串流大數據預測式視覺化系統架構方法與實現

謝名峰　吳秀陽

Department of Computer Science and Information Engineering

National Dong Hwa University

Hualien, Taiwan, ROC

coco860731@gmail.com　showyang@gms.ndhu.edu.tw

摘要

隨著資訊時代的日新月異，大數據系統逐漸邁向巔峰，每天都有數以兆計的資料產生，而我們是否可以從這些制式化的數字中看出彼此之間的關聯性，甚至是整體未來的走向是一個重要的議題。目前現有的視覺化系統多侷限在即時亦或是過去的資料形式，而較無法讓使用者了解更多未來可能出現的狀況。另外對於多數一般的使用者其操作是否足夠簡單讓其可以自行操作得到想要的視覺化圖形，是否能夠在提升視覺化的多元性的前提下讓使用者可以用更少的操作難度與耗費時間，卻可以得到更多資料的相關性，甚至是特定數值的區域分布。本文的目的在探討現有的資料視覺化系統應用及擴展性，並提出3種具有預測性資料視覺化的方法達到更加多元化的搜尋及繪圖，包含了軌跡預測視覺化、預測軌跡密度視覺化、IoT 預測資料視覺化。

Keywords: 大數據、視覺化、資料預測、IoT、MongoDB、PySpark、Kafka、Google Map API、Apache Echarts

I. 研究目的與方法

本文主要探討預測式資料視覺化之應用，現有的視覺化系統多侷限在即時亦或是過去的資料形式，而較無法讓使用者了解更多未來可能出現的狀況。另外希望能夠簡化使用者的操作並提升視覺化的多元性，讓使用者可以用更少的操作難度與耗費時間，卻可以得到更多資料的相關性，甚至是特定數值的區域分布。 因此，在這樣的需求下我們提出數種預測式資料視覺化的方法，結合了資料的實時串流以及數據分析結果，從而生成未來的資料點並進行視覺化，並且建立一個戰情儀錶板的方式，讓使用者可以選擇操作各式各樣的資料視覺化，形成完整的視覺化系統，達到簡化操作並提升使用者對於資料的掌握度，不管是過去、現在甚至是未來。

II. 相關研究

串流資料是由數千個資料來源持續產生的資料，按順序以遞增的方式處理，並用於相互關聯、彙總、篩選和取樣等多種分析，並在需要時即時做出反應。使用地理地圖來可視化[1]，這種方法結合了圖形佈局和圖形聚類以及適當的聚類著色，並基於原始圖形中的聚類和連通性創建邊界。這些基礎數據都是靜態的，但當我們還想可視化一些底層過程時，問題變得更加困難。城市智能交通狀況監測系統[2]。於2020年由 Dujuan Zhang 與 Yangyang Jiang 所提出，其目的在解決傳統的城市智能交通擁堵狀況監測系統中，在對數據進行分析計算時，存在數據查詢時間長的問題。基於此上，設計了一種基於大數據的城市智能交通擁堵狀況監測系統，搭建了系統整體框架以及設計硬件系統和軟件系統。基於 Spark 技術實現數據的分佈式存儲和並行計算，完成基於大數據的城市智能交通擁堵狀況監測系統的設計。大數據分析可視化系統[3]，於2020年由 Tingting Liang, Shan Lu, Quansheng Liu 所提出，其探討了不同於傳統的數據處理工具，數據視覺化技術是指將數據信息以視覺的形式呈現。同時基於大數據感知技術和可視化技術，從本質上說數據視覺化的出現進一步推動了大數據的發展，也使得數據的表現形式多樣化。在數據視覺化的應用和發展過程中，在醫學、軍事領域、教學領域等多個領域也取得了巨大的成就，也實現了當代科學界所強調的人機交互功能。數據視覺化和大數據分析技術增強了數據信息的呈現效果，提高了數據信息的表現效率。

