目号：61170156）和国家自然科学基金 -青年科学基金项目“结合社会网络的网络信息传播分析研究”(项目号：61202337) 的资助，另外感谢国家留学基金委对本人在英国爱丁堡大学留学访问的经费资助。经济基础决定上层建筑，没有钱一切白搭。

感谢我的导师王挺教授对本人的精心指导和悉心培养，您严谨治学的态度让我受益终生。感谢在我英国爱丁堡大学留学期间指导我的 Miles Osbonre 博士，您对科研的敏锐洞察力以及极强的逻辑思维能力让我明白科研可以轻松地“玩”。

感谢国防科技大学自然语言处理组的张晓艳、刘伍颖、唐晋韬、魏登萍、周云、李岩、麻大顺、谢松县、刘培磊、岳大鹏、刘海池、汝承森、张文文、姜仁会、胡长龙、李欣奕，和你们一同探索自然语言处理的未知领域让人回味；感谢英国爱丁堡大学信息学院 348 办公室的 Saša Petrovic、Desmond Elliott、Diego Frassinelli、Eva Hasler、Michael Auli 和 Luke Shrimpton，和你们仔细讨论我的课题细节以及英文的论述让我十分收益。

感谢戴波、雷鸣、林正帅、毛先领、张湘莉兰、任洪广、王鹤、吴诚堃标注 Twitter 观点检索相关数据，感谢王铮标注 Twitter 传播观点检索相关数据。没有你们的无偿帮助，我相信我的课题研究不会如此顺利。感谢 Victor Lavrenko 博士和 Micha Elsner 博士给予我 Twitter 观点检索课题宝贵的意见，和你们讨论是我的荣幸。

感谢我从硕士到博士的室友邹丹、何明、陆化彪，一起研究课题慢慢“变老”的过程值得怀念；感谢我在爱丁堡期间最好的朋友王鹤和黄轩，人生得不多的知己足以；感谢在英国一同留学的杨俊刚、雷鸣、Chee-Ming Ting、陆亮、王铮、刘哲、马瑞、冯翌尧、卢恒、林正帅、曾旋、杨国利、何鑫、胡尽力、贺建森、杨丽莎、魏杰、罗家希，谢谢你们让我拥有那些留学的“回忆”，谢谢你们陪我那时的“孤独”。

感谢我的父母、奶奶、岳父、岳母对我生活上无微不至的照顾；最后感谢我的爱人柳意，谢谢你的支持，爱你！

参考文献

[1] Kaplan A M, Haenlein M. Users of the world, unite! The challenges and opportunities of Social Media [J]. Business horizons. 2010, 53 (1): 59–68.

[2] Eisenstein J. What to do about bad language on the internet [C]. In Proceedings of NAACL-HLT. 2013: 359–369.

[3] Jensen D, Neville J. Linkage and autocorrelation cause feature selection bias in relational learning [C]. In ICML. 2002: 259–266.

[4] Taskar B, Abbeel P, Wong M-F, et al. Label and link prediction in relational data [C]. In Proceedings of the IJCAI Workshop on Learning Statistical Models from Relational Data. 2003.

[5] Stringhini G, Kruegel C, Vigna G. Detecting spammers on social networks [C]. In Proceedings of the 26th Annual Computer Security Applications Conference. 2010: 1–9.

[6] Xiang R, Neville J, Rogati M. Modeling relationship strength in online social networks [C]. In Proceedings of the 19th international conference on World wide web. 2010: 981–990.

[7] Rossion B, Delvenne J-F, Debatisse D, et al. Spatio-temporal localization of the face inversion effect: an event-related potentials study [J]. Biological psychology. 1999, 50 (3): 173–189.

[8] Speriosu M, Sudan N, Upadhyay S, et al. Twitter polarity classification with label propagation over lexical links and the follower graph [C]. In Proceedings of the First workshop on Unsupervised Learning in NLP. 2011: 53–63.

[9] Mislove A, Viswanath B, Gummadi K P, et al. You are who you know: inferring user profiles in online social networks [C]. In Proceedings of the third ACM international conference on Web search and data mining. 2010: 251–260.

[10] Cambria E, White B, Durrani T, et al. Computational Intelligence for Natural Language Processing [Guest Editorial] [J]. Computational Intelligence Magazine, IEEE. 2014, 9 (1): 19–63.

[11] Cambria E, White B. Jumping NLP curves: A review of natural language processing research [J]. IEEE Computational Intelligence Magazine. 2014, 9 (2): 48–57.

[12] Zinko C. What is Biz Stone doing [J]. San Francisco Chronicle. 2009.

[13] Ellison N B, et al. Social network sites: Definition, history, and scholarship [J]. Journal of Computer-Mediated Communication. 2007, 13 (1): 210–230.

[14] Marlow C A. The structural determinants of media contagion [D]. [S. l.]: Massachusetts Institute of Technology, 2005.

[15] O'Reilly T, Milstein S. The twitter book [M]. O'Reilly, 2011.

[16] Honey C, Herring S C. Beyond microblogging: Conversation and collaboration via Twitter [C]. In System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on. 2009: 1–10.

[17] Golder S A, Huberman B A. Usage patterns of collaborative tagging systems [J]. Journal of information science. 2006, 32 (2): 198–208.

[18] Hong L, Dan O, Davison B D. Predicting popular messages in twitter [C]. In Proceedings of the 20th international conference companion on World wide web. 2011: 57–58.

[19] Toutanova K, Klein D, Manning C D, et al. Feature-rich part-of-speech tagging with a cyclic dependency network [C]. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. 2003: 173–180.

[20] Gimpel K, Schneider N, O'Connor B, et al. Part-of-speech tagging for twitter: Annotation, features, and experiments [R]. 2010.

[21] Owoputi O, O'Connor B, Dyer C, et al. Improved part-of-speech tagging for online conversational text with word clusters [C]. In Proceedings of NAACL-HLT. 2013: 380–390.

