文章编号：1005-8451（2018）07-0109-06

基于旅客出行意图的潜在高价值航线挖掘

卢　敏，李照宇，刘康超，李　纯
（中国民航大学，天津　300300）

摘　要：借助Map-reduce平台对旅客订票日志进行挖掘，并采用LDA算法挖掘旅客出行意图，进而计算航线的潜在价值。在2011年中航信民航旅客订票日志上的实验结果表明：采用LDA算法挖掘的航线与实际热门航线的Jacarrd 相似系数达92%，比基于航班次数统计的传统方法高出2%，能够更准确、更有效地预测航线的未来价值。

关键词：Map-reduce；航线；LDA；Gibbs；出行意图
中图分类号：U8：TP39　　**文献标识码：**A

Exploration of potential high-value airlines based on passenger travel intentions

LU Ming, LI Zhaoyu, LIU Kangchao, LI Chun
(Civil Aviation University of China, Tianjin　300300, China)

Abstract: This paper used the map-reduce platform to mine the passenger booking logs, used the LDA algorithm to mine passenger's travel intentions, and then calculated the potential value of the airline. The experimental results on the passenger name records in 2011 indicate that the similarity(Jacarrd Index) between the airlines excavated and the actual hot airlines is up to 92%, which is 2% higher than the conventional method based on the number of flights. It can predict the future value of the airline more accurately and effectively.

Keywords: Map-reduce; airline; LDA; Gibbs; travel intention

我国虽然已形成覆盖沿海发达城市以及重要省会城市的20家区域性枢纽空港，但航线布局呈东密西疏、沿海密内陆疏的发展态势，且航空运输干支网络缺乏有效衔接，使得各个航线不能充分的体现其价值，也无法发挥枢纽机场的中转功能和航空网络的整体效能[1]。因此，如何帮助航空公司合理安排航班航线，减少各航空公司间航班的重叠率，成为航空公司收益管理的重要组成部分。

民航航线挖掘任务是发现具有潜在高收益的航线，进而为航空公司航线开辟提供理论依据，其核心问题是如何准确评估航线的价值。已有的方法可以概括为3大类：

（1）传统航线收益评价方法[2]，其核心思想是只考虑单航线历史的收益，航线收益好，公司就继续经营甚至投放更多的运力扩大经营规模，反之，减少运力投放，缩小经营规模；

（2）在航线收益基础上，融入起飞机场和目的机场的经济特性等知识[3]；

（3）设计融入航线运营成本的多指标决策问题模型[4]。

由于旅客乘坐的航线本质由旅客的出行需求决定，为此本文提出基于旅客出行意图发现的航线价值计算方法。其核心初衷是：旅客出行受出行目的、季节性、年龄段、出差、旅游等因素的影响，而上述因素最终表现为旅客出行意图，因此航线价值较大程度上取决于旅客的出行意图。

1　基于旅客出行意图发现的航线价值计算

1.1　LDA算法（定义）及其核心思想

概率主题模型最早起源于潜在语义分析（LSI，Latent semantic Indexing）[5]，旨在解决信息检索中面临的一词多义和多词同义的语义问题。在此基础之上，研究者提出更多的概率主题模型，其中经典的

收稿日期：2018-05-10
基金项目：国家自然科学基金项目（61502499）；大学生创新创业训练计划项目（201710059047）；中国民航大学科研基金(2013QD18X)；中山大学机器智能与先进计算教育部重点实验室开放课题(MSC-201704A)；中央高校基本科研业务费科研专项（3122013C005）。
作者简介：卢　敏，助理研究员；李照宇，在读本科生。

