基于 PAM 概率主题模型的微博热点挖掘 *

余淼淼[1],周志平[1],赵晓东[1],岳晓冬[2]
(1.同济大学 企业数字化技术教育部工程研究中心,上海 201804;
2.上海大学 计算机工程与科学学院,上海 200444)

摘 要:针对微博本身的语言特点,提出采用 PAM(Pachinko Allocation Model)这种能够提取文本隐含主题的产生式模型,对微博的非结构化文本信息进行热点提取。采用吉布斯抽样方法计算模型参数,获取微博热点的分类分析以及关键词。在真实数据集上的实验表明,PAM 模型能够有效地对微博热点进行挖掘。

关键词:PAM;吉布斯抽样;微博;热点挖掘
中图分类号:TP18 **文献标识码**:A **文章编号**:1674-7720(2013)15-0086-04

PAM-based microblog hot spot mining

Yu Miaomiao[1], Zhou Zhiping[1], Zhao Xiaodong[1], Yue Xiaodong[2]
(1.Engineering Research Center of Enterprise Digitalization Technology, Tongji University, Shanghai 201804, China;
2.School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China)

Abstract: Targeting at solving the above problem, with considering the linguistic characteristics of microblog, a PAM-based approach is proposed to automatically recognize the hotspot from the unstructured text information of microblog. Gibbs sampling is used to calculate the PAM model parameters. And then the classified hot spot and key words can be extracted. Experimental results on the dataset show that the PAM probability topic model can offer an effective solution to hot spot mining for microblog.

Key words: PAM; Gibbs sampling; microblog; hot spot mining

随着互联网的迅速发展,微博作为一种新兴网络媒体,正处于高速发展阶段,注册用户一直在持续增长,微博已经成为互联网上最重要的信息源之一。随着用户数量的增多,微博中的话题越来越广泛,涉及政治、经济、文化以及社会等各个领域的热点。在当今信息爆炸的时代,从海量的网络信息中挖掘出有效的观点主题信息,分析主题关联显得尤其重要。如何有效地获取微博信息中的热点成为目前的一个研究方向。

微博本身是一种实时性较强的文本信息载体,包含联系人信息等结构化数据和文本信息等非结构化数据。联系人信息等结构化数据广泛应用于用户关系和社区结构分析等方面的研究。而文本信息中包含大量的最新社会网络信息,目前对这种非结构化文本信息的研究较少,因此针对微博发布内容进行热点挖掘的研究是非常有意义的。能否准确识别出具体的主题是热点挖掘的关键任务。近年来,关于文本主题挖掘的方法受到了人们广泛的关注和研究。

传统聚类算法。通过 VSM(Vector Space Model)将文本里的非结构化数据映射到向量空间中的点,再使用传统的聚类算法实现文本聚类。有基于划分的算法(如 K-means 算法)、基于层次的算法(自顶向下和自底向上算法)、基于密度的算法等[1]。聚类结果可以近似认为满足同一个主题,但这种基于聚类的算法依赖文本之间的距离,这种距离在大量的文本中是难以确定的[2]。而且,聚类的结果只起到区分类别的作用,并没有给出任何语义上的信息,不利于理解。

概率主题模型。主题模型是将主题看成是词项的概率分布,而文本则看成是主题的随机混合。它与聚类方法相比,利用词的分布,将文本信息转化成为易于建模的数字信息,具备识别大规模文本集中潜在主题信息的

* 基金项目:国家自然科学基金项目(61105047);国家科技支撑计划课题(2012BAF10B12)

能力。能够更直观地表达主题，大大简化了问题的复杂性，主题模型的应用日益广泛。

主题模型的起源是 Hofmann 在 LSI（Latent Semantic Indexing）[3] 的基础上提出的概率隐性语义索引 PLSI（probabilistic Latent Semantic Indexing）[4]，通过对训练集中的有限文档进行拟合，得到特定文档的主题混合比例，该过程导致模型参数随着训练集中文档数目线性增加，出现过度拟合现象，而且对于训练集以外的文档很难分配合适的概率。

