基于微博的时空事件识别研究

郑喆君 金蓓弘 崔艳玲
(中国科学院软件研究所 北京 100190) (中国科学院大学 北京 100190)

摘 要 微博是一种社交网络服务,它主要基于用户的关注关系进行信息分享和传播,具有时效性强、传播迅速等特点。将微博看成是反映城市动态的一类感知器,从识别微博的主题入手,检测微博中反映的时空事件。为此,首先提出了一种用于分析微博主题的主题模型ST-LDA,并应用该模型将具有语义相似性、时空聚集性的微博归属于同一主题下;然后给出了从主题中检测时空事件的方法。基于真实的新浪微博数据进行实验,结果表明此方法比基于LDA的方法、基于TimeLDA的方法在事件识别上有更高的查全率和查准率。

关键词 微博,时空事件,主题模型
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.10.041

Study on Recognition of Spatial-Temporal Events Based on Microblogs

ZHENG Zhe-jun JIN Bei-hong CUI Yan-ling
(Institute of Software, Chinese Academy of Sciences, Beijing 100190, China)
(University of Chinese Academy of Sciences, Beijing 100190, China)

Abstract As a kind of social networking service, microblog service can share and broadcast information mainly through the microbloggers' followers and features strong timeliness of topics and rapid spread. This paper viewed the microblogs as a kind of event sensor which can perceive the dynamic behaviors in the city, and started with identifying the topics in microblogs, and then detecting the spatial-temporal events in microblogs. This paper presented a topic model named ST-LDA for analyzing the topics in microblogs. Applying this model, the microblogs with similar semantics and close spatial-temporal nature can be classified into a same topic. Then, this paper gave a method of discovering the spatial-temporal events from topics. Experimental results based on the real data from weibo.com show that our method has a higher recall and precision ratio than LDA-based and TimeLDA-based methods.

Keywords Microblogs, Spatial-temporal events, Topic model

1 引言

随着社会经济的快速发展,城市运行、管理和发展面临诸多问题,交通拥堵、疾病传播、环境污染、资源紧缺等事件经常发生。在构建智慧城市的背景下,如何依靠大数据进行城市状态和活动的观测,成为当务之急,它是下一步利用大数据进行城市运行机理分析和城市管理机制决策的前提。

到目前为止,已有一些工作关注利用大数据识别城市中发生的事件。例如,考虑到在城市中出租车是一种重要的交通工具[1],利用出租车GPS轨迹来评估城市范围内的社交活跃度,进而检测出社交事件;而文献[2]利用社交网络来预测用户是否参与城市活动。

我们注意到微博服务可以作为城市事件的一类感知器。微博服务最初是作为一种信息分享和交流的平台出现,其用户可以及时发布简短信息并通过用户之间的关注关系传播信息。随着微博使用的普及和用户数量的增加,现实生活中发生的一些事件会反映在微博数据集中,这使得微博服务正在成为观察城市状态和活动的一个重要平台。因此,通过对大量微博的分析,可以发现现实生活中发生的多种事件(例如:体育赛事、恶劣天气等)。

本文关注如何有效地从大量微博中识别时空事件,从而帮助确定城市所处的状态和所发生的活动。与传统的文本(如报纸、杂志、期刊上的文本等)相比,微博数量大,每条微博的文本内容短但噪声强,微博通常会包含发布时间、发布地点、发布者等额外属性。微博的这些特点给微博的分析和挖掘带来了挑战。本文通过改进LDA(Latent Dirichlet Allocation)模型,提出了一种识别微博主题的主题模型ST-LDA(Spatio-Temporal LDA),该模型对微博数据中的文本属性、时间属性和空间属性进行统一建模,使得隶属于同一主题下的微博数据在具有语义相似性的同时,也具有时间和空间上的聚集性。进而,本文给出了从微博主题中识别时空事件的方法。

到稿日期:2015-08-29 返修日期:2016-03-05 本文受国家自然科学基金(61472408,61372182)资助。
郑喆君(1989-),男,硕士生,主要研究方向为分布式计算、普适计算,E-mail:zhengzhejun13@otcaix.iscas.ac.cn;金蓓弘(1967-),女,研究员,博士生导师,主要研究方向为分布式计算、移动与普适计算、中间件,E-mail:jbh@otcaix.iscas.ac.cn;崔艳玲(1991-),女,硕士生,主要研究方向为分布式计算、普适计算,E-mail:cuiyanling13@otcaix.iscas.ac.cn。

