面向时间序列的微博话题演化模型研究

王振飞 刘凯莉 郑志蕴 王 飞
(郑州大学信息工程学院 郑州 450001)

摘 要 话题演化研究有助于追踪用户的喜好和话题的发展趋势,对于舆情预警具有重要意义。目前,话题演化方法注重运用话题生成模型实现话题演化分析,忽略了话题中时间因素和背景词的存在。以传统话题生成模型LDA为基础,将其扩展为微博话题生成模型MTLDA。MTLDA模型增加了对背景词的考虑,提高了话题生成的效率,同时对微博话题集进行时间片划分,利用KL距离计算相邻时间片话题距离,分析话题演化情况。以新浪微博数据为例进行实验,结果表明,MTLDA模型通过时间片划分完成了微博话题的生成,话题演化结果与实际情况吻合。

关键词 微博,话题演化,社交网络,MTLDA模型,KL距离
中图法分类号 TP399 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.08.046

Research on Evolution Model of Microblog Topic Based on Time Sequence

WANG Zhen-fei LIU Kai-li ZHENG Zhi-yun WANG Fei
(School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China)

Abstract *Topic evolution research is helpful to track the user preferences and development trend of topics, and it is of great significance for public sentiment warning. Current topic evolution methods focus on using topic generation model to achieve the topic evolution analysis, and ignore the time factors of topic and background word. Based on the traditional topic generation model LDA, this paper extended it to the micro-blog topic generation model MTLDA. Considering the background word, MTLDA model improves the efficiency of the topic generation. Meanwhile, the micro-blog topic set is divided into time slices, KL divergence is used to calculate the distance between adjacent time slices, and topic evolution is analyzed. Taking Sina Micro-blog data as an example, the experimental results show that the MTLDA model completes the generation of micro-blog topic by using the time slice, and the topic evolution results are tally with the actual situation.*

Keywords Microblog, Topic evolution, Social network, MTLDA model, Kullback Leibler(KL) divergence

1 引言

互联网、物联网的快速发展导致数据呈爆炸式增长,根据互联网数据中心的报告,2012年全球数据总量为2.7ZB,到2020年将达到35ZB,庞大的数据量宣告了大数据时代的到来。医疗、交通、购物等各个领域都面临着大数据时代所带来的各种挑战[1]。特别是社交网络的兴起,使得人们面临着更加庞大、复杂的数据,同时社交网络大数据成为目前研究的重点。社交网络是指人与人之间、组织与组织之间进行信息交流而形成的关系网[2]。社交网络大数据具有较强的实时性和多样性,包含人们对各种时事的看法,因此社交网络大数据成为信息爆炸时代一个亟待研究的热点。

微博因其内容的实时性、多样性和庞大的用户数目而成为热门的社交网络平台,其通过点赞、评论、转发等丰富的参与方式吸引了越来越多的使用者,微博话题的出现也将微博热度再次提升,微博用户通过微博话题实时参与各种社会现象的讨论。随着时间推移,人们对话题的关注点也会发生变化,而及时掌握不同时刻的话题关注点有助于追踪用户的喜好和掌握话题的发展趋势,同时对某时刻演化出的敏感话题给予及时控制对社会舆情预警有很大的帮助。

话题演化是对已有话题随着时间演化情况进行的分析,随着越来越多的人参与到网络话题的讨论中,众多研究者开始对网络话题演化进行研究。文献[3]利用特征计算模型,提出新闻话题演化方法,通过对已有话题文档和最新话题文档的对比来完成话题信息的动态增量扩展,有效解决了话题演化偏斜问题。崔凯等人利用文本生成模型LDA实现在线主题演化的挖掘,将文本进行时间片划分,建立并实现了在线LDA模型,并利用KL相对熵(Kullback Leibler)来衡量主题

到稿日期:2016-07-12 返修日期:2016-10-01 本文受郑州大学新媒体公共传播学科招标课题阶段性成果(XMTGGCBJSZ11),河南省科技攻关项目(142102310531)资助。

王振飞(1973-),男,博士,副教授,CCF会员,主要研究方向为社交网络、大数据;刘凯莉(1991-),女,硕士生,主要研究方向为社交网络、数据挖掘;郑志蕴(1962-),女,博士,教授,主要研究方向为分布式计算、智能信息处理;王 飞(1987-),男,硕士,主要研究方向为无线传感网,E-mail:iezfwang@zzu.edu.cn(通信作者)。