针对这些问题 Blei 等人在 2003 年提出 LDA（Latent Dirichlet Allocation）模型[5]，它是目前应用最为广泛的概率主题模型。将每个文档表示成主题的混合，而每个主题是单词上的多项式分布。用一个服从狄利克雷分布的 K 维隐含随机变量表示文档的主题混合比例，模拟文档的生成过程。LDA 模型中发现的主题可以捕获词之间的相关性，但由于基于 Dirichlet 分布的抽样假设主题之间相互独立，不能够获取主题之间的关系。然而主题的相关性在真实的数据集合中是普遍存在的，忽略这些相关性将限制模型对大规模数据集合的表示能力以及对新数据的预测能力[6]。

随后提出的 CTM（Correlated Topic Model）模型[7]，与 LDA 类似，但它的主题混合比例是从对数正态分布中抽样获得的。CTM 只能描述成对主题间的相关性，基于这个局限性，Li 等人在 2006 年进一步提出了 PAM（Pachinko Allocation Model）模型[8]，用一个有向无环图（DAG）表示语义结构，不仅可以描述词之间的相关性，而且可以灵活地描述主题之间的相关性，较 LDA 和 CTM 具有更强的文本表示能力[6]。

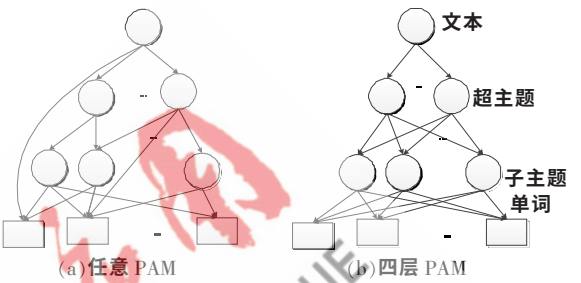
随着主题模型的不断发展，其在文本分类、信息检索以及自然语言处理等方面都有很广泛的应用。PAM 概率主题模型由于其结构的灵活性，获取语义关联的丰富性，且不易产生过度拟合现象等优点，已经成为主题模型的研究热点之一，目前在图像检索、文档分类、目标识别等方面都有一定的应用[9-12]。本文将 PAM 模型应用在微博热点挖掘领域，基于 PAM 模型和 Gibbs 抽样为微博中文本信息建模，并计算模型参数，实验得到分类热点以及关键主题词。

1 PAM 概率主题模型

1.1 PAM 模型框架

PAM 模型是 Li 等人[8]在 2006 年提出的一个具有文本主题表示能力的非监督产生式概率模型，使用一个 DAG 结构去学习和表现主题相关性，其拓扑结构可以是任意嵌套的，如图 1（a）所示。这种有向无环图结构非常灵活，可以是最基本的三层结构，也可以是任意嵌套的，节点间可以是全关联也可以是稀疏关联。

PAM 模型的命名起源是一个日本的“弹珠机”游戏，该游戏中金属球从顶部进入机器，跌入一组复杂的指针中，碰撞到指针会改变小球下落的路径，直到小球落入机器底部。这一过程可以形象地描述 PAM 模型的采样过程，如图 1（b）所示，每个叶子节点为词表中的一个词，非叶子内部节点代表一个主题，每个主题是基于它的孩子节点的狄利克雷分布，通过每个狄利克雷采样一个多项式，从根节点开始，根据多项式分布对其子节点进行采样，沿着 DAG 的路径采样直到叶子节点产生词为止。



（a）任意 PAM　（b）四层 PAM

图 1　PAM 拓扑结构

1.2 四层 PAM 模型

Li 等人[8]提出了一种特殊形式的四层 PAM（4LPAM），四层分别是指：文本、超主题、子主题、单词；如图 1（b）所示。每个矩形代表一个单词，每个圆形代表一个主题，主题上的箭头表示对其孩子的分布。文本由主题随机混合而成，主题由词汇表中所有的单词随机组合。第一层是根节点 r 代表文本；第二层有 s 个主题 $T=\{t_1,t_2,\cdots,t_s\}$，称为超主题；第三层有 s' 个主题 $T'=\{t'_1,t'_2,\cdots,t'_s\}$，称为子主题；最底层是单词。4LPAM 中，根节点与所有超主题相关联，每个超主题与所有子主题相关联，子主题与所有单词相关联。