本文第 2 节介绍相关工作；第 3 节介绍提出的 ST-LDA 主题模型；第 4 节给出了从微博主题中识别时空事件的方法；第 5 节是实验评估；最后总结全文。

2 相关工作

目前，已有一些研究工作关注于从微博中识别出事件。文献[3]提出在给定地理位置和时间的情况下，可根据微博集统计出微博数量、发微博的人数、进入该区域的流动人数。它将历史微博集中的上述 3 个属性与新产生的微博集中的对应属性进行比较，若发现有较大差异，则认为该区域发生了异常事件。但该算法并没有解决微博中的噪声问题。文献[4,5]使用了关键词的概念。它们在对微博进行分析之前，先预定义与社交事件相关的关键词，然后通过统计微博中关键词的出现次数来判断是否有异常事件发生。该方法在一定程度上解决了微博噪声的问题，但是关键词选择得是否恰当会直接影响事件检测的正确率。文献[6]将微博的文本属性、时间属性以及地理位置属性作为衡量微博之间相似度的 3 个维度。通过比较微博之间的相似度来对微博进行聚类。最后通过聚类结果中微博的数量来判断该类是否与异常事件有关。上述方法都是基于微博数量或者微博消息集中某一属性与历史数据相比是否出现突然性增长作为是否发生异常事件的判断标准。文献[7]提出了对微博数据流进行在线聚类的方法，并分析聚类结果中的每一个类，为每一个类提取时间特征、社交特征、主题特征和微博中心特征。然后，根据上述特征，利用 SVM 分类器判断每个类是否与异常事件有关。我们发现上述方法都没有充分考虑文本语义在事件检测中的重要性，并且忽略了微博各属性之间的关系。

微博本质上是一种文本文档，因此识别微博中的事件可以从识别文档的主题入手，而后者通常依赖主题模型（例如 LDA 模型）来完成[14]。

LDA 模型是根据文档的单词建立一个生成过程来确定文档的主题，可用于从大规模文档集和语料库中识别出潜藏的主题[8,9]，进行主题分析[15]。LDA 模型本质上是发现文档集中频繁共现的单词，将这些频繁共现的单词聚成一类形成一个主题。因此，LDA 模型能保证热门词项在主题内具有相对较高的语义相似度，在主题之间具有较大的语义差异。在将主题模型应用于微博时，需要考虑微博的特点。与传统的文本相比，微博的文本内容比较短，在很多情况下，一条微博就是一句话，甚至是一个词。因此，面向微博的主题模型为了解决微博短文本造成的高维稀疏问题，通常摒弃一篇文档包含多个主题的假设，基于每条微博仅包含一个主题的假设。文献[10-12,16,17]的工作都是基于此对 LDA 模型进行了改进。此外，文献[11]还在其主题模型中，把微博的发布时间作为生成对象，即在建模过程中，不仅考虑了文本内容对主题的影响，还考虑了微博发布时间对主题分布的影响。而文献[12]认为同一时间段发布的微博很有可能指向同一个主题，故它将时间分成若干段，根据发布时间将微博分配到每个时间段上，并认为同一个时间段中的所有微博服从同一个主题分布[12]。基于此提出了 TimeLDA 模型。同时，文献[12]认为同一作者发布的微博主题的分布是一定的，并基于此提出了 UserLDA 模型。但是，文献[12]所提的模型均没有考虑发布微博的地理位置信息对微博主题分布的影响。因此该模型在分析发生在同一时间、不同地点的事件时偏向于将这些事件归为同一个主题下，从而降低了对同一时间段内多个事件的辨识能力[16]。考虑到地理位置对微博主题分布的影响，提出将城市划分为多个区域，依据各兴趣点类型及数量对区域赋予权重以表达区域社会功能对微博主题的影响程度，据此改进微博主题建模的方法。但是该方法依赖于城市中的 POI 数据，因此对于如交通事故、突发火灾这样的不依赖于地理位置的异常事件不能进行有效的检测[17]。将已有的 TwitterLDA，AuthorLDA，UserLDA 3 种主题模型应用于微博数据，并比较了这 3 种方法在构建微博用户兴趣模型方面的性能。

3 ST-LDA 主题模型

在现实生活中，有些微博在发布时间和地理位置上呈现一定的聚集性。例如，在某场明星演出或者体育赛事中，大量微博用户会在现场发布与该主题相关的微博。这些微博除了在文本内容上具有语义相似性外，在时间、地理位置信息上也呈现出显著的聚集性。考虑到这些微博的存在，本文提出了一种新的主题模型 ST-LDA。该模型与文献[12]类似，为每一条微博分配一个主题，特别地，在给每个微博分配主题时，不仅考虑了微博发布时间对主题的影响，还增加了微博发布的地理位置对主题的影响。因此，ST-LDA 模型趋向于将邻近时空发布的具有语义相似性的微博归入同一主题，使主题的可理解性得以提高。

3.1 ST-LDA 的基本思想

在 ST-LDA 主题模型中，每一条微博 d 用一个三元组表示，$d=(c,t,g)$。其中，c 表示微博 d 的文本内容，t 表示发布微博的时间，g 表示发布微博的地理位置。

ST-LDA 主题模型基于以下 3 个假设：

(1)每条微博仅包含一个主题。同时，微博中的每一个词一定概率由背景噪声主题产生，后者的引入是为了解决微博数据噪声比较大的问题。

(2)同一时间段发布的微博很有可能指向同一个主题。为此，本文将时间分段，并认为同一个时间段中的所有微博服从同一个主题分布[12]。

(3)发布位置相对靠近的两条微博比发布位置相距较远的两条微博更有可能属于同一主题。为此，本文使用一个 Gaussian 分布来为每个主题刻画其在地理位置上的分布状况，该 Gaussian 分布的参数 μ 表示该主题发生的中心位置，σ^2 表示该主题在空间上的聚集程度。

表 1 列出了本文所使用的符号及含义。

表 1 符号及其含义

符号	描述
z_d	微博 d 的主题，$z_d \sim \text{Multi}(\theta_t)$
g_d	微博 d 的地理位置，$g_d \sim \text{Gaussian}(\mu,\sigma^2)$
w_{dn}	微博 d 的词项序号为 n 的单词
θ_t	t 时间段的微博主题所服从的多项分布的参数，$\theta_t \sim \text{Dirichlet}(\alpha)$
φ_k	隶属主题 k 的微博单词所服从的多项分布的参数，$\varphi_k \sim \text{Dirichlet}(\beta)$
φ_b	隶属噪声主题的微博单词所服从的多项分布的参数，$\varphi_b \sim \text{Dirichlet}(\beta)$
x_{dn}	标记 w_{dn} 是否是噪声的变量，$x_{dn} \sim \text{Bernoulli}(\rho)$，$\rho \sim \text{Beta}(\lambda)$

ST-LDA 的生成过程如算法 1 所示。在算法 1 中，K 表

示预定义的主题个数，T 表示时间分段个数，D_t 表示数据集中第 t 个时间段内的微博个数，N_d 表示微博 d 中的词项个数，N_g 表示在为每条微博分配主题的过程中地理位置信息的重要性。N_g 越大，表示微博越偏向于分配到与自身地理位置相似的主题，反之，微博越偏向于分配到与自身语义比较相似的主题。

算法 1 ST-LDA 的生成过程

1. 生成噪声主题的单词分布 $\varphi_b \sim \text{Dirichlet}(\beta)$，生成词项噪声比例 $\rho \sim \text{Beta}(\lambda)$
2. for t=1 to T do
3. 　生成 $\theta_t \sim \text{Dirichlet}(\alpha)$
4. end for
5. for k=1 to K do
6. 　生成 $\varphi_k \sim \text{Dirichlet}(\beta)$
7. end for
8. for t=1 to T do
9. 　for d=1 to D_t do
10. 　　生成 $z_d \sim \text{Multi}(\theta_t)$
11. 　　for n=1 to N_d do
12. 　　生成 $x_{d,n} \sim \text{Bernoulli}(\rho)$
13. 　　if $X_{d,n}=1$ then 生成 $w_{d,n} \sim \text{Multi}(\varphi_{Z_d})$
14. 　　else 生成 $w_{d,n} \sim \text{Multi}(\varphi_b)$
15. 　　end for
16. 　　for g=1 to N_g
17. 　　$g_d \sim \text{Gaussian}(\mu_{Z_d}, \sigma^2_{Z_d})$
18. 　　end for
19. 　end for
20. end for

3.2 ST-LDA 模型推导

在 ST-LDA 模型中存在两个隐藏变量，即一个微博的主题和微博中每个单词是否由背景噪声产生的标记。利用 ST-LDA 模型识别微博的主题就是在以 ST-LDA 模型方式生成微博的前提下求该模型中所有隐藏变量的条件概率，即一个微博 i 被分配给主题 k 的概率和一个单词由背景噪声主题生成的概率，进而通过 Gibbs 采样获得该微博的主题。

在 ST-LDA 模型中隐藏变量的条件概率表示如下：首先，一个微博 d 被分配给主题 k 的概率是 $p(z_d=k|z_{\neg d},x,g,w,\alpha,\beta,\lambda,\mu,\sigma)$，其中 z_d 表示微博 d 的主题，x 表示每个单词是否由背景噪声产生，g 表示每个微博的地理位置属性，w 表示所有词项；其次，微博 d 的第 n 个单词是否由背景噪声主题生成的概率为 $p(x_{dn}|x_{\neg dn},z,w,\alpha,\beta,\lambda,\mu,\sigma)$。由式(1)、式(2)可知，可以通过计算联合分布 $p(z,x,g,w|\alpha,\beta,\lambda,\mu,\sigma)$ 来得到一个微博 d 被分配给主题 k 的概率，以及通过计算联合分布 $p(x,w|z,\alpha,\beta,\lambda,\mu,\sigma)$ 来得到每个单词是否由背景噪声主题生成的概率。如式(1)、式(2)所示，隐藏变量的条件概率可以表示为两个联合概率相除，通过求联合概率分布来计算出模型中隐藏变量的条件概率。

$$p(z_d=k|z_{\neg d},x,g,w,\alpha,\beta,\lambda,\mu,\sigma)\propto \frac{p(z,x,g,w|\alpha,\beta,\lambda,\mu,\sigma)}{p(z_{\neg d},x_{\neg d},g_{\neg d},w_{\neg d}|\alpha,\beta,\lambda,\mu,\sigma)} \tag{1}$$

$$p(x_{dn}|x_{\neg dn},z,w,\alpha,\beta,\lambda,\mu,\sigma)\propto \frac{p(x,w|z,\alpha,\beta,\lambda,\mu,\sigma)}{p(x_{\neg dn},w_{\neg dn}|z,\alpha,\beta,\lambda,\mu,\sigma)} \tag{2}$$

由 ST-LDA 模型的生成方式可知，联合概率分布中的各部分是相互独立的，因此联合概率分布可以表示成各相互独立部分的乘积形式，如式(3)所示。

$$p(z,x,g,w|\alpha,\beta,\lambda,\mu,\sigma)=p(z|\alpha)\times p(x|\lambda)\times p(w|z,x,\beta)\times p(g|z,\mu,\sigma) \tag{3}$$

因此，需要推导出 $p(z|\alpha)$，$p(x|\lambda)$，$p(w|z,x,\beta)$，$p(g|z,\mu,\sigma)$。

首先，根据 Dirichlet 分布与多项分布的共轭性质，推导出 $p(z|\alpha)$，如式(4)所示，其中，N^{topic} 表示主题个数，$N^{document}_{time=t,z=k}$ 表示在时间段 t 内主题为 k 的微博个数，$N^{document}_{time=t}$ 表示在时间段 t 内的微博个数。

$$p(z|\alpha)=\prod_{t=1}^{T}\left(\frac{\Gamma(\alpha\cdot N^{topic})}{\Gamma(\alpha)^{N^{topic}}}\times\frac{\prod_{i=1}^{N^{topic}}\Gamma(\alpha+N^{document}_{time=t,z=k})}{\Gamma(\alpha\cdot N^{topic}+N^{document}_{time=t})}\right) \tag{4}$$

其次，推出 $p(x|\lambda)$，如式(5)所示，其中，$N^{word}_{x=1}$ 表示微博数据集中由正常主题产生的单词个数，$N^{word}_{x=0}$ 表示微博数据集中由噪声产生的单词个数，N^{word} 表示数据集中单词的个数。

$$p(x|\lambda)=\frac{\Gamma(2\lambda)}{\Gamma(\lambda)^2}\times\frac{\Gamma(\lambda+N^{word}_{x=0})\cdot\Gamma(\lambda+N^{word}_{x=1})}{\Gamma(2\lambda+N^{word})} \tag{5}$$

而 $p(w|z,x,\beta)$ 如式(6)所示，其中 $N^{vocabulary}$ 表示数据集中不同词的个数，$N^{word}_{x=1,z=k,w=j}$ 表示数据集中由主题 k 生成的词项序号为 j 的单词个数，$N^{word}_{x=1,z=k}$ 表示由主题 k 生成的单词个数，$N^{word}_{x=0,w=i}$ 表示由噪声主题生成的词项序号为 i 的单词个数。

$$\begin{aligned}&p(w|z,x,\beta)\\&=\frac{\Gamma(\beta\cdot N^{vocabulary})}{\Gamma(\beta)^{N^{vocabulary}}}\times\frac{\prod_{i=1}^{N^{vocabulary}}\Gamma(\beta+N^{word}_{x=0,w=i})}{\Gamma(\beta\cdot N^{vocabulary}+N^{word}_{x=0})}\times\\&\prod_{i=1}^{N^{topic}}\left(\frac{\Gamma(\beta\cdot N^{vocabulary})}{\Gamma(\beta)^{N^{vocabulary}}}\times\frac{\prod_{j=1}^{N^{vocabulary}}\Gamma(\beta+N^{word}_{x=1,z=k,w=j})}{\Gamma(\beta\cdot N^{vocabulary}+N^{word}_{x=1,z=i})}\right)\end{aligned} \tag{6}$$

最后，推导出 $p(g|z,\mu,\sigma)$，如式(7)所示。其中，$N^{document}_{z=k}$ 表示主题为 k 的微博个数，$g_{document=j}$ 表示第 j 个微博的地理位置信息，$\mu_{z=k}$，$\sigma_{z=k}$ 表示第 k 个主题的地理位置分布参数。

$$p(g|z,\mu,\sigma)=\prod_{i=1}^{N^{topic}}\prod_{j=1}^{N^{document}_{topic=i}}p(g_{document=j}|\mu_{z=k},\sigma_{z=k})^{N_g} \tag{7}$$

根据式(1)—式(7)，可以计算出 $p(z,x,g,w|\alpha,\beta,\lambda,\mu,\sigma)$ 的联合概率。我们发现该联合概率中除了模型中的先验参数以外，其余都是数据集中的各种统计项。类似地，式(1)中的分母项 $p(z_{\neg d},x_{\neg d},g_{\neg d},w_{\neg d}|\alpha,\beta,\lambda,\mu,\sigma)$ 可以通过统计微博 d 以外的其他微博而求得。这样，可以通过化简推导出对于每一个微博 d 分配主题 k 的概率，即 $p(z_d=k|z_{\neg d},x,g,w,\alpha,\beta,\lambda,\mu,\sigma)$，如式(8)所示。

$$\begin{aligned}&p(z_d=k|z_{\neg d},x,g,w,\alpha,\beta,\lambda,\mu,\sigma)\\&=(\alpha+N^{document}_{time=t,z=k}-1)\times\\&\frac{\Gamma(\beta\cdot N^{vocabulary}+N^{word}_{x=1,z=k}-N^{word}_{x=1,d=i})}{\Gamma(\beta\cdot N^{vocabulary}+N^{word}_{x=1,z=k})}\times\\&p(g_{document=d}|\mu_{z=k},\sigma_{z=k})^{N_g}\end{aligned} \tag{8}$$

