图书情报工作
第58卷 第5期 2014年3月

基于LDA模型和微博热度的热点挖掘*

唐晓波 向坤

[摘 要] 分析传统LDA模型在进行微博热点挖掘时所得概率结果抽象且难以结合实际解释的缺点;考虑到微博本身的数据特点和信息论中信息量的观点,提出微博热度的概念,并将其引入到LDA模型的热点挖掘研究中,构建基于微博热度的LDA模型;通过API采集微博数据上的实验,证明新方法与旧方法具有相同的性能,而且能得到更直观的微博热度表,并得出更具有说服力的挖掘结论。

[关键词] LDA 微博热度 主题模型 热点挖掘
[分类号] G203
DOI: 10.13266/j.issn.0252-3116.2014.05.010

1 引 言

随着Web 2.0的互联网产品模式的迅速发展,微博这一新时代的互联网社交网络工具也越来越深入到人们的生活之中。中国的微博从2007年发展至今,已经拥有上亿用户群。不仅是个人,越来越多的机构,包括政府部门、企业单位、社会群体等都建立了自己的官方微博,以此来发表自己的观点和最新动态等信息。新浪微博自2009年8月推出以来,截至2012年12月底,其注册用户已超过5亿人,日活跃用户达到4 620万人,用户每日发博量超过1亿条。正是因为微博有着如此庞大的用户群和信息量,而且对人们生活的方方面面都具有深刻影响,所以对微博信息作数据挖掘,以发现其中有价值的热点信息也愈加显得迫切和意义重大。

对于微博文本的挖掘,应用主题模型是一个很好的方法。主题模型相对于传统的文本挖掘方法,能够高效地完成一些基本的工作,如发掘出文本的潜在关系、判断关联性、分类等。但微博文本的挖掘,面临很多困难。微博用户之间具有关注和被关注的关系,微博本身具有转发、评论的关系,由此形成了庞大、复杂的网状社会网络,而且微博通常是由少于140字的短文本组成,所含信息较少,各种网络用语导致的噪声较大,语义结构不规范,在进行文本挖掘时形成的文本矩阵极为稀疏,维度非常高,易导致维数灾难,所以常规的分析方法并不适用。

由于微博信息构成的文本矩阵的稀疏性和高维性,本文使用基于潜在语义分析的文本挖掘方法来进行微博主题的挖掘,主要使用LDA模型——一种基于潜在狄利克雷分布的主题生成模型。在这个主题模型中,一系列主题以服从多项式分布的形式生成每个文本,再从这些主题中同样以服从多项式分布的方式抽样出每个单词,由此构成该模型围绕主题生成文本的过程。

在传统的LDA主题模型[1]中,分析计算的基数是词频。词频可以被看作是微博的一项元数据特征,而微博具有多项元数据[2],在微博这个社会网络环境下,该方法缺乏对于其他元数据的考虑,如微博的评论、转发等元数据。这样所得主题模型的最终某一主题下的词项的分布仅仅只从语义上表示了词的出现概率,不能充分体现出某一主题下人们所关注的信息,即具有高热度的词。本文中所提的热度,逻辑上的概念指的是人们的观点、话题或者是某一词受关注的程度,从信息论的角度作出的解释是微博所包含的信息量。研究基于微博热度的LDA主题模型正是基于此考虑:将微博热度作为分析计算的基数,由此能得到微博主题热度的一个分布,而不是原始的LDA主题模型通过词频分析得出的主题分布。当人们能够直观地看到微博中相关主题的热度和主题下相关词的热度时,便能更简单地发现高热度的主题和词。

本文所做的主要工作如下:基于微博的评论数、转发数等特征量,构建描述微博热度的模型;将微博热度

* 本文系国家自然科学基金项目"社会化媒体集成检索与语义分析方法研究"(项目编号:71273194)研究成果之一。
[作者简介] 唐晓波,武汉大学信息系统研究中心教授,博士生导师;向坤,武汉大学信息系统研究中心硕士研究生,通讯作者,E-mail:350109583@QQ.com。
收稿日期:2014-01-19 修回日期:2014-02-14 本文起止页码:58-63 本文责任编辑:王善军

