基于LDA模型的WEB文本分类

孟海涛[1],陈 思[2],周 睿[2]
(1.盐城工学院 信息工程学院,江苏 盐城 224051;2.北京大兴区第一中学国际部,北京 102600)

摘要:提出了基于LDA(Latent Dirichlet Allocation)主题模型的Web文本分类方法,利用MCMC方法中的Gibbs抽样获得模型参数从而获取词汇的概率分布,使隐藏于WEB文本内的不同主题与WEB文本字词建立关系。将LDA算法应用于WEB文本分类识别领域,在实验中与k均值聚类和贝叶斯网络方法进行了对比,其结果表明LDA与其他同类算法相比具有一定的优势。

关键词:LDA;主题模型;WEB分类
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1671-5322(2009)04-0056-03

WEB文本分类是根据面向Internet的分布式信息资源的特点的一种模式抽取过程,它不仅能查找到分布式信息资源中已存在的信息,还能识别出大量存在于数据中的隐含的、有效的规律。

WEB文本分类同传统的文本分类相比的不同之处包括:首先,Web文档的数量非常大。使得一些研究人员致力于存储Web上的数据的研究[1],因为传统的数据仓库不能满足这样巨大的数据量的存储。其次,Web文档因为数据的异构,信息来源动态更新等原因,结构十分复杂。再次,由于Web文本是一个半结构化或无结构化,又缺乏机器所能理解的语义,因此现有的文本分类方法不完全适用于WEB文本分类。因而,开发新的WEB文本分类技术以及对Web文本进行预处理,以提取该文本的特征,便成为WEB文本分类研究的重点[2]。

当前研究人员关于WEB文本分类的相关工作主要着眼于用WEB文本内容构建过滤模型,特别是机器学习、模式识别、模式分类和数据挖掘等方法的使用[3-4]。使用的算法也往往是文本分类的算法,比如支持向量机、贝叶斯网络、神经网络、聚类等。但是由于WEB文本分类与文本分类存在着很大区别,所以效果还有改进的余地。

我们进行了基于主题模型的WEB文本分类应用。主题模型的核心思想是认为一个文档是由一系列的主题分布组成的,而每个主题又是由一系列的关键词组成,区别于传统模型。主题模型强调文档是由文档—主题—关键词3层关系组成,主题模型是一种生成概率模型,可以应用于文本数据、图像、生物图像以及其它多维数据的识别、分类和数据挖掘。一个最成功的生成主题模型是Blei,Ng and Jordan发明的LDA模型,这一模型是一个完全的生成图形化模型,在各种应用中都有很好的表现。

1 LDA模型

LDA模型是全概率生成模型,其内在结构清晰,可以利用高效的概率推断算法进行计算。LDA模型参数空间的规模与训练文档数量无关,因此更适合处理大规模语料库[5,6]。

1.1 模型介绍

LDA模型是一种利用概率对文本主题信息进行建模的方法。如图1,一个文本通常由若干

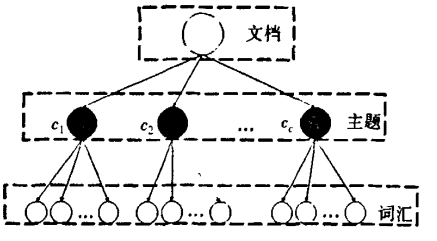

图1 LDA的隐含主题拓扑结构

Fig.1 network topology of LDA latent topics

收稿日期:2009-09-18

作者简介:孟海涛(1971-),男,江苏盐城人,讲师,硕士,主要研究方向为中文信息处理及嵌入式系统。

隐含主题组成，而这些主题由文本中特定词汇体现。因此可将隐含主题看做词汇的概率分布，单个文档表示为这些隐含主题特定比例的随机混合。

LDA模型是由词层、文档层和文档集合层组成有向概率图模型。(α,β)是文档集合层的参数，其决定了LDA模型。文档集合中α用于描述隐含主题间的相对强弱，隐含主题自身的概率分布用β表示。随机变量θ是文档层参数，θ的分量表示目标文档中每个隐含主题的权重。(z,w)是词层的参数，z表示目标文档的隐含主题在每个词上份额，w是目标文档的特征词向量(如图2)。