III. 串流大數據預測式視覺化系統架構

A. 串流大數據視覺化架構與流程

我們透過對過去的歷史資料進行分析，建立其頻繁模式的資料庫，以及對於即時的現有資料進行行為的辨識以及追蹤，讓我們對過去以及現在的資訊已有了一定程度的掌握。於此基礎上我們可以基於資料的頻繁模式並且辨識與追蹤從而對未來資料有了一定的預測手段。透過對未來資料的預測可以協助企業、消費者以及一般用戶對於未來可能發生的不確定性預作準備。它甚至能夠協助管理層級的人員有效率的因應變化、控制整體營運並做出可帶動未來成長並且避開風險的決策，其整體流程如圖 1。

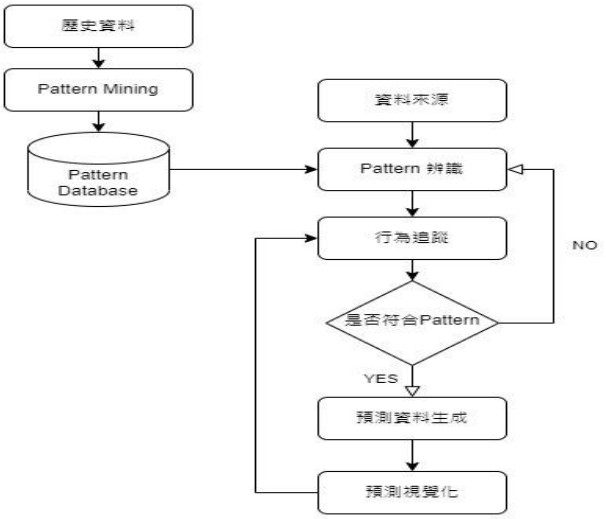


圖 1 串流大數據視覺化架構與流程

B. 串流大數據預測式視覺化系統架構

本文內包含了數種視覺化方法，其架構如圖 2所示，從可以分成幾個部分，資料來源、資料分流、資料庫儲存、資料處理生成再到視覺化以及最後的儀錶板彙整所有的視覺化資訊。首先，我們會收集各式各樣的 IoT 感測型資料或是物件的 GPS 點座標，傳入 Kafka 進行資料的分流，分別存入於MongoDB 的 Collection，而 Collection 內又分為兩個部分，各 IoT、載具收集儲存的現有過去資料以及透過 Pyspark 生成的未來資料，這些未來資料所依賴的是我們架構外部預設的部分，將過去的歷史資料收集並且進行 Pattern Mining 並且存入 Pattern Database，我們將依其配對到的 Pattern Data 透過 Pyspark 進行未來資料的生成並且存入回 MongoDB 內，其整體架構依照我們的資料以及視覺化方法進行區分為三種，再來透過我們所提出的三種視覺化功能提取相對應的這些資料進行視覺化，最後統整匯集到我們的戰情中心的儀錶板，讓使用者可以一覽所有資訊的即時現況，針對整體資料的過去走向以及未來可能的變化做出最好的決斷。