[22] Finkel J R, Grenager T, Manning C. Incorporating non-local information into information extraction systems by gibbs sampling [C]. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. 2005: 363–370.

[23] Ritter A, Clark S, Etzioni O, et al. Named entity recognition in tweets: an experimental study [C]. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2011: 1524–1534.

[24] Foster J, Cetinoglu O, Wagner J, et al. From news to comment: Resources and benchmarks for parsing the language of web 2.0 [J]. 2011.

[25] Han B, Baldwin T. Lexical Normalisation of Short Text Messages: Makn Sens a# twitter. [C]. In ACL. 2011: 368–378.

[26] Han B, Cook P, Baldwin T. Lexical normalization for social media text [J]. ACM Transactions on Intelligent Systems and Technology (TIST). 2013, 4 (1): 5.

[27] Han B, Cook P, Baldwin T. Automatically constructing a normalisation dictionary for microblogs [C]. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2012: 421–432.

[28] Liu F, Weng F, Wang B, et al. Insertion, Deletion, or Substitution? Normalizing Text Messages without Pre-categorization nor Supervision. [J]. ACL (Short Papers). 2011, 11: 71–76.

[29] Liu X, Zhou M, Wei F, et al. Joint inference of named entity recognition and normalization for tweets [C]. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. 2012: 526–535.

[30] Liu F, Weng F, Jiang X. A broad-coverage normalization system for social media language [C]. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. 2012: 1035–1044.

[31] Hassan H, Menezes A. Social text normalization using contextual graph random walks [C]. In Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics. 2013.

[32] Finin T, Murnane W, Karandikar A, et al. Annotating named entities in Twitter data with crowdsourcing [C]. In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. 2010: 80–88.

[33] Liu X, Zhang S, Wei F, et al. Recognizing Named Entities in Tweets. [C]. In ACL. 2011: 359–367.

[34] Li C, Weng J, He Q, et al. TwiNER: named entity recognition in targeted twitter stream [C]. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. 2012: 721–730.

[35] Liu X, Wei F, Zhang S, et al. Named entity recognition for tweets [J]. ACM Transactions on Intelligent Systems and Technology (TIST). 2013, 4 (1): 3.

[36] Liu X, Zhou M. Two-stage NER for tweets with clustering [J]. Information Processing & Management. 2012.

[37] Ritter A, Cherry C, Dolan B. Unsupervised modeling of Twitter conversations [C]. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2010: 172–180.

[38] Sharifi B, Hutton M-A, Kalita J. Summarizing microblogs automatically [C]. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2010: 685–688.

[39] Chakrabarti D, Punera K. Event Summarization Using Tweets. [C]. In ICWSM. 2011.

[40] Takamura H, Yokono H, Okumura M. Summarizing a document stream [C]. In Proceedings of the 33rd European conference on Advances in information retrieval. 2011: 177–188.

[41] Weng J-Y, Yang C-L, Chen B-N, et al. IMASS: an intelligent microblog analysis and summarization system [C]. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations. 2011: 133–138.

[42] Harabagiu S M, Hickl A. Relevance Modeling for Microblog Summarization. [C]. In ICWSM. 2011.

[43] Ren Z, Liang S, Meij E, et al. Personalized time-aware tweets summarization [C]. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA, 2013: 513–522.

[44] Shen C, Liu F, Weng F, et al. A Participant-based Approach for Event Summarization Using Twitter Streams [C]. In Proceedings of NAACL-HLT. 2013: 1152–1162.

[45] Judd J, Kalita J. Better Twitter Summaries? [C]. In Proceedings of NAACL-HLT. 2013: 445–449.

[46] Chang Y, Wang X, Mei Q, et al. Towards Twitter context summarization with user influence models [C]. In Proceedings of the sixth ACM international conference on Web search and data mining. 2013: 527–536.

[47] Brody S, Diakopoulos N. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs [C]. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2011: 562–570.

[48] Rocchio J J. Relevance feedback in information retrieval [J]. 1971.

[49] Joachims T. Optimizing search engines using clickthrough data [C]. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. 2002: 133–142.

[50] Xue G-R, Zeng H-J, Chen Z, et al. Optimizing web search using web click-through data [C]. In Proceedings of the thirteenth ACM international conference on Information and knowledge management. 2004: 118–126.

[51] Joachims T, Granka L, Pan B, et al. Accurately interpreting clickthrough data as implicit feedback [C]. In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. 2005: 154–161.

[52] Radlinski F, Joachims T. Query chains: learning to rank from implicit feedback [C]. In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. 2005: 239–248.

[53] Agichtein E, Brill E, Dumais S. Improving web search ranking by incorporating user behavior information [C]. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. 2006: 19–26.

[54] Agichtein E, Brill E, Dumais S, et al. Learning user interaction models for predicting web search result preferences [C]. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. 2006: 3–10.

[55] Radlinski F, Joachims T. Active exploration for learning rankings from clickthrough data [C]. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. 2007: 570–579.

[56] Bao S, Xue G, Wu X, et al. Optimizing web search using social annotations [C]. In Proceedings of the 16th international conference on World Wide Web. 2007: 501–510.

[57] Duh K, Kirchhoff K. Learning to rank with partially-labeled data [C]. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. 2008: 251–258.

[58] Aslam J A, Kanoulas E, Pavlu V, et al. Document selection methodologies for efficient and effective learning-to-rank [C]. In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. 2009: 468–475.

[59] Qin T, Liu T-Y, Xu J, et al. LETOR: A benchmark collection for research on learning to rank for information retrieval [J]. Information Retrieval. 2010, 13 (4): 346–374.

[60] Xu J, Chen C, Xu G, et al. Improving quality of training data for learning to rank using click-through data [C]. In Proceedings of the third ACM international conference on Web search and data mining. 2010: 171–180.

[61] Burges C, Shaked T, Renshaw E, et al. Learning to rank using gradient descent [C]. In Proceedings of the 22nd international conference on Machine learning. 2005: 89–96.

