学号 1306010103

年级 2013 级

河 海 大 學

本科毕业设计

事件热度趋势的建模和可视化分析

专　　业 计算机科学与技术

姓　　名 徐丽

指导教师 唐彦

评 阅 人

2017 年 6 月

中国　南京

BACHELOR'S DEGREE THESIS
OF HOHAI UNIVERSITY

Modeling and Visualization Analysis of Event Popularity Trend

College： Hohai University

Subject： Computer Science and Technology

Name： Li Xu

Directed by： Associate Professor, Yan Tang

NANJING CHINA

学术声明：

郑 重 声 明

本人呈交的毕业设计，是在导师的指导下，独立进行研究工作所取得的成果，所有数据、图片资料真实可靠。尽我所知，除文中已经注明引用的内容外，本设计（论文）的研究成果不包含他人享有著作权的内容。对本设计（论文）所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确的方式标明。本设计（论文）的知识产权归属于培养单位。

本人签名：　　日期：

摘　要

在如今的大数据时代，突发性事件在各种社交媒体和搜索引擎等平台上迅速广泛传播，产生大量的相关数据。目前的许多研究专注于特定媒体平台上事件的传播趋势分析，很少的工作是放在如何进行热词识别和联系不同的社交平台来进行事件热度预测的。相反，本文的研究主要是针对这两个方面，提出了两个模型用于分析和预测事件热度：事件热度模型，事件预测模型。

事件热度模型：本文提出一个不区分社交平台的能够计算任意时间内的事件热度的算法，其中包括热词选取和事件热度计算两个板块。热词选取板块设计了一个新的算法用以选出最能代表事件的热词。事件热度计算板块，结合第一个板块得到的最适合的热词集，计算任意时间间隔的事件热度。最后，将这些时间单元整合归一化就可以得到某段时间的事件热度向量。

事件预测模型的输入是第一个模型的输出：事件热度向量。事件预测模型的主体是混合概率预测模型，用不同的函数模型拟合已知的事件热度趋势并借此预测下一个点的事件热度。为了弥补混合概率预测模型在峰值捕捉方面的缺陷，本文引入了跨平台的思想，即对于在不同平台有着大致相似发展趋势的事件，发展较快的平台对于发展较慢的平台的事件热度预测有着很好的借鉴意义。

在提出两个模型后，本文选取了 8 个真实事件在微博和百度两个媒体平台上的 16 个数据集，进行了 7 个实验。前 4 个实验是针对事件热度模型的，通过实验我们可以验证热词选取算法的有效性和事件热度模型具有较高的灵敏度。后 3 个实验是针对事件预测模型，实验表明跨平台混合概率预测模型是比较准确的，跨平台分析对事件热度预测是有效的，能够有效帮助混合概率预测模型提高预测精度。

关键词：社交网络；搜索引擎；事件热度；热词选取；事件热度预测；

ABSTRACT

In today's big data era, sudden events spread quickly on platforms such as social media and search engines, producing large volume of relevant data. Many of the current studies focus on the dissemination trend analysis of events on a specific media platform, but very little work has been carried out on identifying the trendy words and combine different social platforms to predict events popularity. Therefore, the research of this paper mainly focuses on these problems, and puts forward two models for analysis and prediction of the event popularity: the event popularity model and the event prediction model.

The event popularity model: this paper proposes an algorithm that can calculate the heat of events in any time, regardless of the social platforms. And it includes two modules: hot word selection and event popularity calculation, The hot word selection module proposes a new algorithm to select the most representative hot words of the event. Event popularity computing module, combined with the most suitable hot word set obtained by the first module, calculates heat of the event over any time interval. Finally, the vector of the event popularity can be gained through the integration normalization process over time unit.

The output results of the first model can be used as the input data of the second one. The subject of the event prediction model is a hybrid probability prediction model, which is used to fit the known event popularity trend with different function models and to predict the event popularity of the next point. In order to improve the defects of the hybrid probability prediction model in the peak capture, this paper introduces the idea of cross-platform. As for the events which have similar trends of development on the different platforms, the platform on which the events have a faster development shows a better reference value for the event popularity forecast than the opposite.

After proposing the two models,16 data set acquired from 8 real event are chosen, based on the Weibo and Baidu platforms, and seven experiments are correspondingly carried out. The first four experiments are on the event popularity model. And through

the experiment, we can verify the effectiveness of the hot word selection algorithm and high sensitivity of the event popularity model. The last three experiments are on the event prediction model. It shows that the cross-platform hybrid probability prediction model is accurate and the cross-platform introduction is effective for event prediction, which can help the hybrid probability prediction model to improve the accuracy of its prediction.

Key words: social network; search engine; events popularity; hot word selection; development popularity prediction

目 录

第 1 章 绪论

1.1 研究背景和意义

近些年来互联网越来越普遍，受众众多，随之而来的是新媒体发展迅猛。传统的信息行业比如新闻，报纸等受到了越来越大的冲击。传统媒体的衰落，新媒体渐渐登上历史舞台。日常生活中，大众已经逐渐习惯于从网络中诸如微博，朋友圈、网易新闻、今日头条等平台获取社会信息。

新媒体是网络时代的产物，迎合了人们时间日趋碎片化的需求，让追求个性化的群众有一个自己做生产者的渠道。大众可以自己发布消息，也可以转发消息，这种自己做生产者也是消费者的方式，尤其是在如今明星效应和大 V 的推动下，都是加快了信息传播的速度，致使网络上每时每刻都有海量的信息在产生和传播。网络上时时刻刻传播着关于现实生活中发生的事件的丰富和详细的信息，如何高效地、有效地利用这些信息得到我们需要的数据已经成为了人们关注的问题。而如何将这些数据清晰明了显示出来，从而更好的帮助决策者分析处理问题也是众多科学家在考虑的问题。

人们都说，凡事都具有两面性。随着网络的高速发展，事件传播的速度越来越快，网络上众多纷纭的言论充斥，关于事件的真实性也越来越不是一干群众关注的重点，那么网络上一件被误传的事可能就会引起现实生活中极大动荡。古人有言“防民之口甚于防川”，由此可见在信息极其匮乏的古代都是极为重视社会舆情的引导，那么在大数据时代的今天，我们必须有一定的方法能够知道并且预知群众的舆论导向，群众的舆论导向直接关乎到事件的发展，如果能后在一定程度上预测舆论风向就可以提前做出准备，有备无患。

在此基础上，本文选取微博和百度两个社交平台上发生的事件的产生，传播，衰落、结束这一完整过程探究使用事件热度模型对事件的发展进行分析，并在事件的发展过程中进行一定的预测。本文的贡献在于提出了一种新的描述网络事件的事件热度模型，并且提出了一种新的跨平台混合概率模型用以预测事件的发展趋势。

1.2 国内外研究现状

关键词发现。S.Siddiqi 等人[1]具体简述了目前研究关于关键词查找的方法。主要分为：基于规则的语言方法、统计方法、机器学习方法和领域特定方法。Xie.F 等人[2]具体提出用通配符提取序列模式进行关键词提取。Marujo 等人[3] Liu.J 等人 [4]都从语料库着手，Marujo 等人[3]是无监督学习过程，Liu.J 等人 [4]有监督学习的过程。从语料库本身出发，所能应用的场合是有限制的。Yang.K[5]基于主题的 Text Rank 图形排序算法加强语义性。然而，TextRank 的使用效果并不会优于 TF-IDF,并且复杂度较高，速率低下。Liu.P 等人 [6]针对语境广告有历史搜索记录的特点进行关键字提取。Gollapalli 等人[7]使用引文网络从科研论文中提取关键字。这都是针对特定的领域提出的关键词检索方式，并不适用于社交网络和搜索引擎。

事件发现：Zhang.X [8] T.Sakaki 等人[9] Nguyen 等人 Zhou.D 等人[10]着手于事件的演变过程进行分析。Zhou.D 等人[11]提出使用基于词典的方式过滤无关词、利用贝叶斯模型进行事件提取和分类。M.Adedoyin-Olowe 等人[12]通过学习的方式对不同领域应用不同的事件窗口来检测事件，然而，这个方法偏向于持续时间短的事件。Zhang.C 等人[13] Guo.J 等人[14]从地理位置方面着手对地点鲜明的事件进行探测，适用范围太过狭窄。

事件预测：社交网络的事件发展主要由内容和用户构成。所以，大家的一个方向是内容：Zhang.X 等人[8] M.Mathioudakis 等人[15]利用高频词进行分析，但是词频的方法对于拥有海量数据和信息稀疏性的社交网络，这种传统的词频方法并不适合；M.Okazaki 等人[16] P.Saleiro 等人[17]侧重于语义分析来预测事件流行趋势；Lymperopoulos 等人[18]分别考虑了事件内容的线性和非线性的动态过程分别提出了对应的算法；另一个方向是用户及所处的网络环境：Lin 等人[19] Zhao.J 等人[20]利用网络拓扑分析传播过程，或是设计概率模型或是短期预测，敏感度低，无法做到实时预测；S.Mishra 等人[21] Wang.S 等人[22]考虑了社交媒体的级联特性，提出一种基于分类和时间窗口的预测方法；有一个有趣的新方向就是很多的论文 Zhang.X 等人[23] Yuan.K 等人**错误!未找到引用源。** S.Ardon 等人[24] 都重视拥有大量粉丝用户的权重，他们认为关注度大的用户是推动事件流行的主要推手。

跨平台应用：目前的跨平台方面的研究侧重于主题检测，跨域信息利用等。Bao 等人[26]利用跨平台对多媒体信息流进行检测，提取三个平台：Twitter，纽约时报和 Flickr 的优势进行互补性操作达到新兴主题的检测。Roy 等人[27]利用 twitter 中的信息进行主题学习来模拟视频的流行程度。最近的，Tang .Y 等人[28]探索了社交媒体和搜索引擎事件发展之间的关系和研究。但是，这些都没有针对性的研究跨平台事件热度。

但是，迄今为止，对于数据进行关键词发现，并进行统一的跨平台的事件热度分析和预测事件热度这两个方面的研究是很少的。

1.3 本文工作

微博和百度是国内极具备代表性的两个大数据平台。本文所采用的的数据就来自于百度和微博。

本文主要解决下面几个问题：

1) 如何从海量的数据中选出最能代表事件的热词？
2) 如何用一个统一的方法描述事件在微博和百度上的流行度？
3) 如何利用获得的事件热词来预测事件的发展趋势？

对于以上的问题，本文提出了两个模型：事件热度模型，事件预测模型。

1) 事件热度模型：

   事件热度模型是在 TF-IDF 算法的基础上改进的，提出了动态热词权重算法来提高跟主题相关度更高的词语的权重，再应用本文热词累积函数自适应地选出适合的热词数目，接着，利用选出的热词集，综合这些热词在时间单元内的权重和出现的次数来计算时间单元内的事件热度，最后，根据每个时间单元构成事件热度向量。

2) 事件预测模型：

   第一个事件热度模型的输出事件热度向量作为输入，在T_i时间，我们利用混合概率模型来预测T_{i+1}时间的事件热度。此外，为了弥补混合概率预测模型峰值捕捉的缺陷，考虑事件在微博和百度上发展趋势的相似性，我们可以利用发展速度快的平台帮助发展速度慢的平台来辅助预测。

本文采用了 8 个数据集，每个数据集分为百度和微博的数据来对本文提出的

方法进行交叉验证两个模型的有效性。希冀对后来的人的研究有一定的帮助。总体而言，本文的贡献主要由下面几点：

1) 本文提出了一种新选词方法，能够选出具有代表性的新词。
2) 本文提出了一种事件热度模型，该模型可以具有较高的描述事件的发展热度，具有很好的灵敏度可以从中分析出相关子事件。此外，这个模型并没有用到网络的拓扑结构和用户的社会关系。
3) 本文提出了一个新的预测方向：跨平台思想的引入。对于在不同平台有着大致相似发展趋势的事件，发展较快的平台对于发展较慢的平台的事件热度预测有着很好的借鉴意义。

1.4 组织结构

第一章绪论。论述了本文研究事件热度趋势及可视化分析的研究背景和意义，并查阅了大量相关资料分析现有的研究现状，最后，介绍了本文的主要工作和组织结构。

第二章背景知识。介绍了本文用的相关的一些基础知识：中文分词和 TF-IDF 算法。

第三章方法介绍。对事件热度模型和事件预测模型进行了具体的详述。

第四章实验及可视化分析。针对提出的两个模型设计了 7 个实验进行验证与分析。

第五章总结与展望。对本文工作进行总结性陈述，分析优缺点，并对未来可改进的地方进行说明。

第 2 章 背景知识

2.1 中文分词

中文分词是指将一段中文分成一个个词语的过程**错误!未找到引用源。**。在英文中，很明显可以知道空格是其分隔符，然而对于中文而言，字、句和段落都是简单易分的，唯独是词的分割，比之英文，要复杂很多。具体而言，中文分词的具体难点在于

