DOI:10.16356/j.1005-2615.2018.05.003

基于 LDA 的航线潜在价值挖掘模型

徐 涛[1,2,3] 徐召朋[1] 卢 敏[1,2,3]

(1.中国民航大学计算机科学与技术学院,天津,300300; 2.中国民航信息技术科研基地,天津,300300;
3.民航旅客服务智能化应用技术重点实验室,北京,101300)

摘要:在分析了传统的主题模型后提出了一种基于 LDA 的航线潜在价值挖掘模型。该模型将旅客出行行为的分析划分成两个阶段,出行意图的确定及出行意图下航线的选择,并与旅客价值进行融合来挖掘航线的潜在价值。出行意图采用 Gibbs sampling 方法从旅客出行记录中获取,航线则在旅客确定出行意图后由出行意图的航线向量获得,旅客价值则结合出行中的舱位信息进行提取。在中国民航旅客订票数据集上的实验表明,本文模型在 2010 年和 2011 年两个数据集上获得的两组航线潜在价值序列比 pLSI 模型和 senLDA 模型获得的两组航线潜在价值序列都拥有更好的有序相关性,且在挖掘排名前 5 的航线潜在价值时,本文模型在该两个数据集上获得了两组完全一致的航线潜在价值序列,表明其在挖掘高潜在价值航线方面的优势。

关键词:航线价值;主题模型;潜在价值;出行意图

中图分类号:TP399 **文献标志码**:A **文章编号**:1005-2615(2018)05-0595-06

Air Routes Potential Value Mining Model Based on LDA

XU Tao[1,2,3], *XU Zhaopeng*[1], *LU Min*[1,2,3]

(1. College of Computer Science and Technology, Civil Aviation University of China, Tianjin, 300300, China;
2. Information Technology Research Base of Civil Aviation Administration of China, Tianjin, 300300, China;
3. Key Laboratory of Intelligent Application Technology for Civil Aviation Passenger Services, Beijing, 101300, China)

Abstract: Aiming at the problem of the value of air routes in the civil aviation route network, this paper proposes an air routes potential value mining model based on LDA by analyzing the traditional theme model. This model divides the analysis of passengers' travel behavior into two stages: The determination of travel intentions, and the selection of air routes implied in travel intentions, and they are incorporated with passenger value to mine air routes potential value. Travel intentions are extracted from passenger booking data by Gibbs sampling method, and air routes are obtained from the air routes vector through determining the travel intentions of passengers. Passenger values are obtained from the information of cabin. Experiments on passenger booking data sets of China Civil Aviation in 2010 and 2011 respectively show that the two air routes potential value sequences obtained by proposed model have better orderly correlation than the pLSI model and senLDA model, and when mining the potential value of the top 5 air routes, we get two identical air routes potential value sequences on two data sets in 2010 and 2011. Therefore, the proposed model has superiority in mining high potential value of air routes.

Key words: air routes value; theme model; potential value; travel intention

基金项目:国家自然科学基金(61502499)资助项目;中国民航科技创新引导资金重大专项(MHRD20140105)资助项目。
收稿日期:2018-05-30;**修订日期**:2018-08-30

作者简介:徐涛,男,教授,博士生导师,现任中国民航信息技术科研基地(由中国民用航空局设立)主任。公开发表学术研究论文 90 余篇,其中 SCI/EI 收录 50 篇,编写出版教材(专著)2 部,申请国家发明专利 10 项(其中已授权 4 项)。获教育部霍英东基金会第六届全国高等院校青年教师奖等。
通信作者:徐涛,E-mail:xutao@nuaa.edu.cn。

引用格式:徐涛,徐召朋,卢敏. 基于 LDA 的航线潜在价值挖掘模型[J]. 南京航空航天大学学报,2018,50(5):595-600. XU Tao, XU Zhaopeng, LU Min. Air routes potential value mining model based on LDA[J]. Journal of Nanjing University of Aeronautics & Astronautics,2018,50(5):595-600.