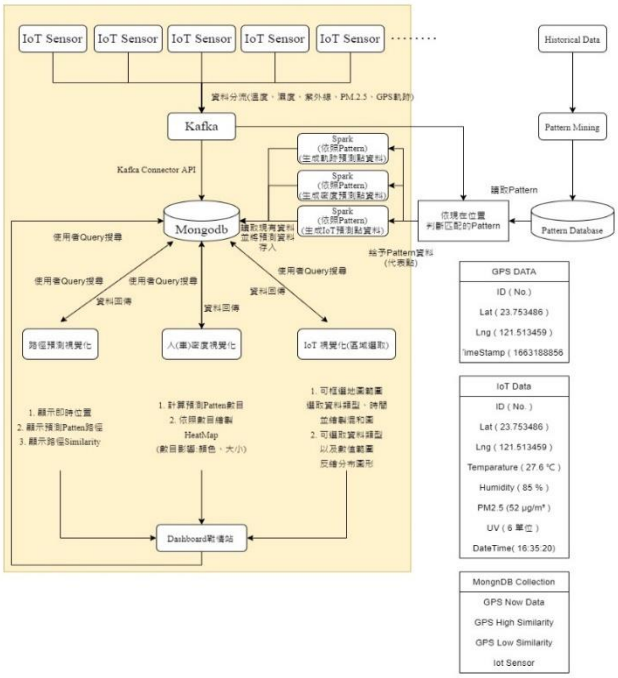


圖 2 串流大數據預測式視覺化系統架構方法與實現

C. 視覺化方法拓展

基於我們上述的架構以及視覺化方法，我們希望可以讓使用者透過此架構產生屬於自己的視覺化或是預測式視覺化系統。使用者們必須具有以下幾種條件以便於整體的構建，第一是物件資料的來源的資料標的，像是 ID、數值等等，第二則是儲存的數據資料庫，第三個是透過過去資料進行 Pattern Mining 並且存入 Pattern Database，用以辨識其當下走向，第四進行資料處理的大數據系統工具，最後就可以透過自己想要的方式進行資料的視覺化。

IV. 串流大數據預測式視覺化方法

A. 軌跡預測視覺化策略與實現

提出軌跡預測視覺化方法讓使用者除了了解物件現有的軌跡移動方式，甚至是延伸到未來的可能運動方式所提出的想法。具體的實作方式可以分成幾個部分。蒐集物件的 ID、經緯度座標以及時間(Timestamp)，透過 Kafka 進行分流儲存到 MongoDB 並發送到外部的 Pattern 預測模組，接著外部的預測模組提供我們物件可能的行為模式，透過 PySpark 我們可以將這些行為模式作為其未來可能的方向從而生成未來點座標並進行視覺化。使用者可以透過網頁上的控制按鈕對於搜尋的 Query 進行選擇，我們展示了所有物件其當下現有的軌跡路徑(紅線)、相似度較高的軌跡預測路徑(藍線)以及相似度較低的軌跡預測路徑(綠線)如圖 3。