用类似的方法求得 $p(x_{dn}|x_{\neg dn},z,w,\alpha,\beta,\lambda,\mu,\sigma)$。

接着，根据推导出的隐藏变量的条件概率，使用 Gibbs 采样估计出模型中的参数，在历经足够多次的迭代(实验中迭代了 500 次)后停止[13]。最后一次迭代时，该微博的主题采样结果被作为该微博的主题。然后，基于 $N^{document}_{time=t,z=k}$，$N^{document}_{time=t}$，

$N^{word}_{x=1,z=k,w=j}$，$N^{word}_{x=1,z=k}$，$N^{document}_{z=k}$ 利用最大似然估计的方法计算 θ_t 与 φ_k，同时，利用矩估计的方法求解每个主题地理位置的分布参数 μ 和 σ^2，计算公式见式(9)、式(10)。

$$\mu_{z=k}=\frac{1}{Num^{document}_{z=k}}\sum_{j=1}^{Num^{document}_{z=k}}X_j \tag{9}$$

$$\sigma^2_{z=k}=\frac{1}{Num^{document}_{z=k}}\sum_{j=1}^{Num^{document}_{z=k}}(X_j-\mu_{z=k})(X_j-\mu_{z=k})^{\mathrm{T}} \tag{10}$$

由于 φ_k 表示所有词项在主题 k 中的分布情况，因此可以根据 φ_k 推测出该主题所包含的语义描述。例如，隶属某一个主题的微博中，“足球，加油，比赛”等词项所占比例比较大，那么可以赋予该主题的语义为“足球赛事”。

4 识别时空事件

4.1 异常主题的检测

利用 ST-LDA 模型，可以分析出微博集中每个主题在每个时间段内所包含的微博数量。在此基础上，构建一个隐马尔科夫模型来刻画微博主题状态的变化。

用 X_n 表示一个时间段内的微博主题的状态，那么，$\{X_n, n=1,2,\cdots\}$ 是一个两状态的马尔科夫链，其中状态 0 代表该主题在该时间段内表现正常，1 代表该主题在该时间段内表现异常。设该马氏链的初始状态概率向量 $\pi=(p_0,p_1)=(P(X_1=0),P(X_1=1))$，其状态转移概率矩阵为 $A=\begin{bmatrix}\rho & 1-\rho\\ 1-\upsilon & \upsilon\end{bmatrix}$，该矩阵表示如果上一个时间段该主题表现正常，则以 ρ 的概率在下一个时刻也表现正常，以 $1-\rho$ 的概率在下一个时刻表现异常；如果在上一个时刻表现异常，那么该主题会以 υ 的概率在下一个时刻表现异常，以 $1-\upsilon$ 的概率在下一个时刻表现正常。其中，ρ 和 υ 可以通过历史数据得到。

由于仅可以观察到一个主题在 T 个时间段内的微博数量 c_i，而主题的状态 X_i 不能被观测到，即为一个隐含状态，因此构建一个隐马尔科夫模型 $H=(A,B,\pi)$，其中，A 和 π 如上所述，而 B 为观测概率矩阵，其元素 $B_{ij}=P(c_i\mid X_j)$ 表示在隐含状态是 X_j 的条件下，观察状态为 c_i 的概率。具体地，分别使用两个 Poisson 分布来描述正常和异常两个状态下微博发布的数量 c_i，其中，正常状态下的 Poisson 分布的参数 μ_0 取该主题在 T 个时间段内的平均微博数，即 $\mu_0=avg(c_1,c_2,\cdots,c_T)$，而异常状态下的 Poisson 分布的参数 μ_1 设为 $\mu_1=\mu_0*3$。

给定一个主题在 T 个时间段内的微博数量 $(c_1,c_2,\cdots,c_T)$，采用 Viterbi 算法即可预测出 T 个时间段内最有可能发生的主题状态序列 $(X_1,X_2,\cdots,X_T)$，若预测结果中 X_i 为 1，那么表示在时间段 i 中，该主题是表现异常的。

4.2 时空事件的描述

将一个时空事件定义为一个五元组，即 $e=(topic,time,location,report\text{-}time,report\text{-}location)$，其中，$topic$ 表示该事件所属主题，$time$ 和 $location$ 表示该事件发生的时间段和位置，$report\text{-}time$ 和 $report\text{-}location$ 表示报告该事件的时间和位置。

按如下方法描述从微博集中识别出的一组时空事件：将用 4.1 节所述方法识别出的发生在时间区间 i 的异常主题作为事件的主题，时间区间 i 的中间值作为事件的报告时间，并用该异常主题的中心位置作为事件的报告位置。查找该异常主题对应的微博，并对微博进行分词与词性标注，若在微博中发现标注为时间元素的词项(即表示时间的词项，例如周五)，那么将它作为事件发生的时间，否则将报告该事件的时间当作事件发生的时间，若在微博中发现标注为地理位置元素的词项，那么将它作为事件发生的位置，否则将报告该事件的位置当作事件发生的位置。对于所识别出的时空事件，按照事件所属异常主题所包含的微博个数的降序给事件排序。排序在先的事件被认为具有较高的热门程度。