日益激烈的市场竞争迫使航空公司争相开辟新航线或加大热门航线的运力来发展航线网络以增强市场竞争力。当前,航线价值计算常用的方法是统计航线上的客流量,而仅具有统计特征的航线客流量方法并没有将旅客按照其出行特征和消费能力进行细分,导致航线价值的计算过于片面。因此,为了避免航空公司因航线后期运力不足引起的效益降低,建立一个合理的航线潜在价值挖掘模型对航空公司具有十分重要的意义。

目前,围绕民航运输中航空公司所涉及的业务领域内的价值挖掘研究工作主要有:(1)航空公司的收益管理。文献[1]利用两阶段法来评估航空公司期望收益的合理性,帮助航空公司建立一个更好的期望收益模型来提高航空公司的营收能力。(2)航空服务评价模型。文献[2]利用 Twitter 上的数据构建了一个基于多重分类方法的多数投票模型来验证航空公司的服务效果,文献[3]则主要从旅客满意度方面来评价航空公司的服务质量,上述两种方法都是从旅客的角度出发,利用旅客对航空公司的评价来监督航空公司的服务质量,进而为改善航空公司服务质量提供更好的建议。(3)航线网络研究。文献[4]分别从全球航线网络、航空公司联盟以及航空公司 3 个层面来分析全球航线网络的鲁棒性,并给出了优化航线网络的建议;文献[5]将民航与其竞争行业相融合,讨论了高铁对民航未来发展的冲击,分析了航空公司如何优化航线网络结构以促进共同发展。(4)民航旅客的研究。文献[6]利用 DBSCAN 算法对旅客的消费行为进行细分,文献[7]则是与 Hadoop 技术相结合提出了一种 TCSDG 的旅客聚类算法来分析旅客群体特征,以使航空公司能够更好地为旅客提供个性化服务。

上述研究都是通过分析民航运输中航空公司涉及的业务领域特点,并针对其特点做出相应的价值挖掘以提高航空公司的效益和服务能力。然而,这些研究并没有考虑从航线的角度出发进行价值挖掘研究,从而为航空公司的航线价值评估与决策问题提供参考。针对这一现状,本文利用中国民航旅客订票数据建立基于 LDA(Latent Dirichlet allocation)[8]的航线潜在价值挖掘模型来评估航线的价值,以期给航空公司在航线价值的评估和决策中提供有意义的参考。

1 基于 LDA 的航线潜在价值挖掘模型

主题模型的起源是隐性语义索引(Latent semantic indexing,LSI)[9]。隐性语义索引通过奇异值分解(Singular value decomposition,SVD)构造一个新的隐性(Latent sematic)语义空间[10]。该空间比原空间维度低,使文档可以降维,找到更简单的表达。实际上,隐性语义索引并不是真正意义上的主题模型,但其基本思想为主题模型的发展奠定了基础。

在 LSI 的基础上,Hofmann 提出了概率隐性语义索引(Probabilistic latent semantic indexing,pLSI)[10],该模型被视为真正意义的主题模型。概率隐性语义索引在 LSI 的基础上引入概率论知识,将其扩展为一个概率生成模型,因此可以用文档生成过程来解释 LSI,将不同类型的语义结构和语法角色引入到 LSI 模型中。此外,相比于 LSI 模型,pLSI 模型具有更好的扩展性。

Blei 等在 pLSI 的基础上,用一个服从 Dirichlet 分布的 K 维隐含随机变量表示文档的主题概率分布来模拟文档的产生过程,进而提出了 LDA 主题模型[8]。LDA 模型可以看作是对 pLSI 进行了贝叶斯化,使得参数变成了具有概率分布的随机变量,即对 pLSI 的参数添加了先验分布。LDA 模型的特点是利用参数的先验分布对其做最大后验概率估计,以提高模型的准确率。

