Computer Science and Application 计算机科学与应用, 2018, 8(6), 976-985
Published Online June 2018 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2018.86109

Hans汉斯

Analysis of Non-Parametric Event Evolution in Social Networks

Yingying Li

Beihang University, Beijing
Email: liyy@act.buaa.edu.cn

Received: Jun. 7th, 2018; accepted: Jun. 22nd, 2018; published: Jun. 29th, 2018

Abstract

Social networks, such as Weibo and Twitter, have become important platforms where billions of individuals follow events. People not only concern about what happened, but also pay more attention to how the event gradually progresses. Therefore, it is crucial to monitor the development of events in social networks. There are some research mining event evolutions in news articles and short texts (texts in social networks). Methods designed for news articles cannot be directly applied to social network due to the fact that short texts are more ill-formed and shorter than news articles. Some methods that apply to short texts do not take into account semantic information, and some cannot discover evolution of long-term spanning events, especially intermittent events. In light of this, we propose a non-parametric method to discover event evolution (storylines). Firstly, a bayesian model is used to measure semantic correlation of short texts. Secondly, an embedded representation-based algorithm is used to generate storylines for long-term and short-term events. We further use dirichlet process to automatically learn an appropriate number of topics. In comparison with other methods, detailed experimental results on three manually labeled data sets demonstrate the effectiveness of our method.

Keywords

Event Evolution, Social Network, Probabilistic Graphical Model

社会网络中非参数化的事件演化分析

李莹莹

北京航空航天大学，北京
Email: liyy@act.buaa.edu.cn

收稿日期：2018年6月7日；录用日期：2018年6月22日；发布日期：2018年6月29日

文章引用：李莹莹. 社会网络中非参数化的事件演化分析[J]. 计算机科学与应用, 2018, 8(6): 976-985.
DOI: 10.12677/csa.2018.86109

摘　要

社会网络，如微博和Twitter，已经成为数十亿人关注事件的重要平台。人们不仅关注所发生的事情，更关注事件的演化。因此，监控社交网络中事件的发展是至关重要的。在新闻文章和短文本(社交网络中的文本)中有一些挖掘事件演化的研究。由于短文的形式比新闻文章短，适用于新闻文章的方法不能直接应用于社交网络。一些应用于短文本的方法不考虑语义信息，有些方法无法发现长期跨度事件的演化，特别是中间有间断的事件。鉴于此，我们提出了一种非参数的方法来发现事件演化(故事情节)。首先使用贝叶斯模型测量短文本的语义相关性。其次，使用基于嵌入表示的算法来生成长期和短期事件的故事线。我们进一步使用Dirichlet过程自动学习适当数量的主题。与其他方法相比，三个人工标记数据集的详细实验结果证明了我们方法的有效性。

关键词

事件演化，社会网络，概率图模型

1. 引言

社会网络已经成为人们分享和传播新闻的主流平台。社会网络每天都会产生大量的用户产生内容(User-Generated Contents)。对于 UGC 流，人们很容易感到不知所措。因此迫切需要从社会网络自动生成事件的演化。

目前有挖掘事件演化的相关研究。这些研究可大致分成两类。一类是基于词相似度的方法。Pei 等人[1]提取用基于子图的增量追踪框架追踪事件演化。该方法用 Jaccard 相似度作为文本相似度的度量方式。词相似度基于文本重合的词越多，文本越相似的假设。这种假设在现实中有可能不成立。很多时候文本的相似度依赖于潜在的语义关联，而不是表面的词重合度。为弥补词相似度的不足，另一类基于隐式语义相似度的方法被提出。Kalyanam 等人[2]用非负矩阵分解(NMF)捕捉事件随着时间的变化。该方法只能发现相邻时刻的事件的演化，不能发现长期跨度的事件演化。Zhou 等人[3]提出一个无监督的贝叶斯模型(DSDM)提取故事情节的结构化表示和演化模式。DSDM 和 NMF 捕捉文档级别的词共线模式。因为短文本中文档级别词共线的稀疏性，这些方法不能直接用于短文本中。

为解决上述方法存在的不足。我们提出一个非参数的方法构造社会网络中的事件演化。首先，我们从微博中检测子事件。然后，我们提出一个非参数的方法提取子事件的隐式语义信息。最后我们基于子事件的隐式语义信息生成事件演化。本文的主要贡献如下所示：

- 我们提出一个非参数化的概念图模型，Biterm Topic Model with Dirichlet Process (BTMDP)，提取子事件的语义信息。
- 我们提出基于子事件语义信息的故事情节生成算法(LineGen)。
- 通过在三个数据集上的实验证明我们方法较已有方法的有效性。

