构建微博用户兴趣模型的主题模型的分析

陈文涛 张小明 李舟军

(北京航空航天大学计算机学院 北京 100191)

摘 要 分析了不同的主题模型,通过实验比较了3种主题模型构建的微博用户兴趣模型的性能。实验结果表明:TwitterLDA适用于新文档或新用户的预测,AuthorLDA产生的主题具有较高的区分度,而UserLDA和AuthorLDA能更好地反映出用户的社交网络关系。上述工作为进一步研究主题模型如何应用于微博的个性化信息推荐、情感分析和话题检测与跟踪等文本挖掘应用奠定了基础。

关键词 主题模型,用户兴趣,个性化服务

中图法分类号 TP391 **文献标识码** A

Analysis of Topic Models on Modeling MicroBlog User Interestingness

CHEN Wen-tao ZHANG Xiao-ming LI Zhou-jun

(School of Computer Science and Engineering, Beihang University, Beijing 100191, China)

Abstract This paper analysed different topic models, and compared three extended topic models' performance on modeling microblog user interestingness via three experiments. Experimental results show that TwitterLDA can apply to predict words on new unseen docuemnts and users, that the topics generated by AuthorLDA have a higher degree of differentiation, and that UserLDA and AuthorLDA can better reflect the users' relationships in real social network. The work in this paper lays the foundation for further studying how the topic model is applied to the text mining applications of microblogs such as personalized recommendation, sentiment analysis and topic detection and tracking.

Keywords Topic model, User interest, Personalized service

1 引言

随着Web2.0技术和无线网络技术的发展,社交网络对人类社会和生活的影响越来越大。微博作为当今流行的一种社交网络平台,为其用户提供了一个实时交流平台。微博用户可以通过电脑或者移动终端关注自己感兴趣的信息,实时地获取各种网络资源,并发表个人观点等。

2012年7月发布的《中国新媒体发展报告(2012)》蓝皮书[1]揭示了我国微博用户数量由2010年底的6311万猛增至2012年6月的2.74亿,使用率增长近300%,中国网民使用微博的比例已经过半。如此庞大的用户数量群,其知识层次差别很大,所产生的网络信息良莠不齐、形式各异。同时,不同用户的信息需求也各不相同。当前已有许多研究基于不同的需求向用户提供不同的服务,例如,基于内容过滤的个性化服务就是通过收集和分析用户信息来学习用户的兴趣和行为,从而实现主动推荐。其中,基于内容的用户兴趣模型的构建在该类研究中至关重要。

由于微博消息的文本长度较短(一条微博的长度不超过140个字),并且带有大量的网络用语、表情符、缩略简称及错别字等不规范用语,因此微博具有"噪声"多、特征词少等特点,从而导致基于VSM(Vector Space Model,向量空间模型)的微博用户兴趣模型的表示面临着维度高、数据稀疏等问题。然而,主题模型作为一种非监督学习的生成模型,不用事先对语料库进行特征抽取,因而对数据维度不敏感。本文将利用3种概率主题模型来对用户的兴趣进行建模,然后通过大量实验来分析和比较这些模型的性能。

本文第1节阐述了概率主题模型的核心思想和关键技术;第2节介绍了主题模型构建基于内容的微博用户兴趣模型的方法和3种常用的微博用户兴趣模型;第3节介绍了评估主题模型性能的方法并对比了3种微博用户兴趣模型的性能;最后总结全文。

2 概率主题模型

概率主题模型[1]是一种非监督的机器学习技术,主要被用来识别大规模文档集或语料库中潜在的主题信息。目前,已经有很多数据挖掘研究人员使用各种概率主题模型来分析文本内容和单词的潜在含义。这些模型都是基于一个同样的思想:每一篇文档都是由若干主题混合生成,并且被表示成主

到稿日期:2012-09-07 返修日期:2012-11-24 本文到国家自然科学基金项目(61170189,60973105,61202239),教育部博士点基金(20111102130003)资助。