1) 歧义判断。歧义导致语句分割方式的多样性，即可能对一句话有许多种分割方式
2) 新词发现。类似人名、地名、专业名等未登录词让计算机无法判断是否切分

目前分词方面的算法有：字符匹配、理解法和统计法。

1) 字符匹配。用一个现有的充分大现有的语料库与所要分词的词条进行匹配，若是找到字符就比对成功；否则，失败。
2) 理解法。模拟人类分词的过程进行智能化处理，简单来说，在分词的过程中应用语法语义的分析。
3) 统计法。在基于成对出现的字的频率越高越有可能组成一个词语的前提下，统计法对语料库中邻近字的组合频率进行统计，按一定算法计算精密度，当超过一定的阈值的时候就判断这是一个词语。

遗憾的是，当前的方法仍旧没有解决上述提到的两个难点：歧义判断和新词的发现。

常用的分词工具有 ICTCLAS、Paoding、HTTPCWS、SCWS 等，本文采用的是 Ansj 分词工具，它是 ICTCLAS 的 Java 实现版本。

2.2 TF-IDF 统计方法

TF-IDF[29]是指一种统计方法用于评估一个词对于一个语料库中的一个文件的重要程度。一般而言，一个词的重要程度与它在文件中出现的频率成正比，与它在语料库中出现的频率成反比。

TF（term frequency, TF）是指词频，指一个词语在一个文本中出现的频次。一般来说，计算结果需要归一化以防止其偏向较长的文件。对于在某一个特定文件里面的词语t_i来说，它的词频可以用如下公式表示：

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{2-1}$$

其中，$n_{i,j}$表示词t_i在文件d_j的出现次数，$\sum_k n_{k,j}$表示文件d_j中所有单词出现的总次数。

逆向文件频率（inverse document frequency, IDF）是对一个词普遍性的衡量。即用语料库文档的数目与包含该词的文档数目的商。其计算公式如下：

$$IDF_{\mathrm{i}} = log\left(\frac{|D|}{|j:t_i \in d_j|+1}\right) \tag{3-2}$$

其中，$|D|$表示语料库 D 中的文件的总量，$|j: t_i \in d_j|$表示包含词语t_i的文件数目，分母加 1 的目的是防止有一个词语从来没出现在语料库中，分母变为 0 无意义。

最后， $\mathrm{TF\text{-}IDF} = \mathrm{TF} * \mathrm{IDF}$。该算法倾向于过滤掉常见的词，保留文件重要的词。

第 3 章 事件热度分析和预测模型

3.1 方法综述

下图 3.1 描述了两个模型的具体流程。输入是动态的从微博和百度获得的数据，有关事件的原始数据存储在数据库中，且将相对应的领域知识也存储在数据库中，作为后续处理的知识库。原始数据中的每个记录都非常简单，只包含了时间戳，用户注释和搜索记录。

事件热度模型应用知识库对原始数据进行文本分割（使用 ansj），进行分词和简单的预处理得到候选热词集 CHW。其次，对 CHW 中的热词计算热词动态权重，再根据一个累计函数自适应地选择热词数目得到最终的热词集合 HWS。最后，依据 HWS 中的热词计算事件热度，再将这些时间段整合归一化就可以得到某段时间的事件热度向量。

第二个模型事件预测模型的输入是事件热度向量。对于任意一个时间T_i，利用混合概率模型来预测T_{i+1}的事件热度，为了弥补混合概率预测模型峰值捕捉的缺陷，引入了跨平台的思想，对于在不同平台有着大致相似发展趋势的事件，发展较慢的平台可以借鉴发展较快的平台的事件热度发展趋势。

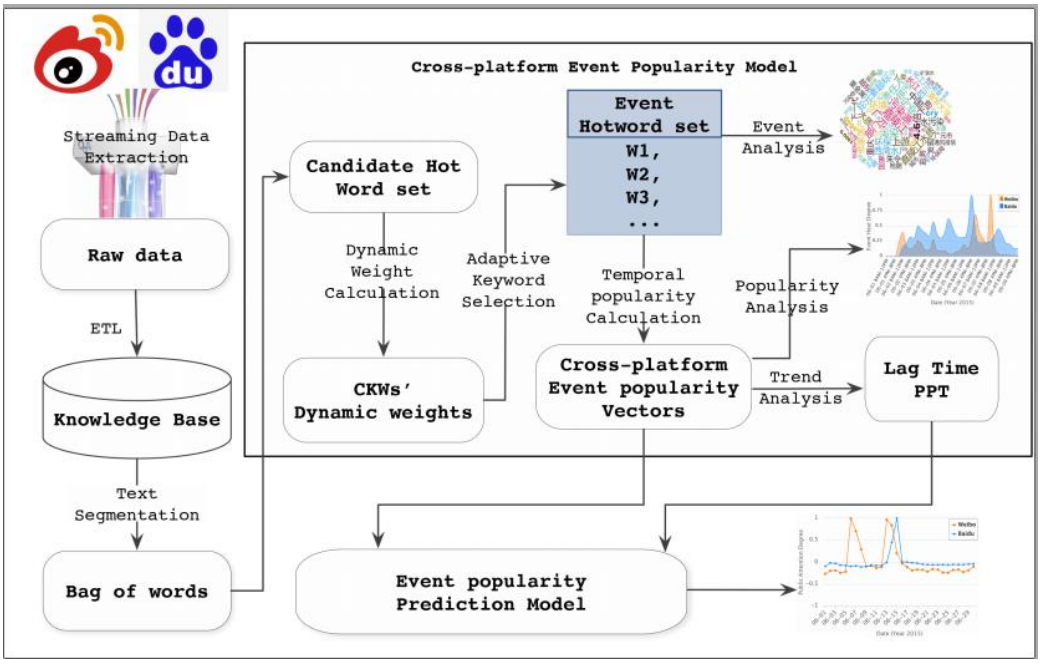

图 3.1 事件热度分析和预测模型流程图

3.2 数学符号的定义

表 3.1 数学符号一览表

符号	描述
E	事件
T_n	由 n 个单位时间构成的一段时间
t	时间单元 t∈[T_{i-1},T_i]
CHW	在时间T_n内的热词候选集合
PMI(w_i,w_j)	单词w_i和w_j的成对互信息
CPMI(w_i)	单词w_i与其他单词总互信息的系数
DW$_T$(w_i)	一个单词w_i和一个时间 T 的动态权重
HWS	热词集合
hw_i	热词集合里面的第 i 个热词
$RTF_{t_i}(hw_i)$	hw_i的相对词频
EP_t	在单位时间 t 内事件 E 的流行度
$EP^P_{T_n}$	在时间T_n内平台 p 上的事件热度向量
$NEP^P_{T_n}$	归一化的在时间T_n内平台 p 上的事件热度向量

3.3 事件热度模型

3.3.1 模型概述

每个事件E_i都是由一组信息来表示的，这组信息通过对事件的数据库筛选过滤得到与关键词相关的信息获得。在微博和百度平台上，这组信息可以是任意长度的文本、单词等。每条信息由平台 P，时间 T，内容 body 组成。这个模型首先处理从数据库中选取热词的问题，本文指出热词的三个特征：TF-IDF 特征，单词间的相关度和长度。因此，我们定义热词的动态权重如下：

定义 3.1 热词的动态权重。

给定一个单词w_i和一个时间 T，其动态权重（DWT(w_i)）是词的重要性 IMPT(w_i)、词相关度 CDT(w_i)和长度的乘积，由公式(3-1)计算。

$$\mathrm{DW}_T(\mathrm{w}_i) = IMP_T(\mathrm{w}_i) * CD_T(\mathrm{w}_i) * \mathrm{length}(\mathrm{w}_i) \quad (3\text{-}1)$$

所以，在时间 T 的热词选取问题定义如下。对数据库进行预处理得到 N 个常用候选热词集 CHW，对 CHW 中单词的动态权重进行降序排列，我们可以得到一个函数：$\mathrm{F}(X) = Y$,其中，Y 表示 CHW 中前 X 个词的累计动态权重，通过图形变换和极值来求得最适合的热词数目对 CHW 进行一次筛选得到其子集 HWS。

在获得热词集 HWS 后，本文通过以下方式计算事件热度向量。

定义 3.2 事件热度向量。

给定时间段T_n，热词集 HWS，事件热度向量表示为$EP_{T_n}^P$由公式（3-3）计算，时间间隔t_i的事件热度由公式(3-2)计算。

$$EP_{T_n}^P = \left[EP_{t_1}, EP_{t_2}, EP_{t_3} \ldots\ldots EP_{t_n}\right] \quad (3\text{-}2)$$

$$EP_{\mathrm{t}_i} = \sum_{hw_i \subset HWS} w_{\mathrm{t}_i}(hw_i) * Fre_{t_i}(hw_i) \quad (3\text{-}3)$$

其中，$w_{t_i}(hw_i)$和$Fre_{t_i}(hw_i)$表示热词hw_i的权重和在时间间隔t_i内出现的次数。

下面分为两个部分得到事件热度向量：热词选取和事件热度计算

3.3.2 热词选取

热词的选取无论是对计算动态权重还是事件热度向量的都是非常重要的。本文认为的热词具有的特征如下：

1) 重要性。热词应该足够重要的词。这里借用 TF-IDF 的概念，热词应该是出现的频率足够高又不是那些类似于“今天”、“那些”、“是”等的常用无关词。
2) 相关度。在事件数据中，经常会出现描述事件的关联性很强的热词，比如，“悲剧”“救援”经常是成对性出现的词。在本文中，我们使用成对词相互信息（PMI）量化相关度。
3) 完整性。完整性高热词一般代表的含义更多，在上下文中占据更重要额地位，本文认为词的长度是热词的最后一个指标。

3.3.2.1 候选热词集

将存储在数据库中的原始数据按时间排序后，输入时间 T 内的所有文档，先将这些文档通过 Ansj 处理得到分词结果集。建立一个停止词库，所谓的停止词指的是中文里面的虚词，介词、冠词、连词、副词（的、得、地）等一些无意义但是在文档中高频出现的词，这些词影响分词的效率和准确性。根据停止词库，将分词结果集中的这些停止词删除。

此外，我们还需要一个知识词典，将这个事件的一些专业词语放入，以防分词切分错误，提高准确率。

通过预处理，我们得到一个单词包，这个集合很庞大，数据稀疏。我们需要对它进行一个简单的操作来获得热词候选集（CHW）。

本文使用公式(3-4)选出 CHW，里面有 N 个单词。假设单词t_i,包含该词的记录的数目是n_1，总记录数时 M，那么该词如果满足：

$$n_1 \geq \sqrt{M} \tag{3-4}$$

那么，该词就添加到集合 CHW 中，否则，不加。

3.3.2.2 热词动态权重

为了计算动态权重，我们需要计算热词的重要性 IMPT(w_i)、词相关度$CD_T(\mathrm{w}_i)$。

关于计算词的重要性 IMPT(w_i)是基于 TF-IDF 算法的，给定单词w_i和时间 T，其由公式(3-5)计算。

$$IMP_T(\mathrm{w}_i) = TF_{w_i} \times IDF_{w_i} \tag{3-5}$$

假设候选热词w_i在 CHW 中出现的次数是n_i，那么它的词频由公式(3-6) 计算。

$$TF_{w_i} = \frac{n_i}{\sum_{w_k} n_{w_k}} \tag{3-6}$$

其中，$\sum_{W_k} n_{W_k}$表示所有候选热词出现的次数之和

候选热词w_i的逆向文件频率由公式(3-7)计算。

$$IDF_{w_i} = log\left(\frac{R_t}{|j:w_i \in R_t|}\right) \tag{3-7}$$

其中，定义R_t表示时间 T 内的所有记录数目，$|j: W_i \in R_t|$表示R_t中包含候选热词w_i的记录数目。

关于计算词相关度$CD_T(w_i)$，需要先引入两个词之间的互信息 PMI 和一个词与其他所有单词的相关度 CPMI。

假设有单词w_i，w_j，包括单词w_i的记录有N_1个，包括单词w_j的记录有N_2个。包括单词w_i和单词w_j的记录数N_3。所以单词w_i出现的概率由公式(3-8)计算。

$$\text{Prob}(w_i) = \frac{N_1+1}{N} \tag{3-8}$$

单词w_j出现的概率由公式(3-9)计算。

$$\text{Prob}(w_j) = \frac{N_2+1}{N} \tag{3-9}$$

其中，N 是 CHW 中所有的记录数。加 1 的目的是为了防止出现 0 的情况。

所以 PMI 由公式(3-10)计算。

$$\text{PMI}(w_i, w_j) = \log\left(\frac{(\text{Prob}(w_i), \text{Prob}(w_j))}{\text{Prob}(w_i) * \text{Prob}(w_j)}\right) \tag{3-10}$$