作为 LDA 模型的分析基数，由此得到基于热度的微博主题分布、微博主题热度表和微博热词表；将该结果与传统 LDA 模型得到的主题分布做对比，总结分析使用基于热度的 LDA 模型的优势。

2 理论基础

2.1 LDA 模型原理

LDA 模型的主要思想是将每篇文章看作所有主题的一个混合概率分布，而将其中的每个主题看作在单词上的一个概率分布。由此，当有 D 篇文档、T 个主题和 W 个单词时，在一篇文档中的第 i 个单词的概率可以表示为：

$$P(wi) = \sum_{j=1}^{T} P(wi|zi=j)\,P(zi=j) \quad (1)$$

在 LDA 模型中，参数 z 代表主题，参数 w 代表单词，则式(1)中 $P(zi=j)$ 表示的是从文档中取出一个单词属于主题 j 的概率，而 $P(wi|zi=j)$ 代表的是当取出单词属于主题 j 时该单词为 i 的概率。可以将 $P(zi=j)$ 表示为文档在主题上的一个多项分布，记为 $\theta_j^d = P(z=j)$，将 $P(wi|zi=j)$ 表示为主题在单词上的一个多项分布，记为 $\varphi_w^j = P(w|z=j)$。

在以上生成式文档思想加入 Dirichlet 先验后，便得到人们熟知的 LDA 模型，其中 θ 表示文档在主题上的分布，φ 表示主题在单词上的分布，再加入 θ 和 φ 的先验分布(分别服从参数 α、β 的 Dirichlet 分布)，这样就能得到 LDA 模型的各层参数之间的依赖关系的数学表述[3]：

$$wi|zi,\varphi^{(zi)} \sim Discrete(\varphi^{(zi)})$$
$$\varphi \sim Dirichlet(\beta)$$
$$zi|\theta^{(di)} \sim Discrete(\theta^{(di)})$$
$$\theta \sim Dirichlet(\alpha)$$

LDA 模型是一个概率生成式模型，其中一条文本生成的过程如下：①对于文档 d，从 $Dirichlet(\alpha)$ 抽样得 $\theta^{(d)}$；②对于主题 z，从 $Dirichlet(\beta)$ 抽样得 $\varphi^{(z)}$；③对于每个单词 wi 及所属主题 zi，从多项式分布 θ 中抽样得 $zi = P(zi|\theta)$，从多项式分布 φ 中抽样得 $wi = P(wi|zi,\varphi)$。

2.2 常用方法

常用的对 LDA 模型的参数进行估计的方法有变分推断(variational inference)、EM(expectation maximization)算法和 Gibbs 采样(Gibbs sampling)。Gibbs 采样是一个 MCMC(Markov chain Monte Carlo)过程的抽样方法，相较于 EM 算法，此方法更易于实现，计算复杂度较小，在速度和结果上都有不弱于前两种方法的表现[4]，本文也将使用此方法，不过在引入热度原理后将对原方法进行一些修改后再使用。

3 基于微博热度的 LDA 模型

3.1 微博热度原理

在文档生成式模型中，一篇文档的产生是基于某个规则下对单词在一定概率分布下的抽样，其中的规则包括一些潜在变量的假设。在主题模型中，这些潜在变量即为文档的主题。为了拟合所假设的模型，就要根据可观察到的变量，即文档中出现的单词，来选择合适的主题分布以更好地描述这个模型。在该模型中，可观察变量所涉及到的信息即为文档中单词出现的次数。

对于文档潜在变量主题的挖掘，如果是在以上模型的思维中，只涉及到文档单词出现的次数，其潜在的意义即单纯地认为某些词出现的次数较多，便是热度高的词。在传统的文档分析中，此方法是适用的。但在微博环境下，由于用户之间的高度互联性和信息的实时互通性，某些热门的微博通常能吸引成千上万的人一起参与讨论，因此能够用来描述词热度的特征量不仅仅是单词出现的次数，更在于微博受到关注的程度，如评论数、转发数等特征量。本文基于此考虑，引入对微博热度的描述。

