文章编号：1003-0077(2018)05-0105-09

基于主题模型的新疆暴恐舆情分析

张绍武[1,2]，邵 华[1]，林鸿飞[1]，杨 亮[1]

(1. 大连理工大学 计算机科学与技术学院，辽宁 大连 116024；
2. 新疆财经大学 计算机科学与工程学院，新疆 乌鲁木齐 830013)

摘 要：随着互联网的飞速发展，网络舆情引发的问题也越发突出。尤其是近年来发生的新疆暴恐事件，已成为公众关注的焦点。主题演化是网络舆情分析的重要内容之一，为了把握关于新疆的舆情动态，该文从主题热度变化、内容变化及关键词等多方面进行了研究。该文首先抓取了 2013 年 1 月到 2015 年 12 月互联网中关于新疆暴恐事件的新闻，并以此作为数据集建立了动态主题模型，实现对新闻的主题演化分析。该模型采用两次非负矩阵分解来生成主题，以层级式狄利克雷过程为对比实验，通过可视化分析与比较，总结出新疆暴恐事件的一些规律。

关键词：动态主题模型；层级式狄利克雷过程；主题模型；可视化

中图分类号：TP391 **文献标识码**：A

Public Opinion Analysis for Xinjiang Violence News Based on Topic Model

ZHANG Shaowu[1,2], SHAO Hua[1], LIN Hongfei[1], YANG Liang[1]

(1. Dalian University of Technology, Institute of Computer Science and Technology, Dalian, Liaoning 116024, China;
2. Xinjiang University of Finance and Economics, Institute of Computer Science and Engineering, Urumqi, Xinjiang 830013, China)

Abstract: With the explosive growth of networks, the internet public opinion becomes a non-negligible issue. A typical example is focus on the events about Xinjiang Violence happened in recent years. In order to examine the corresponding public opinion trends, this paper investigates the key words of topic and the change of both topic strength and its content. On the crawled news about Xinjiang Violence from 2013.01 to 2015.12, we apply the dynamic topic model (DTM) which generate topics by applying NMF twice. Compared to HDP, we reveal some properties by visualized analysis.

Key words: DTM; HDP; topic model; visualization

0 引言

近几年，随着互联网的飞速发展，新闻在网络平台上得到广泛的传播。由于网络新闻是网络舆情传播的主要途径之一，随之而来的由网络舆情引发的问题也日益突出。尤其是近年来发生的新疆暴恐事件，引起了公众的热议和关注，并形成了强大的舆论动向。随着时间的推移不断演化发展，网络舆论会给人们的现实生活带来一些影响，同时也会给社会管理工作提出一些要求，带来一些困难。所以，准确把握舆论动向，有助于政府对新疆暴恐问题采取及时有效的监管和处理措施。

主题演化是网络舆情分析的重要部分。主题模型作为新的一种统计方法，用来发现文本中蕴含的主题，已被广泛地运用在文本挖掘和信息检索等领域中，并且在主题演化方面也得到了广泛的发展。本文以层级式狄利克雷过程(Hierarchical Dirichlet Processing, HDP)模型作为对比实验，运用动态主题模型(Dynamic Topic Model，DTM)，通过分析和比较，总结出新疆暴恐主题在演化中存在的某些规律。

本文的主要贡献是：针对新疆暴恐网络舆情问题，结合数据集涵盖暴恐这一特点，对基于 NMF 的

收稿日期：2017-03-03 **定稿日期**：2017-05-26

基金项目：国家自然科学基金(61562080，71561025，61632011)

DTM主题演化模型进行了改进，在主题一致上取得了更好的效果，并通过比较和分析给出了新疆暴恐主题演化中存在的某些规律。

本文组织如下：第一节将对相关工作进行介绍；第二节介绍实验用到的方法及实验过程；第三节进行实验结果分析；第四节总结并规划未来工作。

1 相关工作

主题模型旨在从海量文本数据中挖掘出有价值的主题，然后对主题进行检测、跟踪和预测。主题演化就是从主题的产生、发展、再到成熟，最后到消失的一系列过程。当前主题演化模型主要分为两大类：基于LDA概率模型和基于矩阵分解模型。此外，本节也在1.3中对新疆暴恐舆情分析相关研究进行了介绍。

1.1 基于概率模型的主题演化方法

TOT(Topic Over Time)模型[1]最早被提出，它是在LDA模型中引入时间因素构建而成，实现简单。TOT将时间也作为可观测变量，然后与文档和单词一起生成主题。DTM(Dynamic Topic Model)[2]先根据时间窗分割文本集合，并假设每个时间窗口的文本都由K个话题的LDA模型生成。上述模型都是在LDA基础上，扩展改进后得到的。其思路及方法都较为简单，而且在主题个数方面都缺少灵活性。