5 实验评估

通过实验来评估用不同的方法从微博中识别时空事件的能力。使用查全率和查准率作为度量指标，其中查全率为识别出的事件数量占全部事件的比例，查准率为正确识别出的事件数量占识别出的所有事件数量的比例。

选用最为广泛使用的主题模型 LDA，在其基础上使用隐马尔科夫模型识别异常的方法增加事件检测功能，形成基于 LDA 的方法。类似地，以增加了时间维度的主题模型 Time-LDA 为基础，形成基于 TimeLDA 的方法，将其与所提出的基于 ST-LDA 的方法进行对比。

实验使用的数据集是从新浪微博开放平台下载的 2015 年 3 月份在北京五环内发布的微博，共有 314939 条。经过分词，去掉标点符号、停用词、微博中的链接、转发标记、@用户名、表情符号、高频词(如我们)和低频词后[17]，该数据集包含了 117043 条不同的词项。

采用基于 LDA 的方法、基于 TimeLDA 的方法和本文方法检测上述微博集中的事件。

在实验参数的设置上，将一天划分为 4 个时间段，即0 点－6 点，6 点－12 点，12 点－18 点，18 点－24 点。这样，3 月份的 31 天被划分成了 124 个时间段。将主题个数 K 设为 100。参考文献[13]设置 Dirichlet 分布的先验参数，即 $\alpha=50/K$，$\beta=0.01$，$\lambda=1$。并且，设 $p_0=0.5$，$p_1=0.5$，$\rho=0.9$，$\upsilon=0.6$。与 LDA 模型、TimeLDA 模型不同，ST-LDA 模型还有一个额外的初始化参数 N_g，将其初始化为 5。

记录实验中检测到的事件。在基于 ST-LDA 的方法和基于 TimeLDA 的方法中，统计每个主题在处于异常状态的时间段时所包含的微博数量并按降序进行排序，然后取前 20 个作为事件集。而在基于 LDA 的方法中，由于 LDA 模型是为一个单词分配一个主题，因此无法直接统计每个主题在其异常状态的时间段中包含的微博数量。为此，按式(11)定义一个事件 $e(topic=i,time=t)$ 的热门程度。

$$H(e)=H(topic_{i,t})=\sum_{d\in D_t}\frac{Num^{word}_{z=i,doc=d}}{Num^{word}_{doc=d}} \tag{11}$$

其中，$H(topic_{i,t})$ 表示主题 i 在时间段 t 中的热门程度，D_t 表示时间段 t 内的所有微博，$Num^{word}_{z=i,doc=d}$ 表示微博 d 中主题为 i 的单词数量，$Num^{word}_{doc=d}$ 表示微博 d 中总单词数量。依据式(11)，对基于 LDA 方法检测出的事件进行排序，并取前 20 个为事件集。

表 2 列出了在 ST-LDA 模型得到的前 5 个异常事件相对应主题的 Top10 词项。通过每个主题的 Top10 词项，可以推断出每个主题对应的社会热门事件。接着，对 3 个实验各自获得的前 20 个事件进行人工标注，判断不同方法得到的每个事件是否能映射到现实生活中确实发生过的事件。称所有能

映射到现实事件的为从该微博集中识别出的真实事件集合。将3个实验各自识别出的事件的并集作为微博集中包含的全部事件。表3列出了2015年3月微博数据集中包含的真实事件。表4列出了所提方法所识别出的前20个事件。

表2 ST-LDA模型主题的Top10词项

主题语义	Top10词项
篮球	首钢,北京,恭喜,牛,冠军,赢,场,加油,决赛,比赛
足球	国安,足球,球迷,工体,三里屯,比赛,胜利,主场,亚冠,VS
315晚会	315,晚会,呷哺呷哺,曝光,鸭血,央视,消费者,电话,安全,WiFi
地铁出轨	列车,线,地铁,亦庄,乘客,影响,受伤,调试,出轨,试车
华能电厂	华能,电厂,发生,朝阳区,大火,伤亡,火灾,直升机,王四营,起火

表3 数据集中的真实事件

事件编号	日期	事件内容
1	3号	北京首钢vs广东宏远,篮球赛,万事达中心
2	15号	北京首钢vs辽宁药都,篮球赛,万事达中心
3	17号	北京首钢vs辽宁药都,篮球赛,万事达中心
4	19号	北京首钢vs辽宁药都,篮球赛,万事达中心
5	4号	北京国安vs水原三星,足球赛,工人体育场
6	13号	北京国安vs河南建业,足球赛,工人体育场
7	17号	北京国安vs浦和红宝石,足球赛,工人体育场
8	15号	315晚会
9	25号	地铁亦庄线出现故障
10	13号	华能电厂起火
11	2号	威廉王子访华
12	16号	王府井抢劫案件
13	5号	元宵节
14	8号	三八妇女节
15	14号	白色情人节
16	26号	电影《速度与激情》北京发布会,三里屯
17	22号	北京首钢CBA夺冠

表4 ST-LDA模型识别出的事件

主题编号	主题语义	对应的事件编号
1	篮球	(4)
2	足球	(7)
3	足球	(6)
4	篮球	(3)
5	篮球	(1)
6	315晚会	(8)
7	足球	(5)
8	topic:48	感情类事件
9	CBA冠军	(17)
10	topic:54	感情类事件
11	地铁出轨	(9)
12	华能电厂	(10)
13	威廉王子	(11)
14	Topic:55	感情类事件
15	王府井抢劫	(12)
16	元宵节	(13)
17	速度与激情	(16)
18	Topic:81	感情类事件
19	篮球	(2)
20	Topic:36	感情类事件