本文的其余部分组织如下。首先，我们介绍相关工作。然后，我们介绍我们的方法。随之，我们展

示实验与结果。最后，我们总结提出的方法并展望未来的工作。

2. 相关工作

我们的工作与三方面的研究相关。1) 主题检测与追踪(TDT)。TDT 旨在基于主题对文本分组，检测异常的和以前未报道的事件，和追踪某主题下事件的发展[4]。2) 时间线(timeline)生成。时间线是一种可使分析任务更简单和更快速的可视化技术[5]。3) 故事情节(storyline)生成。故事情节生成旨在提取特定新闻主题下的事件并展示事件随着时间的演化过程[6]。

故事情节生成相关的研究大致可分为两类：基于新闻文本和基于社会网络文本。1) 基于新闻文本。Wang 等人根据用户给定的关键词形式化故事情节为最小权重连通适配集问题[4]。这类方法的结果严重依赖于用给的查询词。ASG [7]、MEA [8]和 CHARCOAL [9]基于概率图模型在新闻文本上生成故事情节。新闻社会网络文本的口语化、错别字和短文本等特性使得基于新闻文本的方法直接用于社会网络文本可能得不到理想的效果。2) 基于社会网络文本。基于社会网络文本的方法大制可分为两种：基于词相似度的方法和基于隐式语义信息的方法。Lin 等人基于文本的余弦相似度用图优化的方法生成故事情节[10]。词相似度不足以发现子事件的相关性。LTECS [11]基于非负矩阵分解(NMF)学习相邻时刻的事件演化。Lee 等人[12]基于 LDA 的 KL 散度发现事件间的关联关系。由于社会网络文本的稀疏性，捕捉文档级别词花线模式的模型，LDA 和 NMF，不能直接用于社会网络中。

3. 论文方法

基于已有算法的不足，我们提出一种挖掘事件演化的算法。社会网络文本具有不规范、文本较短等特性。单个微博可能只包含事件的部分信息。我们用已有的成熟的子事件检测算法检测能表示完成信息的子事件。事件间存在演化关系，文本间的词相似度不能有效挖掘演化关系。我们用主题模型挖掘子事件内在的主题结构。依据子事件的主题结构信息有效挖掘子事件的演化关系。

本节我们首先介绍挖掘子事件内在主题结构的融合狄利克雷过程的词对主题模型。然后，我们介绍基于子事件主题结构挖掘子事件演化关系的过程，即故事情节生成过程。

3.1. 融合狄利克雷过程的词对主题模型(BTMDP)

两个文本之间的相似性不仅取决于词共线程度，还取决于语义关联。主题模型是一种挖掘语义关联的方式。但传统的主题模型(如 LDA 和 PLSA)遭受严重的数据稀疏短文本[13]。BTMDP 在语料库级别建模词共现模式，因此 BTM 适用于社会网络的短文本。为了方便推导，我们先解释 BTMDP 使用到的符号，如表 1 所示。

其中，词对是由两个词构成的组合(w_1, w_2)。假设一个文章的词集合为{计算机，应用，发展}，该文章的词对集合为{(计算机，应用)，(计算机，发展)，(应用，发展)}。

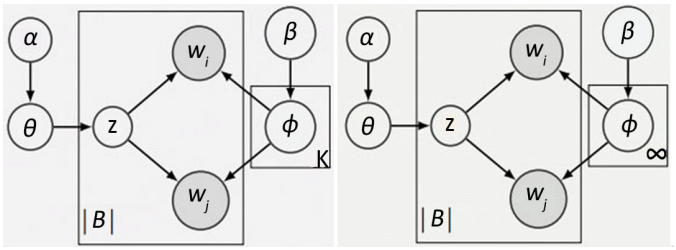
3.1.1. BTMDP 的概率图模型和生成过程

Cheng 等人[14]提出一个对整个语料库上词共现模式的生成过程建模的词对主题模型(BTM)。BTM 需要提前指定主题数。BTM 的时间复杂度为 O ($K|B|$)。主题数 K 越大，模型需要的时间越多。主题数越小，主题的粒度越粗。最好的方式是模型自动学习合适的主题数。基于这个想法，我们提出一个融合狄利克雷过程的词对主题模型(BTMDP)。因为狄利克雷过程的特性，主题数不需要人工干预即可自动学习。BTMDP 可以看成是 BTM 的无限扩展模型。BTM 和 BTMDP 的概率图模型如图 1 所示。