在 LDA 的基础上 Yohan 等以文章中的句子为单位,提出了 senLDA 模型[11]。该模型假设了句子之间的词对文章潜在主题有很强的依赖性。使主题模型从词的维度扩展到了句的维度。

航线潜在价值挖掘模型的关键是通过旅客出行过程中的出行意图来计算旅客对航线的偏好,而出行意图并不能在旅客订票数据集中直接获得。因此借鉴主题模型的思想,从旅客订票数据集中抽取旅客的出行记录并处理成短文本来挖掘旅客的出行意图,从而实现对航线潜在价值的挖掘。

1.1 航线潜在价值的定义

借助贝叶斯公式将旅客出行时对航线偏好产生的价值以及选择舱位时产生的价值整合到航线潜在价值挖掘中,于是,航线 r 的潜在价值 $P(r)$ 可定义为

$$P(r)=\sum_{u}P(r\mid u)P(u) \tag{1}$$

式中:$P(r|u)$ 表示旅客 u 对航线 r 的偏好所产生的价值;$P(u)$ 表示旅客 u 选择舱位时所产生的价值。

由于旅客对航线的偏好所产生的价值并不能从旅客订票数据集中直接获取,然而通过从该数据集中抽取具有旅客特征的出行记录,可以将旅客出行记录表示为具有旅客特征的短文本,然后借鉴 LDA 模型中文本主题的思想引入出行意图的概念将旅客出行行为的分析划分为两个阶段,出行意图的确定以及在出行意图下航线的选择。于是,将式

(1)扩展为

$$P(r)=\sum_{u}\sum_{z}P(r\mid z)P(z\mid u)P(u) \quad (2)$$

式中:$P(r|z)$表示出行意图 z 中出现航线 r 的概率;$P(z|u)$表示旅客 u 拥有出行意图 z 的概率。

1.2　航线潜在价值挖掘求解

1.2.1　旅客价值计算

通常,航空公司会对航班中的座位进行分类,也即划分舱位等级。不同舱位的票价有高低之分,旅客在出行过程中选择不同的舱位不仅体现了旅客的消费能力还会对航线的价值产生一定的影响。因此,为更加全面计算旅客价值,不仅需要考虑其出行次数,还需考虑其对舱位的选择。式(2)中 $P(u)$的计算可表示为

$$P(u)=\frac{\sum_{c}\gamma_c n_u^c}{\sum_{u}\sum_{c}n_u^c} \quad (3)$$

式中:n_u^c 表示旅客 u 乘坐舱位 c 的次数;γ_c 表示舱位 c 的系数。

为使考虑舱位选择的旅客价值计算更符合航空公司的业务特征,式(3)中的舱位系数 γ_c 由航空公司对不同舱位的里程累积系数来代替。

1.2.2　旅客出行意图的挖掘及求解

旅客对不同航线的偏好是由其出行目的所决定的,不同的出行目的下旅客会选择不同的航线。旅客的出行目的便是本文所提出的旅客出行意图。出行意图由航线以及航线出现在该出行意图下的概率表示。

旅客的出行意图在订票数据集中无法直接获取,需要从旅客订票数据集中生成旅客出行记录并采样获得。假设每位旅客 u 都具有包含自身特点的出行意图向量,记为 $\boldsymbol{\theta}_u$,该向量中的元素是旅客 u 选择不同出行意图的概率值;此外,假设共有 K 种出行意图,出行意图集合为 Z,出行意图 $z\in Z$ 下的航线向量记为 $\boldsymbol{\varphi}_z$,由不同航线在出行意图 z 中所占比率表示。为方便起见,设 R 表示所有航线集合,将每条航线 $r\in R$ 都进行编号。于是,旅客 u 在出行中选择某条航线 r 时产生的价值为

$$P(r\mid u)=\sum_{z=1}^{K}P(z\mid u)P(r\mid z)=\sum_{z=1}^{K}\boldsymbol{\theta}_{uz}\boldsymbol{\varphi}_{zr} \quad (4)$$

