·先进计算与数据处理· 文章编号:1000-3428(2018)03-0055-05 文献标志码:A 中图分类号:TP391

基于贝叶斯网络的故事线挖掘算法

佘玉轩,熊 赟

(复旦大学 计算机科学技术学院 上海市数据科学重点实验室,上海 201203)

摘 要: 目前的多数故事线挖掘研究侧重新闻文献和事件的相似性分析,忽略了故事线的结构化表述及新闻具有的延时性,无法直观地从模型结果看出不同新闻话题的发展过程。为此,提出一种基于贝叶斯网络的无监督故事线挖掘算法。将故事线看成日期、时间、机构、人物、地点、主题和关键词的联合概率分布,并考虑新闻时效性。在多个新闻数据集上进行的实验和评估结果表明,与 K-means、LSA 等算法相比,该算法模型具有较高的故事线挖掘能力。

关键词: 故事线挖掘;事件;贝叶斯网络;时效性;新闻;主题

中文引用格式: 佘玉轩,熊 赟. 基于贝叶斯网络的故事线挖掘算法[J]. 计算机工程,2018,44(3):55-59.

英文引用格式: SHE Yuxuan, XIONG Yun. Storyline Mining Algorithm Based on Bayesian Network[J]. Computer Engineering,2018,44(3):55-59.

Storyline Mining Algorithm Based on Bayesian Network

SHE Yuxuan, XIONG Yun

(Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University, Shanghai 201203, China)

【Abstract】 At present, most of the research on story line mining focuses on the similarity analysis of news documents and events, while ignoring the structured expression of stories and the delay of news. It is difficult to intuitively see the development of different news topics from the model results. Therefore, an unsupervised storyline mining algorithm based on Bayesian network is proposed, which considers the story line as the joint probability distribution of date, time, organization, person, place, topic and key words and considers the timeliness of news in inside. Experiments and evaluations results on multiple news datasets show that this algorithm model has a higher mining potential than the K-means and LSA algorithms.

【Key words】 storyline mining; event; Bayesian network; timeliness; news; topic

DOI: 10.3969/j.issn.1000-3428.2018.03.009

0 概述

随着新闻媒体网站的迅速发展,每天都会产生大量的新闻。新闻之间具有很强的关联性,很多新闻都是针对同一个话题,只是时间不同,事件发展的阶段不同。虽然有百度、谷歌等搜索引擎,但是它们返回的结果仍然是繁杂的,很难从如此庞大的搜索结果中看出新闻事件的发展脉络。因此,能够自动地从新闻文献中挖掘事件和故事线,使用户清晰快速地获取新闻事件的结构化信息和发展过程很有必要。

然而,故事线挖掘面临很多挑战。目前的很多工作将不同的故事线看作不同的簇,将故事线挖掘转化成聚类问题[1-2]。但该方法有个很明显的缺点,就是将明显不相关的事件能够区分出来,比如美国总统竞选和火灾,但是对于相似的故事线,比如地震和火灾,模型识别效果很差。文献[3]利用文本摘要算法计算不同时间段的文本相似性,对大量的文献进行故事线抽取。文献[4]提出一种“话题关注度”的量化表示方法,构建了热点话题发现模型。文献[5]基于主题模型为故事线中的事件建模。文献[6]基于主题模型,提出一种时间情境依赖的时序微博话题检测方法。文献[7]利用隐主题分析技术来挖掘故事线。文献[8]利用最小权重支配集和有向斯坦纳树在微博数据集上生成故事线。文献[9]首先利用词共现方法定义不同类型的子事件,然后运用最优化方法将子事件串成时间线来提取故事

基金项目: 国家自然科学基金(91546105,71331005);国家高技术研究发展计划项目(2015AA020105);上海市科委项目(16JC1400801,16511102204);NSFC-广东联合基金(第二期)超级计算科学应用研究专项;国家超级计算广州中心支持项目。

作者简介: 佘玉轩(1992—),男,硕士研究生,主研方向为文本挖掘;熊 赟,教授、博士生导师。

收稿日期: 2017-02-27 **修回日期:** 2017-03-29 **E-mail:** 14210240020@fudan.edu.cn

线。文献[10]利用图论知识,通过构建有向树将故事线和事件联系起来。文献[11]提出了基于多中心模型的网络热点话题发现算法。这些研究大多侧重新闻文献和事件的相似性分析,而忽略了故事线的结构化表述。因此,很难直观地从模型结果看出不同新闻话题的发展过程。

贝叶斯网络能够有效地对建模对象进行结构化描述。非常经典的贝叶斯网络,比如 LDA[12],就可以结构化描述文档、主题、词三者之间的关系。将贝叶斯网络用于故事线挖掘的研究较少,且目前的贝叶斯网络结构较为简单。它们或者将事件看作一个整体,而不区分时间、地点、人物、机构等要素,或者忽略了新闻的时效性,将新闻发布时间当成新闻的发生时间。文献[13]构建了贝叶斯网络挖掘故事线,并且利用 Twitter 数据辅助学习故事线。文献[14]构建了 EHDP(Evolutionary Hierarchical Dirichlet Process)来挖掘故事线中的事件模式。这些研究缺乏对事件的详细拆分与描述。文献[15-16]将故事线看作命名实体、主题的分布,但是模型忽略了新闻文献的时效性,且对事件的描述较为粗糙。

本文构建一个新的贝叶斯网络来识别事件和故事线。假设每篇新闻属于一个故事线 s。故事线包含事件,事件由日期 d、时间 t、机构 o、人物 p、地点 l 等 5 个因素组成。故事线 s 由事件按时间顺序串联而成,且故事线有自己的主题分布和关键词分布。

1 故事情节挖掘模型

1.1 问题描述

给定新闻语料库 A,每篇新闻包括发布时间和新闻内容 2 个部分。即:

$$A=\{<e_1,c_{1,1}>,<e_1,c_{1,2}>,\cdots,<e_1,c_{1,n_1}>,\cdots,<e_f,c_{f,1}>,\cdots,<e_f,c_{f,n_f}>\}$$

新闻语料库共有 f 天数据,第 i 天共有 n_i 篇新闻,e_i 表示第 i 天,$c_{i,j}$ 表示第 i 天第 j 篇新闻。每篇新闻都在描述一个事件,而事件对应于某一新闻话题(如“2016 年美国总统大选”)的某一个发展阶段。本文的目标就是将描述同一个新闻话题的新闻放在一起,按照时间顺序排列,并将能代表新闻内容的特定词语抽取出来。将抽取出的这些词语按照新闻话题分组,以时间先后顺序连接且能代表新闻内容的词语,称为故事线。可以看出,有多少个新闻话题,就对应多少条故事线。

1.2 基于贝叶斯网络的故事情节挖掘模型

本文提出一种基于贝叶斯网络的故事线挖掘算法的基于贝叶斯网络的故事情节挖掘模型(Storyline Mining Model based on Bayesian Network,SMBN)。SMBN 的图模型如图 1 所示。

该模型具有以下特点:

1)图模型是针对每一天的新闻文献,即新闻文献按照发布时间分组,每一天的新闻文献对应的概率图模型如图 1 所示。

2)每篇新闻属于一个故事线。故事线由多项式分布 π_s 生成。

3)对故事线进行非常细致的结构化描述。故事线由若干日期 d、时间 t、机构 o、人物 p、地点 l、主题 z 以及关键词 w 组成。

4)为了描述相同故事线中不同故事之间的关联性,故事线-日期分布、故事线-时间分布、故事线-机构分布、故事线-人物分布、故事线-地点分布、故事线-关键词分布均依赖于过去 M 个时间段的概率分布。

5)文档 a 中的每个词 $w_{a,n}$ 的选取来源由指示器 x 确定。当 $x=0$ 时,从故事线-关键词分布中选取一个词;当 $x=1$ 时,从背景-关键词分布中选取一个词;当 $x=2$ 时,从主题-关键词分布中选取一个词,而主题 z 则由多项式分布 θ 生成。