在信息分类中，人们常使用向量空间模型(vector space model，VSM)来描述一篇文档，并基于对文档的向量化来进行检索、分类等操作[5]。而在建立向量空间模型时，对于向量特征的选择和权重的计算是影响整个模型系统输出结果的重要因素。随着相关研究的深入，人们越来越发现单纯的词频计算方式存在缺陷，如高频的词并不一定能代表文档，低频的词并不一定不具有意义，所以人们使用逆文档频率(inverse document frequency，IDF)来给单词赋予新的权重。郭红钰通过引入信息论中的熵的定义来计算权重[6]，鲁松等对 TF-IDF 进行改进[7]，Yang Yiming 等通过计算互信息、信息增益、Chi 等其他方式来计算权重[8]，这些对不同的词项的权重计算方式的研究，从本质上来说都是为了发现一种合适的文档的向量表示方式。值得一提的是，关于在 LDA 模型中使用其他的权重计算方式的意义和作用，T. Wilson 等也作出过详细说明和论证[9]。

本文引入了微博的评论数和转发数这两个最具有代表性的微博元数据来作为表征微博热度的参数。关于热度的计算方式，通常进行热点发现研究的文章提出过相关理论，如陆铭在评价网页的热度时使用心理学中的艾宾浩斯遗忘曲线来进行计算[10]，赵迎光等使

用 TF-PDF 算法来作为文档热点词的评价指标[11]，Xu Weili 等使用自己设计的 Growth-TD 算法来评价微博热点词[12]。

本文对于热度计算方式的思考来源于信息论中自信息量的描述：一个事件信息量的大小与该信息发生的概率有关，概率小的事件所包含的信息量大，概率大的事件包含的信息量小，则事件 A 的信息量的计算式为：

$$I(A) = -logP(A) \tag{2}$$

类比于信息量的计算，假设某条微博 m 评论数为 c，转发数为 r，则此微博的热度的计算方式定义为：

$$Heat(m) = -log\frac{1}{c+r+1} \tag{3}$$

式(3)中的 $1/(c+r+1)$ 项可以理解为：在微博网络中，c 为评论人数，r 为转发人数，加上微博作者本身，便有总人数为 $(c+r+1)$ 的人群有对微博 m 的信息表示关注，而在这个人群中，作为作者发表该微博的概率便是 $1/(c+r+1)$。代入信息量的计算公式便得到这条微博的热度值。由式(3)可得，当 $(c+r)=0$ 时，$Heat(m)=0$，即该微博热度为零，同时这种评论转发数与微博热度正相关的性质也符合现实生活中对于高热度微博的认识。值得注意的是，根据信息论中的观点，此处对数以 2 为底，计算单位为比特(bit)，本模型便是通过这种方式来计算。

微博 m 的热度等于该微博中词项的热度之和，而在本文所使用的文本词袋模型中不考虑文本之间的语义关系和出现次序，忽略这些因素之后，可以认为每个词项出现的概率是等同的。

因此，当微博中有 N 个词时，单个词的热度 $Heat(w)$ 计算式为：

$$Heat(w) = \frac{Heat(m)}{N} \tag{4}$$

由图 1 可知，对每个微博关键词热度进行分析时，加入对微博本身评论数、转发数的综合处理，并在期间对于各种特征量做出预处理，可得到微博－热度矩阵。

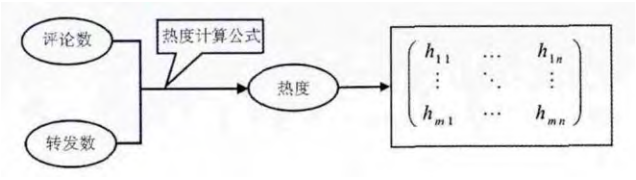

图 1　微博热度矩阵生成过程

3.2　基于热度的 LDA 模型

在引入微博热度的概念后，基于热度的 LDA 模型(heat based LDA，HLDA)的分布中 z 代表的意义变为某个主题的热度 $z'=Heat(z)$，同理 $w'=Heat(w)$ 代表某个单词的热度，总体服从的函数分布不变。以下为主题热度概率 $P(z')$ 和词的热度概率 $P(w'|z')$ 的计算方式：