陈文涛(1988-),男,硕士生,CCF会员,主要研究方向为数据挖掘,E-mail:kangcwt@gmail.com;张小明(1980-),男,博士生,主要研究方向为文本挖掘、话题检测与跟踪;李舟军(1963-),男,教授,博士生导师,主要研究方向为网络与信息安全、数据挖掘与文本挖掘。

[1] 社科文献出版社2012年7月在京发布。

题所构成的一个概率多项分布。其中，每个主题都是一个基于文本单词项的概率多项分布。所以，概率主题模型同时也是一种生成模型，其生成过程如下：对于文本中的每一个单词项 w_i（文本的第 i 个单词项），首先从文本的潜在主题分布中随机采样出一个主题 z，然后从主题 z 的单词概率多项分布中随机采样出一个单词作为 w_i。

2.1 LDA 模型

目前，LDA(Latent Dirichlet Allocation)模型[2]已经成为概率主题模型的一个实现标准。LDA 是一个层次贝叶斯模型[5]，如图 1 所示，共有 3 层：

1. 单词层：单词集 $V=\{w_1,w_2,\cdots,w_V\}$，是从语料库中提取出来的词频次数大于 10 的单词集合。

2. 主题层：主题集 $\Phi=\langle z_1,z_2,\cdots,z_K\rangle$ 中的每一个主题 z_k，都是一个基于单词集 V 的概率多项分布，可以被表示成向量 $\varphi_k=\langle p_{k,1},p_{k,2},\cdots,p_{k,V}\rangle$，其中 $p_{k,j}$ 表示单词 w_j 在主题 z_k 中的生成概率。

3. 文档层：就单词层而言，采用词袋方法，每一篇文档被表示成一个词频向量 $d_i=\langle tf_{i,1},tf_{i,2},\cdots,tf_{i,V}\rangle$，其中 $tf_{i,j}$ 表示单词 j 在文档 i 中的出现次数；就主题层而言，文档集可表示成 $\Theta=\langle\theta_1,\theta_2,\cdots,\theta_D\rangle$，其中每一篇文档由一个向量 $\theta_d=\langle p_{d,1},p_{d,2},\cdots,p_{d,K}\rangle$ 表示，$p_{d,z}$ 是主题 z 在文档 d 中的生成概率。

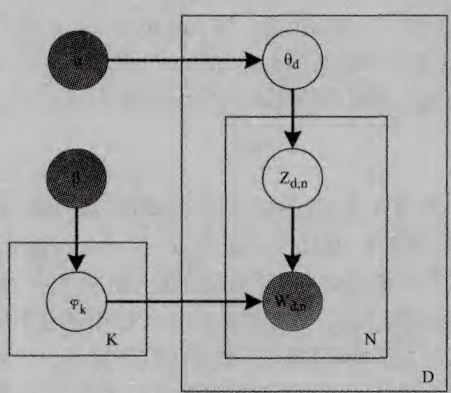


图 1 LDA 模型的图表示

LDA 模型采用 Dirichlet 分布作为概率主题模型中多项分布的先验分布，图 1 中的 α 和 β 则分别是文档-主题概率分布 Θ、主题-单词概率分布 Φ 的先验知识。

其中，LDA 模型的生成过程描述如图 2 所示。

1. 对于每一个主题 z_k，k=1,…,K
   (1) 对主题下的单词概率进行多项采样：$\varphi_k\sim$ Dir(β)
2. 对每篇文档 d_i，i=1,…,D
   (1) 对文档 d_i 下的主题概率进行多项采样：$\theta_d\sim$ Dir(α)
   (2) 对文档 d_i 下的每个单词 $w_{d,n}$，n=1,…,N_i
      (a) 对文档的主题多项分布进行采样得到 $w_{d,n}$ 的主题 $z_{d,n}$，$z_{d,n}\sim$Multi(θ_d)
      (b) 对主题 $z_{d,n}$ 下的单词多项分布进行采样得到单词 $w_{d,n}$，$w_{d,n}\sim$Multi($\varphi z_{d,n}$)

图 2 LDA 模型的生成过程

2.2 参数估计

Θ 和 Φ 分别代表了文档-主题、主题-单词的概率分布，是模型最终要求解的参数。然而要准确计算 Θ 和 Φ 是非常棘手的[2]，因此，Griffiths 和 Steyvers 提出了使用简单而又有效的 Gibbs 采样的方法对这两个参数进行近似估算[3]。Gibbs 采样是一种马尔卡夫蒙特卡洛近似算法：