当主题数 K 成为无穷时，BTM 成为 BTMDP。BTMDP 的生成过程如下所示：

Table 1. Notations of symbols
表 1. 符号的注释

符号	注释
B	语料库中所有词对
$B_{\neg b}$	语料库除去词对 b 的所有词对
$B_{z,\neg b}$	主题 z 的除去词对 b 的所有词对
Z	各词对的主题值
$Z_{\neg b}$	除去词对 b 的各词对的主题值
M	词库中词的数量
n_z	主题 z 的词对数
$n_{w\|z}$	主题 z 下词 w 的出现次数
n_t	BTMDP 学到的主题数



(a) PGM of BTM (b) PGM of BTMDP

Figure 1. PGM of BTM and BTMDP
图 1. BTM 和 BTMDP 的概率图模型

$$\theta \mid \alpha \sim GEM\left(1,\alpha\right) \tag{1}$$

$$z_b \mid \theta \sim Multi\left(\theta\right) \tag{2}$$

$$\phi_k \mid \beta \sim Dir\left(\beta\right) \tag{3}$$

$$b \mid z_b, \{\phi\}_{k=1}^{\infty} \sim p\left(b \mid \phi_{z_b}\right) \tag{4}$$

其中“$X \sim S$”表示 X 服从于 S 的分布。$Multi()$表示多项分布。$Dir()$表示狄利克雷分布。

3.1.2. BTMDP 的推断过程

b 是由两个词(w_i, w_j)组成的词对。$p\left(b \mid \phi_{z_b}\right) = p\left(w_i \mid \phi_{z_b}\right) p\left(w_j \mid \phi_{z_b}\right)$。与 BTM 相同，BTMDP 没有直接对文章的生成过程建模。在主题学习过程我们不能直接推断文章的主题分布。Yan 等人[13]认为文章的主题分布等于文章生成的词对的主题分布的期望。即：

$$P\left(z \mid d\right) = \sum_b P\left(z \mid b\right) P\left(b \mid d\right) \tag{5}$$

$$P\left(z \mid b\right) = \frac{P\left(z\right) P\left(w_i \mid z\right) P\left(w_j \mid z\right)}{\sum_z P\left(z\right) P\left(w_i \mid z\right) P\left(w_j \mid z\right)} \tag{6}$$

在 BTMDP 中不能直接推断。我们采用吉布斯抽样近似推断。在讨论 BTMDP 的抽样算法前，我们先了解一个 BTM 的抽样过程。BTM 中，词对 b 选择一主题的概率如式(7)所示。

$$p\left(z_b=z\mid Z_{\neg b},B,\alpha,\beta\right)\propto p\left(z_b=z\mid Z_{\neg b},B_{\neg b},\alpha,\beta\right)p\left(b\mid z_b=z,Z_{\neg b},B_{\neg b},\alpha,\beta\right)$$

$$\propto p\left(z_b=z\mid Z_{\neg b},\alpha\right)p\left(b\mid z_b=z,B_{z,\neg b},\beta\right)\propto\frac{n_z+\alpha/K}{|B|-1+\alpha}\frac{\left(n_{w_i|z}+\beta\right)\left(n_{w_j|z}+\beta\right)}{\left(\sum_w n_{w|z}+M\beta\right)\left(\sum_w n_{w|z}+M\beta+1\right)} \tag{7}$$

$$p\left(z_b=z\mid Z_{\neg b},B,\alpha,\beta\right)\propto\frac{n_z}{|B|-1+\alpha}\frac{\left(n_{w_i|z}+\beta\right)\left(n_{w_j|z}+\beta\right)}{\left(\sum_w n_{w|z}+M\beta\right)\left(\sum_w n_{w|z}+M\beta+1\right)} \tag{8}$$

$$p\left(z_b=K+1\mid Z_{\neg b},B,\alpha,\beta\right)\propto p\left(z_b=K+1\mid Z_{\neg b},\alpha\right)p\left(b\mid z_b=K+1,\beta\right)$$

$$\propto\left(1-\sum_{k=1}^{K}p\left(z_b=k\mid Z_{\neg b},\alpha\right)\right)p\left(b\mid z_b=K+1,\beta\right)\propto\frac{\alpha}{|B|-1+\alpha}\frac{\beta^2}{\left(M\beta\right)\left(M\beta+1\right)} \tag{9}$$