$$P(z'=j) = \frac{Heat(z'=j)}{\sum_{i=1}^{T} Heat(z'=i)} \tag{5}$$

$$P(w_i'|z'=j) = \frac{Heat(w_i',z'=j)}{Heat(z'=j)} \tag{6}$$

此时，$P(z'=j)$ 表示某条微博在主题 j 上的热度概率，$P(z')$ 即表示该微博在主题热度上的概率分布；$P(wi'|zi'=j)$ 表示主题 j 中单词 i 的热度概率，$P(w'|z')$ 即表示该主题下各个关键词热度的分布。

3.3　模型的计算

根据分布的定义代入 Dirichlet 分布函数可得 $P(w'|z')$，$P(z')$：

$$P(w'|z') = \left(\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\right)^T \prod_{j=1}^{T}\frac{\prod_w \Gamma(h_j^w+\beta)}{\Gamma(h_j^*+W\beta)} \tag{7}$$

其中 h_j^w 代表主题 j 下单词 w 的热度，h_j^* 代表主题 j 的热度。

$$P(z') = \left(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\right)^D \prod_{d=1}^{D}\frac{\prod_j \Gamma(h_j^d+\alpha)}{\Gamma(h_*^d+T\alpha)} \tag{8}$$

其中 h_d^j 代表微博 d 中主题 j 的热度，h_*^d 代表在全部微博文本中微博 d 的热度。

由于上式不可直接进行计算，在采用 Gibbs 采样时，通过构造马尔科夫链进行 burn-in 次迭代后使 Gibbs 采样接近于目标分布[13]，此后的 Gibbs 采样便可以用来代表目标分布的样本值，以此来计算后验概率 $P(z_i|z_{-i},w_i)$：

$$P(z_i=j|z_{-i},w_i) \propto \frac{h_{-i,j}^{w_i}+\beta}{h_{-i,j}^{*}+W\beta}\frac{h_{-i,j}^{d_i}+\alpha}{h_{-i,*}^{d_i}+T\alpha} \tag{9}$$

上式中下标中有 $-i$，如 z_{-i} 代表在当前分配中没有分入 z_i 的单词，$h_{-i,j}^{w_i}$ 代表没有被分入主题 z_i 的单词 w_i 的热度，$h_{-i,j}^{*}$ 代表没有被分入主题 z_i 的全部单词的热度，$h_{-i,j}^{d_i}$ 代表微博 d_i 中没有被分入 z_i 的单词热度，$h_{-i,*}^{d_i}$ 代表微博 d_i 全部没有分配主题的单词热度。

由此可知，分式 $(h_{-i,j}^{w_i}+\beta)/(h_{-i,j}^{*}+W\beta)$ 代表单词 w_i 在主题 j 中的热度分布，分式 $(h_{-i,j}^{d_i}+\alpha)/(h_{-i,*}^{d_i}+T\alpha)$ 代表主题 j 在微博 d_i 中的热度分布。

在进行 Gibbs 采样时取得一部分后验分布值 $P(z|w)$，可以计算 θ、φ 的样本估计值：

$$\hat{\theta}_j^d = \frac{h_j^d+\alpha}{h_*^d+T\alpha} \tag{10}$$

$$\hat{\varphi}_j^w = \frac{h_j^w + \beta}{h_j^* + W\beta} \quad (11)$$

由于计算量是热度，因此在使用 Gibbs 采样时需要进行一些修改。Gibbs 采样通过对 LDA 中的两个关键的矩阵——文档主题矩阵、主题词矩阵与 $P(z_i \mid z_{-i}, w_i)$ 之间的循环迭代计算，当数值变化收敛时便终止，此时矩阵和各参数的值便是最终结果。修改之处在于矩阵和 $P(z_i \mid z_{-i}, w_i)$ 之间的循环迭代时，矩阵之中每次变化的不是主题或词出现的次数，而是主题或词的热度，这样便符合本文基于热度的对 LDA 模型的修改。具体模型修改环节如图 2 所示：