[62] Cao Y, Xu J, Liu T-Y, et al. Adapting ranking SVM to document retrieval [C]. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. 2006: 186–193.

[63] Xu J, Li H. Adarank: a boosting algorithm for information retrieval [C]. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. 2007: 391–398.

[64] Quoc C, Le V. Learning to rank with nonsmooth cost functions [J]. Proceedings of the Advances in Neural Information Processing Systems. 2007, 19: 193–200.

[65] Xu J, Liu T-Y, Lu M, et al. Directly optimizing evaluation measures in learning to rank [C]. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. 2008: 107–114.

[66] Valizadegan H, Jin R, Zhang R, et al. Learning to rank by optimizing ndcg measure [C]. In Advances in neural information processing systems. 2009: 1883–1891.

[67] Wang L, Lin J, Metzler D. Learning to efficiently rank [C]. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. 2010: 138–145.

[68] Dai N, Shokouhi M, Davison B D. Learning to rank for freshness and relevance [C]. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. 2011: 95–104.

[69] Chapelle O, Chang Y, Liu T-Y. Future directions in learning to rank. [J]. Journal of Machine Learning Research-Proceedings Track. 2011, 14: 91–100.

[70] boyd d, Golder S, Lotan G. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter [C]. In System Sciences (HICSS), 2010 43rd Hawaii International Conference on. 2010: 1–10.

[71] Yang Z, Guo J, Cai K, et al. Understanding retweeting behaviors in social networks [C]. In Proceedings of the 19th ACM international conference on Information and knowledge management. 2010: 1633–1636.

[72] Macskassy S A, Michelson M. Why do people retweet? anti-homophily wins the day! [C]. In ICWSM. 2011.

[73] Starbird K, Palen L. (How) will the revolution be retweeted?: information diffusion and the 2011 Egyptian uprising [C]. In Proceedings of the acm 2012 conference on computer supported cooperative work. 2012: 7–16.

[74] Comarela G, Crovella M, Almeida V, et al. Understanding factors that affect response rates in twitter [C]. In Proceedings of the 23rd ACM conference on Hypertext and social media. 2012: 123–132.

[75] Kupavskii A, Umnov A, Gusev G, et al. Predicting the Audience Size of a Tweet [C]. In Seventh International AAAI Conference on Weblogs and Social Media. 2013.

[76] Jenders M, Kasneci G, Naumann F. Analyzing and predicting viral tweets [C]. In Proceedings of the 22nd international conference on World Wide Web companion. 2013: 657–664.

[77] Ahmed M, Spagna S, Huici F, et al. A peek into the future: predicting the evolution of popularity in user generated content [C]. In Proceedings of the sixth ACM international conference on Web search and data mining. 2013: 607–616.

[78] Bao P, Shen H-W, Huang J, et al. Popularity prediction in microblogging network: a case study on sina weibo [C]. In Proceedings of the 22nd international conference on World Wide Web companion. 2013: 177–178.

[79] Stajner T, Thomee B, Popescu A, et al. Automatic selection of social media responses to news [J]. ACM WSDM. 2013.

[80] Kothari A, Magdy W, Kareem Darwish A M, et al. Detecting Comments on News Articles in Microblogs [J]. ICWSM 2013. 2013.

[81] Li C, Sun A, Weng J, et al. Exploiting hybrid contexts for Tweet segmentation [C]. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. 2013: 523–532.

[82] Zhang J, Minami K, Kawai Y, et al. Personalized Web Search Using Emotional Features [M] // Zhang J, Minami K, Kawai Y, et al. Availability, Reliability, and Security in Information Systems and HCI. Springer, 2013: 2013: 69–83.

[83] Luo Z, Osborne M, Petrovic S, et al. Improving Twitter Retrieval by Exploiting Structural Information. [C]. In AAAI. 2012.

[84] Luo Z, Osborne M, Wang T. Opinion Retrieval in Twitter. [C]. In ICWSM. 2012.

[85] Luo Z, Wang T. Propagated Opinion Retrieval in Twitter. [C]. In WISE. 2013.

[86] Luo Z, Osborne M, Tang J, et al. Who will retweet me?: finding retweeters in twitter [C]. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. 2013: 869–872.

[87] Efron M. Hashtag retrieval in a microblogging environment [C]. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. 2010: 787–788.

[88] Duan Y, Jiang L, Qin T, et al. An empirical study on learning to rank of tweets [C]. In Proceedings of the 23rd International Conference on Computational Linguistics. 2010: 295–303.

[89] Massoudi K, Tsagkias M, de Rijke M, et al. Incorporating query expansion and quality indicators in searching microblog posts [M] // Massoudi K, Tsagkias M, de Rijke M, et al. Advances in Information Retrieval. Springer, 2011: 2011: 362–367.

[90] Naveed N, Gottron T, Kunegis J, et al. Searching microblogs: coping with sparsity and document quality [C]. In Proceedings of the 20th ACM international conference on Information and knowledge management. 2011: 183–188.

[91] Callan J P. Passage-level evidence in document retrieval [C]. In Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval. 1994: 302–310.

[92] Ahnizeret K, Fernandes D, Cavalcanti J M, et al. Information retrieval aware web site modelling and generation [M] // Ahnizeret K, Fernandes D, Cavalcanti J M, et al. Conceptual Modeling–ER 2004. Springer, 2004: 2004: 402–419.

[93] Fernandes D, de Moura E S, Ribeiro-Neto B, et al. Computing block importance for searching on web sites [C]. In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. 2007: 165–174.

[94] de Moura E S, Fernandes D, Ribeiro-Neto B, et al. Using structural information to improve search in Web collections [J]. Journal of the American Society for Information Science and Technology. 2010, 61 (12): 2503–2513.

[95] Cai D, Yu S, Wen J-R, et al. Block-based web search [C]. In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. 2004: 456–463.