当 K 趋于无穷时，词对 b 选择一个存在的主题的概率如式(8)所示。词对 b 选择一个新主题的概率如式(9)所示。根据词和词对的主题分布。我们可根据式(10)和式(11)评估主题的词分布和主题分布。

$$p\left(w\mid z\right)=\frac{n_{w|z}+\beta}{\sum_w n_{w|z}+M\beta} \tag{10}$$

$$p\left(\theta_z\right)=\frac{n_z+\alpha/n_t}{|B|+\alpha} \tag{11}$$

BTMDP 的吉布斯抽样过程如算法 1 所示。因为空间限制，我们略去了吉布斯抽样的详细推导。原理和推导过程请参考[13] [14]。

算法 1BTMDP 的抽样过程

输入：超参数 α 和 β，词对集合 B;

输出：主题的词分布 θ 和主题分布 ϕ.

//初始化

所有的计数变量($n_t,n_z,n_{w|z},B_z$)置为零
for each $b\in B$ do
$z_b=random\left(n_t+1\right)$;
If is_a_new_topic (z_b) then
Create_a_new_topic();
$n_t=n_t+1$;
End if
$B_z=B_z\cup\{b\}$;
$n_z=n_z+1$;
for each $word\in b$ do
$n_{w|z}=n_{w|z}+1$;
End for
End for
For each *iteration* do
For each $b\in B$ do
$z_{b\_old}=get\_topic\left(b\right)$;
$B_{z_{b\_old}}=B_{z_{b\_old}}-\{b\}$;
$n_{z_{b\_old}}=n_{z_{b\_old}}-1$;
For each $word\in b$ do
$n_{w|z_{b\_old}}=n_{w|z_{b\_old}}-1$;
End for
If $n_{z_{b\_old}}=0$ then
remove_topic(z_{old});
$n_t=n_t-1$;
End if
compute the probability of b belong to topic(existed topics and a new topic) with(8) and (9);
sample z_{b_new} with the topic probability computed;
If is_a_new_topic (z_{b_new}) then
Create_a_new_topic();
$n_t=n_t+1$;
End if
$B_{z_{b\_new}}=B_{z_{b\_new}}\cup\{b\}$;
$n_{z_{b\_new}}=n_{z_{b\_new}}+1$;
for each $word\in b$ do
$n_{w|z_{b\_new}}=n_{w|z_{b\_new}}+1$;
End for
End for
End for
Compute φ and θ with (10) and (11)

3.2. 故事情节生成(LineGen)

观察发现，相同故事情节下，一些子事件的相似度更高。如表 2 所示，子事件 se1 和 se2 关于美国总统候选人选择竞选搭档。子事件 se3 和 se4 是关于总统候选人的支持率。首先，我们用 *pline*（定义 1）表示相似度更高的子事件集合。然后，我们基于 DBSCAN 的思想将 *pline* 聚成故事情节。

定义 1：*pline* 是一个子事件集合。*Pline* 中所有子事件有相同的 top one 维度。我们用二元组<*centroid*, *selist*>表示。其中(1) *centroid* 是 pline 的中心，(2) *selist* 是 *pline* 的子事件集合。

$$centroid = \left(cen_1, cen_2, \cdots, cen_K\right)$$
$$cen_i = \frac{\sum_{s \in selist} s_i}{\sum_{m=1}^{K} cen_m} \tag{12}$$

高维数据经常包含大量噪音特征，这不利于数据处理[15]。为了降低噪音和后序计算需要的时间。我们为每个子事件提取 top-k 个特征。假设我们子事件的嵌入表示有 K 维。t_{se} 表示子事件的 top-k 个特征。子事件的表示如式(13)所示。

$$s_i = \begin{cases} \dfrac{s_i}{\sum_{j \in t_{se}} s_j} & i \in t_{se} \\ 0 & \text{others} \end{cases} \tag{13}$$

其中，子事件由前 top-k 个维度的特征表示。除前 top-k 维度外，其它维度的值为 0。例，K 为 4 的一个子事件的嵌入表示为 embed_representation (*subevent*) = (0.2, 0.3, 0.4, 0.1)，如果 topk = 2，则 representation (*subevent*) = (0,3/7,4/7,0)。

4. 实验和评价

针对上一节中提出的方法，我们在本节进行实验验证。首先，我们介绍数据集。然后我们评价故事情节生成的性能，并展示我们提出的方法较于已有方法的优势。我们的算法包括两步，子事件检测和故事情节生成。有大量关于子事件检测的相关文献。我们从中选择一个成熟的子事件检测算法。因此我们不再评价子事件检测的性能。

