

Discovering Traffic Congestion Patterns in Big Transportation Networks

Importance

1. Japan recorded 2.61 thousand fatalities in 2022
2. 381 billion hours are lost annually in Japan.
3. 12 trillion yen lost annually.
4. Improving traffic safety is crucial to achieve SDGs.
5. Thus, it is important to discover traffic congestion patterns in large-scale transportation networks.

Table 1: Traffic incidents in Fukushima 2022 and 2023

	Total Incidents	Total Deaths (Including Elderly)	Total injuries	Total Property Incidents
This Year	1,678	30	33 (17)	1,949
Last Year	1,559	25	25 (17)	1,790
Change of Quantity	119	5	8 (0)	159
Change of rate	7.6%	20.0	32.0% (0.0%)	8.9%

Challenge

1. How to model traffic congestion data?
2. What should be the mathematical model to define traffic congestion patterns?
3. What is the algorithm to find all traffic congestion patterns?

Our model and Algorithm

Georeferenced Sequential Pattern Mining

Neighborhood itemset, $NG_i \subseteq O$
 $\forall O_i, O_j \in NG_i, dist(O_i, O_j) \leq maxDist$
 Neighborhood sequence, $NS_j = \bigcup_{i=1}^p NG_i$
 If frequency of NS_j is greater than minimum support ($minSup$), we call it traffic congestion pattern

Object	Coordinates
a	(0,0)
b	(0,1)
c	(3,0)
d	(1,1)
e	(3,2)
f	(3,3)
g	(4,4)

Table 4: Location database of road segments

If $mimsup=2(50\%)$,
 The pattern $\langle \{ab\} \{d,f\} \rangle$ appeared in two out of four sequences (50% of time)

but it is not Georeferenced Sequential Pattern because a,b,d is not neighbor of f

Item	Neighbors
a	a, b, d
b	a, b, d
c	c, d
d	a, b, c, d

Table 5: neighbor database

Meaning of above pattern:

1. Congestion was first observed on the roads a and b.
2. Later, congestion was observed on road d,f.

Minimum Support Based Sequential Pattern Mining

A minimum support is set for each items
 $MIS(i) = \text{minimum item support for } i$
 If X is int constant
 $MIS(i) = \text{Sup}(i) - X$
 If X is float constant
 $MIS = \text{Sup}(i) * X$
 $\text{MinSup}(s_x) = \max(MIS(s_x), LS)$
 *LS=limit of minimum support
 The smallest MIS of the items in the pattern is used as the minimum support.
 $MIS(i_1, i_2) = \min(MIS(i_1), MIS(i_2))$

GFSP	support
$\langle \{a\} \rangle$	4
$\langle \{b\} \rangle$	3
$\langle \{c\} \rangle$	4
$\langle \{d\} \rangle$	4
$\langle \{e\} \rangle$	3
$\langle \{f\} \rangle$	3

Table 6: support of each items

GFSP	MIS
$\langle \{a\} \rangle$	3
$\langle \{b\} \rangle$	2
$\langle \{c\} \rangle$	3
$\langle \{d\} \rangle$	3
$\langle \{e\} \rangle$	2
$\langle \{f\} \rangle$	2

Table 7: MIS

If $mimsup=2(50\%)$,
 The pattern $\langle \{ab\} \{d,f\} \rangle$ appeared in two out of four sequences (50% of time)

it is Minimum Support Based Sequential Pattern because $MIS(\langle \{ab\} \{d,f\} \rangle) = \min(MIS(a), MIS(b), MIS(d), MIS(f)) = 2$

Sequential Pattern Mining

$I = \{i_1, i_2, i_3, \dots, i_n\}$ be the set of items
 $G = \{i_1, i_2, i_3, \dots, i_m\}$ $m \leq n$ be the itemset
 $s = \{G_1, G_2, G_3, \dots, G_k\}$ be the sequence
 $SDB = \{s_1, s_2, s_3, \dots, s_j\}$ be the sequences database
 $|SDB| = j$ is database size
 s_x is subsequence of s in SDB
 $\text{Support of } s_x, \text{sup}(s_x) = |s|s \sqsubseteq s_x \wedge s \in SDB|$
 s_x is a frequent pattern if $\text{sup}(s_x) \geq \text{minSup}$
 $\text{minSup} = \text{minimum support}$

Day	hour	a	b	c	d	e	f	g
1	1	1	1	0	0	0	0	0
1	2	0	0	1	0	0	0	0
1	3	0	0	0	1	1	1	0
2	1	1	0	0	1	0	0	0
2	2	0	0	1	1	0	0	0
2	3	1	0	0	0	0	1	0
3	1	0	1	1	0	0	0	0
3	2	0	0	0	1	0	0	0
3	3	1	1	0	1	1	0	0
4	1	0	1	0	1	0	0	0
4	2	0	0	1	0	0	0	0
4	3	1	0	0	1	0	1	0
4	4	0	0	0	1	0	1	0

Table 3: sequence database

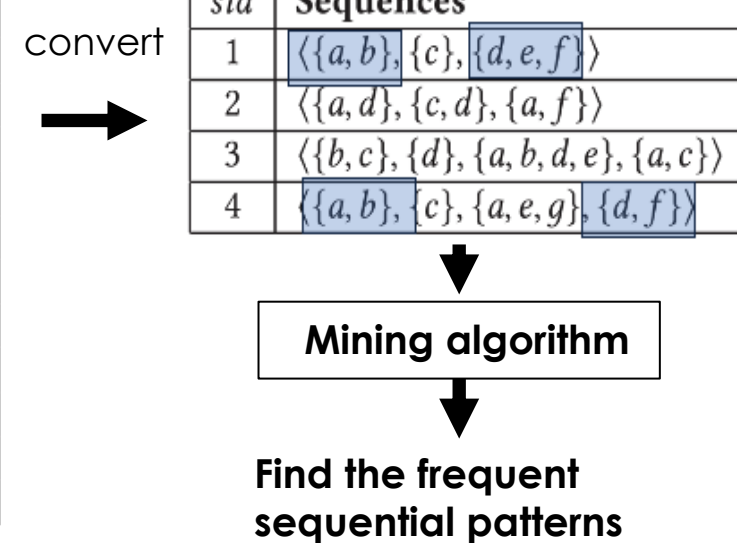
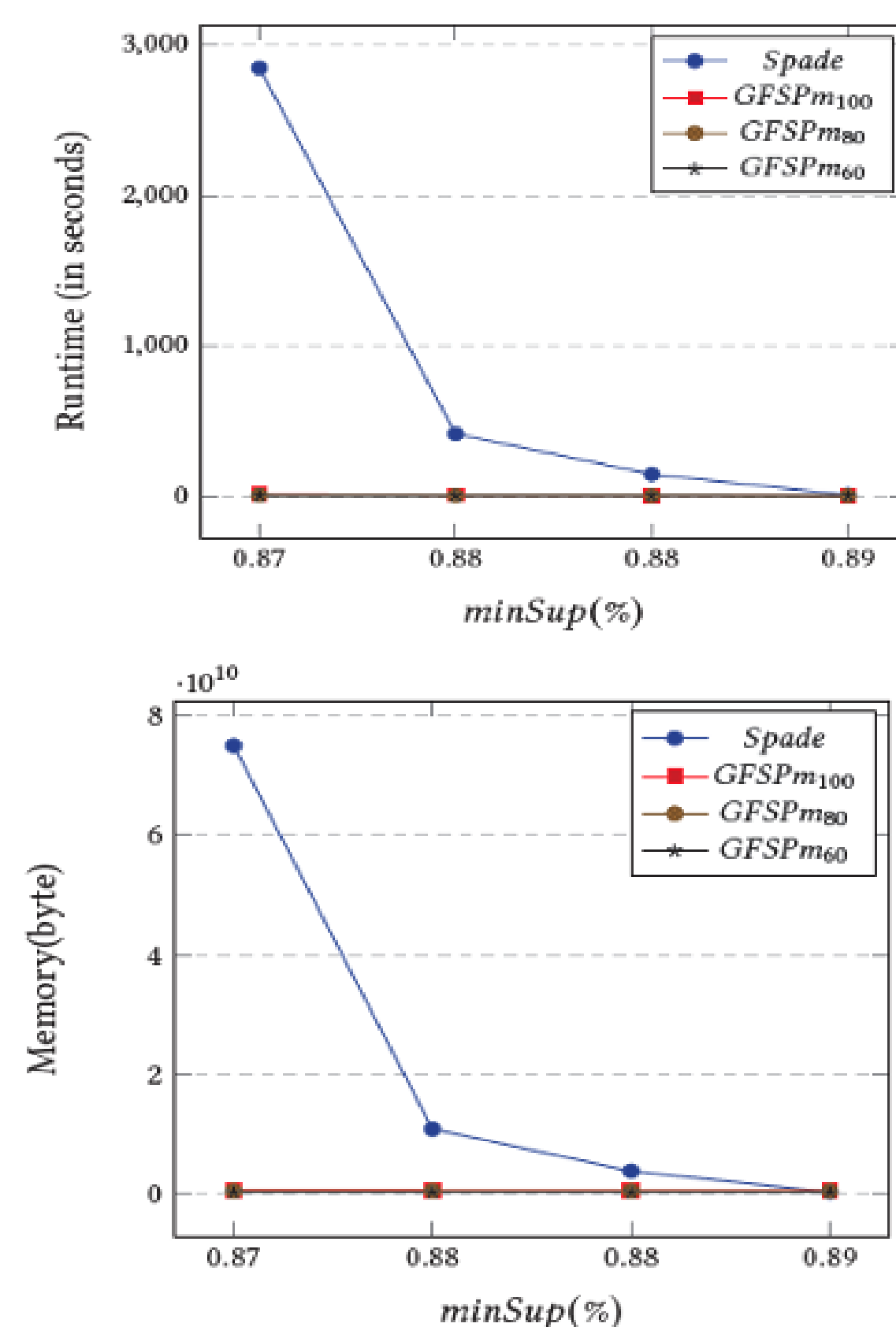
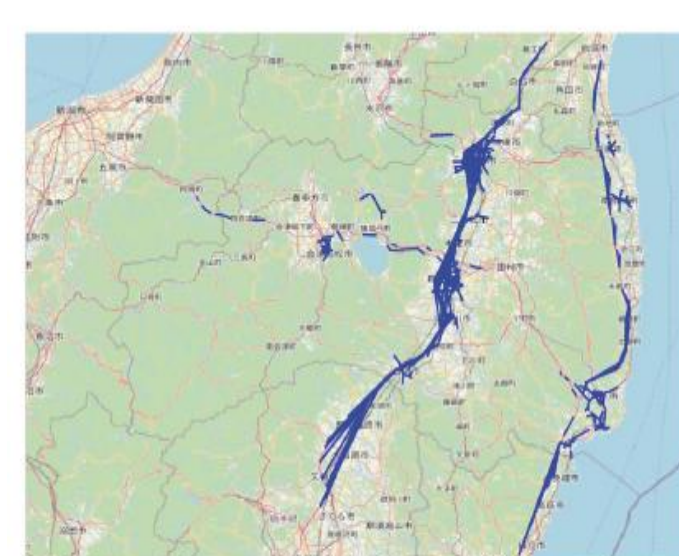