1. 初始化：为语料库中的每篇文档的每个单词项随机赋予一个主题，构成马尔卡夫链的初始状态。

2. 下一个状态的求解：对语料库中的每篇文档 d_i 的每个单词项 w_n 进行迭代，在固定语料库中其他单词项的主题分配(Z_{-n})的前提下，根据式(1)计算当前单词项下各个主题的概率 $p(z_n=k\mid w_n)$；采用概率随机采样获取当前项的分配主题；当迭代结束时，就得到了马尔卡夫链的下一个状态。

$$P(z_n=t|w_n,\alpha,\beta,Z_{-n})\propto\frac{\alpha+N_{t,-n}^{DK}}{\sum_{k=1}^{K}(\alpha+N_{k,-n}^{DK})}\cdot\frac{\beta+N_{t,w,-n}^{CW}}{\sum_{v=1}^{V}(\beta+N_{t,v,-n}^{CW})}\quad(1)$$

式中，$N_{t,-n}^{DK}$ 为文档 d_i 中不考虑当前项 w_n 时由主题 t 生成的单词数；$N_{t,w,-n}^{CW}$ 为语料库中不考虑当前项 w_n 时主题 t 生成单词 w 的次数。

3. 重复步骤 2 中的迭代，直至足够次数后，马尔卡夫链可以达到稳定状态。

当 Gibbs 采样过程结束时，Θ 和 Φ 也并不能准确地估算出来，而是使用文档下主题的分配情况和主题下单词的分配情况的后验概率近似估算 φ(主题在文档中的生成概率 $p(u|z)$)和 θ(单词在主题中的生成概率 $p(z|w)$)，其计算方法如下：

$$\theta_d^t=\frac{N_{dt}^{DK}+\alpha}{\sum_{k=1}^{K}[N_{dk}^{DK}+\alpha]},\varphi_t^w=\frac{N_{tw}^{CW}+\beta}{\sum_{v=1}^{V}[N_{tv}^{CW}+\beta]}\quad(2)$$

式中，N^{DK} 和 N^{CW} 分别为 $V*K$ 和 $D*K$ 的矩阵，N_{tw}^{CW} 是整个语料库中单词 w 分配为主题 t 的数目，N_{dt}^{DK} 是文档 d 中分配为主题 t 的单词数目。

3 基于内容的微博用户兴趣模型

在使用主题模型构建用户兴趣时，用户兴趣普遍被定义为用户对各个主题的喜好程度。用户下某个主题的生成概率反映了用户对该主题的喜好程度。因此，主题模型下用户-主题的生成概率多项分布表示了用户的兴趣。

主题模型构建基于内容的微博用户兴趣模型时，需要将一个用户下的所有微博合并成一个文档进行主题生成，从而得到用户生成主题的概率多项分布，即用户的兴趣模型。而 LDA 模型中的文档层则对应到了兴趣模型中的用户层。

在用户层中，用户集合 $U=\{u_1,u_2,\cdots,u_N\}$ 中的每一个用户 u_i 由微博集合 $\{t_{u,1},t_{u,2},\cdots,t_{u,m}\}$ 组成，而微博集合中的 $t_{u,j}$ 是由用户 u_i 发布或转载的微博的词频向量。同样，从主题层面而言，用户可以被表示成向量 $\theta_u=\{p_{u,1},p_{u,2},\cdots,p_{u,K}\}$，其中 $p_{u,z}$ 表示主题 z 在用户 u 中的生成概率，也就是用户 u 对主题 z 的喜好程度。由此可知，用户层可以生成用户与主题的生成关系，从而构成用户兴趣模型。

目前已经有很多研究使用扩展的主题模型来构建用户的兴趣模型。

TwitterRank 模型[7]利用主题模型生成微博用户兴趣，通过在 PageRank 算法中引入基于用户兴趣的用户相似度，从而找出 Twitter 中某个话题下具有影响力的用户。值得一提的是，TwitterRank 在实现主题模型时，将一个用户发表的所有微博合并成一个单独的文档进行处理。通过这种方式，TwitterRank 将一个用户映射到一个文档，将用户兴趣的生成映射到文档的生成。这种方法被 Liangjie Hong 和 Davison