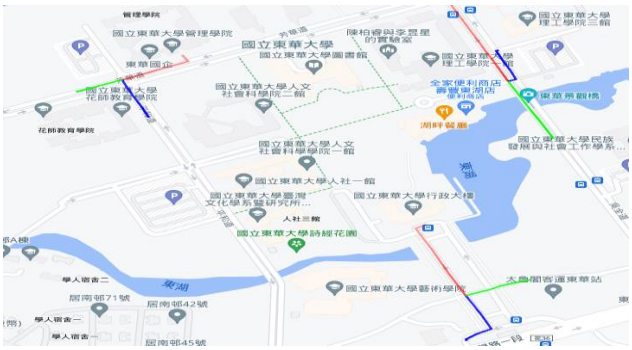


圖 3 所有物件軌跡預測

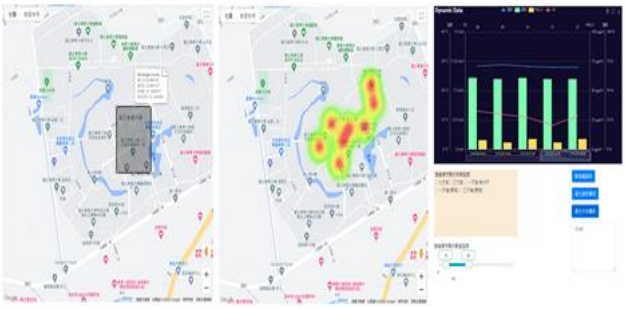
B. 密度預測視覺化策略與實現

密度預測視覺化方法是希望可以讓使用者除了了解物件或行人現有的軌跡移動方式，甚至是看出整體的多寡變化從而預先知曉前往或是避開該處。整體的運作從最一開始的物件資料收集 ID、經緯度座標以及時間戳記，透過 Kafka 分流儲存到 MongoDB 內部的表單，並進行預測點生成，最後預測密度視覺化網頁會透過 Query 向 MongoDB 內的表單搜尋使用者所選的時間戳記，並將這些資料進行處理計算以及轉換成方便視覺化的資料格式，最後回傳到前端的 Google Map API 依造我們預先設計的形式進行視覺化。如圖 4所示，我們展示了所有物件其當下現有的密度分布情況和預測的密度分布情況其視覺化的樣貌，讓使用者可以從而選擇前往或是避開該處。

圖 4 依最大可能樣式預測的密度變化(時間 1 分鐘後)

C. IoT 監測資料預測視覺化策略與實現

IoT 監測資料預測視覺化的實現流程包含收集各 IoT 感測器所產生的資料、透過 Kafka 分流儲存到 MongoDB、發送到外部的 Pattern 預測模組、透過外部的預測模組提供我們數值的變動模式、使用 PySpark 將這些行為模式作為其未來可能走向從而生成的數值並賦予時間戳記，最後我們前端的視覺化網頁會透過使用者所下達的 Query 指令搜尋向 MongoDB 要求使用者興趣區間的資料，並以 Json 的方式回傳進行視覺化。我們分別展示了當使用者於左側框選所感興趣的地圖範圍後，將該範圍內的所有 IoT 感測器數值進行搜尋處理以及視覺化顯示，以及使用者於下方的拉條選取感興趣的數值範圍，從而反向搜尋繪製出該數值範圍的 IoT 分布狀況如**錯誤! 找不到參照來源。**。



(a) (b) (c)

圖 5(a)框選範圍(b) IoT 感測器分布(c)資料視覺化

V. 系統實作與效能評估

在實作方面，我們選擇了多種語言在構建整體的系統，

詳如表 I。

表 I 實作系統環境設置

Software Version		
Apache Echarts	**Visual Studio**	**Apache Http Server**
3.2	16.9.4	2.4.53
MongoDB	**PySpark**	**Kafka**
6.0	3.3.0	3.2.1
Desktop		
CPU	**cores**	**Memory**	**GPU**	**Video Memory**
Intel i5-9600K	6	48 GB	NVIDIA GTX 1660	6 GB

A. 實作效果與效能評估

本文的實驗測試資料中採用了兩種不同的資料集，都是透過 Data Generate 的方式生成。第一是在校園內行人或是汽車的軌跡，模擬路線實際資料，並以此作為 Pattern 的基礎，並以 50%的資料集做為測試，比對預測生成資料與實際資料的準確度，第二種資料集則是生成 IoT 監測資料生成 2 萬筆的資料，模擬 IoT 感測器接收到實際資料的狀況，以此做為 Pattern 的基礎，後以 50%的資料集作為測試，比對預測生成資料以及實際資料的準確度。本文所採用的實驗資料格式如表 II 及表 III。

表 II 物件資料格式描述

Attribute	Description
ID	物件 ID，型別為 string
Lat (Latitude)	緯度，型別為 double
Lng (Longitude)	經度，型別為 double
Timestamp	當下時間，型別為 int

表 III IoT 感測資料格式描述

Attribute	Description
ID	Sensor ID，型別為 string
Lat (Latitude)	緯度，型別為 double
Lng (Longitude)	經度，型別為 double
Temperature	溫度，型別為 float
Humidity	濕度，型別為 float
UV	UV 數值，型別為 int
PM2.5	PM2.5 數值，型別為 int
Datetime	時間，型別為 Datetime