4.1. 数据集和注解

数据集是依据新浪微博用子事件检测算法检测的子事件。子事件用 6 元组<*时间*、*地点*、*参与者*、*核心词*、*描述词*，*描述*>表示。我们请志愿者标注具有演化关系的子事件，即标注子事件所属的故事情节，具有演化关系的子事件为一个故事情节。我们用两种方式标注数据。两种方式的不同在于选择子事件数据集的方式。方式一为将某时间段内所有子事件作为标注使用的子事件数据集。方式二是与关键词相关的子事件作为子事件数据集。依据两种标注方式我们构造了三个数据集，基于方式一构造一个故事情节集，基于方式二构造两个故事情节集。DS11 表示基于方式一从时间段 2016 年 6 月到 2016 年 8 月标注的故事情节集，该故事情节集包含 416 个子事件。DS21 和 DS22 表示基于方式二构造的两个故事情节集。DS21 是基于关键词丝绸之路标注的故事情节集，该故事情节集包含 97 个子事件。DS21 是基于关键词萨德和亲信干政门标注的故事情节集，该故事情节集包含 85 个子事件。标注统计如表 3 所示。

4.2. 实验结果与分析

在本小节，我们首先介绍使用的评价指标。然后，我们介绍所有的对比方法。最后，我们呈现实验

Table 2. The example that some subevents are closer than others in a storyline
表 2. 相同故事情节下，一些子事件相似度更高的例子

Se_id	Description
*se*1	希拉里选择弗吉尼亚州联邦参议员蒂姆·凯恩（Tim Kaine）为副总统竞选搭档
*se*2	特朗普宣布彭斯为竞选搭档
*se*3	从希拉里和特朗普的民意调查来看，很难预测谁是赢
*se*4	美大选希拉里支持率全面领先，特朗普被批有偏见

Table 3. Annotation statistics
表 3. 标注统计

方式	故事情节集名称	故事情节集描述
时间段	DS11	从 2016 年 6 月到 2016 年 8 月标注的 416 个子事件
关键词	DS21	基于关键词丝绸之路标注的 97 个子事件
	DS22	基于关键词萨德和亲信干政门标注的 85 个子事件

结果并展示我们方法的优势。

4.2.1. 评价指标

在三个故事情节集中，子事件标注有故事情节。真实的故事情节集称为 $C\left(\left|C\right|=L\right)$。基于算法生成的故事情节集成为 $M\left(\left|M\right|=T\right)$。一般来讲，$T\neq M$。我们的方法可生产任意数量的故事情节。依据真实的故事情节集 C 和生成故事情节集 M，我们构造一个混乱矩阵(confusion matrix) CM。$CM\left(l,t\right)$ 是真实故事情节 C_l 和生成故事情节 M_t 交集的基，即 $CM\left(l,t\right)=\left|C_l\cap M_t\right|$。对任意标注故事情节 $C_l\in C$，我们用 $top_k\left(C_l\right)$ 表示生成故事情节集中与其交集最大的 top-k 个故事情节。我们基于式(14)计算准确率和召回率。

$$Precision_K=\frac{\sum_{l=1}^{L}\left|C_l\cap top_K\left(C_l\right)\right|}{\sum_{l=1}^{L}\left|top_K\left(C_l\right)\right|}$$
$$Recall_K=\frac{\sum_{l=1}^{L}\left|C_l\cap top_K\left(C_l\right)\right|}{\sum_{l=1}^{L}\left|C_l\right|} \tag{14}$$

其中，C_l 表示一个标注故事情节。$top_k\left(C_l\right)$ 表示生成故事情节集中与标注故事情节 C_l 交集最大的 top-k 个故事情节。|.|表示基，即集合中的元素个数。$\cap$ 表示集合交集。计算准确率和召回率时，我们设置 $K=2$。

4.2.2. 对比方法

我们的方法(BTMDP + LineGen)首先用 BTMDP 学习子事件的主题结构，即为子事件生成主题嵌入表示。然后，我们的方法用 LineGen 生成故事情节。我们的对比方法如下所示：

- DBSCAN：该方法用 DBSCAN 对子事件聚类。一个簇称为一个故事情节。
- NMF + DBSCAN：该方法首先用 NMF 生成子事件的嵌入表示，然后基于嵌入表示用 DBSCAN 对子事件聚类。一个簇称为一个故事情节。
- BTM + DBSCAN：该方法首先用 BTM 生成子事件的主题嵌入表示，然后基于主题嵌入表示用 DBSCAN 对子事件聚类。一个簇称为一个故事情节。
- BTMDP + DBSCAN：该方法首先用 BTMDP 生成子事件的主题嵌入表示，然后基于主题嵌入表示用 DBSCAN 对子事件聚类。一个簇称为一个故事情节。

