●余传明，张小青，陈　雷　(上海理工大学　管理学院，上海　200093)

基于 LDA模型的评论热点挖掘：原理与实现*

摘　要：本文提出了潜在狄利克雷分布模型与自然语言处理技术相结合的一种挖掘用户评论热点的方法。为验证该方法的有效性，以 22 157篇餐馆评论为样本，利用 Gibbs抽样计算模型参数，获取了评论热点及相应的热点词语。实验获得的 9个主题内容较好地反映了餐馆评论中的热点，与现实生活中用户所关心的餐饮热点基本吻合，表明该模型具有较好的热点识别效果。

关键词：热点话题识别；热点挖掘；用户评论；模型

Abstract: This paper presents an approach to mining the hot topics of user comment which combines the Latent Dirichlet Allocation (LDA) model with natural language processing technologies. To verify the validity of the proposed approach, 22 157 comments on restaurants are taken as samples to obtain the hot topics of user comment and their relevant hot words by the use of the Gibbs sampling-computed parameters. The obtained 9 topics reflect the hot topics of user comments on restaurants relatively satisfactorily, and are basically consistent with what the users care about in restaurants in their real life, which shows that this model has a good effect in mining hot topics.

Keywords: topic detection; hot topic mining; user comment; model

1　热点话题识别

随着网络技术的快速发展，互联网已经成为人们日常生活中获取信息、发布信息的重要平台，对这些信息进行分析，可以及时了解社会各个领域所关心的热点。热点信息的识别一方面有助于人们获取当前重要资讯，了解社会动态及关注焦点；另一方面可以促进企业与用户之间的沟通，帮助企业有效地进行用户偏好分析，提高自身竞争能力。在这样的情况下，热点话题识别 (Topic Detection, TD) 成为了一个非常重要的研究问题[1-3]。

所谓热点话题识别，就是对一定的网络数据源进行分析，综合利用统计、聚类等方法从中识别出被用户广泛讨论的热点。依据网络数据源的不同，可以分为基于新闻媒体的热点话题识别和基于用户评论的热点话题识别。

在新闻媒体的热点话题识别方面，曾依灵等人在对语料自动分词后对切分词进行拼接获得新闻事件中的热点词语，并利用特定的噪声库与多级滤噪策略控制拼接过程[4]。该方法容易产生大量冗余词串，且没有将获得的热点词串与相关的事件或话题对应，只孤立地显示了热点词语。周亚东等人采用 DBSCAN (Density Based Spatial Clustering of Applications with Noise) 聚类算法将具有较大相关度的热点词语聚合为簇，并结合热点词语簇相关的网页标题及网站地址信息，得出网络热点话题的属性描述，避免了孤立地显示热点词语，但由于以网络数据流为研究对象，使得大量不具有语义信息的内容被当作热点识别出来，因而效果较差[5]。罗亚平等人则以新华网上的时事要闻为研究对象，采用单遍聚类 (Single-pass) 算法对话题进行聚类，该方法能够通过话题指数来描述话题在一段时间内的发展过程，但需要大量人工因素介入[6]。何婷婷等人则将语料按时间分组，对每天的语料采用凝聚聚类 (Agglomerative Clustering) 得到微类 (Macro Clustering)，再选取某个时间段内所有天数的微类，采用单遍聚类得到事件列表，对候选事件进行过滤和排序后得到最终的热点事件[7]。这种方法能自动发现任意一段时间内网络上的热点事件，并且降低了人工参与的成分。

在用户评论的热点话题识别方面，M. Oka等人提出一种利用 Blog网页上的特征词项抽取相关话题的方法[8]。该方法采用词频片段 (Frequency Segments) 表示每个特征项在一段时间内的变化情况，然后对这些特征项进行排序提取描述话题的主要特征词，并结合与其相关的共现特征词完整描述话题。戴冠中等人首先采用单遍聚类将内容相似的文档归并到同一个主题中，然后结合话题持续时间、话题文档数以及话题来源 (BBS或者其他网站) 等因素，根据 TF-PDF (Term Frequency-Proportional Document

* 本文为国家自然科学基金资助项目 (项目编号：70903047)，上海市重点学科建设项目 (项目编号：S30501，J50504) 和上海市第三期本科教育高地建设项目 (电子商务) 的研究成果之一。

Frequency) 原理计算话题关注度，继而得到热点话题[9]。

在上述背景下，笔者以用户对餐馆的评论为例，采用一种基于主题模型的无监督机器学习方法——潜在狄利克雷分布 (Latent Dirichlet Allocation, LDA) 识别用户评论中的热点以及对应的热点词语。