B. 軌跡預測視覺化效果實驗

預測軌跡視覺化方法所能產生的不一樣視覺化效果，透過使用者的不同參數選擇，只顯示軌跡相似度較高的預測路線如圖 6。

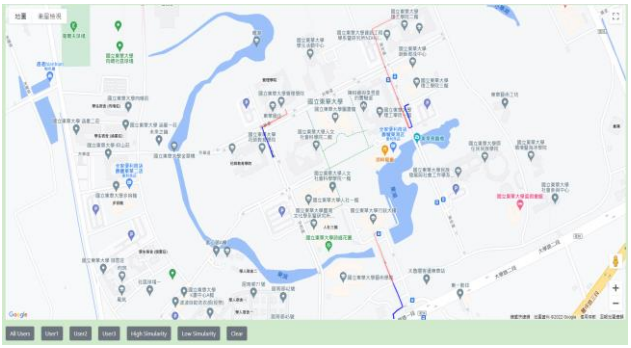


圖 6 全物件高相似度軌跡

C. 軌跡預測視覺化正確性實驗

預測未來點生成進行準確性上的實驗測試，依照 Data Generate 的方式生成在校園內行人或是汽車的軌跡作為實際資料，而後與依照辨識出的 Patten 進行預測點生成軌跡並進行 ST-LCSS[4]的計算，其絕對相似度閥值 ε1 設為 0.0001 而條件相似度閥值 ε2 我們設為 0.0003，依此作為分數的加權進行計算。其結果如圖 7。

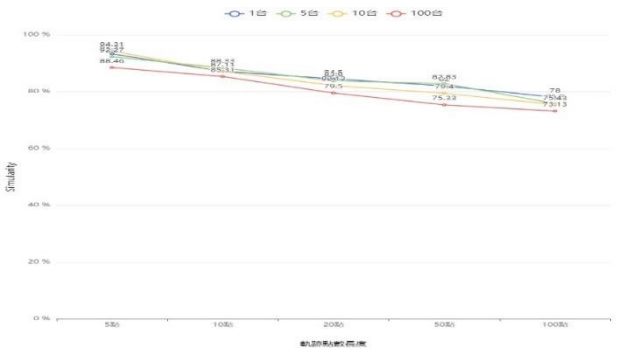


圖 7 不同車數軌跡生成預測準確率比較

D. 軌跡預測視覺化效能實驗

效能的實驗我們則分別測試了 1 台、5 台、10 台、50 台、100 台甚至是 1000 台物件在生成不同未來點數量上所需的時間如圖 8。

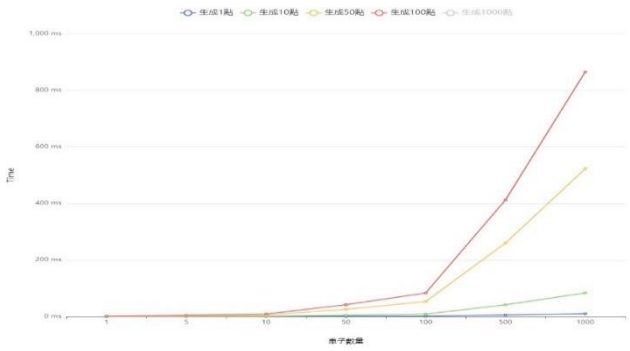


圖 8 預測點生成效能

E. 密度預測視覺化效果實驗

原先展現相似度較高的軌跡路徑其密度預測可能的變化流向，而後將可以透過使用者的不同參數選擇，只顯示軌跡相似度較低時的密度預測變化如圖 9。

圖 9 依較低可能樣式預測的密度變化(時間 1 分鐘後)

F. 密度預測視覺正確性實驗

依照前面所述 Data Generate 的方式生成在校園內行人或是汽車的資料模擬實際資料，而後與依照 Patten 進行預測點生成預測點並進行 ST-LCSS[4]的計算，計算生成的密度座標序列與實際資料序列之間的相似度作為其準確率的判斷。其絕對相似度閥值 ε1 設為 0.0001 而條件相似度閥值 ε2 我們設為 0.0003，而密度預測點的分布其時間為同一當下，依此在序列上的時間設為 1，作為分數的加權進行計算其結果如圖 10。

