摘 要

近年来，随着移动设备飞快普及与硬件存储、计算能力的飞快提升，每天都有海量的轨迹数和带地点标签的签到数据以惊人的速度产生。这些数据蕴含着着人们的移动规律以及出行模式，因而高效地对轨迹数据进行存储、压缩、表征以及知识挖掘将对现有经济、环境、交通等领域产生深远影响。在商业方面，探索如何从海量的用户轨迹或签到数据中挖掘出用户喜好信息，进而向用户推荐潜在的感兴趣地点，将使人们的生活得到极大的提升，也能为不同规模的商业经济带来巨大的效益。

针对轨迹数据的特征，本文提出了一种全局的轨迹压缩、表征方法。区别于传统的将轨迹数据逐条压缩处理，本文将对海量的轨迹数据进行全局处理。这种做法不仅非常高效，还能通过结合都市中丰富的已知POI信息来抓住传统方法忽略的全局统计以及语义信息。这样一方面使得轨迹稀疏的区域的轨迹表征得到矫正，另一方面也能借助大量轨迹数据来探索和理解关键地点的语义信息。之后，本文指出，传统的地点表征仍然是直观的基于地图的距离的，没有对潜藏的各种语义进行探索，借助词向量的嵌入表征方法，本文创新性地将地点表征为隐向量，隐向量之间的相似度即可提现地点间的语义相似度，这将为后续应用场景提供非常有效的表征基础。最后，在地点推荐算法层面上，本文指出了传统协同过滤不能够产生有解释性的推荐结果，进而结合了集成学习的的思想，提出一种局部分步矩阵分解的协同过滤算法，弥补了基于矩阵分解的推荐算法在解释性上的空白。总观全文，本文的创新点体现在以下三方面：

第一、本文提出了一种全局的轨迹压缩表征方式，将整个轨迹数据集表征为一个多粒度的地点网络。这个网络可根据应用的需求将已知的地点信息包含进来以增强轨迹的表征压缩效果。

第二、结合多样化的地点挖掘需求，本文提出了一种将地点表征为隐向量的方法，使得地点之间的高层语义相似度可以直接从隐向量之间的相似度中获得。这将大大提高地点检索的效率。

第三、针对传统矩阵分解在地点推荐中缺乏解释性的缺陷，本文提出一种改进型的局部分步矩阵分解。这种方法应用在地点数据集上将让产生的推荐隐因子具有更具体的含义，从而获取用户的信赖程度，也增强了研究者对于算法的理解。

本文通过实验说明了提出方法的可行性，其贡献填补了轨迹数据挖掘与地点推荐一些空白。

关键词：轨迹压缩，轨迹表征，地点推荐

ABSTRACT

Recently, with the pervasive use of mobile devices and the improvement of store and computing capacity, massive amount of trajectory data and check-in data have been generating at a dazzling speed. There are human mobility patterns waiting to be exploited behind those data. Therefore, trajectory mining topics such as storing, compression and representation draw growing attention to provide profound influences in economy, environment and transportation. In business, the user preference and other knowledge extracted from the sea of trajectory data and check-in data can benefit the location recommendation process. Thus human life quality can be improved, and business with different scales can make a profit out of it.

To handle the trajectory data, this work proposes a trajectory compression and representation method. Different from traditional methods which deal with each trajectory individually, this work globally processes the whole data set. Not only is the process very efficient, but the semantic information that those traditional methods ignore is also captured by integrating the auxiliary urban point of interest (POI) information. By doing this, for one thing, the trajectories located on the spares areas can be regulated and corrected. And for another, the places of interest can be understood well. Furthermore, to overcome the defect that places are represented by coordinate pairs on the map and thus the semantic information cannot be revealed, this work proposes a representation method that project a place to a distributed latent vector. The similarity between two latent vectors stands for the semantic similarity of the two places, and thus the downstream applications can be expedited. At last, this work provides a novel explainable place recommendation method. By locally applying a forwarding stagewise manner matrix factorization on the rating data, the result factors are enriched with meanings and the recommendation results become easy to explain. To conclude, the contributions of this work is as bellows.

1. A semantic trajectory compression model is proposed by considering both global trajectory structure information and available contextual information. This method provides a new perspective for compressing trajectories with semantics.
2. Utilizing the geometric property and semantic information (network structures, temporal information, and domain knowledge), this work proposes a hierarchical embedding model to embed each region or trajectory as a continuous vector in a

semantic vector space. Thereby, the semantic similarity between two regions or trajectories can be measured by computing the Euclidean distance of two vectors directly.

3. A boosted local rank-one matrix approximation (BLOMA) model is proposed. It has three major differences comparing to traditional matrix approximation-based collaborative filtering methods. In BLOMA, the topics of latent factors are more distinct, which makes the recommendation result explainable.

It is through the experiment results that we demonstrate the effectiveness of our methods, which fill the blank of the related research area.

Keywords: Trajectory Compression, Trajectory Representation, POI Recommendation

目 录

第一章 绪论

从轮子的发明，到登月火箭的实现，上千年来的科技的进展给人类的交通方式带来的非常大的改变。越来越多的通信技术开始关移动物体的轨迹。现如今，我们在每一天我们使用计算机、手机时候，都有大量数据产生，接着被以各种形式记录、保留下来。这其中，大量数据是带有地点标签的，这些数据随着时间扩展，便可得到人们日常出行的轨迹以及这个过程中的各种信息。这些信息记录了我们在什么时间，什么地点做了什么东西，利用好这些移动数据，将给未来人们的生活带来革命性的改变。

那么，如何利用好记录了移动模式的轨迹数据进行挖掘和探索呢，本文在将首先介绍轨迹数据挖掘中的一些任务，明确目前轨迹数据中存在的问题，再介绍目前国际上存在的主流方法与技术以及这些技术存在的问题。针对这些问题，本文将提出创新的的方法来对目前的算法进行改进以填补这部分应用的空白。接下来，轨迹的产生的以及记录形式将被正式的介绍。

1.1 轨迹数据挖掘

1.1.1 数据挖掘流程

数据挖掘（Data Mining）是一门综合性的学科。通常来说，数据挖掘是数据库知识发现（Knowledge-Discovery in Databases）中的一个步骤，其目的是在大量的数据中自动搜索隐藏于其中的特殊信息，从而为之后的分析决策提供理论依据。数据挖掘的主要步骤为：

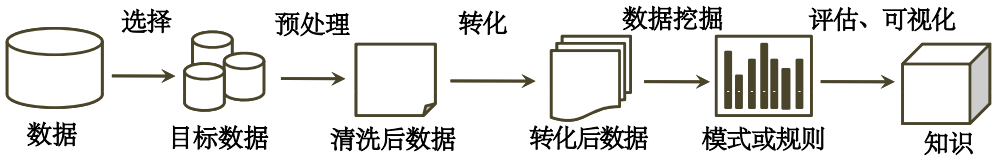

图 1-1 数据挖掘主要步骤图（来源：Synchronization Inspired Data Mining[1]）

- **数据采集** 所有工作开始之前，首先需要采集数据，包括确定数据种类、范围等，然后对数据进行初步选择，挑选出合适的数据。
- **数据预处理** 该过程包括对原始数据的处理，包括数据整合、去除噪声等。
- **数据转化** 对数据进行完预处理后，需要决定数据合适表示，例如特征选筛等。

- **数据挖掘** 这个过程中，人们采用各种方法，例如聚类、分类、关联规则分析等方法来发掘数据中的有用的信息。
- **结果评估与可视化** 最后，需要对得到的结果进行解释与评估，并可视化为易于人理解的形式，在这之后有可能需要重新进行挖掘。

这其中，**数据挖掘**是从数据中学习知识的最关键的步骤，因此很多时候，数据挖掘泛指从数据中学习知识的过程。关于数据挖掘的更多信息，读者可查阅经典的综述[2]和[3]。

1.1.2 轨迹数据的形式

由于记录设备的不同，轨迹的存在形式可以由多种。Spinsanti等人[4]将轨迹数据的形式区分为基于GPS（global positioning system），GSM（global system for mobile communications）和基于社交网络的轨迹这三种。Pelekis和Theodoris[5]又追加了两种轨迹数据，分别为基于RFID（radio frequency identification）的和基于Wi-Fi数据的。这其中，基于GPS系统的轨迹数据由一系列带时间戳的二维地理坐标构成。基于GSM的轨迹数据由一系列带时间戳的物体经过的细胞标号组成。而基于RFID的轨迹数据包含物体经过的一系列RFID接收器的标号组成，基于Wi-Fi数据的轨迹也一样由物体连接通讯过的一系列Wi-Fi基站标号组成。不同形式的轨迹数据的精度是不同的，它们在不同的应用场景下有着不同的应用。

通常，一条轨迹总是可以被表示为一下形式：

$$T = \{\langle s_1, s_2, \cdots, s_n \rangle \,|\, s_i = (P_i, t_i)\}, \tag{1-1}$$

其中P代表一个位置或者一个区域的标号，在基于GPS的系统中，$P_i = (x_i, y_i)$ 表示一个GPS经纬度坐标，也是地图上的一个采样点。而t_i是P_i的采样时间。n代表了轨迹T的采样数目。例如，图1-2可视化了大西洋在1851年-2018年间，每年飓风的移动轨迹①。图中每一条线即为一个轨迹，而轨迹上每一个圆圈代表一个采样点，在此处即为飓风在某一个月内的定位。

1.1.3 轨迹数据挖掘

除了这些经典的数据挖掘以外，也有研究者提出了在地理数据上进行数据挖掘的的综述[6–8]。区别于传统数据挖掘，它们主要考虑了方法在空间上的依赖性以及空间属性与非空间属性的结合。值得一提的是，这些工作都没有很好的考虑轨迹的时间特征。

① 数据来源：`https://www.nhc.noaa.gov/data/`

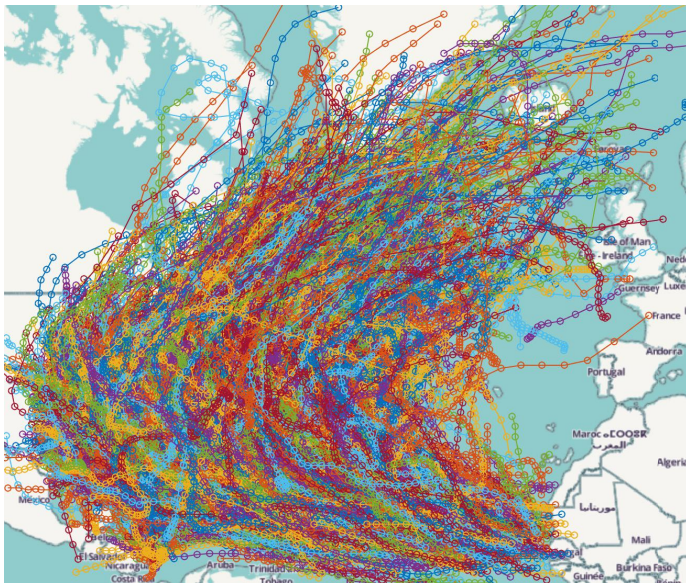

图 1-2 大西洋飓风1851年-2018年轨迹可视化图

根据Zheng等人[9]的划分，在轨迹数据上进行的数据挖掘步骤一般有四个任务：

1. **消除轨迹不确定性：**现实生活中物体的运动轨迹是连续的，但是由于内存、采样率等原因，我们得到的轨迹数据往往是离散的，因此在两个离散时间点之间的轨迹的运动状态就存在一定的不确定性。针对这个问题，一方面，为了提高轨迹数据挖掘的质量，一些工作致力于降低轨迹的不确定性[10, 11]，另一方面，为了保护用户的隐私，一些工作致力于扩大轨迹的不确定性[11, 12]。
2. **轨迹模式挖掘：**我们可以从大量的轨迹中挖掘轨迹的模式(单个用户的轨迹模式或者是群体的轨迹模式)。与常规的数据挖掘任务类似，轨迹模式挖掘任务包括运动模式挖掘[13]、轨迹聚类模式挖掘[14]、轨迹运动周期挖掘[15]等等。
3. **轨迹分类：**针对大量的轨迹数据，我们可以使用传统的监督学习方法对轨迹进行分类。按照不同的分类目的可以将轨迹分为不同的类别[16]，利用按照交通方式将轨迹分为“步行”、“开车”、“火车”等类别[17]，或者按照轨迹出行目的将轨迹分为“工作路线”、“旅游路线”等。
4. **轨迹异常检测：**轨迹异常检测就是寻找与其他轨迹显著不同的轨迹[18]。例如交通事故发生的路段产生的轨迹就可以被看作异常轨迹，对同一个用户来说，其某一天的轨迹因为与大部分时间的轨迹差异较大也有可能会被判定为异常轨迹。

现如今，大多数工作把重心放在轨迹数据挖掘上，却忽略了底层轨迹的表征问题。因为不同于其他领域的数据，轨迹数据是非结构数据，其长度不确定，其采样率不确定，就连轨迹的相似度计算都是一个问题。于是轨迹的存储和表征将是一个问题。为了解决这些问题，本文将在下一章节中介绍轨迹数据常用的表征方式。接下来，另一个高层次的应用：基于地理位置的推荐系统将在这里引入介绍。

1.2 地点推荐系统

推荐系统旨在为用户推荐其潜在感兴趣的物件、信息或者服务，已经得到了大量学术界和工业界的关注。如今的生活里，拿起手机，打开音乐App，当日刚兴趣的歌单已经生成；打开购物App，令人眼花杂乱的首页展现的结果已经经过了算法的高度定制，从而很大概率上是用户会喜爱继而购买的物品；打开新闻App，今日新闻列表里的主题都是接着昨天用户感兴趣的话题而展开的。对于信息消费者，也就是用户，从大量信息中找到自己感兴趣的信息变得越来越困难；对于信息生产者，让自己生产的信息在众多信息中脱颖而出也变得越来越困难。推荐系统正是为了解决这一矛盾而应运而生的。基本上，推荐系统已经成功地融入了各个领域：电子商务（Amazon、阿里巴巴、Netflix）、信息检索（Google、百度、Bing）、社交网络（Facebook、Twitter、微信、QQ）、位置服务（大众点评、高德地图、Yelp、Foursquare）、新闻推送（Google News、今日头条、腾讯新闻）等等。

1.2.1 常见推荐方法分类

传统的推荐方法主要包括基于用户的推荐方法、协同过滤与混合的推荐方法。由于对理解有帮助，以下对三种方法的思想做简要介绍：

1. **基于内容的推荐**：主要根据用户已经评分商品，找到其他内容上相似的物件作推荐结果，属于中的物件到物件关联(Item-to-Item Correlation)的方法。这要求算法首先通过一些显式反馈（例如评分、喜欢或者不喜欢）或隐式反馈（例如观看、搜索、点击、购买等行为）的方式获取曾与用户有过交互的物件，之后从这些物件的特征中学习用户的偏好，就能计算用户与待预商品在内容(由特征刻画)上的匹配度(或相似度)，最后根据匹配度对所有待预商品进行排序，从而为用户推荐潜在感兴趣商品。基于内容的推荐方法依赖于关于用户偏好商品的特征信息，不需要大量的评分记录，因此不存在评分数据稀疏的问题。同时，对于商品，只需要进行特征提取就可以

向用户进行推荐，可以解决了商品的冷启动（Cold Start）问题，但这种方式会遭遇到特征提取困难这一个问题。

2. **协同过滤：**这个方法类似于现实中口口相传(word-of-mouth)的过程，协同过滤利用相似用户之间具有相似兴趣偏好的方法的这一特点来发现用户对物件的偏好。主要包括基于记忆和基于模型这两种类型，基于记忆式的方法首先通过用户的历史评分差异来对用户（或商品）之间的相似度进行计算，然后再根据用户的历史评分和用户之间的相似度来计算效用值，基于模型的方法则主要是通过构建一个用户偏好模型来对用户商品的潜在偏好进行预测。协同过滤的方式则只要利用用户的历史评分数据，因此更加简单有效，是目前应用最为广泛应用的推荐算法。在2008年，当Koren[?]提出了基于矩阵分解的潜因子模型（Latent Factor Model）后，这种方式成了工业界最通用的模型。但这种方式也由于用户商品的评分数据相对于商品的总数量太少，遭遇数据稀疏的问题。此外，对于新用户或者商品，由于没有评分数据而无法进行推荐，也存在冷启动问题。
3. **混合推荐方法：**这一方式考虑以上方法存在各自的不足，故通过组合不同的推荐算法进行混合推荐，往往能够产生更好的推荐性能。常见的组合策略主要分为:后结合、中结合和前结合。后结合是指将两种或两种以上的推荐算法产生的推荐结果，以投票、线性结合或可信度组合等方式来产生最终的推荐结果，是决策层上的混合。而中结合的思想是以一种推荐算法为基础，结合另一种推荐算法，如:以协同过滤算法为基础，融入基于内容的推荐算法，这样做可以有效缓解数据稀疏问题，这是一种模型层面上的结合。前结合则是直接将多种推荐算法安插到同一模型里，将各种特征作为模型输入来产生推荐结果。例如，在这个同一模型里，将所有用户属性以及行为数据作为输入来产生推荐结果。其本质上是特征层的结合。

以上方法是传统也是最经典的推荐方法，而现如今，随着需求的增多，研究者们纷纷考虑结合更多除了交互信息以外的额外信息来辅助推荐，这样一方面将提升推荐的准确性，另一方面在数据融合后，应用的可能性也大大增多了。通常来说，传统的推荐算法只利用了用户—商品的交互或者评分矩阵，而在地点推荐领域，其特殊性在于有大量的时空信息可用以辅助推荐。接下来，本文将介绍地点推荐系统的基本知识。

1.2.2 地点推荐系统

由于目前科技的发展，人们的各种活动数据、社交数据以及出行数据都可联系起来，形成一个网络，被称为LBSN（Location-based Social Network）。地点推

荐，也被称为POI（Point of Interest）推荐，即为用户提供下一个落脚的地点，或者探索可能感兴趣的地方，是LBSN中非常重要的一环。虽然传统的推荐方法已经很成熟，但其无法完全移植到地点推荐的场景汇总。追根揭底，地点推荐场景与传统推荐场景的不同因素体现在三个不同方面：

1. **地理因素**：如同Tobler第一定律[19]中描述的：“所有事物都与其他事物有有着关联，但通常与更近的事物关联更紧密”。在LBSN中，Tobler第一定律意味着用户会更倾向于访问其附近的地点而非遥远的地点，而且用户可能会访问其喜欢的地点附近的地点。地理因素是地点推荐中最重要的一个因素，其体现了现实中用户的心理特点与行为模式。
2. **隐式反馈与数据稀疏问题**：在传统的推荐系统里，用户通常会显示地对商品（如书籍、电影等）提交一个分数评价，这个分值通常有一个范围限制，如[1,5]，体现出了用户对各种商品的喜爱程度，更高的评分意味着更强烈的喜爱。但是，在签到数据集中，用户与地点的交互是没有评分的，而是用一个频率计数来代替。这个频率是没有上限限制的，比如用户有可能访问一个地方上千次，而其他地方仅仅访问几次。另外，数据的稀疏性也是大问题，比如传统推荐数据集Netflix的稀疏程度为99%左右，而Gowalla数据集有数值的地方仅占整个数据集的2.08×10^{-4}左右。
3. **社交影响**：在LBSN中，一个很容易想到的假设是，朋友之间的偏好会互相收到影响，很多工作会利用这个假设来用用户朋友们的行为轨迹来影响该用户[20, 21]。一些研究者的工作中表明，社交因素将会影响推荐系统。然而这种印象也不是绝对的，Ye等人的工作表明96%的用户共享不到10%的兴趣，绝大多数用户是没有共同兴趣点的。故社交信息在地点推荐中并不像地理位置那样占重要因素。

在下一章节，本文将介绍更多地点推荐算法的细节。现在，我们将总结本文的贡献：

1.3 本文主要贡献与创新点

本文的核心工作围绕着轨迹数据与用户签到数据展开，其贡献与创新点如下：

- **全局轨迹压缩方法**：随着轨迹数据的数量级增长，轨迹的压缩成为了存储、处理的必要预处理步骤。通常的轨迹压缩算法通常将轨迹进行单条压缩，这样的缺陷是效率低，会丢失轨迹数据整体的统计信息。本文利用了Sync算法，将轨迹数据集进行整体压缩，从下到上压缩得到不同粒度的结果。根据

需求不同，不同粒度的结果蕴含着不同的语义。这样压缩还有一个好处，城市中已知的重要节点可以直接插入到压缩结果中，成为压缩后轨迹网络的一个节点。根据轨迹随时间增长的这一特点，本文还将这一轨迹压缩方法扩展到了数据流上。

- **区域与轨迹分布式表征：**至今，轨迹的相似度都是一个难题，而如何表示轨迹的语义级别的相似度更是一个问题。受到自然语言处理中的词向量嵌入的启发，本文将轨迹作为上下文信息将压缩得到的轨迹网络上的区域上的节点嵌入为一个隐向量。在这个嵌入的过程中，根据实际场景需求，各种语义相似度都可以考虑进来。让后续的地点推荐等应用更为便利。
- **提出一种可解释的地点推荐算法：**传统的地点推荐多是基于协同过滤中的矩阵分解模型，虽然这种方式准确率很高，但其分解出的隐因子是没有含义的。从某种意义上来说，这种模型是一种黑箱模型。在本文中，为了克服这个问题，我提出了一种可解释的矩阵分解算法。具体的做法是将原始的矩阵分解修改为了局部分步分解，让每次分解出的隐因子都有对应的含义。且由于每次分解的秩的降低，矩阵分解的效率将大大增加。

1.4 本文的结构组织与章节安排

本章从轨迹数据挖掘这一应用开始介绍，介绍了轨迹数据的形式，以及轨迹数据集上的一些挖掘任务。进一步地，关注点从原始轨迹到用户签到数据后，地点推荐任务和常见思想也被做了介绍。接下来的几章的安排如下：

- 第二章为相关工作，将分别对国际上主流的轨迹数据集上的压缩、表征算法，以及签到数据上的的地点推荐算法进行介绍。并对现有问题进行概括，为后面正文的动机做出铺垫。
- 第**??**章为本文的第一个主体方法，提出了全局轨迹压缩算法，并将其扩展到数据流上，最后用实验来证明我们全局轨迹压缩算法的可行性。
- 第**??**章紧接着上一张接着对轨迹的表征进行探索，在介绍其丰富的应用潜力后，本文将区域以及轨迹表示成为了带语义的隐向量。本文用轨迹上的检索任务来证明了提出算法的可行性。
- 第**??**章的探索领域从原始轨迹上升到了特定地点的推荐。在这章中，一种具有解释性的算法被提了出来，并结合LBSN中丰富的社交信息、地理信息，综合成了一种全面的，能让用户信任的算法。在这一章中，我将我的算法与国际主流的几种算法进行对比，证明了提出了算法的可行有效性。

- 第**??**章为总结和展望部分，总结了这篇文章的主要工作，给出客观的评价。最后给出了本工作没有涉及的部分和之后可以继续深入做下去的一些工作。

第二章 相关工作简介

随着轨迹数据的大量产生，获取轨迹数据难度的降低，对轨迹数据的挖掘和表征得到了越来越多科研工作者的注意。挖掘轨迹中的语义信息对分析人类、车辆、动物和气候变化等行为有重要的意义，对基于位置的应用也提供了很多帮助。通常我们获取到的原始轨迹数据都是不包含太多语义信息的时空数据，这些数据通常数目巨大，格式、内容不同，长短不一，因此在挖掘轨迹数据之前寻求合适的压缩方法及合理的轨迹表征方法就显得非常重要。本章节将首先介绍目前轨迹数据挖掘中轨迹表征领域的相关工作，再针对性的对已有方法的不足做出总结。之后，本章节将探讨地点推荐的常用科技以及现状，一些重要的问题将在这里被揭示，为后面本文的创新点做出铺垫。

2.1 轨迹压缩方法小结

根据Renso等人[22]关于轨迹压缩的全面的综述，轨迹压缩的目标大体可以分为三个：（1）减少轨迹数据集的规模，从而使得海量数据达到方便处理的程度；（2）提升轨迹数据集上的计算效率；（3）确保压缩后的轨迹对比压缩前的轨迹没有太大的偏差，或者偏差在一个可接受的阈值内。下面是一些主流的轨迹压缩方法，其大体分为两类：单纯轨迹压缩方法与考虑语义的轨迹压缩方法。下面将展开介绍：

2.1.1 单纯轨迹压缩

为了解决轨迹数据增长迅速的问题，很多研究者提出了各种轨迹压缩的科技。Sun等人[23]给了一个轨迹压缩的综述。其中，早期的工作多是把轨迹视为一些直线的衔接组合，这就把轨迹压缩变为了欧式空间的线条的简化问题。例如，早期最出名的Douglas-Peucker算法[24]的思想是，迭代的选出轨迹中间的一些点，使得这些点到去除这个点的线段的垂线距离不超过一个给定阈值范围，这个过程将不断进行，直到剩下的所有点都满足条件。如图2-1所示，链接A,F两点，在一定阈值范围里（红局限的短边长），D点是第一个超过阈值的的点，于是把轨迹划分为AD与DF两段，继续找到阈值外的C点。其余点都不超出阈值，于是在简化中被去除。最终，压缩后的轨迹如图2-1(d)所示。

在Douglas-Peucker算法的基础上，一些算法[25, 26]将原始的距离度量改为了同步的欧式距离（Synchronized Euclidean Distance，SED），这样轨迹的时间信息

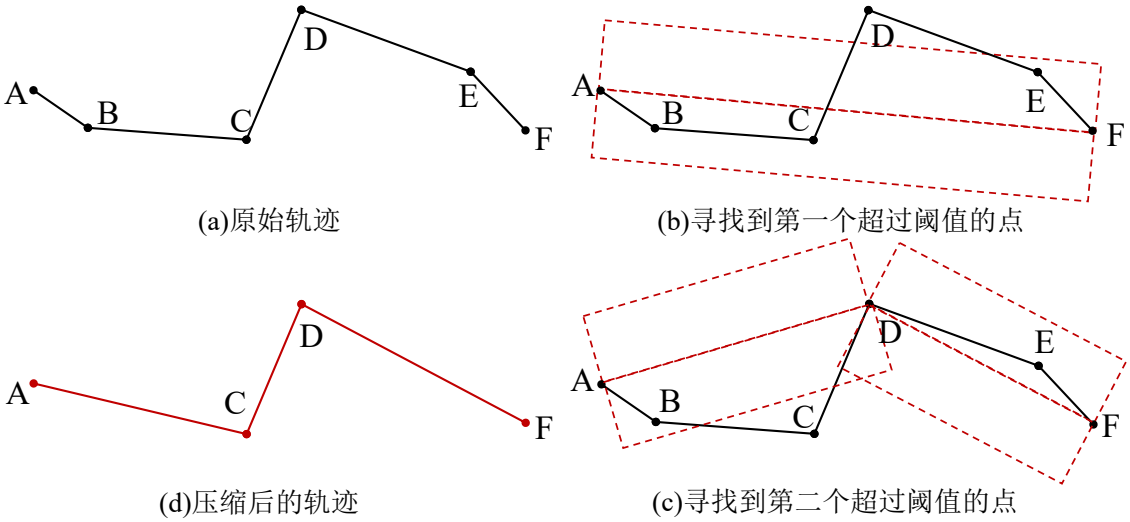

图 2-1 Douglas-Peucker算法示意图

也被考虑了进去了。例如Meratnia等人[25]就在SED的度量上提出了一个从顶向下的时间率（TD-TR）和一个打开窗口的时间率算法。在算法的开始，从第一到第三个采样点被一根线段练了起来，如果SED度量从中间第二个点到这个线段投影的距离随着时间流逝始终小于某一个阈值的话，则算法向前移动一个采样点继续运行；否则，若这个距离超过一个阈值的话，这个使得阈值超过的点成为原起点段的终点，和后一段的起点。

此外，Potamias等人[26]提出了两种算法，分别是Thresholds和STTrace，来对在线的轨迹流进行处理。这两个算法利用了GPS坐标、速度、目前位置的方向来计算出一个安全的区域，以估计下一时刻的所在区域。如果下个位置确实来到了估算中的安全区域，那算法将继续，否则将处理异常来对现有参数进行矫正。Lee等人[27]以及Soares等人[28]用最小描述长度（MDL）来对轨迹压缩的程度进行控制，其基本原理是找寻一个准确度与压缩率的折中。Muckell等人提出了空间质量简化的启发方法（SQUISH）,其基本的原理是将优先度付给那些轨迹流中重要的采样点，将冗余的采样点永久地去除。之后他们又改进了这一方法，取名为了SQUISH-E,让调节压缩率和错误率更为方便[29]。还有一些方法提出了处理在线轨迹的压缩模式。例如，Dead Reckoning算法用现在目标的位置和速度来估计之后目标的位置信息[30]。而Liu等人[31]提出的BQS算法通过计算新点与一个在维护的线段的距离来进行保留点选择。而Lin等人[32]提出了一种只过一遍的策略来计算轨迹中每个点的误差以确定压缩轨迹的点保留情况。

值得一提的是，这些在线算法仅仅是输出一个压缩后的轨迹集合，这任然没有解决原始轨迹杂乱无章的格式的问题，含时间的检索仍然成问题。更重要的是，

这些方法都是将轨迹视为独立的线条组合来考虑的，并没有考虑城市中复杂的交通情况、丰富的重要节点。所以这些方式只能视为是轨迹降采样的方法，并不适合用来建立全局的城市轨迹档案。

2.1.2 语义轨迹压缩

如同Parent等人的研究指出那样，在轨迹上的大多数应用研究分析都要求更多的额外信息了。举个例子，如果要分析解释某个用户的行为动机，那么城市中的额外的信息例如交通地图或者重要的POI点就会很重要。建立在这些额外信息的基础上，则原本时空坐标序列可以被代替为街道的组合，或者重要商店、餐厅的序列。这样，原本普通的原始轨迹就被赋予了丰富的语义信息。

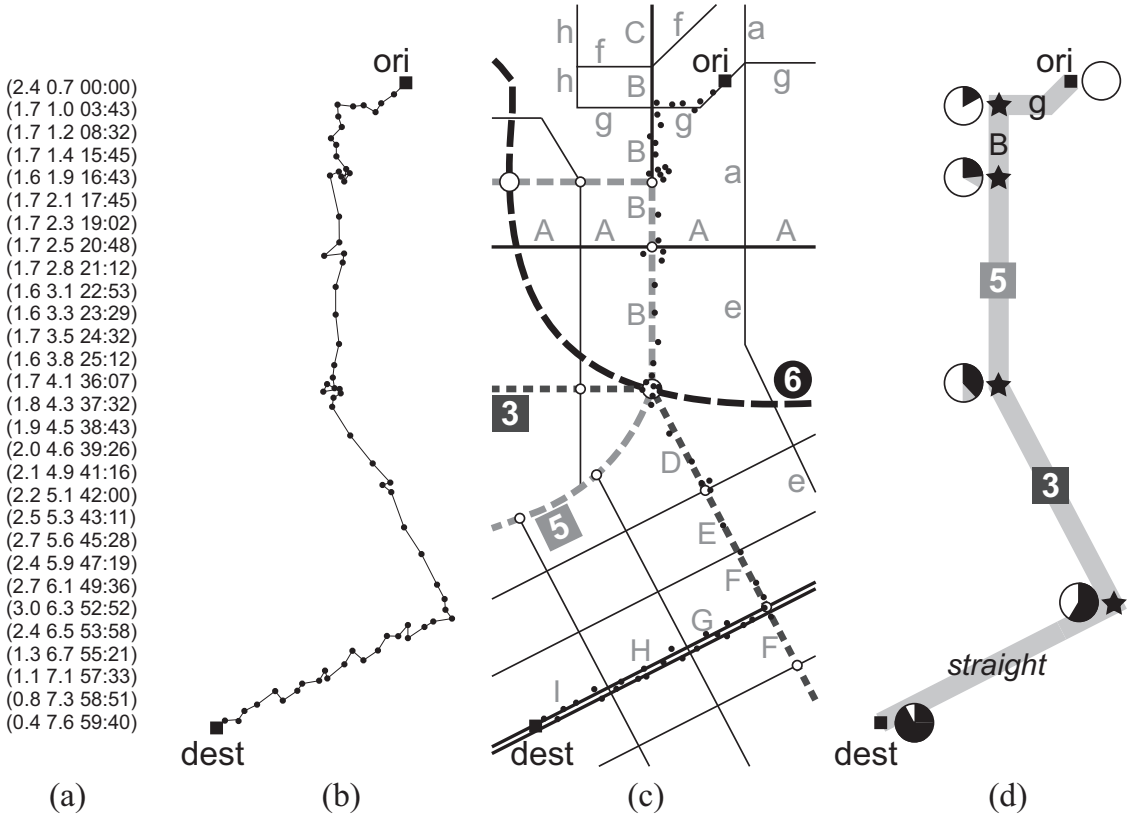

图 2-2 语义轨迹压缩（STC）算法示意图[33]

在这种思路的指引下，很多研究者纷纷加入合适的额外信息。比如Schmid等人[33, 34] 提出了一种语义轨迹压缩（STC）的方式，其利用城市的道路网，将原始轨迹直接代替成为了这个网络上的节点和边，大大简化了轨迹的表示。STC的基本思路可由图2-2表示。其中图2-2(a)是一条原始轨迹序列，(b)中可视化了其在空间中的样貌。在(c)中城市的路网被一起加入可视化，可看到轨迹顺着街道穿

梭。在(d)中，这条原始轨迹被用街道衔接组合了起来，成为一个街道的串联。并且每一个节点的用时信息也被保留下来。

这也是一种全局的表示方法，本文采用的方法就是这种思路。然而，STC这个方法也不是没有缺陷，其完全依赖于城市交通网络，也就是说，当这个网络的信息有缺失或者稀疏，则这种表征的性能将大大收到影响。其次，交通网络只能给出汽车的形势轨迹，其对骑行轨迹以及步行轨迹的刻画也不全面。之后，Liu等人[35]延续这个思路进一步地提出了基于速度和锚点的轨迹简化算法，其目标也是同时将所有轨迹同时压缩简化为固定锚点与速度的表示。此外，还有另外的方法着重考虑轨迹数据中的一些行为模式，例如Chen等人[36]以及Zheng等人[37]利用了轨迹的速度、加速度还有方向改变快慢的信息，将轨迹切割为了行走段和非行走段，并维护了轨迹的骨架信息以及语义信息。在这之后，他们进一步调节了非行走段的部分，使之被更细节的交通方式所刻画，比如自行车、公交车以及自驾。这个过程用了一系列不同的科技，比如监督学习到决策树推理，再比如一个增强准确度刻画的后续处理流程。

此外，一些研究者开始更加注重轨迹的方向信息，强调着轨迹的方向信息里蕴含保留了更多的轨迹结构信息。比如，Song等人[38]提出了一个叫做PRESS的框架来将轨迹分为时间部分和空间部分，并分别针对这两部分提出两种两个算法，一个是混合空间压缩（HSC）算法，另一个是误差限制的时间压缩（BTC）算法。

也有研究者开始强调城市路网的重要性，认为轨迹应该投影到路网上。而这个过程中，路网匹配的技术开始被广泛研究和改进[39–42]，更多的相关信息可以参阅综述[9, 22, 43]。

值得一提的，现有大多数的方法都是假设城市的交通网络和重要语义节点是全面的，没有考虑这些信息部分缺失或者稀疏的情况，比如路网在某些部分没有收录信息。若是碰到这种情况，则上述的算法将失去作用，其性能将大大收到影响。基于这个动机，我们利用提出了一种全新的轨迹聚类压缩算法，使得算法能很好地工作在无论是有路网还是路网确实的环境中。

2.2 语义轨迹表征

为了从轨迹数据中提取有用的知识，最主要的步骤是生成一个良好的轨迹表示。然而，由于轨迹数据的无序性，例如，不同的采样率和实际应用中不同的轨迹长度，这仍然是一项具有挑战性的任务。

2.2.1 轨迹表征方法总结

目前，根据相关的轨迹挖掘场景，主流的轨迹表示主要分为三种：基于轨迹采样点的表征[44–46]，基于轨迹线段的表征[27, 47, 48]和基于轨迹特征的表示[49–51]。

- **基于点的表征**：其总体思想是识别轨迹中的关键点，然后使用这些点来表征这条轨迹。关键点包括驻点和停留点等。Zheng等人[52]在2009年定义在一个给定的时间和空间阈值内的一系列GPS点的中心点是一个驻点，由此将一条由GPS点连成的轨迹划分为一系列的驻点，并根据用户的历史轨迹来衡量用户之间的相似性，接着他们在2011年使用同样的表征方法[45]，实现了用户的旅游推荐。Alvares等人[53]在2007年提出了新的轨迹划分的方式，将轨迹表示为一系列的停留点和连接这些停留点的“运动”。基于点的轨迹表征很简单直观，但是使用少量点表征一整条轨迹能表达的轨迹语义信息是很有限的，轨迹的时间信息也被忽略，因此难以很好的表征轨迹的时空语义信息。
- **基于分段的表征**：其核心思想是将轨迹分割为若干部分，然后使用这些部分来表征这条轨迹。这种方法一般假设每一个轨迹分段在几何形状比较平滑，或者有特殊的语义标注。Chen等人[36]在2008年，根据移动物体位置和速度的区别，将轨迹表征为行走（walk）和非行走（non-walk）。Lee等人[48]在2011年使用了基于最小描述长度的轨迹分割方法，考虑到轨迹在现实世界中受到道路路径的限制，Song等人[54]在2014年提出了一种基于路网匹配的轨迹分段方法，将GPS轨迹映射到路网中并进行压缩存储。基于分段的轨迹表征可以表达一定的轨迹语义信息，例如基于“行走”，“非行走”的分割方法可以反映出轨迹的运动状态的变化，基于路网匹配的轨迹表征方法可以反映轨迹在真实路段上的行走状态，但是它们难以表达轨迹点序列的时间信息，对轨迹整体语义的表达也是很有限的。
- **基于特征的表征**：其希望通过提取轨迹中的某些特征来表征轨迹，例如轨迹的速度、方向、角度、形状，或者轨迹的周期性[49]等。其中Annoni等人[49]将轨迹从原始空间转换到了谱空间来进行表示，将二维的轨迹转换为了一维的表示。Gariel等人[55]用不同的采样率对轨迹进行重新采样，这样就得到了长度相同的轨迹，之后再用主成分分析（PCA）来对重采样后的轨迹进行进一步压缩。还有一些方法利用了目前热门的神经网络，将轨迹喂给LSTM网络来学习轨迹的隐藏特征[50, 56]。这类方法的不足之处在于它大多只能提取到轨迹的地理特征或者几何特征等简单的特征，对轨迹语义的

表达能力依旧是有限的。

值得一提的是，上面提到的主流轨迹表征的关键点主要集中在地理信息上，几乎没有考虑具体的语义知识。因此，在轨迹表征上建立的索引系统会受到多种类型的检索任务的限制，从而产生了大量孤立的研究。例如，没有为多角度查询检索而设计的系统。给定特定的查询需求，例如：（1）检索城市中与给定犯罪区域最为相似的某些可疑区域。（2）通过给予犯罪分子的典型行动路线，从轨迹数据库中找出最可疑的运动轨迹。对于这种复杂的检索任务，用户必须手动将问题分解为单独的部分，然后分别构造查询。下文将描述主流的轨迹相似性度量及检索方法：

2.3 轨迹相似度度量与轨迹检索

2.3.1 轨迹的相似性度量

对轨迹进行表征，一项最基本的概念是定义了轨迹上的相似性度量。目前轨迹上相似性度量的算法很多，如：DTW距离，它允许一些点重复多次以获得最佳对齐[57, 58]。还有最长公共子序列（LCSS）距离[59]和实际序列编辑距离（EDR）距离[60]，他们的原理是消除噪声点引起的影响。Fréchet距离是另一种新颖的曲线之间的相似性度量，其考虑了沿着曲线[61]的点的位置和顺序。在这里，本文简要介绍最具有代表性的DTW距离度量方式。

动态时间规整（Dynamic Time Warping）[57, 62]，简称为DTW，是一种非常经典的比较时间序列的相似性（距离）的度量方式。它的基本想法是为两个时间序列找到“最好的”匹配，为了实现这一点，时间序列在时间轴上被拉长或压缩(”Warping”)[63]，之后两个时间序列上的点再“最合适地”相互匹配。

假设两个时间序列$X=(x_1,x_2,\ldots,x_m)$和$Y=(y_1,y_2,\ldots,y_n)$的长度分别为m和n，时间序列规整的目标是找到X和Y的最佳的匹配路径W，记W为

$$W=(w_1,w_2,\ldots,w_K). \tag{2-1}$$

K为匹配路径W的长度，K满足条件$\max(m,n)<K<m+n-1$；匹配路径W中的每个元素$W_k(0\le k\le K)$表示X中的某个点x_i与Y中的某个点y_j 匹配到了一起，记作$w_k=(i,j)$，W应满足以下约束:

1. **边界条件：**$w_1=(1,1)$以及$w_K=(m,n)$, 即X的第一个点和Y的第一个点要匹配到一起，X的最后一个点和Y的最后一个点要匹配到一起。

2. **匹配顺序**：对于W中的任意两个点$w_k=(i,j)$和$w_{k+1}=(i',j')$，要求$i'\geq i$并且$y'\geq y$。实际上，这个约束限制了X和Y只能线性地(在时间轴上拉长或缩短)匹配。
3. **连续性**：对于W中的任意两个点$w_k=(i,j)$和$w_{k+1}=(i',j')$，要求$i'-i\leq 1$并且$y'-y\leq 1$，这个约束要求X和Y中的每个点都有匹配。

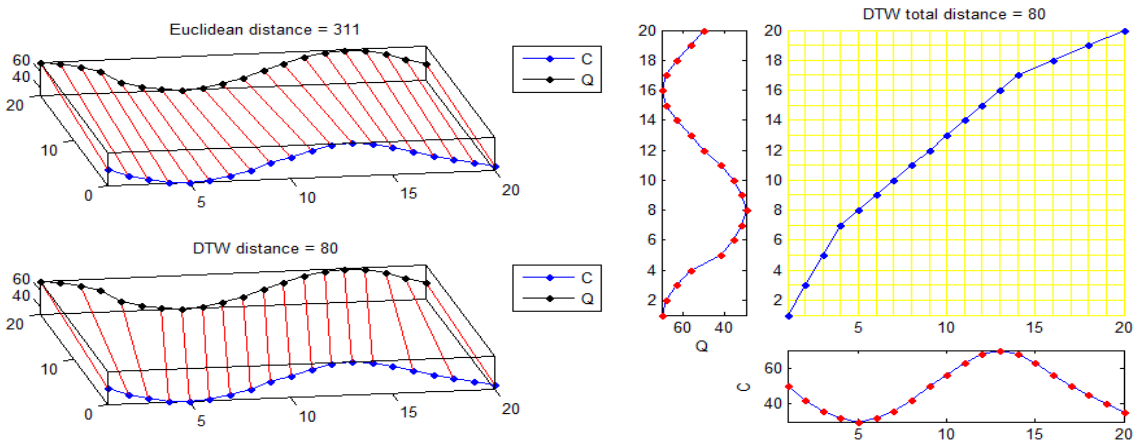

图 2-3 基于DTW的轨迹相似度示意图[64]

满足以上约束的匹配有许多可能，动态时间规整是要找出其“最好”的匹配，也就是使X和Y距离最小的匹配:

$$dtw(X,Y)=\min\left(\sum_{k=1}^{K}d(w_k)\right) \tag{2-2}$$

其中$d(\cdot)$是距离函数，例如$d(w_k)=d(x_i,y_j)=|x_i-y_j|$。在求解两个时间序列的DTW距离时，动态规划技术尝尝被用来加快寻找最佳匹配的速度。图2-3展示了DTW的原理示意图。

2.3.2 轨迹检索

通常，轨迹的查询是启发式的：一般会查询感兴趣的重要节点（POI）或与指定的POI或轨迹最相似的轨迹。最关键的任务是定义相似性度量。几乎所有主流指标都只关注地理相似性。早期研究人员[65]使用两条轨迹所有采样点的总和距离，这要求轨迹具有统一长度，这在现实世界数据中是不现实的。动态时间扭曲（DTW）距离的提出就可以克服这一缺陷。

在轨迹的表示与各种度量的基础上，有研究者提出了许多轨迹索引结构。例如，STR-tree[66]，TB-tree[66]和HR-tree[67]这三种树结构都概括了R-tree这种将空间和时间维度存储在一起的有效的空间数据库。之后，Chakka等人提出SETI[68]来区分时空信息与空间索引系统，以提高检索效率。

然而，几乎所有传统的轨迹表示模型和索引系统都建立在轨迹数据的地理和时间特征上，因此语义检索和挖掘任务难以得到执行。

2.4 词向量表征算法

在NLP(自然语言处理)任务中，我们需要将自然语言交给机器算法来处理，寻找一种将自然语言数学化的表示方法对算法实现非常重要。词向量就是一种将自然语言数学化的一种方式。

最简单的一种词向量是独热向量(one-hot vector)[69]，即是用一个非常长的向量来表示一个词。向量的长度是词典D的大小N，向量的分量只有一个1，其他全是0。1的位置对应该词在词典中的索引。这种词向量的缺点就是容易受到维数灾难的困扰，难以刻画词之间的相似性。

另一种词向量是分布式向量(distributed vector)，也就是本工作使用的向量表示方法。它最早是由Hinton[70]在1986年提出的，可以克服独热向量的上述缺点。它的基本思想是:通过学习训练，将某种语言中的每一个词映射成一个固定长度的向量(向量长度通常比独热向量短)，所有向量构成一个向量空间，每一个向量可以看作是该向量空间中的一个点。在该向量空间上定义“距离”，就可以衡量该向量空间中每个点之间的相似性，也就是对应语言中每一个单词之间的相似性。

2.4.1 word2vec模型

word2vec[71]是Google在2013年开源推出的一个用于获取词向量的工具包，获取到的词向量就是上文提到的分布式向量。通过在向量空间上定义距离计算方法，就可以计算向量之间的距离，也就是相似性。word2vec模型通过在给定语料上进行训练，最终可以实现将词典中的每一个单词映射到同一个向量空间中，定义向量之间的余弦相似性作为向量之间的距离，模型学习结果显示相似的单词之间距离较小，而不相似的单词之间的距离较大。例如“王后”和“国王”的距离比较小，“男孩”，“男人”对应的向量之间距离较近，但是“北京”和“女孩”对应的向量之间的距离较远。

如图2-4是对word2vec模型学习结果的直观展示。将学习到的词向量通过主成分分析(Principle Component Analysis, PCA)降维之后，提取最重要的两个特征进行可视化。可见“国家”类别的单词之间距离较近，而“首都”类型的单词之间距离较近，但是两种类型的单词之间的距离较远。

除此之外，word2vec模型学习到的向量还满足“线性可加性”。即如果按照余弦相似性在全语料中搜索，可以发现，与v(中国)-v(北京)+v(巴黎)距离最近的向量对应的单词是“法国”。

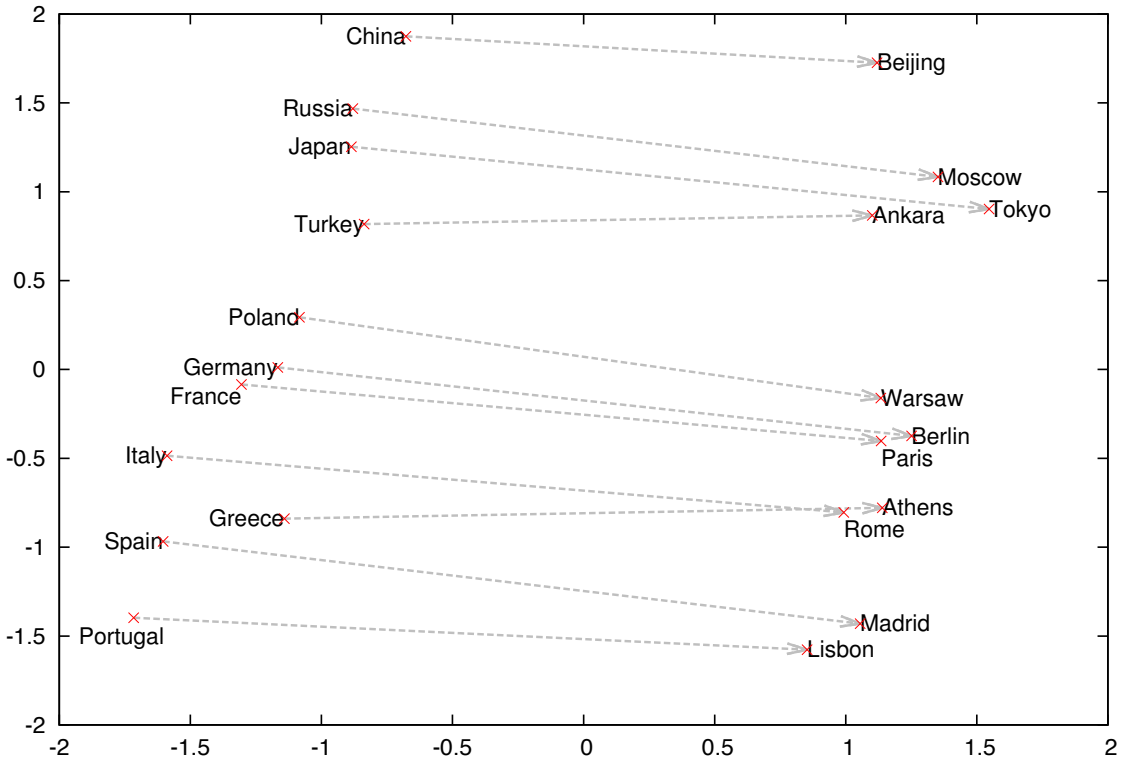

图 2-4 word2vec学习出的词向量的二维展示图[71]

2.5 同步聚类算法

由于我们的方法用了同步聚类算法，所以在这里提出来单独介绍。同步聚类算法（Synchronization-based Clustering，Sync）是根据物理、生物学、化学以及社会科学中提到的同步现象所提出的一种聚类算法。而同步现象最早是有Acebron等人提出来的[72]。

2.5.1 同步现象及应用

试想，在漆黑的丛林中，各种昆虫开始为了生存而进行各项活动。到处都是蝉鸣，却找不到声源的具体来源。此时，如果有一只萤火虫开始若隐若现的闪光，那这闪光马上就不会孤单：一瞬间，练成片的火光反复一下子被这气氛点燃，开始热烈地一起闪耀。这闪光，是由成百上千的萤火虫一起发出的，但它们并不是各自成灯，而是整齐划一的遵循着某种规律一起变亮，一起变暗，反复一堆节日彩灯被看不到的电网练了起来而随着时强时弱的电流周期性的闪耀。这就是同步现象引申出的地方，之后，更多的研究者发现了同步现象存在的其他领域[73–82]。

Kuramoto在1975年提出了Kuramoto模型来对同步现象进行了一个刻画，成为此类现象中最被认同的模型。之后Seliger等人[73]也对Kuramoto模型进行了讨论，并改进提出基于弹簧振子刻画的一般化Kuramoto模型。Arenas等人[74]将Kuramoto模型推广到了网络分析，并研究了网络的拓扑结构和动态时间范围之间的关系。他们的研究对网络的拓扑结构、谱分析、同步动态机制之间的关系进行了联系。在生物学中，Kim等人[83]也利用了同步模型来对细胞进行检测，并从中发现了成组的新基因。与之相似的，Shao等人[81]也改进了同步算法并提出了一种双边聚类算法，使相似的蛋白质以及相似的基因能够能够同时得到发掘。近年来，更多基于同步的算法被提了出来[75, 76, 78, 80, 84]，Kuramoto模型也在这过程中被一步一步被改进。

2.5.2 同步聚类以及优势

在同步算法Sync[75]中，一种基于Kuramoto的数据挖掘聚类被提了出来，在这个工作中，一种基于局部聚类的动态机制是整个工作的核心，而最小描述长度（MDL）则用来决定动态交互的程度。整体来说，这种算法支持高质量的社区发现，其对噪声数据也更加敏感。

基于GPS的轨迹数据通常由一组点表示，由于不同的轨迹通常具有不同的长度和不同的采样率，这并不是一个直接挖掘轨迹数据的好方法。基于此，一种直观的改进方法是将这些点人为分组或者说聚类为具有语义信息的不同区域。但是，传统的聚类方法不适合这项任务。例如，k-means样式聚类算法需要人为给定聚类数k，这是一个依靠经验的步骤，并不能广泛适用于各种情况。其次，k-means算法得到的聚类不是均匀分布的，因此其表示错误不能被误差界阈值ζ界定。与基于k-means的聚类算法相比，同步聚类的优势在于它可以自动生成具有有意义数目的聚类结果。具体而言，簇的数量受到交互范围ϵ的影响：簇的数量将随着ϵ的增加而减少。此外，确定交互范围比提供正确数量的聚类更直观。前者可以控制聚类过程，这促使我们在某种情况下使用ϵ作为误差界限ζ的指示符，即，当大多数点或者几乎所有点都具有小于ζ的代表性误差时，可认为此时误差界阈值ζ就是交互范围ϵ。相反地，如果我们使用k-means作为轨迹压缩的聚类方法，则只能全局地执行聚类，这并不利于控制压缩过程。

聚类模型的另一个分支是基于DBSCAN[85]或OPTICS[86]的搜索。通过寻找密集区域来聚类。然而，这种算法的结果中，聚类簇是任意形状的，这可能违反现实世界区域约束，并可能在我们的压缩情况下产生无意义的聚类簇。例如，城市中的一条主要道路可能被检测为单个聚类簇，这不适合轨迹压缩。相比之下，

基于同步的Sync聚类方法倾向于生成均匀分布的聚类簇，这将使这些有意义的聚类簇成为地图上的压缩点。因此，这些聚类中心可以直观且合理地呈现轨迹压缩结果。

此外，由于所有采样点由本地交互驱动同步，基于同步的聚类算法提供了更直观的压缩轨迹点的方法。与传统的聚类算法不同，基于同步的聚类方法将每个数据对象视为相位振荡器，并模拟对象随时间的动态行为。通过与相似对象的交互，对象的相位逐渐与其相邻对象对齐，从而导致由本地簇结构驱动的非线性对象进行移动。最后，群集中的对象将同步在一起并具有相同的相位。因此，基于同步的集群可以识别由本地数据结构驱动的集群，更重要的是，可以使用同步对象很好地对全局数据结构进行保留。

在这些工作的推动下，在本文中我们从动态角度观察轨迹压缩，并基于同步原理提取多分辨率轨迹数据抽象。据我们所知，本文是第一个将同步概念应用于数据压缩的工作。接下来，本文将叙述轨迹的表征问题。重点在强调目前表征算法的缺陷，然后引入本文的方法。

2.6 本章小结

本章介绍了双边聚类中联合簇的各种类别，

致 谢

大学四年来，我经历了很多，也成长了很多，这个过程少不了很多帮助我的人、改变我的人，借此机会，我要对你们表示我最真诚的感谢。

谢谢我的父亲和母亲，是你们给了我最无微不至的照顾和无条件的支持。当我得意时，是你们分享我的喜悦，并嘱咐我不要骄傲；当我落寞时，是你们鼓励我，让我重整旗鼓。我的每一个重大决定都能得到你们的支持，我取得的每一项成就都离不开你们。如今，我常年不在你们身边，当我一天一天强大，你们却一天一天老去，这是我心头的最痛。你们的恩情，我此生难报，只希望自己变得更强大，有能力保护你们、照顾你们，如同当初你们对我那样。

其次，我必须对我的科研导师邵俊明老师表达我由衷的敬意和谢意。自从大三我跨入教研室，邵老师就成了我最尊敬的人。邵老师以身作则，让我懂得了什么叫做科研，因此我奋斗，我每一天的努力都为缩小自己和邵老师之间的差距。邵老师的人格也让我肃然起敬，可以说作为邵老师的学生，他的高尚、包容与正直能让我们每一个人自惭形秽。我从邵老师身上学到的远远不止做学问，还有做人。我大学最庆幸的一件事情之一就是自己能找到邵老师作为自己的科研导师。

我也要感谢给我上课的每一个老师，你们传授我知识，你们让我看到世界。我忘不了徐全智老师对每一个学生的鞭策与鼓励，忘不了胡建浩老师每节课的“库式论坛”，忘不了每一个含辛茹苦的老师！你们是真正的大师，大学如果没有你们则不能称之为大学。

最后，我也要感谢四年的同学们，我庆幸我们能在一起生活，一起交流学习，同时互相竞争。四年来，我们共同仰望星空，脚踏实地。如今大家各奔东西，祝愿大家都能追到自己的梦想。大学中我最好的朋友们，互相关心互相为对方着想的、能称为兄弟姐妹的各位，我不担心毕业后会失去你们，我相信友谊天长地久。保重，各位，但愿人长久，千里共婵娟。

参考文献

[1] J. Shao. Synchronization Inspired Data Mining[M]. 2011

[2] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth. From data mining to knowledge discovery in databases[J]. AI magazine, 1996, 17(3):37

[3] J. Han, J. Pei, M. Kamber. Data mining: concepts and techniques[M]. Elsevier, 2011

[4] L. Spinsanti, M. Berlingerio, L. Pappalardo. Mobility and geo-social networks[M]. 2013

[5] N. Pelekis, Y. Theodoridis. Mobility data management and exploration[M]. Springer, 2014

[6] J. Mennis, D. Guo. Spatial data mining and geographic knowledge discovery—An introduction[J]. Computers, Environment and Urban Systems, 2009, 33(6):403–408

[7] J. Han, H. J. Miller. Geographic data mining and knowledge discovery[M]. CRC Press, 2009

[8] H. J. Miller. Geographic data mining and knowledge discovery[M]. Blackwell Publishing Malden, MA, 2008

[9] Y. Zheng. Trajectory data mining: an overview[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2015, 6(3):29

[10] R. Cheng, D. V. Kalashnikov, S. Prabhakar. Querying imprecise data in moving object environments[J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(9):1112–1127

[11] T. Emrich, H.-P. Kriegel, N. Mamoulis, et al. Querying uncertain spatio-temporal data[M]. 2012, 354–365

[12] A. Y. Xue, R. Zhang, Y. Zheng, et al. Destination prediction by sub-trajectory synthesis and privacy protection against such prediction[M]. 2013, 254–265

[13] J. Gudmundsson, M. van Kreveld, B. Speckmann. Efficient detection of motion patterns in spatio-temporal data sets[M]. 2004, 250–257

[14] A. Kharrat, I. S. Popa, K. Zeitouni, et al. Clustering algorithm for network constraint trajectories[M]. Springer, 2008, 631–647

[15] H. Cao, N. Mamoulis, D. W. Cheung. Discovery of periodic patterns in spatiotemporal sequences[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(4):453–467

[16] T. Sohn, A. Varshavsky, A. LaMarca, et al. Mobility detection using everyday GSM traces[M]. 2006, 212–224

[17] Y. Zheng, Q. Li, Y. Chen, et al. Understanding mobility based on GPS data[M]. 2008, 312–321

[18] J.-G. Lee, J. Han, X. Li. Trajectory outlier detection: A partition-and-detect framework[M]. 2008, 140–149

[19] W. R. Tobler. A computer movie simulating urban growth in the Detroit region[J]. Economic geography, 1970, 46(sup1):234–240

[20] H. Ma, H. Yang, M. R. Lyu, et al. Sorec: social recommendation using probabilistic matrix factorization[M]. 2008, 931–940

[21] M. Jamali, M. Ester. A matrix factorization technique with trust propagation for recommendation in social networks[M]. 2010, 135–142

[22] D. C. Renso, D. S. Spaccapietra, D. E. Zimnyi. Mobility Data: Modeling, Management, and Understanding[M]. New York, NY, USA: Cambridge University Press, 2013

[23] P. Sun, S. Xia, G. Yuan, et al. An Overview of Moving Object Trajectory Compression Algorithms[J]. Mathematical Problems in Engineering, 2016, 2016

[24] D. H. Douglas, T. K. Peucker. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature[J]. Cartographica: The International Journal for Geographic Information and Geovisualization, 1973, 10(2):112–122

[25] N. Meratnia, A. Rolf. Spatiotemporal compression techniques for moving point objects[M]. 2004, 765–782

[26] M. Potamias, K. Patroumpas, T. Sellis. Sampling trajectory streams with spatiotemporal criteria[M]. 2006, 275–284

[27] J.-G. Lee, J. Han, K.-Y. Whang. Trajectory clustering: a partition-and-group framework[M]. 2007, 593–604

[28] A. Soares Júnior, B. N. Moreno, V. C. Times, et al. GRASP-UTS: an algorithm for unsupervised trajectory segmentation[J]. International Journal of Geographical Information Science, 2015, 29(1):46–68

[29] J. Muckell, P. W. Olsen, J.-H. Hwang, et al. Compression of trajectory data: a comprehensive evaluation and new approach[J]. GeoInformatica, 2014, 18(3):435–460

[30] G. Trajcevski, H. Cao, P. Scheuermanny, et al. On-line data reduction and the quality of history in moving objects databases[M]. 2006, 19–26

[31] J. Liu, K. Zhao, P. Sommer, et al. Bounded quadrant system: Error-bounded trajectory compression on the go[M]. 2015, 987–998

[32] X. Lin, S. Ma, H. Zhang, et al. One-pass error bounded trajectory simplification[J]. Proceedings of the VLDB Endowment, 2017, 10(7):841–852

[33] K.-F. Richter, F. Schmid, P. Laube. Semantic trajectory compression: Representing urban movement in a nutshell[J]. Journal of Spatial Information Science, 2012, 2012(4):3–30

[34] F. Schmid, K.-F. Richter, P. Laube. Semantic trajectory compression[J]. Advances in Spatial and Temporal Databases, 2009:411–416

[35] K. Liu, Y. Li, J. Dai, et al. Compressing large scale urban trajectory data[M]. 2014, 3

[36] Y. Chen, K. Jiang, Y. Zheng, et al. Trajectory simplification method for location-based social networking services[M]. 2009, 33–40

[37] Y. Zheng, Y. Chen, Q. Li, et al. Understanding transportation modes based on GPS data for web applications[J]. ACM Transactions on the Web (TWEB), 2010, 4(1):1

[38] R. Song, W. Sun, B. Zheng, et al. PRESS: A novel framework of trajectory compression in road networks[J]. Proceedings of the VLDB Endowment, 2014, 7(9):661–672

[39] A. Civilis, C. S. Jensen, S. Pakalnis. Techniques for efficient road-network-based tracking of moving objects[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(5):698–712

[40] R. Gotsman, Y. Kanza. A Dilution-matching-encoding compaction of trajectories over road networks[J]. GeoInformatica, 2015, 19(2):331–364

[41] I. S. Popa, K. Zeitouni, V. Oria, et al. Spatio-temporal compression of trajectories in road networks[J]. GeoInformatica, 2015, 19(1):117–145

[42] Y. Dong, D. Pi. Novel Privacy-preserving Algorithm Based on Frequent Path for Trajectory Data Publishing[J]. Knowledge-Based Systems, 2018

[43] J. D. Mazimpaka, S. Timpf. Trajectory data mining: A review of methods and applications[J]. Journal of Spatial Information Science, 2016, 2016(13):61–99

[44] N. J. Yuan, Y. Zheng, L. Zhang, et al. T-finder: A recommender system for finding passengers and vacant taxis[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(10):2390–2403

[45] Y. Zheng, X. Xie. Learning travel recommendations from user-generated GPS traces[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(1):2

[46] A. Anagnostopoulos, R. Atassi, L. Becchetti, et al. Tour recommendation for groups[J]. Data Mining and Knowledge Discovery, 2017, 31(5):1157–1188

[47] R. Bellman. On the approximation of curves by line segments using dynamic programming[J]. Communications of the ACM, 1961, 4(6):284

[48] W.-C. Lee, J. Krumm. Trajectory preprocessing[M]. 2011, 3–33

[49] R. Annoni, C. H. Forster. Analysis of aircraft trajectories using fourier descriptors and kernel density estimation[M]. 2012, 1441–1446

[50] I. S. Ardakani, K. Hashimoto. Encoding bird's trajectory using Recurrent Neural Networks[M]. 2017, 1644–1649

[51] N. Pelekis, P. Tampakis, M. Vodas, et al. On temporal-constrained sub-trajectory cluster analysis[J]. Data Mining and Knowledge Discovery, 2017, 31(5):1294–1330

[52] Y. Zheng, L. Zhang, X. Xie, et al. Mining interesting locations and travel sequences from GPS trajectories[M]. 2009, 791–800

[53] L. O. Alvares, V. Bogorny, B. Kuijpers, et al. A model for enriching trajectories with semantic geographical information[M]. 2007, 22

[54] J. Paefgen, F. Michahelles, T. Staake. GPS trajectory feature extraction for driver risk profiling[M]. 2011, 53–56

[55] M. Gariel, A. N. Srivastava, E. Feron. Trajectory clustering and an application to airspace monitoring[J]. IEEE Transactions on Intelligent Transportation Systems, 2011, 12(4):1511–1524

[56] Q. Gao, F. Zhou, K. Zhang, et al. Identifying human mobility via trajectory embeddings[M]. 2017, 1689–1695

[57] B.-K. Yi, H. Jagadish, C. Faloutsos. Efficient Retrieval of Similar Time Sequences Under Time Warping[M]. 1998

[58] M. Shokoohi-Yekta, B. Hu, H. Jin, et al. Generalizing DTW to the multi-dimensional case requires an adaptive approach[J]. Data mining and knowledge discovery, 2017, 31(1):1–31

[59] M. Vlachos, G. Kollios, D. Gunopulos. Discovering Similar Multidimensional Trajectories[M]. 2002

[60] L. Chen, M. T. Özsu, V. Oria. Robust and Fast Similarity Search for Moving Object Trajectories[M]. 2005

[61] T. Eiter, H. Mannila. Computing Discrete Fréchet Distance[R]

[62] P. Senin. Dynamic time warping algorithm review[J]. Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA, 2008, 855:1–23

[63] S. Salvador, P. Chan. Toward accurate dynamic time warping in linear time and space[J]. Intelligent Data Analysis, 2007, 11(5):561–580

[64] P. Chen, J. Gu, D. Zhu, et al. A dynamic time warping based algorithm for trajectory matching in LBS[J]. International Journal of Database Theory and Application, 2013, 6(3):39–48

[65] R. Agrawal, C. Faloutsos, A. Swami. Efficient similarity search in sequence databases[J]. Foundations of data organization and algorithms, 1993:69–84

[66] D. Pfoser, C. S. Jensen, Y. Theodoridis, et al. Novel approaches to the indexing of moving object trajectories.[M]. 2000, 395–406

[67] M. A. Nascimento, J. R. Silva. Towards historical R-trees[M]. 1998, 235–240

[68] V. P. Chakka, A. C. Everspaugh, J. M. Patel. Indexing Large Trajectory Data Sets with SETI[M]. 2003

[69] J. Turian, L. Ratinov, Y. Bengio. Word representations: a simple and general method for semi-supervised learning[M]. 2010, 384–394

[70] G. E. Hinton. Distributed representations[J]. 1984

[71] T. Mikolov, I. Sutskever, K. Chen, et al. Distributed representations of words and phrases and their compositionality[M]. 2013, 3111–3119

[72] J. A. Acebrón, L. L. Bonilla, C. J. P. Vicente, et al. The Kuramoto model: A simple paradigm for synchronization phenomena[J]. Reviews of modern physics, 2005, 77(1):137

[73] P. Seliger, S. C. Young, L. S. Tsimring. Plasticity and learning in a network of coupled phase oscillators[J]. Physical Review E, 2002, 65(4):041906

[74] A. Arenas, A. Diaz-Guilera, C. J. Pérez-Vicente. Synchronization reveals topological scales in complex networks[J]. Physical review letters, 2006, 96(11):114102

[75] J. Shao, X. He, C. Böhm, et al. Synchronization-Inspired Partitioning and Hierarchical Clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(4):893–905

[76] J. Shao. Synchronization on data mining: a universal concept for knowledge discovery[J]. LAP LAMBERT Academic Publishing, Saarbrücken, 2012

[77] J. Shao, Z. Han, Q. Yang, et al. Community detection based on distance dynamics[M]. 2015, 1075–1084

[78] W. Ying, F.-L. Chung, S. Wang. Scaling up synchronization-inspired partitioning clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(8):2045–2057

[79] J. Shao, Q. Yang, H.-V. Dang, et al. Scalable clustering by iterative partitioning and point attractor representation[J]. ACM Transactions on Knowledge Discovery from Data, 2016, 11(1):5

[80] J. Shao, F. Huang, Q. Yang, et al. Robust Prototype-based Learning on Data Streams[J]. IEEE Transactions on Knowledge and Data Engineering, 2017

[81] J. Shao, C. Gao, W. Zeng, et al. Synchronization-inspired Co-clustering and Its Application to Gene Expression Data[M]. 2017

[82] J. Shao, X. Wang, Q. Yang, et al. Synchronization-based scalable subspace clustering of high-dimensional data[J]. Knowledge and Information Systems, 2017, 52(1):83–111

[83] C. S. Kim, C. S. Bae, H. J. Tcha. A phase synchronization clustering algorithm for identifying interesting groups of genes from cell cycle expression data[J]. BMC bioinformatics, 2008, 9(1):56

[84] J. Shao, C. Böhm, Q. Yang, et al. Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part III[M]. Springer Berlin Heidelberg, 2010

[85] M. Ester, H.-P. Kriegel, J. Sander, et al. A density-based algorithm for discovering clusters in large spatial databases with noise.[M]. 1996, 226–231

[86] M. Ankerst, M. M. Breunig, H.-P. Kriegel, et al. OPTICS: ordering points to identify the clustering structure[M]. 1999, 49–60

[87] R. Agrawal, R. Srikant, et al. Fast algorithms for mining association rules[M]. Proc. 20th int. conf. very large data bases, VLDB, 1994, 487–499

[88] J. Bao, Y. Zheng, D. Wilkie, et al. Recommendations in location-based social networks: a survey[J]. GeoInformatica, 2015, 19(3):525–565

[89] Y. Zhang, X. Chen. Explainable Recommendation: A Survey and New Perspectives[J]. arXiv preprint arXiv:1804.11192, 2018

[90] C.-Y. Tsai, B.-H. Lai. A location-item-time sequential pattern mining algorithm for route recommendation[J]. Knowledge-Based Systems, 2015, 73:97–110

[91] Y. Si, F. Zhang, W. Liu. CTF-ARA: An adaptive method for POI recommendation based on check-in and temporal features[J]. Knowledge-Based Systems, 2017, 128:59–70

[92] M. Lv, L. Chen, Y. Shen, et al. Measuring cell-id trajectory similarity for mobile phone route classification[J]. Knowledge-Based Systems, 2015, 89:181–191

[93] L. Wang, K. Hu, T. Ku, et al. Mining frequent trajectory pattern based on vague space partition[J]. Knowledge-based systems, 2013, 50:100–111

[94] J. A. Acebrón, L. L. Bonilla, C. J. P. Vicente, et al. The Kuramoto model: A simple paradigm for synchronization phenomena[J]. Reviews of modern physics, 2005, 77(1):137

[95] A. Arenas, A. Diaz-Guilera, C. J. Pérez-Vicente. Synchronization reveals topological scales in complex networks[J]. Physical review letters, 2006, 96(11):114102

[96] C. S. Kim, C. S. Bae, H. J. Tcha. A phase synchronization clustering algorithm for identifying interesting groups of genes from cell cycle expression data[J]. BMC bioinformatics, 2008, 9(1):56

[97] Y. Kuramoto. Chemical oscillations, waves, and turbulence[M]. Springer Science & Business Media, 2012

[98] J. Shao. Synchronization on data mining: a universal concept for knowledge discovery[J]. LAP LAMBERT Academic Publishing, Saarbrücken, 2012

[99] W. Ying, F.-L. Chung, S. Wang. Scaling up synchronization-inspired partitioning clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(8):2045–2057

[100] J.-G. Lee, J. Han, X. Li, et al. TraClass: trajectory classification using hierarchical region-based and trajectory-based clustering[J]. Proceedings of the VLDB Endowment, 2008, 1(1):1081–1094

[101] Y. Tao, D. Papadias. The mv3r-tree: A spatio-temporal access method for timestamp and interval queries[M]. 2001

[102] E. Frentzos, K. Gratsias, N. Pelekis, et al. Algorithms for nearest neighbor search on moving object trajectories[J]. Geoinformatica, 2007, 11(2):159–193

[103] Z. Chen, H. T. Shen, X. Zhou, et al. Searching trajectories by locations: an efficiency study[M]. 2010, 255–266

[104] K. Deng, K. Xie, K. Zheng, et al. Trajectory indexing and retrieval[M]. Springer, 2011, 35–60

[105] J. H. Friedman, J. L. Bentley, R. A. Finkel. An algorithm for finding best matches in logarithmic expected time[J]. ACM Transactions on Mathematical Software (TOMS), 1977, 3(3):209–226

[106] C.-C. Hung, W.-C. Peng, W.-C. Lee. Clustering and aggregating clues of trajectories for mining trajectory patterns and routes[J]. The VLDB Journal—The International Journal on Very Large Data Bases, 2015, 24(2):169–192

[107] C. Long, R. C.-W. Wong, H. Jagadish. Direction-preserving trajectory simplification[J]. Proceedings of the VLDB Endowment, 2013, 6(10):949–960

[108] C. Long, R. C.-W. Wong, H. Jagadish. Trajectory simplification: on minimizing the direction-based error[J]. Proceedings of the VLDB Endowment, 2014, 8(1):49–60

[109] P. D. Grünwald, I. J. Myung, M. A. Pitt. Advances in minimum description length: Theory and applications[M]. MIT press, 2005

[110] J. Muckell, J.-H. Hwang, V. Patil, et al. SQUISH: an online approach for GPS trajectory compression[M]. 2011, 13

[111] L. E. Kavraki, P. Svestka, J.-C. Latombe, et al. Probabilistic Roadmaps for Path Planning in High-dimensional Configuration Spaces[M]. 1996

[112] J. Yuan, Y. Zheng, X. Xie, et al. Driving with knowledge from the physical world[M]. 2011, 316–324

[113] J. Yuan, Y. Zheng, C. Zhang, et al. T-drive: driving directions based on taxi trajectories[M]. 2010, 99–108

[114] Y. Zheng, X. Xie, W.-Y. Ma. Geolife: A collaborative social networking service among user, location and trajectory.[J]. IEEE Data Eng. Bull., 2010, 33(2):32–39

[115] A. Shaker, E. Hüllermeier. IBLStreams: a system for instance-based classification and regression on data streams[J]. Evolving Systems, 2012, 3(4):235–249

[116] H. Wang, W. Fan, P. S. Yu, et al. Mining concept-drifting data streams using ensemble classifiers[M]. 2003, 226–235

[117] F. Cao, M. Estert, W. Qian, et al. Density-based clustering over an evolving data stream with noise[M]. 2006, 328–339

[118] G. Castelli, M. Mamei, A. Rosi. The whereabouts diary[M]. 2007, 175–192

[119] L. Liao, D. J. Patterson, D. Fox, et al. Learning and inferring transportation routines[J]. Artificial Intelligence, 2007, 171(5-6):311–331

[120] L. Liao, D. Fox, H. Kautz. Extracting places and activities from gps traces using hierarchical conditional random fields[J]. The International Journal of Robotics Research, 2007, 26(1):119–134

[121] J. Shao, X. Wang, Q. Yang, et al. Synchronization-based scalable subspace clustering of high-dimensional data[J]. Knowledge and Information Systems, 2017, 52(1):83–111

[122] R. Gotsman, Y. Kanza. Compact representation of GPS trajectories over vectorial road networks[M]. 2013, 241–258

[123] J. Jiang, C. Xu, J. Xu, et al. Route planning for locations based on trajectory segments[M]. 2016, 6

[124] J. Shao, X. He, C. Böhm, et al. Synchronization-Inspired Partitioning and Hierarchical Clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(4):893–905

[125] D. Aeyels, F. De Smet. A mathematical model for the dynamics of clustering[J]. Physica D: Nonlinear Phenomena, 2008, 237(19):2517–2530

[126] C. S. Kim, C. S. Bae, H. J. Tcha. A phase synchronization clustering algorithm for identifying interesting groups of genes from cell cycle expression data[J]. BMC Bioinformatics, 2008, 9

[127] S. Ando, T. Thanomphongphan, Y. Seki, et al. Ensemble anomaly detection from multi-resolution trajectory features[J]. Data Mining and Knowledge Discovery, 2015, 29(1):39–83

[128] A. Eckstein. Automated flight track taxonomy for measuring benefits from performance based navigation[M]. 2009, 1–12

[129] A. Naftel, S. Khalid. Motion trajectory learning in the DFT-coefficient feature space[M]. 2006, 47–47

[130] W. Blacoe, M. Lapata. A comparison of vector-based representations for semantic composition[M]. 2012, 546–556

[131] S. Feng, G. Cong, B. An, et al. POI2Vec: Geographical Latent Representation for Predicting Future Visitors[M]. 2017, 102–108

[132] S. Feng, X. Li, Y. Zeng, et al. Personalized Ranking Metric Embedding for Next New POI Recommendation[M]. 2015, 2069–2075

[133] P. Yanardag, S. Vishwanathan. Deep graph kernels[M]. 2015, 1365–1374

[134] A. Grover, J. Leskovec. node2vec: Scalable feature learning for networks[M]. 2016, 855–864

[135] J. Wieting, M. Bansal, K. Gimpel, et al. Towards universal paraphrastic sentence embeddings[J]. arXiv preprint arXiv:1511.08198, 2015

[136] Y. Adi, E. Kermany, Y. Belinkov, et al. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks[J]. arXiv preprint arXiv:1608.04207, 2016

[137] T. Mikolov, K. Chen, G. Corrado, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013

[138] J. J.-C. Ying, W.-C. Lee, T.-C. Weng, et al. Semantic Trajectory Mining for Location Prediction[M]. 2011, 34–43

[139] Z. Yan, D. Chakraborty, C. Parent, et al. Semantic Trajectories: Mobility Data Computation and Annotation[J]. ACM Transactions on Intelligent Systems and Technology, 2013, 4(3):49:1–49:38

[140] T. Eiter, H. Mannila. Computing discrete Fréchet distance[R]. CD-TR 94/64, Information Systems Department, Technical University of Vienna

[141] M. Paterson, V. Dančík. Longest common subsequences[J]. Mathematical Foundations of Computer Science 1994, 1994:127–142

[142] L. Chen, R. Ng. On the marriage of lp-norms and edit distance[M]. 2004, 792–803

[143] D. J. Berndt, J. Clifford. Using dynamic time warping to find patterns in time series.[M]. 1994, 359–370

[144] E. Cho, S. A. Myers, J. Leskovec. Friendship and mobility: user movement in location-based social networks[M]. 2011, 1082–1090

[145] G. Yuan, P. Sun, J. Zhao, et al. A review of moving object trajectory clustering algorithms[J]. Artificial Intelligence Review, 2017, 47(1):123–144

[146] D. Phan, L. Xiao, R. Yeh, et al. Flow map layout[M]. 2005, 219–224

[147] S. Gupta, A. Campa, S. Ruffo. Kuramoto model of synchronization: equilibrium and nonequilibrium aspects[J]. Journal of Statistical Mechanics: Theory and Experiment, 2014, 2014(8):R08001

[148] Y. Kuramoto. Self-entrainment of a population of coupled non-linear oscillators[M]. 1975, 420–422

[149] Y. Kuramoto. Self-entrainment of a population of coupled non-linear oscillators[M]. 1975, 420–422

[150] A. Arenas, A. Díaz-Guilera, J. Kurths, et al. Synchronization in complex networks[J]. Physics reports, 2008, 469(3):93–153

[151] Y. Liu, C. Liu, B. Liu, et al. Unified Point-of-Interest Recommendation with Temporal Interval Assessment[M]. 2016, 1015–1024

[152] Y. Wang, Y. Zheng, Y. Xue. Travel time estimation of a path using sparse trajectories[M]. 2014, 25–34

[153] L. Wang, Y. Zheng, X. Xie, et al. A flexible spatio-temporal indexing scheme for large-scale GPS track retrieval[M]. 2008, 1–8

[154] A. Guttman. R-trees: a dynamic index structure for spatial searching[M]. ACM, 1984

[155] Y. Wang, N. J. Yuan, D. Lian, et al. Regularity and conformity: Location prediction using heterogeneous mobility data[M]. 2015, 1275–1284

[156] B. Zheng, N. J. Yuan, K. Zheng, et al. Approximate keyword search in semantic trajectory database[M]. 2015, 975–986

[157] N. J. Yuan, Y. Zheng, X. Xie, et al. Discovering urban functional zones using latent activity trajectories[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(3):712–725

[158] J. Shang, Y. Zheng, W. Tong, et al. Inferring gas consumption and pollution emission of vehicles throughout a city[M]. 2014, 1027–1036

[159] L.-Y. Wei, Y. Zheng, W.-C. Peng. Constructing popular routes from uncertain trajectories[M]. 2012, 195–203

[160] C. Zhou, D. Frankowski, P. Ludford, et al. Discovering personally meaningful places: An interactive clustering approach[J]. ACM Transactions on Information Systems (TOIS), 2007, 25(3):12

[161] D. Ashbrook, T. Starner. Learning significant locations and predicting user movement with GPS[M]. 2002, 101–108

[162] D. Ashbrook, T. Starner. Using GPS to learn significant locations and predict movement across multiple users[J]. Personal and Ubiquitous computing, 2003, 7(5):275–286

[163] Z. Chen, H. T. Shen, X. Zhou. Discovering popular routes from trajectories[M]. 2011, 900–911

[164] H. Jeung, M. L. Yiu, X. Zhou, et al. Discovery of convoys in trajectory databases[J]. Proceedings of the VLDB Endowment, 2008, 1(1):1068–1080

[165] Z. Li, B. Ding, J. Han, et al. Mining periodic behaviors for moving objects[M]. 2010, 1099–1108

[166] X. Xiao, Y. Zheng, Q. Luo, et al. Inferring social ties between users with human location history[J]. Journal of Ambient Intelligence and Humanized Computing, 2014, 5(1):3–19

[167] Y. Ye, Y. Zheng, Y. Chen, et al. Mining individual life pattern based on location history[M]. 2009, 1–10

[168] Y. Zheng, L. Zhang, Z. Ma, et al. Recommending friends and locations based on individual location history[J]. ACM Transactions on the Web, 2011, 5(1):5

[169] Z. Li, J.-G. Lee, X. Li, et al. Incremental clustering for trajectories[M]. 2010, 32–46

[170] J. Yin, X. Chai, Q. Yang. High-level goal recognition in a wireless LAN[M]. 2004, 578–584

[171] D. J. Patterson, L. Liao, D. Fox, et al. Inferring high-level behavior from low-level sensors[M]. 2003, 73–89

[172] N. Adrienko, G. Adrienko. Spatial generalization and aggregation of massive movement data[J]. IEEE Transactions on Visualization and Computer Graphics, 2011, 17(2):205–219

[173] G. Kellaris, N. Pelekis, Y. Theodoridis. Trajectory compression under network constraints[M]. 2009, 392–398

[174] C. Parent, S. Spaccapietra, C. Renso, et al. Semantic trajectories modeling and analysis[J]. ACM Computing Surveys, 2013, 45(4):42

[175] Y. Zheng, L. Liu, L. Wang, et al. Learning transportation mode from raw gps data for geographic applications on the web[M]. 2008, 247–256

[176] L. O. Alvares, G. Oliveira, C. A. Heuser, et al. A Framework for Trajectory Data Preprocessing for Data Mining.[M]. 2009, 698–702

[177] S. Spaccapietra, C. Parent, M. L. Damiani, et al. A conceptual view on trajectories[J]. Data & Knowledge Engineering, 2008, 65(1):126–146

[178] A. T. Palma, V. Bogorny, B. Kuijpers, et al. A clustering-based approach for discovering interesting places in trajectories[M]. 2008, 863–868

[179] G. Hu, J. Shao, F. Liu, et al. IF-Matching: Towards Accurate Map-Matching with Information Fusion[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(1):114–127

[180] Q. Li, Y. Zheng, X. Xie, et al. Mining user similarity based on location history[M]. 2008, 34

[181] D. Sankoff, J. B. Kruskal. Time warps, string edits, and macromolecules: the theory and practice of sequence comparison[J]. Reading: Addison-Wesley Publication, 1983, edited by Sankoff, David; Kruskal, Joseph B., 1983, 1

[182] F. Giannotti, M. Nanni, F. Pinelli, et al. Trajectory pattern mining[M]. 2007, 330–339

[183] L. Chen, M. T. Özsu, V. Oria. Robust and fast similarity search for moving object trajectories[M]. 2005, 491–502

[184] C. Böhm, C. Plant, J. Shao, et al. Clustering by synchronization[M]. 2010, 583–592

[185] J. Shao, Z. Ahmadi, S. Kramer. Prototype-based learning on concept-drifting data streams[M]. 2014, 412–421

[186] S. Maurus, C. Plant. Skinny-dip: Clustering in a Sea of Noise[M]. 2016, 1055–1064

[187] X. Li, J. Han, J.-G. Lee, et al. Traffic Density-based Discovery of Hot Routes in Road Networks[M]. Berlin, Heidelberg: Springer-Verlag, 2007, 441–459

[188] C. Böhm, C. Plant, J. Shao, et al. Clustering by Synchronization[M]. New York, NY, USA: ACM, 2010, 583–592

[189] V. Bogorny, B. Kuijpers, L. O. Alvares. ST-DMQL: A Semantic Trajectory Data Mining Query Language[M]. 2009

[190] L. O. Alvares, V. Bogorny, B. Kuijpers, et al. Towards Semantic Trajectory Knowledge Discovery[M]. 2007

[191] K. S. Tai, R. Socher, C. D. Manning. Improved Semantic Representations from Tree-structured Long Short-term Memory Networks[M]. 2015

[192] N. Kalchbrenner, E. Grefenstette, P. Blunsom. A Convolutional Neural Network for Modelling Sentences[M]. 2014

[193] R. Socher, E. H. Huang, J. Pennin, et al. Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection[M]. 2011

[194] Q. Le, T. Mikolov. Distributed representations of sentences and documents[M]. 2014

[195] M. Iyyer, V. Manjunatha, J. Boyd-Graber, et al. Deep Unordered Composition Rivals Syntactic Methods for Text Classification[M]. 2015

[196] J. Mitchell, M. Lapata. Composition in Distributional Models of Semantics[M]. 2010

[197] R. Collobert, J. Weston, L. Bottou, et al. Natural Language Processing (Almost) from Scratch[M]. 2011

[198] Y. Bengio, R. Ducharme, P. Vincent, et al. A Neural Probabilistic Language Model[M]. 2003

[199] J. Pennington, R. Socher, C. Manning. Glove: Global Vectors for Word Representation[M]. 2014

[200] S. Arora, Y. Liang, T. Ma. A Simple but Tough-to-beat Baseline for Sentence Embeddings[M]. 2017

[201] H. Cho, I. S. Dhillon. Coclustering of human cancer microarrays using minimum sum-squared residue coclustering[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 2008, 5(3):385–400

[202] I. S. Dhillon, S. Mallela, D. S. Modha. Information-theoretic co-clustering[M]. Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, 2003, 89–98

[203] R. Agrawal, T. Imieliński, A. Swami. Mining association rules between sets of items in large databases[J]. ACM SIGMOD Record, 1993, 22(2):207–216

[204] C. Huygens. Horologium oscillatorium: 1673[M]. Dawson, 1966

[205] B. Frisch, N. Koeniger. Social synchronization of the activity rhythms of honeybees within a colony[J]. Behavioral ecology and sociobiology, 1994, 35(2):91–98

[206] P. Carmona-Saez, R. D. Pascual-Marqui, F. Tirado, et al. Biclustering of gene expression data by non-smooth non-negative matrix factorization[J]. BMC bioinformatics, 2006, 7(1):1

[207] Y. Kluger, R. Basri, J. T. Chang, et al. Spectral biclustering of microarray data: coclustering genes and conditions[J]. Genome research, 2003, 13(4):703–716

[208] Q. Sheng, Y. Moreau, B. De Moor. Biclustering microarray data by Gibbs sampling[J]. Bioinformatics, 2003, 19(suppl 2):ii196–ii205

[209] L. Lazzeroni, A. Owen. Plaid models for gene expression data[J]. Statistica sinica, 2002:61–86

[210] G. Li, Q. Ma, H. Tang, et al. QUBIC: a qualitative biclustering algorithm for analyses of gene expression data[J]. Nucleic acids research, 2009:gkp491

[211] J. Shi, J. Malik. Normalized cuts and image segmentation[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2000, 22(8):888–905

[212] W.-H. Yang, D.-Q. Dai, H. Yan. Finding correlated biclusters from gene expression data[J]. Knowledge and Data Engineering, IEEE Transactions on, 2011, 23(4):568–584

[213] C. Cano, L. Adarve, J. López, et al. Possibilistic approach for biclustering microarray data[J]. Computers in biology and medicine, 2007, 37(10):1426–1436

[214] G. P. Coelho, F. O. de França, F. J. Von Zuben. Multi-objective biclustering: When non-dominated solutions are not enough[J]. Journal of Mathematical Modelling and Algorithms, 2009, 8(2):175–202

[215] L. N. De Castro, J. Timmis. Artificial immune systems: a new computational intelligence approach[M]. Springer Science & Business Media, 2002

[216] J. Kennedy. Particle swarm optimization[M]. Springer, 2011, 760–766

[217] J. Liu, Z. Li, X. Hu, et al. Biclustering of microarray data with MOSPO based on crowding distance[J]. BMC bioinformatics, 2009, 10(4):1

[218] A. Das, B. K. Chakrabarti. Quantum annealing and related optimization methods[M]. Springer Science & Business Media, 2005

[219] K. Bryan, P. Cunningham, N. Bolshakova. Application of simulated annealing to the biclustering of gene expression data[J]. Information Technology in Biomedicine, IEEE Transactions on, 2006, 10(3):519–525

[220] J. Yang, H. Wang, W. Wang, et al. An improved biclustering method for analyzing gene expression profiles[J]. International Journal on Artificial Intelligence Tools, 2005, 14(05):771–789

[221] K. Y. Yip, D. W. Cheung, M. K. Ng. Harp: A practical projected clustering algorithm[J]. Knowledge and Data Engineering, IEEE Transactions on, 2004, 16(11):1387–1397

[222] A. Mukhopadhyay, U. Maulik, S. Bandyopadhyay. A novel coherence measure for discovering scaling biclusters from gene expression data[J]. Journal of Bioinformatics and Computational Biology, 2009, 7(05):853–868

[223] J. S. Aguilar-Ruiz. Shifting and scaling patterns from gene expression data[J]. Bioinformatics, 2005, 21(20):3840–3845

[224] H.-P. Kriegel, P. Kröger, A. Zimek. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2009, 3(1):1

[225] B. Pontes, R. Giráldez, J. S. Aguilar-Ruiz. Biclustering on expression data: A review[J]. Journal of biomedical informatics, 2015, 57:163–180

[226] Y. Cheng, G. M. Church. Biclustering of expression data.[M]. Ismb, 2000, 93–103

[227] J. A. Hartigan. Direct clustering of a data matrix[J]. Journal of the american statistical association, 1972, 67(337):123–129

[228] A. Tanay, R. Sharan, R. Shamir. Discovering statistically significant biclusters in gene expression data[J]. Bioinformatics, 2002, 18(suppl 1):S136–S144

[229] A. Tanay, R. Sharan, R. Shamir. Biclustering algorithms: A survey[J]. Handbook of computational molecular biology, 2005, 9(1-20):122–124

[230] V. Estivill-Castro. Why so many clustering algorithms: a position paper[J]. ACM SIGKDD explorations newsletter, 2002, 4(1):65–75

[231] C. Cortes, V. Vapnik. Support-vector networks[J]. Machine learning, 1995, 20(3):273–297

[232] J. R. Quinlan. Induction of decision trees[J]. Machine learning, 1986, 1(1):81–106

[233] L. E. Peterson. K-nearest neighbor[J]. Scholarpedia, 2009, 4(2):1883

[234] J. Han, J. Pei, Y. Yin. Mining frequent patterns without candidate generation[M]. ACM Sigmod Record, 2000, 1–12

[235] A. Y. Ng, M. I. Jordan, Y. Weiss, et al. On spectral clustering: Analysis and an algorithm[J]. Advances in neural information processing systems, 2002, 2:849–856

[236] J. A. Hartigan, M. A. Wong. Algorithm AS 136: A k-means clustering algorithm[J]. Journal of the Royal Statistical Society. Series C (Applied Statistics), 1979, 28(1):100–108

附录 A 人工数据集上的CoSync运行结果

以下是本次在人工数据集上进行的部分结果，每一幅图的左边为原始数据矩阵，每一个色块的轮廓代表一个联合簇；右边为CoSync完成联合后的结果。

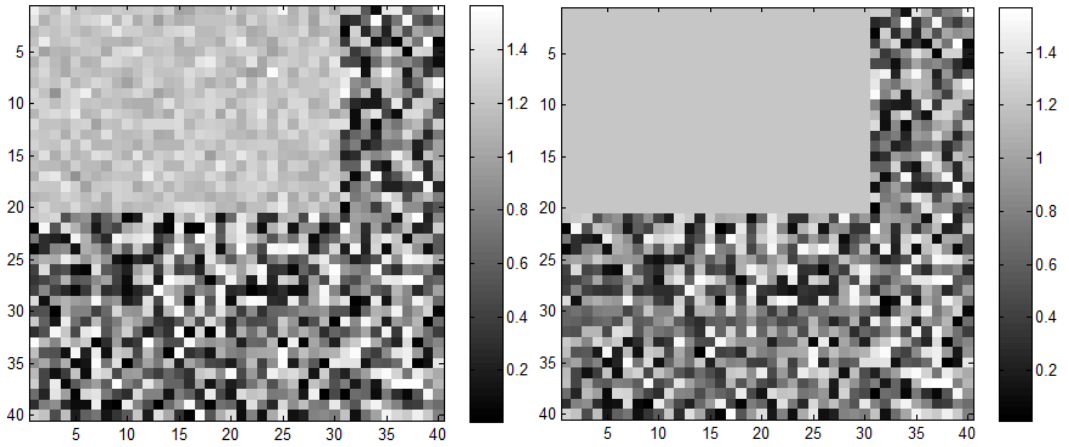

图 A-1 单个联合簇

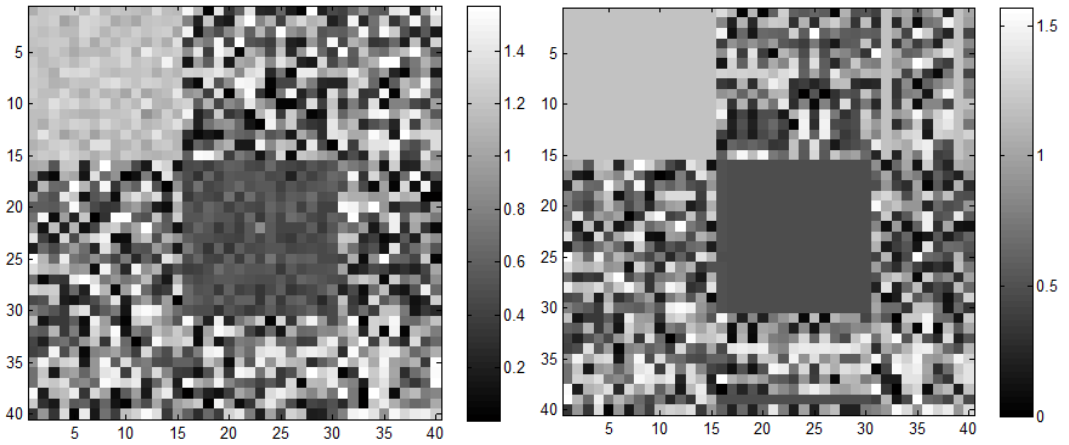

图 A-2 双联合簇，对角分布

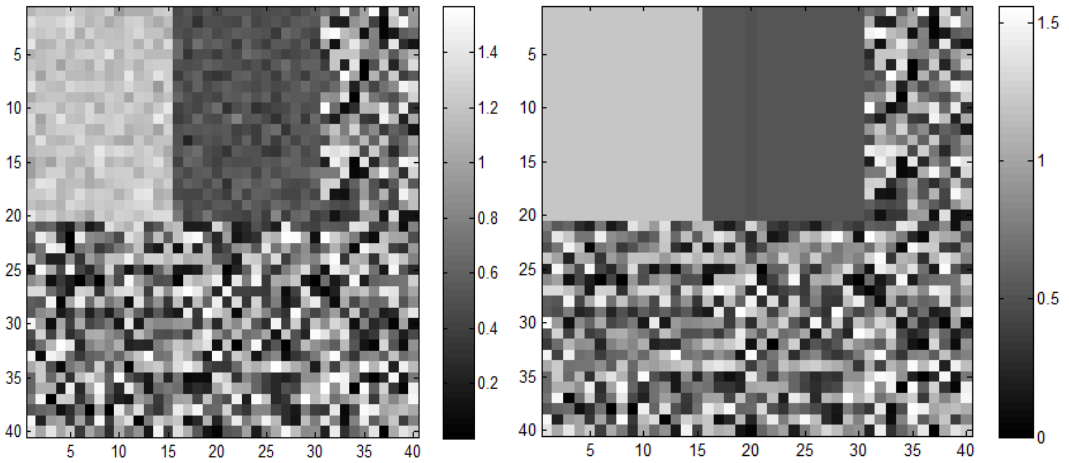

图 A-3 双联合簇，并排分布

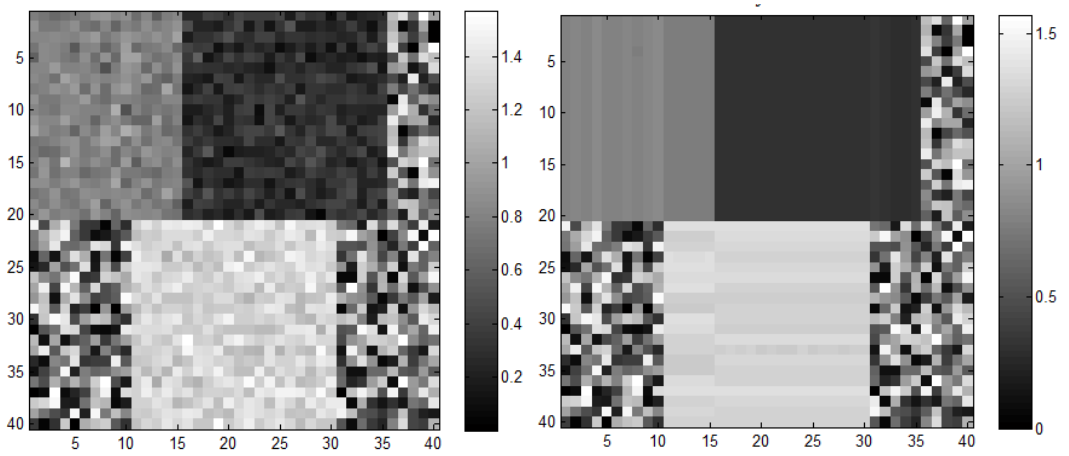

图 A-4 三联合簇，不规则分布

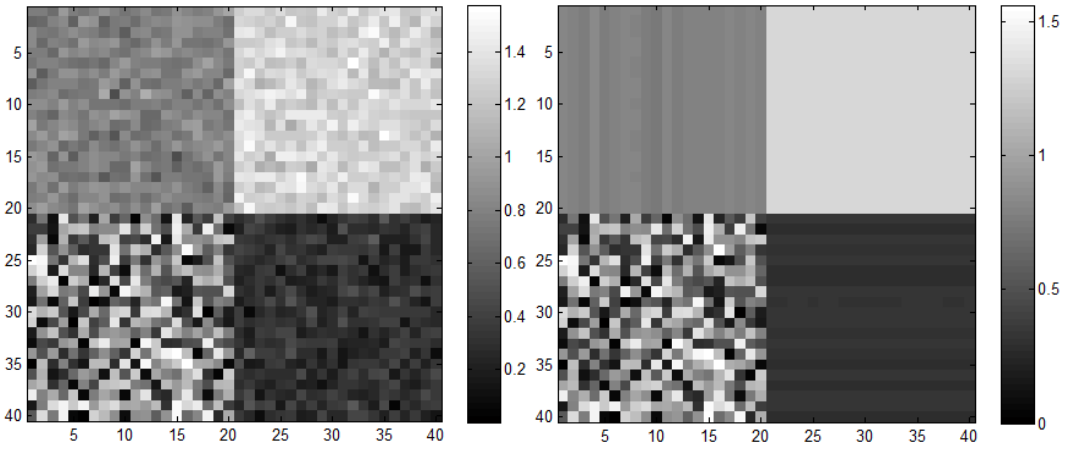

图 A-5 三联合簇，规则分布

攻硕期间取得的研究成果

[121] J. Shao, C. Gao, W. Zeng, et al. Synchronization-Inspired Co-Clustering and Its Application to Gene Expression Data[M]. 2017, 1075–1080