[96] O'Connor B, Krieger M, Ahn D. Tweetmotif: Exploratory search and topic summarization for twitter [J]. Proceedings of ICWSM. 2010: 2–3.

[97] Metzler D, Cai C. USC/ISI at TREC 2011: Microblog Track. [C]. In TREC. 2011.

[98] Miyanishi T, Okamura N, Liu X, et al. TREC 2011 Microblog Track Experiments at Kobe University. [C]. In TREC. 2011.

[99] Zhang X, He B, Luo T, et al. Query-biased learning to rank for real-time twitter search [C]. In Proceedings of the 21st ACM international conference on Information and knowledge management. 2012: 1915–1919.

[100] Darwish K, Magdy W, Mourad A. Language processing for arabic microblog retrieval [C]. In Proceedings of the 21st ACM international conference on Information and knowledge management. 2012: 2427–2430.

[101] Zhang X, He B, Luo T. Transductive Learning for Real-Time Twitter Search. [C]. In ICWSM. 2012.

[102] Miyanishi T, Seki K, Uehara K. Combining recency and topic-dependent temporal variation for microblog search [M] // Miyanishi T, Seki K, Uehara K. Advances in Information Retrieval. Springer, 2013: 2013: 331–343.

[103] Choi J, Croft W B, Kim J Y. Quality models for microblog retrieval [C]. In Proceedings of the 21st ACM international conference on Information and knowledge management. 2012: 1834–1838.

[104] Ferguson P, O’ Hare N, Lanagan J, et al. An investigation of term weighting approaches for microblog retrieval [M] // Ferguson P, O’ Hare N, Lanagan J, et al. Advances in Information Retrieval. Springer, 2012: 2012: 552–555.

[105] Amati G, Amodeo G, Gaibisso C. Survival analysis for freshness in microblogging search [C]. In Proceedings of the 21st ACM international conference on Information and knowledge management. 2012: 2483–2486.

[106] Tjong Kim Sang E F, De Meulder F. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition [C]. In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4. 2003: 142–147.

[107] Lafferty J, McCallum A, Pereira F C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data [J]. 2001.

[108] Gimpel K, Schneider N, O’Connor B, et al. Part-of-speech tagging for Twitter: annotation, features, and experiments [C]. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2. 2011: 42–47.

[109] Pang B, Lee L. Opinion mining and sentiment analysis [J]. Foundations and trends in information retrieval. 2008, 2 (1-2): 1–135.

[110] Gouws S, Metzler D, Cai C, et al. Contextual bearing on linguistic variation in social media [C]. In Proceedings of the Workshop on Languages in Social Media. 2011: 20–29.

[111] Wei Z, Zhou L, Li B, et al. Exploring Tweets Normalization and Query Time Sensitivity for Twitter Search [J]. TREC-2011. 2011.

[112] Liu T-Y. Learning to rank for information retrieval [J]. Foundations and Trends in Information Retrieval. 2009, 3 (3): 225–331.

[113] Page L, Brin S, Motwani R, et al. The PageRank citation ranking: bringing order to the web. [J]. 1999.

[114] Voorhees E M. Natural language processing and information retrieval [M] // Voorhees E M. Information Extraction. Springer, 1999: 1999: 32–48.

[115] Voorhees E M. Variations in relevance judgments and the measurement of retrieval effectiveness [J]. Information processing & management. 2000, 36 (5): 697–716.

[116] Jansen B J, Zhang M, Sobel K, et al. Twitter power: Tweets as electronic word of mouth [J]. Journal of the American society for information science and technology. 2009, 60 (11): 2169–2188.

[117] O'Connor B, Balasubramanyan R, Routledge B R, et al. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. [J]. ICWSM. 2010, 11: 122–129.

[118] Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market [J]. Journal of Computational Science. 2011, 2 (1): 1–8.

[119] Zhang W, Yu C, Meng W. Opinion retrieval from blogs [C]. In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. 2007: 831–840.

[120] He B, Macdonald C, He J, et al. An effective statistical approach to blog post opinion retrieval [C]. In Proceedings of the 17th ACM conference on Information and knowledge management. 2008: 1063–1072.

[121] Zhang M, Ye X. A generation model to unify topic relevance and lexicon-based sentiment for opinion retrieval [C]. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. 2008: 411–418.

[122] Gerani S, Carman M J, Crestani F. Proximity-based opinion retrieval [C]. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. 2010: 403–410.

[123] Gerani S, Carman M, Crestani F. Aggregation methods for proximity-based opinion retrieval [J]. ACM Transactions on Information Systems (TOIS). 2012, 30 (4): 26.

[124] Ounis I, de Rijke M, Macdonald C, et al. Overview of the TREC blog track 2006 [C]. In Proceedings of the 15th Text REtrieval Conference (TREC 2006). 2006.

[125] Macdonald C, Ounis I, Soboroff I. Overview of the TREC 2007 Blog Track. [C]. In TREC. 2007: 31–43.

[126] Ounis I, Macdonald C, Soboroff I. Overview of the TREC-2008 blog track [R]. 2008.

[127] Macdonald C, Santos R L, Ounis I, et al. Blog track research at TREC [C]. In ACM SIGIR Forum. 2010: 58–75.

[128] Robertson S E, Walker S, Jones S, et al. Okapi at TREC-3 [J]. NIST SPECIAL PUBLICATION SP. 1995: 109–109.

[129] Robertson S E, Walker S, Beaulieu M, et al. Okapi at TREC-4 [C]. In Proceedings of the fourth text retrieval conference. 1996: 73–97.

[130] Salton G, Wong A, Yang C-S. A vector space model for automatic indexing [J]. Communications of the ACM. 1975, 18 (11): 613–620.

[131] Jiang L, Yu M, Zhou M, et al. Target-dependent Twitter Sentiment Classification. [C]. In ACL. 2011: 151–160.

