Short and Long-term Pattern Discovery Over Large-Scale Geo-Spatiotemporal Data

大规模地理时空数据的短期以及长期模式发现

Department of Computer Science and Engineering

计算机科学与工程

The Ohio State University

俄亥俄州立大学

ABSTRACT

摘要

Pattern discovery in geo-spatiotemporal data (such as traffic and weather data) is about finding patterns of collocation, co-occurrence, cascading, or cause and effect between geospatial entities.地理时空数据（如交通和天气数据）中的模式发现是关于寻找地理实体之间的搭配、共生、级联或因果模式。Using simplistic definitions of spatiotemporal neighborhood (a common characteristic of the existing general-purpose frameworks) is not semantically representative of geo-spatiotemporal data.使用简单的时空领域定义（现有的通用框架的共同特征）不能在语义上代表地理时空数据。We therefore introduce a new geo-spatiotemporal pattern discovery framework which defines a semantically correct definition of neighborhood; and then provides two capabilities, one to explore propagation patterns and the other to explore influential patterns.我们因此~~介绍~~应用一种新的地理时空模式发现框架，它定义了一种语义上正确的领域定义；然后提供了两种功能，一种是探索传播模式，另一种是探索影响模式。Propagation patterns reveal common cascading forms of geospatial entities in a region.传播模式揭示了一个区域中地理空间试题的常见级联形式。Influential patterns demonstrate the impact of temporally long-term geospatial entities on their neighborhood.影响模式显示了时间上长期地理空间实体对其领域的影响。We apply this framework on a large dataset of traffic and weather data at countrywide scale, collected for contiguous United States over two years.我们将这个框架应用在全国范围内的大型交通和天气数据集上，这些数据时在连续两年内为美国连续收集的。Our important findings include the identification of 90 common propagation(传播) patterns of traffic and weather entities (e.g., rain -> accident -> congestion), which results in identification of four categories of states within the US; and interesting influential patterns with respect to the “location”, ”duration”, and “type” of long-term entities (e.g., a major construction -> more traffic incidents). 我们的重要发现包括对交通和天气实体的90种共同传播模式的确定（例如，雨->事故->塞车），这导致了在美国四类地区的确定。还有关于长期实体的“地点”、“持续时间”和“类型”的有趣的影响模式（例如，一个主要的建筑 -> 更多的交通意外）。These patterns and the categorization of the states provide useful insights on the driving habits and infrastructure characteristics of different regions in the US, and could be of significant value for applications such as urban planning and personalized insurance. 这些模式和州的分类为美国不同地区的驾驶习惯和基础设施特征提供了有用的建议，对城市规划和个性化保险有重要的价值。

1 INTRODUCTION

Spatiotemporal pattern discovery has seen considerable interest over the past decade, with various frameworks were proposed to process the data to find interesting patterns. 在过去十年间，时空模式发现引起了相当大的兴趣，人们提出各种框架来处理数据以找到有趣的模式。The application domains(域) of relevance(相关的) include public safety, transportation, earth science, epidemiology(流行病学), climatology(气候学), and environmental management. 相关的应用领域包括公共安全、交通、地球科学、流行病学、气候学和环境管理。These frameworks can be used to discover patterns of collocation and co-occurrence, interactions and correlations, cascading, sequential, or cause and effect relationship patterns. 这些框架可以用来发现搭配和共生、交互和相关性、级联、顺序或因果关系模式的模式。However, they all rely on a simplistic definition of spatiotemporal neighborhood, essentially(本质上) spatial closeness(空间接近度/空间封闭性) based on an Euclidean or Cartesian system and temporal overlap(时间重叠), which often makes their use impractical for applications such as traffic, transportation, or weather analyses. 然而，他们都依赖对时空领域的简单定义，本质上是基于欧几里得或笛卡尔系统的空间接近度和时间重叠度，这往往使得它们在交通、运输或天气分析等应用中的使用变得不切实际。For example, a traffic accident on one lane of a freeway has no impact on traffic flow on an opposite lane, yet general-purpose frameworks will locate both lanes in a single neighborhood. 例如，在高速公路一条车道上的交通事故对对面车道的交通流浪没有影响，但是通用框架将这两条车道定位在一个社区。Another example arises when studying the impact of a snow event (on traffic flow) which continues well past when the snow event has ended. 另一个例子出现在研究研究雪事件（对交通流量）的影响，这个事件一直持续到雪事件结束。The time overlap constraint required by existing frameworks would hinder such a study. 现存的框架所要求的时间重叠显示将妨碍研究。Note that there may not be any trivial changes to be made to make the existing frameworks semantically(语义上) applicable for this type of data. 请注意，可能不会有任何琐碎的更改以使现有框架在语义上适用于此类数据。Because, their basis is on a specific way of defining spatiotemporal neighborhood, which changing that would make them unusable (e.g., regarding their pruning step) or expensive to be employed. 因为它们的基础是一种定义时空领域的特定方式，这种变化将使它们无法使用（例如，有关其修剪步骤）或使用成本很高。