由于向量 $\boldsymbol{\theta}_u$ 和 $\boldsymbol{\varphi}_z$ 本身属于变量,依据贝叶斯理论,其自身具有先验分布。选择 Dirichlet 分布[8]作为 $\boldsymbol{\theta}_u$ 和 $\boldsymbol{\varphi}_z$ 的先验分布,将 $\boldsymbol{\theta}_u$ 和 $\boldsymbol{\varphi}_z$ 进行贝叶斯化。设 $\boldsymbol{\theta}_u$ 和 $\boldsymbol{\varphi}_z$ 的先验参数分别为 $\boldsymbol{\alpha}$ 和 $\boldsymbol{\beta}$。旅客出行记录的生成是基于旅客具有出行意图以及旅客在出行意图下具有航线偏好的假设,因此只要求解出参数 $\boldsymbol{\theta}_u$ 和 $\boldsymbol{\varphi}_z$,便可获得旅客选择航线时对航线产生的价值,进而与旅客自身价值相结合来挖掘航线的潜在价值。

旅客出行记录中已知的数据是旅客 u 出行时选择的航线 r,出行意图 z 是隐含的变量,即旅客订票数据集中抽取的旅客出行记录是航线和出行意图的联合分布。由于旅客出行意图向量 $\boldsymbol{\theta}_u$ 和出行意图下的航线向量 $\boldsymbol{\varphi}_z$ 都以 Dirichlet 分布作为先验分布,所以航线与出行意图的联合分布可表示为

$$P(R,Z\mid\boldsymbol{\alpha},\boldsymbol{\beta})=\prod_{u}\frac{\Delta(\boldsymbol{\theta}_u+\boldsymbol{\alpha})}{\Delta(\boldsymbol{\alpha})}\prod_{z}\frac{\Delta(\boldsymbol{\varphi}_z+\boldsymbol{\beta})}{\Delta(\boldsymbol{\beta})} \quad (5)$$

式中:$\Delta(\boldsymbol{x})=\dfrac{\prod_{i=1}^{\dim\boldsymbol{x}}\Gamma(x_i)}{\Gamma(\sum_{i=1}^{\dim\boldsymbol{x}}x_i)}$;$\Gamma(\cdot)$是 Gamma 函数,$\dim\boldsymbol{x}$ 表示向量 $\boldsymbol{x}$ 的维数。

从上述联合分布中获得旅客的出行意图需要通过 Gibbs sampling[10,11]方法从旅客出行记录中进行采样。通过 Dirichlet-multinomial 共轭[8]关系推导出的 Gibbs sampling 公式为

$$P(z\mid Z_{\neg i},R)\propto\frac{\text{theme\_line}_{z,\neg i}^{(r)}+\beta_r}{\sum_{r=1}^{|R|}(\text{theme\_line}_{z,\neg i}^{(r)}+\beta_r)}\times\frac{\text{user\_theme}_{u,\neg i}^{(z)}+\alpha_z}{\sum_{z=1}^{K}(\text{user\_theme}_{u,\neg i}^{(z)}+\alpha_z)} \quad (6)$$

式中:$\text{theme\_line}_{z,\neg i}^{(r)}$ 表示出行意图 z 中剔除第 i 条航线($i=1,2,\cdots,|R|$)后航线 r 出现的次数;$\text{user\_theme}_{u,\neg i}^{(z)}$ 表示旅客 u 中剔除第 i 条航线后出行意图 z 出现的次数。利用 Gibbs sampling 方法挖掘旅客出行意图的过程如表 1 所示。

