中国科学: 信息科学 2017 年 第 47 卷 第 8 期: 1023–1035
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
从大数据到大知识工程专刊 · 论文

基于用户在线查询行为的民航异常需求发现

许强永[1], 林友芳[1], 万怀宇[1*], 吴丽娜[2], 贾旭光[2]

1. 北京交通大学计算机与信息技术学院交通数据分析与挖掘北京市重点实验室, 北京 100044
2. 中国民航信息网络股份有限公司, 北京 100010
* 通信作者. E-mail: hywan@bjtu.edu.cn

收稿日期: 2016–11–27; 接受日期: 2017–03–06; 网络出版日期: 2017–07–04
国家自然科学基金 (批准号: 61403028)、教育部 – 中国移动科研基金 (批准号: MCM20150513) 和中央高校基本科研业务费 (批准号: 2016JBM017) 资助项目

摘要 在线机票预订网站上的用户查询量变化反应了民航市场需求的变化. 通过对用户在线查询大数据的分析, 可以及时准确地发现异常的民航需求, 有利于机票代理及航空公司做出快速的市场反应. 本文提出了一种新颖的民航异常需求发现方法, 基于不同航线的用户查询量时间序列, 并利用全国航线网络, 从网络整体而非单条航线的视角来检测民航需求异常. 基于某在线订票网站的真实查询数据集进行了实验, 表明本文提出的方法能够有效地从用户查询行为记录中及时发现民航异常需求.

关键词 民航需求, 在线机票查询, 用户行为分析, 异常行为检测, 时间序列曲线

1 引言

近年来, 我国民航业的发展十分迅猛. 从 2012 年到 2016 年, 我国民航旅客运输量从 3.19 亿人次增加到 4.88 亿人次, 增加了 52.98%, 年均增长达到 10.60% [1]. 随着民航市场的持续增长, 对航空公司、机场和机票代理等民航相关企业的管理和运营水平, 尤其是市场反应能力提出了更高的要求. 这就要求这些民航企业能够及时准确地掌握民航市场需求的变化, 及时采取相应的市场对策, 从而提高企业的运营能力和服务质量, 提高收益, 改善用户出行体验.

从长期来看, 民航市场需求总体上呈现自然增长态势, 但在短期范围内, 周末、节假日、大型活动、社会事件和自然灾害等, 都有可能给民航市场需求带来较大的波动或异常. 在大多数情况下, 当人们有出行需求时, 通常会提前一定时间进行机票查询. 因此, 机票的查询量能够在很大程度上反应出真实的民航市场需求. 如果能够对机票查询量的变化情况进行有效的分析, 将会及时准确地掌握民航需求异常.

引用格式: 许强永, 林友芳, 万怀宇, 等. 基于用户在线查询行为的民航异常需求发现. 中国科学: 信息科学, 2017, 47: 1023–1035, doi: 10.1360/N112016-00268
Xu Q Y, Lin Y F, Wan H Y, et al. Discovering abnormal civil aviation requirements by analyzing users' online query behaviors (in Chinese). Sci Sin Inform, 2017, 47: 1023–1035, doi: 10.1360/N112016-00268

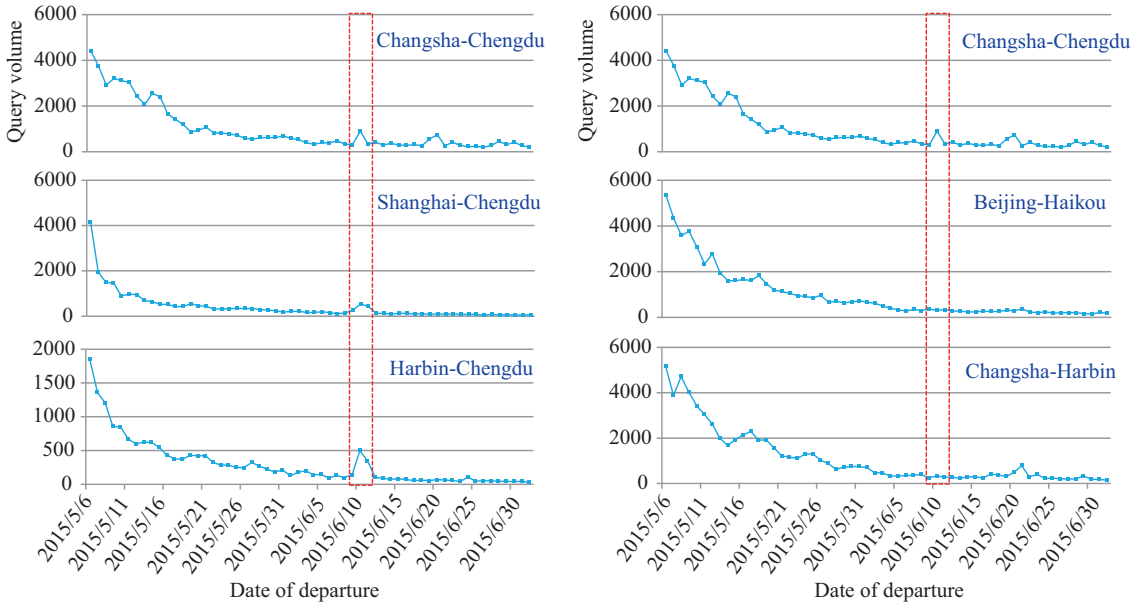


图 1　(网络版彩图) 用户在线机票查询量时间序列曲线示例

Figure 1　(Color online) Examples of time series curves of users' online query volumes

国内的机票订票渠道主要分为传统代理订票 (即 MCSS, message center switch system) 和互联网订票 (即 IBE, Internet booking engine) 两种. 随着互联网和移动智能终端技术的发展, 用户通过 IBE 渠道进行机票查询和预订所占的比例越来越高, 这给采集和分析用户查询数据带来了极大的方便. 大量用户在一段时间内对某条航线在未来不同起飞日期的查询量, 构成了一条时间序列曲线. 通常情况下, 对越临近的起飞日期查询量越大, 对越遥远的起飞日期查询量越小, 因此在理想状况下查询量时间序列曲线应该呈现出一个平缓下降的趋势. 但在实际情况中, 经常会有一些航线在某些起飞日期出现异常的民航需求, 这些异常也会反映到查询量曲线上, 因而查询量曲线并非总是平滑的. 例如, 图 1 展示了多条航线在 2015 年 4 月 29 日至 5 月 5 日一周之内对未来 60 个起飞日期的查询量时间序列曲线. 从图中可以看出, 曲线在多处出现异常波动. 因此, 可以通过对这些查询量曲线的异常检测, 来提前感知民航需求的异常.

传统的时间序列异常检测方法通常只是单独对一条时间序列曲线进行分析 [2~9]. 当存在某些偶然因素 (如恶意查询等) 带来单条航线的查询量异常时, 可能导致该航线的需求异常检测产生误差. 事实上, 在民航领域, 一个目的地的需求异常通常会导致多条相关航线的查询异常. 在图 1 左图中, 长沙 – 成都、上海 – 成都以及哈尔滨 – 成都这 3 条航线在 2015 年 6 月 10 日都出现了疑似查询量异常. 而与此同时, 观察其他与成都无关的航线在这一天的查询量情况, 如图 1 右图所示, 发现其他航线在这一天没有出现类似的查询异常. 这就表明, 存在某个 (或某些) 只与成都有关的原因, 导致多条以成都为目的地的航线在 2015 年 6 月 10 日这一天都出现了需求异常.

基于上述观察分析, 本文提出了一种新颖的基于用户在线查询行为分析的民航异常需求发现方法. 该方法首先针对每一条航线的查询量时间序列曲线, 从多维度计算每个起飞日期窗口的异常值. 具体的维度包括曲线自身的不稳定性、与本航线历史同期纵向对比的异常程度以及与其他航线当前同期横向对比的异常程度. 然后, 在全国范围的航线网络 (即由全国有机场的城市及其之间的航线构成的网络) 上, 利用同出发地或者同目的地航线之间的需求相关性, 对单条航线的异常值进行基于网络的

迭代优化. 这一方法综合利用了单条曲线的多维特征和多条曲线之间的相关性, 因此预期将可以发现更加细微的异常, 从而提高民航异常需求发现的召回率. 同时, 经过网络迭代优化之后的异常值将会更加精确, 从而提高民航异常需求发现的准确率.

在来自某在线订票网站的用户查询真实数据集上进行实验, 实验结果表明, 所提出的多维度民航需求异常值计算方法与当前基于单条时间序列曲线的异常检测方法相比, 极大地提高了准确率. 而基于航线网络的迭代优化策略, 进一步提高了异常发现的民航异常发现的效果.

接下来的章节安排如下: 第 2 节介绍与本文相关的旅客行为分析和时间序列异常检测方面的工作. 第 3 节给出相关的定义. 第 4 节详细描述本文提出的方法. 第 5 节进行实验描述和结果分析. 第 6 节给出全文总结.

2 相关工作

随着近年来大数据相关技术的迅猛发展, 民航领域的大数据分析尤其是旅客行为分析也已迅速展开. 当前已有的一些研究主要基于旅客的出行记录 (即机票订单), 挖掘用户的出行模式, 从微观的视角对旅客进行画像. 例如, Barlés-Arizón 等 [2] 研究了不同家庭结构 (如两口之家、三口之家和三代同堂等) 的出行模式, 有助于为不同的家庭提供个性化的出行服务. Ma 等 [3] 对旅客群体在机场各个场所 (如候机厅、咖啡厅、咨询台、电话厅和值机台等) 的空间分布和流动模式建立了基于 agent 的仿真模型, 以提高机场吞吐能力和服务质量. Budesca 等 [4] 通过研究旅客的行为习惯和群体关系, 提出一种新的登机策略, 可以有效节省登机时间. Lin 等 [5] 利用旅客的历史共同出行记录构建旅客社交网络, 并提出一种迭代分类算法来预测大规模旅行团体的出行目的. Wan 等 [6] 提出了一种基于旅客社交网络的家庭结构发现方法, 可以准确地从海量旅客中识别家庭结构. 这些工作将有利于民航企业为旅客家庭或群体提供个性化的出行服务或针对性的产品推荐. Lin 等 [7,8] 还研究了民航旅客的出行偏好和价值预测问题, 提出了基于旅客社交网络的解决方法, 为民航领域的旅客画像问题提供新的思路.

上述研究主要从微观的视角对民航旅客出行行为进行分析, 其主要目的在于为旅客提供个性化、差异化以及精准化的推荐或服务. 本文的目的在于通过对大量用户的整体行为进行分析, 从宏观上把握各航线市场需求的异常变化情况. 大量用户的在线查询行为形成了各航线机票查询量时间序列曲线, 本文将在此基础上采取时间序列异常检测方法来发现民航需求中的异常点.

常见的时间序列异常检测方法主要包括基于相似性的方法、基于统计的方法、基于聚类的方法, 以及基于密度的方法等 [9]. Protopapas 等 [10] 将时间序列划分为多个子序列, 然后通过计算子序列之间的两两相似性来衡量某个子序列的异常值. Das 等 [11] 提出了一种时空多维时间序列异常检测方法, 在不同维度采用不同的相似性度量方法来进行异常检测. Chandola 等 [12] 首先对子序列利用 k-medoids 方法进行聚类, 然后将子序列的异常值转化为与其最近的类中心之间的距离. Yao 等 [13] 提出基于邻近图和 PageRank 的异常检测算法 ADPP, 该算法首先构造子序列的 ε- 邻域图, 之后利用 PageRank 算法发现图中的异常子序列. Izakian 等 [14] 将时间序列中的异常定义为振幅异常和形态异常, 然后采用模糊聚类的方法对序列中的每一个子序列窗口计算一个异常值.

上述时间序列异常检测方法主要针对单个时间序列进行分析, 不同的方法都有其特定的适用场景. 由于各航线的查询量曲线之间存在一定的相关性, 实验表明, 直接将现有的方法应用于民航机票查询量时间序列的异常检测难以取得理想的效果. 事实上, 已经有学者尝试利用多个时间序列之间的相关性分析来进行异常检测. 例如, Qiao 等 [15] 对异构且相关的时间序列通过聚类方法进行趋势和相关性的比较, 从而发现时间序列中的异常点. Akoglu 等 [16] 综述了动态图结构上的时间序列异常检测方法,

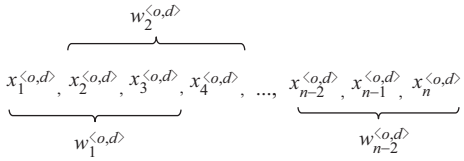


图 2　由查询量时间序列生成滑动窗口子序列

Figure 2　Sliding window sub sequences generated by query volume time series

为从网络整体而非单个时间序列的视角进行异常检测提供了借鉴意义. 本文将利用相同出发地或目的地航线之间的自然相关性, 采用基于航线网络的迭代优化策略来提高异常检测的准确性.

3　问题定义

本节首先对涉及到的一些相关概念给出形式化定义, 然后对所研究的问题进行形式化描述.

定义1 (航线网络)　用 $G=(V,E)$ 表示全国范围的航线网络, 其中 V 为有机场的城市集合, E 为航线集合, 一条边 $\langle o,d\rangle$ 代表一条航线, 其中 $o\in V$ 表示出发地, $d\in V$ 表示目的地.

定义2 (查询量时间序列)　用 $X^{\langle o,d\rangle}=x_1^{\langle o,d\rangle},x_2^{\langle o,d\rangle},\ldots,x_n^{\langle o,d\rangle}$ 表示过去一段时间内, 某个订票网站所有用户对航线 $\langle o,d\rangle$ 未来 n 天每天的机票查询量形成的时间序列, 其中 $x_i^{\langle o,d\rangle}$ 表示时间序列上的第 i 个点, 即用户对第 i 天的机票查询量.

要检测查询量时间序列上的异常, 不能只分析序列上的单个点, 而是要对前后相邻的几个点构成的子序列进行分析. 用一个固定长度的时间窗口在查询量时间序列 $X^{\langle o,d\rangle}$ 上滑动, 形成子序列.

定义3 (滑动窗口子序列)　用 $w_j^{\langle o,d\rangle}$ 表示查询量时间序列 $X^{\langle o,d\rangle}$ 上的一个由长度为 t 的滑动窗口所形成的子序列, 显然有 $1\leqslant j\leqslant(n-t+1)$. 这样的子序列即为文本的研究对象.

图 2 给出了由查询量时间序列 $X^{\langle o,d\rangle}$ 生成滑动窗口子序列 $w_j^{\langle o,d\rangle}$ 的示意图, 其中时间窗口长度 t 设为 3.

为了在后续异常检测中对任意航线 $\langle o,d\rangle$ 的当前查询量与其历史同期查询量进行纵向对比, 还将针对每条航线通过一定的方法计算其查询量标准曲线.

定义4 (查询量标准曲线)　用 $S^{\langle o,d\rangle}=s_1^{\langle o,d\rangle},s_2^{\langle o,d\rangle},\ldots,s_n^{\langle o,d\rangle}$ 表示航线 $\langle o,d\rangle$ 在通常情况下的查询量时间序列标准曲线. 该标准曲线由该航线的近期历史查询量数据经过去噪和加权求平均计算而得. 标准曲线上的滑动窗口子序列用 $\mathrm{ws}_j^{\langle o,d\rangle}$ 表示.

问题定义: 民航需求异常发现. 用 $\phi_j^{\langle o,d\rangle}$ 表示航线 $\langle o,d\rangle$ 的查询量序列曲线 $X^{\langle o,d\rangle}$ 上第 j 个滑动窗口子序列 $w_j^{\langle o,d\rangle}$ 的异常值, 则民航需求异常发现的目标就是检测某航线查询量时间序列上每个滑动窗口子序列的异常情况, 即计算异常值序列 $\Phi^{\langle o,d\rangle}=\phi_1^{\langle o,d\rangle},\phi_2^{\langle o,d\rangle},\ldots,\phi_{n-t+1}^{\langle o,d\rangle}$.

4　民航异常需求发现算法

本节将对提出的基于用户查询量时间序列分析和航线网络迭代优化的民航异常需求发现算法进行详细的描述. 算法首先从多个维度对单条航线查询量时间序列 $X^{\langle o,d\rangle}$ 上各个滑动窗口子序列 $w_j^{\langle o,d\rangle}$ 计算其异常值 $\phi_j^{\langle o,d\rangle}$, 然后基于全国范围的航线网络对单条航线上的异常值进行迭代优化, 得到最终优化后的异常值序列 $\widetilde{\Phi}^{\langle o,d\rangle}=\{\widetilde{\phi}_j^{\langle o,d\rangle}\}$.

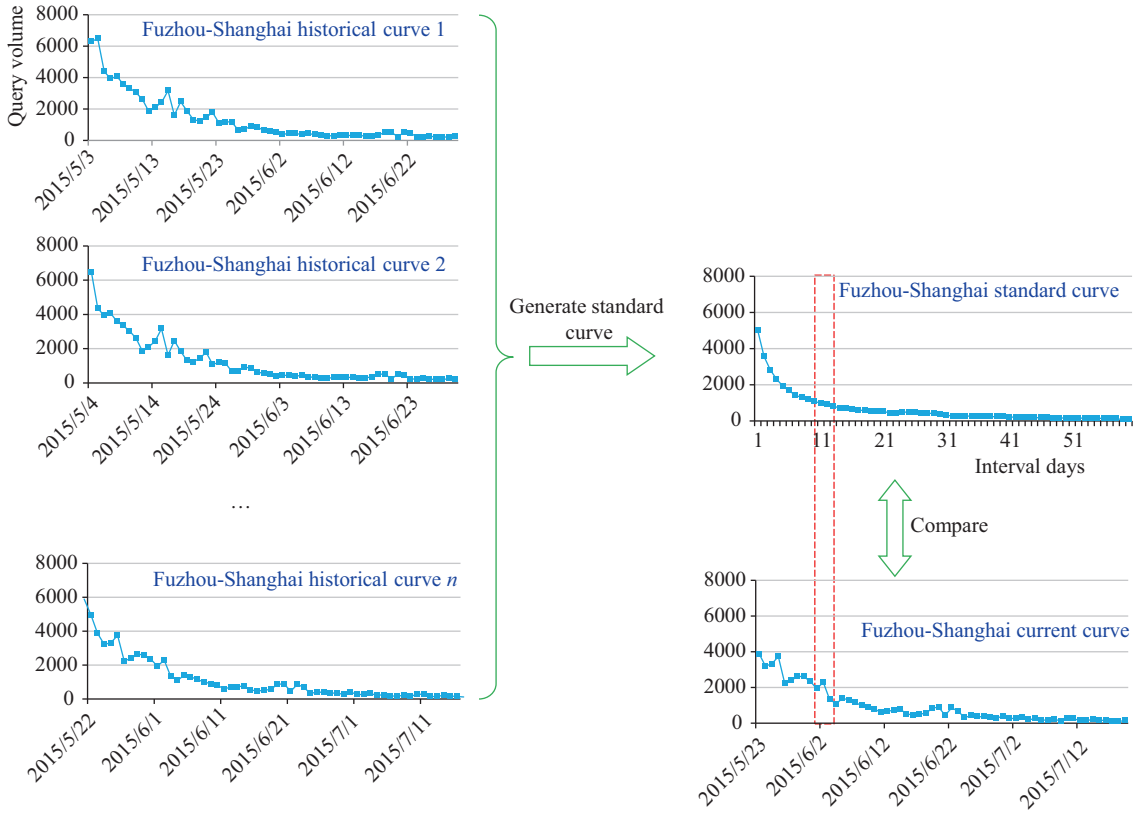


图 3 (网络版彩图) 当前窗口子序列与本航线历史曲线的纵向比较

Figure 3 (Color online) Vertical comparison between the sub sequence of the current window and its historical curves

4.1 单条航线上的多维度民航需求异常检测

针对某条航线的一个滑动窗口子序列 $w_j^{\langle o,d\rangle}$, 分别从 3 个维度来衡量其异常程度, 分别为: 该窗口与本航线历史查询量曲线同期窗口子序列之间的差异、该窗口与其他航线当前查询量曲线同期窗口子序列之间的差异以及该航线查询量时间序列曲线的自身复杂性. 最后通过将这 3 个维度计算所得的异常值相结合, 得到子序列 $w_j^{\langle o,d\rangle}$ 的异常值 $\phi_j^{\langle o,d\rangle}$.

(1) 与本航线的历史同期差异. 通常情况下, 大量用户在历史同期对于同一条航线的查询量应该呈现相似的趋势. 因此, 通过将当前查询曲线与同一条航线的历史曲线进行对比, 可以从纵向的角度反映当前曲线的某种异常程度. 如图 3 所示, 首先将某条航线 (如福州 – 上海, 当前时间为 2016/5/22, 查询未来 60 天的查询量时间序列) 的多条近期历史查询量时间序列曲线经过去噪和加权平均 (由于不同时段航班查询量差异较大, 距当前时间较远的查询量时间序列与当前查询量序列相差可能较大, 所以越久远的曲线权值也越低), 得到一条该航线的查询量标准曲线 $S^{\langle o,d\rangle}$. 然后, 将当前查询量曲线 $X^{\langle o,d\rangle}$ 上第 j 个滑动窗口子序列 $w_j^{\langle o,d\rangle}$ 与标准曲线 $S^{\langle o,d\rangle}$ 上同位置的子序列 $\mathrm{ws}_j^{\langle o,d\rangle}$ 进行比较.

用当前子序列 $w_j^{\langle o,d\rangle}$ 与标准曲线子序列 $\mathrm{ws}_j^{\langle o,d\rangle}$ 的欧式距离来描述这种差异, 记为 $\phi_{1j}^{\langle o,d\rangle}$. 同时, 考虑到不同航线的查询总量存在较大差别, 为了后续对比不同航线的子序列的异常程度, 用当前查询量曲线的总量对差异值进行规范化, 得到

$$\phi_{\mathrm{I}j}^{\langle o,d\rangle}=\frac{\mathrm{Eud}\left(w_j^{\langle o,d\rangle},\mathrm{ws}_j^{\langle o,d\rangle}\right)}{\sum_{i=1}^{n}x_i^{\langle o,d\rangle}},\tag{1}$$

其中 Eud(·) 为求序列欧式距离的函数.

(2) 与其他航线的当前同期差异. 通常情况下, 大量用户在当前同期对于大多数航线的查询量应该呈现相似的趋势. 因此, 通过将当前查询曲线与其他不同航线的当前查询曲线进行对比, 可以从横向的角度反映当前查询曲线的某种异常程度. 用当前航线 $\langle o,d\rangle$ 的查询量曲线 $X^{\langle o,d\rangle}$ 上第 j 个滑动窗口子序列 $w_j^{\langle o,d\rangle}$ 与其他任意航线 $\langle o',d'\rangle$ 的查询曲线 $X^{\langle o',d'\rangle}$ 上同位置的子序列 $w_j^{\langle o',d'\rangle}$ 进行比较. 用子序列的余弦距离来描述这种差异, 得到

$$\phi_{\mathrm{II}j}^{\langle o,d\rangle}=\frac{\sum_{\langle o',d'\rangle\in E\&\langle o',d'\rangle\neq\langle o,d\rangle}1-\cos\left(w_j^{\langle o,d\rangle},w_j^{\langle o',d'\rangle}\right)}{|E|-1},\tag{2}$$

其中 cos(·) 为求序列余弦距离的函数.

(3) 本航线自身的查询量曲线的复杂性. 通常情况下, 如果一条查询量曲线本身的波动性较大, 那么其中的一个普通的波动就很平常. 反之, 如果一条曲线自身比较平滑, 那么偶尔出现一个小的波动都很可疑. 因此, 在计算滑动窗口子序列的异常程度时, 其所在曲线本身的复杂性也要考虑在内. 利用经典的时间序列分解方法 STL [17] 来计算曲线的复杂性. 首先使用 STL 将一条航线查询量时间序列 $X^{\langle o,d\rangle}$ 分解得到该序列的趋势项 $T^{\langle o,d\rangle}$ 和周期项 $C^{\langle o,d\rangle}$. 其中趋势项反映了该序列的整体发展趋势, 周期项反映了该序列的周期变动情况. 原序列减去趋势项和周期项, 就可得到该序列的随机波动项 $I^{\langle o,d\rangle}$, 它反映了原序列的不规则波动情况. 用随机波动项 $I^{\langle o,d\rangle}$ 的标准差来度量查询量时间序列的自身复杂性, 同样用该序列的查询总量进行规范化, 得到

$$\phi_{\mathrm{III}j}^{\langle o,d\rangle}=\frac{\mathrm{StdDev}\left(I^{\langle o,d\rangle}\right)}{\sum_{i=1}^{n}x_i^{\langle o,d\rangle}}=\frac{\mathrm{StdDev}\left(X^{\langle o,d\rangle}-T^{\langle o,d\rangle}-C^{\langle o,d\rangle}\right)}{\sum_{i=1}^{n}x_i^{\langle o,d\rangle}},\tag{3}$$

其中 StdDev(·) 为求序列标准差的函数.

(4) 综合异常值计算. 接下来, 将子序列的上述 3 个维度的异常值组合起来, 得到一个综合的异常值. 显然, $\phi_{\mathrm{I}j}^{\langle o,d\rangle}$ 和 $\phi_{\mathrm{II}j}^{\langle o,d\rangle}$ 与子序列的整体异常程度 $\phi_j^{\langle o,d\rangle}$ 正相关, 而 $\phi_{\mathrm{III}j}^{\langle o,d\rangle}$ 与 $\phi_j^{\langle o,d\rangle}$ 是负相关的, 因此, 得到航线 $\langle o,d\rangle$ 的第 j 个滑动窗口子序列的综合异常值如下:

$$\phi_j^{\langle o,d\rangle}=\frac{\phi_{\mathrm{I}j}^{\langle o,d\rangle}\phi_{\mathrm{II}j}^{\langle o,d\rangle}}{\phi_{\mathrm{III}j}^{\langle o,d\rangle}}.\tag{4}$$

式 (4) 计算得到的 $\phi_j^{\langle o,d\rangle}$ 是单条航线的需求异常值, 其完全依赖于单条航线的查询量. 事实上, 某一条航线的查询量常常会受到一些偶然因素的影响. 例如, 一些恶意的查询行为或者自动爬虫导致某些航线的查询量在短时间内急剧增长. 一些机票代理仅通过少量的查询就为多个团体旅客订购机票, 从而导致某些航线的查询量看起来偏低. 这些偶然因素的存在, 会降低单条航线需求异常值的准确性. 因此, 基于全国范围内的航线网络, 从网络整体的视角对各单条航线的需求异常值进行迭代优化.

4.2 基于航线网络的迭代优化

一个城市的需求异常通常会导致其作为出发地或者目的地的多条相关航线的查询异常. 例如, 在 2014 年 11 月 APEC 中国峰会期间, 北京旅客有较大的外出需求, 造成了以北京为出发地的多条航线

在该时期的查询量均远高于其他航线. 2015 年 8 月的广州博览会则带来全国各地到广州的出行需求增加, 从而导致以广州为目的地的多条航线在该时期的查询量同时升高.

由以上分析可见, 航线的需求异常大多都是由其出发地或者目的地的特定事件触发的, 由此假设: 有相同出发地或目的地的航线具有相似的需求异常倾向. 全国范围内拥有机场的城市之间通过航线关系组成了一个复杂的航线网络 $G=(V,E)$. 基于这个整体航线网络, 提出一种需求异常值的网络迭代优化算法, 利用上面的假设来不断地迭代调整需求异常值, 直到最后在整个航线网络上达到一种平衡.

对于某条航线 $\langle o,d\rangle$ 的一个滑动窗口子序列 $w_j^{\langle o,d\rangle}$, 以第 4.1 小节中计算得到的需求异常值作为本节迭代算法中该子序列的初始状态. 用 k 表示迭代次数, 则初始状态 $(k=0)$ 为

$$\phi_{(0)}^{\langle o,d\rangle}=\phi_j^{\langle o,d\rangle}. \tag{5}$$

注意, 为了书写简洁, 此处省略了下标 j, 本节中出现的所有子序列均为某航线查询量曲线上的第 j 个子序列.

在接下来的每一次迭代过程中, 首先针对每一个城市 $c\in V$, 利用上一次的迭代结果分别计算其作为出发地和目的地的总体异常情况. c 为出发地的总体异常情况用以 c 为出发地的全部航线的平均异常值来描述:

$$\phi_{(k)}^{\langle c,\cdot\rangle}=\frac{\sum_{d\in V,d\neq c}\phi_{(k-1)}^{\langle \mathrm{c,d}\rangle}}{|E_{c\cdot}|}, \tag{6}$$

其中, $\langle c,\cdot\rangle$ 表示 c 作为出发地, $E_{c\cdot}$ 表示以 c 为出发地的全部航线集合.

类似地, c 作为目的地的总体异常情况用以 c 为目的地的全部航线的平均异常值来描述:

$$\phi_{(k)}^{\langle \cdot,c\rangle}=\frac{\sum_{o\in V,o\neq c}\phi_{(k-1)}^{\langle o,c\rangle}}{|E_{\cdot c}|}, \tag{7}$$

其中, $\langle\cdot,c\rangle$ 表示 c 作为目的地, $E_{\cdot c}$ 表示以 c 为目的地的全部航线集合.

对于航线 $\langle o,d\rangle$, 根据相同出发地或目的地的航线具有相似需求异常倾向的假设, 用出发地 o 的异常值 $\phi_{(k)}^{\langle o,\cdot\rangle}$ 和目的地 d 的异常值 $\phi_{(k)}^{\langle \cdot,d\rangle}$ 来共同修正 $\langle o,d\rangle$ 的异常值 $\phi_{(k)}^{\langle o,d\rangle}$, 修正值为

$$\phi_{(k)}^{\prime\langle o,d\rangle}=\alpha\phi_{(k)}^{\langle o,\cdot\rangle}+\beta\phi_{(k)}^{\langle \cdot,d\rangle}, \tag{8}$$

其中, α 和 β 分别表示目的地和出发地异常值的权重, 且 $\alpha+\beta=1$.

接下来, 用修正值 $\phi_{(k)}^{\prime\langle o,d\rangle}$ 来更新航线 $\langle o,d\rangle$ 的最新异常值:

$$\phi_{(k)}^{\langle o,d\rangle}=\phi_{(k-1)}^{\langle o,d\rangle}+\eta\mathrm{sign}\left(\phi_{(k)}^{\prime\langle o,d\rangle}-\phi_{(k-1)}^{\langle o,d\rangle}\right)\ln\left(\left|\phi_{(k)}^{\prime\langle o,d\rangle}-\phi_{(k-1)}^{\langle o,d\rangle}\right|+1\right), \tag{9}$$

其中, η 为迭代更新速率, $\mathrm{sign}(\cdot)$ 为符号函数, 自然对数函数用来降低相同出发地或目的地的航线的异常值方差过大带来的影响.

利用式 (9) 更新完整个航线网络中所有航线的异常值后, 本次迭代完成, 然后进入下一次迭代. 如此反复, 直到整个网络趋于收敛, 即每条航线的异常值不再发生较大的改变, 记最终的异常值为 $\widetilde{\phi}_j^{\langle o,d\rangle}$.

整个网络迭代优化算法框架描述如算法 1 所示.

算法 1 航线异常值的网络迭代优化算法

Input: 所有航线异常值序列初始值集合 $\Phi = \{\Phi^{\langle o,d\rangle}\}$, 航线网络 $G = (V, E)$.
Output: 所有航线异常值序列优化值集合 $\widetilde{\Phi} = \{\widetilde{\Phi}^{\langle o,d\rangle}\}$.

```
k ⇐ 1;
repeat
    for all c ∈ V do
        根据式 (6) 计算出发地异常值序列 Φ_(k)^⟨c,·⟩;
        根据式 (7) 计算目的地异常值序列 Φ_(k)^⟨·,c⟩;
    end for
    for all ⟨o, d⟩ ∈ E do
        根据式 (8) 计算航线异常修正值序列 Φ'_(k)^⟨o,d⟩;
        根据式 (9) 更新航线异常值序列 Φ_(k)^⟨o,d⟩;
    end for
    k++;
until k 达到最大迭代次数 ‖ 所有航线的异常值不再发生较大改变.
```

5 实验及结果分析

在某在线订票网站提供的真实查询数据集上进行实验, 将本文提出的异常发现算法与一种常用的基于曲线相似性度量的时间序列异常模式检测方法进行了对比, 证明了所提出的方法具有良好的民航异常需求发现能力.

5.1 数据集

本文使用的数据来自某在线订票网站提供的真实机票查询数据集, 数据中包含的信息包括航线 (即出发地和目的地)、起飞日期、查询日期以及查询次数 (即查询量). 数据集中一共涉及到了 159 个有机场的城市, 包含 23416 条航线, 构成全国范围的航线网络.

由于单独某一天对某个航线的查询量存在一定的偶然性, 用连续一周对某航线的每个起飞日期的累积查询量作为查询量时间序列上的一个数据点. 共选取了 4 个查询周期 (分别为 2015/5/5 ∼ 2015/5/11、2015/5/7 ∼ 2015/5/13、2015/5/9 ∼ 2015/5/15 和 2015/5/11 ∼ 2015/5/17) 对全部航线的查询数据作为本文的实验数据集. 一个查询量时间序列包含 60 个起飞日期, 即一共有 60 个数据点. 根据民航业务经验, 设置滑动时间窗口 $t = 3$, 则每个时间序列有 58 个滑动窗口子序列.

为了对本文提出的异常需求发现算法的效果进行验证, 标注了一个测试数据集. 选取北京、昆明和西宁 3 个城市之间的 6 条航线, 对其 4 个查询周期的 24 条查询量时间序列上的 $24 \times 58 = 1392$ 个滑动窗口子序列进行人工标注. 邀请在线订票网站的民航收益分析专家对测试集进行了标注, 在 1392 个样本点中分别标注了 378 个正例 (即有异常需求) 和 1014 个负例 (即无异常需求).

实验数据集的统计信息如表 1 所示.

5.2 实验设置

采用一种常用的基于曲线相似度的时间序列异常检测方法 [18] 作为本文的基准方法, 其具体做法是将滑动窗口子序列与同序列中的全部其他子序列进行相似度比较. 用余弦距离来描述子序列的相似

表 1 实验数据集统计信息

Table 1 Experimental data set statistics

Item	Value
Number of cities in the entire data set	159
Number of air routes in the entire data set	23416
Query intervals in the entire data set	2015/5/5 ∼ 2015/5/11, 2015/5/7 ∼ 2015/5/13, 2015/5/9 ∼ 2015/5/15, 2015/5/11 ∼ 2015/5/17
Number of air routes in the testing set	Beijing-Kunming, Kunming-Beijing, Beijing-Xining, Xining-Beijing, Kunming-Xining, Xining-Kunming
Number of positive examples in the testing set	378
Number of negative examples in the testing set	1014

度, 对于航线 $\langle o,d\rangle$ 上的第 j 个子序列 $w_j^{\langle o,d\rangle}$, 其异常值计算如下:

$$\phi_j^{\langle o,d\rangle}=\frac{1}{n}\sum_{i=1}^{n-t+1}\left(1-\cos\left(w_j^{\langle o,d\rangle},w_i^{\langle o,d\rangle}\right)\right), \tag{10}$$

其中, $n-t+1$ 为航线 $\langle o,d\rangle$ 所对应的查询量时间序列上滑动窗口子序列的个数, $w_i^{\langle o,d\rangle}$ 为任意一个子序列, $\cos(\cdot)$ 为余弦相似度函数. 用这一方法计算测试集中的所有样本点的异常值, 然后与本文提出的方法进行比较.

在本文提出的异常需求发现算法中, 涉及多个参数需要提前设置. 首先是式 (8) 中的出发地异常值权重 α 和目的地异常值权重 β, 通过实验发现当设置 $\alpha=0.4$ 和 $\beta=0.6$ 时实验结果最好, 这说明目的地的影响略大于出发地. 然后是式 (9) 中的迭代更新速率 η, 通过实验发现当设置 $\eta=10$ 时能够最快地完成迭代步骤. 采用前 K 个预测异常样本的准确率 (即 Precision@K) 和召回率 (即 Recall@K) 作为实验结果的评价标准. 对每次实验, 将测试集中的所有样本按照计算所得异常值从大到小排序, 排序中前 K 个样本中正例 (真实异常样本) 所占的百分比即为 Precision@K, 而排序中前 K 个样本中正例占全部 378 个正例的百分比即为 Recall@K. 分别取 $K=50,100,150,200,250,300$ 和 378 来观察准确率和召回率的变化情况.

5.3 实验结果及分析

图 4 和 5 分别给出了本文提出的多维度民航需求异常检测方法 (未经过航线网络迭代优化) 与基于曲线相似度的时间序列异常检测方法在准确率和召回率上的对比. 从图中可以看出, 本文的多维度检测方法明显优于传统的基于曲线相似度的方法. 随着 K 的增大, 基于曲线相似度的方法准确率迅速下降、召回率缓慢上升, 而本文的多维度检测方法准确率下降速度相对缓慢、召回率上升相对迅速. 当 K 值较大时, 本文的方法将准确率和召回率均提高了约 20%. 这一实验结果表明, 本文提出的结合本航线的历史查询情况、其他航线的同期查询情况以及当前查询曲线本身的复杂性等多个维度来检测民航异常需求的方法是行之有效的.