To address(应对) these challenges, we propose a new framework for finding patterns in geo-spatiotemporal data. 为了应对这些挑战，我们提出了一个新的框架来寻找地理时空数据中的模式。This framework consists of two parts, one to explore propagation patterns, and the other to reveal influential patterns. 这种框架由两部分组成，一部分来探索传播模式，另一部分来揭示影响模式。Identifying(识别) propagation patterns requires the exploration of partially ordered sets of geospatial entities, that are spatially co-located and temporally co-occurring, with potential(潜在的) “cause and effect” relationships between the entities. 识别传播模式需要探索部分有序的地理空间实体集，这些实体在空间和时间上是共存的，实体之间还存在着潜在的“因果”关系。An example of this type is a rain event, which causes an accident, with the accident then causing congestion. 这种类型的一个例子是雨事件，它会导致事故，随着事故发生还会导致拥堵。Identifying influential patterns, on the other hand, requires studying the impact of temporally long-term geospatial entities (e.g. a major construction) on their spatial neighborhoods. 另一方面，识别影响模式要求研究长期地理空间实体（例如，一个主要建筑）对其空间领域的影响。An example of this type of pattern is the increase in number of congestion events in a region because of a long-term snowing event. 这种模式的一个例子是，由于长期下雪这个事件，一个区域的拥堵事件的数量增加了。

To explore propagation patterns – also referred as “cascading patterns” or “spatiotemporal couplings”, we propose a tree-pattern-mining-based process, we term short-term pattern discovery, which employs a strict definition of spatial neighborhood to ensure spatial collocation, and a definition of temporal co-occurrence specific to geo-spatiotemporal data and application domain constraints. 为了探索传播模式——也称为“级联模式”或者“时空耦合”，我们提出一种基于树模式挖掘的过程，我们称之为短期模式发现，这个过程采用对空间领域的严格定义来确保空间搭配（spatial collocation），以及针对地理时空数据和应用程序域约束的时间共现的定义。To explore influential patterns – also referred as “tele-couplings” – we propose a new process, we term long-term pattern discovery, to examine the effect of long-term entities on their neighborhood to reveal any significant impact. 为探索影响模式——也称为“远程耦合”——我们提出了一个新的过程，我们称之为长期模式发现，用来检查长期实体对其邻居的影响以揭示任何重大影响。As in, and drawing from, this process may be used to study impacts with respects to different types, different locations, and duration(持续时间) of long-term geospatial entities. 正如从中得出的，该过程可用于研究对长期地理空间实体的不同类型，不同位置和持续时间的影响。

To evaluate(评估) our framework, we used a large-scale, real-world geo-spatiotemporal dataset of traffic and weather data. 为了评估我们的框架，我们使用了一个大规模的，现实世界的地理时空数据集的交通和天气数据。This dataset covers the contiguous United States, includes data collected from August 2016 to August 2018, and contains about 13.1 million instances of traffic entities (e.g., accident, congestion, and construction), and about 2.2 million instances of weather entities (e.g., rain, snow, and storm). 这个数据集包括了连续的美国，包括从2016年8月到2018年8月收集的数据，并且包含大约1310万个交通实体实例（例如，交通事故，拥堵和建筑），以及大约220万个气象实体实例（例如，降雨、降雪和风暴）Through the process mentioned above, we found 90 common patterns of propagation of relatively(相对) short-term traffic or weather entities, and identified four categories of states based on these patterns. 通过上面讲的过程，我们发现了相对短期的交通或天气实体的常见传播模式，并在这些模式的基础上定义了四种状态类别。In addition, we carefully studied the impact of relatively long-term traffic or weather entities on traffic, and identified a variety of insights with respect to “location”, “type”, and “duration” of the entities. 此外，我们还仔细地研究了相对长期的交通或天气实体对交通的影响，并且确定了关于实体的“位置”、“类型”和“持续时间”的各种见解。The main contribution of this paper are as follows: 本文主要贡献：