表 1　旅客出行意图挖掘的 Gibbs sampling 方法

Tab. 1　Gibbs sampling method for passenger travel intention mining

输入:由旅客订票数据集生成的旅客出行记录,出行意图 z,航线集合 R
输出:$\boldsymbol{\theta}_u$,$\boldsymbol{\varphi}_z$
步骤 1:遍历航线集合 R,对每一旅客 u 所乘坐的每条航线 r,根据出行意图 z,更新 $\text{user\_theme}_{u,\neg i}^{(z)}$ 和 $\text{theme\_line}_{z,\neg i}^{(r)}$,($i=1,2,\cdots|R|$)。
步骤 2:扫描旅客出行记录集合,用式(6)对每条航线 r 更新其在出行意图 z 下出现的概率。
步骤 3:重复步骤 1,2,直到每条航线 r 在出行意图 z 下出现的概率不再变化。
步骤 4:生成每一旅客 u 的出行意图向量 $\boldsymbol{\theta}_u$ 以及出行意图 z 下的航线向量 $\boldsymbol{\varphi}_z$。

通过表 1 给出的 Gibbs sampling 过程可获得任意旅客 u 选择某一出行意图 z 的概率 $\theta_{uz}\in\boldsymbol{\theta}_u$ 以及某一出行意图 z 下航线 r 出现的概率 $\varphi_{zr}\in\boldsymbol{\varphi}_z$,即

$$\theta_{uz}=\frac{\text{user\_theme}_{u,\neg i}^{(z)}+\alpha_z}{\sum_{z=1}^{K}(\text{user\_theme}_{u,\neg i}^{(z)}+\alpha_z)} \tag{7}$$

$$\varphi_{zr}=\frac{\text{theme\_line}_{z,\neg i}^{(r)}+\beta_r}{\sum_{r=1}^{|R|}(\text{theme\_line}_{z,\neg i}^{(r)}+\beta_r)} \tag{8}$$

将式(7,8)融入到式(4)中,可得旅客 u 选择航线 r 时所产生的潜在价值,即

$$P(r\mid u)=\sum_{z}\frac{\text{them\_line}_{z,\neg i}^{(r)}+\beta_r}{\sum_{r=1}^{|R|}(\text{them\_line}_{z,\neg i}^{(r)}+\beta_r)}\times\frac{\text{user\_theme}_{u,\neg i}^{(z)}+\alpha_z}{\sum_{z=1}^{K}(\text{user\_theme}_{u,\neg i}^{(z)}+\alpha_z)} \tag{9}$$

最终,结合式(3,9),由式(1)可得基于 LDA 的航线潜在价值挖掘模型为

$$P(r)=\sum_{u}\sum_{z}\frac{\text{them\_line}_{z,\neg i}^{(r)}+\beta_r}{\sum_{r=1}^{|R|}(\text{them\_line}_{z,\neg i}^{(r)}+\beta_r)}\times\frac{\text{user\_theme}_{u,\neg i}^{(z)}+\alpha_z}{\sum_{z=1}^{K}(\text{user\_theme}_{u,\neg i}^{(z)}+\alpha_z)}\times\frac{\sum_c\gamma_c n_u^c}{\sum_u\sum_c n_u^c} \tag{10}$$

基于 LDA 的航线潜在价值挖掘模型相较于传统的基于旅客客流量统计的航线价值挖掘方法,不仅通过旅客出行时的舱位选择来计算其自身的价值,还将旅客出行行为的分析划分为旅客出行意图的确定及出行意图下航线的选择两个阶段,并将这两个阶段分别进行量化后与旅客价值相结合最终实现对航线潜在价值的挖掘计算。

2 实验及结果分析

2.1 实验数据及预处理

实验选取中国民航旅客订座系统中 2010 年 1 月 1 日至 2011 年 12 月 31 日两年的旅客订票数据,其数据量是 48.9 GB。包含订票记录数 102 305 312条,旅客 96 298 451 人,航线 1 634 条。该数据的具体内容包含身份证号、性别、所选航空公司、航班号、舱位、起飞机场以及到达机场等 17 个属性。