2008年，Ahmed等[3]人提出TDPM (Temporal Dirichlet Process Mixture Model)，通过Dirichlet Process确定演化过程中每个时间窗中的主题个数。2010年，Ahmed等[4]人又提出iDTM (infinite Dynamic Topic Models)，引入HDP[5-7]方法，解决了单纯使用LDA过程中各时间窗内主题数固定的问题。

HDP[5]模型是一个层级式的狄利克雷过程。它相比LDA，使主题个数的选取变得灵活。该模型可以用中国餐馆过程(CRP)来描述，如式(1)～(2)所示。在CRP中，每个文本对应一个餐馆，单词对应顾客。在餐馆中顾客围绕不同的桌子而坐，每张桌子供应一道菜。这里的菜相当于主题。用 m_{tk} 表示在时间窗口 t 下所有餐馆中供应第 k 道菜的桌子数，用 n_{ib}^{t} 表示在时间窗口 t 下坐在餐馆(文本)i 里的第 b 张桌子的单词数。当新的顾客进来时，他选择桌子的过程服从分布(1)。

$$\theta_{ij}^{t} \mid \theta_{i1}^{t},\ldots,\theta_{ij-1}^{t},\alpha \sim \sum_{\varphi_{itb}=\theta_{ij}} \frac{n_{ib}^{t}}{n_{i}^{t}-1+\gamma}\delta_{\varphi_{jb}} + \frac{\gamma}{n_{i}^{t}-1+\gamma}\delta_{\varphi_{jb}^{new}} \quad (1)$$

当顾客选择一张新桌子的时候，还需要给新桌子供应一道菜，选菜的过程服从分布(2)。

$$\varphi_{ti}^{new} \mid \varphi,\alpha \sim \sum_{k} \frac{m_{tk}}{\sum_{i} m_{ti}+\alpha}\delta_{\varphi_k} + \frac{\alpha}{\sum_{i} m_{ti}+\alpha}G_0 \quad (2)$$

1.2 基于矩阵分解的主题演化方法

非负矩阵分解(Non-negative Matrix Factorization，NMF)[8]是一种新的矩阵分解方法。一般的矩阵分解，如SVD(奇异值分解)，PCA(主成分分析)等都会出现分解结果中出现负值的情况。而负值在某些环境下是没有意义的，比如文本中单词的统计，数字图像中的像素等。

NMF是另一种有效的提取主题的方法[9-11]。处理大规模数据更快更便捷，且实现简便、占用存储空间少。Saha & Sindhwani[12]提出了在社交媒体上运用NMF做主题演化的方法。Derek Greene[13]在NMF的基础上对欧洲政治议程做了主题演化分析。

针对本实验数据类型是网络新闻这一特点，本文借鉴了Derek Greene提出的基于NMF的主题演化方法。由于本实验的数据集是结合新疆暴恐的，所以本实验在进行NMF分解时，对单词的权重进行了改进，可以看到改进后实验效果有一定的提升。最后为了验证该模型的有效性，和HDP模型进行了对比分析。

1.3 新疆暴恐舆情分析相关研究

近年来对新疆暴恐舆情研究的论文也有很多。如戴继诚[14]对当前新疆暴恐活动新变化的探析，发现受国内外各种因素影响，当前新疆的暴恐活动出现一些新的变化。如活动主体的年轻化、活动范围的扩大化、活动方式的小团体化、暴恐手段的激烈化等。王定等[15]对当前全球化背景下新疆暴恐活动呈现特点进行了研究。研究指出自2008年“七五事件”以来，一些极端的外来宗教渗透到新疆各地。西方敌对势力与反华势力对新疆分裂势力的支持，使中国境内的暴恐活动呈现高发状态。全球化时代的恐怖活动给我国的社会发展带来了严峻挑战。

然而，这些作者都是在传统的统计方法上，从政治，社会学的角度出发看待和研究问题。本文尝试通过机器学习的方法，即通过主题演化模型来描述

和分析数据，从而发现规律并给予验证。

2　基于 NMF 的 DTM 主题模型

本节主要介绍基于 NMF 的 DTM 方法。该方法基于新疆暴恐数据采用 NMF 两次分解，进行主题演化和舆情分析。目的是通过舆情分析，发现规律从而对舆情的把握提供一定的借鉴意义。实验首先从互联网中爬取新疆暴恐相关的新闻，然后对数据进行预处理，接着两次运用 NMF 生成窗口和动态主题，最后对实验结果进行可视化和分析。具体如框架图 1。