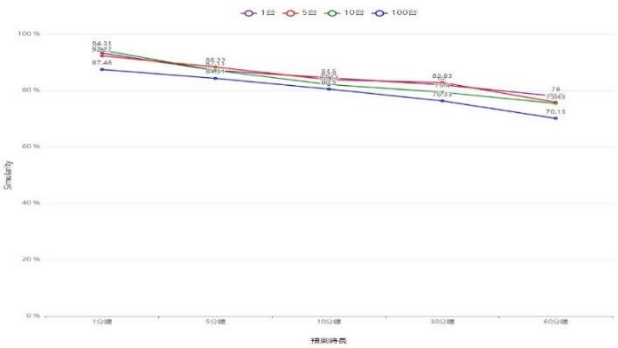


圖 10 不同車數密度預測比對準確率比較

G. 密度預測視覺化效能實驗

密度預測效能實驗我們分別測試了 1 點、5 點、10 點、50 點、100 點甚至是 10000 點物件在搜尋資料、資料處理以及繪製視覺化所需花費的時間如圖 11。

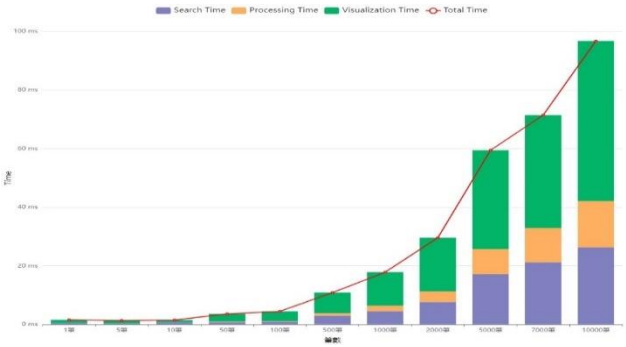


圖 11 密度預測視覺化花費時間

H. IoT 監測資料預測視覺化效果實驗

原先展現所有 Features 搜尋並且繪製的樣貌，而後透過使用者選取的不同 Features 可以繪製不同的數據資料圖如圖 12 所示。

圖 12 使用者選取較少 Features 的視覺化效果

I. IoT 監測資料預測視覺化正確性實驗

在溫度的相似度實驗中，我們的參數絕對相似度閥值 ε1 設為 1 而條件相似度閥值 ε2 設為 3。濕度的絕對相似度閥值 ε1 設為 5 而條件相似度閥值 ε2 設為 10。PM2.5 的絕對相似度閥值 ε1 設為 5 而條件相似度閥值 ε2 設為 10。UV 的絕對相似度閥值 ε1 設為 1 而條件相似度閥值 ε2 設為 2。分別依此作為分數的加權進行計算其結果如圖 13。