概率主题模型是潜在狄利克雷分配（LDA）[6]。它可以将文档集中每篇文档的主题以概率分布的形式给出，从而通过分析一些文档抽取出它们的主题（分布）出来后，便可以根据主题（分布）进行主题聚类或文本分类。同时，它是一种典型的词袋模型，即一篇文档是由一组词构成，词与词之间没有先后顺序的关系。针对词条件独立的约束，研究者分别提出考虑 syntax 的主题模型[7]、引入 word correlation 的主题模型[8]以及 term selection 的主题模型[9]。针对文档相互独立的假设，由于很多实际应用中文档间存在引用和链接关系，研究者分别提出 link-based LDA[10]以及 author-topic LDA[11]。针对主题相互独立的约束，提出 topic-link LDA[12]、correlated LDA[13]、hierarchy LDA[14]等。针对没有充分利用文档标注信息，研究者提出 supervised LDA[15]。除上述研究之外，研究者还研究了大规模数据上的 LDA 近似计算及推理模型[16]，以及将 LDA 模型应用于社交网络分析[17-18]。

1.2 问题建模

采用 P(a) 描述航线 a 未来旅客的乘坐概率，取值越大表示航线潜在价值越高，其中航线是由起始机场和目的机场决定的，如 PEK#SHA 表示北京 PEK 到上海虹桥 SHA 的航线。航线的价值体现为现在和将来选择乘坐该航线上航班的旅客数量，而航班记录仅是由已乘坐的旅客组成，很多旅客当前未乘坐该航线上航班，并不代表这些旅客将来不会选择该航线，故而不能根据航班记录简单的统计直接得出结论。因此，航线 a 的潜在价值应该为所有旅客乘坐航线 a 的概率和，即：$P(a)=\sum_u p(a, u)$。

1.3 航线价值的计算方法

根据用户出行意图和贝叶斯全概率公式，将航线概率 $P(a)$ 展开为：

$$P(a)=\sum_u P(u,a)=\sum_u\sum_{z_u} P(a|z_u)P(z_u|u)P(u) \quad (1)$$

其中：

a—表示航线，如 PEK#SHA 表示北京到上海的航线；

$P(a)$—表示航线 a 的价值；

u—表示具体某个旅客；

$P(u)$—表示用户乘坐所有航线的次数，即在航班记录中的乘坐次数，也是用户的重要度；

z_u—表示用户 u 的潜在出行意图 z，在 LDA 中，每位旅客都有对应的出行意图 $z \in \{1, 2, \cdots, K\}$，即 LDA 中总有 K 个出行意图，每条航线只能选择其中一个意图，意图标号依次是 $1, 2, \cdots, K$；

$P(a|z_u)$—表示潜在出行意图 z_u 下的航线 a 分布；

$P(z_u|u)$—表示用户 u 的潜在出行意图 z_u 分布。

公式的物理含义为：用户 u 乘坐航线 a 的概率 $P(a, u)$，直接由旅客出行意图影响。由于旅客的出行意图 z_u 是不可见的，使用 LDA 主题模型从预处理过的民航旅客订票日志中挖掘，并根据贝叶斯全概率公式 $P(a, u)=\sum_{z_u} p(a, z_u, u)$ 和条件独立的假设，进一步展开求解 $p(a, z_u, u)=P(z_u|u)P(u)$。

旅客 u 在潜在出行意图 z_u 下选择航线 a 的概率 $p(a, z_u, u)$ 本质是两阶段形成的：旅客 u 首先确定潜在出行意图 z_u，然后才会根据出行意图再选择航线 a。在此基础之上，使用贝叶斯条件概率公式将上述概率进一步展开，则 $p(a, z_u, u)=p(a|z_u)p(z_u|u)$。其中，$p(z_u|u)$ 表示用户 u 选择潜在出行意图 z_u 的可能性，而 $p(a|z_u)$ 为当前意图 z_u 下选择航线 a 的概率。

基于上述假设，式 (1) 中需要求解的是 $P(u)$，$p(a|z_u)$ 和 $p(z_u|u)$，其中，$p(a|z_u)$ 和 $p(z_u|u)$ 可通过主题模型 LDA 进行挖掘。