使用公式(3-10)和热词候选集 CHW，我们可以得到一个表示为 PMI 的矩阵，其中$\text{PMI}[i,j] = \text{PMI}(w_i, w_j)$。给定一个单词$w_i$，其 PMI 的和表示为$\text{SPMI}(w_i)$,由公式(3-11)计算：

$$\text{SPMI}(w_i) = \sum_{i \neq j} \text{PMI}[i,j] \tag{3-11}$$

所以，词的相关度$CD_T(w_i)$由公式(3-12)计算如下。

$$CD_T(w_i) = \text{SPMI}(w_i) \Big/ MAX(\text{SPMI}(w_i)) \tag{3-12}$$

其中$MAX(\text{SPMI}(w_i))$表示最大的一个$\text{SPMI}(w_i)$。

以上，在计算出 IMPT(w_i)和 CDT(w_i)后，我们就可以根据公式(3-1)计算出热词动态权重。

3.3.2.3 热词累计权重函数

在计算出所有热词的动态权重后。如 3.3.1 节所示，时间 T 内的热词集合 HWS 是热词候选集 CHW 的子集，应该选出动态权重最大的热词添加到 HWS 中。对 CHW 中的热词按照其动态权重进行降序排序，我们可以得到热词的累计

权重函数由公式(3-13)计算。

$$\mathrm{Cum}(x) = \sum_{i\in[1,x]} \mathrm{DW}_T(\mathrm{w}_i) \tag{3-13}$$

$\mathrm{Cum}(x)$表示前 x 个最高的热词动态权重的和。如图 3.2(a)所示，他是一个单调递增函数。我们将曲线进行一系列操作得到如图 3.2(b)所示。

a) 将曲线平移到原点

b) 曲线起点 A，终点 B 的连线与 x 轴的夹角为θ，逆时针旋转曲线θ度

c) 将变换后的曲线上的每个点的横坐标变为和原坐标一致，纵坐标仍旧使用变换后的

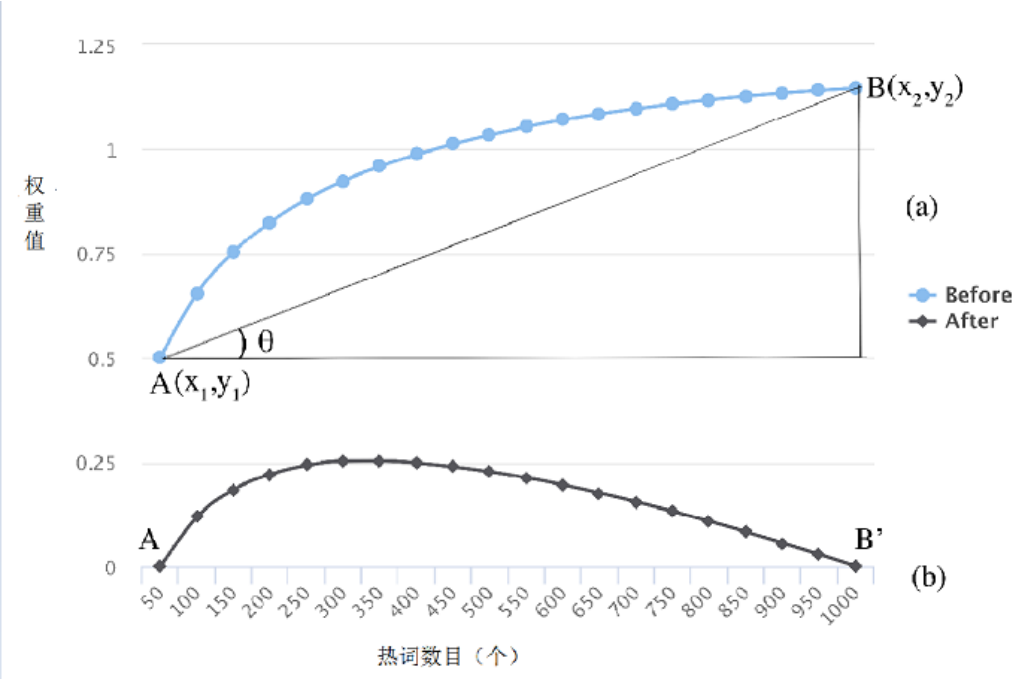

图 3.2 累计权重函数变换

函数变换过程如下：

假设起点 $\mathrm{A}(\mathrm{x}_1, y_1)$，终点 $\mathrm{B}(\mathrm{x}_2, y_2)$, $\tan\theta = \frac{y_2 - y_1}{\mathrm{x}_2 - \mathrm{x}_1}$。

先平移到原点，平移后坐标(x^1, y^1)

$$x^1 = x - x_1$$

$$y^1 = y - y_1$$

顺时针旋转θ度，设曲线上任意一点(x^2, y^2)

$$x^2 = x^1 \cos\theta + y^1 \sin\theta$$

$$y^2 = -x^1 \sin\theta + y^1 \cos\theta$$

再将 x 坐标与原来一致，最后

$$x^2 = \mathrm{x}$$

$$y^2 = -(x - x_1)\sin\theta + (y - y_1)\cos\theta$$

所以，变换后的函数$\mathrm{Tcum}(x)$由公式(3-14)计算。

$$\mathrm{Tcum}(x) = -(x - x_1)\sin\theta + (\mathrm{Cum}(x) - y_1)\cos\theta \tag{3-14}$$

3.3.2.4 热词集合

给定时间 T 和 CHW 热词候选集，里面有 N 个单词，热词集 HWS 是 CHW 中动态权重最高的 k 个词，其中 k 满足以下条件：

$$\forall \mathrm{p} \in [0, \mathrm{N}], \exists \mathrm{k}, |k| = 1 \wedge \mathrm{Tcum}(x) \geq \mathrm{Tcum}(p)$$

即是存在唯一的一个极值 k。

证明如下：

由图 3.2(a)可知，$\mathrm{Cum}(x)$函数连续可导，由拉格朗日中值定理得，在区间$[x1, x2]$，至少有一点ϵ，满足公式(3-15)

$$\frac{\mathrm{f}(x_2) - f(x_1)}{x_2 - x_1} = \dot{\mathrm{Cum}}(\mathrm{\mathcal{E}}) \tag{3-15}$$

其几何意义就是至少有一个点的切线与直线 AB 平行

又因为，f(x)是一个递增函数，所以：

在图 3.1 中，有且只有一条这样的切线与 AB 平行，

又因为，平移和旋转不改变图 3.1 的特性，最后一步的 x 坐标的改变不影响 y 坐标的变化趋势，所以，图 3.2 中只有一条切线与 x 轴（即是 AB 连线）平行。由拉格朗日定理可知此时的拉格朗日公式变为(3-16)所示。

$$0 = \dot{\mathrm{Tcum}}(\mathrm{\mathcal{E}}) \tag{3-16}$$

所以这个点就是唯一一个切线与 x 轴（即是 AB 连线）平行的点，又因为这个点的导数是 0，所以他是极值点。

又因为，变换后的曲线连续可导，所以极值点都是导数为 0 的点，几何意义就是切线与 x 轴平行。

所以，图 3.1(b)中有唯一的一个极值点.

完毕证明。

我们可以通过以下方式找到 k 值。

$$\mathrm{Tcum}(x) = -(x - x_1)\sin\theta + (\mathrm{Cum}(x) - y_1)\cos\theta$$

$$\dot{\mathrm{Tcum}}(x) = -\sin\theta + \dot{\mathrm{Cum}}(x)\cos\theta$$

$$\dot{\mathrm{Cum}}(x) = \tan\theta$$

所以，这个 k 值满足$\dot{\mathrm{Cum}}(k) = \tan\theta$。从图 3.1(b)中，我们可以看到这个 k 值是新增热词带来的优势开始下降的点，本文通过这种自适应算法选择热词的数目，这 k 个热词的集合就是 HWS。

3.3.3 事件热度

在获得时间 T 内的 HWS 后，根据公式（3-2）我们需要计算热词hw_i的权重$w_{t_i}(hw_i)$是相对词频和 IDF 逆向文件频率及单词长度的乘积。

3.3.3.1 相对词频

给定 HWS 中的热词hw_i和时间间隔t_i,相对词频表示为$RTF_{t_i}(hw_i)$，其由公式(3-17)计算。

$$RTF_{t_i}(hw_i) = \frac{\mathrm{n}{hw_i}^{t_i}}{\sum_{\mathrm{k}} \mathrm{n}{hw_k}^{t_i}} \tag{3-17}$$

其中，$\mathrm{n}{hw_i}^{t_i}$表示热词hw_i在时间间隔t_i出现的次数，$\sum_{\mathrm{k}} \mathrm{n}{hw_k}^{t_i}$表示所有热词出现在时间间隔$t_i$内的总次数，$hw_k \subset HWS$

3.3.3.2 热词权重

给定 HWS 中的热词hw_i和时间间隔t_i,热词权重$w_{\mathrm{t}_i}(hw_i)$由公式(3-18)计算。

$$w_{t_i}(hw_i) = RTF_{t_i}(hw_i) \times IDF_{t_i}(hw_i) \times log\big(len(hw_i)\big) \tag{3-18}$$

其中，

$$IDF_{t_i}(hw_i) = log\left(\frac{R_{t_i}}{|j: hw_i \in R_{t_i}|}\right) \tag{3-19}$$

定义R_{t_i}表示时间间隔t_i内的所有记录数目，$|j: hw_i \in R_{t_i}|$表示R_{t_i}中包含热词hw_i的记录数目。

$len(hw_i)$表示热词的长度。

3.3.3.3 事件热度向量

根据公式(3-2)(3-3) 可以计算事件热度向量。$EP_{T_n}^P$是一个向量。将其由公式(3-21)进行归一化处理。

$$NEP_{T_n}^P = \frac{[EP_{t_1},EP_{t_2},EP_{t_3}.....EP_{t_n}]}{max(EP_{t_i}^P)} \tag{3-21}$$

其中，$max(EP_{t_i}^P)$表示向量中元素值最大的一个。

3.4 事件预测模型

3.4.1 模型概述

3.4.1.1 混合概率预测模型

一个事件的发展趋势有些阶段呈现线性，而有些是非线性的，这种不确定性往往就是事件预测的难点，让大家对于事件的发展趋势没有一个比较具体的判断。在对已经发生过的事件画出事件热度曲线，我们利用现有的函数分布模型对热度曲线进行拟合发现，变化的曲线也是存在一定的函数关系，而且在某些时间段有着很高的拟合程度，由此，从这些事件的分析中我们想出了混合概率模型。

但是我们也很显然的可以分析出混合概率模型的缺点：真实事件发展的不确定性，事件热度变化幅度无法确定，给我们的预测带来很大的困难；其次，当现实世界的事件已经到达峰值点，但是根据概率模型我们并不能捕捉这个峰值点，对于峰值之后的预测会存在一定的误差。

混合概率模型（HPM）的输入是事件热度向量。对于任意一个时间T_i，混合概率模型用来预测T_{i+1}的事件热度。

我们使用指数模型，指数二项式模型，正态分布模型和高斯二项式模型来对已知的点进行曲线拟合，在依据动态选择标准来选择最适合的模型进行预测。

3.4.1.2 跨平台混合概率预测模型

网络的发达让一个事件在微博和百度两个平台的发展趋势具有相关性。通过之前的实验我们也可以看出在某些情况下两个平台时间事件的发展趋势是类似

的。为了克服混合概率模型存在的峰值捕捉和趋势发展不确定的情况，这里引入跨平台事件混合概率模型。

对同一个事件的发展，两个平台之间有些事件是发展趋势相似但存在一定的时间差（lagt）。所以，在这种情况下，发展较慢的平台就可以借鉴发展较快的平台的发展趋势，这对于峰值捕捉有着很强的借鉴意义。我们在利用混合概率模型时，对选择的最优的概率模型应用跨平台进行微调从而求得比较符合实际的预测热度。即在应用选择标准时兼之参考另一个平台发展趋势选择出更适合的概率模型。

3.4.2 模型详述

3.4.2.1 混合概率模型预测算法

混合概率预测模型使用了指数模型、指数二项式模型、正态分布模型和高斯二项式模型。各个模型的表达式如下所示。

指数模型：

$$\mathrm{f}(x) = a * e^{bx} \tag{3-22}$$

指数二项式模型：

$$\mathrm{f}(x) = a * e^{bx} + c * e^{dx} \tag{3-23}$$

正态分布模型：

$$\mathrm{f}(x) = a_1 * e^{-\left(\frac{x-b_1}{c_1}\right)^2} \tag{3-24}$$

高斯二项式模型：

$$\mathrm{f}(x) = a_1 * e^{-\left(\frac{x-b_1}{c_1}\right)^2} + a_2 * e^{-\left(\frac{x-b_2}{c_2}\right)^2} \tag{3-25}$$