4LPAM 的模型结构以及产生过程都和 LDA 类似，主要的差别在于另外有一个服从狄利克雷分布的超主题层，该层是获取主题间关联的关键。4LPAM 的图模型如图 2 所示。图中只有单词 w 是可观测变量，其他的变量都是隐藏变量，箭头方向表示条件概率方向，矩形表示可重复过程。4LPAM 通过总主题和超主题的狄利克雷分布选取多项式 θ。对于每个单词 w，通过 θ 选取一个超主题 z 和一个子主题 z'，单词服从的多项式分布 $\phi_{z'}$ 是和 z'，相关联的。对这种结构的另外一种解释是给出了子主题，每一个超主题实质上就是一个独立的 LDA。因此，它可以看作是一系列 LDAs 的组合。

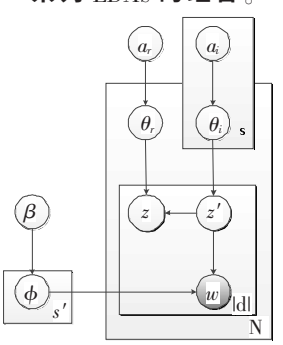


图 2　4LPAM 图模型

1.3 文档产生过程

4LPAM 中总主题服从多项式分布 $g_r(\alpha_r)$，超主题服从狄利克雷多项式分布 $\{g_i(\alpha_i)\}_{i=1}^{s}$，子主题服从固定的多项式分布 $\{\phi_{t'j}\}_{j=1}^{s'}$。文档的产生过程如下：

(1)根据总主题的分布 $g_r(\alpha_r)$ 选取 $\theta_r^{(d)}$，$\theta_r^{(d)}$ 是超主题服从的多项式分布。

(2)对于每个超主题 t_i，根据主题的分布 $g_i(\alpha_i)$ 选取 $\theta_{t_i}^{(d)}$，$\theta_{t_i}^{(d)}$ 是子主题服从的多项式分布。

(3)对于文档中的每个单词 w，

①通过 $\theta_r^{(d)}$ 选取一个超主题 z_w。

②通过 $\theta_{zw}^{(d)}$ 选取一个子主题 z'_w。

③通过 $\phi_{z'w}$ 选取单词 w。

根据以上过程，得到产生一个文档 d 的概率可以表示为：

$$P(d|\alpha,\phi)=\int P(\theta_r^{(d)}|\alpha_r)\times\prod_w\Big(\sum_{z_w z'_w}P(z_w|\theta_r^{(d)})P(z'_w|\theta_{z_w}^{(d)})P(w|\phi_{z'w})\Big)\mathrm{d}\theta^{(d)} \quad (1)$$

根据每个文档的生成概率，对子主题 ϕ 的多项式分布进行积分，可以得到整个语料库的产生概率。

2 Gibbs 抽样

模型参数估计的方法有很多，通常采用的有 Gibbs 抽样[13]、EM 算法[5]、Expectation-Propagation 方法[14]。Gibbs 抽样是 MCMC (Markov Chain Monte Carlo)[14] 的一种实现形式，最早由 GEMAN S 和 GEMAN D 讨论图像恢复时提出，其目的是从收敛于目标函数的马尔可夫链中抽取接近某概率分布的样本。Gibbs 抽样易于实现且速度较快，所以本文采用 Gibbs 抽样算法计算模型参数 α 和 β。

2.1 Gibbs 抽样方法

Gibbs 抽样是一种基于条件分布的迭代取样方法。通过总体分布的条件分布族来构建一个以该总体分布为平稳分布的马尔可夫链。在 4LPAM 中，需要对超主题和子主题的词汇分布，也就是变量 z_w 和 z'_w 进行抽样。记后验概率为：$P(z_w=t_i,z'_w=t'_j|D,z_{-w},z'_{-w},\alpha,\beta)$，计算公式如下：

$$P(z_w=t_i,z'_w=t'_j|D,z_{-w},z'_{-w},\alpha,\beta)\propto$$

$$P(w,z_w,z'_w|D_{-w},z_{-w},z'_w,\alpha,\beta)=\frac{P(D,z,z'|\alpha,\beta)}{P(D_{-w},z_{-w},z'_{-w}|\alpha,\beta)}$$

$$=\frac{n_i^{(d)}+\alpha_{ri}}{n_r^{(d)}+\sum_{i=1}^{s}\alpha_{ri}}\cdot\frac{n_{ij}^{(d)}+\alpha_{ij}}{n_i^{(d)}+\sum_{j=1}^{s'}\alpha_{ij}}\cdot\frac{n_{jk}+\beta_k}{n_j+\sum_{k=1}^{n}\beta_k} \quad (2)$$

