交通数据可视化调查分析

摘要：以数据为驱动的智慧交通系统会自动分析利用系统内部产生的数据，以此来提高交通系统的性能，并且提供方便可靠的服务。交通数据指的是由移动的车辆行人产生的数据集。数据可视化是一种用来展现分布式和结构化数据集的有效方式，并且能够挖掘出数据中不易被察觉的信息。本文介绍了交通数据可视化的基本概念和流程，提供了相关数据处理技术的概述，并且总结了现有的描述交通数据的时间，空间，数值和分类特性的方法。

关键词：交通，交通数据可视化，可视化分析，数据驱动的智慧交通系统

1. 动机

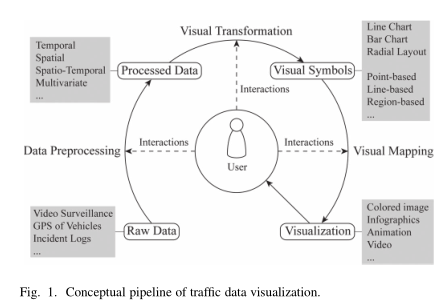
所谓交通是指机动车辆，非机动车辆和行人在道路上的通行或通行，或乘客（例如地铁交汇处）的移动【1】。交通无处不在，可以发生在城市地区，土地，海洋，空中，甚至地下。随着运输系统的快速发展，交通已经成为人类生活的重要组成部分并且对人们的生活质量产生了深远的影响。例如，大约平均有40%的人每天至少需要花费1小时来通勤【2】。

现代城市里，巨大的人口和成千上万的机动车引发了交通堵塞、交通事故、空气污染等一系列问题。为了解决这些问题，人们采取了一些措施，并且颇具成效，比如智能交通系统（ITS），公共交通系统，安全座椅安全带和安全气囊 。然而，越来越多的私家车，大大中和交通管制和控制的实现。在这些解决方案中，ITS格外引人注目，因为它使用先进的信息技术提高了交通系统的效率和功能【3】。尤其值得一提的是，随着收集的数据越来越多，数据本身所扮演的角色在ITS中显得愈发关键。这些数据包含了大量信息，同时，这些数据也可用于在ITS中生成新的功能和服务。数据驱动的ITS允许用户交互式地利用与交通系统相关的数据资源，以及通过更方便和可靠的服务来访问和使用数据，以提高交通系统的性能。

数据可视化的优势在于将人的能力融合到直观的可视化界面中，从而将机器智能与人机界面相结合。科学可视化，信息可视化和可视化分析是数据可视化的三大领域。科学可视化展示了空间域中物理或化学性质的结构和演变。信息可视化侧重于抽象的，非结构化的和高维数据的表示，其中包括业务数据，社交网络数据和文本数据等【31】。随着人工智能和数据分析的不断发展和迭代，形成了一种新颖的分析策略，即可视化分析【32】。交通数据集往往是高维度的，或者说它是时变的，因此，交通数据可视化主要就是指用户信息可视化和可视化分析。

可视化和可视化分析对于一个高效的数据驱动的ITS而言是非常重要的。具体来说，交通数据可视化可以促进对移动物体（车辆）的行为的理解，甚至可以促进交通，社会，地理空间乃至经济模式的发掘。一般来说，分析系统由四部分组成：数据收集，数据预处理，数据查询和数据分析。每个部分都需要专门的可视化技术。例如，可视化数据清理可以帮助用户转换数据，使其可用于下游分析任务【33】。其他过程如聚合和聚类也可以通过可视化界面来增强【34】。一个友好的用户查询界面必须要有自动纠错的功能【14】。此外，交通情况监测和交通模式识别也被广泛地研究，以实现智能轨迹分析的目的【11】【35】。考虑到现有的交通数据分析应用的任务，交通数据可视化的任务需求可以被分类如下：

* 可视化的交通情况监控 我们可能在交通堵塞等交通数据中发现一些有趣的信息。通过实时监测数据（例如隧道或交叉口的视频监控），可以观察和跟踪实时交通状况和综合事件，以了解长途交通堵塞的原因和机制。基于出租车轨迹的视觉分析系统就是一个很好的例子。
* 运行模式的发掘与集群化的探索 交通数据可视化的一个重要目标是发现对象



的移动模式并将这些模式集群化。这些模式反映了个体运动的特点，演变以及与其它参数的关联。例如，Gennady Andrienko等人[36]提出了一种交互式视觉聚类方法来对大质量物体的轨迹进行分类。

* 情境意识的探索和预测 数据分析任务可以分为两类:描述和预测。许多分析系统都能够探索和解释交通状况，例如，在城市中以视觉方式查询出租车行程[14]，并预测城市中大型汽车的轨迹[37]。
* 路线规划和建议 交通法规和路线建议是ITS的重要组成部分。数据驱动的控制和规划已被证明是有效的，并且取得了令人满意的结果。将人力资源纳入分析过程可以进一步提高效率，如可视化辅助路线推荐系统[10]所示。