2 基于 LDA模型的评论热点挖掘

2.1 LDA模型

LDA模型是指由 Blei等人在 2003年提出的一个三层贝叶斯产生式概率模型，该模型假设文档是由一系列潜在主题混合而成，主题是由词汇表中所有的词汇混合而成，不同文档的主要区别在于它们的主题混合比例不同[10]。

假定给定一个评论集 R 包含 M 篇评论 $\{r_1, r_2, \cdots, r_M\}$，在 M 篇评论中分布着 K 个主题 $\{t_1, t_2, \cdots, t_K\}$，评论中的所有特征词构成一个词汇表 V，对评论集中的特征词进行词汇标记，记为 $\{w_1, w_2, \cdots, w_N\}$，词汇记号与该词所在的评论在整个评论集中的位置有关，N 为评论集中所有的词汇记号个数。词汇 w_i 在评论 r_m 中的概率可以表示为：

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

其中 z_i 是潜在变量，表示词汇 w_i 的主题序号，$p(w_i \mid z_i = j)$ 表示词汇 w_i 被分配到第 j 个主题的概率，$p(z_i = j \mid r_m)$ 表示第 j 个主题在评论 r_m 中的概率。

模型假定每篇评论是由各个主题随机混合而成，混合比例服从多项式分布，记为：

$$z_i \mid \theta^{(r_i)} \sim Multi(\theta^{(r_i)}) \quad (2)$$

每个主题又是词汇表中所有词汇的随机混合，混合比例也服从多项式分布，记为：

$$w_i \mid z_i, \phi^{(z_i)} \sim Multi(\phi^{(z_i)}) \quad (3)$$

在 LDA模型中，假定所有主题具有相同的先验概率，因此可以为多维随机变量 θ指定一个超参数为 α的对称狄利克雷先验分布[10]。由于新评论中一些特征词有可能在词汇表中没有出现过，为避免对这些特征词赋予零概率，可以为变量 ϕ 指定超参数为 β的对称狄利克雷先验分布[11]。模型假设如图 1所示。

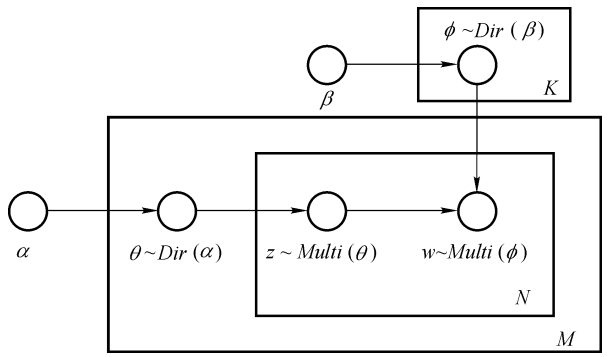

图 1 LDA模型

2.2 基于 LDA模型的评论热点识别

在本文的评论热点识别研究中，评论中的特征词是整个模型中唯一的可观察变量，LDA模型在已知主题数目的情况下，通过调节特征词语在潜在主题上的概率分布完成每篇评论的生成过程。在此过程中，可以获得每个特征词语在各个潜在主题上的概率分布情况以及每篇评论在这些潜在主题上的概率分布情况。如果一篇评论在某个潜在主题上的概率分布值越高，那么它成为该评论主题的可能性就越大，如果此潜在主题同时又是其他多篇评论的主题，那么它就可能是整个评论集中的评论热点。如果某个特征词在某个主题上的概率分布值越高，说明此特征词对该主题的贡献也就越大，也就越有可能成为该主题的热点特征词语。例如，针对评论“一进店门，就被温暖的橙色充溢着，镂空雕花的灯饰，白色的沙发软座，创意独特的隔断墙，无不让人融入一个时尚而新颖的用餐氛围。薄底的芝士披萨烘烤后上面点缀了生菜叶和烟熏三文鱼。清爽的生菜叶和三文鱼解了点芝士的腻。蛮不错的”，通过 LDA模型对整个评论语料库进行拟合，发现该评论在“环境”和“西餐”这两个潜在主题上的混合比例较大，可以判定它们是该评论的主题。同时，可以得到特征词语在这两个主题上的分布情况，反映环境的特征词语包括“氛围”、“灯饰”、“沙发”、“墙”等，反映西餐的特征词语包括“芝士”、“披萨”、“三文鱼”等。统计整个评论语料库中的主题分布情况，对各个主题出现的次数从高到低进行排序，通过设定阈值，选择排在前若干位的主题作为评论集中的评论热点，然后根据词汇表中的特征词语在这些主题上的概率分布得到所对应的热点词语。