预测流程：

因为高斯二项式模型需要 6 个点才能够实现曲线的拟合，我们统一选择 6 个点作为拟合的已知点。

假设已知点$T_1, T_2, T_3, T_4, T_5, T_6$对应的热度分别是$NEP_1, NEP_2, NEP_3, NEP_4, NEP_5, NEP_6$。$\widehat{NEP}$为预测热度。$\overline{NEP}$为热度均值。

1. 对现有的EP_i，利用 MATLAB 进行曲线拟合，分别使用指数模型、指数二项式模型、正态分布模型和高斯二项式模型进行拟合得到拟合曲线的预测值；

2. 计算 SSR。SSR 指的是预测数据与真实数据均值的差值的平方和。由公式(3-26)计算。

$$\text{SSR} = \sum_{i=1}^{M}\left(\widehat{NEP_i} - \overline{NEP}\right)^2 \tag{3-26}$$

3. 计算 SST。SST 指的是真实数据和真实数据均值差值的平方和。由公式(3-27)计算。

$$\text{SST} = \sum_{i=1}^{M}(NEP_i - \overline{NEP})^2 \tag{3-27}$$

4. 根据 SSR 和 SST 计算 R-square 确定系数。确定系数越大表示拟合程度越高。由公式(3-28)计算。

$$\text{R-square} = \frac{SSR}{SST} \tag{3-28}$$

5. 计算 4 个模型的 R-square，选择值最大的模型定为最优模型。
6. 输入时刻T_7代入到模型方程中，计算对应的NEP_7，这就是混合概率预测模型预测的数据NEP_7。

3.4.2.2 跨平台混合概率预测模型

这里提出的跨平台算法是为了辅助混合概率预测模型，但是只有当一个事件在两个平台的发展有着相似性的时候才可以应用，这个模型有着他的局限性。跨平台的引用首先就需要判定一个事件在微博和百度这两个平台是否有着发展相似性，此外，亟待解决的还有在两个平台如果已经确定具备发展相似性的情况下，如何计算两个平台的时间差。

本文对于两个平台具备相似性的判断和 lagt 的计算方法是有人工参与的，还没有很好的算法来计算。

lagt 是根据第一个模型的输出事件热度，按时间单位 t 向前平移发展较慢的平台的事件热度图，应用余弦相似度找到两个平台相似度最大的时候的时间差就是所求的 lagt。余弦相似度由公式(3-29)计算。

$$\text{Sim}\left(NEP_T^P, NEP_T^Q\right) = \frac{\sum_{i=1}^{n}\left(NEP_{t_i}^P, NEP_{t_i}^Q\right)}{\sqrt{\left(\sum_{i=1}^{n}\left(NEP_{t_i}^P\right)\right)^2} \times \sqrt{\left(\sum_{i=1}^{n}\left(NEP_{t_i}^Q\right)\right)^2}} \tag{3-29}$$

这个方法加入了人工的判断过程，并不是非常的好，在以后的研究方向，应

该考虑将求出 lagt 的过程更加的优化，依靠算法与计算机来完成而非人工。在基于 lagt 人工求解出来后，本文提出下面的操作方法来辅助混合概率预测模型。

假设有一个事件 E 在平台 A，B 上的发展是有关联的。A 平台响应快先发生，B 平台后发生。给定事件的时间跨度T_n。

跨平台辅助混合概率模型应用在下面两个方面：

1. 跨平台分析-有概率模型选择

   A 平台在时刻 t-lag-time 的事件热度是EP^A_{t-lagt}，在 t+1-lagt 时刻，热度是$EP^A_{t+1-lagt}$。平台 B 在时刻 t 的事件热度是EP^B_t，待预测的 t+1 时刻的热度是EP^B_{t+1}。当$EP^A_{t-lagt} > EP^A_{t+1-lagt}$时，跨平台分析选取模型时选择满足$EP^B_t > EP^B_{t+1}$的模型；反之，则相反。

2. 跨平台分析-无概率模型选择

   A 平台在时刻 t-1-lagt，t-lagt，t+1-lagt 的事件热度分别是$EP^A_{t-1-lagt}$，EP^A_{t-lagt}，$EP^A_{t+1-lagt}$。平台 B 已知 t-1，t 时刻的事件热度分别是EP^B_{t-1}，EP^B_{t+1}，待预测的 t+1 时刻的热度是EP^B_{t+1}。当$EP^A_{t-lagt} \geq (\leq) EP^A_{t+1-lagt}$时，所有的模型都没有满足$EP^B_t \geq (\leq) EP^B_{t+1}$的模型。那么，就计平台 A 的 t-1-lagt，t-lagt，t+1-lagt 的热度平均变化率，V=$\frac{EP^A_{t+1-lagt}-EP^A_{t-lagt}}{EP^A_{t-lagt}-EP^A_{t-1-lagt}}$。所以，$EP^B_{t+1}$=V × $(EP^B_t - EP^B_{t-1}) + EP^B_t$。

当两个平台 A、B 之间并没有关联时，我们仍旧使用混合概率预测模型。

3.5 本章小结

本章具体讲述了事件分析两个模型：事件热度模型和事件预测模型。事件热度模型。

事件热度模型利用单词之间的相关度综合 TF-IDF 方法设计了衡量热词的动态权重的算法，据此画出热词的累积算法根据图形变换和极值得到适合的热词集合 HWS，最后根据 HWS 内的热词设计计算事件热度向量的算法。

事件预测模型分为两个方案：混合概率预测模型和跨平台混合概率预测模型。混合概率预测模型是使用不同的函数模型对事件热度曲线进行拟合，选择拟合程

度最高的当做最优模型，并根据这个最优模型预测下一个点的数据。而跨平台的引用是为了弥补混合概率模型在峰值捕捉方面的不足，以达到更精确预测的目的。

第 4 章 实验及可视化分析

本章的实验是针对第三章的两个模型：事件热度模型和事件预测模型。本章一共设计了 7 个实验，其中前四个是针对事件热度模型，后三个是针对事件预测模型。本章的实验的所有数据集是第 3 章提过的 7 个数据集，每个数据集分为百度和微博两个平台。

4.1 事件热度模型实验及可视化分析

本节进行了 4 个实验。第 1 个实验是比较使用本文的选词算法和 TF-IDF 算法哪个词集更具备代表性，这里选取的事件是“南海仲裁案”和“科比退役事件”。第 2 个是对比根据本文的事件热度模型画出的热度曲线和使用 TF-IDF 画出来的事件热度曲线哪个更灵敏，这个实验选取的事件是“股市大跌”、“东方之星沉船事件”、“Alpha Go”和“Pokémon Go”。第 3 个实验是选取“东方之星沉船事件”和“Pokémon Go”这两个事件按照天为单位，依据关键词的变化和每天的事件热度曲线来分析事件的每天的发展情况是否一致。第 4 个实验是选取“南海仲裁案”和“科比退役事件”按小时画出事件热度曲线并进行事件分析。

4.1.1 实验 1：热词选取算法对比

实验 1 选取“南海仲裁案”、“科比退役”在微博和百度两个平台上的发展进行对比。对比结果如下，其中图 4.1 是南海仲裁案，图 4.2 是篮球运动员科比退役事件。这两个表下划线是横线表示是本文选词算法特有的热词；下划线是波浪线表示是 TF-IDF 算法选出来的热词。

分析表 4.1，我们发现在微博平台，排名前 5 的热词，本文的热词选取算法和 TF-IDF 算法都有的是“南海仲裁”，“菲律宾”，“一点都不能少”，而明显可以看出 TF-IDF 包含了“转发微博”和“哈哈哈哈”这种无意义的高频词,本文认为“转发微博”并不具实际意义。

再分析表 4.1，我们发现在百度平台，排名前 5 的热词并没有明显的差异，但是我们关注 26 以后的热词会发现，本文的热词选取算法包含的热词类似“南沙”，“舰队”，“演习”更能够从中窥探出事情的梗概；而 TF-IDF 的“控制”，“最终”这类词却是没有实际的意义。

表 4.1 南海仲裁案

	微博平台		百度平台	
编号	热词选取	TF-IDF	热词选取	TF-IDF
1	南海仲裁	转发微博	南海	南海仲裁
2	菲律宾	哈哈哈	中国	南海
3	中国	南海仲裁	南海仲裁	中国
4	一点都不能少	一点都不能少	菲律宾	公布
5	南海	菲律宾	仲裁	菲律宾
…	…	…	…	…
26	主权	仲裁	自古以来	法庭
27	我国	祖国	法庭	海牙
28	祖国	穿越	南沙	外交部
29	爱国	中华	舰队	控制
30	承认	拳头	演习	最终
…	…	…	…	…
46	12 日	和平	韩国	承认
47	捍卫	共青团	日本	控制
48	寸土不让	历史	局势	海战
49	人民日报	威武	明日	空军
50	政府	香蕉	政府	几点
…	…	…	…	…

对比两种方法各自有的热词，很明显，TF-IDF 算法独有的热词类似“穿越”，“共青团”，“控制”，“几点”都不是有重要意义的热词；反之，本文热词选取算法独有的热词类似“爱国”，“12 日”，“人民政府”等都可以让观者从中获取到一定的信息。

由此可以发现，选词算法比之 TF-IDF 在南海仲裁案这件事上选出的热词无论是在微博还是百度平台都具有更好的代表性。微博平台凸显的优势要比百度更加的明显。

表 4.2 科比退役事件

	微博平台		百度平台	
编号	热词选取	TF-IDF	热词选取	TF-IDF
1	Thank you	转发微博	Kobe	告别
2	科比	李易峰	退役	60 分
3	504	Thank you	视频	Kobe
4	李易峰	504	NBA	退役
5	Kobe	微博	比赛	布莱恩特
…	…	…	…	…
26	致敬	60 分	致敬	谢幕
27	感谢	投票	81 分	球衣
28	篮球	霍建华	篮球	潘玮柏
29	黑曼巴	泪流满面	2016	凌晨
30	谢谢	比赛	乔丹	乔丹
…	…	…	…	…
46	青春	传奇	背景	黑曼巴
47	永远	查看	奥尼尔	Nike
48	洛杉矶	完美	11 代	瓦妮莎
49	结束	20 年	洛杉矶	职业
50	偶像	永远	20 年	Ins
…	…	…	…	…

通过对比微博平台，两个方法选出的热词，我们发现，在微博平台，前 5 个热词，TF-IDF 仍旧出现了“转发微博”和“微博”这类无意义的热词。此外，我们发现“李易峰”也在两种方法都出现了，本人重新搜索过相关新闻，发现“李易峰”这个中国明星是科比的粉丝，并在科比退役赛那场出现告别科比，但是，“李易峰”这个名词是明星效应的结果，跟本事件并没有太大的联系，而本文的热词选取算法尽力降低该词的重要程度这方面做得比 TF-IDF 要好。对比 26-30 这几个热词，我们发现 TF-IDF 仍旧出现“霍建华”这个无关热词，而本文的热词选取算

法降低了他的重要程度。

对比百度平台，我们发现，前 5 个热词并没有太大的差异，而在后面的热词中，我们可以明显的看到“篮球”，“81 分”，“2016”这些词，科比这名优秀的篮球运动员是 2016 年退役的，拿过的最好的分数是“81 分”，显然，这些词是与科比本身息息相关的；反之，我们观察 TF-IDF 所选出来的热词，出现了“潘玮柏”，“球衣”这类重要程度明显很低的名词。

对比两者独有的热词，TF-IDF 都是类似于“投票”，“霍建华”“查看”这类无关事件的词。

从这两个表我们可以看出，热词选取算法比之 TF-IDF 算法在科比退役这件事上选出的热词无论是在微博还是百度平台都具有更好的代表性。微博平台凸显的优势要比百度更加的明显。

还有其余的 7 个事件的对比实验本文就不再这里展开，总的说来，热词选取算法和 TF-IDF 算法的对比是非常明显的，热词选取算法在选词方面比 TF-IDF 有更好的代表性。我们也有趣的发现，热词选取算法的优势在微博这个平台能够得到更好的展示。本人猜测是因为百度的特性是搜索相关，群众带着更明显的目的与搜索，或者是被新闻媒体引导去思考这件事，所以他的热词会更加的凸显出来，主题更加明确，这样也导致两者的对比不是很明显。

4.1.2 实验 2：事件热度模型算法对比

实验 2 是对比根据本文的选词算法得到的事件热度曲线和根据 TF-IDF 算法得到的热词画出的事件热度曲线。我们选取“股市大跌”、“东方之星沉船事件”、“Alpha Go”和“Pokémon Go”进行试验。结果如图 4.1,4.2,4.3,4.4。