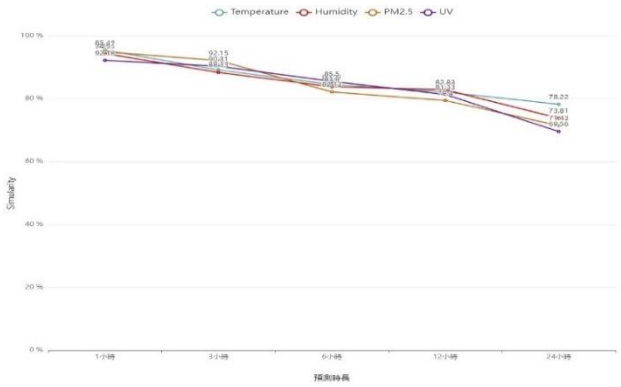


圖 13 多 IoT Sensor 預測資料與實際資料比對準確率

J. IoT 監測資料預測視覺化效能實驗

IoT 監測資料預測效能實驗，如圖 14 所示我們分別測試了在不同 Features 數下所需花費的連線時間、搜尋時間以及繪製視覺化所需花費的時間，並進行比較。

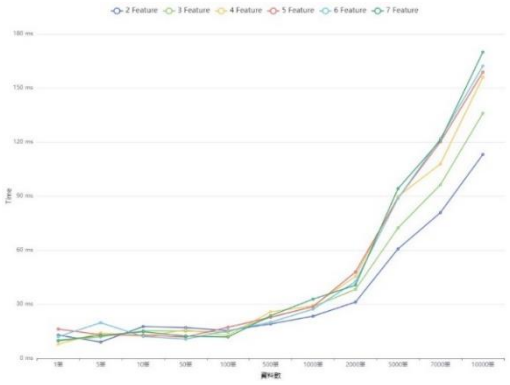


圖 14 多個 Features 視覺化效能比較

K. 架構效能實驗

我們針對整體架構上進行的實驗測試，如圖 15 所示我們分別測試了在不同的資料庫架構下去搜尋相同資料並且進行視覺化繪製所需花費的時間。

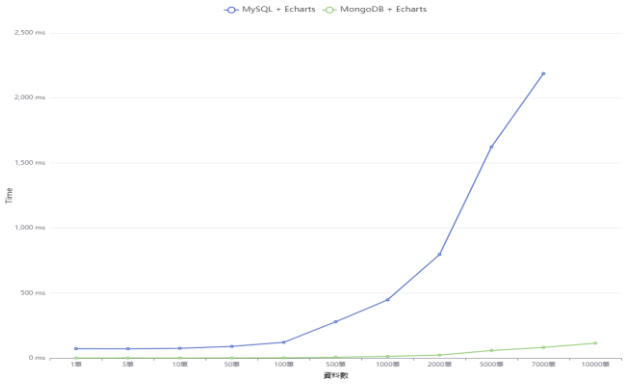


圖 15 資料庫架構效能比較

VI. 結論與未來工作

在本文中，介紹了串流大數據預測式視覺化系統架構以及三種視覺化方法的實現，針對整體預測式視覺化架構的設計以及後續拓展方法進行了實作與說明。透過方法及實驗我們可以知道:

1. 各項視覺化方法都有別於現有的視覺化工具，具有較多複合性以及預測性，效能上亦可以符合在1-2秒內達到即時顯示的部分。
2. 基於外部 Pattern 生成的預測點正確率亦有相當程度的正確性，讓使用者在對未來決策時有所幫助。

在未來工作上，對於預測點的生成方法，或許有更好的計算方式，像是有辦法得知 Google Map 上的道路實際路徑邊界範圍，在生成的時候預先判斷符合現有實際道路，進而提高軌跡的預測生成準確度是可以值得探討的議題。

REFERENCES

[1]. D. Mashima, S. G. Kobourov and Y. Hu, "Visualizing dynamic data with maps," 2011 IEEE Pacific Visualization Symposium, 2011, pp. 155-162, doi: 10.1109/PACIFICVIS.2011.5742385.

[2]. D. Zhang and Y. Jiang, "Design of Urban Intelligent Traffic Congestion Situation Monitoring System Based on Big Data" 2020 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), 2020, pp. 12-15, doi: 10.1109/ICITBS49701.2020.00011.

[3]. T. Liang, S. Lu and Q. Liu, "Data Visualization System Based on Big Data Analysis," 2020 International Conference on Robots & Intelligent System (ICRIS), 2020, pp. 76-79, doi: 10.1109/ICRIS52159.2020.00027.

[4]. 陳昱衡(2020)。以整合式形狀與符號聚合近似法為基礎之時間序列資料分析與行為辨識。國立東華大學資訊工程學系碩士論文，花蓮縣。

[5]. Ardi Imawan and Joonho Kwon. “A timeline visualization system for road traffic big data.” 2015 IEEE International Conference on Big Data (Big Data), 2015, pp.2928 – 2929, doi: 10.1109/BigData.2015.7364125