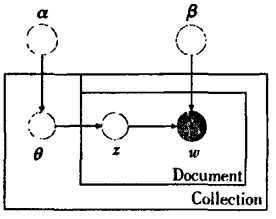

图2　LDA的图模型

Fig.2　LDA model

假设隐含主题有T个，则在所给文本中的第i个词汇w_i出现的概率为：

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

其中，w_i是文本的第i个特征词，j表示第j个隐含主题，$z_i=j$表明w_i取第j个隐含主题。$P(w_i|z_i=j)$是w_i属于j的概率，$P(z_i=j)$是j属于所给文本的概率。假定D个文本具有T个隐含主题，形成以W个唯一性词汇表示。令$\varphi_w^{(z=j)}=P(w|z=j)$，$\psi_{z=j}^{(d)}=P(z=j)$，则在文本$d$中词汇$w$出现的概率为：

$$P(w \mid d) = \sum_{j=1}^{T} \varphi_w^{(z=j)} \cdot \psi_{z=j}^{(d)} \quad (2)$$

为使得模型易于处理训练语料之外的新文本LDA模型[7]在$\psi^{(d)}$上作Dirichlet(α)的先验概率假设；文中还在$\varphi^{(z)}$上亦作对称的Dirichlet(χ)的先验概率假设[8]，以便于对模型参数的进行推理。如下：

$w_i \mid z_i,\varphi^{(z)} \sim \text{Discerte}(\varphi^{(z_i)}),\varphi^{(z_i)} \sim \text{Dirichlet}(\chi)$

$z_i \mid \psi^{(d)} \sim \text{Discerte}(\psi^{(d)}),\psi^{(d_i)} \sim \text{Dirichlet}(\alpha)$

式中的χ是根据文本的主题抽取样本获得的词汇出现频率，而不是统计文本中的词汇计算得出；α是根据文本的分类获得的主题被抽样的次数。主题及词汇被使用的程度由α和χ的具体取值决定，但是不同的词汇和主题被使用的方式基本相同。对此可以假定所有的α取相同的值，所有的χ取相同的值，即对称的Dirichlet分布。

1.2　Gibbs抽样

MCMC提供了从后验分布直接抽样本值的近似迭代方法，而Gibbs抽样简化了实现MCMC算法。从词汇对于主题的后验概率$P(w|z)$出发，使用Gibbs抽样计算出φ和ψ的值，得到词汇的概率分布。使用Gibbs抽样的关键是构造目标概率分布函数。LDA模型在本文中，只需要对变量z_i进行抽取样本。计算后验概率$P(z_i=j|z_{-i},w_i)$的公式如下：

$$P(z_i=j|z_{-i},w_i)=\frac{\frac{n_{-i,j}^{(w_i)}+\chi}{n_{-i,j}^{(\cdot)}+W\chi}\cdot\frac{n_{-i,j}^{(d_i)}+\alpha}{n_{-i}^{(d_i)}+T\alpha}}{\sum_{j=1}^{T}\frac{n_{-i,j}^{(w_i)}+\chi}{n_{-i,j}^{(\cdot)}+W\chi}\cdot\frac{n_{-i,j}^{(d_i)}+\alpha}{n_{-i,\cdot}^{(d_i)}+T\alpha}} \quad (3)$$

其中，$z_i=j$表示主题j包含词汇w_i(w_i是词汇w在所给的文本中位置权值)；z_{-i}表示所有$w_k(k\neq i)$对该主题的分配；$n_{-i,\cdot}^{(d_i)}$是d_i中所有主题包含的词汇个数；$n_{-i,j}^{(d_i)}$是文本d_i中主题j包含的词汇个数；$n_{-i,j}^{(\cdot)}$是主题j包含除$z_i=j$的所有词汇个数；$n_{-i,j}^{(w_i)}$是主题j包含w_i的个数。

Gibbs抽样算法：

(1)初始化马尔科夫(Markov)链。$z_i\in[1,T]$，$i\in[1,N]$，T为主题个数，N是文本特征词个数。z_i初值随机选取：

(2)for$(i=1,i<=N,i++)$|根据公式(3)将词汇分配给主题，计算*Markov*链下一个状态|；

(3)对第(2)步迭代多次，使Markov链接近目标分布，取z_i的当前值作为样本记录下来。

对于每个单一样本，按下式估算φ和ψ的值：

$$\hat{\varphi}_w^{(z=j)}=\frac{n_j^w+\chi}{n_j^{\cdot}+W\chi},\hat{\psi}_{z=j}^{(d)}=\frac{n_j^{(d)}+\alpha}{n_{\cdot}^{(d)}+T\alpha} \quad (4)$$