首先，本文对含有 n 个文档的语料集，建立一个矩阵 $A \in \mathrm{IR}n \times m$，其中 m 是语料集中不同单词数目。运用 NMF 方法，就是把 A 近似成两个非负矩阵乘积的形式 $A \approx WH$，并且最小化 A 与 WH 间的误差。其中的 $H \in \mathrm{IR}k \times m$，每一行代表一个 topic，行上的元素代表了不同单词在该 topic 下的相对权重。$W \in \mathrm{IR}n \times k$ 的每一列，表示了文档对该 topic 的贡献情况。

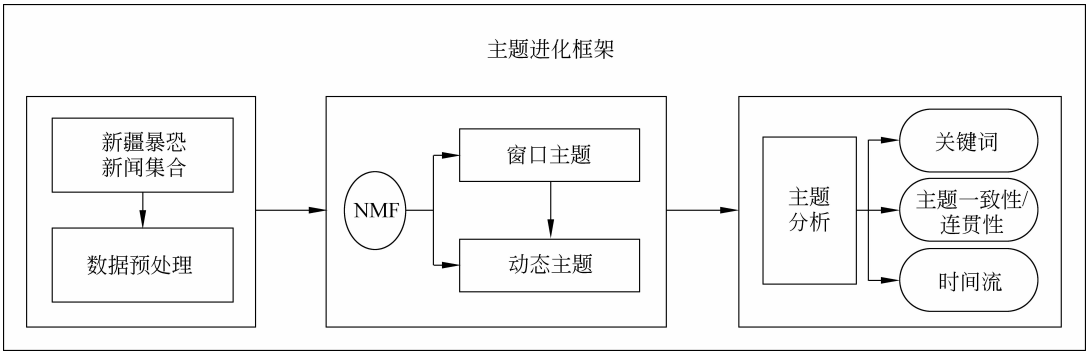

图 1　实验框架图

本文通过设定主题个数区间来增加主题个数的灵活性，然后选取主题连贯性取值最大的个数作为最终的主题个数。这里的 topic coherence，计算公式如式(3)、式(4)所示。

$$\mathrm{coh}(t_h) = \frac{1}{\binom{t}{2}} \sum_{j=2}^{t} \sum_{i=1}^{j-1} \cos(wv_i, wv_j) \tag{3}$$

$$\mathrm{coh}(T) = \frac{1}{k} \sum_{h=1}^{k} \mathrm{coh}(t_h) \tag{4}$$

对每个主题，文中采用 TC-W2V[8]方法来计算主题连贯性。即用主题的关键词集合在 word2vec 词向量空间上的相似度，来表征该主题关键词间的相关程度。实验中基于权重较大的前 t 个单词两两之间的余弦相似度均值来表示。最后该模型的主题连贯性用所有 topic 的 coherence 均值来表示。

由于主题动态演化的特性，某些主题可能分布在多个时间窗口下，所以还需要生成动态主题。整体计算过程如下：

1) 计算窗口主题(window topics)

① 删除窗口下出现文档数小于 5 的单词。

② 构造文本和单词输入矩阵，计算单词的 TF-IDF 权重，计算式如式(5)所示。

$$w(d,r) = \frac{(1+\log_2 TF(d,r)) \times \log_2 IDF(r) \times s(r)}{C} \tag{5}$$

C 是归一化系数，$w(\mathrm{d},\mathrm{r})$ 对应文本 d 中单词 r 的权重，$s(\mathrm{r})$ 对应单词 r 与暴恐的相关程度，实验中用该单词与暴恐词集合的相似程度来表示，即在 word2vec 词向量空间上的余弦相似度均值。

③ 选取要生成的主题个数 k 的区间，本实验选取 4～25。

④ 对每个 k 运用 NMF 生成主题，计算该 k 个主题下的主题连贯性，然后选取取值最大的 k 作为窗口最终的主题个数。

该算法伪代码如算法一所示。

算法 1：generating window topics：

```
Input:
A              : a matrix of document-word weights
k_min,k_max    : interval of topic numbers
w2v-bin        : word2vec for the words
Output:
H              : a matrix of topic-word weights
for every time window T:
  mx=0,best_k=-1
  for K in the range(k_min,k_max):
    do NMF using the inputs above and outputs H
    sum=0
    for topic k in the range(1,K):
      compute topic coherence coh using H
      sum+=coh
```