根据3个实验各自得到的前20个事件结果,计算出3个模型的查准率和查全率,如表5、表6所列。

表5 3个模型的查准率比较

事件个数	Top5	Top10	Top15	Top20
LDA	1.0	0.8	0.67	0.5
TimeLDA	0.8	0.6	0.6	0.6
ST-LDA	**1.0**	**0.8**	**0.87**	**0.8**

表6 3个模型的查全率比较

事件个数	Top5	Top10	Top15	Top20
LDA	0.29	0.47	0.59	0.59
TimeLDA	0.24	0.35	0.53	0.71
ST-LDA	**0.29**	**0.47**	**0.71**	**0.88**

我们发现3种方法所识别出的前20个事件除了真实事件外,还包含感情类事件。基于LDA的方法所识别出的感情类事件最多(10个),基于TimeLDA的方法其次(9个),本文的方法最少(4个)。这是由于感情类事件通常不具有地理位置聚集性,按照本文方法,这类微博往往不能聚集在一个主题下,而另两种方法不能避免这一点。

从实验结果看,本文的方法能区分两个在相同时间、不同地点发生的同语义的事件(如体育比赛),例如,本文方法能区分出3月17日发生的两个事件:万事达中心的篮球赛(北京首钢vs辽宁药都)和工人体育场的足球赛(北京国安vs浦和红宝石)。而基于TimeLDA的方法则将这两场赛事的微博纳入了一个主题下,从而增加了主题理解的困难。类似地,本文方法能区分两个在不同时间但同一地点发生的同语义的事件(如体育比赛),例如,本文方法能区分3月的不同时间和在万事达中心发生的多个篮球赛以及在工人体育场发生的多个足球赛。

结束语 从文本短、噪声大的大量微博中准确识别出事件的内容、时间和位置是一个挑战性任务。从分析微博的主题入手,通过改进LDA主题模型提出了一个面向微博数据的主题模型ST-LDA,进而从微博的异常主题中获得时空事件。特别是以生成方式识别事件,为此首先构造生成模型,将那些既在文本语义上相似又在时空上聚集的微博数据归到一个主题,然后判断这样的主题的异常情况,进而提取主题的语义和时间、空间信息用于生成时空事件。实验结果表明,本文方法比基于LDA的方法、基于TimeLDA的方法能更准确、更全面地识别出时空事件。

参考文献

[1] Zhang W, Qi G, Pan G, et al. City-scale Social Event Detection and Evaluation with Taxi Traces[J]. ACM TIST, 2015, 6(3): 1-20

[2] Du R, Yu Z, Mei T, et al. Predicting activity attendance in event-based social networks: content, context and social influence[C]// UbiComp. 2014: 425-434

[3] Lee R, Wakamiya S, Sumiya K. Discovery of unusual regional social activities using geo-tagged microblogs[J]. World Wide Web, 2011, 14(4): 321-349

[4] Sakaki T, Okazaki M, Matsuo Y. Earthquake shakes Twitter users: real-time event detection by social sensors[C]// Proceedings of the 19th International Conference on World Wide Web. ACM, 2010: 851-860

[5] Sankaranarayanan J, Samet H, Teitler B E, et al. Twitterstand: news in tweets[C]// Proceedings of the 17th ACM Sigspatial International Conference on Advances in Geographic Information Systems. ACM, 2009: 42-51

[6] Becker H, Naaman M, Gravano L. Learning similarity metrics for event identification in social media[C]// Proceedings of the Third ACM International Conference on Web Search and Data Mining. ACM, 2010: 291-300

[7] Becker H, Naaman M, Gravano L. Beyond Trending Topics: Real-World Event Identification on Twitter[J]. ICWSM, 2011, 11:438-441

[8] Blei D M, Ng A Y, Jordan M I, et al. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3:993-1022

[9] Blei D M. Introduction to Probabilistic Topic Models[J]. Signal Processing Magazine IEEE, 2011, 27(6):55-65

[10] Ramage D, Dumais S T, Liebling D J. Characterizing Microblogs with Topic Models[J]. ICWSM, 2010, 5(4):130-137

[11] 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

[12] Diao Q, Jiang J, Zhu F, et al. Finding bursty topics from microblogs[C]// Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, Association for Computational Linguistics. 2012:536-544

[13] Heinrich G. Parameter estimation for text analysis[R]. Technical Report, 2004

[14] Xu Ge, Wang Hou-feng. The Development of Topic Models in Natural Language Processing[J]. Chinese Journal of Computer, 2011, 34(8):1423-1436(in Chinese)
徐戈, 王厚峰. 自然语言处理中主题模型的发展[J]. 计算机学报, 2011, 34(8):1423-1436

[15] Shi Jing, Fan Meng, Li Wan-long. Topic Analysis Based on LDA Model[J]. Acta Automatica Sinica, 2009, 35(12):1586-1592(in Chinese)
石晶, 范猛, 李万龙. 基于 LDA 模型的主题分析[J]. 自动化学报, 2009, 35(12):1586-1592