其中 $z_w=t_i,z'_w=t'_j$ 表示将标记为 w 的单词分配给超主题 i 和子主题 j'。$n_r^{(d)}$ 是根节点 r 在文档 d 里面出现的次数，这与文件中标记的个数相等。$n_i^{(d)}$ 是超主题 t_i 在 d 中出现的次数，$n_{ij}^{(d)}$ 是 d 中抽样出的子主题 t'_j 之于超主题 t_i 的倍数。n_j 是整个语料库中子主题出现的次数。n_{jk} 是在子主题 t'_j 中单词 w_k 出现的次数。有三种类型的狄利克雷参数，α_r 是一个与根节点相连的 s 维向量，α_i 是一个与超主题 t_i 相连的 s' 维向量。β 是所有子主题的先验。$-w$ 表示除了词 w 以外的所有观测值和主题分布。

2.2 Gibbs 抽样过程

4LPAM 中 Gibbs 抽样过程：

(1)随机初始化。对于每个超主题 z_w 和 z'_w 子主题和被初始化为 $1\sim T$ 之间的随机数，得到初始马尔可夫链。

(2) 开始迭代。单词总数为 N，w 从 1 循环到 N，按照后验概率公式(2)将单词进行分配，得到马尔可夫链的下一个状态。

(3)计算 α 和 β 的值。重复执行步骤(2)到一定的次数，即马尔可夫链逐渐接近目标分布时。

3 实验及分析

3.1 数据集及预处理

本实验采用数据集的原始数据来源于 Twitter，包含了 876 名用户于 2012 年 12 月发布的 2 000 条微博。

在使用 PAM 模型进行热点提取之前，对收集的原始数据进行预处理。去除每条微博中用户联系人等结构化数据，仅保留结构化的微博正文文本信息，再对所有可用信息去除停用词。

3.2 实验设置及结果

本文的实验环境为 Pentium Dual-Core 2.50 Hz 的 CPU，2 GB 的内存，250 GB 硬盘的 PC 机。操作系统为 Window 7。超参数的设置为 α=1.0，β=0.01，子主题数为 50，超主题数为 10。

(1)获取主题及热点词汇。实验结果如表 1 所示，列举一个超主题，这个超主题下的子主题和子主题的概率分布，以及每个子主题下的 5 个热点词汇。

表 1 主题关联与词汇

超主题	子主题	主题概率分布	高频词
Supertopic0	Subtopic0	0.363 17	family, school, country, shopping, pride
	Subtopic1	0.250 14	Love, life, parents, hungry, girlfriend
	Subtopic2	0.226 03	food, lunch, restaurant, party, bread
	Subtopic3	0.208 52	heart, hand, girl, kid, children
	Subtopic4	0.176 78	cooking, holidays, son, wifi, rule

(2)列举 Supertopic0 下的 Subtopic0 里的 10 个 top 主题词，以及每个主题词的概率分布。如表 2 所示。

(3)留存测试数据的似然性。将数据集分成 75%和 25%的 2 个数据子集进行实验，对较大数据集进行建模，对较小数据集计算似然值，进行定量测度，对比 PAM 和 LDA 模型的实验结果。

PAM 在 50 个超主题时得到最好结果，受子主题数目的影响较大，因此设定超主题的个数为 50，子主题的

表 2 子主题词汇概率

高频词	概率
family	0.466 67
school	0.025
country	0.02
shopping	0.017 39
pride	0.016 67
plan	0.016 67
simple	0.016 67
officer	0.014 71
cooking	0.016 53
video	0.010 31

数目为 20 到 180 之间，采用一个基于 EL（Empirical Likelihood）的方法计算留存数据的似然值。最优主题数目如图 3 所示，PAM 的对数似然值随着主题数目的增加而增加，在 60 个子主题左右处于峰值，且始终优于 LDA 的实验结果。