1.4 模型优化及参数求解

为了快速求解参数 $P(u)$，$p(a|z_u)$ 和 $p(z_u|u)$，提出基于 Map-reduce 和 LDA 的航线价值计算方法。

Map-reduce 是一种编程模型，用于大规模数据集（大于 1 TB）的并行运算。概念“Map（映射）”和“Reduce（归约）”，是它们的主要思想，都是从函数式编程语言和矢量编程语言里借来的特性。它极大地方便了编程人员在不会分布式并行编程的情况下，将自己的程序运行在分布式系统上。 当前的软件实现是指定一个 Map（映射）函数，用来把一组键值对映射成一组新的键值对，指定并发的 Reduce（归约）函数，用来保证所有映射的键值对中的每一个共享相同的键组。

1.4.1 基于Map-reduce和LDA的航线价值计算方法

（1）输入航班数据；

（2）Map 处理：将乘客的每一条乘坐记录单独

分割；

（3）reduce处理：以每一个乘客的id作为键，乘坐记录作为值，将旅客所乘坐的所有航班的飞行航线筛选出来合并到一起；

（4）基于LDA的旅客出行意图挖掘，计算$p(a|z_u)$和$p(z_u|u)$；

（5）根据公式(1)，计算航线乘坐概率P(a)。

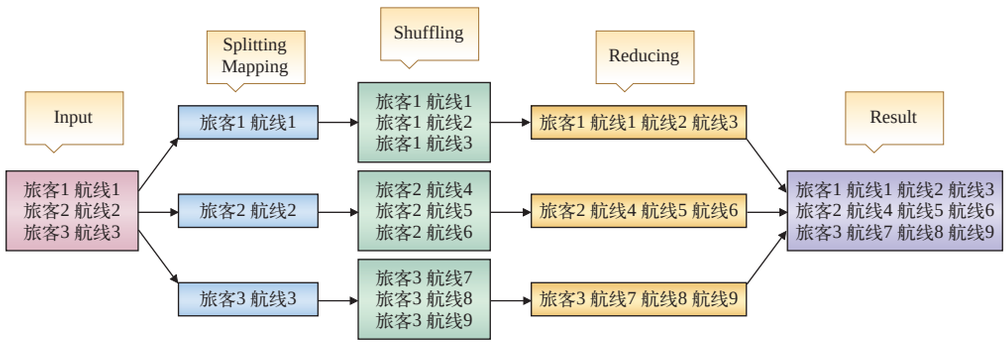

图1　Map-reduce处理过程

1.4.2　基于Map-reduce的民航旅客订票日志数据处理

民航旅客订票日志（PNR）是旅客乘坐航班的信息，对于如此庞大的数据量，一般算法用于有限内存，难以得出LDA的数据源。Map-reduce是基于分布式的针对这种场景的算法，将航班数据文件切割成小段，再对其每一段进行运算，统计每位旅客乘坐的所有航线的集合，得出结果后以旅客身份证(加密)为键值进行归并得到最终的LDA输入数据源，即一行为一个旅客的所有乘坐记录，而每一行内单个词则是旅客的一次乘坐记录，Map-reduce处理过程如图1所示。

1.4.3　基于LDA的旅客出行意图挖掘

LDA有两个先验参数α和β。参数α决定了旅客的意图概率先验分布，而参数β则描述某出行意图下的航线概率先验分布。最终通过LDA模型训练得到旅客-意图概率θ和意图-航线概率φ。用Gibbs采样估计两个未知参数，主要思想是贝叶斯估计。贝叶斯估计把待估计参数看作是服从某种先验分布的随机变量。

学习过程：给定一个旅客集合，旅客乘坐航线是可以观察到的已知变量，α和β根据经验给定，其他的变量Z(出行意图)、θ和φ都是未知的隐含变量，需要根据观察到的变量来学习估计的。根据LDA的图模型，可以写出所有变量的联合分布：

$$P(\vec{w}_m,\vec{Z}_m,\vec{\vartheta}_m,\Phi|\vec{\alpha},\vec{\beta})=\prod_{n=1}^{N_m}P(w_{m,n}|\vec{\varphi}_{Z_{m,n}})\cdot P(Z_{m,n}|\vec{\vartheta}_m)\cdot P(\vec{\vartheta}_m|\vec{\alpha})\cdot P(\Phi|\vec{\beta}) \quad (2)$$