* Short-term pattern discovery(短期模式发现): We propose a new process for discovering propagation patterns in geo-spatiotemporal data, which models spatiotemporal collocation and co-occurrence in terms of tree structures, and adopts an existing tree pattern mining approach to reveal prevalent(流行) patterns. 我们提出了一个在地理时空数据中发现传播模式的新过程，它从树结构的角度对时空搭配和共现进行建模，并且采用一个现有的树模式挖掘方法来解释流行的模式。In comparison to the general purpose frameworks, this method better suits application domain requirements of a stricter definition of spatiotemporal neighborhood. 与通用流行的框架相比，这个发现更适合时空领域更加严格的应用领域要求。
* Long-term pattern discovery(长期模式发现): We propose a new process for discovering influential patterns in geo-spatiotemporal data, which examines the impact of long-term geospatial entities on their neighborhood in order to reveal significant influential patterns. 我们提出了一个新的用来发现在地理时空数据中发现影响模式的方法，他考察长期地理空间对其领域的影响，以揭示重要的影响模式。Exploring such patterns with existing frameworks is not feasible(可行的), due to lack of effective spatiotemporal neighborhood metrics(指标) to explore longer-term (or lagging) impacts. 因为缺少有效的时空领域度量标准来探索更长期的模式（或者滞后）的影响，所以在现有的框架下探索此类模式是不可行的。
* Data collection and processing(数据收集和处理): We present a set of processes for collecting real-time traffic and historic weather data, using which we built a publicly available “research dataset” of 13.1 million traffic entities (e.g., accident, congestion, and construction), and 2.2 million weather entities (e.g., rain, snow, and storm). 我们提出了一套收集实时交通和历史天气数据的过程，使用它建立了一个公开可用的“研究数据集”，其中包含1310万交通实体（包含交通事故、拥堵和建筑）和220万天气实体（例如下雨、下雪和风暴）。This dataset is accessible from <https://smoosavi.org/datasets/lstw>.
* Findings and insights(研究结果和见解): By applying our new framework on the above dataset, we present a range of insights for different regions in the United States. 通过将我们的新框架应用于上述数据集，我们提供了针对美国不同地区的一系列见解。These insights may be further utilized for applications such as urban planning, exploring flaws in transportation infrastructure design, traffic control and prediction, impact prediction, personalized insurance, potentially(潜在地) with relevance to the creation of smart cities. 这些见解可以进一步用于诸如城市规划，探索交通基础设施设计，交通控制和预测，影响预测，个性化保险等方面的应用，这些应用可能与智慧城市的创建有关。

The rest of this paper is organized as follows: We review the related work in Section 2, and provide preliminaries(初步) in Section 3. Section 4 describes the dataset preparation, followed by description of framework in Section 5. Experiments and results are presented in Section 6, and Section 7 concludes the paper. 本文其余部分组织如下：我们在第二节回顾相关工作，并在第三节中提供了初步的知识。在第四节介绍数据集的准备工作，然后在第五节中介绍了框架。第六节介绍了实验和结果，在第七节总结了本文。