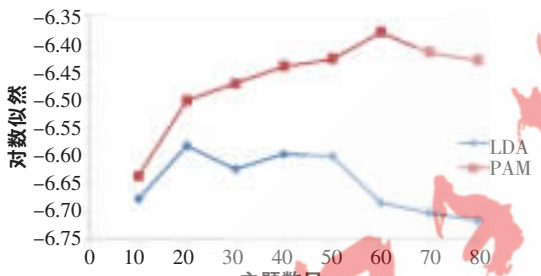


图 3 最优主题数目

综上所述，使用 PAM 概率主题模型可以有效地进行微博热点的挖掘，得到热点及其概率分布。与传统的 LDA 模型相比，PAM 模型具有更强的表达能力。

本文针对微博的特殊文本结构，提出使用基于 PAM 概率主题模型的方法，通过采用 Gibbs 抽样算法进行模型参数的计算和模型的迭代推导，从而获取微博文本信息中的隐含观点主题及相应的热点词汇。实验结果证明了该方法的有效性。

然而在使用 4LPAM 模型对微博观点主题进行挖掘的同时还存在一些问题：（1）4LPAM 模型的主题数需要人为地事先确定，不能自动得到最优的主题结构。（2）得到超主题和子主题之间的关联，不能够从子主题对应的热点词汇中抽象出具体的主题描述。因此，如何动态确定主题数目，以及如何准确根据热点词得出主题描述是下一步研究的重点。

参考文献

[1] XU R, WUNSCH D. Survey of clustering algorithms [J]. IEEE Trans on Neural Networks, 2005, 16(3):645-678.

[2] 张晨逸，孙建伶，丁轶群.基于 MB_LDA 模型的微博主题挖掘[J].计算机研究与发展，2001，48(10)：1795-1802.

[3] DEERWESTER S C, DUMAIS S T, LANDAUER T K, et al. Indexing by latent semantic analysis [J]. Journal of the American Society for Information Science, 1990,41 (6): 391-407.

[4] HOFMANN T. Probabilistic latent semantic indexing [C]. Proceedings of the 22nd Annual International SIGIR Conference. New York: ACM Press, 1999.

[5] BLEI D, NG A, JORDAN M. Latent dirichlet allocation[J]. Journal of Machine Learning Research 2003(3):993-1022.

[6] 许戈，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.

[7] BLEI D M, LAFFERTY J D. Correlated topic models [C]. Advances in Neural Information Processing Systems 18. Cambridge, MA: MIT Press, 2006.

[8] LI W, MCCALLUM A, ALLOCATION P. DAG-structured mixture models of topic correlations [C]. Proceedings of the International Conference on Machine Learning (ICML). Pittsburgh, Pennsylcania, 2006.

[9] BOULEMDEN A, TLILI Y. Image indexing and retrieval with pachinko allocation model: Application on local and global features [J]. Lecture Notes in Computer Science, 2012:140-146.

[10] BAKALOV A, MCCALLUM A, WALLACH H, et al. Topic models for taxonomies [C]. Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries. 2012.

[11] MA H, CHEN E, XU Linli, et al. Capturing correlations of multiple labels: A generative probabilistic model for multi-label learning[J]. Neurocomputing. 2012,92(9):116-123.

[12] LI Y, WANG W, GAO W. Object recognition based on dependent pachinko allocation model [J]. In: IEEE ICIP, 2007 :337-340.

[13] STEYVERS M, GRIFFITHS T. Probabilistic topic models[M]. Handbook of Latent Semantic Analysis. New Jersey: Springer, 2007.

[14] MINKA T, LAFFERTY J. Expectation-propagation for the Generative Aspect Model [C]. Proc of the 18th Conf on Uncertainty in Artificial Intelligence (UAI). [s.l.]: [s.n.], 2002.

（收稿日期：2013-04-01）

作者简介：

余淼淼，女，1989 年生，硕士研究生，主要研究方向：概率主题模型。

周志平，男，1961 年生，讲师，主要研究方向：语义信息获取。

赵晓东，男，1968 年生，高工，主要研究方向：模型可视化。