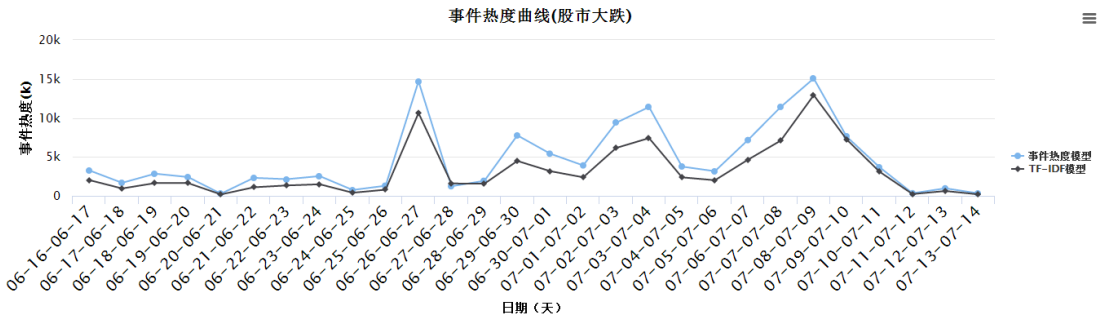

图 4.1 股市大跌事件热度曲线

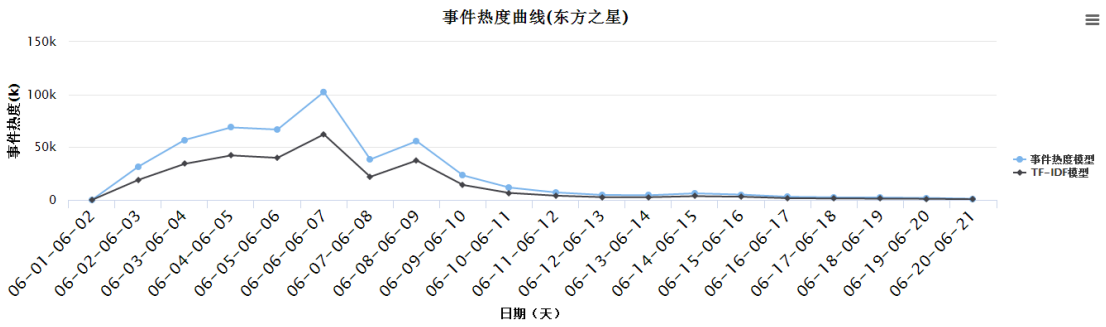

图 4.2 东方之星沉船事件热度曲线

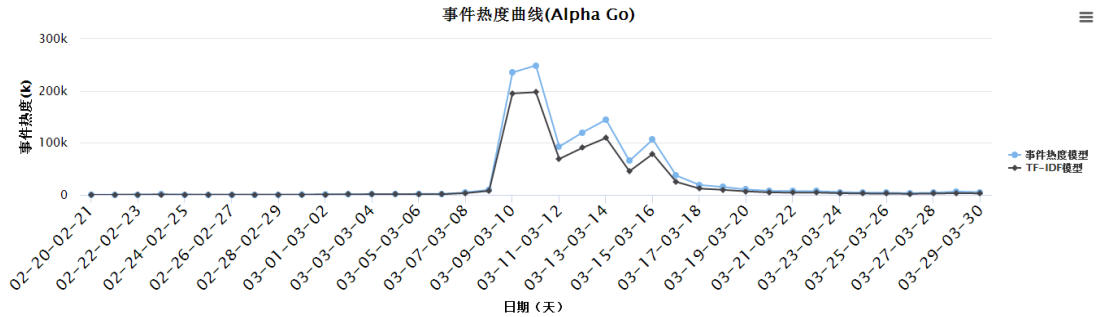

图 4.3 Alpha Go 事件热度曲线

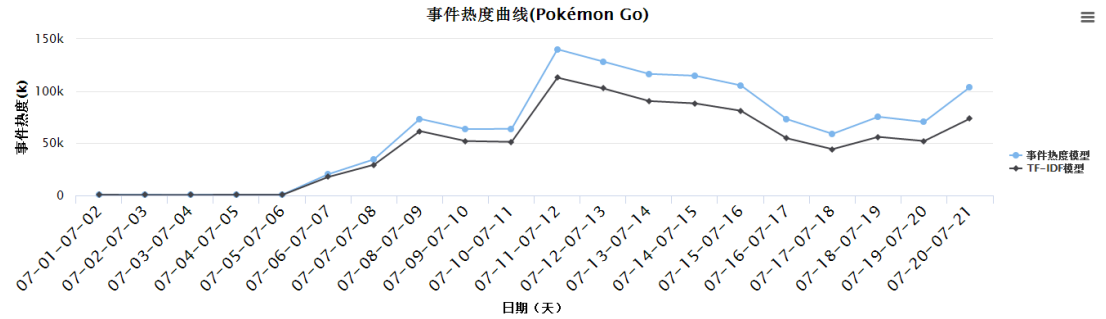

图 4.4 Pokémon Go 事件热度曲线

分析这 4 个曲线对比图，蓝色的线代表本文提出的事件热度模型画出的事件热度曲线，黑色曲线代表使用 TF-IDF 方法选词后画出的事件热度曲线。我们可以很明显的看出蓝色的线一直在黑色的线上方，其显示的热度一般是 TF-IDF 方法计算出的热度的 1.5 倍。故而，本文提出的方法具有更高的灵敏度。

4.1.3 实验 3：按天分析事件

实验 3 交叉性验证了实验 1 的热词选取的代表性，本文针对东方之星沉船事件和 Pokémon Go 按天为单位分析关键词的变化，进而分析整个事件的过程并发现子事件。

下表 4.3 是东方之星沉船事件在两个平台每天的新增热词表，图 4.5 是事件在每天的热度曲线。事件是在 2015.6.1，但是事件为人所知是 6.2。这里将 6.2 看

作事件可追踪的第一天。

表 4.3 东方之星沉船事件

时间	新增的热词	
	微博平台	百度平台
6.3	蜡烛、希望、救起、6 月、人生	幸存者、直播、搜狐、14、国内、生死、原因、打捞
6.4	打捞、事件、小时、生命	事件、一天、工作、曝光
6.5	船体、遇难、14、水面	船体、6 月、百度
6.6	搜寻、船上、船舶、指挥部	曝光、状况、人数、396、遗体
6.7	头七	原因、一天
6.8	遗体、上海	搜狐

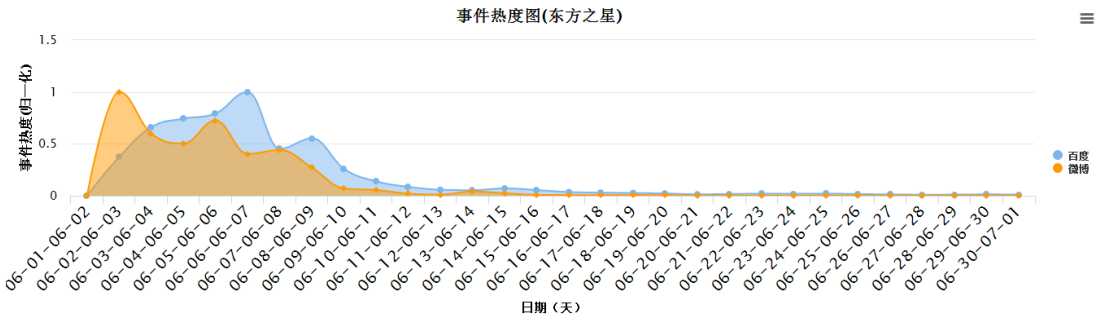

图 4.5 东方之星沉船事件热度曲线图

这里先对事件进行一次回顾：6 月 1 日，客轮“东方之星”沉船；6 月 2 日到 6 月 4 日，事件为社会广大群众所知，打捞工作在紧锣密鼓的执行；6 月 5 日船体扶正；6 月 7 日，遇难人员的“头七”悼念活动，6 月 8 日上海遇难者遗体告别仪式启动。通过对照，我们可以发现每天新增加的热词都是合理的，符合当天的情况。

通过对照，我们可以看到无论是微博还是百度的数据都是与事件真相是符合的。但是我们也发现他们之间的差异：百度有更精确的事件数据。比如 6 月 3 日这天百度新增词“14”，这个数字是当天已经被获救的人数。再看 6 月 6 日，百度新增词“396”，这个数字是当天 12 点官方统计出的遇难者人数。

再根据图 4.5 的微博和百度的事件热度曲图。我们可以看出事件热度在 6 月 8 日后事件热度急剧降低。主要原因是 8 号前是集中性的打捞救援工作和头七祭奠仪式。可以看到事件在发生后，微博平台立刻就给出了响应，并且事件热度一

直是居高不下到“头七”结束热度才开始减弱。而百度的响应过程较慢，并且是在8号前世明显的多峰事件，在船体扶正和“头七”两天有明显的峰。

下表4.4 Pokémon Go入驻中国市场的事件在微博平台和百度平台的新增热词表，图4.6是事件每天的热度曲线。

表4.4 Pokémon Go事件

时间	新增的热词	
	微博	百度
7.6	514、IOS、训练、中心、通缉、VPN、下载	Pokémon go、宝可梦、精灵、GPS、signal、
7.7	bubble、nintendokyo、国服、5公里、种草、	iOS、神奇
7.8	垃圾桶、想象、力量、隐藏、关键、黑暗	Pokémon、中国、安卓、google、iPhone
7.11	中国、东北、VPN、妖怪	定位
7.13	GPS、国内	地图、虚拟、教程、破解、账号
7.17	日本	
7.18	喜欢	国内

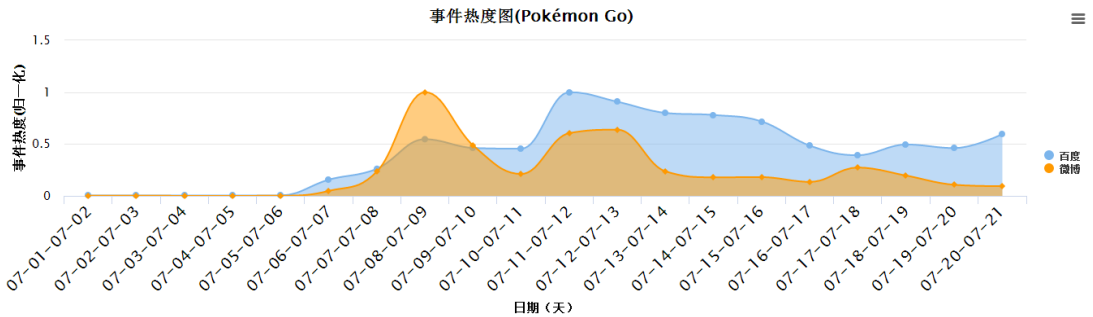

图4.6 Pokémon Go事件热度曲图

这里先对事件进行一个简单的回顾：2016年7月6日，Pokémon Go正式上架，安卓和iOS系统都可以玩，但是国内的安卓机需要VPN才能登陆；2016年7月20日进入日本市场；2016年7月25日，正式进入中国香港。但是就关于何时进入亚洲区有过不少的传闻，当时网络传言7月11日中国区开服，后来又有

网传亚洲区 7 月 15 开服，这些的虚假信息都导致了当时的一阵热潮。

结合新词对比我们可以看出，新词的增加状态是与事件的发展紧密结合的。一部分是关于游戏内容的探讨，一部分是关于何时在中国上线的讨论。但是通过对比，我们可以看到，百度平台上的热词更加的具有针对性，比如“iOS”、“android”、“google”、“iPhone”，“domestic”等热词。

再根据图 4.6，可以看出，两个平台在 7 月 6 日后两天热度直升，在谣传的 7 月 11 日热度也有所增加，还有个小高峰是在日本正式上线 7 月 20 日左右。

综合两个实验东方之星沉船事件和 Pokémon Go 事件，我们发现，事件热度模型每天找出的的热词都是符合事件的发展过程的，这个方法可以帮助按天找到事件的子事件。