[132] Liu K-L, Li W-J, Guo M. Emoticon Smoothed Language Models for Twitter Sentiment Analysis. [C]. In AAAI. 2012.

[133] Meng X, Wei F, Liu X, et al. Entity-centric topic-oriented opinion summarization in twitter [C]. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. 2012: 379–387.

[134] 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.

[135] Davidov D, Tsur O, Rappoport A. Enhanced sentiment learning using twitter hashtags and smileys [C]. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters. 2010: 241–249.

[136] Barbosa L, Feng J. Robust sentiment detection on twitter from biased and noisy data [C]. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters. 2010: 36–44.

[137] Kouloumpis E, Wilson T, Moore J. Twitter sentiment analysis: The Good the Bad and the OMG! [C]. In ICWSM. 2011.

[138] Agarwal A, Xie B, Vovsha I, et al. Sentiment analysis of twitter data [C]. In Proceedings of the Workshop on Languages in Social Media. 2011: 30–38.

[139] Tan C, Lee L, Tang J, et al. User-level sentiment analysis incorporating social networks [C]. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. 2011: 1397–1405.

[140] Wang X, Wei F, Liu X, et al. Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach [C]. In Proceedings of the 20th ACM international conference on Information and knowledge management. 2011: 1031–1040.

[141] Hu X, Tang L, Tang J, et al. Exploiting social relations for sentiment analysis in microblogging [C]. In Proceedings of the sixth ACM international conference on Web search and data mining. 2013: 537–546.

[142] Mukherjee S, Malu A, AR B, et al. TwiSent: a multistage system for analyzing sentiment in twitter [C]. In Proceedings of the 21st ACM international conference on Information and knowledge management. 2012: 2531–2534.

[143] Saif H, He Y, Alani H. Semantic Smoothing for Twitter Sentiment Analysis [C]. In Proceeding of the 10th International Semantic Web Conference (ISWC). 2011.

[144] Marchetti-Bowick M, Chambers N. Learning for microblogs with distant supervision: Political forecasting with twitter [C]. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. 2012: 603–612.

[145] Aisopos F, Papadakis G, Tserpes K, et al. Content vs. context for sentiment analysis: a comparative analysis over microblogs [C]. In Proceedings of the 23rd ACM conference on Hypertext and social media. 2012: 187–196.

[146] MARTÍNEZ-CÁMARA E, MARTÍN-VALDIVIA M T, UREÑA-LÓPEZ L A, et al. Sentiment analysis in Twitter [J]. Natural Language Engineering: 1–28.

[147] Kontopoulos E, Berberidis C, Dergiades T, et al. Ontology-based sentiment analysis of twitter posts [J]. Expert Systems with Applications. 2013.

[148] Hu X, Tang J, Gao H, et al. Unsupervised sentiment analysis with emotional signals [C]. In Proceedings of the 22nd international conference on World Wide Web. 2013: 607–618.

[149] Bollen J, Mao H, Pepe A. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. [C]. In ICWSM. 2011.

[150] Thelwall M, Buckley K, Paltoglou G. Sentiment in Twitter events [J]. Journal of the American Society for Information Science and Technology. 2011, 62 (2): 406–418.

[151] Bermingham A, Smeaton A F. Classifying sentiment in microblogs: is brevity an advantage? [C]. In Proceedings of the 19th ACM international conference on Information and knowledge management. 2010: 1833–1836.

[152] Paltoglou G, Thelwall M. Twitter, MySpace, Digg: Unsupervised sentiment analysis in social media [J]. ACM Transactions on Intelligent Systems and Technology (TIST). 2012, 3 (4): 66.

[153] Chung J E, Mustafaraj E. Can collective sentiment expressed on twitter predict political elections? [C]. In AAAI. 2011.

[154] Conover M, Ratkiewicz J, Francisco M, et al. Political Polarization on Twitter. [C]. In ICWSM. 2011.

[155] González-Ibáñez R, Muresan S, Wacholder N. Identifying Sarcasm in Twitter: A Closer Look. [C]. In ACL (Short Papers). 2011: 581–586.

[156] Eguchi K, Lavrenko V. Sentiment retrieval using generative models [C]. In Proceedings of the 2006 conference on empirical methods in natural language processing. 2006: 345–354.

[157] Huang X, Croft W B. A unified relevance model for opinion retrieval [C]. In Proceedings of the 18th ACM conference on Information and knowledge management. 2009: 947–956.

[158] Gerani S, Carman M J, Crestani F. Investigating learning approaches for blog post opinion retrieval [M] // Gerani S, Carman M J, Crestani F. Advances in Information Retrieval. Springer, 2009: 2009: 313–324.

[159] Mishne G. Multiple ranking strategies for opinion retrieval in blogs [C]. In Online Proceedings of TREC. 2006.

[160] Na S-H, Lee Y, Nam S-H, et al. Improving opinion retrieval based on query-specific sentiment lexicon [M] // Na S-H, Lee Y, Nam S-H, et al. Advances in Information Retrieval. Springer, 2009: 2009: 734–738.

[161] Santos R L, He B, Macdonald C, et al. Integrating proximity to subjective sentences for blog opinion retrieval [M] // Santos R L, He B, Macdonald C, et al. Advances in Information Retrieval. Springer, 2009: 2009: 325–336.

[162] Zhang W, Jia L, Yu C, et al. Improve the effectiveness of the opinion retrieval and opinion polarity classification [C]. In Proceedings of the 17th ACM conference on Information and knowledge management. 2008: 1415–1416.

[163] Vechtomova O. Facet-based opinion retrieval from blogs [J]. Information processing & management. 2010, 46 (1): 71–88.

[164] Vechtomova O. Using Subjective Adjectives in Opinion Retrieval from Blogs. [C]. In TREC. 2007.

[165] Bermingham A, Smeaton A F. A study of inter-annotator agreement for opinion retrieval [C]. In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. 2009: 784–785.