称为用户视图[8]，同时也等同于单一作者的 Author-Topic 模型，其图模型表示如图 3(a)所示。

Author-Topic 模型[9,10]通过引入作者协作关系，在 LDA 模型的基础上增加了作者层：由多个作者共同完成的文档可以描述成一个由作者构成的概率多项分布；而作者则被表示成一个基于主题的概率多项分布。微博的评论和转发消息共有原作者和评论作者两人，其中原微博内容可以认为是由两个作者协作完成的。因此 Author-Topic 模型可应用于建模微博用户兴趣，如图 3(b)所示。

由于一条微博的文本内容比较短，TwitterLDA 模型[6]认为一条微博通常只包含一个主题，即一条微博中的所有单词仅由一个主题生成。TwitterLDA 模型如图 3(c)所示，它也是基于用户视图的扩展主题模型。该实验表明，TwitterLDA 在微博消息的主题发现应用中比 Author-Topic 模型和普通的 LDA 模型效果更好。

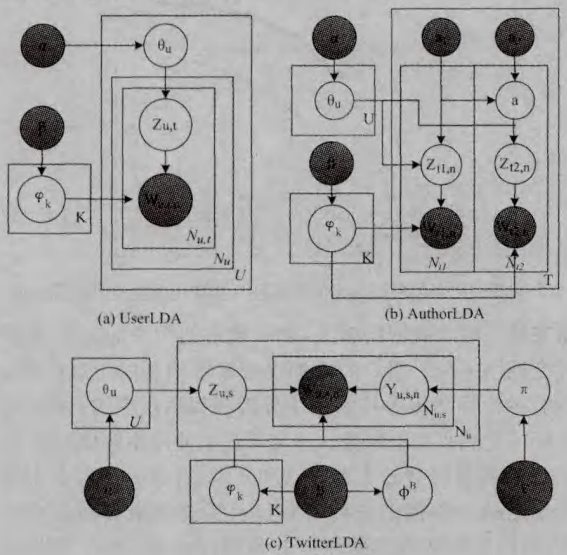


图 3　3 种微博用户兴趣模型

以上 3 种微博用户兴趣模型在主题模型的基础上引入了一些适当假设，甚至考虑了作者的协作关系。还有一些其他主题模型的扩展研究，例如，在模型中引入时间[12,14]、地理位置[16]、文本间的引用关系[17,18]等。

其中，TOT(Topics Over Time Model)模型[14]引入时间概念对主题模型进行扩展。TAT(Temporal-Author-Topic Model)模型[12]则通过考虑单词与时间的共现关系扩展了 Author-Topic 模型，从而形成随着时间变化而变化的临时用户兴趣模型。然而 TOT 模型和 TAT 模型主要用于在文献语料库中挖掘主题的变化，其时间片一般以月和年为单位，并不适用于实时性很强的微博平台。

TAR(Topic-Author-Recipient Model)模型[15]通过引入消息的发送-回应关系来扩展 LDA 模型，它主要用于挖掘社交网络中的主题和角色。然而稍微热门的微博，其转发条数经常达到成千上万条，这些转发中大部分评论是空的。因而，在 TAR 模型中，发送者和回应者(微博中对应的是原创者和评论转发者)的条件概率分布将会非常稀疏，不适用于构建微博用户的兴趣模型。

4　实验与结果

4.1　实验数据

实验数据来自腾讯微博[2]的开放平台 API[3]，共采集了 9540 位用户 2012 年 7 月 21 日至 8 月 1 日的所有带'#'符号的微博。本实验通过正则表达式去除微博中的链接、转发标志、@用户名和表情符号等噪音信息之后，对其进行分词，并去除停用词、高频词和低频词，最后得到表 1 所列数据集合。