4.1.4 实验 4：小时分析事件

本实验是选取事件以小时为单位分析事件热度。这里选取南海仲裁案，科比退役事件来分析。

图 4.7 是南海仲裁案的小时图，选取的是 2016 年 7 月 12 日当天上午 9 点到第二天下午 3 点的小时事件热度曲线。

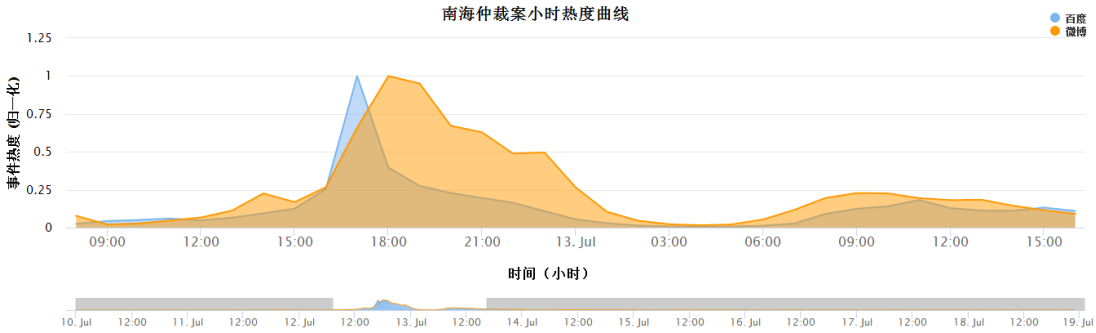

图 4.7 南海仲裁案小时热度曲线图

南海仲裁案是起源已久，最早可以追溯到 2014 年，但是在 2016 年 7 月 12 日北京时间 17 点，海牙国际仲裁法庭对“南海仲裁案”做出最终的“判决”，判菲律宾“胜诉”，中国方面拒不参与，不接受，不执行！

从图 4.7 我们可以看到，在 2016 年 7 月 12 日当天距离海牙法庭的判决结果出来前，事件的热度就开始发酵。在 16 点，事件热度增速加快，并一直激增到 17 点。海牙法庭做出判决后，引起国内激烈的讨论，所以从图中我们可以看出事件热度在 17 点的高峰后一直居高不下延续到 7 月 13 日凌晨 2 点才消退。

在 7 月 13 日 2 点之后是群众睡眠时间，热度降为 0，直到早晨 6 点开始，12 日事件余音未消仍旧在讨论中。在上午时间 10 点左右，事件热度开始增加，后来经过笔者的了解，当天 7 月 13 日上午 10 点中国国务院新闻办公室发表白皮书，并举办发布会又一次将事件热度引向高潮，从图中可以看出事件在 11 点达到顶峰。

下图 4.8 是科比退役事件的小时图，这里选取科比退役时间 2016.04.14 当天的小时图。

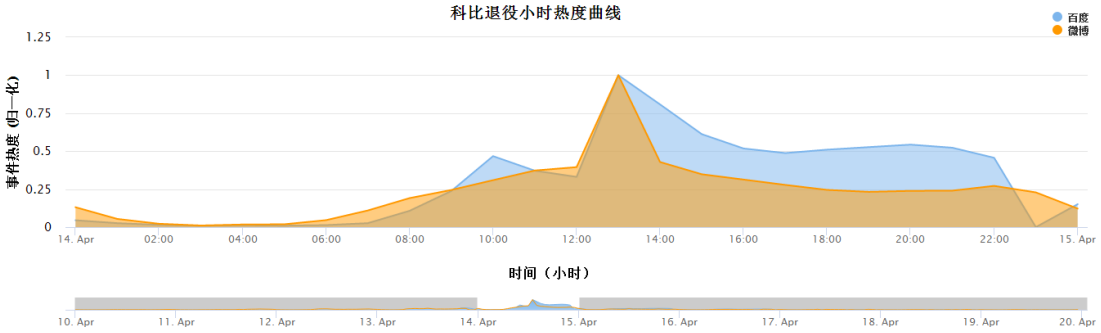

图 4.8 科比退役事件小时热度曲线

科比·布莱恩特最后一场比赛是在北京时间 2016 年 4 月 14 日当天 10:30 开始的直播。显然，由图可知事件热度在当天就一直是直线增加，并且在 10:30 前后有一个增加加速的趋势，比赛全程 1 小时 45 分钟，大约是在中午 12 点左右结束比赛，很明显的，我们可以看到在 12 点后事件热度处于一个激增的状态。这符合大家在比赛结束后进入激烈讨论状态的情况。事件热度居高不下维持到了 14 点，在此之后，科比退役事件就维持一个比较稳定的热度持续到了当天结束。

4.1.5 实验小结

事件热度模型主要做了 4 个实验，对比依据事件热度向量画出的事件热度曲线能不能比原始的 TF-IDF 算法有所优势，再者，本章想要通过画出事件热度曲线了解事件的发展流程以及发现子事件。根据以上的实验结果，我们可以得到以下几个结论。

1. 基于热词动态权重的热词选取算法要比 TF-IDF 的选词算法更加的好，选出的热词更具备代表性，可以将无意义或者跟事件无关的热词降低排名或者删除；
2. 实验 2 验证事件热度向量画出的事件热度曲线与 TF-IDF 选出的热词经

过相同操作画出的热度曲线的对比具有更好的灵敏度，对数据的反应更好；

3. 实验 3 对事件每天的新增词和每天的事件热度曲线进行了具体性分析，对应现实事件发展，验证了事件热度模型热词选取算法可以很好的提取出最能代表事件的热词，并且根据分析可以找到每天的子事件；
4. 实验 4 按小时对事件进行了分析，可以得出事件热度模型能灵敏地对每小时的事件进行捕捉分析。

4.2 事件预测模型及可视化分析

这章节的实验目的是为了验证混合概率预测模型的可用性以及他的局限性，并验证我们的跨平台的应用可以弥补到混合概率模型对于峰值捕捉的局限性。

对于实验结果的比较这里使用均方根误差（RMSE）。所谓均方根误差[29]就是计算所有预测值和真实值的差值的平方和和观察次数 n 的比值的平方根。均方根误差是为了描述预测数据与真实值的离散程度。均方根误差的计算公式如下：

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n}\left(y_{pre,i} - y_{true,i}\right)^2 \Big/ n} \tag{4-1}$$

其中，$y_{pre,i}$表示预测值，$y_{true,i}$表示真实值。

实验 5 我们进行混合概率预测模型（HPM）和单一概率模型（SPM）的对比，这里的单一概率模型使用的是高斯二项式模型。实验 6 是跨平台混合概率预测模型（CPHPM）的验证。实验 7 是误差对比混合概率预测模型，单一概率预测模型和跨平台混合概率预测模型。

4.2.1 实验 5：对比 HPM 和 SPM

这里本文选取了事件南海仲裁案、Pokémon Go 中国上线的事件、东方之星沉船事件以及英国脱欧事件进行分析。

1. 南海仲裁案预测分析

下图 4.1 是南海仲裁案百度预测分析图，4.2 是微博预测分析图。

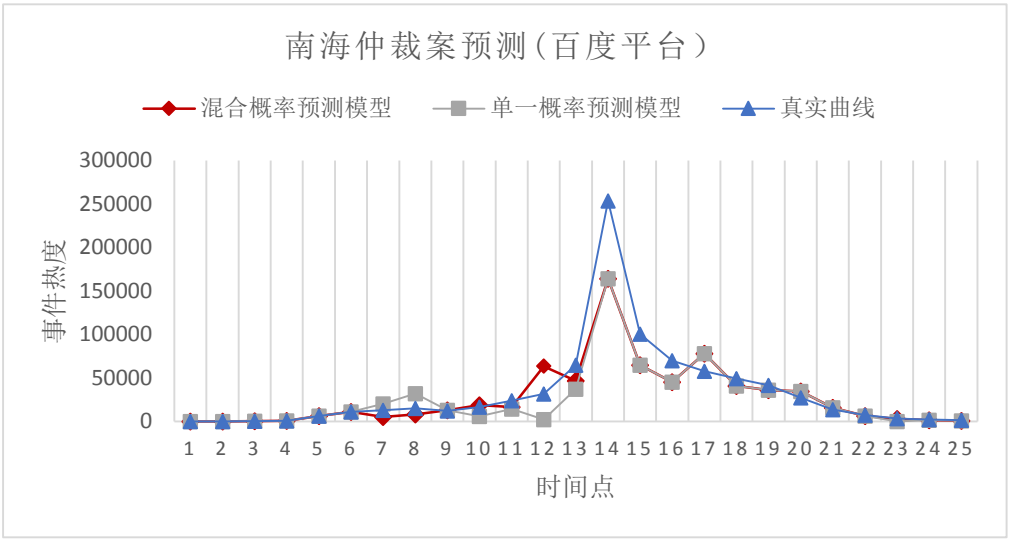

图 4.9 南海仲裁案百度预测分析图

从图中可以看出，三角形结点的曲线是事件的实际热度曲线，圆形结点的曲线是混合预测概率模型能够较好的拟合实际热度曲线的发展。在曲线的上升过程中，混合概率模型基本都是可以捕捉到的，而单一概率预测模型的波动就比较大；在曲线下降的过程，两种概率预测模型走势一致，基本重合，这是因为在下降的过程中混合概率预测模型所选取的模型跟单一概率预测模型是一样的。总而言之，在整个区间，混合概率预测模型拟合度更高。

计算均方根误差 RMSE 结果如下：

$$RMSE_{HPM} = 21845$$

$$RMSE_{SPM} = 22449$$

显然，混合概率预测模型的结果误差更小，模型更优。

下面是南海仲裁案微博预测。

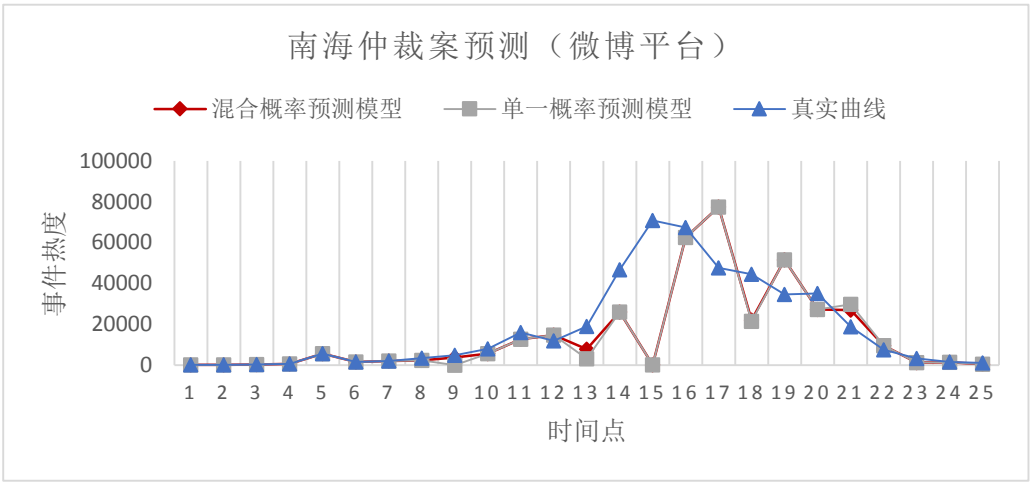

图 4.10 南海仲裁案微博预测分析图

从图中我们可以看出在曲线的上升趋势的前半段，混合概率预测模型是比单一模型要拟合度高一点；但是二者在峰值那里都没有很好的预测到；在曲线的下降趋势的预测，混合概率预测模型比单一模型又是稍微好一点。