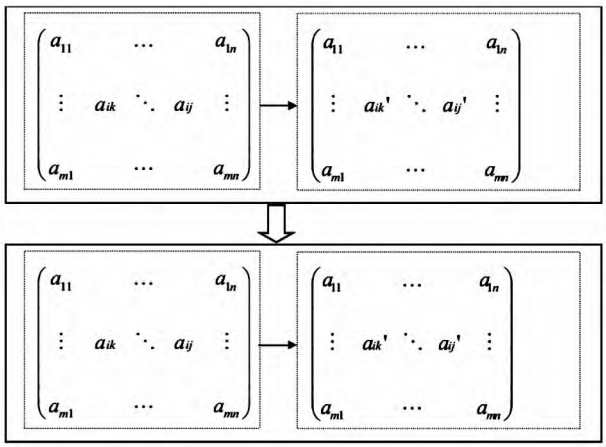

图 2 模型修改重要环节

图 2 中 m 表示微博总数，n 表示主题数，a_{ij} 表示第 i 篇微博中的词被分配入主题 j 的权重，横向过程表示的是 Gibbs 采样进行某一次对主题 z 的取样时矩阵的变化，上下两部分为旧模型与新模型之间的对比显示以突出修改之处：图上半部分表示原 LDA 模型，由于原模型中分析的权重即为词频，所以进行一次取样后 $a_{ij}' = a_{ij} - 1$，$a_{ik}' = a_{ik} + 1$；图下半部分表示基于热度的 LDA 模型对权重进行基于热度 h 的加减，其中 $h = Heat(m)/n$，Heat(m) 表示微博 m 的热度，n 表示微博 m 的词数，则 $a_{ij}' = a_{ij} - h$，$a_{ik}' = a_{ik} + h$。同时可知，当热度 h 为 1 时，新模型等同于旧模型。因此新旧模型之间在原理上并不冲突，可以把新模型看做旧模型的一个扩展，旧模型是新模型在 h 为 1 时的特例。

4 实 验

4.1 数据准备

本实验数据集通过新浪 API 在网络上自动抓取，主要采集的是各大门户网站的官方微博数据，如“央视新闻”、“环球时报”、“凤凰财经”等。在删除了一些空白噪声数据后，留下 10 000 条作为原始分析数据。为了保证最后结果的可靠性和良好的可识别性，需要事先建立一个停用词表，该词表中包括常见的语气词、助词、无意义的网络用语等。使用该词表可以在分词的过程中便剔除掉大部分的噪声词语。同时，D. M. Blei 等[1]也说明去除停用词的预处理在使用 LDA 模型过程中是必要的。

4.2 实验步骤

本实验主要通过模拟的 LDA 模型和 Gibbs 采样的 Java 程序来实现，使用 IKAnalyzer2012 这个开源分词类库来进行分词。在基于热度的 LDA 模型实验中，各个参数的设置参考了文献[3]中的说明，Topic = 10，α = 0.5，β = 0.1，循环迭代抽样的次数设为 100 次。实验的主要步骤如图 3 所示：