表 1　实验数据统计信息

用户数(个)	9,540
微博数(条)	440,328
单词种类(种)	40,229
单词总个数(个)	13,019,554
时间跨度(天)	10

4.2　实验方法和结果

通常使用基于 Perplexity 的方法[18]评估主题模型的生成性能，有时也会通过计算模型下的主题相似度和文本相似度等间接任务比较主题模型的好坏。本文将利用 3 个实验，分别对比 UserLDA、AuthorLDA 和 TwitterLDA 模型的生成能力、内部相似度和用户相似度。在这些实验中，主题模型的运行参数统一设为：$\alpha=50/K$，$\beta=0.1$，$\gamma=0.1$，迭代次数 800。

4.2.1　Perplexity

Perplexity 是一种评估语言模型生成性能的标准测量指标[4,9]。Perplexity 值表示模型生成测试集中新文本的似然估计，它用来衡量模型对新文本的预测能力。Perplexity 值越小，似然估计就越高，也就表示模型的生成性能越好。Perplexity 值的计算如下列公式所示。其中式(6)适用于 AuthorLDA 中有两个作者的转发微博。

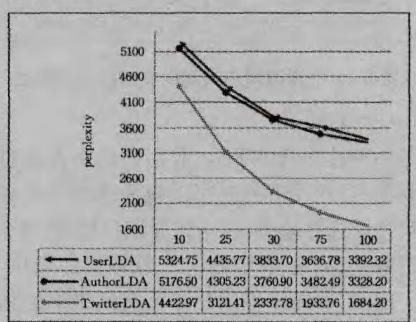
$$\text{Perplexity}(D_{test})=\exp\{-\frac{\sum_{u=1}^{U}\log p(w_u)}{\sum_{u=1}^{U}N_u}\} \tag{3}$$

$$p(w_u)=\prod_{n=1}^{N_u}\sum_{k=1}^{K}p(w_n|z_n=k)\cdot p(z_n=k|u) \tag{4}$$

$$\log\ p(w_u)=\sum_{n=1}^{N_u}\log(\sum_{k=1}^{K}\theta_{uk}\cdot\varphi_{k,w_n}) \tag{5}$$

$$p_2(w_u|u_1u_2)=\prod_{n=1}^{N_t}\frac{1}{2}\sum_{u,k=1}^{K}p(w_n|z_n=k)\cdot p(z_n=k|u) \tag{6}$$

式(3)中的 $p(w_u)$是 D_{test}中新用户的微博在兴趣模型下的生成概率。计算 $p(w_u)$时，首先需要用模型对 D_{test}中的测试用户进行推理得到用户下主题的生成概率分布，再使用相应公式进行计算。



	10	25	30	75	100
UserLDA	5324.75	4435.77	3833.70	3636.78	3392.32
AuthorLDA	5176.50	4305.23	3760.90	3482.49	3328.20
TwitterLDA	4422.97	3121.41	2337.78	1933.76	1684.20

图 4　3 个微博用户兴趣模型的 Perplexity 性能对比

2) http://t.qq.com

3) http://open.t.qq.com

在本文的实验中，语料库被分成 9000 个训练用户 D_{train} 和 540 个测试用户 D_{test}，并利用式(3)计算各个模型分别在主题数为 10、25、50、75 和 100 时的 Perplexity(D_{test})的值，其结果如图 4 所示。从图中的数据可以看出，TwitterLDA 的模型生成能力要远远大于其它两个模型。这也同时说明了：一条微博通常只包括一个主题的假设在很大程度上是可行的。

4.2.2 主题相似度

主题模型可以从语料库中挖掘出潜在的主题。根据主题模型的思想，模型中生成的各个主题之间应该存在较大差异，即各个主题间的相似度要尽可能地小。基于这一点，本实验将模型下主题间的相似度作为模型性能的一个评估标准。