[16] Duan Lian, Guo Wei, Zhu Xin-yan, et al. Constructing Spatio-Temporal Topic Model for Microblog Topic Retrieving[J]. Geomatics and Information Science of Wuhan University, 2014, 39(2):210-213(in Chinese)
段炼, 吴维, 朱欣焰, 等. 基于时空主题模型的微博主题提取[J]. 武汉大学学报(信息科学版), 2014, 39(2):210-213

[17] Chen Wen-tao, Zhang Xiao-ming, Li Zhou-jun. Analysis of Topic Models on Modeling MicroBlog User Interestringness[J]. Computer Science, 2013, 40(4):127-130(in Chinese)
陈文涛, 张小明, 李舟军. 构建微博用户兴趣模型的主题模型的分析[J]. 计算机科学, 2013, 40(4):127-130

(上接第 192 页)

大小，比较结果如图 4 所示。

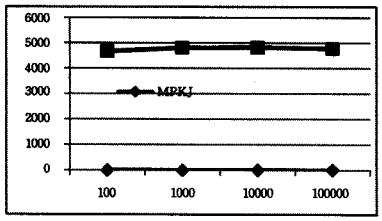

图 4 固定 M、N、D、K，改变 MAX 时 MPKJ 所用时间

从图 4 中可以明显看出，MAX 值的大小与查询时间并无太大关系，MAX 的值只是限定了所产生数据中的最大值，但对于最大值个数等并无限定，所以出现 MAX 增大但计算时间减少的情况，也就不足为奇了。

通过以上实验可以得出结论，随着数据规模的不断增大，串行 KNNJ 所用时间呈十倍以上趋势上升，但 MPKJ 变化缓慢，甚至达到上百倍的加速比，与现有加速比只有几十倍的改进算法相比，优势非常明显，因此 MPKJ 算法具有明显的高效性和可扩展性。

结束语 针对 GPU 的特性，本文设计了一种基于 CUDA 的并行 KNNJ 加速方法 MPKJ，MPKJ 能高效地处理大规模空间数据的查询问题，也可用于其它相似性连接查询问题。实验表明，随着数据规模的不断增大，MPKJ 并行优化的效果也越来越明显，加速比呈成倍增长的趋势。但本文在 CPU 与 GPU 之间的传输问题上并没有过多优化，因此，下一步需要尽可能缩短 CPU 与 GPU 之间的传输时间，进一步优化算法性能，增强 MPKJ 的可扩展性和可适用性。

参 考 文 献

[1] Bohm C, Krebs F. The k-nearest neighbor join: Turbo charging the KDD process[J]. Knowledge Information System, 2004: 6(6):728-749

[2] Xia C Y, Lu H J, Coi B C, et al. Gorder: An efficient method for KDD joins processing[C]// Proc. of the 30th Int'l Conf. on VeryLarge Data Bases. 2004:756-767

[3] Yu C, Cui B, Wang S G, et al. Efficient index-based KNN join processing for high-dimensional data[J]. Information and Software Technology, 2007, 49(4):332-344

[4] Yao B, Li F F, Kumar P. K nearest neighbor queries and KNN-joins in large relational databases (almost) for free[C]// Proc. of the 26th Int'l Conf. on Data Engineering (ICDE). 2010:4-15

[5] Wu En-hua. Technology, current situation and challenge of graphics processor are used for general computing[J]. Journal of Software, 2004, 15(10):1493-1504(in Chinese)
吴恩华. 图形处理器用于通用计算的技术、现状及其挑战[J]. 软件学报, 2004, 15(10):1493-1504

[6] Xu Xue-gui, Zhang Qing. High efficiency parallel remote sensing image processing based on CUDA[J]. Geospatial Information, 2011, 9(6):47-54(in Chinese)
许雪贵, 张清. 基于 CUDA 的高效并行遥感影像处理[J]. 地理空间信息, 2011, 9(6):47-54

[7] Dong Luo, Ge Wan-cheng, Chen Kang-li. Application Research of CUDA parallel computing [J]. Information Technology, 2010, 4(5):11-15(in Chinese)
董荦, 葛万成, 陈康力. CUDA 并行计算的应用研究[J]. 信息技术, 2010, 4(5):11-15

[8] Liu Yi, Jing Ning, Chen Luo, et al. Algorithm for Processing k-Nearest Join Based on R-Tree in MapReduce[J]. Journal of Software, 2013, 24(8):1836-1851(in Chinese)
刘义, 景宁, 陈荦, 等. MapReduce 框架下基于 R-树的 k-近邻连接算法[J]. 软件学报, 2013, 24(8):1836-1851

[9] Sosutha S, Mohana D. Heterogeneous parallel computing using CUDA for chemical process [J]. Procedia Computer Science, 2015, 47(1):237-246

[10] Leutenegger S T, Lopez M A, Edgington J. STR: A Simple and Efficient Algorithm for R-tree Packing [C]// The 13th International Conference on Data Engineering. Birmingham, England, 1997:497-506

[11] Hoare C A R. Quicksort [J]. The Computer J., 1962, 15(1):10-15