[166] Jia L, Yu C, Meng W. The effect of negation on sentiment analysis and retrieval effectiveness [C]. In Proceedings of the 18th ACM conference on Information and knowledge management. 2009: 1827–1830.

[167] Li B, Zhou L, Feng S, et al. A unified graph model for sentence-based opinion retrieval [C]. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. 2010: 1367–1375.

[168] Lee S-W, Lee J-T, Song Y-I, et al. High precision opinion retrieval using sentiment-relevance flows [C]. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. 2010: 817–818.

[169] Gerani S, Keikha M, Crestani F. Aggregating multiple opinion evidence in proximity-based opinion retrieval [C]. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. 2011: 1199–1200.

[170] He B, Peng J, Ounis I. Fitting score distribution for blog opinion retrieval [C]. In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. 2009: 688–689.

[171] Amati G, Amodeo G, Capozio V, et al. On performance of topical opinion retrieval [C]. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. 2010: 777–778.

[172] Seki K, Uehara K. Adaptive subjective triggers for opinionated document retrieval [C]. In Proceedings of the Second ACM International Conference on Web Search and Data Mining. 2009: 25–33.

[173] Orimaye S O. Sentence-level contextual opinion retrieval [C]. In Proceedings of the 20th international conference companion on World wide web. 2011: 403–408.

[174] Guo L, Wan X. Exploiting syntactic and semantic relationships between terms for opinion retrieval [J]. Journal of the American Society for Information Science and Technology. 2012, 63 (11): 2269–2282.

[175] Xu X, Tan S, Liu Y, et al. Find me opinion sources in blogosphere: a unified framework for opinionated blog feed retrieval [C]. In Proceedings of the fifth ACM international conference on Web search and data mining. 2012: 583–592.

[176] Orimaye S O, Alhashmi S M, Siew E-G. Can predicate-argument structures be used for contextual opinion retrieval from blogs? [J]. World Wide Web. 2012: 1–29.

[177] Amati G, Ambrosi E, Bianchi M, et al. Automatic construction of an opinion-term vocabulary for ad hoc retrieval [M] // Amati G, Ambrosi E, Bianchi M, et al. Advances in Information Retrieval. Springer, 2008: 2008: 89–100.

[178] Jijkoun V, de Rijke M, Weerkamp W. Generating focused topic-specific sentiment lexicons [C]. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. 2010: 585–594.

[179] Joachims T. Advances in kernel methods [M] // Schölkopf B, Burges C J C, Smola A J: chapter Making large-scale support vector machine learning practical, 169–184. Cambridge, MA, USA: MIT Press, 1999: 1999: 169–184.

[180] Manning C D, Raghavan P, Schtze H. Introduction to Information Retrieval [M]. New York, NY, USA: Cambridge University Press, 2008.

[181] Go A, Bhayani R, Huang L. Twitter Sentiment Classification using Distant Supervision [J]. Processing. 2009: 1–6.

[182] West T, Turner L. Introducing communication theory: analysis and application with powerweb [J]. 2003.

[183] 朱海青. 微博传播中“沉默的螺旋” [J]. 青年记者. 2013 (12): 43–44.

[184] Petrovic S, Osborne M, Lavrenko V. RT to Win! Predicting Message Propagation in Twitter. [C]. In ICWSM. 2011.

[185] Pang B, Lee L. Opinion Mining and Sentiment Analysis [J]. Found. Trends Inf. Retr. 2008, 2 (1-2): 1–135.

[186] Agarwal D, Chen B-C, Pang B. Personalized recommendation of user comments via factor models [C]. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA, 2011: 571–582.

[187] Suh B, Hong L, Pirolli P, et al. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network [C]. In Social Computing (SocialCom), 2010 IEEE Second International Conference on. 2010: 177–184.

[188] Zaman T R, Herbrich R, Van Gael J, et al. Predicting information spreading in twitter [C]. In Workshop on Computational Social Science and the Wisdom of Crowds, NIPS. 2010: 17599–601.

[189] Naveed N, Gottron T, Kunegis J, et al. Bad News Travel Fast: A Content-based Analysis of Interestingness on Twitter [C]. In WebSci '11: Proceedings of the 3rd International Conference on WebScience. 2011.

[190] Artzi Y, Pantel P, Gamon M. Predicting responses to microblog posts [C]. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2012: 602–606.

[191] Gupta M, Gao J, Zhai C, et al. Predicting future popularity trend of events in microblogging platforms [J]. Proceedings of the American Society for Information Science and Technology. 2012, 49 (1): 1–10.

[192] Feng W, Wang J. Retweet or not?: personalized tweet re-ranking [C]. In Proceedings of the sixth ACM international conference on Web search and data mining. 2013: 577–586.

[193] Kong S, Feng L, Sun G, et al. Predicting lifespans of popular tweets in microblog [C]. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. 2012: 1129–1130.

[194] Hong L, Doumith A S, Davison B D. Co-factorization machines: modeling user interests and predicting individual decisions in Twitter [C]. In Proceedings of the sixth ACM international conference on Web search and data mining. 2013: 557–566.

[195] Stieglitz S, Dang-Xuan L. Political Communication and Influence through Microblogging–An Empirical Analysis of Sentiment in Twitter Messages and Retweet Behavior [C]. In System Science (HICSS), 2012 45th Hawaii International Conference on. 2012: 3500–3509.

[196] Paltoglou G, Buckley K. Subjectivity annotation of the microblog 2011 realtime adhoc relevance judgments [M] // Paltoglou G, Buckley K. Advances in Information Retrieval. Springer, 2013: 2013: 344–355.

[197] Kim S-M, Pantel P, Chklovski T, et al. Automatically assessing review helpfulness [C]. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. 2006: 423–430.

[198] Zhang Z, Varadarajan B. Utility scoring of product reviews [C]. In Proceedings of the 15th ACM international conference on Information and knowledge management. 2006: 51–57.