交通数据可视化系统通常包含四个数据状态和三个流程阶段，如图1中的可视化管道[38]表示。数据流包括四个状态，即原始数据、处理数据、可视化符号和可视化。原始数据可以从不同的数据源收集，包括视频监控，车辆的GPS和事件日志。预处理的数据包含时间，空间，时空等多变量属性。然后，进行视觉转换，将数据转换为适当的设计和放置视觉符号，例如线形图，条形图，圆点图和星形图等等。最后，视觉符号和隐喻用不同的视觉通道(颜色、透明度、纹理等)绘制成各种可视化的形式，如彩色图像、信息图形、动画和视频。用户可以使用用户界面调制每个阶段的参数。在数据预处理阶段，用户可以通过迭代调整参数来优化预处理功能。在视觉转换阶段，用户可以对数据进行转换和筛选，以确定所显示的数据。在视觉映射阶段，用户可以操作视觉映射类型并与视觉符号交互。这样，隐藏在数据中的模式和知识就可以很容易地被理解和发现。

本文的其余部分组织如下:第2节介绍了交通数据和预处理技术的详细内容。第三节详细阐述了各种可视化技术，包括时间、位置和其他聚合或减除的变量。第四节展示了可视化如何与分析技术相结合，以提高对交通数据的理解和挖掘。最后，对该调查进行了总结，并在第五节中进行了重点研究。

ii．交通数据预处理

不同类型的交通数据需要不同的可视化和分析方法。实际捕获的数据通常是原始的、错误的，包含不确定性、异常值、丢失的值或不匹配的项。必须处理原始数据以进行可视化分析。

1. 交通数据

交通数据是指在道路上安装的交通车辆或监视器中的传感器产生和收集的数据集。 交通数据的例子包括车辆的GPS数据，GSM位置或人员移动的小区站记录，以及监视设备的视频/图像/计数记录。 传感器的工作模式大致可以分为：

以下三类[39]，[40]：

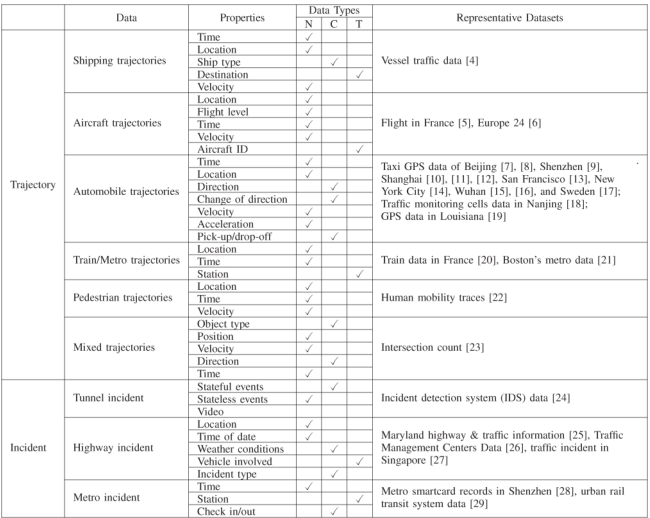
* 基于定位信息的 进入传感器范围后记录物体的位置。 例如，在交叉的情况下，视频监视装置只有当他/她经过监视器时才捕捉行人的位置和移动方向。
* 基于运动信息的 当一个对象进行某种活动时，记录相关或派生的信息。 例如，当他/她拨打电话时，GSM用户的位置被自动记录。
* 基于设备信息的 由对象携带的设备主动记录并发回位置信息和其他信息。例如，带GPS设备的出租车每20秒向数据中心发送一个信号。

轨迹是最常见的交通数据形式。轨迹包含时间信息，记录运动的时间轴和空间信息，记录每个时间点的位置。以往的研究集中于对轨迹的形象化和分析，包括[4], [7],[13],[18],[23],[41]-[43]。其他类型的轨迹信息也可以利用[39]，包括运动方向，改变方向[23]，运动速度[11]，速度变化[13]。

事件日志基于事件，并包含事件类型、事件位置和相关实体的其他信息等属性。

表格1

交通数据的例子。在这里，N、C和T分别代表数值、分类和文本。



除上述两种数据类型之外，还可以从轨迹或事件记录中导出其他多变量数据，或由特殊传感器记录，包括速度，方向和加速度。 表I总结了现有数据集及其各自的属性。

1. 数据预处理

在数据分析之前，进行一系列数据预处理操作是必需的，包括清理数据，匹配、组织数据和数据聚合。

* 数据清理 错误数据，异常值和原始数据的冲突值必须清除[33]。 在[7]中，应用了一个称为GPS数据清理的步骤来消除GPS错误并滤除无用的记录。

典型的数据清理过程分为三个阶段：审计数据以发现差异，选择转换来修复这些差异，并将转换应用到数据集[44]。

第一阶段检测原始数据中的错误 .Rahm列举了原始数据[45]中的主要问题，包括唯一性，参照完整性，拼写错误，冗余和矛盾价值。 检测原始数据中的错误的常规方法包括数据分析和数据挖掘。

在第二阶段，数据转换是根据数据源的数量和数据的肮脏程度来仔细设计和选择的。 这个阶段可以手动或自动完成。 例如，用户可以编写自定义脚本来控制整个清洁程序或使用提取/转换/加载（ETL）工具来转换数据。

第三阶段执行数据集的转换并用已清理的数据替换脏数据。 在流量可视化系统中，需要对已清理的数据进行进一步处理以适应分析任务。

* 数据匹配 原始交通数据记录是离散的采样点，可能与城市中的道路网络不匹配。地图匹配，即将观察到的用户位置序列与数字地图上的道路网络对齐，是数据预处理中不可或缺的一步[46]。 现有的地图匹配算法可以分为四类，即几何，拓扑，概率和其他先进技术[47]，如表2所示。
* 数据组织 预处理的数据需要在数据库或数据仓库中进行组织。一个经过充分研究的数据库应该支持查询结果的交互式查询和可视化，并且应该与移动对象的数据（如轨迹）兼容。索引方法分为两类。 第一类包括多维索引方法，如3D R-Tree [58]，STR-Tree [59]和HR-Tree [60]。 第二类包括将空间划分为网格并为每个网格建立时间索引的索引方法，如SETI [61]和MTSB-Tree [62]。 数据立方体是另一种标准的数据结构，可以为数据查询提供快速响应[63]。最近，Nanocubes [64]已经发展为支持随着时间的快速索引和聚合空间区域的查询。 一些关系数据库，如PostGis（PostgreSQL的扩展）[65]和MySQL Spatial（MySQL的扩展）[66]提供了空间数据的空间扩展。
* 数据聚合 交通数据集通常包含空间和时间属性，并且跨越大规模的空间和时间。 数据聚合[25]可有效减少数据量，并为随后的分析提供便利。 流量数据的基本聚合操作是空间（S），时间（T），方向（D）和属性（A）相关聚合。它们的组合产生不同类型的聚合：S×T聚合，S×T×A 聚合[25]，[67]，S×S×T×T聚合[68]，S×T×D聚合[69]和S×S聚合[69]。 S聚合主要通过计算区域内每个网格内的数据点的密度来完成。 T聚合用于显示沿时间轴的变化，并通过合并每个时间间隔中的数据点来完成。 对应于T聚合的最常见的可视化是时间直方图。 S×T聚合只是在连续的时间间隔计算密度[69]。时变密度可以通过动画密度图可视化。 S×T×A

聚合[67]首先在定期采样网格的基础上组合空间记录，然后聚合每个网格中的时间属性。 S×S×T×T聚合根据起始位置，结束位置，开始时间和结束时间[68]组合聚合。 它计算在一段时间内从一个地方移动到另一个地方的实体的数量。 S×T×D聚合不仅按空间和时间汇总数据，而且按移动方向聚合[69]。 S×S聚合组具有相同开始位置和相同结束位置的轨迹或运动[69]。 不同的聚合策略满足分析任务的不同要求。

III 交通数据可视化

交通数据包含多个变量，其中最重要的变量是时间和空间。 本节介绍专门针对时间，地点，时空信息和其他交通数据属性设计的可视化技术。

1. 时间的可视化

一般来说，时间可以分为线性时间，周期时间和分支时间。 面向时间的可视化[70]强调沿时间轴显示数据的趋势，周期性和异常性。

1) 线性时间：线性时间将时间视为从开始时间点到结束时间点的线性场。 它是最广泛使用的时间表示，并产生一系列时间线可视化技术。 例如，在折线图中，时间沿X轴表示，另一个变量沿Y轴表示。 图2显示了美国纽约市给定日期的每次出租车行程的提示[14]。 折线图易于阅读，但由于杂波问题，它们不是显示多个变量的正确选择。

线性时间能够表达流量数据如何随时间变化，并指示随时间变化的变化的峰值或谷值。

2）周期时间：许多递归过程发生在我们的自然世界中。许多时间与时间有关，例如季节，周和日的迭代。 可视化周期性的常用方法是使用径向布局，如图3中的交通信息可视化[11]。 在图中，每个圆代表一个星期的一周，而每个圆的代表一个小时。 径向布局的优点是它容易显示周期性图案，缺点是空间效率低。