2 RELATED WORK

Spatiotemporal pattern discovery has been thoroughly discussed in literature [3, 4, 14, 19-21, 24]. 时空模式发现在文献[3, 4, 14, 19-21, 24]中已经有深入的讨论。Earlier work focused more on spatial prevalence(流行；广泛) and paid less attention to temporal aspects, while later work considered both aspects simultaneously. 早期的工作更多地关注空间流行而对时间方面关注很少，而后期的工作则同时考虑两个方面。The common process of spatiotemporal pattern discovery is to first define spatiotemporal co-occurrence and collocation criteria; then introduce an interest measure (e.g., participation index); and finally outline a miner algorithm to find interesting patterns. 时空模式发现的相同过程是首先定义时空co-occurrence和collocation标准；然后引入一个兴趣度量（例如参与度）；最后概述一个来寻找有趣模式的挖掘算法。Techniques in these papers being general purpose solutions, rely on simplistic definitions of collocation (spatial) and co-occurrence (temporal), and unable to reveal complex spatiotemporal correlations(相关性) (such as influential patterns). 这些论文中所用到的技术是通用的解决方案，依赖于collocation（空间）和co-occurrence（时间）的简单定义，无法揭示复杂的时空相关性（例如影响模式）。Further, they have been developed and only tested on small-scale (real-world or synthetic) data. 此外，它们还只是在小规模（现实世界或者合成）数据上开发和测试。To address these challenges with respect to geo-spatiotemporal data, we propose a new framework which provides an appropriate(适当的) and precise(精确的) definition of collocation and co-occurrence criteria. 为了解决地理时空数据方面的这些挑战，我们提出了一个新的框架，为collocation和co-occurrence的标准提供了一个适当准确的定义。Moreover, we outline the process of finding complex spatiotemporal patterns and prove its applicability(适用性) through extensive(广泛的) experiments. 此外，我们还概述了寻找复杂时空模式的过程并且通过广泛的实验证明其适用性。Lastly, we apply our framework on a large-scale, countrywide geo-spatiotemporal dataset of traffic and weather data to explore interesting patterns. 最后我们将我们的框架应用于一个关于交通和天气的大规模、国家级的地理时空数据集来探索一些有趣的模式。

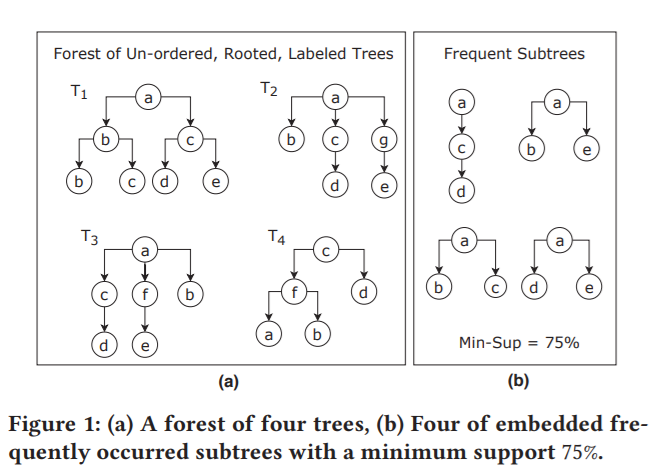
Regarding the application domain, there are numerous studies for finding patterns in traffic and weather data, with the following goals: 在应用领域，有许多寻找交通和天气数据模式的研究，其目标如下：to study the impact of precipitation on likelihood or severity of accident [7, 16, 28]; to explore the impact of weather on traffic intensity [5, 31]; to reveal the effect of climate change and weather condition on road safety [1, 11, 29]; to characterize road accidents locations [18]; or, to discover frequent spatiotemporal patterns in traffic data [15, 17, 19]. 研究降水对事故可能性或严重程度的影响；探讨天气对交通强度的影响；揭示气候变化和天气状况对道路安全的影响；描述道路事故地点；或在交通数据中发现频繁时空模式。The scale of data in most of these studies is limited to one or at most a few cities. 在大多数这些研究中，数据规模仅限于一个或最多几个城市。 Moreover, interactions(相互作用) and correlations(相关性) between the different types of traffic entities (accident, congestion, etc.) has not been studied before. 此外，不同类型的交通实体（事故、拥堵等）之间的相互作用和相关性之前还没有被研究过。Although similar ideas to explore long-term patterns have been preciously suggested [7, 11, 16], we extend them by: 1) examining(examine，检查；调查，测试) a wide range of weather and traffic entity types besides precipitation; 2) exploring properties of different “locations”; and 3) analyzing the impact of “duration length” on traffic flow. 虽然以前曾提出过类似的探索长期模式的想法，但是我们将其扩展：1）测试/检查除降水外的更广泛的天气和交通实体；2）探索不同“地点”的特性；3）分析“持续时间长短”对交通流量的影响。

3 PRELIMINARIES AND PROBLEM

In this section, we first provide preliminaries and definitions, and then present the problem statement. 在这一节中，我们首先提供预备知识和定义，然后给出问题陈述。Note that some of the definitions are customized for our illustration application domain (i.e., traffic and weather data). 请注意，一些定义是针对我们的插图应用程序域（即交通和天气数据）定制的。 However, this will not limit their generalizability to the other related domains.然而，这不会限制它们对其他相关领域的可推广性。