之间的相似度，从而发现主题演化中的主题遗传和主题变异[4]。胡艳丽等人将文本生成模型改进扩展为 OLDA(Online Latent Dirichlet Allocation)，用模型抽取各时间片包含的子话题，通过 Gibbs 抽样对话题模型参数进行估计，对子话题进行关联分析，并定义子话题产生、消亡、继承、分裂和合并 5 种演化类型，完成对话题演化的描述[5]。方莹等人对 LDA 模型进行改进，利用改进后的 ILDA 模型完成面向动态主题数的话题演化分析[6]。文献[7]实现了一种微博转发预测的 MTERR 算法，将微博的转发特性和时间属性考虑在内，构建话题关联函数，生成话题演化拓扑图。国外学者对话题演化也进行了深入研究。M. Jayashri 等人分析了文本文档的数据流趋势，使用话题检测和跟踪的方法来跟踪生成每个主题在到达时间周期的演变[8]。S. Jensen 等人使用受限制的元路径构造一棵基于网络的可视化的主题演化树，为研究人员提供了科学主题并在他们感兴趣的上下文中演化[9]。Y. Jo 等人研究了带时间戳文档集合的主题演化规律。其设计了独特的、捕捉主题语料库内部的拓扑结构，而不是描述固定的时间点的演化主题，该方法实现了非均匀分布话题随着时间推移的演化图[10]。

上述研究没有将时间因素和背景词同时考虑，本文基于 LDA 主题模型和 KL 距离提出微博话题生成模型(Microblog Topic Latent Dirichlet Allocation，MTLDA)来实现对话题演化的分析。MTLDA 模型同时考虑了时间和背景词因素，增加了对背景词的处理，提高了主题发现的效率。将带有时间戳的微博话题内容作为研究对像，发现每个时间片中的话题，并通过 KL 距离计算相邻时间片间话题的相似情况，对微博话题随时间的演化进行准确描述。

2　微博话题演化分析

将预处理后的微博话题数据集输入 MTLDA 模型从而得到话题及词的分布，计算相邻时间片微博话题的 KL 距离，通过 KL 距离与阈值的比较来分析微博话题演化。

2.1　文本生成模型 LDA

LDA(Latent Dirichlet Allocation)是一种文档主题生成模型，是 Blei 等人于 2003 年提出的基于概率模型的主题模型算法，它以概率分布的形式给出文档集中每篇文档的主题[11]。通过对文本建模并进行主题分类来判断相似度。LDA 将文本映射到主题空间，即认为一篇文章由若干主题随机组成，从而获得文本间的关系。LDA 文本生成过程如下，其中包含两个分布：*Dir*(表示狄利克雷分布)和 *Mult*(表示多项式分布)。

(1)对每一篇文档 d_i，利用 $\theta_i \sim Dir(\alpha)$ 获得“文档-主题”分布 θ_i，θ_i 作为多项式分布的参数。

(2)对于文档 d_i 中的每个词 w_i：

1)利用 $z_i \sim Mult(\theta_i)$ 获得一个 topic 主题；

2)从 $\varphi_k \sim Dir(\beta)$ 中获取单词的“主题-词”分布 φ_k，φ_k 作为多项式分布的参数；

3)利用 $w_i \sim Mult(\varphi_k)$ 获取主题 z_i 下的 word 词。

(3)重复上述过程直至遍历完文档中的每一个单词。

对于 D 个文本，T 个主题，W 个词汇，文本中第 i 个词汇 w_i 可表示为：

$$p(w_i)=\sum_{j=1}^{T}p(w_i \mid z_i=j)p(z_i=j) \tag{1}$$

其中，z_i 表示词 w_i 所属的主题，$p(z_i=j)$ 表示主题 j 属于当前文本的概率，$p(w_i \mid z_i=j)$ 表示词汇 w_i 属于主题 j 的概率。

令 $\theta_w^{z=j}=p(w_i \mid z_i=j)$ 代表编号为 j 的主题词分布，令 $\varphi_{z=j}^{d}=p(z_i=j)$ 表示文本 d 上编号为 j 的主题分布，则文本 d 中词汇 w 的概率分布为：

$$p(w \mid d)=\sum_{j=1}^{T}\theta_{(w)}^{z=j}\times\varphi_{z=j}^{(d)} \tag{2}$$

LDA 的贝叶斯网络图如图 1 所示。其中，阴影圆圈表示可观测变量(observed variable)，非阴影圆圈表示潜在变量(latent variable)，箭头表示两变量间的条件依赖性(conditional dependency)，方框表示重复抽样，重复次数显示在方框的右下角。

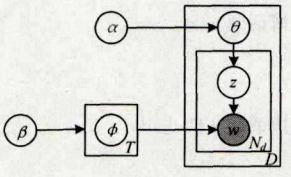

图 1　LDA 贝叶斯网络图

M 表示训练语料中的文章数，K 表示设置的主题数，V 表示训练语料库中出现的所有词的词表，θ 是一个 $M\times K$ 的矩阵，θ_m 代表第 m 篇文章的主题分布，φ 是一个 $k\times v$ 的矩阵，φ_k 代表编号为 k 的主题之上的词分布，α 是每篇文档的主题分布的先验 Dirichlet 分布参数(也被称为超参数)，其中 $\theta\sim Dir(\alpha)$，β 是每个主题的词分布的先验 Dirichlet 分布参数(也被称为超参数)，其中 $\varphi_k \sim Dir(\beta)$，$W$ 是可被观测的词。

2.2　微博话题演化方法

本文以新浪微博话题为研究对象，其作为国内影响最大的社交网络平台，截止 2015 年用户量已经达到 2.22 亿。在新浪微博中两个“#”之间的文字被定义为某个用户的微博话题，用户可以参与自己感兴趣的话题，也可以自己创建话题。话题随着时间或衍生出新的话题，或消亡，或持续不变，称此现象为微博话题的演化。通过对微博话题演化的分析可以更好地预测出话题的走向，以便对可能出现的负面现象采取必要的措施。

2.2.1　预处理

(1)去除停用词。将出现频率高、没有太大检索意义的词定义为停用词。抓取参与同一微博话题的用户所发表的微博评论并将其组合成一个文档，使用停用词表去除微博话题文档中的停用词。

(2)对微博话题文档进行分词。采用中国科学院计算技术研究所研制的汉语词法分析系统(Institute of Computing Technology Chinese Lexical Analysis System，ICTCLAS)进行微博数据分词。

(3)剔除垃圾用户发布的微博。结合用户发布微博的周期频率、提及其他用户的比例、包含 URL 的比例、用户好友数目及其粉丝数目的比例这 4 个因素来判断其是否为垃圾用户[12]。

2.2.2 微博话题获取模型

考虑到微博话题去除停用词之后仍存在一些背景词，本文对 LDA 模型改进为微博主题模型(MTLDA)来实现对微博话题演化的分析。将对某微博话题的背景信息进行描述的词定义为背景词，背景词在微博话题中大量重复出现，从而影响微博话题获取的效率。形成微博主题模型的过程为：首先对于每一个微博话题文档，按照主题分布抽取一个主题；然后按照主题词分布计算每个词属于某主题的概率，若该词是背景词，则选择背景词超参数作为主题词分布的超参数；最后选取概率最大的 n 个词来描述主题。微博话题演化模型算法的描述如算法 1 所示。

算法 1 微博话题主题分析算法

Begin

输入：超参数 $\alpha,\beta,\gamma,\beta_1$

输出：主题词概率

1. $\pi\sim Dir(r),\Omega\sim Dir(\beta_1)$

2. For 每一个微博话题 z_i，do

$\theta_i\sim Dir(\alpha)$

End for

3. For 每一个微博话题文档 d_i，do

$\varphi_k\sim Dir(\beta)$

$z_{m,n}\sim Mult(\theta_m)$

For 微博文档中的每一个词 w_i，do

$Y\sim Bernoulli(\pi)$

If Y=1，do

$w_{m,n}\sim Mult(\Omega)$

Else do

$w_{m,n}\sim Mult(\varphi_{z_{m,n}})$

End for

End for

End

算法 1 通过 3 个步骤来完成微博话题的获取。首先，设定超参数，确定狄利克雷分布 $\pi\sim Dir(r),\Omega\sim Dir(\beta_1)$。$\pi$ 同时存在为算法步骤 3 中二项分布的超参数；利用 $\Omega\sim Dir(\beta_1)$ 来计算背景词所属主题的概率。步骤 2 中，针对每一个微博话题，按照狄利克雷分布得出“微博文档-主题”的分布 $\theta_i\sim Dir(\alpha)$。步骤 3 中，确定“主题词”的分布 $\varphi_k\sim Dir(\beta)$，并根据步骤 2 的“微博文档-主题”分布，按照 $z_{m,n}\sim Mult(\theta_m)$ 获取一个主题，之后循环遍历每一个词，运用步骤 1 中的超参数 π 来计算并判断背景词的值 $Y\sim Bernoulli(\pi)$，若 $Y=1$ 则为背景词，对背景词按照 $w_{m,n}\sim Mult(\Omega)$ 获得每个话题对应的词；否则为非背景词，按照 $w_{m,n}\sim Mult(\varphi_{z_{m,n}})$ 获取主题词。通过算法 1 获得微博话题文档的主题。

微博话题提取模型的概率图如图 2 所示。图中方框表示重复抽样，重复次数显示数在方框的右下角，其中 T 为话题个数，N_d 为第 d 个文档的单词个数，D 表示文档数。β 和 β_1 是每个主题下词的多项分布的 Dirichlet 先验参数，α 是每个文档下主题的多项分布的 Dirichlet 先验参数，r 是判断背景词的 Y 值的二项分布的 Dirichlet 先验参数。隐含变量 φ 和 θ 分别表示第 m 个文档下的 Topic 分布和第 k 个主题下词的分布。$z_{m,n}$ 是第 m 个文档中第 n 个词的主题，$w_{m,n}$ 是 m 个文档中的第 n 个词。π 是服从参数为 r 的 Dirichlet 分布，Y 是服从 π 的伯努利分布，用来判断是否为背景词。若 $Y=0$，则从参数 φ 的多项分布中抽取主题下的词；否则 $Y=1$，该词是背景词，从参数 Ω 的多项式分布中抽取主题下的词。

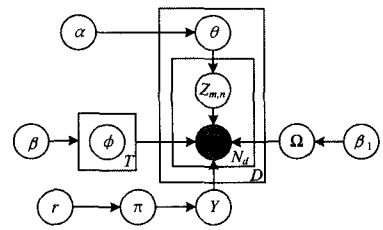

图 2 微博话题获取模型概率图

由于 LDA 中的变量 θ_m 和 φ_k 都是未知的隐含变量，因此需要根据观察到的文档集合中的词来学习估计。本文采用 Gibbs sampling 间接求得 θ_m 和 φ_k 的值[13]。

2.2.3 KL 距离

设定时间片长度为 1 天，将抓取到的微博内容进行时间片的划分。对于每一个时间片内的微博内容，运用 LDA 模型来获得当前时间片的 k 个话题。将微博话题的演化定义为在不同时间片中微博主题的演化过程。对于相邻的时间片循环，使用 KL 距离计算其话题的演化。

KL 距离也称相对熵，用于衡量相同事件空间里两个概率分布的相似情况[14]。本文采用 KL 距离来对相邻时间片的微博话题进行衡量。设 $Z_1=\{w_{11},w_{12},\cdots,w_{1n}\}$ 和 $Z_2=\{w_{21},w_{22},\cdots,w_{2n}\}$ 是两个相邻时间片中的子话题，$P(i)$ 是子话题 Z_1 中第 i 个词的概率分布，$Q(i)$ 是子话题 Z_2 中第 i 个词的概率，则两个话题的 KL 距离为：

$$D(P\parallel Q)=\sum_i \ln\left(\frac{P(i)}{Q(i)}\right)P(i) \tag{3}$$

由式(3)可知，若 $P(i)$ 和 $Q(i)$ 两个概率分布越接近，则两个话题的 KL 距离越小，说明两个话题越相近。两个相邻的时间片中，若上一个时间片 t_{i-1} 的话题与下一个时间片 t_i 的所有子话题之间的 KL 距离都大于给定的阈值，则定义为新话题产生。若同时存在 t_{i-1} 的话题与 t_i 的所有子话题的 KL 距离大于给定的阈值以及小于给定的阈值的情况，则定义为子话题分裂。

3 实验分析

3.1 微博话题数据处理

本文采用八爪鱼爬虫工具，以新浪微博热门话题板块的数据为原始数据集，抓取参与话题讨论的用户的用户名、发布微博的内容、发布时间作为数据集元素。数据集选择 2015 年 7 月到 2016 年 7 月之间比较热门的 30 个话题，抓取所有话题的内容以及微博用户参与讨论的内容。

其中，微博数据包括原创微博、转发微博、提及他人等几种形式。在采集数据时，对于不同形式的微博，需要抓取的内容各不相同。对于原创微博，直接抓取发布的内容作为微博话题数据。对于转发其他用户的微博，若转发的同时也对原发布内容进行了评论，则抓取评论内容作为实验所需数据；若转发但未评论，则仍抓取原发布内容。对于提及其他用户的微博内容，抓取除所提及用户名之外的内容。

按照去除停用词、分词、剔除垃圾用户的方法对微博话题数据集进行预处理。实验共抓取微博数据 211393 条，去除垃圾数据 3358 条，有效实验数据共 208035 条，包括 30 个微博

热门话题。首先将预处理过的微博话题内容进行时间片划分，时间片长度为 1 天；然后将所有时间片中的微博话题内容输入到 MTLDA 模型中进行微博话题提取；最后计算相邻时间片的微博话题的 KL 距离，判断微博话题的演化情况。

3.2 主题提取结果分析

设置 MTLDA 模型参数为 $\alpha=1,\beta=0.01,\beta_1=0.01,T=30$。实验中每个主题下的关键词取 10 个，通过 MTLDA 模型计算各主题的主题词概率，并将主题词概率按照从大到小的顺序排序，提取前 10 个关键词作为话题描述。将提取到的微博话题划分成 10 个时间片，选择第一个时间片段的微博话题来完成主题提取结果的分析。图 3 示出了其中 5 个话题的描述。根据各个主题对应的关键词可以看出，Topic1 是关于“该不该赊借地铁票”的主题，Topic2 是描述“薛之谦搞笑背后的深情”的主题，Topic3 是关于“生病未让座被骂快滚”的主题，Topic4 是关于“地铁 5 号线小偷被暴揍”的主题，Topic5 是“鹰爸开学堂，培养 13 岁上清华的神童”的主题。将主题提取结果与人工标注的结果进行比较，MTLDA 微博主题提取模型的效果与人工标注真实情况基本一致。

Top1	Top2	Top3	Top4	Top5
税	薛之谦	老人	小偷	孩子
素质	段子手	倚老卖老	打得好	内容实用
地铁	温柔以待	道德绑架	报警	清华
公道	喜欢	孩子	以暴制暴	幸福
照片	自己长大	生病	人权	因材施教
职责	坚强	让座	勇敢	童年
遵守	心疼	尊重	冤枉	社会关系
陌生人	深情	义务	抓错	倒霉
支付宝	善待	讨厌	暴力	适应社会
讨厌	世界和平	反驳	倒霉	压抑天性

图 3　某个时间片的 5 个话题的描述

3.3 对比实验

对比 MTLDA 模型和传统的主题生成模型 LDA，采用 Perplexity 指标对结果进行评估。Perplexity 是一种信息理论的测量方法，一个量 b 的 Perplexity 值定义为基于该量熵的能量（b 可以是一个概率分布，或者概率模型），通常用于概率模型的比较[15]。Perplexity 指标的值越小表示其性能越好。Perplexity 的定义为：

$$Perplexity(W)=\exp\{-\frac{\sum_m \ln p(w_m)}{\sum_m N_m}\} \tag{4}$$

其中，W 表示文本集，w_m 表示文本集中的第 m 个词，N_m 表示文本集中词的数量。

实验设定在相同的迭代次数下将传统文本生成模型 LDA 和微博主题生成模型 MTLDA 的 Perplexity 指标进行比较，结果如表 1 所列。

表 1　LDA 和 MTLDA 模型的 Perplexity 值

迭代次数	LDA 的 Perplexity 值	MTLDA 的 Perplexity 值
50	7083.4	6976.9
150	6460.2	6255.6
250	6057.1	6031.7

由表 1 可以看出，在迭代次数逐渐增大的过程中，相对于传统的文本生成模型，MTLDA 模型的 Perplexity 值一直处于较小的水平，说明相对于传统主题生成模型，MTLDA 模型有较好的性能。

3.4 话题演化结果分析

通过计算相邻时间片间话题的 KL 距离来描述某话题随着时间变化的话题演化。根据研究，本文定义一个给定的 KL 距离阈值，当计算出的相邻时间片话题的 KL 距离大于给定的阈值时，称其为新话题的产生。下面针对话题“‘鹰爸’开学堂，培养 13 岁上清华的‘神童’”，给出各时间片间的 KL 距离，如图 4 所示。表 2 分别列出了各时间片中的话题词。

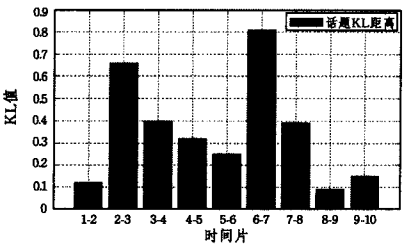

图 4　话题 KL 距离直方图

表 2　指定话题演化

时间片	话题描述
1	孩子、清华、内容实用、幸福、因材施教、童年、社会关系、倒霉、适应社会、压抑天性
2	孩子、童年、清华、教育、倒霉、循序渐进、悲哀、特殊、情商、快乐
3	孩子、幸福、不切实际、童年、拔苗助长、社交、脱离社会、清华、朋友、飞机
4	天才、学校、疯狂、道具、开心、拔苗助长、童年、人际交往、可怜、教育
5	孩子、催化童年、目标、个性、适合、快乐、喜欢、拔苗助长、社交、尊师重教
6	孩子、炒作、拔苗助长、快乐、意见、学校、教育、社交、童年、情商
7	学校、教育、问题、情商、童年、同学、群体生活、强迫、长大、拔苗助长
8	应试教育、弊端、快乐、政治性、改善、基本功、社会、反思、拔苗助长、剥夺
9	孩子、成熟、拔苗助长、性格、闭门造车、学校、教育、社会、人际交往、全面发展
10	孩子、培养、适合、愿意、童年、拔苗助长、清华、倒霉、学校教育、社会生活

通过 KL 距离图和话题描述表可以看出该话题的演化情况。在话题刚出现时，话题的重心主要集中在对“鹰爸”的教育和孩子意愿的讨论上；当话题发展到第 3 个时间片时，开始出现新的话题，即孩子将来是否会脱离社会；而在话题发展到第 7 个时间片时，出现基于现代教育的讨论的新话题。

结束语　在传统文本话题生成模型 LDA 的基础上，提出适合微博话题发现的 MTLDA 模型，该模型将背景词考虑在内，极大地提高了话题抓取的效率。将抓取到的微博文本按照时间片划分为 10 个阶段，并通过 KL 距离计算相邻时间片之间的相似程度，完成话题演化分析。通过实验比较可以看出，MTLDA 模型较 LDA 模型有更好的性能。并且本文的话题演化方法相比于实际话题演化具有更高的准确性。

本文从微博话题内容方面对话题演化进行分析，分析的结果无法清晰地显示话题的演化强度变化，下一步将对微博话题演化强度进行深入研究。

（下转第 279 页）

2013,54(1):149-165.

[16] LIU B X,LI Y. Construction Principles and Algorithms of Concept Lattice Generated by Random Decision Formal Context[J]. Computer Science,2013,40(6A):90-92. (in Chinese)
刘保相,李言. 随机决策形式背景下的概念格构建原理与算法[J]. 计算机科学,2013,40(6A):90-92.

[17] WEI L,Q J J,ZHANG W X. Attribute Reduction of Concept Lattice in Decision Formal Context[J]. Science in China (Series E):Information Science,2008,38(2):195-208. (in Chinese)
魏玲,祁建军,张文修. 决策形式背景的概念格属性约简[J]. 中国科学 E 辑:信息科学,2008,38(2):195-208.

[18] LI J H. Rule Acquisition Oriented Reduction Methods for Concept Lattices and Their Implementation Algorithms [D]. Xi'an:Xi'an Jiaotong University,2012. (in Chinese)
李金海. 面向规则提取的概念格约简方法及其算法实现[D]. 西安:西安交通大学,2012.

[19] GANTER B,WILLE R. Formal Concept Aanlysis [M]. Mathematical Foundations. New York:Springer-Verlag,1999.

[20] 张文修,仇国芳. 基于粗糙集的不确定决策[M]. 北京:清华大学出版社,2005:185-205.

[21] 张文修,梁怡. 不确定性推理原理[M]. 西安:西安交通大学出版社,1996:56-76.