6)区分新闻发布时间 e 和新闻发生时间 d,修正新闻的时效性问题。

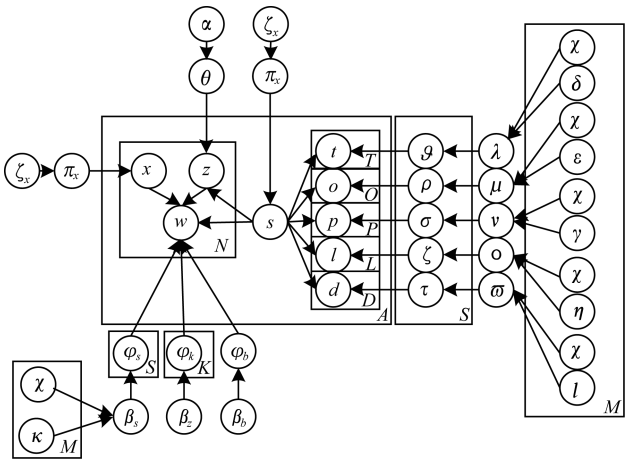

图 1 SMBN 概率图模型

SMBN 的生成算法如算法 1 所示。

算法 1 SMBN 生成算法

1. Draw$\theta \sim Dir(\alpha)$
2. Draw$\pi_s \sim Dir(\zeta_s)$
3. for each storylines$\in\{1,2,\cdots,S\}$ do
4. Draw$\varphi_s \sim Dir(\beta_s)$
5. for each topick$\in\{1,2,\cdots,K\}$ do
6. Draw$\varphi_k \sim Dir(\beta_z)$
7. Draw$\varphi_b \sim Dir(\beta_b)$
8. Draw$\pi_x \sim Dir(\zeta_x)$
9. for each storylines$\in\{1,2,\cdots,S\}$ do
10. Draw a distribution over
11. time$\vartheta_s \sim Dir(\lambda_s)$
12. organization $\rho_s \sim Dir(\mu_s)$
13. person$\sigma_s \sim Dir(\nu_s)$

14. location$\zeta_s \sim \mathrm{Dir}(o_s)$
15. date$\tau_s \sim \mathrm{Dir}(\varpi_s)$
16. for each documenta $\in \{1,2,\cdots,A\}$ do
17. chooses$_a \sim \mathrm{Multi}(\pi_s)$
18. choose t,o,p,l,d
19. for other word positionsn $\in \{1,2,\cdots,N\}$ do
20. choosex$_{a,n} \sim \mathrm{Multi}(\pi_x)$
21. if $x_{a,n}=0$
22. then draww$_{a,n} \sim \mathrm{Multi}(\varphi_s)$
23. if $x_{a,n}=1$
24. then draww$_{a,n} \sim \mathrm{Multi}(\varphi_b)$
25. if $x_{a,n}=2$
26. then drawz$_{a,n} \sim \mathrm{Multi}(\theta)$,
27. draww$_{a,n} \sim \mathrm{Multi}(\varphi_{z_{a,n}})$

在算法 1 中,θ、π_s、π_x、φ_s、φ_k、φ_b、ϑ、ρ、σ、ζ、τ 为多项式分布的参数,α、ζ_s、ζ_x、β_s、β_z、β_b、λ、μ、ν、o、$\overline{w}$ 为对应多项式分布的 Dirichlet 先验分布的参数。故事线 s_a 由多项式分布 π_s 生成,代表一篇文档 a 所属的故事线。每个故事线针对日期、时间、机构、人物、地点都有一个多项式分布,且每个分布的先验依赖于过去 M 个时间段的概率分布。新闻文献中的词分 3 种:故事线关键词,主题关键词和背景关键词,具体选取哪种类型的词依赖于指示器 x。

2　模型推导

模型推导的核心是估计 3 个隐变量的后验概率:1)文档 a 所属的故事线 s_a;2)文档 a 中词 $w_{a,n}$ 对应的主题 $z_{a,n}$;3)指示器 x。本文采用 Gibbs 采样方法来进行模型推导。

故事线 s_a 的 Gibbs 采样公式为:

$$p(s_a=i|\overrightarrow{s_{-a}},\vec{z},\vec{w},\vec{t},\vec{o},\vec{p},\vec{l},\vec{d}) \propto \frac{n_{i,-a}+\zeta_s}{n_{-a}+S\zeta_s} \cdot \frac{\prod_{t}^{T}\prod_{c=1}^{n_{i,t}^{(*)}}(n_{i,t}+\lambda_{i,t}-c)}{\prod_{c=1}^{n_i^{(*)}}(n_i^T+\sum_{t=1}^{T}\lambda_{i,t}-c)} \cdot \frac{\prod_{l}^{L}\prod_{c=1}^{n_{i,l}^{(*)}}(n_{i,l}+o_{i,l}-c)}{\prod_{c=1}^{n_i^{(*)}}(n_i^L+\sum_{l=1}^{L}o_{i,l}-c)} \cdot \frac{\prod_{o}^{O}\prod_{c=1}^{n_{i,o}^{(*)}}(n_{i,o}+\mu_{i,o}-c)}{\prod_{c=1}^{n_i^{(*)}}(n_i^O+\sum_{o=1}^{O}\mu_{i,o}-c)} \cdot \frac{\prod_{p}^{P}\prod_{c=1}^{n_{i,p}^{(*)}}(n_{i,p}+\nu_{i,p}-c)}{\prod_{c=1}^{n_i^{(*)}}(n_i^P+\sum_{p=1}^{P}\nu_{i,p}-c)} \cdot \frac{\prod_{d}^{D}\prod_{c=1}^{n_{i,d}^{(*)}}(n_{i,d}+\overline{w}_{i,d}-c)}{\prod_{c=1}^{n_i^{(*)}}(n_i^D+\sum_{d=1}^{D}\overline{w}_{i,d}-c)} \cdot \frac{\prod_{v}^{V}\prod_{c=1}^{n_{i,v}^{(*)}}(n_{i,v}+\beta_{i,v}-c)}{\prod_{c=1}^{n_i^{(*)}}(n_i^V+\sum_{v=1}^{V}\beta_{i,v}-c)} \cdot \frac{\prod_{k}^{K}\prod_{c=1}^{n_{i,k}^{(*)}}(n_{i,k}+\alpha_k-c)}{\prod_{c=1}^{n_i^{(*)}}(n_i^K+\sum_{k=1}^{K}\alpha_k-c)} \cdot \frac{\prod_{c=1}^{V}\prod^{n_{k,v}^{(*)}}(n_{k,v}+\beta_{k,v}-c)}{\prod_{c=1}^{n_k^{(*)}}(n_k^V+\sum_{v=1}^{V}\beta_{k,v}-c)} \tag{1}$$

其中,$-a$ 表示除去文档 a 后的数量统计,$n_{i,-a}$ 表示除去文档 a 后属于故事线 i 的文档数目,n_{-a} 表示除去文档 a 后的文档数量,$n_{i,t}$ 表示时间 t 被分配到故事线 i 中的次数,n_i^T 表示所有被分配到故事线 i 中的时间次数,$n_{i,l}$ 表示地点 l 被分配到故事线 i 中的次数,n_i^L 表示所有被分配到故事线 i 中的地点次数,$n_{i,o}$ 表示机构 o 被分配到故事线 i 中的次数,n_i^O 表示所有被分配到故事线 i 中的机构次数,$n_{i,p}$ 表示人物 p 被分配到故事线 i 中的次数,n_i^P 表示所有被分配到故事线 i 中的人物次数,$n_{i,d}$ 表示日期 d 被分配到故事线 i 中的次数,n_i^D 表示所有被分配到故事线 i 中的日期次数,$n_{i,k}$ 表示故事线 i 中主题为 k 的词数,n_i^K 表示故事线 i 中所有主题的词数,$n_{k,v}$ 表示主题 k 中词 v 出现的次数,n_k^V 表示主题 k 中所有词的数目,上标 (a) 表示数目统计仅针对文档 a。

第 a 篇文档中词 $w_{a,n}$ 对应的主题 $z_{a,n}$ 以及指示器 x 的 Gibbs 采样公式如下。

词 $w_{a,n}$ 被添加到故事线 i 关键词的后验概率为:

$$p(x_{a,n}=0|\overrightarrow{x_{a,-n}},\overrightarrow{w_i}) \propto \frac{n_{a,i}+\zeta_x-1}{\sum_{j=0}^{2}(n_{a,j}+\zeta_x)-1} \cdot \frac{n_{i,w_{a,n}}+\beta_{i,w_{a,n}}-1}{\sum_{v=1}^{V}(n_{i,v}+\beta_{i,v})-1} \tag{2}$$

其中,$n_{a,i}$ 为文档 a 中属于故事线 i 的关键词数,$n_{i,w_{a,n}}$ 表示故事线 i 中词 $w_{a,n}$ 出现的次数。

词 $w_{a,n}$ 被添加到背景词的后验概率为:

$$p(x_{a,n}=1|\overrightarrow{x_{a,-n}},\overrightarrow{w_b}) \propto \frac{n_{a,b}+\zeta_x-1}{\sum_{j=0}^{2}(n_{a,j}+\zeta_x)-1} \cdot \frac{n_{b,w_{a,n}}+\beta_{b,w_{a,n}}-1}{\sum_{v=1}^{V}(n_{b,v}+\beta_{b,v})-1} \tag{3}$$

其中,$n_{a,b}$ 为文档 a 中属于背景词的词数,$n_{b,w_{a,n}}$ 表示背景词中 $w_{a,n}$ 出现的次数。

词 $w_{a,n}$ 被添加到故事线 i 的主题 k 中的后验概率为:

$$p(x_{a,n}=2,z_{a,n}=k|\overrightarrow{x_{a,-n}},\overrightarrow{w_{i,z}}) \propto \frac{n_{a,z}+\zeta_x-1}{\sum_{j=0}^{2}(n_{a,j}+\zeta_x)-1} \cdot \frac{n_{i,k}+\beta_k-1}{\sum_{y=1}^{K}(n_{i,y}+\beta_y)-1} \cdot \frac{n_{k,w_{a,n}}+\beta_{k,w_{a,n}}-1}{\sum_{v=1}^{V}(n_{k,v}+\beta_{k,v})-1} \tag{4}$$

其中,$n_{a,z}$ 为文档 a 中属于主题词的数目,$n_{i,k}$ 表示故事线 i 中主题的词数,$n_{k,w_{a,n}}$ 表示主题 k 中词 $w_{a,n}$ 出现的词数。

通过 Gibbs 采样,就可以得出每篇新闻所属的故事线 s 和每个词所属的主题 z。由于日期、时间、机构、人物、地点、关键词等都是可观测变量,接下来只需把属于相同故事线的新闻按照时间顺序排序,提取出日期、时间、机构、人物、地点、主题、关键词等有代表性词语即可。

3　实验结果与分析

3.1　数据集

为了评估本文的故事线挖掘模型性能,从 CNN

和 ABC 这 2 个新闻网站爬取新闻文献。其中在 CNN 爬取新闻 4 203 条,包含故事线 88 个,此为数据集 1;在 ABC 爬取新闻 2 185 条,包含故事线 42 个,此为数据集 2。本文运用斯坦福命名实体识别算法来识别新闻中的命名实体。

3.2 对比算法

本文选择以下 5 种算法进行对比:

1)K-means 算法:通过 K-means 聚类算法对所有新闻文档进行聚类。

2)LSA 算法:LSA 利用 SVD 分解方法达到文本特征的降维,可用于文本中发现隐含的语义维度。

3)LDA 算法:用 Latent Dirichlet Allocation(LDA)主题模型来生成每篇文档的隐含故事线。每个故事线表示为命名实体和关键词的联合分布。

4)DSDM 算法[15]:该方法假设故事线是命名实体和主题的联合概率分布。该算法不对命名实体进行进一步区分。

5)DSEM 算法[16]:该方法假设事件由机构、人物、地点组成。故事线是事件、主题和关键词的联合概率分布。

对于 K-means 和 LSA,数据集 1 和数据集 2 的簇数目分别设置为 100 和 50。对于 LDA,数据集 1 和数据集 2 的主题数分别设置为 100 和 50。对于 DSDM,数据集 1 的主题数和故事线数都设置为 100,数据集 2 的主题数设置为 100,故事线数设置为 50。本文算法 SMBN 参数和 DSDM 设置一样。对于 DSEM 算法,数据集 1 和数据集 2 的主题数都设置为 100。

3.3 实验结果

本文使用准确率、召回率和 F1 值 3 个指标来评价各个模型的结果。故事线抽取正确有 2 点要求:1)算法抽取的故事线中新闻文献的主题、命名实体、关键词正确;2)故事线起始时间正确。