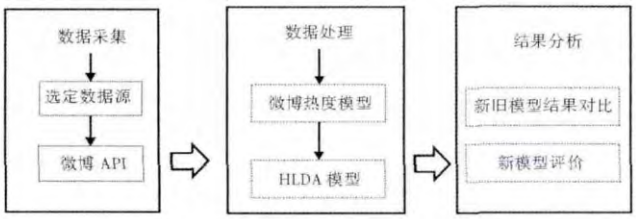

图 3 实验的主要步骤

程序设计的整体架构和主要功能如图 4 所示：

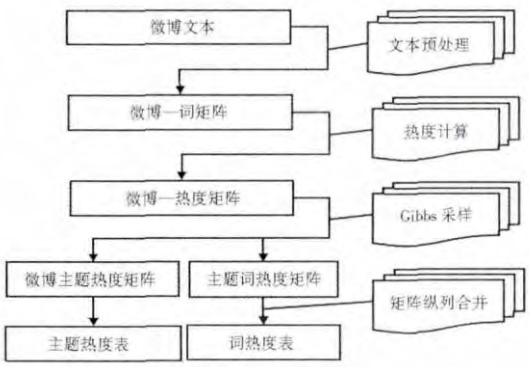

图 4 程序整体架构和主要功能

整体架构中展示了主要架构各模块中数据的由来和去向，实现的 3 个主要功能是文本预处理、热度计算和 HLDA 模型的 Gibbs 采样。

4.3 实验结果

本实验主要得到 3 个结果：

LDA 模型的主题分布和 HLDA 模型的主题分布（分别见图 5、图 6）显示了两个模型各个主题中的热词，并按照各个词的热度概率降序排列。

降序排列热度值得到微博主题热度排行，如表 1 所示：

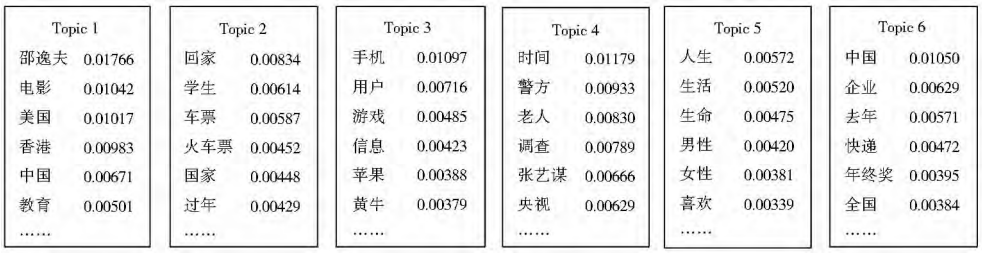

图 5　LDA 主题分布

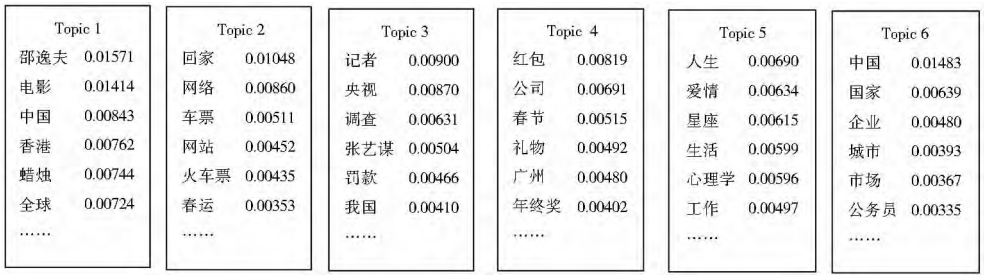

图 6　HLDA 主题分布

表 1　微博主题热度

主题	热度
Topic 1	8 561
Topic 2	7 098
Topic 3	6 897
Topic 4	6 886
Topic 5	6 625
Topic 6	6 583
Topic 7	6 533
Topic 8	6 450
Topic 9	6 229
Topic 10	5 926

降序排列生成微博热词的排行，如表 2 所示:

表 2　微博词热度

词	热度
中国	533.00
时间	323.85
邵逸夫	318.29
电影	290.70
孩子	287.99
记者	231.80
手机	212.13
话筒	191.24
公司	189.93
蜡烛	165.55
调查	164.67
北京	160.68
上海	160.57
工作	156.54
香港	153.98

4.4　实验结论

本文从微博热度的角度出发，重新分析、设计了 LDA 模型的实现过程，实现了基于热度的 LDA 模型，最后能够得到更加直观的、更易于理解分析的结果。由图 5、图 6 的对比可知，在主要的主题词分布中，HLDA 模型与 LDA 模型得到的结果差别不大，同样能够观察到各个主题下哪些词是热词(即出现概率高的词) ; 表 1 中直接显示了各个主题的热度值，这是在基于热度的 LDA 模型中得到的新结果，能以此来说明各个主题具体的热度值; 表 2 同样是基于热度的 LDA 模型得到的热度词表，能发现不同热词的热度之间同样存在差距，以前原模型得到的结论笼统地认为“邵逸夫”和“公司”都是热词，现在知道“邵逸夫”比“公司”更热门。这些结果相比于传统的 LDA 模型得到的并不太直观的概率分布的结果，无疑更容易让人理解、更具有说服力，同时新方法扩充了 LDA 模型在舆情分析、热点发现等多方面的应用。