续表

算法 1：generating window topics：

```
9      sum/=K
10      if best_k==-1 or mx<sum:best_k=K,mx
        =sum
```

2）计算动态主题(dynamic topics)

① 构造一个空矩阵 B，对于每个时间窗口计算出的 H，在每一行选取前 t 个权重较大的单词，其余单词权重设为 0，然后把该行添加到 B 里。最后去掉 B 中只包含 0 元素的列。

② 采用 NMF 对 B 进行分解。B 分解后的 H，其每一行的前 t 个单词，描述了本行的动态主题。B 分解后的 W，其每一列表示了各个时间窗口和该动态主题的相关程度。

该算法的伪代码如算法 2 所示。

算法 2：generating dynamic topics：

```
   Input:
   B              : a matrix of window-topic-word weights
   k_min,k_max    : interval of topic numbers
   w2v-bin        : word2vec for the words
   Output:
   H              : a matrix of dynamic-topic-word weights
1  B=[]
2  for every time window T:
3     select top t words in H,and add the row in B
4  remove the empty columns in B
5  mx=0,best_k=-1
6  for K in the range(k_min,k_max):
7        do NMF using B and outputs H
8        sum=0
9        for topic k in the range(1,K):
10           compute topic coherence coh using H
11           sum+=coh
12     sum/=K
13     if best_k==-1 or mx<sum:best_k=K,mx=sum
```

3 实验与分析

3.1 数据集与预处理

本文的数据集依赖一个谷歌的全球新闻关系数据库(Gdelt)。数据集来源于数据库中从 2013 年开始至今全球每天发生的重大新闻事件。它有两种存储方式，一种是按天划分以 csv 的格式保存在硬盘里，另一种是保存在谷歌的 BigQuery 数据库里。两种形式都记录了事件发生的时间、地点，事件的类别，事件的发起者和承受者，事件新闻所在的网址以及谷歌标注的情感分数等。而不足之处在于数据库没有提供新闻的文本内容。

本文首先按关键字从数据库里导出 2013 年 1 月至 2015 年 12 月关于新疆的新闻网址，然后通过自编爬虫程序，爬取网址上带有＜p＞＜/p＞标签的文本。为了获取关于暴恐相关的新闻，本文利用暴恐相关的关键词如 attack、terrorist、kill 等，采取“或”的方式对文本进行过滤，只留下包含关键词的新闻。由于某些网址失效或网络连接不到，最终得到的文档数量是 14 416，单词集合大小在 131 000 左右。

实验首先对数据进行了清洗工作。先进行了去停用词、词干化处理，并删去了长度小于 4 的单词。本文以季度为单元划分时间窗口，每个时间窗口下的数据集是对应季度的新闻集合。最终划分后的数据集如图 2 所示，横坐标代表时间窗，纵坐标代表时间窗内文本数。

从图中可以看出 2014 年 1～3 月及 7～9 月以及 2015 年下半年的新闻量较大。在这些时间段内发生过一些影响比较大的事件，分别为 2014 年 3 月发生的昆明火车站暴力恐怖袭击事件，7 月发生的

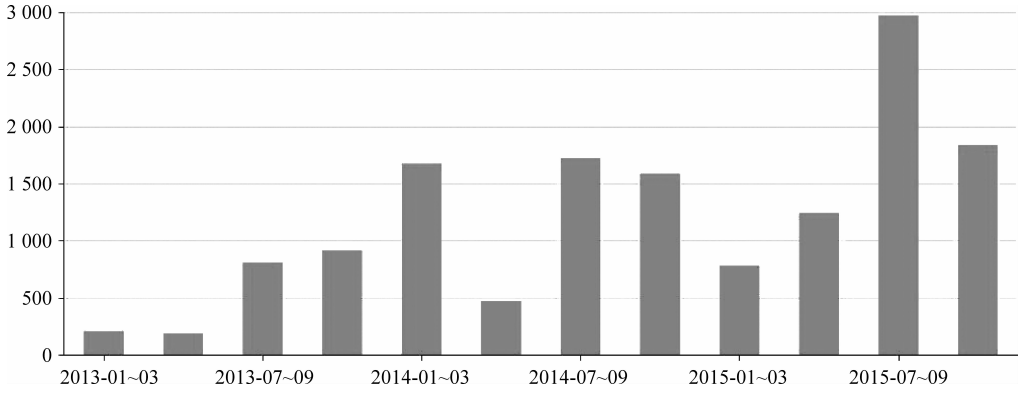

图 2 各时间窗口下文档数分布

莎车县爆恐袭击案，2015 年阿克苏地区拜城县爆恐袭击案以及最近的泰国移民事件。这些事件不仅在新闻上迅速传播，在微博等社交媒体上也迅速蔓延开来，形成强烈的网络舆论。