- NMF + LineGen：该方法首先用 NMF 生成子事件的主题嵌入表示，然后基于主题嵌入表示用 LineGen 生成故事情节。
- LDA + LineGen：该方法首先用 LDA 生成子事件的主题嵌入表示，然后基于主题嵌入表示用 LineGen 生成故事情节。
- BTM + LineGen：该方法首先用 BTM 生成子事件的主题嵌入表示，然后基于主题嵌入表示用 LineGen 生成故事情节。

LDA + DBSCAN 将所有的子事件聚到一个簇，我们展示 LDA + DBSCAN 的效果。在 LDA、BTM 和 BTMDP 中，吉布斯抽样迭代 1000 次。在需要指定主题数的 NMF、LDA 和 BTM 中，我们设定主题数为 250。在 LDA、BTM 和 BTMDP 中，参数 α 和 β 分别为 500 和 0.02。在 NMF + DBSCAN、BTM + DBSCAN 和 BTMDP + DBSCAN 中，我们使用的距离函数如式(15)所示。在 NMF + LineGen、BTM + LineGen 和 BTMDP + LineGen 中，我们使用的距离函数如式(16)所示。在 LineGen 中我们设置参数 top-k 为 5。所有的结果是十次结果的平均值。

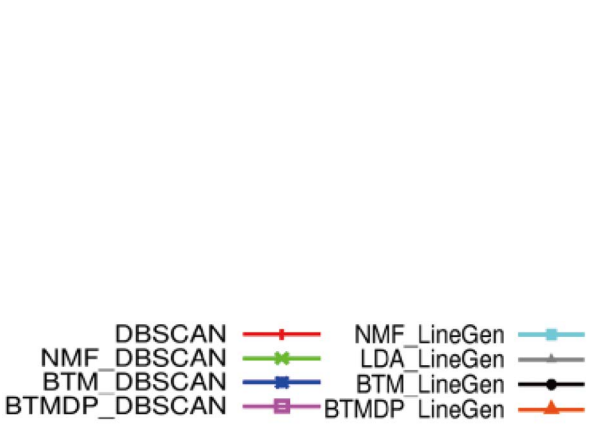
$$dis(se_1, se_2) = 1 - \cos(se_1, se_2) \tag{15}$$

$$dis(p_1, p_2) = 1 - 0.7 * Jaccard(p_1, p_2) - 0.3 * \cos(p_1, p_2) \tag{16}$$

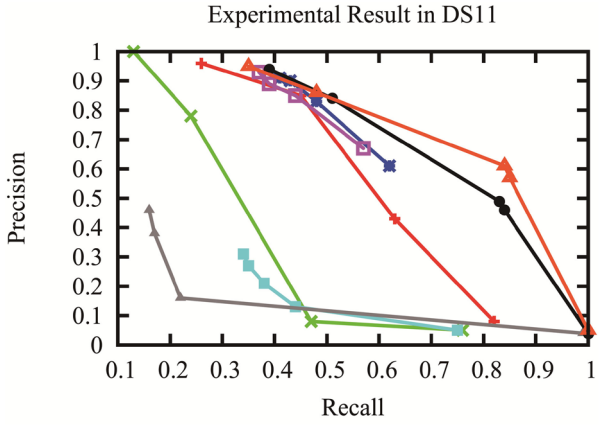
4.2.3. 结果与分析

三个故事情节集，DS11、DS21 和 DS22，的准确率和召回率如图 2 所示。在三个数据集上，BTMDP + LineGen 和 BTM + LineGen 的准确率和召回率的曲线在 BTMDP + DBSCAN 和 BTM + DBSCAN 之上。这说明 LineGen 方法的有效性。LineGen 使用子事件的 top-k 个特征减小子事件特征的噪音和后序计算需要的时间。LineGen 使用故事情节中子事件的特性，一些子事件的相似度更高，来提高准确率和召回率。BTMDP + LineGen 比 BTM + LineGen 有更高或相等的准确率和召回率。BTMDP 可自适应的学习合适的主题数，避免主题数过大或过小。BTMDP + LineGen 和 BTM + LineGen 的准确率和召回率的曲线在 DBSCAN 之上，这说明子事件的主题结构有利于事件演化的挖掘。NMF + DBSCAN，NMF + LineGen and LDA + LineGen 在三个数据集上比其他方法的效果更差，这说明对文档级别词共现模式建模的主题模型在短文本上得不到理想的效果。BTMDP 是适用于短文本的主题模型。