5　结　语

本文主要从信息论的角度，诠释了新的微博热度的概念，并将此代入到 LDA 模型中进行基于微博热度的计算。本质上来说，本文是对 LDA 模型在应用上的一个扩展，并没有对 LDA 模型本身的理念进行修改。这项工作所具有的实际意义在于，通过在初始数据中赋予热度的概念，使最终结果也能用热度来表示。这样的好处不言而喻: 可以发现主题的热度、词的热度以及谁更“热”，对这些概念进行量化之后，相比较于原模型中概率的解释，无疑使人们有了更多的思路去进行更深入的研究。

本文的不足之处在于本实验中微博评论数和转发数的采集并没有考虑时间因素的影响。一般来说，微

博本身的评论数等会随着微博存在时间的增长而增长(也可能停止增长)。因此,为了尽量淡化时间因素对实验结果的影响,在采集数据时采取每隔一段时间(本实验中每隔一天)定时采集前天数据的方案。笔者在今后的研究中,将引入对时间序列的考虑,这也将是一个有价值的研究方向。

参考文献:

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

[2] 蔡淑琴,张静,王旸,等. 基于中心化的微博热点发现方法 [J]. 管理学报,2012(6):874 - 879.

[3] Griffiths T, Steyvers M. Finding scientific topics [J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(1): 5228 - 5235.

[4] Steyvers M, Griffiths T. Probabilistic topic models [J]. Handbook of Latent Semantic Analysis, 2007, 427(7): 424 - 440.

[5] 薛德军. 中文文本自动分类中的关键问题研究 [D]. 北京: 清华大学,2004.

[6] 郭红钰. 基于信息熵理论的特征权重算法研究 [J]. 计算机工程与应用,2013,49(10):140 - 146.

[7] 鲁松,李晓黎,白硕,等. 文档中词语权重计算方法的改进 [J]. 中文信息学报,2000,14(6):8 - 13.

[8] Yang Yiming, Pedersen J O. A comparative study on feature selection in text categorization [C]//Proceeding of the Fourteenth International Conference on Machine Learning(ICML'97). San Francisco: Morgan Kaufmann Publishers Inc, 1997: 412 - 420.

[9] Wilson A T, Chew P A. Term weighting schemes for latent dirichlet allocation [C]//Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Los Angeles: Association for Computational Linguistics, 2010: 465 - 473.

[10] 陆铭. Web 2.0 网络热点发现与个性化检索 [D]. 合肥: 中国科学技术大学,2012.

[11] 赵迎光,安新颖,李勇,等. 一种基于生命周期理论的文献热点发现方法 [J]. 现代图书情报技术,2012(11):86 - 91.

[12] Xu Weili, Feng Shi, Wang Lin, et al. Detecting hot topics in Chinese micro-blog streams based on frequent patterns mining [M]//Web Information Systems and Mining. Heidelberg: Springer, 2012: 637 - 644.

Hotspot Mining Based on LDA Model and Microblog Heat

Tang Xiaobo Xiang Kun

Center for Studies of Information System, Wuhan University, Wuhan 430072

[Abstract] This paper analyses shortcomings in the traditional LDA (Latent Dirichlet Allocation) model when performing microblog hotspot mining, which include that excavated probability results is abstract and is difficult to interpret. Taking into account the characteristics of the microblog and the viewpoint of the information quantity in information theory, it proposes the concept of microblog heat, introduces it into the hotspots mining research of the LDA model, and frams the LDA model based on microblog heat. With experiments on microblog data collected through API, this paper proves that the new method has the same performance compared to the old one, furthermore, it can express a more intuitive table of microblog heatand draw a more convincible conclusion.

[Keywords] LDA microblog heat topic model hotspot mining