3.2　结果分析

本文共做了 3 组实验，第一组是不同时间窗下的主题一致性分析，通过改进后的 NMF 和原始 NMF 及 HDP 作对比，验证了改进后的 DTM 模型的效果。见表 1。第二组实验和第三组实验是对整个时间序列下动态主题的分析，其中第二组实验分析了动态主题的演化，见表 2。第三组实验分析了主题关键词的演化，见表 3。

3.2.1　不同时间窗下主题一致性分析

本节实验通过主题分析，来验证改进后的 DTM 方法其在主题连贯性值上的提升，同时与 HDP[5] 方法作对比，验证了针对本实验数据，该方法比 HDP 更加适用。

表 1　DTM 与 HDP 每个时间窗下的主题连贯性

model	2013-01～03	2013-04～06	2013-07～09	2013-10～12	2014-01～03	2014-04～06	2014-07～09	2014-10～12	2015-01～03	2015-04～06	2015-07～09	2015-10～12
DTM	0.182 6	0.171 5	0.170 8	0.174 5	0.182 4	0.172 1	0.182 4	0.179 0	0.197 0	0.186 0	0.189 8	0.180 4
HDP	0.159 3	0.142 0	0.162 0	0.154 5	0.161 1	0.142 4	0.153 4	0.170 0	0.148 2	0.109 5	0.173 8	0.139 8
DTM1	0.178 5	0.169 2	0.168 3	0.169 5	0.176 8	0.171 0	0.179 5	0.177 3	0.189 0	0.181 2	0.184 5	0.181 0

表 1 是 3 个方法在各个时间窗下生成最优主题数后的主题连贯性值，表 1 里的 DTM1 表示 NMF 的输入单词权重矩阵是原始的 TF-IDF，DTM 表示本实验改进后的模型。通过对比可以发现，结合新疆暴恐的特点，对主题的连贯性有一定的提升。

表 2　DTM 与 HDP 在每个时间窗最优主题数

model	2013-01～03	2013-04～06	2013-07～09	2013-10～12	2014-01～03	2014-04～06	2014-07～09	2014-10～12	2015-01～03	2015-04～06	2015-07～09	2015-10～12
DTM	4	4	15	4	8	15	10	5	4	5	7	6
HDP	19	23	25	19	19	17	25	24	20	9	19	25

表 2 是 DTM 与 HDP 在各个时间窗下生成的最优主题数。从表 1 和表 2 可以看出，DTM 产生的最优主题数相对 HDP 较少，但主题连贯性是相对较高的。而 HDP 的优点是其生成的主题较多，覆盖范围广，生成能力较强。

表 3 和表 4 分别是 HDP 和 DTM 在 2013-01～03 时间窗下生成的主题，这组实验是为了验证 DTM 模型在本实验数据集下的主题一致性比 HDP 更强。表中每行代表一个主题，每列是该主题相关的关键词。

表 3　HDP 在 2013-01～03 时间段下生成的主题

1	said	china	people	police	chinese	attack	region	govemment	group	terrorist
2	yuan	yang	reward	drug	tourist	tuesday	terror	explosive	paid	extremism
3	reserve	site	biosphere	area	specie	forest	world	island	take	marine
4	post	explosion	comment	consistently	washington	train	thought	railway	provoking	editor
5	full	event	kenya	story	nairobi	former	harmony	commission	governor	nigerian
6	house	billion	yuan	million	bulid	rural	subsidized	household	grain	ukraine
7	chine	vous	dans	chinois	mais	cette	pour	sont	avec	empire
8	pakistan	china	country	chinese	question	afghanistan	india	sharif	said	taliban

续表

9	globe	content	bangkok	digital	full	leam	mail	post	main	shave
10	largely	telegraph	concerned	addresed	freedom	unprecedented	industry	excluded	immigrant	trouble
11	read	afternoon	thing	today	started	independently	focus	ordered	disputed	raided
12	japan	kong	japanese	tokyo	hong	island	monitoring	mainland	land	real
13	bono	kramer	levin	year	court	work	case	voter	individual	firm
14	sport	draw	announced	video	lead	comment	festival	success	tech	candidate
15	march	dissent	levy	rachel	student	worker	commission	church	congregation	blank
16	vietnamese	shootout	immigrant	attention	prepared	plan	month	cross	crossing	guard
17	della	alla	sono	stato	paese	nella	hanno	dell	dove	iraq
18	free	brown	time	agent	game	team	four	said	point	wednesday
19	cheimical	philippine	iraqi	northern	hussein	islamic	based	save	forced	thru