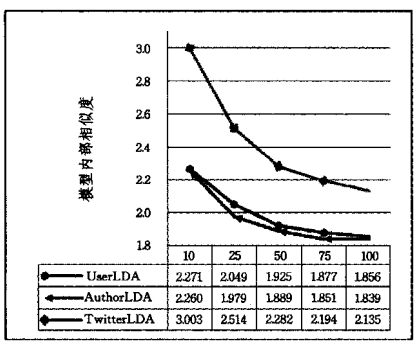
首先，计算模型下各个主题间的相似度；接着，计算每一个主题与其他主题的平均相似度；最后，将所有主题的平均相似度的平均值作为模型的内部相似度。显然，模型内部相似度越大，模型下各个主题重叠的可能性越大，性能也就越差。本实验使用 Jensen-Shannon(JS)距离计算主题相似度。JS 距离，如式(7)所示，是一种可以测量一对随机变量的概率分布相似性的方法[3]。其中，T_{m1} 和 T_{m2} 分别是两个随机变量的概率分布，R 是两者的概率分布平均值；$D_{KL}(A \parallel B)$ 表示随机变量 $A=(p_1, p_2, \cdots, p_T)$ 和 $B=(q_1, q_2, \cdots, q_T)$ 的 Kullback Leibler(KL)距离，其计算方法如式(9)所示。

$$D_{JS}=[D_{KL}(T_{m1} \parallel R)+D_{KL}(T_{m2} \parallel R)]/2 \tag{7}$$

$$R=(T_{m1}+T_{m2})/2 \tag{8}$$

$$D_{KL}(A \parallel B)=\sum_{j=1}^{T}(p_j \log_2 \frac{p_j}{q_j}) \tag{9}$$

JS 距离越小，其相似度越大。故将主题的概率多项分布的 JS 距离的倒数当作主题间的相似度。图 5 给出了 3 个主题模型在主题数为 10、25、50、75 和 100 时的模型内部相似度，可以看出，AuthorLDA 在模型内部相似度上占有一定的优势。同时也可以看出，随着主题数目的增加，模型内部相似度减少。这可能是因为当主题数目增多时，主题的内容被逐步细化，从而导致主题与主题间的差异变得越来越大。



	10	25	50	75	100
UserLDA	2.271	2.049	1.925	1.877	1.856
AuthorLDA	2.260	1.979	1.889	1.851	1.839
TwitterLDA	3.003	2.514	2.282	2.194	2.135

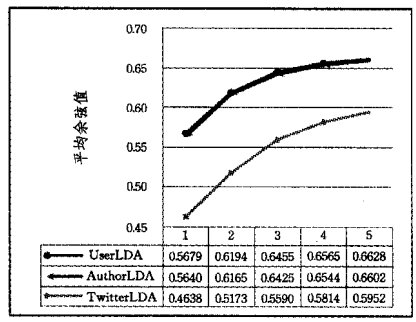
图 5 3 个微博用户兴趣模型的内部相似度

4.2.3 用户相似度

微博是一种社交网络平台，其用户的个人信息(标签、地点和自我介绍等)和收听列表可以视为用户个人兴趣的一种表现。使用相同标签越多、收听相同用户越多，两个用户就越相似。所谓“物以类聚、人以群分”，越相似的用户，其兴趣也就越接近。因此，本实验将用户在社交网络下的相似度作为一个基准，对比各个模型下用户相似度与基准的差异，以此衡量模型的好坏。该过程可描述如下：

首先，根据用户的收听列表、标签和自我介绍等计算两两之间用户的相似度，从而得到用户的社交网络相似度矩阵 $netSim_{U*U}$。接着，根据用户的主题概率分布，使用 JS 距离的倒数计算同一模型下两两用户之间的相似度，从而得到用户的模型相似度矩阵 $modelSim_{U*U}$。其中，矩阵 $netSim_{U*U}$ 和矩阵 $modelSim_{U*U}$ 的元素 $[i,j]$ 表示用户 i 与用户 j 的相似度，第 i 行则表示了用户 i 与其他用户的相似度向量。最后，使用余弦值计算同一个用户在矩阵和矩阵中相似度向量的拟合度。

余弦值越大，拟合度就越高，同时表示模型所计算出来的用户与用户之间的关系就越符合社交网络中的关系。从图 6 中可以看出，UserLDA 和 AuthorLDA 刻画用户的社交网络关系的能力都比 TwitterLDA 更优越。这可能是因为 TwitterLDA 的模型假设约束了用户的主题分配情况(用户的主题数与用户发表的微博数一一对应)，从而它的主题分配远没有其他两种模型的主题分配(用户的主题数与用户发表微博中的单词数一一对应)粒度细。