(上接第 273 页)

参 考 文 献

[1] REN L,DU Y,MA S. Visual Analytics Toward Big Data[J]. Journal of Software,2014,25(9):1909-1936. (in Chinese)
任磊,杜一,马帅. 大数据可视分析综述[J]. 软件学报,2014,25(9):1909-1936.

[2] XU J,WANG G Y,YU H. Review of Big Data Processing Based on Granular Compu-ting[J]. Chinese Journal of Computers,2015,38(8):1497-1517. (in Chinese)
徐计,王国胤,于洪. 基于粒计算的大数据处理[J]. 计算机学报,2015,38(8):1497-1517.

[3] ZHAO X J,YANG C M,LI B. A Topic Evolution Minning Algorithm of News Text Based on Feature Evolving[J]. Chinese Journal of Computers,2014(4):819-832. (in Chinese)
赵旭剑,杨春明,李波. 一种基于特征演变的新闻话题演化挖掘方法[J]. 计算机学报,2014(4):819-832.

[4] CUI K,ZHOU B,JIA Y. LDA-based Model for Online Topic Evolution Mining[J]. Computer Science,2010,37(11):156-193. (in Chinese)
崔凯,周斌,贾焰. 一种基于 LDA 的在线主题演化挖掘模型[J]. 计算机科学,2010,37(11):156-193.

[5] HU Y L,BAI L,ZHANG W M. Modeling and Analyzing Topic Evolution[J]. Acta Automatica Sinica,2012,38(10):1690-1697. (in Chinese)
胡艳丽,白亮,张维明. 一种话题演化建模与分析方法[J]. 自动化学报,2012,38(10):1690-1697.

[6] FANG Y,HUANG H Y,XIN X. Topic Evolutionary Analysis for Dynamic Topic Number[J]. Journal of Chinese Information Processing,2014,28(3):142-149. (in Chinese)
方莹,黄河燕,辛欣. 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] XU W,ZHAO B,JI G L. Microblog Topic Evolution Algorithm Based on Retweeti-ng Relationship[J]. Computer Science,2016,43(2):79-100. (in Chinese)
徐伟,赵斌,吉根林. 基于转发关系的微博话题演化算法[J]. 计算机科学,2016,43(2):79-100.

[8] JAYASHRI M,CHITRA P. Topic Clustering and Topic Evolution Based On Temporal Parameters[C]// International Conference on Recent Trends in Information Technology. Chennai,India:IEEE,2012:559-564.

[9] JENSEN S,LIU X Z,YU Y G. Generation of topic evolution trees from heterogeneous bibliographic networks[J]. Journal of Informetrics,2016,4(2):606-621.

[10] JO Y,HOPCROFT J E,LAGOZE C. The Web of Topics:Discovering the Topology of Topic Evolution in a Corpus[C]// WWW 2011-Session:Spatio-Temporal Analysis. Hyderabad,India:ACM,2011:257-266.

[11] ZHAO A H,LIU P U,ZHENG Y. Subtopic Division in News Topic Based on Latent Dirichlet Allocation[J]. Journal of Chinese Computer Systems,2013,34(4):732-737. (in Chinese)
赵爱华,刘培玉,郑燕. 基于 LDA 的新闻话题子话题划分方法[J]. 小型微型计算机系统,2013,34(4):732-737.

[12] DING Z Y,ZHOU B,JIA Y. Detecting Spammers with a Bidirectional Vote Algorithm Based on Statistical Features in Microblogs[J]. Journal of Computer Research and Development,2013,50(11):2336-2348. (in Chinese)
丁兆云,周斌,贾焰. 微博中基于统计特征与双向投票的垃圾用户发现[J]. 计算机研究与发展,2013,50(11):2336-2348.

[13] CAI G Y,PENG L B,WANG Y. Topic Detection and Evolution Analysis on Microblog[C]// International Federation for Information Processing. Trondheim,Norway:2014:67-77.

[14] ZAHO B,XU W,JI G L. Discovering Topic Evolution Topology in a Microblog Corpus [C]// Third International Conference on Advanced Cloud and Big Data. YangZhou,JiangSu,China:CBD,2016:7-14.

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

[16] CAO J P,WANG H,XIA Y Q. Bi-path Evolution Model for Online Topic Model Based on LDA[J]. Acta Automatica Sinica,2014,40(12):2877-2886. (in Chinese)
曹建平,王晖,夏友清. 基于 LDA 的双通道在线主题演化模型[J]. 自动化学报,2014,40(12):2877-2886.