计算均方根误差 RMSE 结果如下：

$$RMSE_{HPM} = 17216$$

$$RMSE_{SPM} = 17479$$

根据 RMSE 的值可以看出二者差距不大。但仍旧是混合概率预测模型拟合程度更高。

2. 手游 Pokémon Go 中国上线事件

下图 4.11 是 Pokémon Go 百度预测分析图，4.12 是微博预测分析图。

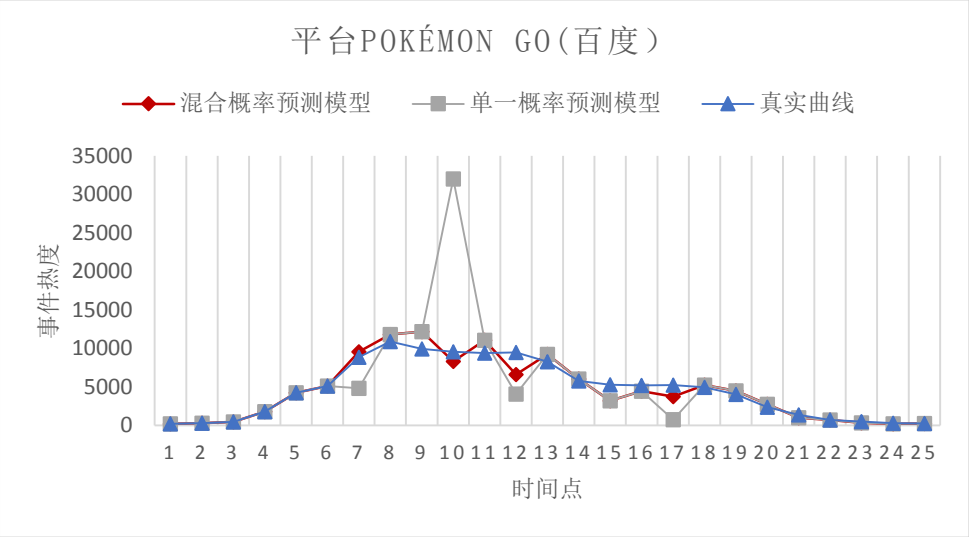

图 4.11 Pokémon Go 百度预测分析图

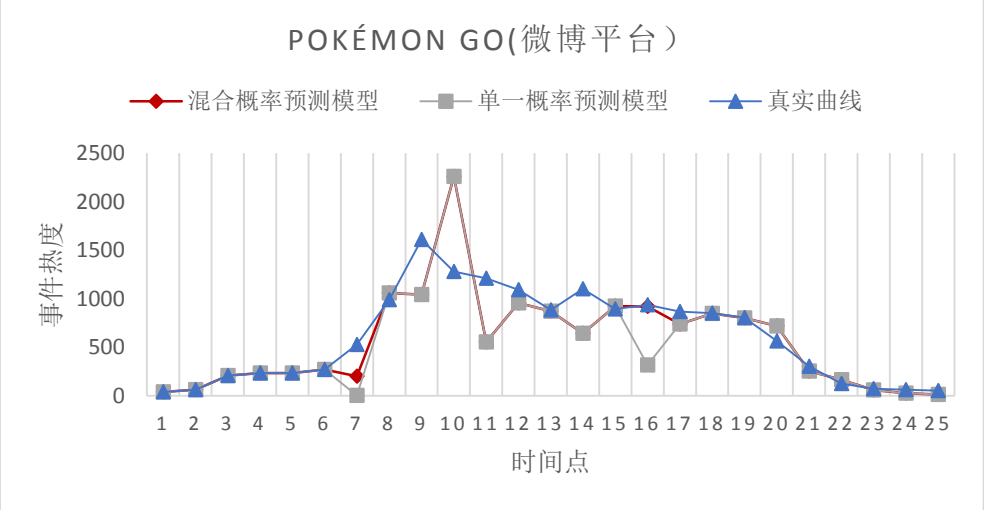

图 4.12 Pokémon Go 微博预测分析图

从图中我们可以看出，同南海仲裁案一样，在事件的上升阶段和下降阶段混合概率预测模型都是比单一概率模型的拟合程度要高。但是，峰值点的预测两种模型都表现的不好。

计算百度平台均方根误差 RMSE 结果如下：

$$RMSE_{HPM} = 1049.6$$

$$RMSE_{SPM} = 4833.4$$

计算微博平台均方根误差 RMSE 结果如下：

$$RMSE_{HPM} = 289.7$$

$$RMSE_{SPM} = 325.9$$

根据 RMSE 的值我们可以看出混合概率预测模型拟合程度高。

继续比较东方之星沉船事件和英国脱欧。我们发现与前文讲述的两个事件的规律是一致的。混合概率预测模型在预测的过程中基本是可以捕捉事件的上升和下降的过程，从整体来看，混合概率预测模型能够紧密的围绕实际热度，对于百度平台的分析如 4.9 和 4.11 表现的更加明显。但是该模型的明显的缺陷是无法捕捉峰值点，对于后面的预测带来误差，这点四张图都清晰的表现出来了。这也是跨平台的引入原因。

总而言之，混合概率预测模型有着它的缺点，但是相比于单一概率模型，混合概率预测模型能够保证预测的稳定性，预测的结果更具参考性。

4.2.2 实验 6：跨平台混合概率预测模型

实验 6 是跨平台混合概率预测模型的验证。选取实验 5 的事件南海仲裁案和手游 Pokémon Go 事件进行跨平台分析。

下图 4.13 是南海仲裁案双平台事件热度图，其中黑色表示微博平台，蓝色是百度平台，图 4.14 是南海仲裁案跨平台微博分析图。

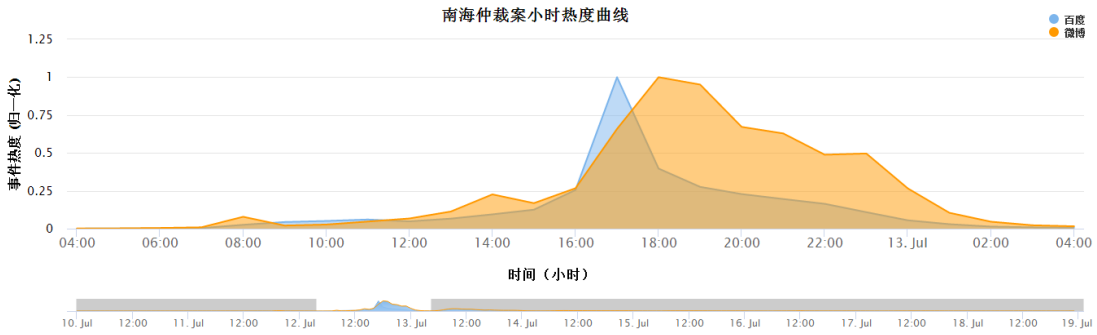

图 4.13 南海仲裁双平台事件热度图

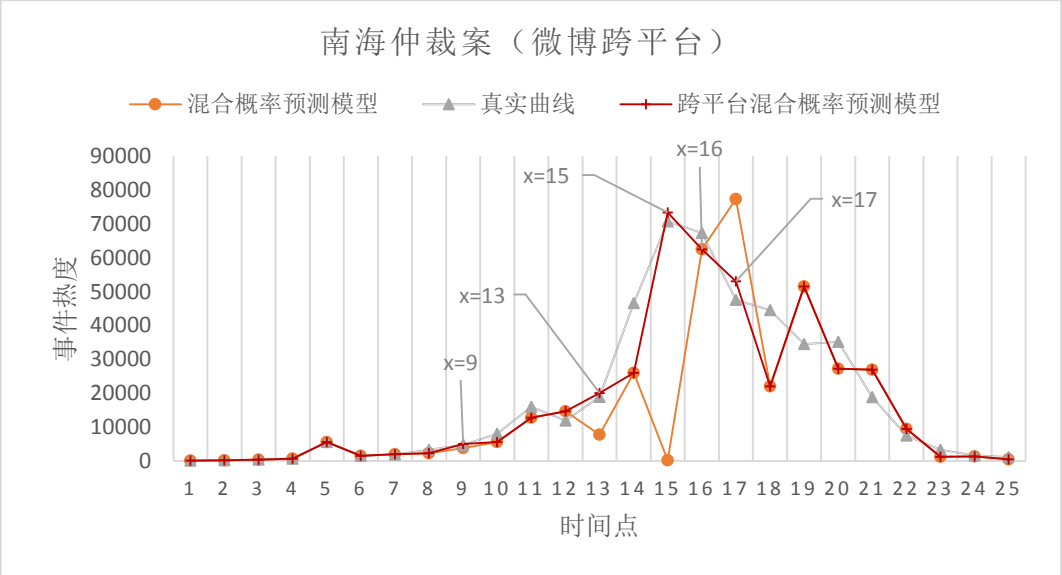

图 4.14 南海仲裁案跨平台微博分析

应用跨平台混合概率预测模型所得的预测曲线如图中“×”形结点的曲线所示。三角形结点的曲线是南海仲裁案事件热度实际曲线图。在图 4.14 中，每个点代表图 4.13 的一个小时，时间从 2016 年 7 月 12 日凌晨 4 点开始。从图中我们看出跨平台事件预测曲线的拟合程度很高。在预测过程中我们应用了跨平台思想来解决混合概率预测模型峰值捕捉的问题。

在预测 x=9,x=13 和 x=15 的过程中，符合跨平台分析预测算法的有概率模型，发现在 lagt 的时间差下，在图 4.13 中，可以看到在发展较快的百度平台中，实际的热度一直呈现上升趋势，因此，在这些点的预测中，考虑到跨平台的影响，我们选择在这点是上升趋势的模型，结果真实事件的趋势果然也是呈上升趋势。同理，我们在 x=17 时，选择下降的趋势的模型。对比，我们可以看到跨平台的考虑成功弥补了混合概率模型的峰值捕捉失败和峰值后什么时候下降的问题。

在预测 x=16 时，从图 4.14 的圆形结点曲线可以看出混合概率预测模型还是上升的趋势，但是对比图 4.13 百度这个发展较快的平台，我们可以看到，百度平台在点 14 到点 15 时就已经开始下降，人工估计 lagt 约是 0.5 小时，那么微博平台的峰值极有可能是百度峰值后的 0.5 小时候达到，推测微博在点 x=15 就是峰值 ，x=16 呈下降趋势，因此选择下降趋势的模型，实验证明，我们的猜想是正确的。

下图 4.15 是 Pokémon Go 的双平台事件热度图，图 4.16 是 Pokémon Go 跨平台微博分析图。

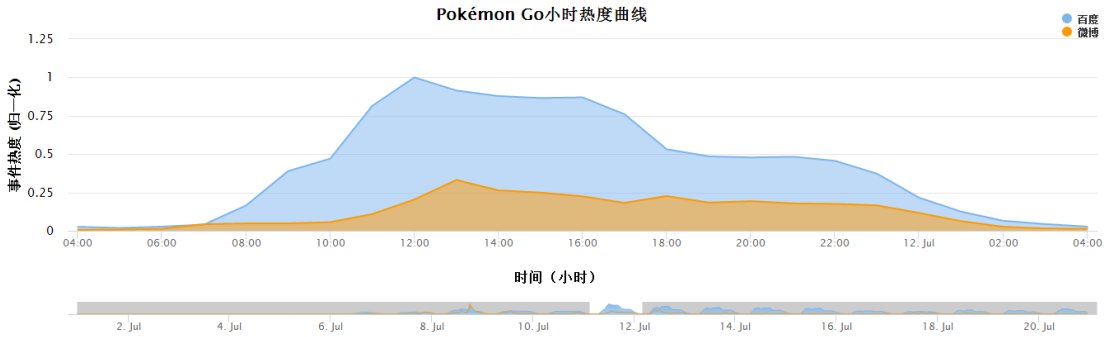

图 4.15　Pokémon Go 双平台事件热度图

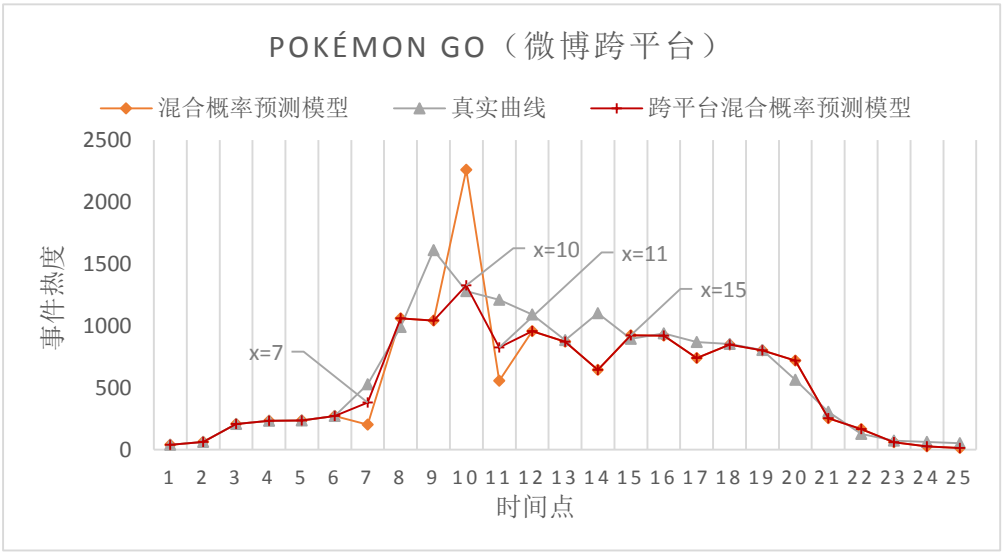

图 4.16　Pokémon Go 跨平台微博分析图

从图 4.15 可以看出，微博平台在此数据集呈现明显的延后趋势，并且到达峰值的时间差是 1 个小时。

在图 4.16 中，每个点代表图 4.15 的一个小时，时间从 2016 年 12 月 7 日凌晨 5 点开始。在预测 x=7 时，混合概率预测模型呈下降趋势。但是结合跨平台考虑，在百度平台事件是呈上升趋势，所以我们选择一个上升的模型来预测热度。最终得到热度也印证了该过程为是上升趋势，预测热度也是在理想范围内。

在预测 x=10 的热度时，混合概率预测模型是上升趋势明显，但是在百度平台中 x 从 6 到 9 过程中出现上升趋势放缓，并且在 x=9 时出现了下降，故推测后一平台即将达到热度峰值，但是此时，并没有一个模型呈现的是下降趋势，所以按照跨平台混合概率预测模型无模型可供选择的情况，利用EP_{t+1}^{B}=V × $(EP_{t}^{B} - EP_{t-1}^{B}) + EP_{t}^{B}$算出跨平台混合概率预测模型预测的值。