其中：

$\vec{\vartheta}_m$—表示从狄利克雷分布α中取样生成旅客m的意图分布；

$Z_{m,n}$—表示从意图的多项分布ϑ_m取样生成旅客m第n个可选航线的意图；

$\vec{\varphi}_{Z_{m,n}}$—表示从狄利克雷分布β中取样生成意图$Z_{m,n}$对应的航线分布；

$w_{m,n}$—表示旅客m最终选择的航线n。

因为意图分布θ，确定具体意图，且β产生的航线分布φ，确定具体航线，所以式(2)等价于式(3)所表达的联合概率分布：

$$P(\vec{w},\vec{z}|\vec{\alpha},\vec{\beta})=P(\vec{w}|\vec{z},\vec{\beta})P(\vec{z}|\vec{\alpha}) \quad (3)$$

公式的物理含义为：第1项因子表示的是根据确定的意图和航线分布的先验分布参数采样航线的过程，第2项因子是根据意图分布的先验分布参数采样意图的过程，这两项因子是需要计算的两个未知参数。

根据推算得到$P(\vec{w},\vec{z})$的联合分布结果为：

$$P(\vec{w},\vec{z}|\vec{\alpha},\vec{\beta})=\prod_{z=1}^{K}\frac{\Delta(\vec{n}_z+\vec{\alpha})}{\Delta(\vec{\beta})}\cdot\prod_{m=1}^{M}\frac{\Delta(\vec{n}_m+\vec{\alpha})}{\Delta(\vec{\alpha})} \quad (4)$$

有了$P(\vec{w},\vec{z})$联合分布，便可以通过联合分布来计算在给定可观测变量w下的隐变量z的条件分布(后验分布)$P(\vec{w},\vec{z})$，进行贝叶斯分析。

先定义几个变量：¬i表示除去i的航线，$\vec{w}=\{w_i=t,\vec{w}_{\neg i}\}$，$\vec{z}=\{z_i=k,\vec{z}_{\neg i}\}$。排除当前航线的意图分配，即根据其他航线的意图分配和观察到的航线来计算当前航线的意图的概率公式为：

$$P(z_i=k|\vec{z}_{\neg i},\vec{w})\propto P(z_i=k,w_i=t|\vec{z}_{\neg i},\vec{w}_{\neg i})=\hat{\theta}_{mk}\cdot\hat{\varphi}_{kt} \quad (5)$$

经推导得到结果：

$$P(z_i=k|\vec{z}_{\neg i},\vec{w})\propto\frac{n_{m,\neg i}^{(k)}+\alpha_k}{\sum_{k=1}^{K}(n_{m,\neg i}^{(k)}+\alpha_k)}\cdot\frac{n_{k,\neg i}^{(t)}+\beta_t}{\sum_{t=1}^{V}(n_{k,\neg i}^{(t)}+\beta_t)} \quad (6)$$

其中，$\vec{n}_m$是旅客m的意图数向量，$\vec{n}_k$是意图k

下的航线项数向量。Gibbs Sampling 通过求解出意图分布和航线分布的后验分布，从而成功解决意图分布和航线分布这两参数未知的问题。

2 试验与分析

2.1 数据及其预处理结果

对中国民航信息股份有限公司（TravelSKY Technology Limited）2010—2011 年共计 48 G 的航班记录做预处理，将乘坐次数 5 次和 10 次以上的旅客记录借助 Map-reduce 算法在 Hadoop 分布式平台上分别筛选出来，然后以旅客身份证号（加密）将筛选过的航班记录的特征性信息（每个旅客乘坐过的航线）提取出来，整个文档中每一行表示一个旅客，行中每一词即是该旅客乘坐过的航班航线。

表 1 是原始数据经过 Map-reduce 平台进行数据预处理，筛选出来的乘坐次数大于等于 5 次的航班记录。第 1 行为旅客总数量，其他每行则是单一旅客乘坐过的航班航线记录。每条航线记录由 6 个大写英文字符构成，前 3 个字符是出发机场三字码，后 3 个则是到达机场三字码，如 NKGSZX 表示从南京（NKG）到深圳（SZX）的航线。