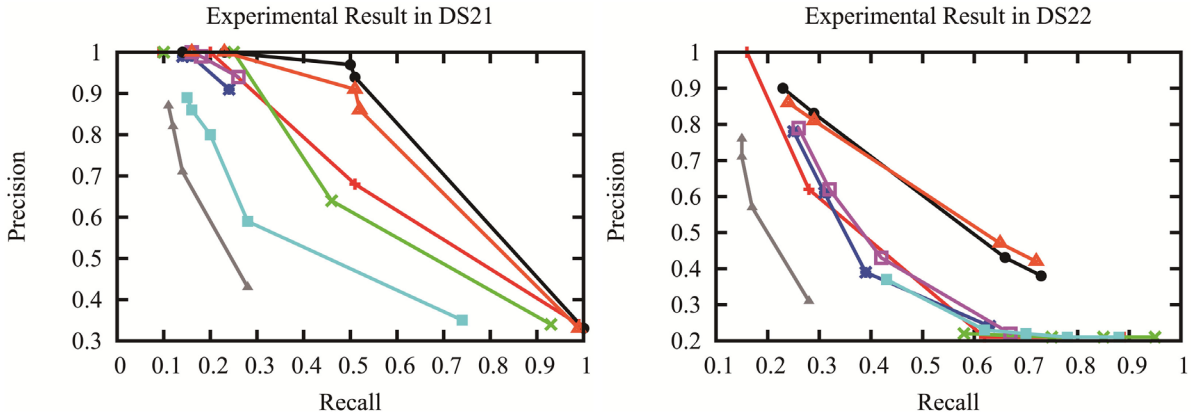
我们也对比了我们的方法在不同的 topk 数下的实验结果，如图 3 所示。top-k 为 5 的性能优于或等于 top-k 为 all 的性能。这说明，选择有区别性的特性对性能的提升有帮助。



(a) DS11、DS21 和 DS22 结果使用的图例

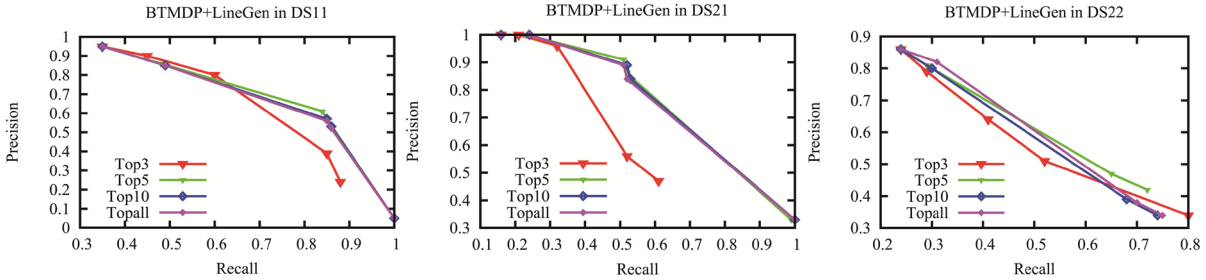


(b) DS11 故事情节集上的实验结果



(c) DS21 故事情节集上的实验结果　(d) DS22 故事情节集上的实验结果

Figure 2. Experimental result in three storyline sets
图 2. 三个故事情节集上的实验结果



(a) DS11 上不同特征数的实验结果　(b) DS21 上不同特征数的实验结果　(c) DS22 上不同特征数的实验结果

Figure 3. Experimental result of BTMDP + LineGen with different top-k numbers
图 3. BTMDP + LineGen 在不同的 top-k 下的实验结果

5. 工作总结与展望

在社会网络挖掘事件演化具有重要意义。我们提出一个非参数化的方法在社会网络上挖掘事件的演化。首先，我们用一个子事件检测算法检测子事件。然后，我们提出非参数化的贝叶斯模型(BTMDP)生成子事件的语义信息。最后，我们用基于子事件嵌入表示的故事情节生成算法(LineGen)生成故事情节。实验结果表明我们的方法比已有的方法有较好的准确率和召回率。考虑到社会网络数据的动态性，我们将设计增量算法在社会网络挖掘事件的演化。我们将设计一个展示系统便于用户对事件演化的理解。

参考文献

[1] Lee, P., Lakshmanan, L.V.S. and Milios, E.E. (2014) Incremental Cluster Evolution Tracking from Highly Dynamic Network Data. *Proceedings of the IEEE 30th International Conference on Data Engineering*, Chicago, IL, 31 March-4 April 2014, 3-14. https://doi.org/10.1109/ICDE.2014.6816635

[2] Janani, K., Sumithra, V., *et al.* (2016) From Event Detection to Storytelling on Microblogs. *Proceedings of the IEEE International Conference on Advances in Social Networks Analysis and Mining*, San Francisco, CA, 18-21 August 2016, 437-442.