3 实验过程与结果分析

3.1 数据预处理

实验以文献 [12] 的研究数据为基础，以餐馆评论作为研究对象。首先从点评网上抽取 300家餐馆信息和相应的 22 157条评论，存入数据库中。针对每一篇评论，进行分词和词性标注。

3.2 输入向量的构造

考虑到评论热点大多表现为名词和名词短语，因此在词性标注之后，只统计名词和名词短语的词频。通过设置频度阈值，过滤掉总频度低于 6次以及所出现的评论数少于 10篇的词汇，得到一个由 3 671个词汇所构成的特征词汇表 V。通过对评论中的词汇进行标记统计 3 671个词汇在 21 580篇评论中的分布情况，构建两个列数为 666 272的行向量 WS和 DS。在行向量 WS中，ws_i 表示序列号为 i

($i \in \{1, 2, \cdots, 666272\}$) 的词语在词汇表 V 中的序号，取值范围为 $\{1, 2, \cdots, 3671\}$；在行向量 DS 中，ds_i 表示序列号为 i ($i \in \{1, 2, \cdots, 666272\}$) 的词语所在的评论在整个评论集 R 中的位置，取值范围为 $\{1, 2, \cdots, 21580\}$。

3.3 模型求解

在输入向量构造完成后，需要对 LDA 模型中的潜在变量 θ 和 ϕ 进行求解，在这里采用了 Gibbs 抽样方法。Gibbs 抽样是马氏链蒙特卡罗 (MCMC) 的一种实现形式，它利用每个变量的条件分布实现从联合分布中抽样[13]。在抽样过程中，每个变量以固定次序从其他变量的条件分布中进行抽样，构造收敛于目标概率分布的 Markov 链，并从链中抽取被认为接近该概率分布值的样本。该方法易于实现、速度快、所需内存较小，其过程如下[11]：

1) 初始化。z_i 被初始化为 1 到 K 之间的某个随机整数，i 从 1 循环到 N，N 为评论集中所有的词汇记号个数。此为马氏链的初始状态。

2) 迭代。i 从 1 循环到 N，将词汇按照公式 (4) 分配给主题，获取马氏链的下一个状态。

$$p(z_i = j \mid z_{-i}, w_i) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + V\beta} \cdot \frac{n_{-i,j}^{(r_i)} + \alpha}{n_{-i}^{(r_i)} + K\alpha} \tag{4}$$

其中，$z_i = j$ 表示将词汇记号 w_i 分配给主题 j，z_{-i} 表示所有 z_N ($N \neq i$) 的分配，$n_{-i,j}^{(w_i)}$ 表示词汇记号 w_i 所对应的唯一性词汇被分配给主题 j 的次数，$n_{-i,j}^{(\cdot)}$ 表示分配给主题 j 的所有词汇数，$n_{-i,j}^{(r_i)}$ 表示评论 r_i 中分配给主题 j 的词汇数，$n_{-i}^{(r_i)}$. 表示评论 r_i 中所有被分配了主题的词汇数之和。

3) θ 和 ϕ 的求解。当迭代足够次数后，马氏链逐渐接近目标分布，记录 z 的当前值，得到两个矩阵："特征词—主题"矩阵和"评论—主题"矩阵。在"特征词—主题"矩阵中，每一行对应词汇表中的特征词，每一列对应各个主题，元素值代表某一特征词分配给对应主题的次数；在"评论—主题"矩阵中，每一行对应各篇评论，每一列对应各个主题，元素值代表评论中的特征词被分配给某一主题的次数。

由上述两个矩阵可以计算参数的值，如下所示：

$$\theta_j^r = \frac{n_j^r + \alpha}{n_{\cdot}^r + K\alpha} \qquad \phi_w^j = \frac{n_j^w + \beta}{n_j^{(\cdot)} + V\beta} \tag{5}$$

其中，n_j^r 表示评论 r 中分配给主题 j 的词数，$n_{\cdot}^r$ 表示评论 r 中所有被分配了主题的词数之和，n_j^w 表示词汇 w 被分配给主题 j 的次数，$n_j^{(\cdot)}$ 表示分配给主题 j 的所有词数。