其中，$n_j^{(w)}$表示主题j包含词汇w的个数；$n_j^{\cdot}$表示主题j包含所有的词数；$n_{\cdot}^{(d)}$表示文本d中所有主题包含的词数；$n_j^{(d)}$表示文本d中分配给主题j的词数。

2　实验

2.1　实验数据

本实验采用了从网上人工下载的16 000篇英文WEB文档进行文本分类实验，它们包括教育、新闻、体育、军事、科技5个类别。将这些分好类的语料平均分成8份，选择其中一份运行LDA

表1 LDA与其他方法比较

Table 1 Comparing LDA with the other algarithms

	k均值聚类			贝叶斯网络			LDA		
维数	指标 Z	指标 L	指标 D	指标 Z	指标 L	指标 D	指标 Z	指标 L	指标 D
100	24.5%	9.5%	-1.86	18.5%	9.6%	-2.01	19.4%	7.8%	-2.08
500	8.3%	8.6%	-2.47	9.2%	6.8%	-2.54	11.7%	5.8%	-2.49
1 000	8.7%	5.2%	-2.70	6.5%	5.6%	-2.81	8.2%	3.4%	-2.94
2 000	4.9%	3.6%	-3.17	5.8%	4.6%	-2.96	5.3%	2.0%	-3.42
3 000	5.1%	3.1%	-3.22	3.9%	4.8%	-3.14	3.6%	1.6%	-3.73
5 000	5.8%	4.1%	-3.02	3.3%	3.9%	-3.33	2.0%	1.2%	-4.17

分类算法,共执行8次分类操作,计算其平均值。因为LDA是一种无监督的分类方法。首先对数据集进行标准化:去除噪音词,过滤文档频率小于2和大于8 000的词;表示出现的词的权重,对每篇WEB文档的主题与WEB文档的正文内容设置不同的权重。同时,删除所有不是在两个字母中间出现的连字符的词。用上述的方法处理后的WEB文档做实验。

2.2 实验判别标准

错分率 Z:是所有判断的文本中与人工分类结果不吻合的文本所占的比率。

误分率 L:是人工分类结果应有的文本中分类系统结果不吻合的文本所占的比率。

平均误分率的对数 D:

$$D = \log it^{-1}(\log it(Z)/2 + \log it(L)/2)$$

其中

$$\log it(x) = \log\left(\frac{x}{1-x}\right)$$

$$\log it^{-1}(x) = \log\left(\frac{e^x}{1+e^x}\right)$$

2.3 实验环境与方法

在我们的实验中,我们使用多种方法与LDA方法进行对比实验。分别是使用k均值聚类、贝叶斯网络方法。由于这些方法是有监督学习,采用已经预处理过的WEB文档集作为这些方法的训练集。

使用的k均值聚类、贝叶斯网络方法的实现软件是weka。LDA的实现工具是GibbsLDA-0.2。硬件条件是DELL PC机(3.0G双核,2G内存)。

2.4 实验结果

从表1中,对每种模型的性能用3个评价指标来表示,从单个指标来看LDA在指标 L 与指标 D 的表现最好,贝叶斯网络次之;贝叶斯网络在指标 Z 上表现最好,LDA次之;3个模型在不同维数上WEB文档过滤的性能指标值的变化都比较明显,而其他两种模型随维数变化的WEB文档过滤性能指标值的变化更大;在WEB文档中指标 D 最为重要的指标,再综合在不同维数上的三个指标,LDA模型优于其他两种模型,尤其是在较多的维数时,表现得更明显。实验也表明了LDA模型在WEB文档过滤中的性能表现良好、稳定。因此总的可以得出,LDA方法作为新的分类方法,在WEB文档过滤中总体性能占优。

3 结论

在本文中,采用新引入的LDA主题模型来进行WEB文本分类。过滤实验结果表明基于LDA的WEB文本分类器,能够较好地实现WEB文本的分类,比其它的经典的机器算法在多个指标上都占优。下一步的工作将针对LDA的算法进行改进和优化,从而使得其在WEB文本分类的各项指标进一步提高。

参考文献:

[1] Konstantin Tretyakov. Machine Learning Techniquesin Spam Filtering[A]. Data Mining Problem-Oriented Seminar, MTAT. 03, 177, May 2004: 60-79.

[2] Nello C, John S T. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods[M]. Cambridge: Cambridge University Press, 2000.

[3] Wegelin J A. A Survey of Partial Least Squares(PLS) Methods, with Emphasis On the Two-block Case[R]. Seattle: Department of Statistics, University of Washington, 2000: 21-28.

[4] Hoskuldsson A. PLs regressiOn methods[J]. Journal of Chemo metrics, 1988, 3(2): 211-228.

[5] Xiaogang Wang, Eric Grimson. Spatial Latent Dirichlet Allocation. Proceedings of Neural Information Processing Systems (NIPS2007). 2007[EB/OL]. Http://books. nips. cc/papers/files/nips20/NIPS2007_0964. pdf.
[6] McCallum A, Corrada - Emmanuel A, Wang X. Topic and role discovery in social networks[A]. Proceedings of 19th Joint conference on artificial intelligence. 2005.
[7] Thorsten Brants, Francine Chen, Ioannis Tsochantaridis. Topic - based document segmentation with probabilistic latent semantic analysis[A]. Proceedings of the eleventh international Conference on Information and knowledge management McLean, Virginia, USA. 2002. 211 - 218.
[8] Thomas Minka, John Lafferty. Expectation - Propagation for the Generative Aspect Model[A]. Uncertainty in Artificial Intelligence, 2002.

Web Text Classification based on LDA Model

MENG Hai-tao[1], CHEN Si[2], ZHOU Rui[2]
(1. School of Information Technology Yancheng of Institute Technology, Jiangsu Yancheng 224051, China;
2. International Department Beijing Daxing No. 1 Middle School, Beijing 102600, China)

Abstract: A kind of web text classification is put forward on the basis of LDA model. Latent Dirichlet Allocation (LDA) is an unsupervised topic learning model which extracts latent topics from text data. Parameters are estimated with Gibbs sampling of MCMC and the word probability is represented. Thus different latent topics are associated with observable words. Contrasting to SVM and Bayesian Network, the result in the experiment shows that LDA has the better performance than any other algorithm.
Keywords: Latent Dirichlet Allocation(LDA); topic model; WEB classification

(责任编辑：沈建新；校对：张英健)

(上接第47页)

参考文献：

[1] Paolo Toth, Daniele Vigo. The Vehicle Routin Problem[A]. Society for Industrial and Applied Mathematics philadephia. 2002.
[2] Clarke G, Wright J. Scheduling of vehicles from central depot to a number of delivery points[J]. Operations Research, 1964, 12: 568 - 581.
[3] 袁庆达，闫昱，周再玲. Tabu Search算法在优化配送路线问题中的应用[J]. 计算机工程，2001(11)：86 - 89.
[4] Osman I H. Meta. strategy simulated annealing an Tabu search algorithms for the vehicle routin problem[J]. Annu Oper Res, 1993, 41: 77 - 86.
[5] 段海滨. 蚁群算法原理及其应用[M]. 北京：科学出版社，2007.
[6] 弓晋丽，程志敏. 基于matlab物流配送路径优化问题遗传算法实现的实现[J]. 物流科技，2005，29(131)：103 - 105.

Research and Practice of VRP Based on the Improved K - means and Ant Colony Algorithm

ZHU Jin-xin
(School of Department of Experiment Teaching, Yancheng Institute of technology, Jiangsu Yancheng 224051, China)

Abstract: A study is made on VRP. To avoid long - time searching、precocity and stagnation and tendency to local optimization of traditional ant colony algorithm. First, improved K - means is applied to divide regions of customs, ant colony algorithm is applied to solve the problem in each region. Experiments indicate that the proposed method has good performance.
Keywords: VRP; Ant Colony Algorithm; Clustering Analysis; improved K - means

(责任编辑：张英健；校对：沈建新)