在计算旅客对航线的偏好时用经济舱信息补全舱位空缺信息。此外,旅客订票数据集中并不包含航线信息,可采用从旅客订票数据信息中进行提取的方案,用起飞机场和到达机场这两个属性来标识航线。

基于 LDA 的航线潜在价值挖掘模型利用旅客的出行意图来挖掘航线的潜在价值,如果旅客的出行次数达不到一定的要求,可视为其出行意图过于单一,进而导致航线潜在价值挖掘性能的降低。因此筛选出行次数 30 次及以上的旅客作为基准实验数据,其筛选后的数据规模如表 2 所示。

表 2 年出行次数 30 次及以上的旅客数据规模

Tab. 2 30 times and above data sets for annual trips

出行次数	年份	记录数	旅客数
30	2010	6 421 320	201 317
	2011	7 071 129	232 173

2.2 实验的评价指标

本文采用 TopN 的方法来评价不同出行意图数目下挖掘的航线潜在价值序列的有序相关性。选取常见的两种有序相关性度量方法:肯德尔相关系数(Kendall rank correlation coefficient)和斯皮尔曼相关系数(Sperman's rank correlation coefficient)。

(1)肯德尔相关系数

记 τ 为肯德尔相关系数,简称肯德尔系数,是衡量两个待测序列之间有序相关性的一种度量标准。其取值范围在 -1 到 1 之间,当 τ 为 1 时,表示两个待测序列拥有一致的有序相关性;当 τ 为 -1时,表示两个待测序列拥有完全相反的有序相关性;当 τ 为 0 时,表示两个待测序列不具有相关性。其计算公式为

$$\tau=\frac{s-d}{n(n-1)/2} \tag{11}$$

式中:s 表示待测序列排序一致的对数;d 表示待测序列排序不一致的对数;n 表示两个待测序列的长度。

(2)斯皮尔曼相关系数

记 r_s 为斯皮尔曼相关系数,是一种非参数的有序相关性的度量标准,其取值范围和含义与肯德尔相关系数一致。其计算公式为

$$r_s=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{12}$$

式中:$x_i\in x$,$y_i\in y$;x,y 表示两个待测的序列,$\bar{x}$,$\bar{y}$ 表示待测序列 x,y 的平均值,n 表示待测序列的长度。

2.3 实验结果及分析

基于不同的出行意图数目,在表 2 筛选出的 2010 和 2011 年旅客订票数据集上,将基于 LDA 的航线潜在价值挖掘模型(简称本文模型)与 pLSI 的模型及 senLDA 模型进行对比实验,并采用肯德尔相关系数和斯皮尔曼相关系数比较衡量两种模型挖掘出的航线潜在价值序列的有序相关性。实验的主要过程如下:

(1)利用本文模型分别计算 2010 年旅客订票

数据集与 2011 年旅客订票数据集上的航线潜在价值序列,并降序排列。

(2)利用基于 pLSI 的模型分别计算 2010 年旅客订票数据集与 2011 年旅客订票数据集上的航线潜在价值序列,并降序排列。

(3)利用基于 senLDA 的模型分别计算 2010 年旅客订票数据集与 2011 年旅客订票数据集上的航线潜在价值序列,并降序排列。

(4)计算步骤(1)中所得两组航线潜在价值序列的肯德尔相关系数和斯皮尔曼相关系数。

(5)计算步骤(2)中所得两组航线潜在价值序列的肯德尔相关系数和斯皮尔曼相关系数。

(6)计算步骤(3)中所得两组航线潜在价值序列的肯德尔相关系数和斯皮尔曼相关系数。

实验中设置基于 LDA 的航线潜在价值挖掘模型的出行意图的先验分布参数 $\boldsymbol{\alpha}$ 固定为 $50/K$,出行意图中航线的先验分布参数 $\boldsymbol{\beta}$ 固定为 0.01,出行意图的数目分别设置为 10,30,50 以及 100。不同出行意图数目下挖掘的航线价值序列的有序相关性实验结果如表 3—6 所示。

表 3　pLSI 模型、senLDA 与本文模型的航线价值的有序相关性(出行意图数目 10)

Tab. 3　The orderly correlation between pLSI model, senLDA and the model of this paper (The number of traval intention 10)

模型	Top5		Top20		Top50		Top100	
	τ	r_s	τ	r_s	τ	r_s	τ	r_s
pLSI	0.632	0.808	0.618	0.634	0.580	0.737	0.612	0.625
senLDA	0.552	0.713	0.531	0.577	0.411	0.633	0.583	0.617
本文模型	1.000	1.000	0.619	0.832	0.668	0.843	0.631	0.806

表 4　pLSI 模型、senLDA 与本文模型的航线价值的有序相关性(出行意图数目 30)

Tab. 4　The orderly correlation between pLSI model, senLDA and the model of this paper (The number of traval intention 30)

模型	Top-N							
	Top5		Top20		Top50		Top100	
	τ	r_s	τ	r_s	τ	r_s	τ	r_s
pLSI	0.637	0.828	0.628	0.664	0.583	0.757	0.622	0.627
senLDA	0.543	0.747	0.537	0.583	0.401	0.581	0.557	0.584
本文模型	1.000	1.000	0.684	0.849	0.797	0.903	0.729	0.851

表 5　pLSI 模型、senLDA 与本文模型的航线价值的有序相关性(出行意图数目 50)

Tab. 5　The orderly correlation between pLSI model, senLDA and the model of this paper (The number of traval intention 50)

模型	Top-N							
	Top5		Top20		Top50		Top100	
	τ	r_s	τ	r_s	τ	r_s	τ	r_s
pLSI	0.627	0.818	0.613	0.652	0.571	0.713	0.608	0.622
senLDA	0.471	0.673	0.491	0.533	0.437	0.646	0.575	0.603
本文模型	1.000	1.000	0.623	0.817	0.703	0.852	0.672	0.839

表 6　pLSI 模型、senLDA 与本文模型的航线价值的有序相关性(出行意图数目 100)

Tab. 6　The orderly correlation between pLSI model, senLDA and the model of this paper(The number of traval intention 100)

模型	Top-N							
	Top5		Top20		Top50		Top100	
	τ	r_s	τ	r_s	τ	r_s	τ	r_s
pLSI	0.620	0.798	0.601	0.631	0.552	0.701	0.592	0.602
senLDA	0.466	0.621	0.509	0.517	0.397	0.611	0.562	0.593
本文模型	0.617	0.831	0.613	0.803	0.672	0.841	0.612	0.827

从表 3—6 的实验结果对比中可以看出,无论出行意图数目多少,基于 LDA 的航线潜在价值挖掘模型挖掘到的航线潜在价值序列比 pLSI 及 senLDA 模型均拥有更高的有序相关性,表明本文模型在 2010 年与 2011 年旅客订票数据集上获得的两组航线潜在价值序列具有更好的一致性。当出行意图数目分别为 10,30,50 时,该模型挖掘的航线潜在价值序列在采用 Top5 方法评价时,肯德

尔相关系数和斯皮尔曼相关系数都达到了1,即本文模型在2010年与2011年旅客订票数据集上获得的两组排名前5的航线潜在价值序列是完全一致的,表明本文模型在挖掘高潜在价值航线方面的优势。这是因为本文模型在挖掘航线的潜在价值时不仅考虑了旅客出行意图的先验分布,也考虑了出行意图下航线的先验分布。而基于pLSI的模型仅考虑出行意图下航线的先验分布,使得在获得旅客出行意图时缺少了先验知识,所以本文模型在挖掘航线潜在价值序列有序相关性方面的性能总体高于基于pLSI的模型。由于senLDA模型从句子的维度来考虑文档的主题,而旅客出行记录所组成的文本并无明确的语义关系,所以其性能相交于其他两种模型反而更差。