表1 经过Map-reduce预处理过的数据示例

旅客数(人)	5 563 527
旅客ID	乘坐航班
User1	NKGSZX PEKNKG NKGXIY NKGINC XIYNKG XIYNKG NKGXIY NKGPEK NKGPEK XIYNKG XIYUYN NKGPEK XILHET INCNKG

2.2 传统方法

基于航班次数统计的航线价值计算方法：统计各航线的旅客乘坐次数，进而计算 $P(a)$。

$$P(a)=\frac{\text{航线}a\text{的乘坐人数}}{\text{所有航线的总乘坐人数}} \tag{7}$$

2.3 评价指标

Jaccard index 又称为 Jaccard 相似系数（Jaccard similarity coefficient），用于比较有限样本集之间的相似性与差异性，Jaccard 系数值越大，样本相似度越高。

给定两个集合 A、B，Jaccard 系数定义为 A 与 B 交集的大小与 A 与 B 并集的大小的比值，定义如下：

$$J(A,B)=\frac{|A\cap B|}{|A\cup B|}=\frac{|A\cap B|}{|A|+|B|-|A\cap B|} \tag{8}$$

当集合 A，B 都为空时，J(A,B) 定义为 1。

真实列表通过统计航线热度（即航线乘坐次数）并倒序处理得到；而本文的航线预测列表则是首先通过公式（1）计算所有航线的价值，然后按照价值从高到低进行降序排列。为了更加彰显算法的性能优势，本文选择排序列表的前 Top K 个航线计算 Jaccard 系数。

2.4 实验设置

（1）参数 α，β 的设置

α 和 β 是 LDA 模型中旅客出行意图分布 θ 和意图下航线分布 φ 的先验分布的先验参数，这两个参数的设置会影响 θ 和 φ 的生成，从而影响最终航线价值，α 和 β 取值范围皆为 0.1 ~ 0.9。

（2）参数 k 的设置

k 值为主题数，其值会影响 θ 矩阵的列数和 φ 矩阵的行数，取值范围根据内存大小而定，实验中取值为：10、20、50、80、100。

2.5 实验结果

（1）热点航线挖掘

以航线潜在价值（预测序列）为键进行排序得到表 2，并与其对应真实航线乘坐次数进行对比，可以发现两者并不完全正比，说明航线潜在价值会受其他因素影响，传统算法具有一定局限性。

（2）性能对比

3 组最优 LDA 模型与传统算法的性能对比见表 3，可以发现，与传统算法相比，本文算法在 top100 内参数设置为 α=0.8, β=0.3, k=20 时指标提升 2%，但在 top500 和 top1 000 范围内性能与传统算法相差无几，可见这组模型在 top100 是性能表现最优。另外两组模型类似，分别在 top500 和 top1 000 内达到性能最优，并且性能指标分别高于传统算法 2% 和 1%，说明 LDA 模型可以通过调整先验参数可以挖掘不同范围内比传统算法更加有效的航线潜在价值。

2.6 实验参数分析

进行多组实验，将旅客乘坐次数大于 10 次以上和 5 次以上的筛选出来，作为 LDA 模型输入，并对结果进行统计分析，得到 top100、top500、top1 000 下各模型的性能，如表 4、表 5 所示。通过表 4 可以发现，本文算法在 top100 时能够取得 92% 的 Jacarrd