[199] Liu J, Cao Y, Lin C-Y, et al. Low-Quality Product Review Detection in Opinion Summarization. [C]. In EMNLP-CoNLL. 2007: 334–342.

[200] Ghose A, Ipeirotis P G. Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics [J]. IEEE Trans. on Knowl. and Data Eng. 2011, 23 (10): 1498–1512.

[201] Liu Y, Huang X, An A, et al. Modeling and predicting the helpfulness of online reviews [C]. In Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on. 2008: 443–452.

[202] Danescu-Niculescu-Mizil C, Kossinets G, Kleinberg J, et al. How opinions are received by online communities: a case study on amazon.com helpfulness votes [C]. In Proceedings of the 18th international conference on World wide web. New York, NY, USA, 2009: 141–150.

[203] Lu Y, Tsaparas P, Ntoulas A, et al. Exploiting social context for review quality prediction [C]. In Proceedings of the 19th international conference on World wide web. New York, NY, USA, 2010: 691–700.

[204] Tsur O, Rappoport A. RevRank: A Fully Unsupervised Algorithm for Selecting the Most Helpful Book Reviews. [C]. In ICWSM. 2009.

[205] Siersdorfer S, Chelaru S, Nejdl W, et al. How useful are your comments?: analyzing and predicting youtube comments and comment ratings [C]. In Proceedings of the 19th international conference on World wide web. 2010: 891–900.

[206] Jindal N, Liu B, Lim E-P. Finding unusual review patterns using unexpected rules [C]. In Proceedings of the 19th ACM international conference on Information and knowledge management. 2010: 1549–1552.

[207] Chen B-C, Guo J, Tseng B, et al. User reputation in a comment rating environment [C]. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. 2011: 159–167.

[208] Tsaparas P, Ntoulas A, Terzi E. Selecting a comprehensive set of reviews [C]. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. 2011: 168–176.

[209] Agarwal D, Chen B-C, Pang B. Personalized recommendation of user comments via factor models [C]. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2011: 571–582.

[210] Shmueli E, Kagian A, Koren Y, et al. Care to comment?: recommendations for commenting on news stories [C]. In Proceedings of the 21st international conference on World Wide Web. 2012: 429–438.

[211] Dalal O, Sengemedu S H, Sanyal S. Multi-objective ranking of comments on web [C]. In Proceedings of the 21st international conference on World Wide Web. 2012: 419–428.

[212] Mishra A, Rastogi R. Semi-supervised correction of biased comment ratings [C]. In Proceedings of the 21st international conference on World Wide Web. 2012: 181–190.

[213] Mahajan D K, Rastogi R, Tiwari C, et al. LogUCB: an explore-exploit algorithm for comments recommendation [C]. In Proceedings of the 21st ACM international conference on Information and knowledge management. 2012: 6–15.

[214] Jain V, Galbrun E. Topical organization of user comments and application to content recommendation [C]. In Proceedings of the 22nd international conference on World Wide Web companion. 2013: 61–62.

[215] Jindal N, Liu B. Analyzing and detecting review spam [C]. In Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on. 2007: 547–552.

[216] Lim E-P, Nguyen V-A, Jindal N, et al. Detecting product review spammers using rating behaviors [C]. In Proceedings of the 19th ACM international conference on Information and knowledge management. 2010: 939–948.

[217] Wang G, Xie S, Liu B, et al. Review graph based online store review spammer detection [C]. In Data Mining (ICDM), 2011 IEEE 11th International Conference on. 2011: 1242–1247.

[218] Mukherjee A, Liu B, Wang J, et al. Detecting group review spam [C]. In Proceedings of the 20th international conference companion on World wide web. 2011: 93–94.

[219] Mukherjee A, Liu B, Glance N. Spotting fake reviewer groups in consumer reviews [C]. In Proceedings of the 21st international conference on World Wide Web. 2012: 191–200.

[220] Li F, Huang M, Yang Y, et al. Learning to identify review spam [C]. In Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Three. 2011: 2488–2493.

[221] Xie S, Wang G, Lin S, et al. Review spam detection via temporal pattern discovery [C]. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. 2012: 823–831.

[222] Morales A, Sun H, Yan X. Synthetic review spamming and defense [C]. In Proceedings of the 22nd international conference on World Wide Web companion. 2013: 155–156.

[223] Katz S. Estimation of probabilities from sparse data for the language model component of a speech recognizer [J]. Acoustics, Speech and Signal Processing, IEEE Transactions on. 1987, 35 (3): 400–401.

[224] Nagarajan M, Purohit H, Sheth A P. A Qualitative Examination of Topical Tweet and Retweet Practices. [C]. In ICWSM. 2010.

[225] Cha M, Mislove A, Gummadi K P. A measurement-driven analysis of information propagation in the flickr social network [C]. In Proceedings of the 18th international conference on World wide web. 2009: 721–730.

[226] Kwak H, Lee C, Park H, et al. What is Twitter, a social network or a news media? [C]. In Proceedings of the 19th international conference on World wide web. 2010: 591–600.

[227] Petrović S, Osborne M, Lavrenko V. Streaming first story detection with application to twitter [C]. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2010: 181–189.

[228] Becker H, Naaman M, Gravano L. Beyond Trending Topics: Real-World Event Identification on Twitter. [C]. In ICWSM. 2011.

[229] Weng J, Lee B-S. Event Detection in Twitter. [C]. In ICWSM. 2011.

[230] Naaman M, Becker H, Gravano L. Hip and trendy: Characterizing emerging trends on Twitter [J]. Journal of the American Society for Information Science and Technology. 2011, 62 (5): 902–918.

[231] Benson E, Haghighi A, Barzilay R. Event discovery in social media feeds [C]. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. 2011: 389–398.

[232] Petrović S, Osborne M, Lavrenko V. Using paraphrases for improving first story detection in news and Twitter [C]. In Proceedings of The 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2012: 338–346.

[233] Kanhabua N, Nejdl W. Understanding the Diversity of Tweets in the Time of Outbreaks [C]. In Proceedings of the 22nd international conference on World Wide Web companion. 2013: 1335–1342.