	1	2	3	4	5
UserLDA	0.5679	0.6194	0.6455	0.6565	0.6628
AuthorLDA	0.5640	0.6165	0.6425	0.6544	0.6602
TwitterLDA	0.4638	0.5173	0.5590	0.5814	0.5952

图 6 3 个微博用户兴趣模型的相似度矩阵与基准相似度矩阵的拟合度

结束语 随着微博的普及，越来越多的用户加入了微博这个社交网络平台。用户的增加也带来了信息的过载问题，基于内容的个性化服务技术用于分析和勾画用户的兴趣，并为其提供可能感兴趣的服务。本文阐述了使用主题模型构建微博用户的兴趣模型的主要方法和技术，并进一步对其中 3 个较为常用的兴趣模型进行了对比。实验结果表明：TwitterLDA 适用于新文档或新用户的预测，AuthorLDA 产生的主题具有较高的区分度，而 UserLDA 和 AuthorLDA 能更好地反映出用户的社交网络关系。

参 考 文 献

[1] Blei D M, Lafferty J. Text Mining: Theory and Applications [M]. Chapter Topic Models, Taylor and Francis, London, 2009

[2] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 2003, 3(4/5): 993-1022

[3] Steyvers M, Griffiths T. Probabilistic Topic Models[M]. Latent Semantic Analysis: A Road to Meaning, Laurence Erlbaum, 2005

[4] Heinrich G. Parameter estimation for text analysis[R]. Technical report. http://www.arbylon.net/publications/textest, Version 2, 2008

[5] Koller D, Friedman N. Probabilistic Graphical Models: Principles and Techniques[M]. MIT Press, 2009

[6] Zhao Xin, Jiang Jing, Weng Jian-shu, et al. Comparing Twitter and traditional media using topic models[C]//Proceedings of the 33rd European Conference on Information Retrieval. Springer-Verlag Berlin, Heidelberg, 2011: 338-349

[7] Weng Jian-shu, Lim E-P, Jiang Jing, et al. TwitterRank: finding topic-sensitive influential twitterers[C]//Proceedings of the 3th ACM International Conference on Web Search and Data Mining. New York City, NY, USA, 2010: 261-270

(下转第 135 页)

结束语 本文提出了一种基于K-S统计的不平衡数据分类方法，该方法通过对数据分片调整数据不平衡度，然后进行分类学习。分片结果表明，多数类与少数类达到了很好的聚集。将多数类或少数类集中于某一分片，或者在分片中使正类与负类的分布差异性最大，可便于分类器区别。同时对于不同程度的不平衡样本，该方法的分类精度有一定程度的提高。可见，本文提出的方法是有效、可行的。本文对于极端不平衡的数据集具有很好的效果，但是如何进一步提高不平衡程度一般的数据集，将是今后需要进一步研究的目标。分片数量对分类精度有所影响，如何自适应地确定决策树的大小也是今后的研究任务。

参 考 文 献

[1] Ling C X, Li C. Data mining for direct marketing: Problems and solutions[C]// Proceedings of the 4th international conference on knowledge discovery and data mining. New York, NY, 1998: 73-79

[2] Sun Yan-min, Kamel M S, Wong A K C, et al. Cost-Sensitive Boosting for Classification of Imbalanced Data [J]. Pattern Recognition, 2007, 40(12): 3358-3378

[3] Estabrooks A, Jo T, Japkowicz N. A multiple resampling method for learning from imbalanced data sets [J]. Computational Intelligence, 2004, 20(1): 18-36

[4] Japkowicz N, Stephen S. The class imbalance problem: A systematic study [J]. Intelligent Data Analysis, 2002, 6(5): 429-450

[5] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic minority over-sampling techniques [J]. Journal of Artificial Research, 2002, 16: 321-357

[6] Drummond C, Holte R C. C4.5, Class imbalance, and cost sensitivity: Why under-sampling beats over-sampling [C]// Proceedings of the ICML'03 Workshop on Learning from Imbalanced Data Sets. 2003