3.4 实验结果与分析

由上文模型求解的过程可知，模型中存在 3 个可变量：Dirichlet 分布中的超参数 α、β 以及主题数目 K，为了有效利用 Gibbs 抽样算法，需要确定这 3 个可变量的最佳取值。本实验根据经验值确定 α 和 β 的取值，令 $\alpha = 50/K$，$\beta = 0.1$，这种取值在实验数据集中有较好的表现。主题数目 K 的最佳值采用统计语言模型中常用的评价标准困惑度 (Perplexity) 来进行选取[14]，其计算公式如下：

$$\text{perplexity}(R) = \exp\left\{ -\frac{\sum_{m=1}^{M} \log(P(r_m))}{\sum_{m=1}^{M} N_m} \right\} \tag{6}$$

其中，N_m 表示第 m 篇评论 r_m 的长度，$P(r_m)$ 表示模型产生评论 r_m 的概率，如公式 (7) 所示。

$$P(r_m) = \prod_{i=1}^{n} \sum_{j=1}^{K} P(w_i \mid z_i = j) P(z_i = j \mid r_m)$$
$$\text{where } r_m = (w_1, w_2, \cdots, w_n) \tag{7}$$

通常情况下，困惑度越低，说明模型产生文档的能力越高，模型的推广性也就越好。实验中测试了 K 为 1, 10, 20, 30, 40 和 50 的情况，困惑度的变化情况见图 2。

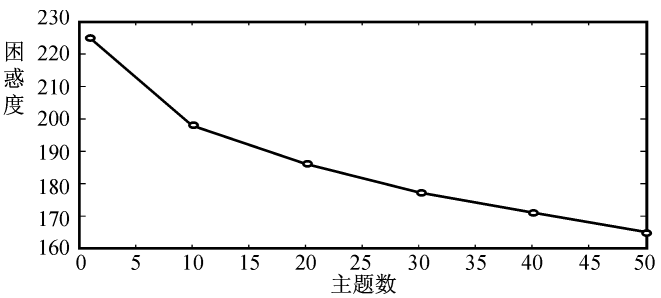

图 2 困惑度曲线

从图 2 可以看出，随着主题数的增加，困惑度值减小，当主题数目为 50 时，困惑度达到最小值，此时模型的性能最佳。因此，确定评论集的最优主题数目 K 取 50。

模型中 3 个可变量的取值确定后，结合输入向量运行 Gibbs 抽样，得到词汇表中的特征词语在 50 个潜在主题上的概率分布以及潜在主题在每篇评论中的概率分布。将每篇评论中概率值排在前两位的潜在主题看作是该篇评论的主题，统计所有评论的主题，然后按照主题出现次数的多少判定该主题是否为评论热点。本文选取出现次数排在前 9 位的主题作为评论热点，特征词语在这 9 个主题上的概率分布情况如图 3 所示，限于篇幅，图 3 中只显示其中的 3 个主题。

图 3 中 Topic 1, Topic 2, Topic 3 代表评论集中的 3 个潜在主题，每个主题下面是该主题在各个特征词上的概率分布情况，这里将主题在各个特征词上的概率分布值从大到小进行排列，列出排在前 10 位的主题特征词。在 Topic 1 中，"服务员 (0.27237)"、"时候 (0.09452)"、"态度 (0.08609)"、"老板娘 (0.08609)" 的概率值较高，可以判定该主题对应的评论热点为"服务"。在 Topic 2 中，

Topic 1		Topic 2		Topic 3	
服务员/n	0.27237	楼/n	0.14279	牛排/n	0.13690
时候/n	0.09452	位置/n	0.08642	汤/n	0.12284
态度/n	0.08609	环境/n	0.06618	餐/ng	0.08209
时/n	0.06253	地方/n	0.04750	甜品/n	0.05561
人/n	0.05386	风景/n	0.04317	主菜/n	0.05323
老板娘/n	0.05263	灯光/n	0.04224	套餐/n	0.05172
经理/n	0.03737	音乐/n	0.03823	奶油/n	0.04501
菜单/n	0.02417	窗/ng	0.03722	鹅肝/n	0.03989
时间/n	0.02240	电梯/n	0.03166	磨菇/n	0.03801
客人/n	0.02156	酒吧/n	0.02788	羊/n	0.03686

图 3 输出结果 (部分)

“楼”、“位置”、“环境”、“地方”、“风景”这些特征词的概率值较高，它们与主题“环境”密切相关，反映了“环境”这一热点。在 Topic 3中，“牛排”、“汤”、“甜品”、“主菜”等特征词都与西餐中的菜品有关，可以判定它们反映的是“西餐”这个主题。