实验结果如表 1 所示。

表 1 在 2 个数据集上的模型效果比较 %

算法	数据集 1			数据集 2		
	准确率	召回率	F1 值	准确率	召回率	F1 值
K-means	20.00	22.73	21.28	24.00	28.57	26.09
LSA	29.00	32.95	30.85	32.00	38.10	34.78
LDA	34.00	38.64	36.17	36.00	42.86	39.13
DSDM	52.00	59.09	55.32	52.00	61.90	56.52
DSEM	64.63	60.23	62.35	63.64	66.67	65.12
SMBN	75.00	85.23	79.79	74.00	88.10	80.43

从表 1 结果可以看出,使用 K-means 聚类取得的实验效果是最差的。运用 LSA 和 LDA 2 个主题模型效果显著高于 K-means。其中 LDA 效果略优于 LSA。DSDM 构建了贝叶斯网络,将故事线、事件等用结构化的信息进行表述。但是 DSDM 不详细区分时间、地点、人物等事件因素。DSDM 效果明显优于 LDA。在数据集 1 中,DSDM 的 F1 值比 LDA 高出将近 20 个百分点。在数据集 2 中,DSDM 的 F1 值比 LDA 高出 17 个百分点。DSEM 将命名实体明确分为机构、人物、地点 3 类,将新闻发布的时间作为事件发生的时间,忽略了新闻也有一定的延时性。DSEM 模型整体优于 DSDM 模型。在 2 个数据集中 DSEM 都比 DSDM 效果要好。

通过考虑新闻的时效性,并对事件、故事线进行更加详细的划分与描述,本文模型 SMBN 将故事线看作日期、时间、地点、人物、机构、主题和关键词的联合概率分布,通过贝叶斯网络将故事线、事件、主题和关键词联系起来。本文算法相比于其他几种算法,实验效果显著提升。在数据集 1 中,本文算法的 F1 值比 DSEM 高出 17 个百分点。在数据集 2 中,本文算法的 F1 值比 DSEM 高出 15 个百分点。

3.4 案例分析

本文通过案例来详细了解模型挖掘出的故事线细节。表 2 描述了算法挖掘的关于“图-154 飞机失事事件”的故事线细节,细节包括故事线中重要事件的日期、时间、地点、组织、人物以及故事线的主题和关键词。

表 2 故事线案例

日期	时间	地点	组织	人物	主题	关键词
Dec 25	5:20	Russia, Syria	defense		aircrash, accident, crash, plane	aircrash
Dec 26		Russia, St. Petersburg		Putin	aircrash, national, mourn, victim	mourning, half-mast
Dec 27		Sochi, Moscow	emercom		aircrash, plane, record, salvage	black, box, fragment, remains
Dec 28		Russia	defense		aircrash, plane, record, salvage	black, box, second, salvage
Dec 29		Moscow	committee	Sokolov	aircrash, salvage, end, victim	press, conference, investigate, end

从模型结果中可以推断出,故事线开始于 12 月 25 号早上 5:20,俄罗斯国防部所属一架飞往叙利亚的飞机发生空难;12 月 26 日,俄罗斯圣彼得堡降半旗志哀飞机失事遇难者,俄罗斯总统普京宣布 12 月 26 号为全国哀悼日;12 月 27 日,通过地点“索契,莫斯科”、组织“紧急情况部”、关键词“黑匣子,遗体,碎片”,可以推断出失事飞机的一个“黑匣子”从索契运至莫斯科,紧急情况部找到部分遇难者遗体和飞机碎片;12 月 28 日,国防部表示,飞机的第 2 个“黑匣子”已被打捞;最终,于 12 月 29 日,国家调查委员会负责人索科洛夫在莫斯科举行的新闻发布会上表示失事飞机现场的主要搜索工作结束。“图-154飞机失事事件”到此告一段落。这样,通过本文的模型,仅通过一张表格,不到 100 个单词,就从关于“图-154 飞机失事事件”的上百篇新闻报道中提取出故事脉络,总结抽象成详细的故事线,大大节省了用户的时间成本和阅读难度。

本文研究主要有以下创新点:

1)提出贝叶斯网络模型来结构化描述故事线和事件。更加合理地定义了故事线、事件、事件要素、主题、关键词之间的关系。

2)模型考虑了新闻的时效性,新闻发生的时间与新闻发布的时间有一定的时间差,部分新闻时间差甚至较大。目前的很多模型将新闻发布时间近似看作事件的发生时间。SMBN 模型对新闻发布时间和事件发生时间进行明显区分和识别,模型更加合理精确。

3)Gibbs 采样[17]在贝叶斯网络中进行参数估计具有效率高和准确率高 2 个优点。对 SMBN 模型的 Gibbs 采样公式进行推导,通过 Gibbs 采样对模型的参数进行估计。

4)本文算法在多个新闻数据集上进行测试和评估。实验结果证明,SMBN 模型明显优于 K-means、LSA 等算法。

4　结束语

本文提出一种基于贝叶斯网络的故事线挖掘算法 SMBN。该算法模型结合事件中的日期、时间、机构、人物、地点等要素,将故事线、事件、主题、关键词紧密关联起来,考虑了新闻的时效性特点。将模型应用在现实新闻数据集中取得了显著效果。下一步将结合微博、Twitter 等数据对模型进行有监督学习,使本文算法模型更加合理精确。

参考文献

[1] SHAHAF D, YANG J, SUEN C, et al. Information Cartography: Creating Zoomable, Large-scale Maps of Information [C]//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, ACM Press, 2013:1097-1105.

[2] 骆卫华,于满泉,许洪波,等. 基于多策略优化的分治多层聚类算法的话题发现研究[J]. 中文信息学报,2006,20(1):29-36.

[3] YAN R, KONG L, HUANG C, et al. Timeline Generation Through Evolutionary Trans-temporal Summarization[C]// Proceedings of Conference on Empirical Methods in Natural Language Processing. [S.l.]: Association for Computational Linguistics, 2011:433-443.

[4] 罗亚平,王　枞,周延泉. 基于关注度的热点话题发现模型[C]//中文信息处理国际会议论文集. 北京:[出版社不详],2007:212-222.

[5] MEI Q, ZHAI C X. Discovering Evolutionary Theme Patterns from Text: An Exploration of Temporal Text Mining[C]//Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. New York, USA: ACM Press, 2005: 198-207.

[6] 庄婷婷,王　平,程齐凯. 一种时间情境依赖的微博话题抽取方法[J]. 信息资源管理学报,2013(3):40-46.

[7] 路　荣,项　亮,刘明荣,等. 基于隐主题分析和文本聚类的微博客中新闻话题的发现[J]. 模式识别与人工智能,2012,25(3):382-387.

[8] 李　培,翁　伟,林　琛. 中文微博故事线生成方法[J]. 中文信息学报,2016,30(3):143-151.

[9] HUANG L, HANG Lian'en. Optimized Event Storyline Generation Based on Mixture-event-aspect Model[C]// Proceedings of IEEE EMNLP'13. Washington D. C., USA: IEEE Press, 2013:726-735.

[10] DEHGHANI N, ASADPOUR M. Graph-based Method for Summarized Storyline Generation in Twitter [EB/OL]. (2015-10-21). http://x-algo.cn/wp-content/uploads.

[11] 王　巍,杨　武,齐海凤,等. 基于多中心模型的网络热点话题发现算法[J]. 南京理工大学学报,2009,33(4):422-426.

[12] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 2003,3:993-1022.

[13] HUA T, ZHANG X, WANG W, et al. Automatical Storyline Generation with Help from Twitter [C]// Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. New York, USA: ACM Press, 2016:2383-2388.

[14] LI J, LI S. Evolutionary Hierarchical Dirichlet Process for Timeline Summarization[C]//Proceedings of IEEE ACL'13. Washington D. C., USA: IEEE Press, 2013: 556-560.

[15] ZHOU D, XU H, HE Y. An Unsupervised Bayesian Modelling Approach for Storyline Detection on News Articles [C]//Proceedings of IEEE EMNLP'15. Washington D. C., USA: IEEE Press, 2015:1943-1948.

[16] ZHOU D, XU H, DAI X Y, et al. Unsupervised Storyline Extraction from News Articles [C]//Proceedings of IEEE AAAI. Washington D. C., USA: IEEE Press, 2016:145-156.

[17] GEMAN S, GEMAN D. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1984,20(6):721-741.

编辑　索书志