3.1 Definitions

* Geospatial Entity(地理空间实体): a geospatial entity is represented by a tuple , which shows an entity of type , happened in time interval , and its location is specified by . 一个地理空间实体 由元组组成，表示实体的类型，表示发生的时段，还有指定了实体的位置。Definition of is related to the application domain. 的定义和应用领域有关。For traffic data, we have , where shows the relative side of a street (i.e., R or L). 对于交通数据，我们有，表示街道的相对侧（即左或右）。For weather data, we have , which represent the “airport” that is reported form its weather station. 对于天气数据，我们有，代表实体从气象站报告的“机场”。A geospatial entity is called , if it takes place over a relatively long time interval (see Section 5.2).如果一个地理空间实体发生在一个相对较长的时间间隔，那么它就称为长实体。
* Weak-Dependency Relationship(弱依赖关系): two co-occurring and co-located geospatial entities are called weakly dependent.两个共同发生（co-occurring）和共存（co-located）的地理空间实体被称为弱依赖。 Co-occurrence for two entities and means , where is a time-threshold. 两个实体和共现是指 ，其中 是一个时间阈值。Collocation for two traffic entities requires location matching as well as spatial closeness. 两个交通实体共现要求位置匹配以及空间紧密度。The former means that all location fields except the GPS coordinates should be the same. 位置匹配意味着除GPS坐标外的所有位置字段都应该相同。By latter, we require that , where is the Haversine distance function based on GPS coordinates, and is a distance threshold. 通过后者，我们要求 ，其中是基于GPS坐标的Haversine距离函数， 是一个距离阈值。With respect to matching a pair of weather and traffic entities, collocation means a match between the “airport station” at which the weather entity is reported and the “airport station” closest to the traffic entity’s location. 关于匹配一对天气和交通实体，collocation意味着在报告天气实体的“机场站”与最接近交通实体位置的“机场站”之间进行匹配。
* Child-Parent Relationship(子父关系): for two weakly dependent geospatial entities and , is a parent for if begins before . 对于两个弱依赖地理空间实体和，如果在之前开始则是的父亲。We treat parent-child relationship as indicative of a and relation. 我们将亲子关系看作是因果关系的象征。A weather entity may only be the parent (or cause) of a traffic entity, and we do not define such a relationship between two weather entities. 天气实体可能只是交通实体的父亲（或者原因），而且我们不定义两个天气实体之间的这种关系。
* Tree Structure(树结构): given a set of vertices , we define tree , where and is a set of edges, and each connects a pair of vertices using an un-directed edge. 给定一组顶点，我们定义树，其中 ，是一组边。每个 通过无向边连接一对顶点 。A tree is an acyclic graph(无环图), and vertices with the same parent are siblings(兄弟姐妹). 树是一种无环图，具有相同父亲结点的节点是兄弟姐妹。Trees in this work have a root node(根结点), sibling nodes are un-ordered(无序的), and nodes are labeled(被标记的). 在本工作中树有一个根结点，兄弟结点是无序的，结点被标记。Figure 1-(a) shows several examples of such tree structure. 图1-(a)显示了这种树结构的几个例子。In this work, each node of a tree is a geospatial entity, and each edge shows a child-parent relationship between two entities.在本工作中，每一个树节点都是一个地理空间实体，每条边都显示两个实体之间的子父关系。



* Embedded Subtree(嵌入子树): given a tree , we define a subtree as , where and . 给定树，我们将嵌入子树定义为，其中 ， 。A subtree is said to be an embedded subtree of if for each edge , is an ancestor (and not necessarily the parent) of in .一棵子树如果它的每条边，其中在树中是的祖先（不一定是父亲结点），那么被称作的嵌入子树。

3.2 Short and Long-term Pattern Discovery

We now formalize(使形式化) the two related problems studied in this paper. 现在我们将本文研究的两个相关问题形式化。

*3.2.1 Short-term Pattern Discovery.* Here we seek to find common short-term propagation patterns that indicate how geospatial entities cause other entities to happen. We represent a set of weakly dependent geospatial entities as un-ordered, rooted, labeled stress, where the entities are nodes, weak dependency relations are the edges, and entity types (e.g., rain, accident, and congestion) are the labels of the nodes. Thus, given a forest of such tress, the short-term pattern discovery problem is about finding all embedded subtrees in which are occurred relatively frequently. Formally, for a subtree and tree we define by Equation 1:

