京津冀旅游大数据分析流程模型研究

Research on Big Data Analysis Process Model of Beijing-Tianjin-Hebei Tourism

李学龙[1] 郝文英[2]

Xuelong Li[1] Wenying Hao[2]

1.北京第二外国语学院
中国·北京 100000；
2.北京物资学院
中国·北京 100000

1.Beijing Second Institute of Foreign Languages, Beijing, 100000, China;
2.Beijing Wuzi University, Beijing, 100000, China

【摘 要】近年来，伴随着“大数据”产业的蓬勃发展，大数据处理技术已经渗透至各个行业。通常，大数据处理流程主要包括数据采集、数据清洗、分析挖掘、可视化、分析报告等重要环节，其中大数据的质量管理伴随整体数据处理流程，每一环节都会对数据质量产生影响，进而决定大数据的分析结果。

【Abstract】In recent years, with the vigorous development of big data industry, big data treatment technology has penetrated into various industries. In general, big data's processing process mainly includes data acquisition, data cleaning, analysis and mining, visualization, analysis and report and other important links, in which big data's quality management is accompanied by the overall data processing process, each link will have an impact on the data quality, and then determines big data's analysis results.

【关键词】旅游大数据；京津冀协同；大数据模型

【Keywords】tourism big data; Beijing-Tianjin-Hebei cooperation; big data model

【DOI】10.36012/sde.v1i3.460

1 引言

优秀的大数据分析前提是：海量基础数据、快速数据清洗、精准数据建模、清晰可视化展示以及通俗易懂的文字报告。本文将大数据分析技术运用于旅游分析中，构建了京津冀旅游大数据分析流程模型，进行了相关的分析研究。

2 京津冀旅游大数据采集与抽选

京津冀旅游大数据的数据源大体可分为两类：传统结构化数据和非结构化数据。其中，结构化数据主要来自购票、住宿、采购等旅行消费数据；非结构化数据主要来自公众号信息收集、旅行评论、晒圈等信息收集数据。

机构化数据通常存储于传统的结构化数据库中，如 Orcle、Sqlserver、DB2 等。非结构化数据通常存储于分布式的数据集中，如 Hadoop 等。在数据处理过程中，结构化数据和非结构化数据也是不太相同的，非结构的数据处理更加复杂。

大数据采集要求在所有相关数据采集完成之后，能够方便地进行增、删、改、查等工作，尤其是对查询的速度要求较高，其影响数据处理的速度和生成结果的质量。

3 京津冀旅游大数据清洗与集成

完成大数据的抽选与采集之后，需要对数据质量进行进一步的整理，数据治理尤为重要。

采集后数据为去除无用、重复、错误等相关无用的数据，需进行简单的数据清洗和预处理。常见的数据预处理方法主要有：数据集成、数据变换、数据归约等[1]。

①数据集成：是将多个数据来源中的数据收集并统一存储，建立数据仓库的过程；②数据变换：通过平滑聚集、数据概化、规范化等方式将数据转换成适用于数据挖掘的形式；③数据归约：数据挖掘时数据量需求非常大，通常在少量数据上进行挖掘分析都需要很长时间，数据归约技术可以用来得到数据集的归约表示，它相对较小，但仍然接近于保持原数据的完整性，且结果与归约前结果相同或几乎相同。

在大数据的清洗与集成过程中，导入的数据量巨大，经常会达到百兆，甚至千兆级，给数据处理带来一定的困难和挑战。通过对京津冀旅游大数据的清洗与集成，可完成数据的审核、筛选、排序等重要工作，为后续的数据分类、分组、分主题等工作进行充分的准备。

4 京津冀旅游大数据分析挖掘

数据挖掘一般是指从海量数据中应用算法搜索隐藏于其中信息的过程。数据挖掘是多个学科知识的融合，包含计算机科学与技术、梳理统计学、组织行为学等。应用现代计算机数

据处理技术，结合数理统计、情景分析、信息检索、机器学习、专家判断和模式识别等技术统一实现数据挖掘的目标。在京津冀旅游大数据中充分利用数据挖掘技术，可分析出京津冀大数据数据中所蕴含的无穷价值。本文认为通过如下几个方面进行数据挖掘，会得到很多优秀成果：

4.1 旅游数据基本项之间必相关

搜集到京津冀旅游大数据之后，可以得知京津冀各地区的城市环境、人口、交通、旅游目的地、消费等基础数据项，可通过相关性分析来计算出数据项之间的关联，可运用的技术有线性分析、回归分析等。分析结果可定量显示，可以更加直观地反映出数据项之间的关联性。

4.2 旅游数据对象聚类处理分析

根据旅游数据项的特征进行标签聚类，可以分成旅行、消费、评价、建议等人群聚类等，根据聚类再进一步分析类群的特色及现骨干型。根据分析结果可看出相应的聚类特色，应用于数据分析报告。

4.3 旅游新数据学习分类

集成新数据后，通过机器学习、智能研判等手段进行数据聚类，可降低基础数据分类时间，提高数据分析效率。

4.4 旅游数据分析最终助力决策

通过旅游大数据挖掘，可以得出京津冀区域旅游的很多分析结论，可助力相关部门进行决策，提高京津冀旅游大数据的应用价值，使得宝贵的数据资产发挥应有价值。

5 京津冀旅游大数据可视化与分析报告

西方很多发达国家在事务处理的过程中总是遵循“让数据说话”的原则，在行业中更加看重数据分析报告。如在 NBA（美国职业篮球联盟）中，大数据分析报告应用较为成熟，球员在选秀进入联盟中都会有一份“球探报告”，在进入 NBA 之后自身会形成“数据本”，记录各方面的数据。

一份优秀的数据分析报告尤为重要，大数据分析报告输出分析成果，可为辅助决策提供数据支撑。

数据分析报告以数据可视化为前提，数据可视化是指将数据挖掘分析结果以计算机图形或图像等以更加直观的方式呈现，并可进行交互互动。数据可视化技术有利于发现大量业务数据中隐含的规律性信息，为管理决策提供依据。数据可视化使得分析结果更加直观，便于理解使用，数据可视化是大数据可用性的关键因素。

一般的饼图、直方图、散点图、柱状图、折线图等图形是数据可视化的最基础可视化图形和常见应用，可在PPT 展示、报表统计、方案汇报中予以引用。

现今市面上有很多成熟的 BI 商业化智能工具，如 FineBI、Tableau 等，借助成熟的 BI 工具可加快大数据处理速度，能做到事半功倍的效果。

遵循如下的制作原则进行京津冀旅游大数据分析报告的设计，可形成一份完备的旅游数据分析报告：①梳理报告框架。架构清晰、主次分明、推理严密、结论精准、建议明确。②结论总结清晰。分析报告结论要清晰明确，简单明了。③精益求精。分析报告结论要精益求精，数据分析重在发现问题和分析问题，不追求多而追求精，力求简单易懂。④推理严密，多图少字，可读性强。数据分析报告结论一定要基于紧密严禁的数据分析推导过程，按发现问题—找出问题原因——解决问题的流程进行。数据分析报告可读性强，可充分考虑决策者关心的问题，适当地用图表代替大量堆砌的文字、数字有助于更直观地阐述相关问题。

6 结语

本文做了一定的基础性研究探索工作，初步构建了京津冀旅游大数据处理的基本模型，将大数据分析技术成功引入到京津冀旅游数据处理之中，构建了基础数据分析模型，为打造京津冀线上无障碍旅游区提供数据服务，以更好地实现三地的无障碍旅游提供数据依据，以充分发挥三地旅游资源的优势，形成三地旅游资源共建、共享、共管的模式做出了初步的数据建模模型。

京津冀旅游大数据分析技术的应用将打破三地行政区域界限，构筑互动机制，在资源互享、信息互通、客源互送、线路互推、政策互惠等方面开展充分的旅游合作，在共同开发精品旅游线路、提升旅游联盟的知名度和影响力、促进京津冀区域旅游融合发展等方面潜力无限[2]。

旅游大数据的分析对实现京津冀智慧旅游提供了新思路，为如何共享旅游资源和信息方面提供了数据决策依据。通过大数据智能分析挖掘，可为建立旅游信息共享、及时发布旅游提醒信息、疑似违规旅游企业信息等方面提供更加丰富的决策依据，构建强大的京津冀一体化大数据分析平台，打造出京津冀智慧旅游新思路，联合智能推介与营销，强化京津冀三地的“大旅游”形象，能为把京津冀旅游打造成为区域旅游一体化的样本和示范区提供良好数据支撑。

参考文献

[1]胡秀.数据挖掘中数据预处理的研究[J].赤峰学院学报(自然科学版),2015,31(5):5-6.

[2]赵慧娟.基于旅游服务供应链联盟的京津冀区域旅游合作策略研究[J].北京经济管理职业学院学报,2017,32(2):21-25.