[234] Sakaki T, Okazaki M, Matsuo Y. Earthquake shakes Twitter users: real-time event detection by social sensors [C]. In Proceedings of the 19th international conference on World wide web. 2010: 851–860.

[235] Paul M J, Dredze M. You Are What You Tweet: Analyzing Twitter for Public Health. [C]. In ICWSM. 2011.

[236] Aramaki E, Maskawa S, Morita M. Twitter catches the flu: Detecting influenza epidemics using twitter [C]. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2011: 1568–1576.

[237] Abel F, Hauff C, Houben G-J, et al. Twitcident: fighting fire with information from social web streams [C]. In Proceedings of the 21st international conference companion on World Wide Web. 2012: 305–308.

[238] Yin J, Karimi S, Robinson B, et al. ESA: emergency situation awareness via microbloggers [C]. In Proceedings of the 21st ACM international conference on Information and knowledge management. 2012: 2701–2703.

[239] Efron M, Golovchinsky G. Estimation methods for ranking recent information [C]. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. 2011: 495–504.

[240] Metzler D, Cai C, Hovy E. Structured event retrieval over microblog archives [C]. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2012: 646–655.

[241] Soboroff I, McCullough D, Lin J, et al. Evaluating real-time search over tweets [J]. Proc. ICWSM. 2012: 943–961.

[242] Choi J, Croft W B. Temporal models for microblogs [C]. In Proceedings of the 21st ACM international conference on Information and knowledge management. 2012: 2491–2494.

[243] Meij E, Weerkamp W, de Rijke M. Adding semantics to microblog posts [C]. In Proceedings of the fifth ACM international conference on Web search and data mining. 2012: 563–572.

[244] Cassidy T, Ji H, Ratinov L-A, et al. Analysis and Enhancement of Wikification for Microblogs with Context Expansion. [C]. In COLING. 2012: 441–456.

[245] Osborne M, Petrovic S, McCreadie R, et al. Bieber no more: First story detection using Twitter and Wikipedia [C]. In Proceedings of the SIGIR Workshop on Time-aware Information Access. 2012.

[246] Dong A, Zhang R, Kolari P, et al. Time is of the essence: improving recency ranking using twitter data [C]. In Proceedings of the 19th international conference on World wide web. 2010: 331–340.

[247] Phelan O, McCarthy K, Bennett M, et al. On using the real-time web for news recommendation & discovery [C]. In Proceedings of the 20th international conference companion on World wide web. 2011: 103–104.

[248] Tsagkias M, de Rijke M, Weerkamp W. Linking online news and social media [C]. In Proceedings of the fourth ACM international conference on Web search and data mining. 2011: 565–574.

[249] Becker H, Iter D, Naaman M, et al. Identifying content for planned events across social media sites [C]. In Proceedings of the fifth ACM international conference on Web search and data mining. 2012: 533–542.

[250] Petrovic S, Osborne M, McCreadie R, et al. Can Twitter replace Newswire for breaking news? [C]. In Proceedings of the 7th International AAAI Conference on Weblogs and Social Media. 2013.

[251] Chang Y, Dong A, Kolari P, et al. Improving recency ranking using twitter data [J]. ACM Transactions on Intelligent Systems and Technology (TIST). 2013, 4 (1): 4.

作者在学期间取得的学术成果

发表的学术论文

[1] Zhunchen Luo, Miles Osborne, Sasa Petrovic and Ting Wang. Improving Twitter Retrieval by Exploiting Structural Information. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence* (**AAAI 2012**), Toronto, Canada, July 2012. （**CCF A 类会议，人工智能领域顶级会议**）

[2] Zhunchen Luo, Miles Osborne, Jintao Tang and Ting Wang. Who Will Retweet Me? Finding Retweeters in Twitter. In *Proceedings of the Thirty-Sixth International ACM SIGIR Conference on Research and Development in Information Retrieval* (**SIGIR 2013**), Dublin, Ireland, July 2013.（**CCF A 类会议，信息检索领域顶级会议，获得会议旅行奖金 1300 美元**）

[3] Zhunchen Luo, Miles Osborne and Ting Wang. Opinion Retrieval in Twitter. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media* (**AAAI-ICWSM 2012**), Dublin, Ireland, June 2012. （**社交媒体领域顶级会议，获得会议旅行奖金 300 美元**）

[4] Zhunchen Luo, Miles Osborne and Ting Wang. An Effective Approach to Tweets Opinion Retrieval. To appear in **World Wide Web Journal**. （**SCI 期刊，影响因子 1.196**）

[5] Zhunchen Luo, Jintao Tang and Ting Wang. Propagated Opinion Retrieval in Twitter. In *Proceedings of the Fourteenth International Conference on Web Information System Engineering* (**WISE 2013**), Nanjing, China, October 2013.（**CCF C 类会议，信息检索与数据挖掘领域重要会议**）

[6] Zhunchen Luo, Jintao Tang and Ting Wang. Improving Keyphrase Extraction from Web News by Exploiting Comments Information. In *Proceedings of the Fifteenth International Asia-Pacific Web Conference* (**APWeb 2013**), Sydney, Australia, April 2013. （**CCF C 类会议，信息检索与数据挖掘领域重要会议**）

[7] Zhunchen Luo, Lan Rao, Chengsen Ru and Ting Wang. Finding High-Quality Posts from Microblogging Conversations. In *the Eighth International Conference on Modeling Decisions for Artificial Intelligence (**MDAI 2011**)*, Changsha, China, July, 2011.

[8] 罗准辰，王挺. 基于分离模型的中文关键词提取算法研究. 中文信息学报, 2009，23 (01): 63-70.

[9] 罗准辰，王挺. 搜索词同现网络研究. 第六届全国信息检索学术会议 (**CCIR 2010**)，镜泊湖，2010 年 8 月.