图 6 和 7 分别给出了本文的多维度异常检测方法经过进一步基于航线网络迭代优化之后准确率和召回率的变化情况. 从图中可以看出, 异常需求发现的准确率和召回率在经过大约 40 次迭代优化之后逐渐趋于稳定, 比优化之前都有了很大的提升, 尤其是当 K 取值较大时, 准确率和召回率都提升

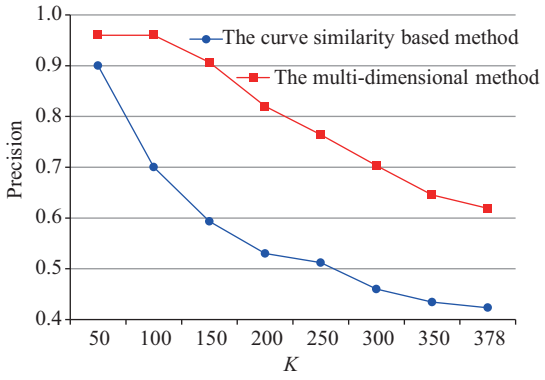


图 4 (网络版彩图) 本文的多维度异常检查方法与基于曲线相似度的异常检测方法准确率对比

Figure 4 (Color online) Precision comparison between the proposed multi-dimensional method and the curve similarity based method

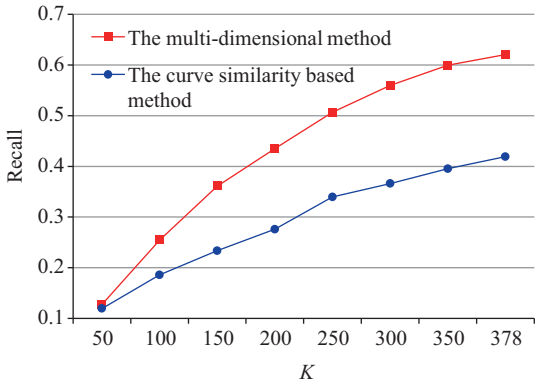


图 5 (网络版彩图) 本文的多维度异常检查方法与基于曲线相似度的异常检测方法召回率对比

Figure 5 (Color online) Recall comparison between the proposed multi-dimensional method and the curve similarity based method

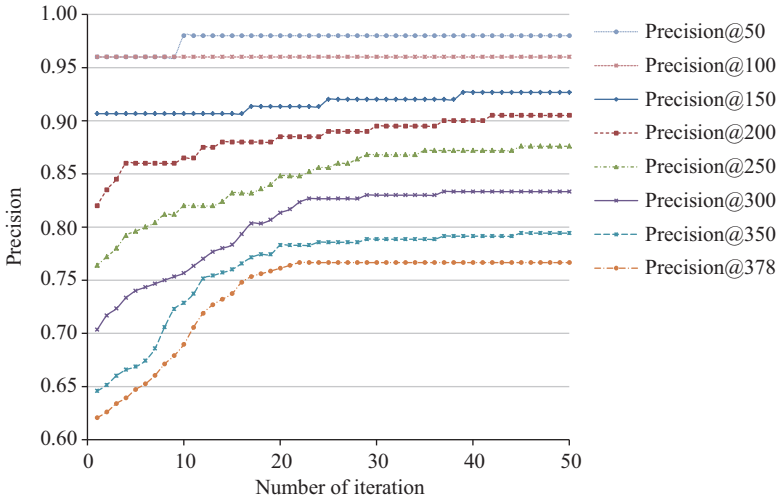


图 6 (网络版彩图) 基于航线网络的民航异常需求发现准确率迭代优化效果

Figure 6 (Color online) Improvement of precision by the route network based iterative method

了 10% 以上. 这一实验结果表明, 基于航线网络的异常检测迭代优化方法, 可以有效地检测出那些常规手段难以发现的潜在异常点, 从而大大提高算法的异常发现能力.

6 结论

本文研究了基于用户在线查询行为分析的民航异常需求发现问题, 提出了一种多维度的时间序列异常检测方法, 结合本航线的历史查询情况、其他航线的同期查询情况以及当前查询曲线本身的复杂性等维度来进行民航异常需求检测, 并利用具有相同出发地或目的地的航线具有相似的需求异常倾向这一假设, 在全国范围的航线网络对不同航线的异常值进行迭代优化. 在某在线订票网站提供的真实

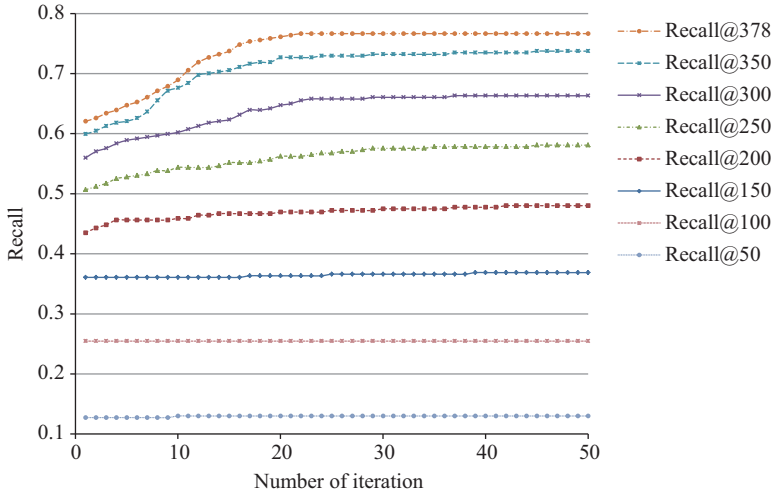


图 7 (网络版彩图) 基于航线网络的民航异常需求发现召回率迭代优化效果

Figure 7 (Color online) Improvement of recall by the route network based iterative method

查询数据集上的实验结果证明, 本文提出民航异常需求检测算法能够有效地从用户查询行为记录中发现民航异常需求, 从而有助于民航相关企业更准确地掌握民航市场需求的变化, 及时采取相应的市场对策, 提高服务质量和收益水平, 改善用户出行体验.