表2　航线潜在价值排名top20

排名	航线	潜在航线价值P(a)	航线真实乘坐次数
1	北京-上海	0.030 707 448 6	760 898
2	上海-北京	0.029 990 224 2	746 931
3	北京-广州	0.013 952 755 3	370 068
4	广州-北京	0.013 816 298 3	365 241
5	成都-北京	0.012 209 438 2	315 494
6	北京-成都	0.012 176 064 1	312 037
7	北京-深圳	0.011 345 224 2	294 627
8	深圳-北京	0.011 249 205 1	289 700
9	北京-西安	0.008 521 798 3	230 798
10	西安-北京	0.008 314 447 8	229 023
11	北京-南京	0.008 224 815 9	170 259
12	上海-广州	0.008 106 5475	212 778
13	南京-北京	0.008 094 153 8	162 955
14	杭州-北京	0.008 075 076 9	197 914
15	北京-杭州	0.008 065 348 1	197 749
16	广州-上海	0.008 022 710 6	212 626
17	北京-沈阳	0.006 579 759 7	179 051
18	大连-北京	0.006 400 538 2	159 566
19	北京-大连	0.006 359 000 4	157 407
20	北京-武汉	0.006 351 005 9	163 135

表3　实验性能对比

算法		Top100	Top500	Top1000
基于航班次数统计的旅客价值计算方法		0.904 762	0.883 239	0.899 335
本文算法	α=0.8, β=0.3, k=20	0.923 077	0.883 239	0.895 735
	α=0.6, β=0.9, k=10	0.869 159	0.901 141	0.886 792
	α=0.1, β=0.9, k=20	0.904 762	0.886 792	0.902 95

相似系数，说明算法有效。

通过实验数据发现，不同范围内算法性能也不一样，α、β、k 三者同时影响航线价值。可以看出，k=10 的记录占 top rank 的大部分（实验中，k=10、50、80、100），意味着旅客出行的潜在意图的数量占 10 个的概率极大。

表4　性能评估，旅客乘坐次数为10次以上

先验参数	Top K	Jaccard 系数
α=0.1, β=0.4, k=80	100	0.923 077
α=0.1, β=0.2, k=80	100	0.923 077
α=0.9, β=0.1, k=10	100	0.904 762
α=0.6, β=0.9, k=10	500	0.901 141
α=0.8, β=0.8, k=10	500	0.897 533
α=0.7, β=0.9, k=10	500	0.897 533
α=0.5, β=0.2, k=10	1000	0.901 141
α=0.2, β=0.6, k=10	1000	0.899 335
α=0.2, β=0.4, k=10	1000	0.899 335

表5　性能评估，旅客乘坐次数5次以上

先验参数	Top K	Jaccard 系数
α=0.5, β=0.1, k=100	100	0.769 912
α=0.5, β=0.1, k=100	500	0.834 862
α=0.5, β=0.1, k=100	1000	0.874 414

3　实验结果分析

通过实验可以发现，虽然传统算法计算的航线价值实现简单，并且准确度也不低，但其挖掘潜在价值的方式是基于航线的历史使用情况，所以预测效果存在偏差；而 LDA 与传统算法比较，准确度较高，能挖掘出更多的航线，并且算法模型可控，能够适应旅客基于多种潜在出行意图下的航线价值，同时具有可扩展性，可以通过词扩充来提高航线的概率。综上所述，LDA 算法对于挖掘潜在模式具有一定的优势。

4　结束语

本文从旅客出行行为的角度出发，将出行的不同因素归结为出行意图，从而利用出行意图得到航线数据。（1）构建了面向大规模民航旅客订票数据分析的 Map-reduce 平台，处理中航信 2010 年和 2011 年的 48 G 订票日志。（2）提出了基于旅客出行意图发现的潜在高价值航线挖掘算法，通过挖掘在大规模旅客订票日志的旅客出行意图，计算航线未来潜在概率，丰富了机场的航线营销和航空公司的航线网络布局技术。（3）面向大规模的 LDA 主题模型构建方法，丰富了主题模型构建方法，可拓展到其他大规模数据的主题模型中。

参考文献：

[1] 杨迎卯．城市轨道交通行为分析与数据挖掘决策系统研究[J]. 铁路计算机应用，2016，25（6）：65-69.

[2] 冯　宇，张　琪．浅谈新航线的开辟（一）[J]. 空运商务，2007（3）：12-15.

[3] 张永莉，张晓全．我国城市间航空客运量影响因素的实证分析 [J]. 经济地理，2007，27（4）：658-660.

[4] 吴薇薇，朱金福．航线开辟优选方案模型的集对分析研究 [J]. 工业技术经济，2009，28（7）：121-124.