在实验结果中也可发现,在不同的出行意图数目下,基于LDA的航线潜在价值挖掘模型与pLSI模型、senLDA模型的肯德尔相关系数和斯皮尔曼相关系数也不相同。随着出行意图数目的增多,其肯德尔相关系数与斯皮尔曼相关系数呈现先增后减的趋势。当出行意图的数目为30时,本文模型和基于pLSI模型的肯德尔相关系数和斯皮尔曼相关系数都达到一个峰值,其后随着出行意图数目的增加,肯德尔相关系数和斯皮尔曼相关系数均随之下降。这是因为从某种程度上讲,旅客的出行意图数目代表了航线潜在属性的类别,当出行意图数目过大时,会造成航线的过度分类,导致航线出现在某些出行意图下的概率值过低,使得旅客在某些出行意图下的航线选择成为小概率事件,从而影响模型挖掘航线潜在价值时的性能。这样的实验结果从侧面也表明了航线的潜在价值并不是无限可挖掘的。

3 结 论

本文针对民航航线网络中航线价值挖掘的问题提出了基于LDA的航线潜在价值挖掘模型。该模型通过Gibbs sampling方法来获得旅客的出行意图,并将这种出行意图赋给航线,成为航线的潜在属性,进而计算航线的潜在价值。从实验结果可见,本文提出的模型不仅可以挖掘出航线的潜在价值,并在挖掘高潜在价值航线方面具有较明显的优势,对航空公司进行航线评估与决策具有较大的参考意义。然而,文中实验假设旅客在选择各条航线的过程中都是相互独立的,并没有考虑先后关系。在实际出行中,通常会出现旅客购买往返机票或联程机票等情形,而本文模型并没有对这一情况进行区分和处理。今后的工作将进一步考虑旅客出行时所选航线的先后关系等情形来挖掘航线的潜在价值,使模型更加符合旅客的实际出行情况。

参考文献:

[1] CHEN L, HOMEM-DE-MELLO T. Resolving stochastic programming models for airline revenue management[J]. Annals of Operations Research, 2010, 177(6):91-114.

[2] WAN Y, GAO Q. An ensemble sentiment classification system of twitter data for airline services analysis [C]//2015 IEEE International Conference on Data Mining Workshop (ICDMW). Atlantic City: IEEE, 2015:1318-1325.

[3] SUKI N M. Passenger satisfaction with airline service quality in malaysia: A structural equation modeling approach[J]. Research in Transportation Business & Management, 2014,10(4):26-32.

[4] LORDAN O, SALLAN J M, SIMO P. Study of the topology and robustness of airline route networks from the complex network approach: A survey and research agenda[J]. Journal of Transport Geography, 2014,37(8):112-120.

[5] JIANG C, ZHANG A. Airline network choice and market coverage under high-speed rail competition [J]. Transportation Research Part A: Policy and Practice, 2016,92(10):248-260.

[6] 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.
PAN Lingling, ZHANG Yuping, XU Tao. Application of kernel DBSCAN algorithm in civil aviation customer segmentation[J]. Computer Engineering, 2012,38(10):70-73.

[7] 冯霞,徐冰宇, 卢敏. 民航旅客订票行为细分及群体特征分析[J]. 计算机工程与设计, 2015,36(8):2217-2222.
FENG Xia, XU Bingyu, LU Min. Booking behavior subdivision and characteristics analysis of civil aviation passenger[J]. Computer Engineering and Design, 2015,36(8):2217-2222.

[8] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003,3(1):993-1022.

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

[10] HOFMANN T. Probabilistic latent semantic analysis [C]//The Fifteenth Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers Inc, 1999:289-296.

[11] YOHAN J, OH A H. Aspect and sentiment unification model for online review analysis [C]//ACM International Conference on Web Search and Data Mining. New York: ACM, 2011:815-824.

(编辑:刘彦东)