参考文献

1 Civil Aviation Administration of China. The civil aviation industry development statistical bulletin in 2016. 2017. http://www.caac.gov.cn [中国民用航空局. 2016 年民航行业发展统计公报. 2017. http://www.caac.gov.cn]

2 Barlés-Arizón M J, Fraj-Andrés E, Martínez-Salinas E. Family vacation decision making: the role of woman. J Travel Tourism Marketing, 2013, 30: 873–890

3 Ma W, Kleinschmidt T, Fookes C, et al. Check-in processing: simulation of passengers with advanced traits. In: Proceedings of the Winter Simulation Conference, Phoenix, 2011. 1783–1794

4 Budesca G C, Juan A A, Casas P F. Optimization of aircraft boarding processes considering passengers' grouping characteristics. In: Proceedings of the Winter Simulation Conference, Savannah, 2014. 1977–1988

5 Lin Y F, Wan H Y, Jiang R, et al. Inferring the travel purposes of passenger groups for better understanding of passengers. IEEE Trans Intell Transport Syst, 2015, 16: 235–243

6 Wan H Y, Wang Z W, Lin Y F, et al. Discovering family groups in passenger social networks. J Comput Sci Tech, 2015, 30: 1141–1153

7 Lin Y F, Wang K K, Zhou C, et al. Modeling the preference of air passengers based on social networks. J Beijing Jiaotong Univ, 2014, 6: 33–39 [林友芳, 王琨琨, 周超, 等. 基于社交网络的民航旅客偏好建模. 北京交通大学学报, 2014, 6: 33–39]

8 Lin Y F, Zhang A S, Wan H Y, et al. Predicting the growth of new passengers in civil aviation based on social networks. J Beijing Jiaotong Univ, 2014, 6: 40–46 [林友芳, 张奥爽, 万怀宇, 等. 一种基于社交网络的民航新旅客成长性预测方法. 北京交通大学学报, 2014, 6: 40–46]

9 Xiao H. Similarity search and outlier detection in time series. Dissertation for Ph.D. Degree. Shanghai: Fudan University, 2005 [肖辉. 时间序列的相似性查询与异常检测. 博士学位论文. 上海: 复旦大学, 2005]

10 Protopapas P, Giammarco J M, Faccioli L, et al. Finding outlier light curves in catalogues of periodic variable stars. Mon Notices Roy Astron Soc, 2006, 369: 677–696

11 Das M, Parthasarathy S. Anomaly detection and spatio-temporal analysis of global climate system. In: Proceedings of the 3rd International Workshop on Knowledge Discovery From Sensor Data, Paris, 2009. 142–150

12 Chandola V, Banerjee A, Kumar V. Anomaly detection for discrete sequences: a survey. IEEE Trans Knowl Data

Eng, 2012, 24: 832–839
13 Yao Z, Mark P, Rabbat M. Anomaly detection using proximity graph and PageRank algorithm. IEEE Trans Inf Foren Secur, 2012, 7: 1288–1300
14 Izakian H, Pedrycz W, Jamal I. Clustering spatio-temporal data: an augmented fuzzy c-means. IEEE Trans Fuzzy Syst, 2013, 21: 855–868
15 Qiao Z, He J, Cao J, et al. Multiple time series anomaly detection based on compression and correlation analysis: a medical surveillance case study. In: Proceedings of the 14th Asia-Pacific Web Conference, Kunming, 2012. 294–305
16 Akoglu L, Tong H, Koutra D. Graph based anomaly detection and description: a survey. Data Min Knowl Discov, 2015, 29: 626–688
17 Cleveland R B, Cleveland W S, McRae J E, et al. STL: a seasonal-trend decomposition procedure based on loess. J Off Stat, 1990, 6: 3–73
18 Faloutsos C, Ranganathan M, Manolopoulos Y. Fast subsequence matching in time-series databases. ACM Sigmod Rec, 2001, 23: 419–429

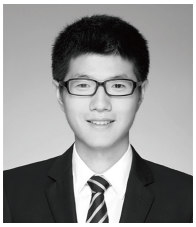
Discovering abnormal civil aviation requirements by analyzing users' online query behaviors

Qiangyong XU[1], Youfang LIN[1], Huaiyu WAN[1*], Lina WU[2] & Xuguang JIA[2]

1. *Beijing Key Laboratory of Traffic Data Analysis and Mining, School of Computer and Information Technology, Beijing Jiaotong University, Beijing* 100044*, China*;
2. *TravelSky Technology Limited, Beijing* 100010*, China*
* Corresponding author. E-mail: hywan@bjtu.edu.cn

Abstract Changes of users' query volume in online fight ticketing systems indicate the changes of requirements in civil aviation markets. By analyzing the big data of users' online query behaviors, we can timely and accurately discover abnormal civil aviation requirements. This ability is very conducive for airlines and agencies for taking immediate and effective marketing actions. In this paper, we propose a novel method to discover abnormal civil aviation requirements based on time-series curves of users' query volumes. In addition, we utilize the domestic airline route network to optimize the anomaly detection results from the perspective of a global network rather than that of a single airline. We conduct experiments on real-world users' query datasets collected from an online ticketing site. The experimental results demonstrate that the proposed method can effectively discover abnormal civil aviation requirements from users' online query logs.

Keywords civil aviation requirements, online flight ticket query, user behavior analysis, abnormal behavior detection, time-series curves



Qiangyong XU received a B.S. degree in computer science from Hainan University, Haikou, China, in 2014. He is currently working toward a master's degree at the School of Computer and Information Technology, Beijing Jiaotong University. His research interests include traffic data analysis and mining.

Youfang LIN received his Ph.D. in computer science and technology from Beijing Jiaotong University, Beijing, China, in 2003. He is a professor at the School of Computer and Information Technology, Beijing Jiaotong University. His research interests include data warehousing, data mining, business intelligence, and complex networks.

Huaiyu WAN received his Ph.D. in computer science and technology from Beijing Jiaotong University, Beijing, China, in 2012. He is an assistant professor at the School of Computer and Information Technology, Beijing Jiaotong University. His research interests include social network mining, user behavior analysis, and recommendation systems.

Lina WU received her Ph.D. in computer science and technology from Beijing Jiaotong University, Beijing, China, in 2014. Her research and development interests focus on artificial intelligence, data mining, and big data technology.