[3] Zhou, D., Xu, H. and He, Y. (2015) An Unsupervised Bayesian Modelling Approach for Storyline Detection on News Articles. *Proceedings of the* 2015 *Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, 17-21 September 2015, 1943-1948. https://doi.org/10.18653/v1/D15-1225

[4] Wang, D., Li, T. and Ogihara, M. (2012) Generating Pictorial Storylines via Minimum-Weight Connected Dominating Set Approximation in Multi-View Graphs. Proceedings of the 26th AAAI Conference on Artificial Intelligence, AAAI, Menlo Park, CA, 683-689.

[5] Yong, C., Hui, Z., Wu, J.J., *et al.* (2016) Modeling Emerging, Evolving and Fading Topics Using Dynamic Soft Orthogonal NMF with Sparse Representation. *Proceeding of the* 2015 *IEEE International Conference on Data Mining*, Piscataway, Atlantic City, NJ, 14-17 November 2015, 61-70.

[6] Zhou, D.Y., Xu, H.Y., Dai, X.Y., *et al.* (2016) Unsupervised Storyline Extraction from News Articles. *Proceedings of the* 25*th International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, Burlington, MA, 3014-3021.

[7] Hua, T., Zhang, X.C., Wang, W., Lu, C.-T. and Ramakrishnan, N. (2016) Automatical Storyline Generation with Help from Twitter. *Proceedings of the* 25*th ACM International Conference on Information and Knowledge Management*, Indianapolis, Indiana, 24-28 October 2016, 2383-2388. https://doi.org/10.1145/2983323.2983698

[8] Huang, L.F. and Huang, L.E. (2013) Optimized Event Storyline Generation Based on Mixture Event-Aspect Model. *Proceedings of the* 2013 *Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, 18-21 October 2013, 726-735.

[9] Tang, S.L., Wu, F., Li, S., Lu, W.M., Zhang, Z.F. and Zhuang, Y.T. (2015) Sketch the Storyline with CHARCOAL: A Non-Parametric Approach. *Proceedings of the* 24*th International Joint Conference on Artificial Intelligence*, Buenos Aires, 25-31 July 2015, 3841-3848.

[10] Lin, C., Lin, C., Li, J.X., Wang, D.D., Chen, Y. and Li, T. (2012) Generating Event Storylines from Microblogs. *Proceedings of the* 21*st ACM International Conference on Information and Knowledge Management*, Maui, Hawaii, October 29-November 2 2012, 175-184. https://doi.org/10.1145/2396761.2396787

[11] Kalyanam, J., Mantrach, A., Saez-Trumper, D., Vahabi, H. and Lanckriet, G.R.G. (2015) Leveraging Social Context for Modeling Topic Evolution. *Proceedings of the* 21*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Sydney, NSW, 10-13 August 2015, 517-526. https://doi.org/10.1145/2783258.2783319

[12] Lee, P., Lakshmanan, L.V.S. and Milios, E.E. (2014) CAST: A Context-Aware Storyteller for Streaming Social Content. *Proceedings of the* 23*rd ACM International Conference on Conference on Information and Knowledge Management*, Shanghai, 3-7 November 2014, 789-798. https://doi.org/10.1145/2661829.2661859

[13] Yan, X.H., Guo, J.F., Lan, Y.Y. and Cheng, X.Q. (2013) A Biterm Topic Model for Short Texts. *Proceedings of the* 22*nd International World Wide Web Conference*, Rio de Janeiro, 13-17 May 2013, 1445-1456. https://doi.org/10.1145/2488388.2488514

[14] Cheng, X.Q., Yan, X.H., Lan, Y.Y. and Guo, J.F. (2014) BTM: Topic Modeling over Short Texts. *IEEE Transactions on Knowledge and Data Engineering*, **26**, 2928-2941. https://doi.org/10.1109/TKDE.2014.2313872

[15] Nie, F.P., Zhu, W. and Li, X.L. (2016) Unsupervised Feature Selection with Structured Graph Optimization. *Proceedings of the* 30*th AAAI Conference on Artificial Intelligence*, Phoenix, 12-17 February 2016, 1302-1308.