[7] Kubat M, Matwin S. Addressing the curse of imbalanced training sets: one-sided selection [C]// Proceedings of the 14th International Conference on Machine Learning. 1997: 179-186

[8] Holte R C, Acker L E, Porter B W. Concept learning and the problem of small disjuncts[C]// Proceedings of the 11th joint international conference on artificial intelligence. 1989: 813-818

[9] Weiss G M. Mining with rarity: A unifying framework [J]. ACM SIGKDD Explorations Newsletter-Special Issue on Learning from Imbalanced Datasets, 2004, 6(1): 7-19

[10] Quinlan J R. Improved estimates for the accuracy of small disjuncts [J]. Machine Learning, 1991, 6(1): 93-98

[11] Ling C X, Sheng V, Yang Q. Test strategies for cost-sensitive decision trees [J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(8): 1055-1067

[12] Veropoulos K, Campbell C, Cristianini N. Controlling the sensitivity of support vector machines [C]// Proceedings of international joint conference on artificial intelligence. 1999: 55-66

[13] Zheng Z, Wu X, Srihari R. Feature selection for text categorization on imbalanced Data [J]. SIGKDD Explorations, 2004, 6(1): 80-89

[14] Larose D T. 数据挖掘方法与模型[M]. 北京：高等教育出版社，2011：143-146

[15] Han H, Wang W Y, Mao B H. Borderline-SMOTE: A New Over-Sampling Method in imbalanced Data Sets Learning[C]// Proceedings of the International Conference on Intelligent Computing. Hefei, China, 2005: 878-887

(上接第 130 页)

[8] Hong Liang-jie, Davison B D. Empirical study of topic modeling in Twitter[C]// Proceedings of the First Workshop on Social Media Analytics. Washington DC, USA, 2010: 80-88

[9] Rosen-Zvi M, Griffiths T, Steyvers M, et al. The author-topic model for authors and documents[C]// Proceedings of the 20th conference on Uncertainty in artificial intelligence. AUAI Press Arlington, Virginia, United States, 2004: 487-494

[10] Steyvers M, Smyth P, Rosen-Zvi M, et al. Probabilistic author-topic models for information discovery[C]// Proceedings of the Tenth ACM SIGKDD International Confe-rence on Knowledge Discovery and Data Mining. Seattle, WA, USA, 2004: 306-315

[11] Ramage D, Dumais S, Liebling D. Characterizing micorblogs with topic models[C]// Proceedings of the 4th International Conference on Weblogs and Social Media. Washington DC, U S A, 2010

[12] Daud A, Li Juan-zi, Zhou Li-zhu, et al. Exploiting temporal authors interests via temporal-author-topic modeling[C]// Proceedings of 5th International Conference on Advanced Data Mining and Applications. Verlag Berlin, Heidelberg, 2009: 435-443

[13] Liu Yan, Niculescu-Mizil A, Gryc W. Topic-link LDA: joint models of topic and author community[C]// Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, QC, Canada, 2009: 665-672

[14] Wang Xue-rui, McCallum A. Topics over time: a non-markov continuous-time model of topical trends[C]// Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, USA, 2006: 424-433

[15] McCallum A, Corrada-Emmanuel A, Wang Xue-rui. Topic and role discovery in social networks[C]// Proceedings of 19th International Joint Conference on Artificial Intelligence, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA: 786-791

[16] Mei Qiao-zhu, Liu Chao, Su Hang, et al. A probabilistic approach to spatiotemporal theme pattern mining on weblogs[C]// Proceedings of the 15th International Conference on Word Wide Web. Edinburgh, Scotland, UK, 2006: 533-542

[17] Su Yi-zhou, Han Jia-wei, Gao Jing, et al. iTopicModel: Information Network-Integrated Topic Modeling[C]// Proceeding of the 9th IEEE International Conference on Data Mining. Miami, USA, 2009: 487-497

[18] Mei Qiao-zhu, Cai Deng, Zhang Duo, et al. Topic modeling with network regularization[C]// Proceeding of the 17th International World Wide Web Conference. Beijing, China, 2008: 101-111