[5] S Deerwester，ST Dumais，GW Furnas，TK Landauer，R

Harshman. Indexing by latent semantic analysis[J]. Journal of the Association for Information Science and Technology, 41(6):391-407, 1990.

[6] DM Blei，AY Ng，MI Jordan. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 3:993-1022, 2003.

[7] J Boydgraber，DM Blei. Syntactic Topic Models[C]//Advances in Neural Information Processing Systems, 185-192, 2010.

[8] J Qiang，P Chen，T Wang，X Wu. Topic Modeling over Short Texts by Incorporating Word Embedding[C]//Advances in Knowledge Discovery and Data Mining, 2017: 363-374.

[9] S Lacoste-Julien，S Fei，MI Jordan. DiscLDA: Discriminative Learning for Dimensionality Reduction and Classification[C]// Proceedings of NIPS Neural Information Processing, 897-904, 2008.

[10] I Bíró，J Szabó. Large scale link based latent Dirichlet allocation for web document classification[C]// Computer Science, 2010.

[11] M. Rosen-Zvi, T. Griffiths, M. Steyvers, P. Smyth. The author-topic model for authors and documents[C]//Proceedings of the 20th Annual Conference on Uncertainty in Artificial Intelligence, 2004.

[12] Yan Liu, Alexandru Niculescu-Mizil. Topic-Link LDA: Joint Models of Topic and Author Community[C]//Proceedings of the 26th International Conference on Machine Learning, 2009.

[13] David M. Blei, John D. Lafferty . Correlated Topic Models[C]// In NIPS, 2005.

[14] J Paisley, C Wang, DM Blei, MI Jordan. Nested Hierar-chical Dirichlet Processes[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 37(2):256-70, 2014.

[15] David Blei, Jon Mcauliffe. Supervised Topic Models[C]//In Advances in Neural Information Processing Systems,2008.

[16] K Than, BH Tu. Inference in topic models: sparsity and trade-off[C]//Statistics, 2015

[17] K Henderson, T Eliassi-Rad. Applying latent dirichlet allocation to group discovery in large graphs[C]//Acm Symposium on Applied Computing,1456-1461,2009.

[18] H Zhang, B Qiu, CL Giles, HC Foley. An LDA-based Community Structure Discovery Approach for Large-Scale Social Networks[C]//IEEE International Conference on Intelligence & Security Informatics. 200-207, 2007.

[19] 于　辉，陈敬光．基于 CVaR 的航空公司机票超售策略 [J]. 工业工程，2012，15（1）：1-7.

[20] 顾兆军，王　伟，李晓红．基于潜在类别模型的航空旅客分类 [J]. 计算机技术与发展，2012，22（4）：182-186.

[21] 潘玲玲，张育平，徐　涛．核 DBSCAN 算法在民航客户细分中的应用 [J]. 计算机工程，2012，38（10）：70-73.

[22] Z.H. Wu, Y.F. Lin, et al. Balanced multi-label propagation for overlapping community detection in social networks[J]. Journal of Computer Science and Technology, 27（3）：468-479, 2012.

[23] 林友芳，王天宇，唐　锐，等．一种有效的社会网络社区发现模型和算法 [J]. 计算机研究与发展，2012，49（2）：337-345.

[24] H.Y. Wan, Y.F. Lin, C.Y. Jia, H.K. Huang. Community-based relational Markov networks in complex networks[C]//In Proceedings of the 6th international conference on Rough sets and knowledge technology, pages 301-310, 2011.

[25] Z.H. Wu, Y.F. Lin, et al. Efficient overlapping community detection in huge real-world networks[J]. Physica A: Statistical Mechanics and its Applications, 391(7):2475-2490, 2012.

[26] X. Feng, B.Y. Xu, M. Lu and C.Y. Zhang. Infrequent Passenger Value Discovery by Random Walk on Passenger-route Heterogenzeous Network[J]. Journal of Computational and Theoretical Nanoscience, 2(1):10-17,2015.

责任编辑　付　思