从对比中可以看出 HDP 生成的主题不仅包含了 DTM 生成的主题，还涉及了一些与暴恐不太相关的主题。如表 3 中第 14 行对 festival 和 sport 的描述，第 18 行对 game 的描述等。并且相比 DTM 模型，其主题关键词描述的主题语义并不是很明显，主题连贯性大部分较低。所以，DTM 模型更适用于本实验数据。

3.2.2 所有时间窗下的动态主题分析

下面是关于 DTM 模型生成的动态主题的实验结果分析，主要对动态主题在时间上的演化，及话题热度的变化和其关键词云做了分析。首先给出了 DTM 模型生成的主题连贯性较大的前 10 个主题，如表 5 所示。表 5 中每行代表一个动态主题，以及该主题的简短描述和前 10 个关键词。从中可见新疆暴恐语料在整个时间序列上生成的主要主题。

表 4 DTM 在 2013-01～03 时间段下生成的主题

1	police	people	region	china	clash	violence	riot	govemment	unrest	chinese
2	attack	terrorist	people	china	chinese	terror	stability	beijing	meng	terrorism
3	india	pakistan	afghanistan	china	taliban	country	chinese	world	foreign	news
4	remark	rioter	china	terrorism	police	township	spokesman	hope	department	injured

DTM 模型生成的最优动态主题数是 10，该主题数下主题连贯性最大。从表 5 中可见主题涉及了种族、恐怖活动、移民、斋月、偷渡等。例如，表 5 中第二行主要描述恐怖活动，其关键词主要包括 attack、police、killed、terrorist 等。表 5 中第四行是描述难民的，尤其指最近的泰国移民事件，其关键词主要包括 thailand、turkey、refugee、immigration 等。

表 5 动态主题对应的前 10 个单词

topic	short label	w1	w2	w3	w4	w5	w6	w7	w8	w9	w10	topic coherence
1	ethnic	china	chinese	terrorism	country	region	people	president	govern-ment	central	ethnic	0.169 411
2	terrorist	police	attack	killed	people	terrorist	station	dead	knife	incident	suspect	0.229 545
3	milltant	pakistan	afghani-stan	china	india	taliban	afghan	pakistanl	militant	country	economic	0.276 93

续表

topic	short label	w1	w2	w3	w4	w5	w6	w7	w8	w9	w10	topic coherence
4	refugee	thailand	thai	turkey	refugee	turkish	bangkok	australia	china	immigra-tion	asylum	0.209 805
5	thomson militant	reuters	news	thomson	china	people	business	beijing	minute	world	militant	0.106 113
6	guantanamo detaine	guantan-am	detainee	transfer	prisoner	release	obama	prison	united	facility	detention	0.180 19
7	toumament	team	group	final	network	uber	thomas	indonesia	china	draw	tourna-ment	0.091 525
8	media	youtube	facebook	china	share	twitter	follow	world	cctv	video	subscribe	0.163 769
9	ramadan banned	ramadan	fasting	muslim	religious	china	chinese	govem-ment	student	party	banned	0.176 43
10	separatism sentence	court	china	sentence	lawyer	right	scholar	separa-tism	sentenced	chinese	beijing	0.132 311

图 3 是前 4 个动态主题在整个时间序列上的演化，图 4 是描述这 4 个动态主题的关键词。结合图 3 和图 4，在下文进行了详细的分析。

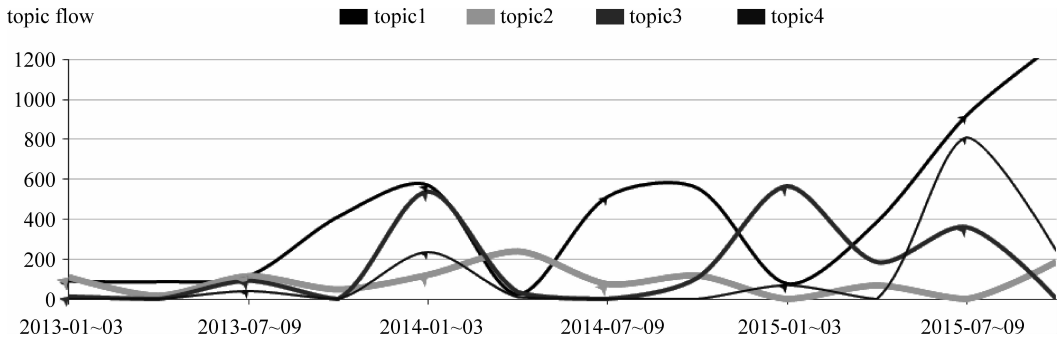

图 3　前 4 个动态主题的演化发展

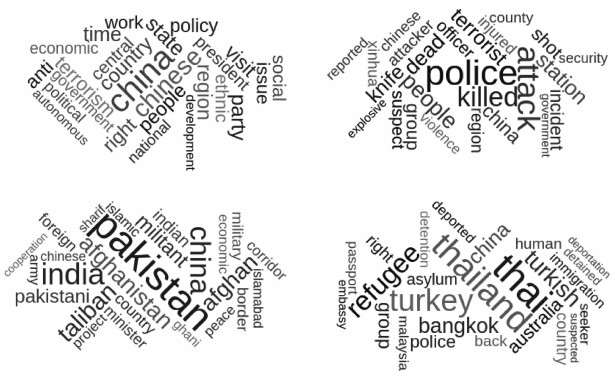

图 4　前 4 个动态主题的关键词云

主题 1 主要跟政府、人权和种族相关。这也说明“疆独”势力是影响中国及新疆社会稳定和安全的重要因素。从图 2 中可以看出它在各个时间段内占的比例都比较大，也说明该主题一直是公众舆论的焦点。尤其是 15 年以来，随着政府加大了对恐怖分子的打击力度，该话题热度也呈现上升的趋势。

主题 2 主要跟暴力事件相关。从图中可以看出它的热度和事件发生的时期基本吻合。例如，2013 年 6 月新疆吐鲁番地区鄯善县袭警事件，2014 年 3 月昆明火车站严重暴恐事件及 5 月份在乌鲁木齐发生的 2 起爆炸案，2015 年 9 月新疆莎车县爆炸案等。发生的时间大部分集中在一年里的正月、5 月及 9 月，也就是穆斯林的斋月期间。这说明斋月期是个敏感时期，政府需在此期间加强社会的保卫工作。

主题 3 主要跟印度、巴基斯坦和阿富汗有关。主题 4 主要跟泰国、土耳其和叙利亚有关。主题 3 和 4 在演化的分布以及关键词云的描述上都十分相似，都反映的是国际势力因素。尤其在 2015 年 7 月发生的泰国向中国遣返非法移民事件表现最为突出。这说明近年来新疆暴恐事件的发生与国外恐怖势力有一定的联系。而且从叙利亚事件上也说明，中国籍

极端分子也参与其中,表明境内的暴恐分子与国际恐怖势力已经合流。所以,打击暴恐犯罪活动就需要切断“疆独”势力与国际恐怖主义之间的联系。

3.2.3 动态主题关键词的演变分析

这组实验是对动态主题随时间推移其关键词的分布变化进行分析。表6和表7是上述动态主题1和2在各个时间窗口下,其关键词分布的变化。表中第一列代表时间,随着序号递增而推移,第二列代表不同时间下描述该主题的关键词。

表6 动态主题1的关键词变化

1	people attack china
2	family land political bejing tibet authority policy sholar minister asian tibetan party government ethnic
3	contry pakistan chinese
4	people region religious ethnic government terrorism chinese
5	public young bank philippine work anti example user central network national stability facility local private
6	crowd city student militant tear state police syria country islamic democracy terrorism protest iraq pakistan
7	thailand kong hong chinese
8	ethnic authority people chinese
9	city autonomous people government rgion cooperation minister chinese turkish president visit china
10	islamic beijing human turkey russian right russia party syrian chinese syria

表7 动态主题2的关键词变化

1	terrorism china people terrorist rioter remark region attack
2	terror police township china people region attack
3	tourist forbidden monday chinese notice people hotel attacker incident attack county attacked beijing knife violence
4	terrorist people attack
5	xinhua terrorist attacker police people knife station
6	authority explosive resident suffered training cousin armed group attacker region majority explosion attack injury
7	injured station china terrorist civilian dead people killed police
8	police killed people death
9	people cinese police
10	checkpoint killed officer attack
11	coat suspect killed mine police

从表6和表7可以看出,主题关键词围绕着事件而变化。例如,表6是关于种族人权主题的描述,从国内的西藏、新疆宗教问题,衍生到巴基斯坦、伊拉克等国际问题上。也说明了国内暴恐活动与国际恐怖势力存在着一定的联系。表7是关于暴恐主题的描述,从北京天安门袭击事件(表中第3行)到昆明火车站袭击事件(表中第5行),以及最近的加油站袭击事件(表中第10行)等。其关键词的变化,也反映了暴恐分子的行动特征。例如,暴恐手段以传统的刀砍、车碾、纵火为主。但最近的枪击爆炸说明暴恐分子对“热兵器”使用频率的上升,对抗性和危害性也随着在增加。所以,政府也应加强对枪支、火药等的管理。

4 结束语

本文针对新疆暴恐事件的舆情进行分析,结合数据自身跟暴恐相关的这一特点,对基于NMF的DTM主题模型进行了改进,实现对新疆暴恐舆情展开了详细的分析,从多个角度分析主题演变的现象和规律;通过与HDP实验方法的比较,可以看出该模型适用于本文所涉及的数据集。再通过可视化

分析，总结了关于暴恐事件的一些现象和规律，对把握舆论动向提供了一定的借鉴意义。在未来研究工作中，会尝试一些复杂的模型，从而发现更多的现象和规律；对于可视化方面，也会尝试更为丰富的可视化数据分析手段。

参考文献

[1] Wang X, McCallum A. Topics Over Time: A Non-Markov Continuous-Time Model of Topical Trends [C]//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2006: 424-433.

[2] Ding W, Chen C. Dynamic Topic Detection and Tracking: A Comparison of HDP, C-word, and Cocitation Methods [J]. Journal of the Association for Information Science and Technology, 2014, DOI: 10.1002/asi.23134.

[3] Ahmed A, Xing E P. Dynamic Non-Parametric Mixture Models and the Recurrent Chinese Restaurant Process: With Applications to Evolutionary Clustering [C]//Proceedings of the SIAM International Conference on Data Mining, Atlanta, Georgia, USA, 2008: 219-230.

[4] Ahmed A, Xing E P. Timeline: A Dynamic Hierarchical Dirichlet Process Model for Recovering Birth/Death and Evolution of Topics in Text Stream [C]//Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence. AUAI Press, 2010.

[5] Teh Y W, Jordan M I, Beal M J, et al. Hierarchical Dirichlet Processes [J]. Journal of the American Statistical Association, 2004, 101(476): 1566-1581.

[6] 方莹，黄河燕，辛欣，等. 面向动态主题数的话题演化分析[J]. 中文信息学报，2014，28(3)：142-149.

[7] Wang C, Paisley J W, Blei D M. Online Variational Inference for the Hierarchical Dirichlet Process [C]//Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, 2011: 752-760.

[8] Lee D D, Seung H S. Learning the parts of objects by non-negative matrix factorization [J]. Nature, 1999, 401: 91-788.

[9] O'Callaghan D, Greene D, Carthy J, et al. An analysis of the coherence of descriptors in topic modeling [J]. Expert Systems with Applications An International Journal, 2015, 42(13): 5645-5657.

[10] Wang Q, Cao Z, Xu J, Li H. Group matrix factorization for scalable topic modeling[C]//Proceedings of the 35th SIGIR Conf. on Research and Development in Information Retrieval, ACM, 2012: 375-384.

[11] 肖永磊，刘盛华，刘悦，等. 社会媒体短文本内容的语义概念关联和扩展[J]. 中文信息学报，2014，28(4)：21-28.

[12] Saha A and Sindhwani V. Learning evolving and emerging topics in social media: A dynamic NMF approach with temporal regularization[C]//Proceedings of the 5th ACM Int. Conf. Web search and data mining, 2012: 693-702 .

[13] Greene, Derek, and James P. Cross. Unveiling the Political Agenda of the European Parliament Plenary: A Topical Analysis [C]//Proceedings of the ACM Web Science Conference. ACM, 2015.

[14] 戴继诚. 当前新疆暴恐活动新变化探析[J]. 科学与无神论，2016(1)：29-34.

[15] 王定，吴绍忠. 去"极端化"背景下的新疆反暴恐情报体系研究[J]. 情报杂志，2016，35(4)：21-26.

张绍武(1967—)，博士，副教授，研究生导师，主要研究领域为搜索引擎、文本挖掘、情感计算等。
E-mail: zhangsw@dlut.edu.cn

林鸿飞(1962—)，博士，教授，博士生导师，主要研究领域为信息检索，社会计算，情感分析，文本挖掘等。
E-mail: hflin@dlut.edu.cn

邵华(1991—)，硕士，主要研究领域为文本挖掘、情感计算。
E-mail: shdut0901@163.com