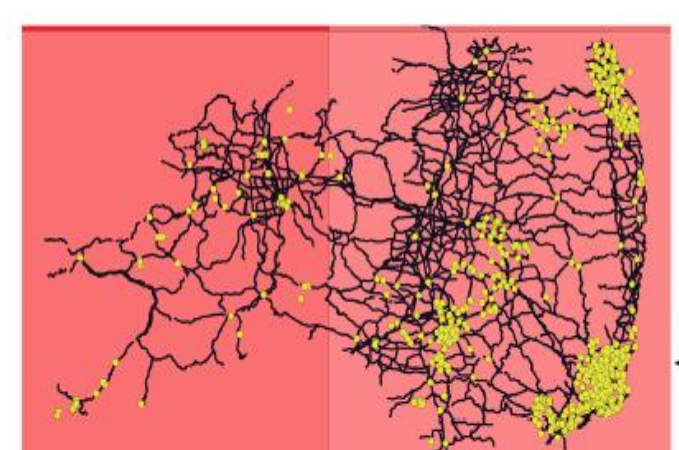
Table 2: Traffic congestion on road segments

If $mimsup=2(50\%)$,
 The pattern $\langle \{ab\} \{d,f\} \rangle$ appeared in two sequences (50% of time)
 So it is sequential pattern

Experimental Results



(a) Sensor network



(f) Visualization

Day	Time	R1	R2	...	Rn
1-1-2020	00:00	1	1	...	0
1-1-2020	00:05	0	1	...	1
...	...	...	...	...	...
1-1-2020	23:55	1	1	...	1
2-2-2020	00:00	0	1	...	0
2-2-2020	00:01	0	1	...	1
...	...	...	...	...	...

Binarized time series

Road ID	Spatial info
R1	line((0 0), (1 1))
R2	line((0 0), (2 2))
...	...

Spatial database

(b) Binarized Geo-reference time series

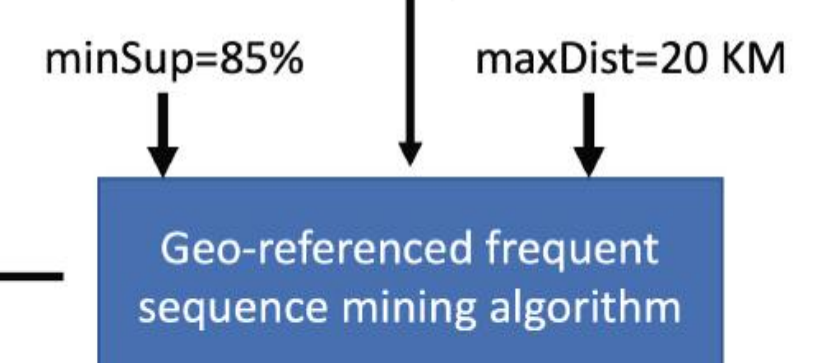
Day	SID	Sequence
1-1-2020	1	$\langle \{R1, R2\}, \{R2\}, \dots, \{R1, R2, \dots, Rn\} \rangle$
2-1-2020	2	$\langle \{R2\}, \{R2, Rn\}, \dots \rangle$
...	...	...

Sequence database

Road ID	Spatial info
R1	line((0 0), (1 1))
R2	line((0 0), (2 2))
...	...

Spatial database

(c) Geo-referenced sequence data



(d) Mining

Visualization by incorporating other data sources

outputs

$\langle \{R15, R37\}, \{R15, R124\}, \{R137\}, \{R88\} \rangle$
 $[support = 87\%] \dots$

(e) geo-referenced frequent sequence patterns

Figure 3: Case-study on Fukushima Traffic Congestion Data