在预测 x=11 和 x=15 的热度时，混合概率预测模型最优模型和发展较快的

百度平台一致处于平稳下降的趋势并且未出现波动情况。所以在选择模型时考虑到跨平台，选择下降的模型，结果也很好的印证了跨平台在选择模型中的运用。

综上，我们可以看出跨平台混合概率模型在捕捉峰值和判断何时从峰值下降有着很好的辅助作用。

4.2.3 实验 7：误差对比分析

实验 7 是将单一概率预测模型、混合概率预测模型和跨平台混合概率预测模型进行对比实验，进行误差分析，使用公式(4-1)计算 RMSE。这里还是用南海仲裁案和 Pokémon Go 事件进行对比。因为这两个事件都是微博平台比百度平台发展较慢，跨平台的应用只能是放在微博平台。下表 4.5 是南海仲裁案和 Pokémon Go 在微博平台用三种模型预测额均方根误差值 RMSE。下图 4.17 是精度误差对比图。

表 4.5 微博平台事件误差 RMSE

	SPM	HPM	CPHPM
南海仲裁案	17479	17216	7595.1
Pokémon Go	325.9	289.7	176.3

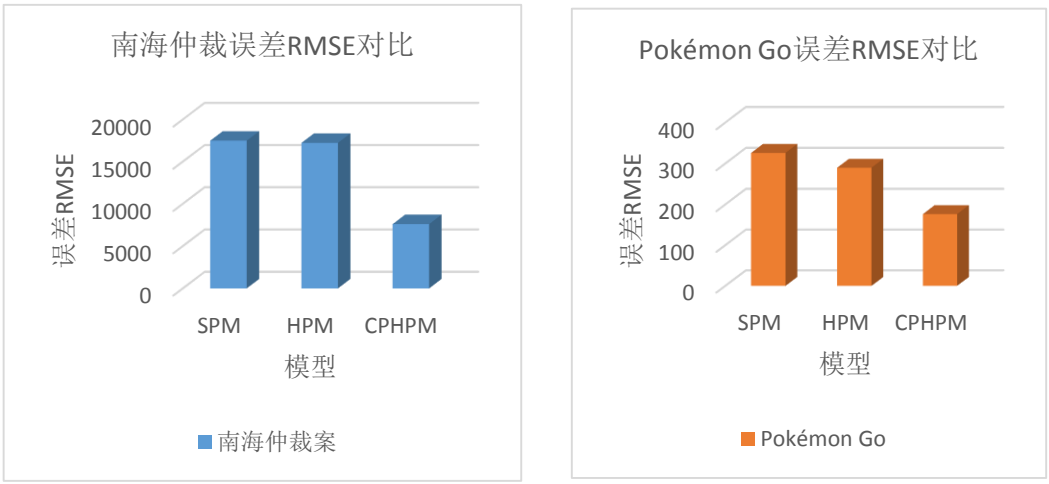

图 4.17 精度误差对比图

由图可知跨平台混合概率预测模型是三种方案里面误差最小的。而混合概率模型比单一概率模型要好，但是可以看出这两个模型之间的误差和跨平台混合概率模型的误差相比很小，这是因为单一概率预测模型使用的是高斯二项式模型，本文的实验证明高斯二项式模型在很多时候就是最优模型。

跨平台混合概率预测模型是三种方案里面误差最小的，这表明本文提出利用跨平台的思想是合理的，它的确能够帮助混合概率模型提高准确率。

4.2.4 实验小结

混合概率预测模型对事件热度进行预测，跨平台的思想是针对混合概率预测模型对峰值捕捉不确定性的缺点提出的。通过本节的三个实验可以得到以下结论：

1) 混合概率预测模型要优于单一概率预测模型。但是，混合概率预测模型的缺点是无法捕捉到峰值点。
2) 跨平台思想的应用的确可以帮助混合概率预测模型的预测，跨平台混合概率预测模型的误差最小。
3) 跨平台混合概率预测模型目前还不完善，不能依据算法计算 lagt 而是靠人工完成。

总而言之，跨平台的应用是有效的合理的，以后的研究完全可以优化跨平台这方面提高预测的准确度。

第 5 章 总结与展望

互联网行业的迅猛发展，网络上事件的发展的不确定性和多变性让了解事件的发展过程和趋势变得尤为重要。

本文具体提出了两个模型。第一个模型是事件热度模型。在这个模型中提出了一个新颖的选词算法来选出最能代表事件的热词，再利用选出的热词计算事件热度向量，用以描述一个事件在一定时间内的热度。

第二个模型事件预测模型是根据第一个模型的输出事件热度向量完成事件预测功能。提出了混合概率预测模型，并利用跨平台的思想弥补了混合概率预测模型的缺陷提出了跨平台混合概率预测模型。经过试验证明，跨平台的引入是创新性的，也是有效的、合理的。

本文通过 8 个来自真实世界的热门事件来对两个模型进行试验，试验结果证明，事件热度模型能够选出具有代表性的热词，对事件的刻画具备灵敏性，能够捕捉到事件的子事件，而事件预测模型也能够比较好的预测事件的发展趋势。

但是，无法否认的是，事件预测模型的跨平台混合概率模型也是有缺陷的。跨平台的引入本文还没有很好的算法计算出平台之间的时间差 lagt，本文还没有准确的算法来量化两个平台间的相似性。目前这两个工作本文是靠人工完成。本文现在做的就是引入跨平台的思想并且证明这个想法是有效的。

在未来的工作中，会主要针对跨平台混合概率预测模型的改进，设计出算法计算 lagt，并确定相似性的阈值条件。相信完成这些工作后，本文的提出的模型才算是完善的。

本科期间发表成果

1. CSCW Poster: “Is Danmaku an effective way for Supporting Event based Social Network?”(第二作者)；
2. CSCW Full paper submitted: “Hey Front Guy, Stop Running!”: Unders tanding EventMediated Crowd Interaction through Danmaku；
3. 创业产品 USee，获得 2016 年“创青春”全国大学生创业大赛电子商务专项赛校内选拔赛金奖；
4. 创业产品 USee，获得 2016 年河海大学创新创业大赛暨第二届中国”互联网+“校内选拔赛一等奖。

致谢

论文写到这里为止了，好像也为大学四年画上了休止符。一时间竟然感慨万千，从一开始写论文的茫然无知，到后来实验遇到问题瓶颈期的急躁，再到现在有条理有目的的跟进论文，不能不说，我从中学到了很多。写论文的过程是对自己的思维一个锤炼的过程，是对自己的逻辑一个验证的过程，也是对我的心境一个提升的过程。大学四年过得太过于急躁喧嚣，这次写论文的过程反而让我有了以前努力奋斗的感觉。书山有路勤为径，学海无涯苦作舟。这是个很累但是我很喜欢的过程。感谢带着我做实验的、教我写论文的唐老师！感谢在我遇到问题时帮助我的同学！

参考文献

[1] Siddiqi S, Sharan A. Keyword and Keyphrase Extraction Techniques: A Literature Review[J]. International Journal of Computer Applications, 2015, 109(2):18-23.

[2] Xie F, Wu X, Zhu X. Document-Specific Keyphrase Extraction Using Sequential Patterns with Wildcards[C]. IEEE International Conference on Data Mining. IEEE, 2014:1055-1060.

[3] Marujo, Luis and Ling, Wang and Trancoso，et al. Automatic Keyword Extraction on Twitter[M]. ACL(2), 2015:637-643.

[4] Liu J, Shang J, Wang C, et al. Mining Quality Phrases from Massive Text Corpora[C]. Acm Sigmod International Conference on Management of Data. Proc ACM SIGMOD Int Conf Manag Data, 2015:1729.

[5] Yang K, Chen Z, Cai Y, et al. Improved Automatic Keyword Extraction Given More Semantic Knowledge[M]. Database Systems for Advanced Applications. Springer International Publishing, 2016.

[6] Liu P, Azimi J, Zhang R. Automatic keywords generation for contextual advertising[C]. International Conference on World Wide Web. ACM, 2014:345-346.

[7] Gollapalli S D, Caragea C. Extracting keyphrases from research papers using citation networks[C]. Twenty-Eighth AAAI Conference on Artificial Intelligence. AAAI Press, 2014:1629-1635.

[8] Zhang X, Chen X, Chen Y, et al. Event detection and popularity prediction in microblogging[J]. Neurocomputing, 2015, 149:1469-1480.

[9] Sakaki T, Okazaki M, Matsuo Y. Tweet Analysis for Real-Time Event Detection and Earthquake Reporting System Development[J]. IEEE Transactions on Knowledge & Data Engineering, 2013, 25(4):919-931.

[10] Nguyen D T, Jung J E. Real-time event detection for online behavioral analysis of big social data[J]. Future Generation Computer Systems, 2017, 66:137-145.

[11] Zhou D, Chen L, He Y. An unsupervised framework of exploring events on twitter: filtering, extraction and categorization[C]. Twenty-Ninth AAAI Conference on Artificial Intelligence. AAAI Press, 2015:2468-2474.

[12] Adedoyin-Olowe M, Gaber M M, Dancausa C M, et al. A rule dynamics approach to event detection in Twitter with its application to sports and politics[J]. Expert Systems with Applications, 2016, 55(C):351-360.

[13] Zhang C, Zhou G, Yuan Q, et al. GeoBurst: Real-Time Local Event Detection in Geo-Tagged Tweet Streams[C]. International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2016:513-522.

[14] Guo J, Gong Z. A Nonparametric Model for Event Discovery in the Geospatial-Temporal Space[C]. ACM International on Conference on Information and Knowledge Management. ACM, 2016:499-508.

[15] Mathioudakis M, Koudas N. TwitterMonitor:trend detection over the twitter stream[C]. ACM SIGMOD International Conference on Management of Data, SIGMOD 2010, Indianapolis, Indiana, Usa, June. DBLP, 2010:1155-1158.

[16] Okazaki M, Matsuo Y. Semantic Twitter: Analyzing Tweets for Real-Time Event Notification[J]. Breslin J, Burg T, Kim HG, Raftery T, Schmidt JH (eds) Recent Trends and Developments in Social Software, Lecture Notes in Computer Science, vol 6045, Springer, Berlin, Heidelberg, 2011:63–74.

[17] Saleiro P, Soares C. Learning from the News: Predicting Entity Popularity on Twitter[M]. Advances in Intelligent Data Analysis XV. Springer International Publishing, 2016.

[18] Lymperopoulos I N. Predicting the popularity growth of online content: Model and Algorithm[J]. Information Sciences, 2016, 369:585-613.

[19] Lin C X, Zhao B, Mei Q, et al. PET: a statistical model for popular events tracking in social communities[C]. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2010:929-938.

[20] Zhao J, Wu W, Zhang X, et al. A Short-Term Prediction Model of

Topic Popularity on Microblogs[M]. Computing and Combinatorics. Springer Berlin Heidelberg, 2013:759-769.

[21] Mishra S, Rizoiu M A, Xie L. Feature Driven and Point Process Approaches for Popularity Prediction[J]. Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, 2016:1069-1078.

[22] Wang S, Yan Z, Hu X, et al. Burst time prediction in cascades[C]. Twenty-Ninth AAAI Conference on Artificial Intelligence. AAAI Press, 2015:325-331.

[23] Zhang X, Li Z, Chao W, et al. Popularity Prediction of Burst Event in Microblogging[J]. 2014, 8485:484-487.

[24] Yuan K, Wu J, Zhao Z. Burst prediction from Weibo: A crowd-sensing and tweet-centric method[C]. International Conference on Service Systems and Service Management. 2016:1-6.

[25] Ardon S, Bagchi A, Mahanti A, et al. Spatio-temporal and events based analysis of topic popularity in twitter[C]. ACM International Conference on Information & Knowledge Management. ACM, 2013:219-228.

[26] Bao B K, Xu C, Min W, et al. Cross-Platform Emerging Topic Detection and Elaboration from Multimedia Streams[J]. Acm Transactions on Multimedia Computing Communications & Applications, 2015, 11(4):1-21.

[27] Roy S D, Mei T, Zeng W, et al. Towards Cross-Domain Learning for Social Video Popularity Prediction[J]. IEEE Transactions on Multimedia, 2013, 15(6):1255-1267.

[28] Tang Y, Ma P, Kong B, et al. ESAP: A Novel Approach for Cross-Platform Event Dissemination Trend Analysis Between Social Network and Search Engine[M]. Web Information Systems Engineering – WISE 2016. Springer International Publishing, 2016.

[29] 曹卫峰. 中文分词关键技术研究[D]. 南京理工大学, 2009.

[30] Chum O, Philbin J, Zisserman A. Near Duplicate Image Detection: min-Hash and tf-idf Weighting[C]. British Machine Vision Conference 2008, Leeds, Septembe

r. DBLP, 2008.

[31] 白丁. 《误差理论与实验数据处理》[J]. 科学通报, 1964, 9(8):752-752.