表 1中列出了实验获得的 9个评论热点以及它们对应的前 10个特征词。

表 1 评论热点提取结果

评论热点	热点词语 (前 10个)
服务	服务员 时候 态度 时 人 老板娘 经理 菜单 时间 客人
口味	口感 味 口味 美味 入口 白色 球 香 颜色 奶
环境	楼 位置 环境 地方 风景 灯光 音乐 窗 电梯 酒吧
价格	价 钱 性 人 价钱 话 环境 价格 地方 印象
火锅	火锅 锅 牛肉 调料 店 丸子 羊肉 菇 菌 丸
海鲜	鱼 文 蚝 海鲜 鳕鱼 金枪鱼 身 扇贝 拼盘 鳗鱼
西餐	牛排 汤 餐 甜品 主菜 套餐 奶油 鹅肝 蘑菇 羊
点心	蛋糕 巧克力 莓 奶油 核桃 饼干 口感 香蕉 乳酪 手工
水果饮料	水果 芒果 布丁 草莓 果 饮料 甜品 牛奶 焦糖 冰

从表 1可以看出，9个主题的内容较好地反映了餐馆评论中的热点：服务、口味、环境、价格这 4个主题与现实生活中用户所关心的餐饮热点基本吻合；火锅、海鲜、西餐表示不同的饮食方式，反映了餐饮业中比较热门的饮食方式，它们成为实验获得的评论热点说明人们对这几种饮食方式可能比较青睐；而点心、水果、饮料是就餐时必不可少的选择，有时更体现了餐馆的特色，因此成为人们的评论热点也在情理之中。

4 结束语

综上所述，本文采用一种产生式概率主题模型挖掘用户评论中的热点，该方法不仅能自动识别出评论中的热点词语，而且可以将热点词语归类到对应的主题中，直观地反映了评论中的热点。

值得说明的是，在本文的实验中，词语切分的不准确性 (例如“性价比”被划分为“性 /ng价 /n比 /p”) 对评论热点的识别造成了一定影响。例如，在反映主题“价格”、“口味”和“海鲜”的热点词语中分别出现了“话”、“性”、“球”、“文”、“身”等语义信息不完整的词语，这些词不能明确地反映相应的主题，影响了实验效果，这将在今后的研究中进行改进。此外，如何自动对评论热点进行情感分析将是笔者下一步研究的重点。□

参考文献

[1] ALLAN J, CARBONELL J, DODDINGTON G, et al. Topic detection and tracking pilot study: final report [C] // Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop. Virginia: Lansdowne, 1998: 194-218.

[2] LEEK T, SCHWARTZ R M, SISTA S. Probabilistic approaches to topic detection and tracking [C] //Topic Detection and Tracking: Event-based Information Organization. Kluwer Academic: Massachusetts, 2002: 67-83.

[3] CHEN K Y, LUESUKPRASERT L, CHOU S C T. Hot topic extraction based on timeline analysis and multidimensional sentence modeling [J]. IEEE Transactions on Knowledge Data Engineering, 2007 (19): 1016-1025.

[4] 曾依灵，许洪波. 网络热点信息发现研究 [J]. 通信学报，2007, 28 (12): 141-145.

[5] 周亚东，孙钦东，管晓宏，等. 流量内容词语相关度的网络热点话题提取 [J]. 西安交通大学学报，2007, 41 (10).

[6] 罗亚平，王枞，周延泉. 基于关注度的热点话题发现模型 [M] //萧国政，何炎祥，孙茂松. 中文计算技术与语言问题研究. 北京：电子工业出版社，2007: 402-408.

[7] 刘星星，何婷婷，龚海军，等. 网络热点事件发现系统的设计 [J]. 中文信息学报，2008, 22 (6): 80-85.

[8] OKA M, ABE H, KATO K. Extracting topics from Weblogs through frequency segments [C] // Proceedings of the WWW2006 Workshop on Web Intelligence, 2006: 22-26.

[9] YE H, CHENG W, DAI G Z. Design and implementation of on-line hot topic discovery model [J]. Wuhan University Journal of Natural Sciences, 2006, 11 (1): 21-26.

[10] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003 (3).

[11] GRIFFITHS T L, STEYVERS M. Finding scientific topics [C] // Proceedings of the National Academy of Science, 2004.

[12] 余传明. 从用户评论中挖掘产品属性——基于 SOM的实现 [J]. 现代图书情报技术，2009 (5): 61-66.

[13] STUART G, DONALD G. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1984 (6): 721-741.

[14] CAO Juan, XIA Tian, LI Jintao, et al. A density-based method for adaptive LDA model selection [J]. Neurocomputing, 2009 (72): 1775-1781.

作者简介：余传明，男，博士，副教授，硕士生导师。
张小青，女，1986年生，硕士生。
陈雷，男，1986年生，硕士生。

收稿日期：2010 - 01 - 11