Western Norway University of Applied Sciences

Utility-Oriented Techniques, Modeling, and Analytics

- CoBotAGV Seminar

Jerry Chun-Wei Lin

Jerry Chun-Wei Lin

jerrylin@ieee.org
IKE Lab: http://ikelab.net

2022.05.18

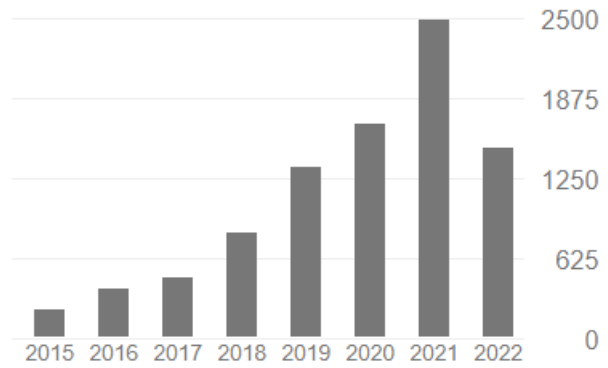
Personal info.

Professor in Western Norway University of Applied Sciences since 2018.7, Bergen Campus

IET Fellow, ACM Distinguished Scientist, IEEE Senior Member

Published 500+ papers in several journals

Cited by		VIEW ALL
	All	Since 2017
Citations	9457	8323
h-index	49	44
i10-index	203	185



Editor-in-Chief
Data Science and Pattern Recognition

Associate Editor (10+ SCI journals)
IEEE TNNLS, IEE TCYB, Information Sciences, …etc.

2

Outline

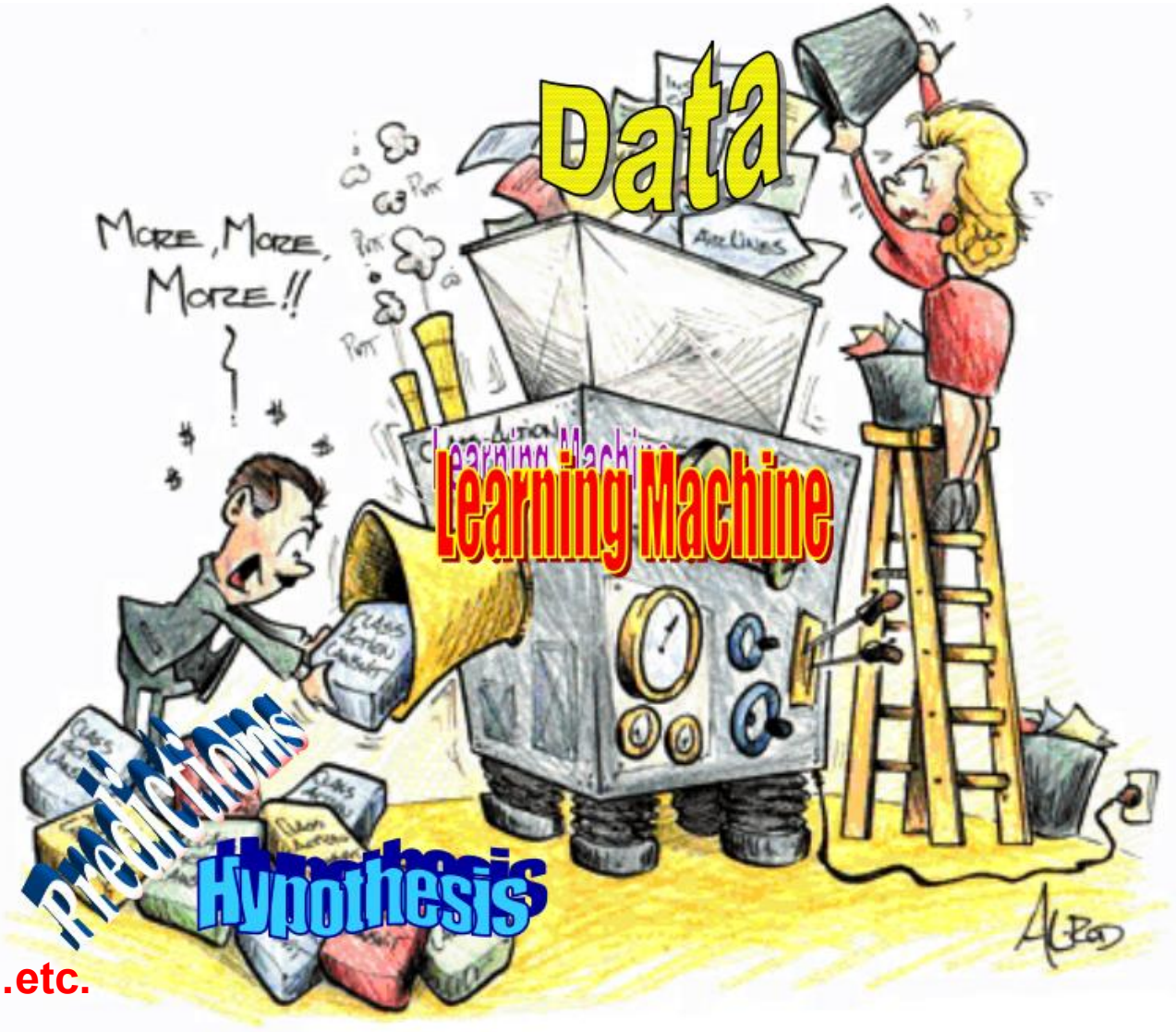
- How do we find the useful patterns?
  - Association rule mining
  - High-utility itemset mining
- Utility-Driven Mining Techniques
  - Data-type-oriented models
  - Constraint-based models
  - Big and dynamic data models
- Projects in progress for ML/DL
- SPMF project, CoBotAGV project
- Conclusions and future works

What is the interesting patterns

- There can be millions or billions of patterns from a large database
- Interestingness measures are used to select interesting patterns:
  - support, confidence, life, up-to-date, utility, periodic timestamp, correlation, importantness, weight, statistical significance, etc.
- To efficiently find patterns, **strategic data mining** is designed to avoid considering all possibilities
  - Effectiveness and efficiency

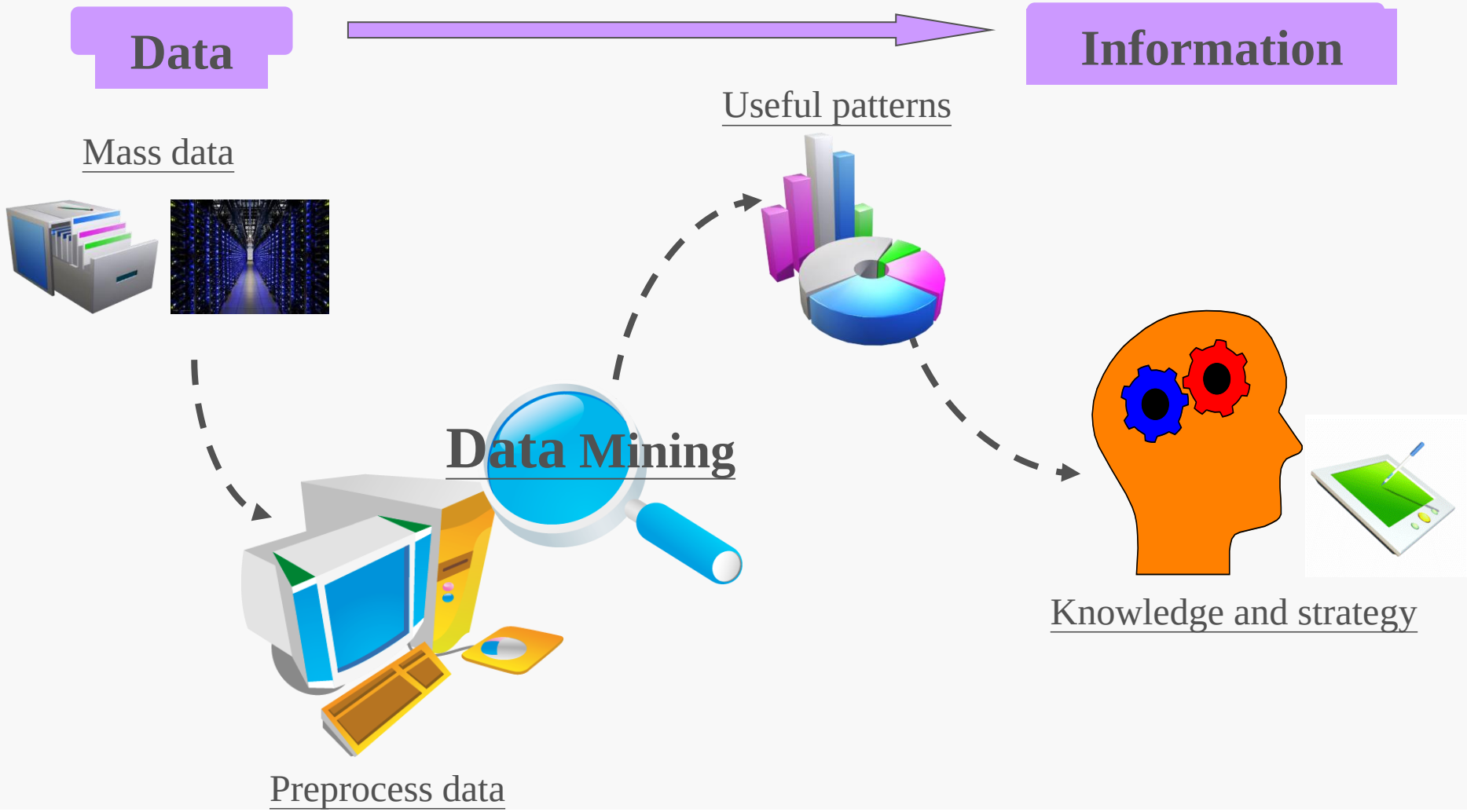
Introduction of background

- Why do we need data mining?



Analytics, prediction, recommendation…etc.

The role of data mining



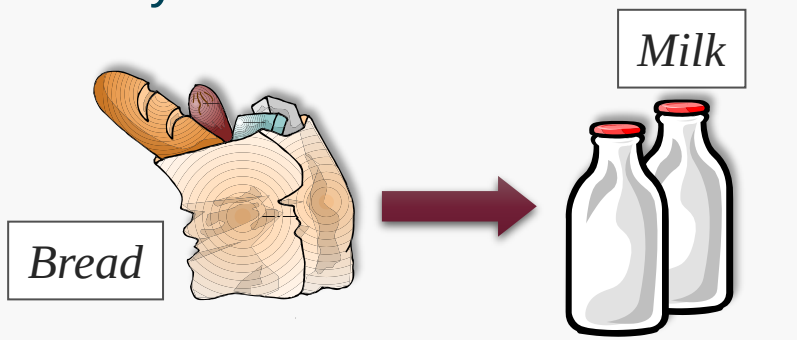
Frequent itemset mining

- Apriori algorithm (proposed by Agrawal et al. in SIGMOD1993)
- FP-growth algorithm (Han et al., 2004)
  - FIM/ARM is a fundamental research topic in DM
  - Many real-world applications, such as market basket analysis

TID	Items
T1	milk, bread, **cookies, beverage**
T2	milk, bread, **cookies**
T3	bread, **cookies, beverage**
T4	milk, bread
T5	milk, bread

minsup

minconf



IF bread is bought

THEN milk is bought

An example of frequent itemset mining

Input:

A transaction database

Transaction	Items
T_1	{a, b, c, d}
T_2	{a, b}
T_3	{a, d, e}
T_4	{a, b, d, e}

Minsup = 2

Output:

frequent itemsets

Itemset	Support
{a}	4
{a, b}	3
{a, b, d}	2
…	…

Possible solutions

- The **naïve approach**
  - scan the database to calculate the frequency of each possible itemset
    - {a}, {a,b}, {a,c}, {a,d}, {a,e}, {a,b,c}, {a,b,d}, {a,b,e}…
      {b}, {b, c}, …. {c}, {c,d}, {c,e} … {d}, {d,e},{e}
      … {a,b,c,d,e}
- But if n items, then $2^n - 1$ possible itemsets.

2^{500} = 327339060789614187001318969682759915221664204604306478948329136809613379206376147488327009232590415715088668412756007100921725654588539305332852758937 6 → This approach is inefficient

Apriori property

Property (monotonicity)

Let two itemsets be **X** and **Y.** If **X** $\subset$ **Y**, then the support of Y is less than or equal to the support of **X**.

Transaction	items
T_1	{a, b, c, d, e}
T_2	{a, b, e}
T_3	{c, d, e}
T_4	{a, b, d, e}

Example

The support of **{a,b}** is **3**.

Thus, supersets of **{a,b}** have a support ≤ **3.**

(abc:1), (abcd:1), (abd:2)….etc.

Introduction of utility mining

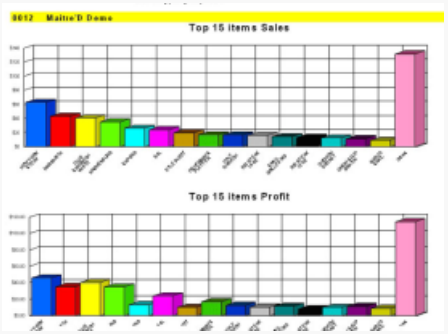
- Traditional FIM or ARM only handles binary dataset
  - In real-life situations
  - Different weight, interestingness, importance…etc.
  - Incomplete information may have wrong decision
  - An itemset with high support may have low utility

- Basket Analytics
  - Each item has a distinct price/profit
  - Each item in a transaction is associated with a distinct quantity

Items	Frequency	Profit
Diamonds	Low	High
Clothes	High	Low

What is high-utility itemset mining?

$ 2 * 100 = 200



$ 1500 * 2 = 3000

coffee	beer	diamond	…	…
$5	$2	$1500	…	…

Not only for basket analytics, but also in other applications (recommendation, prediction,…etc.)

What is high-utility itemset mining?

- High Utility Itemset (HUI)
  - Each item has a weight/unit profit and can appear more than once in each transaction (purchase quantity)
  - The utility of an itemset is measured by the unit profit and purchase quantity (e.g., profit of the itemset)
  - An itemset is a HUI if its utility is no less than a user-specified minimum utility threshold (priori knowledge)

Transactional Database

TID	Transaction
1	(A,1)(C,1)(D,1)
2	(A,1)(B,2)(C,1)(D,6)
3	(B,4)(C,3)(D,3)
4	(B,2)(C,2)

Item	Profit
A	5
B	2
C	1
D	2

minUtil = 30

High-Utility Itemsets
{BD}:30, {BCD}:34

High-utility itemset mining

Input

a transaction database

TID	Transaction
T_1	$(a,1),(b,5),(c,1),(d,3),(e,1),(f,5)$
T_2	$(b,4),(c,3),(d,3),(e,1)$
T_3	$(a,1),(c,1),(d,1)$
T_4	$(a,2),(c,6),(e,2),(g,5)$
T_5	$(b,2),(c,2),(e,1),(g,2)$

a unit profit table

Item	a	b	c	d	e	f	g
Profit	5	2	1	2	3	1	1

***minutil**: a minimum utility threshold set by the user (a positive integer)*

Output: All high-utility itemsets (itemsets having a utility $\geq minutil$)

For example, if *minutil = 33$*, the high-utility itemsets are:

{b,d,e} 36$ 2 transactions	{b,c,d} 34$ 2 transactions
{b,c,d,e} 40$ 2 transactions	{b,c,e} 37 $ 3 transactions

Utility calculation

a transaction database

TID	Transaction
T_1	$(a,1),(b,5),(c,1),(d,3),(e,1),(f,5)$
T_2	$(b,4),(c,3),(d,3),(e,1)$
T_3	$(a,1),(c,1),(d,1)$
T_4	$(a,2),(c,6),(e,2),(g,5)$
T_5	$(b,2),(c,2),(e,1),(g,2)$

a unit profit table

Item	a	b	c	d	e	f	g
Profit	5	2	1	2	3	1	1

The **utility** of the itemset {b,d,e} is calculated as follows:

u({b,d,e}) = (5x2)+(3x2)+(3x1) + (4x2)+(2x3)+(1x3) = **36$**

utility in transaction T_1 | utility in transaction T_2

A difficult task!

Why? the **utility** measure does not hold the monotonic property

TID	Transaction
T_1	$(a,1),(b,5),(c,1),(d,3),(e,1),(f,5)$
T_2	$(b,4),(c,3),(d,3),(e,1)$
T_3	$(a,1),(c,1),(d,1)$
T_4	$(a,2),(c,6),(e,2),(g,5)$
T_5	$(b,2),(c,2),(e,1),(g,2)$

Item	a	b	c	d	e	f	g
Profit	5	2	1	2	3	1	1

u({**b,d,e**}) = **36$**
u({**b,c,d,e**}) = **40$**
u({**a,b,c,d,e,f**}) = 30 **$**

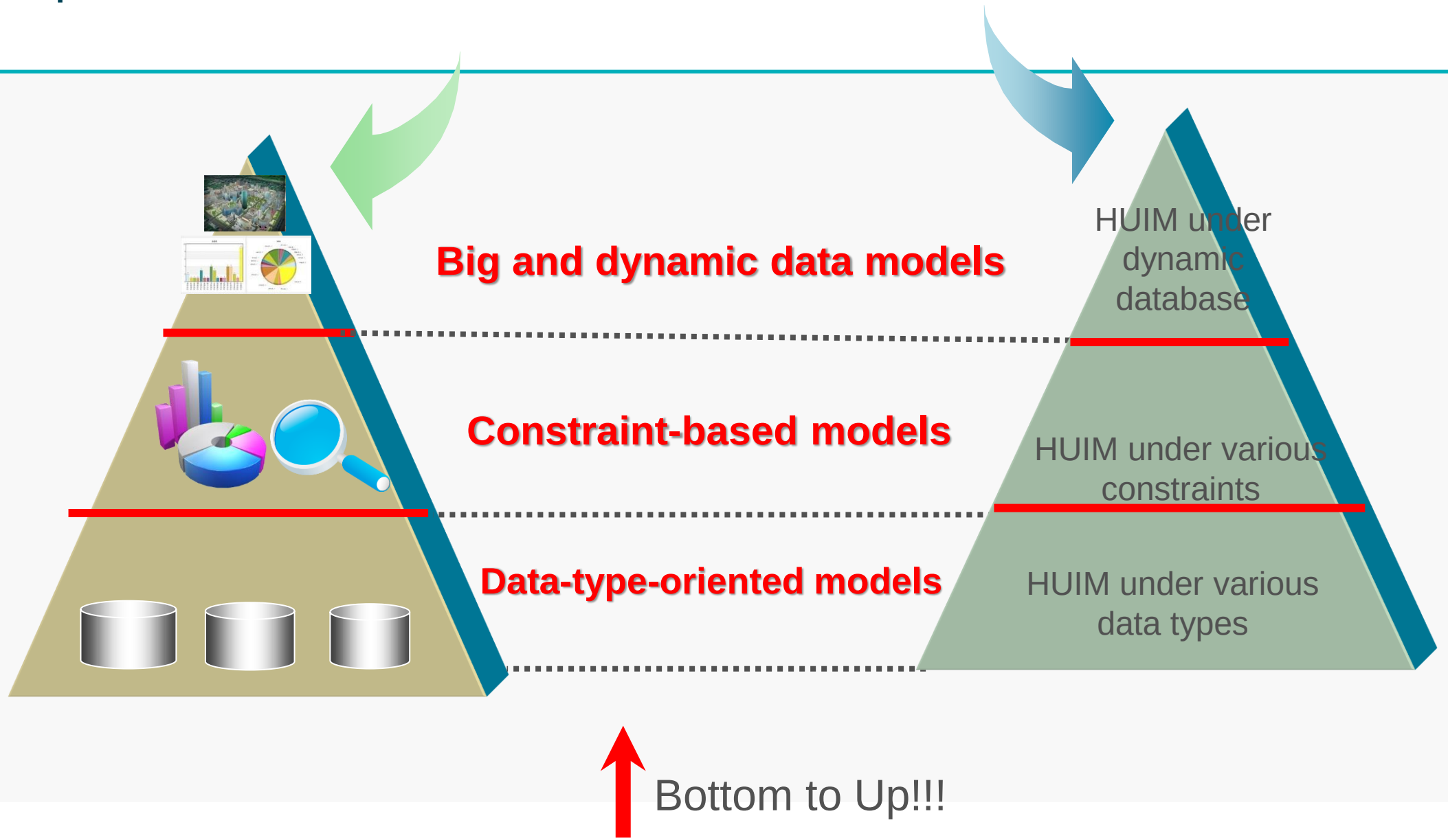
Minutil = 38
(b, c, d, e) would not be found

The utility-driven data mining framework

- **Data-type-oriented models**
  - Sensor or IoT data
  - Temporal data
  - Sequential data
- **Constraint-based models**
  - Weight and correlation
  - Individualized evaluation
- **Big and dynamic data models**
  - Dynamic situation
    - Insertion, deletion, modification
  - Stream data
  - Large-scale, heterogenous, multi-source data

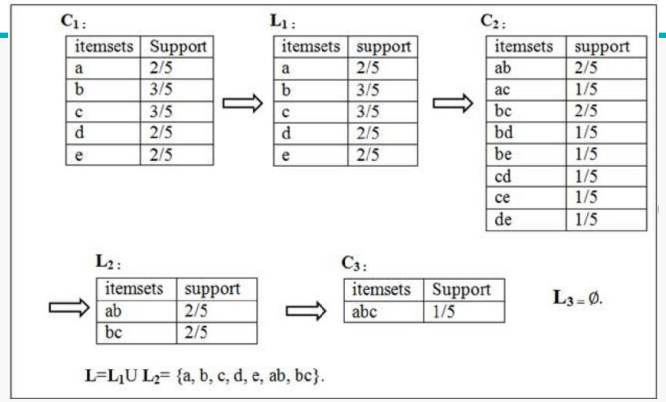
Utiliverse

Proposed framework



What is important in mining progress

- **Effectiveness**
  - Data-type-based
    - New pattern development
  - Constraint-based
    - More specific consideration
- **Efficiency**
  - Mining performance
    - Data structure
      - Apriori
      - Projection
      - Tree
      - List-based
    - Pruning strategy
      - Downward closure property
  - Handling large-scale data



C_1:

itemsets	Support
a	2/5
b	3/5
c	3/5
d	2/5
e	2/5

L_1:

itemsets	support
a	2/5
b	3/5
c	3/5
d	2/5
e	2/5

C_2:

itemsets	support
ab	2/5
ac	1/5
bc	2/5
bd	1/5
be	1/5
cd	1/5
ce	1/5
de	1/5

L_2:

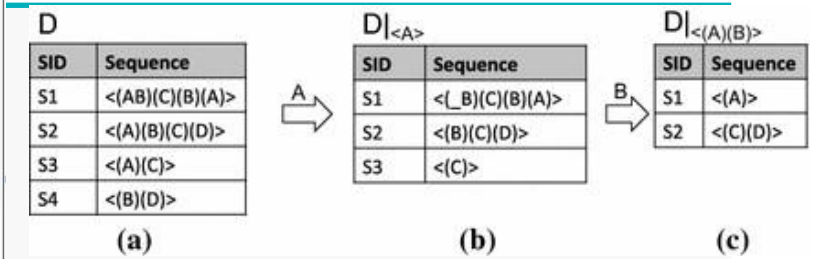
itemsets	support
ab	2/5
bc	2/5

C_3:

itemsets	Support
abc	1/5

$L_3 = \emptyset$.

$L = L_1 \cup L_2 = \{a, b, c, d, e, ab, bc\}$.



D

SID	Sequence
S1	<(AB)(C)(B)(A)>
S2	<(A)(B)(C)(D)>
S3	<(A)(C)>
S4	<(B)(D)>

A

$D|_{<A>}$

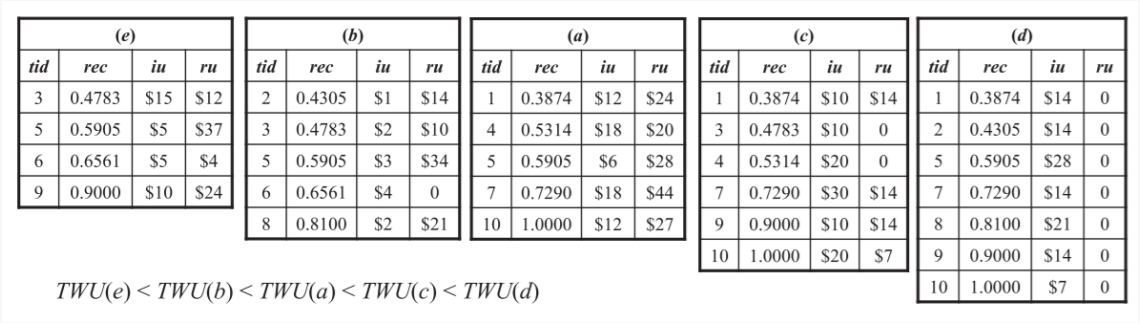
SID	Sequence
S1	<(_B)(C)(B)(A)>
S2	<(B)(C)(D)>
S3	<(C)>

B

$D|_{<(A)(B)>}$

SID	Sequence
S1	<(A)>
S2	<(C)(D)>

(a) (b) (c)



(e)

tid	rec	iu	ru
3	0.4783	$15	$12
5	0.5905	$5	$37
6	0.6561	$5	$4
9	0.9000	$10	$24

(b)

tid	rec	iu	ru
2	0.4305	$1	$14
3	0.4783	$2	$10
5	0.5905	$3	$34
6	0.6561	$4	0
8	0.8100	$2	$21

(a)

tid	rec	iu	ru
1	0.3874	$12	$24
4	0.5314	$18	$20
5	0.5905	$6	$28
7	0.7290	$18	$44
10	1.0000	$12	$27

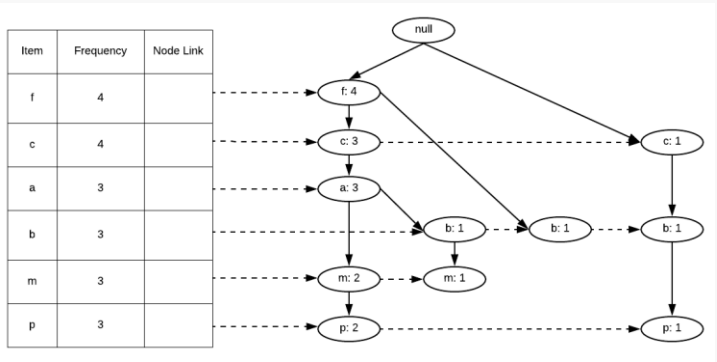
(c)

tid	rec	iu	ru
1	0.3874	$10	$14
3	0.4783	$10	0
4	0.5314	$20	0
7	0.7290	$30	$14
9	0.9000	$10	$14
10	1.0000	$20	$7

(d)

tid	rec	iu	ru
1	0.3874	$14	0
2	0.4305	$14	0
5	0.5905	$28	0
7	0.7290	$14	0
8	0.8100	$21	0
9	0.9000	$14	0
10	1.0000	$7	0

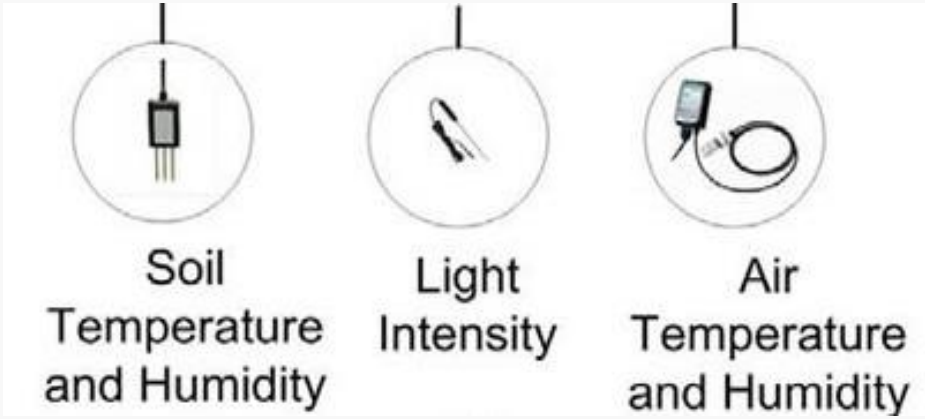
$TWU(e) < TWU(b) < TWU(a) < TWU(c) < TWU(d)$

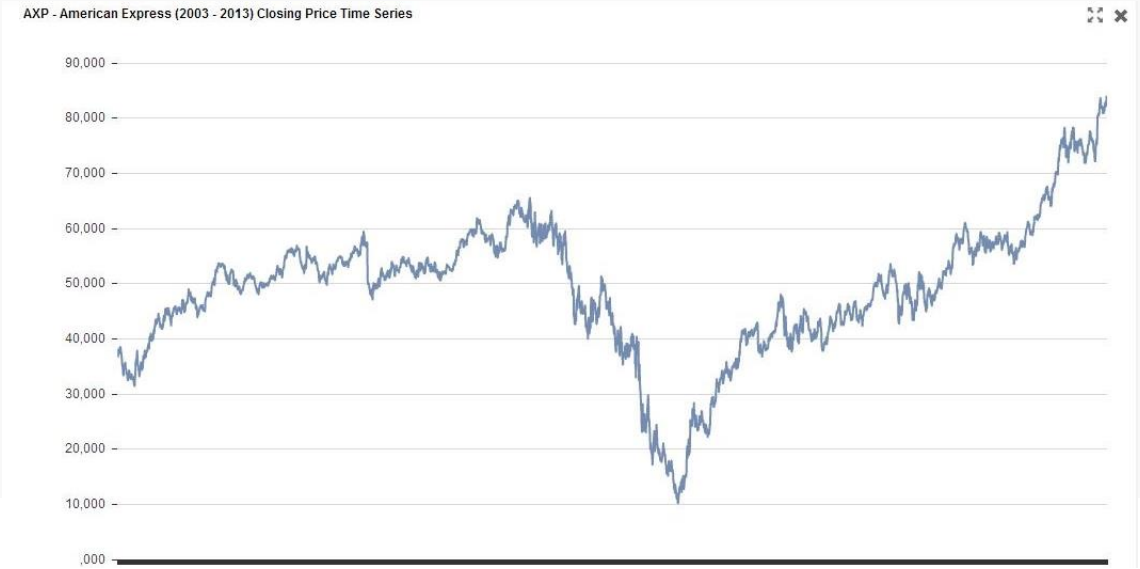


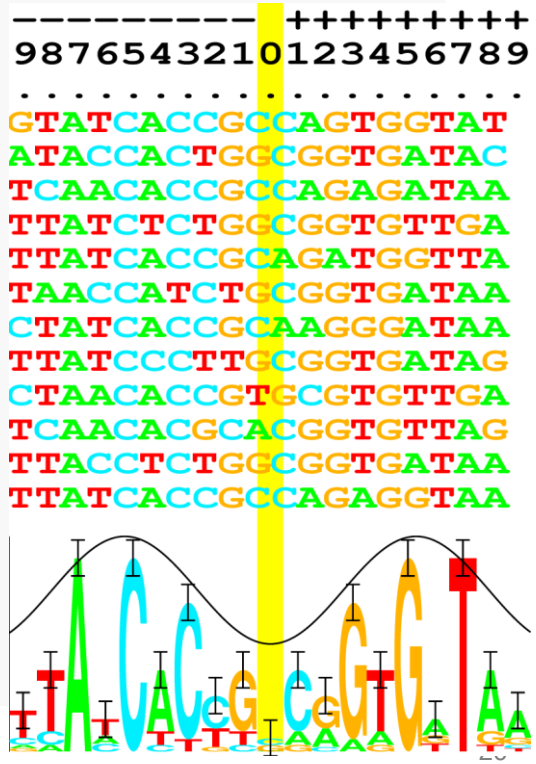
Item	Frequency	Node Link
f	4	
c	4	
a	3	
b	3	
m	3	
p	3	

4 different proposed models

- Data type is vey various
  - Temporal/recency
  - Multi-dimension
  - Uncertain
  - Sequence







Data level (1) – Recency

1. Wensheng Gan, **Jerry Chun-Wei Lin***, Han-Chieh Chao, Philippe Fournier-Viger, XuanWang, and Philip S. Yu, “Utility-Driven Mining of Trend Information for Intelligent System,” **ACM Transactions on Management Information Systems**, Vol. 11(3), Article No.14, 2020

2. Wensheng Gan, **Jerry Chun-Wei Lin***, Jiexiong Zhang, Hongzhi Yin, Philippe Fournier-Viger, Han-Chieh Chao, and Philip S. Yu, “Mining Across Multi-Dimensional Sequences,” **ACM Transactions on Knowledge Discovery from Data**, Vol.15(5), Article No.:82, 2021 (SCI, JCR:Q2, IF:2.713)

3. Wensheng Gan, **Jerry Chun-Wei Lin**, Han-Chieh Chao, Athanasios V. Vasilakos, and Philip S. Yu, “Utility-driven Data Analytics on Uncertain Data,” **IEEE Systems Journal**, Vol. 14(3), pp. 4442–4453, 2020 (SCI, JCR:Q2, IF:3.921)

4. Gautam Srivastava, **Jerry Chun-Wei Lin***, Alireza Jolfaei, Yuanfa Li, and Youcef Djenouri, “Uncertain-Driven Analytics of Sequence Data in IoCV Environments,” **IEEE Transactions on Intelligent Transportation Systems**, Vol. 22(8), pp. 5403–5414, 2021 (SCI, JCR:Q1, IF:6.492)

Recency factor

- More recent patterns could make more significant contributions
  - More recent transactions and the items within them are more important

Definition 3.1. The recency of a q-th transaction T_q is denoted as $r(T_q)$ and defined as $r(T_q) = (1-\delta)^{(T_{current}-T_q)}$, where δ is a user-specified time-decay factor ($\delta \in (0,1]$), $T_{current}$ is the current timestamp that is equal to the number of transactions in D, and T_q is the *TID* of the q-th transaction.

Pattern	$r(X)$	$u(X)$	Pattern	$r(X)$	$u(X)$
(a)	2.9145	\$66	(ce)	1.2405	\$45
(c)	3.6235	\$100	(de)	1.3414	\$57
(d)	4.3626	\$112	(abd)	0.5314	\$37
(e)	2.3624	\$35	(acd)	1.9048	\$137
(ac)	2.3831	\$140	(ade)	0.5314	\$39
(ad)	2.4362	\$111	(bde)	0.5314	\$36
(bd)	1.6479	\$69	(cde)	0.81	\$34
(be)	1.5524	\$34	$(abde)$	0.5314	\$42
(cd)	2.7148	\$119			

Techniques

(e)

tid	rec	iu	ru
3	0.4783	$15	$12
5	0.5905	$5	$37
6	0.6561	$5	$4
9	0.9000	$10	$24

(b)

tid	rec	iu	ru
2	0.4305	$1	$14
3	0.4783	$2	$10
5	0.5905	$3	$34
6	0.6561	$4	0
8	0.8100	$2	$21

(a)

tid	rec	iu	ru
1	0.3874	$12	$24
4	0.5314	$18	$20
5	0.5905	$6	$28
7	0.7290	$18	$44
10	1.0000	$12	$27

(c)

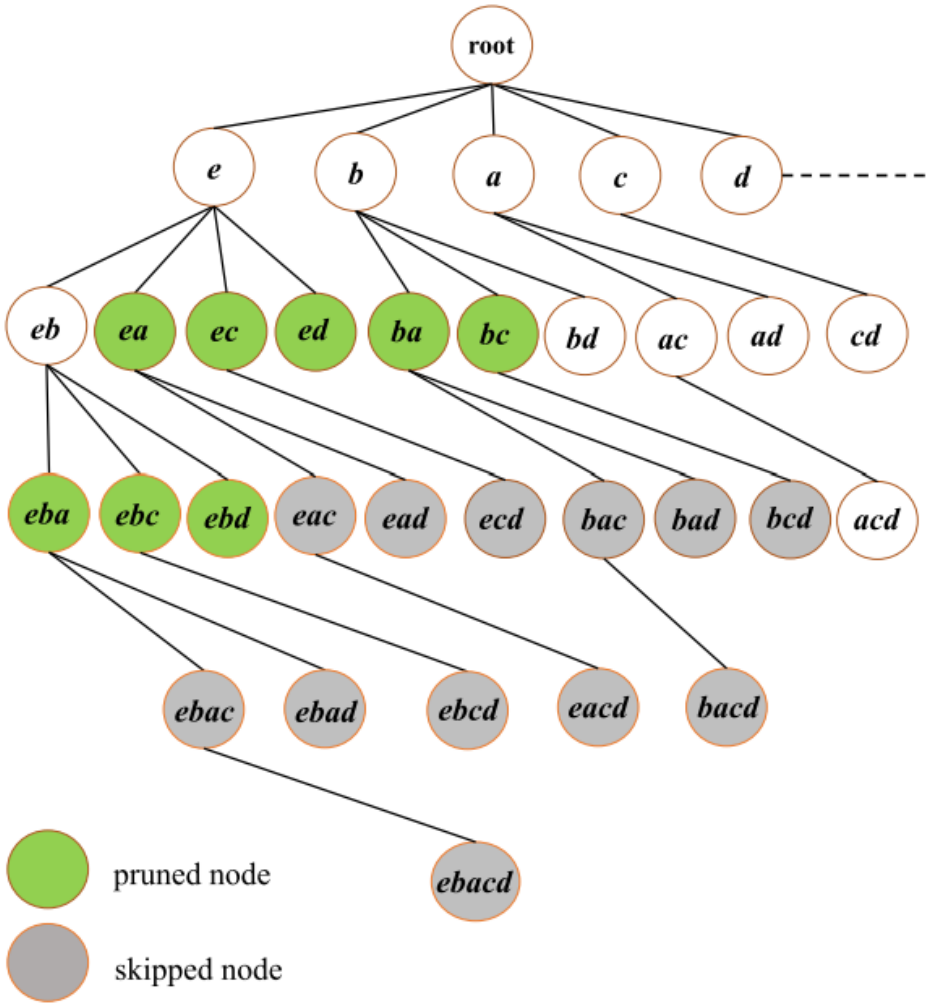
tid	rec	iu	ru
1	0.3874	$10	$14
3	0.4783	$10	0
4	0.5314	$20	0
7	0.7290	$30	$14
9	0.9000	$10	$14
10	1.0000	$20	$7

(d)

tid	rec	iu	ru
1	0.3874	$14	0
2	0.4305	$14	0
5	0.5905	$28	0
7	0.7290	$14	0
8	0.8100	$21	0
9	0.9000	$14	0
10	1.0000	$7	0

$TWU(e) < TWU(b) < TWU(a) < TWU(c) < TWU(d)$

- 4 Pruning strategies to respectively improve the mining performance by reducing the number of redundant and unpromising patterns in the early stage



(d)

tid	re	iu	ru
1	0.3874	$14	0
2	0.4305	$14	0
5	0.5905	$28	0
7	0.7290	$14	0
8	0.8100	$21	0
9	0.9000	$14	0
10	1.0000	$7	0

Data level (2) – Multi-dimension

1. Wensheng Gan, **Jerry Chun-Wei Lin***, Han-Chieh Chao, Philippe Fournier-Viger, XuanWang, and Philip S. Yu, “Utility-Driven Mining of Trend Information for Intelligent System,” **ACM Transactions on Management Information Systems**, Vol. 11(3), Article No.14, 2020

2. Wensheng Gan, **Jerry Chun-Wei Lin***, Jiexiong Zhang, Hongzhi Yin, Philippe Fournier-Viger, Han-Chieh Chao, and Philip S. Yu, “Mining Across Multi-Dimensional Sequences,” **ACM Transactions on Knowledge Discovery from Data**, Vol.15(5), Article No.:82, 2021 (SCI, JCR:Q2, IF:2.713)

3. Wensheng Gan, **Jerry Chun-Wei Lin**, Han-Chieh Chao, Athanasios V. Vasilakos, and Philip S. Yu, “Utility-driven Data Analytics on Uncertain Data,” **IEEE Systems Journal**, Vol. 14(3), pp. 4442–4453, 2020 (SCI, JCR:Q2, IF:3.921)

4. Gautam Srivastava, **Jerry Chun-Wei Lin***, Alireza Jolfaei, Yuanfa Li, and Youcef Djenouri, “Uncertain-Driven Analytics of Sequence Data in IoCV Environments,” **IEEE Transactions on Intelligent Transportation Systems**, Vol. 22(8), pp. 5403–5414, 2021 (SCI, JCR:Q1, IF:6.492)

Multi-dimension data

- We have multi-dimension data in the real world

SID	Time	Place	Customer	Sex	Age	Occupation	Q-sequence
S_1	5/2/2017 09:31	Store	***	Male	Young	Doctor	<[(*a*:1) (*c*:3)], [(*a*:5) (*c*:1) (*e*:4)], [(*c*:2)], [(*b*:1)]>
S_2	5/2/2017 10:02	Supermarket	***	Female	Middle	Lawyer	<[(*c*:1)], [(*b*:4)], [(*b*:9) (*d*:8)], [(*b*:9) (*e*:6)]>
S_3	5/5/2017 12:25	Supermarket	***	Male	Young	Actor	<[(*a*:10) (*d*:5)]>
S_4	5/7/2017 10:30	Drugstore	***	Male	Young	Farmer	<[(*a*:3) (*b*:4) (*d*:2) (*e*:6)], [(*b*:3) (*c*:2)]>
S_5	5/7/2017 16:58	Supermarket	***	Female	Old	Artist	<[(*e*:4)], [(*d*:7)], [(*c*:5)], [(*a*:9) (*b*:3) (*c*:7) (*d*:7)]>

Transfer to the binary representation of the categorical data

SID	Q-sequence
S_1	<[(Male:0)(Young:0)(Doctor:0)], [(*a*:1) (*c*:3)], [(*a*:5) (*c*:1) (*e*:4)], [(*c*:2)], [(*b*:1)]>
S_2	<[(Female:0)(Middle:0)(Lawyer:0)], [(*c*:1)], [(*b*:4)], [(*b*:9) (*d*:8)], [(*b*:9) (*e*:6)]>
S_3	<[(Male:0)(Child:0)(Driver:0)], [(*a*:10) (*d*:5)]>
S_4	<[(Male:0)(Young:0)(Writer:0)], [(*a*:3) (*b*:4) (*d*:2) (*e*:6)] [(*b*:3) (*c*:2)]>
S_5	<[(Female:0)(Old:0)(Artist:0)], [(*e*:4)], [(*d*:7)], [(*c*:5)], [(*a*:9) (*b*:3) (*c*:7) (*d*:7)]>

Techniques

- Take the sequence data into the account in the mining progress
  - Lexicographic (LS)-tree
- Projection model to clearly discover the required patterns
  - Simplified the database
- 2 theorems presented here to maintain the correctness and its mining efficiency

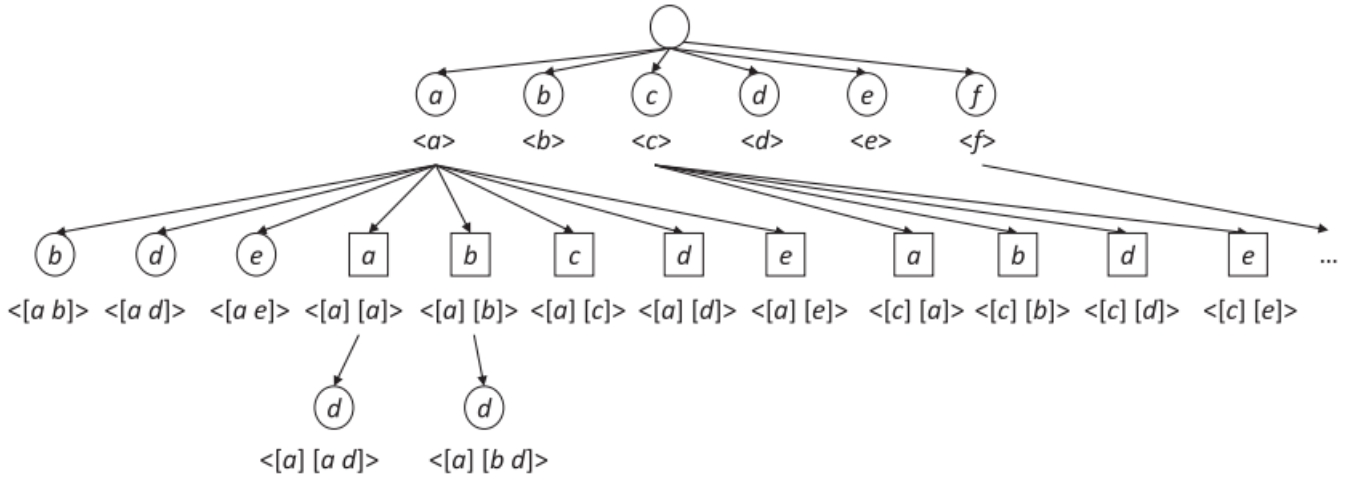


Table 4. The Dimensional Database of <a>

SID	Transaction	*TU*
S_1	(Male Young Doctor)	$20
S_2	(Female Middle Lawyer)	$0
S_3	(Male Child Driver)	$40
S_4	(Male Young Writer)	$12
S_5	(Female Old Artist)	$36

Data level (3) – uncertain data

1. Wensheng Gan, **Jerry Chun-Wei Lin***, Han-Chieh Chao, Philippe Fournier-Viger, XuanWang, and Philip S. Yu, “Utility-Driven Mining of Trend Information for Intelligent System,” **ACM Transactions on Management Information Systems**, Vol. 11(3), Article No.14, 2020

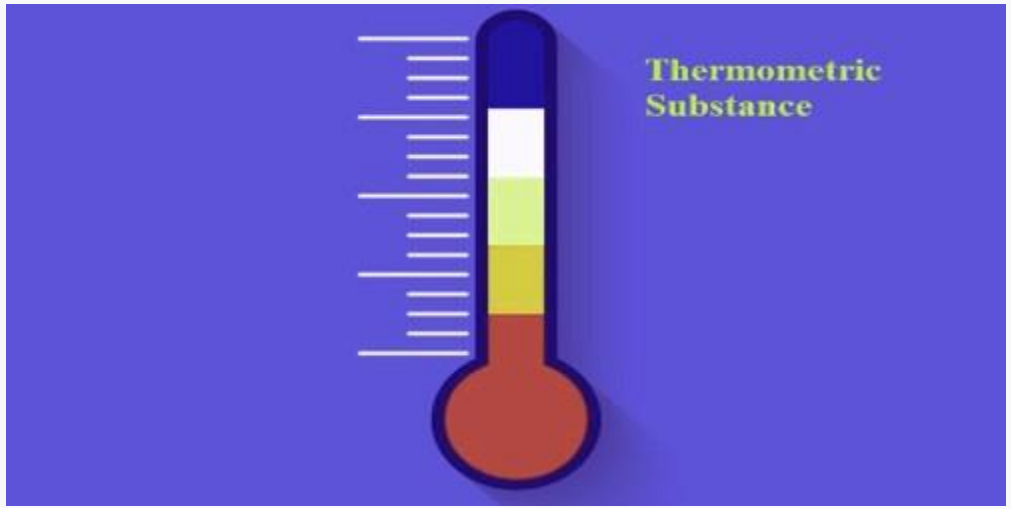
2. Wensheng Gan, **Jerry Chun-Wei Lin***, Jiexiong Zhang, Hongzhi Yin, Philippe Fournier-Viger, Han-Chieh Chao, and Philip S. Yu, “Mining Across Multi-Dimensional Sequences,” **ACM Transactions on Knowledge Discovery from Data**, Vol.15(5), Article No.:82, 2021 (SCI, JCR:Q2, IF:2.713)

3. Wensheng Gan, **Jerry Chun-Wei Lin**, Han-Chieh Chao, Athanasios V. Vasilakos, and Philip S. Yu, “Utility-driven Data Analytics on Uncertain Data,” **IEEE Systems Journal**, Vol. 14(3), pp. 4442–4453, 2020 (SCI, JCR:Q2, IF:3.921)

4. Gautam Srivastava, **Jerry Chun-Wei Lin***, Alireza Jolfaei, Yuanfa Li, and Youcef Djenouri, “Uncertain-Driven Analytics of Sequence Data in IoCV Environments,” **IEEE Transactions on Intelligent Transportation Systems**, Vol. 22(8), pp. 5403–5414, 2021 (SCI, JCR:Q1, IF:6.492)

Uncertain data

- Data is not always precise or clean
  - Probability of temperature
    - Close to window
    - Close to heater



tid	Item: quantity, probability)	total utility
T_1	(a:5, 0.6); (b:3, 0.50); (d:2, 0.9); (e:4, 0.8)	\$107
T_2	(c:1, 0.75); (d:1, 0.9); (e:2, 1.0)	\$24
T_3	(a:4, 1.0); (b:3, 1.0); (c:2, 0.7); (e:1, 0.75)	\$50
T_4	(a:3, 0.9); (c:1, 0.9)	\$22
T_5	(b:2, 1.0); (c:4, 0.95); (d:5, 0.6); (e:4, 1.0)	\$90

Techniques

- 6 pruning strategies to reduce the size of the promising patterns
  - Unpromising ones will be removed
- EUCS (estimated utility co-occurrence structure) helps to remove k>=3-patterns

(a)

tid	pro	pu	nu	rpu
T_1	0.6	40	0	67
T_3	1.0	32	0	22
T_4	0.9	24	0	0

(d)

Tid	pro	pu	nu	rpu
T_1	0.9	24	0	43
T_2	0.9	12	0	14
T_5	0.6	60	0	38

(b)

tid	pro	pu	nu	rpu
T_1	0.5	15	0	28
T_3	1.0	15	0	7
T_5	1.0	10	0	28

(e)

tid	pro	pu	nu	rpu
T_1	0.8	28	0	0
T_2	1.0	14	0	0
T_3	0.75	7	0	0
T_5	1.0	28	0	0

(c)

tid	pro	pu	nu	rpu
T_2	0.75	0	-2	0
T_3	0.7	0	-4	0
T_4	0.9	0	-2	0
T_5	0.95	0	-8	0

Fig. 2. Constructed $PU^{\pm}$-list of the running example.

Item	a	b	c	d	e	f
b	30					
c	65	61				
d	38	50	58			
e	57	61	77	50		
f	30	30	30	30	30	
g	27	38	38	0	38	0

Data level (4) – sequential data

1. Wensheng Gan, **Jerry Chun-Wei Lin***, Han-Chieh Chao, Philippe Fournier-Viger, XuanWang, and Philip S. Yu, “Utility-Driven Mining of Trend Information for Intelligent System,” **ACM Transactions on Management Information Systems**, Vol. 11(3), Article No.14, 2020

2. Wensheng Gan, **Jerry Chun-Wei Lin***, Jiexiong Zhang, Hongzhi Yin, Philippe Fournier-Viger, Han-Chieh Chao, and Philip S. Yu, “Mining Across Multi-Dimensional Sequences,” **ACM Transactions on Knowledge Discovery from Data**, Vol.15(5), Article No.:82, 2021 (SCI, JCR:Q2, IF:2.713)

3. Wensheng Gan, **Jerry Chun-Wei Lin**, Han-Chieh Chao, Athanasios V. Vasilakos, and Philip S. Yu, “Utility-driven Data Analytics on Uncertain Data,” **IEEE Systems Journal**, Vol. 14(3), pp. 4442–4453, 2020 (SCI, JCR:Q2, IF:3.921)

4. Gautam Srivastava, **Jerry Chun-Wei Lin***, Alireza Jolfaei, Yuanfa Li, and Youcef Djenouri, “Uncertain-Driven Analytics of Sequence Data in IoCV Environments,” **IEEE Transactions on Intelligent Transportation Systems**, Vol. 22(8), pp. 5403–5414, 2021 (SCI, JCR:Q1, IF:6.492)

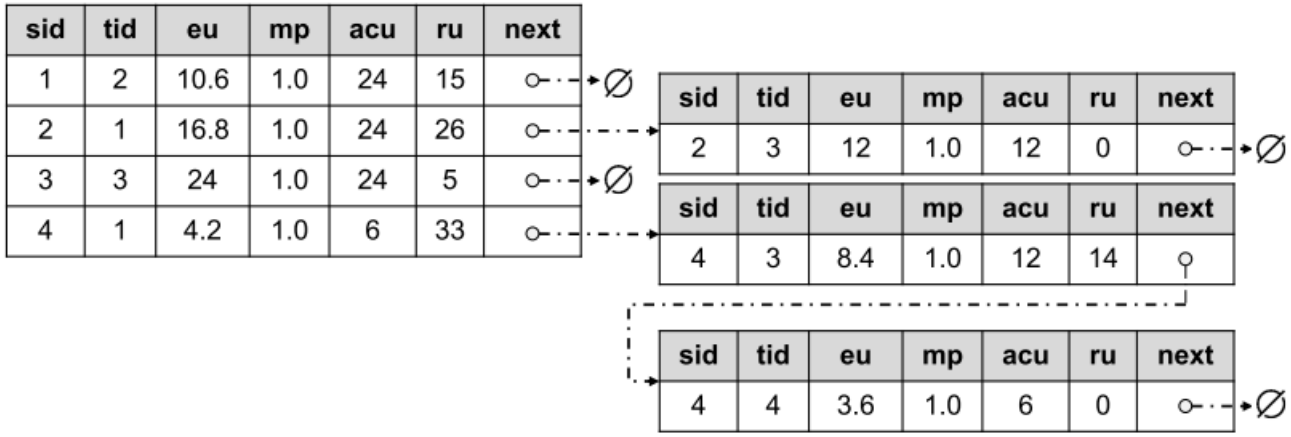
Sequence data

- Sequence plays an important rule in pattern analysis
  - DNA analysis, Web usage analysis…etc.
- Also consider the uncertain characteristic of the data

sid	sequence
s_1	$<[(a,3,0.9)],[(b,4,0.45)],[(a,1,0.7)(c,1,0.6)(e,2,1.0)]>$
s_2	$<[(b,4,0.7)(c,1,0.45)],[(a,1,0.6)(d,1,0.5)],[(a,2,0.9)(b,2,1.0)]>$
s_3	$<[(e,1,1.0)],[(c,1,0.7)(d,1,0.6)],[(b,4,1.0)(c,1,0.7)]>$
s_4	$<[(b,1,0.7)],[(c,1,0.9)],[(a,1,1.0)(b,2,0.7)],[(a,4,0.45)(b,1,0.6)]>$

Techniques

- Satisfy two criteria
  - Utility threshold
  - Uncertain threshold
- 2 algorithms with 2 data chain structures
- 4 theorems to hold the correctness
  - Upper-bound values
- 6 pruning strategies



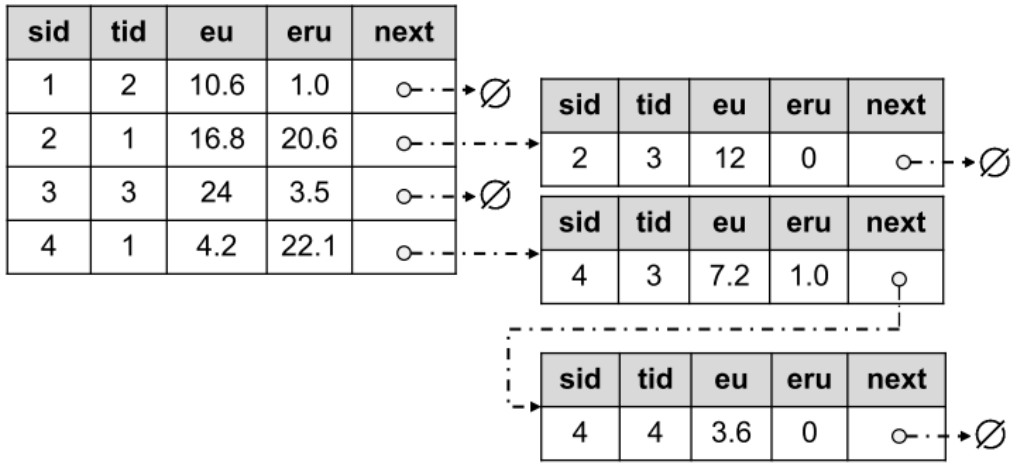
sid	tid	eu	mp	acu	ru	next
1	2	10.6	1.0	24	15	∅
2	1	16.8	1.0	24	26	
3	3	24	1.0	24	5	∅
4	1	4.2	1.0	6	33	

sid	tid	eu	mp	acu	ru	next
2	3	12	1.0	12	0	∅

sid	tid	eu	mp	acu	ru	next
4	3	8.4	1.0	12	14	

sid	tid	eu	mp	acu	ru	next
4	4	3.6	1.0	6	0	∅

Fig. 2. The PUL-Chain of $< b >$.



sid	tid	eu	eru	next
1	2	10.6	1.0	∅
2	1	16.8	20.6	
3	3	24	3.5	∅
4	1	4.2	22.1	

sid	tid	eu	eru	next
2	3	12	0	∅

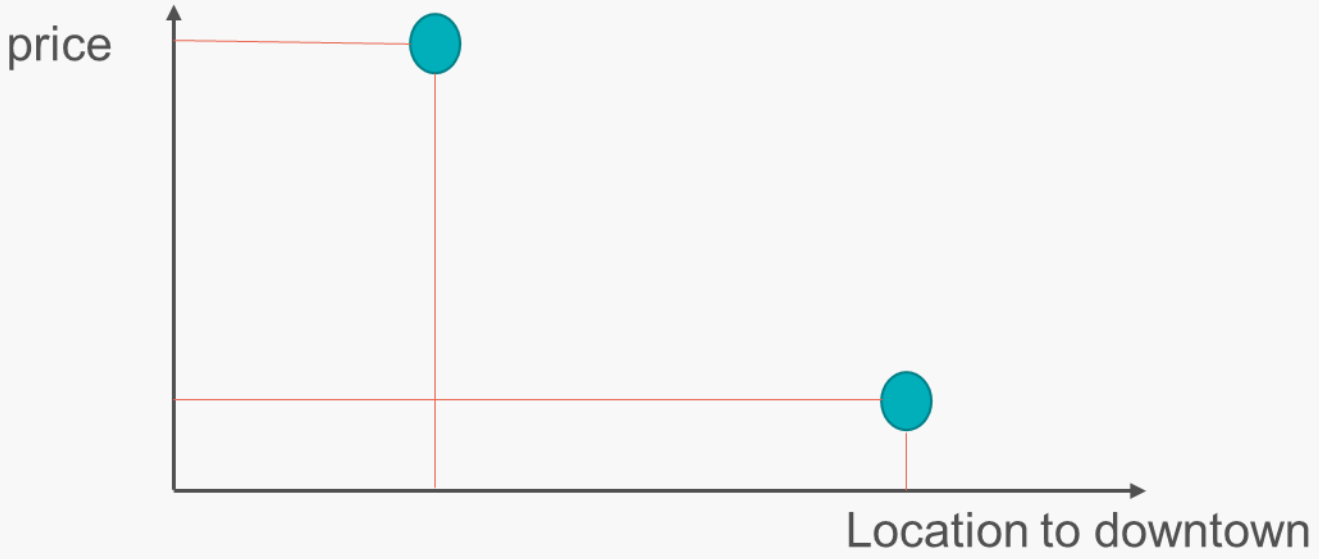
sid	tid	eu	eru	next
4	3	7.2	1.0	

sid	tid	eu	eru	next
4	4	3.6	0	∅

Fig. 3. The EUL-Chain of $< b >$.

3 different proposed models

- Constraint-based model
  - Individualized thresholds evaluation
  - Multi-objective mining progress
  - occupation constraint (other extensions…)



Constraint-based level (1) – individualized threshold

1. Wensheng Gan, **Jerry Chun-Wei Lin***, Jiexiong Zhang, and Philip S. Yu, “Utility Mining Across Multi-Sequences with Individualized Thresholds,” **ACM Transactions on Data Science**, Vol. 1(2), Article No 8, 2020

2. Wensheng Gan, **Jerry Chun-Wei Lin***, Philippe Fournier-Viger, Han-Chieh Chao, and Philip S. Yu, “Beyond Frequency: Utility Mining with Varied Item-Specific Minimum Utility,” **ACM Transactions on Internet Technology**, Vol. 21(1), Article No. 3, 2021 (SCI, JCR:Q1, IF:3.135)

3. Jimmy Ming-Tai Wu, Qian Teng, Gautam Srivastava, Matin Pirouz, and **Jerry Chun-Wei Lin***, “The Efficient Mining of Skyline Patterns from a Volunteer Computing Network,” **ACM Transactions on Internet Technology**, Vol. 21(4), Article No.:89, 2021, (SCI, JCR:Q1, IF:3.135)

4. Wensheng Gan, **Jerry Chun-Wei Lin***, Philippe Fournier-Viger, Han-Chieh Chao, and Philip S. Yu, “HUOPM: High Utility Occupancy Pattern Mining,” **IEEE Transactions on Cybernetics**, Vol. 50(3), pp. 1195–1208, 2019 (SCI, JCR:Q1, IF:11.448)

Individualized threshold concept

- Single minimum utility threshold
  - Assume all items have the same nature and/or have similar utilities
- Not true
  - In many applications, some items bring very high utility in the data, while others utilities are relatively lower

- Rare item problem:
  - If minUtil is set **too high**, those HUIs that involve relatively low-utility items will not be found. To find HUIs that involve both high-utility and relatively low-utility items, minUtil has to be **set very low**. This may cause **combinatorial explosion problem**

Techniques

- Different items cannot be treated with an uniform threshold for the evaluation
- A MMU-table
  - $MMU\text{-}table = \{mu(a), mu(b), mu(c), mu(d), mu(e)\} = \{\$56, \$65, \$53, \$50, \$70\}.$
-

TID	Time	Customer ID	Event (product: quantity)
T_1	05-08-2017 10:45:30	C_1	*a*:1, *c*:2, *d*:3
T_2	05-08-2017 10:59:12	C_2	*a*:2, *d*:1, *e*:2
T_3	05-08-2017 11:05:40	C_3	*b*:3, *c*:5
T_4	05-08-2017 11:40:00	C_4	*a*:1, *c*:3, *d*:1, *e*:2
T_5	05-08-2017 12:55:14	C_3	*b*:1, *d*:3, *e*:2
T_6	05-08-2017 14:08:58	C_2	*b*:2, *d*:2
T_7	05-08-2017 14:40:00	C_5	*b*:3, *c*:2, *d*:1, *e*:1
T_8	05-08-2017 15:01:40	C_4	*a*:2, *c*:3
T_9	05-08-2017 15:04:26	C_6	*c*:2, *d*:2, *e*:1
T_{10}	05-08-2017 15:30:20	C_7	*a*:2, *c*:2, *d*:1

Technique (1)

- A utility-array is used to keep more information
- A LS-tree is applied here for handling the sequence issue

Table 3. The Utility-array Structure of S_3 (Notation "-" Means Empty)

	eid	*item*	*u*	*ru*	*next_pos*	*next_eid*
$array_1$	1	*a*	$12	$84	3	3
$array_2$	1	*b*	$10	$72	4	3
$array_3$	2	*a*	$8	$64	-	6
$array_4$	2	*b*	$15	$49	6	6
$array_5$	2	*c*	$3	$46	7	6
$array_6$	3	*b*	$20	$26	-	9
$array_7$	3	*c*	$15	$11	-	9
$array_8$	3	*e*	$8	$3	-	9
$array_9$	4	*d*	$3	$0	-	-

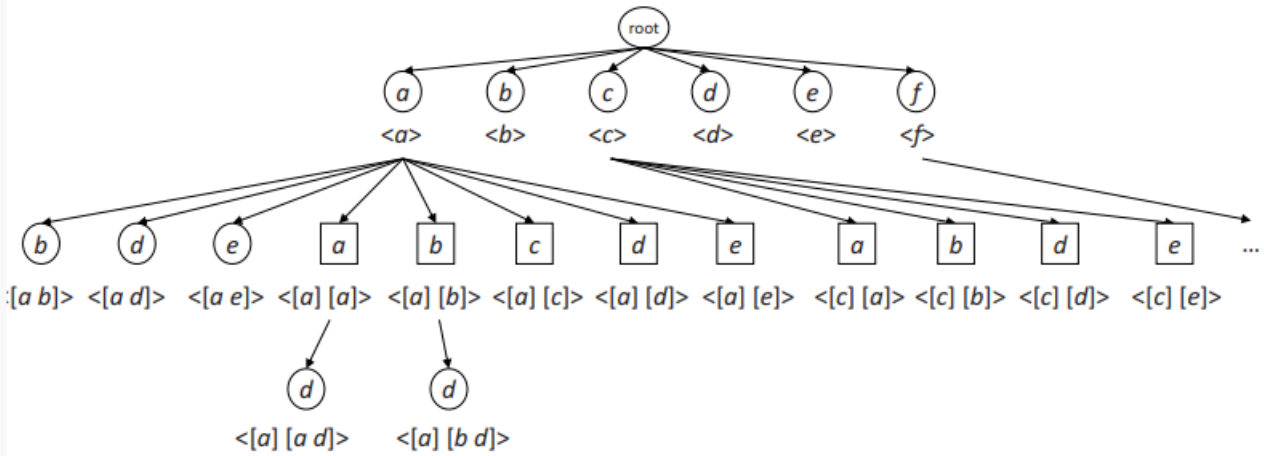


Fig. 1. A lexicographic-sequential (LS)-tree [34].

Technique (2)

- Link-list structure
- EUCS
  - Estimated utility co-occurrence structure
- 2 pruning strategies
  - Based on TWDC

(d)

tid	iu	ru
1	27	8
2	9	18
4	9	15
5	27	42
6	18	24
7	9	41
9	18	5
10	9	14

(c)

tid	iu	ru
1	2	6
3	5	12
4	3	12
7	2	39
8	3	12
9	2	3
10	2	12

(a)

tid	iu	ru
1	6	0
2	12	6
4	6	6
8	12	0
10	12	0

(b)

tid	iu	ru
3	12	0
5	36	6
6	24	0
7	36	3

(e)

tid	iu	ru
2	6	0
4	6	0
5	6	0
7	3	0
9	3	0

Item	*a*	*b*	*c*	*d*	*e*
b	$42	—	—	—	—
c	$175	$27	—	—	—
d	$179	$80	$171	—	—
e	$42	$78	$61	$76	—

Constraint-based level (2) – skyline

1. Wensheng Gan, **Jerry Chun-Wei Lin***, Jiexiong Zhang, and Philip S. Yu, “Utility Mining Across Multi-Sequences with Individualized Thresholds,” **ACM Transactions on Data Science**, Vol. 1(2), Article No 8, 2020

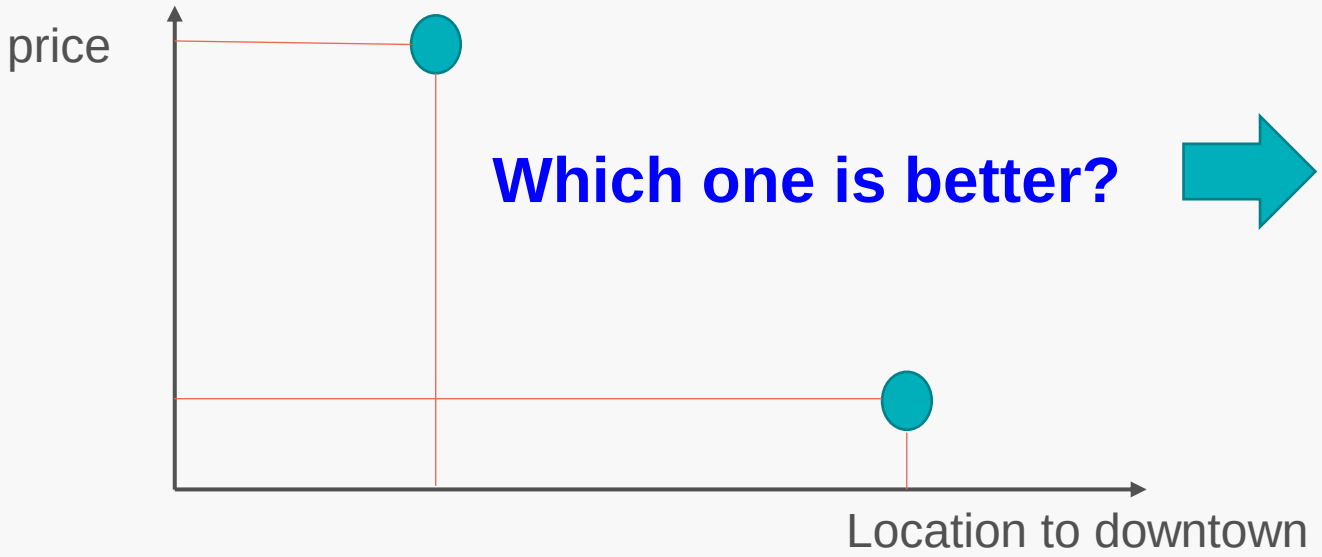
2. Wensheng Gan, **Jerry Chun-Wei Lin***, Philippe Fournier-Viger, Han-Chieh Chao, and Philip S. Yu, “Beyond Frequency: Utility Mining with Varied Item-Specific Minimum Utility,” **ACM Transactions on Internet Technology**, Vol. 21(1), Article No. 3, 2021 (SCI, JCR:Q1, IF:3.135)

3. Jimmy Ming-Tai Wu, Qian Teng, Gautam Srivastava, Matin Pirouz, and **Jerry Chun-Wei Lin***, “The Efficient Mining of Skyline Patterns from a Volunteer Computing Network,” **ACM Transactions on Internet Technology**, Vol. 21(4), Article No.:89, 2021, (SCI, JCR:Q1, IF:3.135)

4. Wensheng Gan, **Jerry Chun-Wei Lin***, Philippe Fournier-Viger, Han-Chieh Chao, and Philip S. Yu, “HUOPM: High Utility Occupancy Pattern Mining,” **IEEE Transactions on Cybernetics**, Vol. 50(3), pp. 1195–1208, 2019 (SCI, JCR:Q1, IF:11.448)

Multi-objective mining progress

- It is possible to consider more than one factor
  - Instead of utility, frequency or other factors (i.e., uncertainty) is also interesting to be considered in the mining progress
- Sometimes, the objectives could have the trade-off relationship
  - Cost and distance to the city

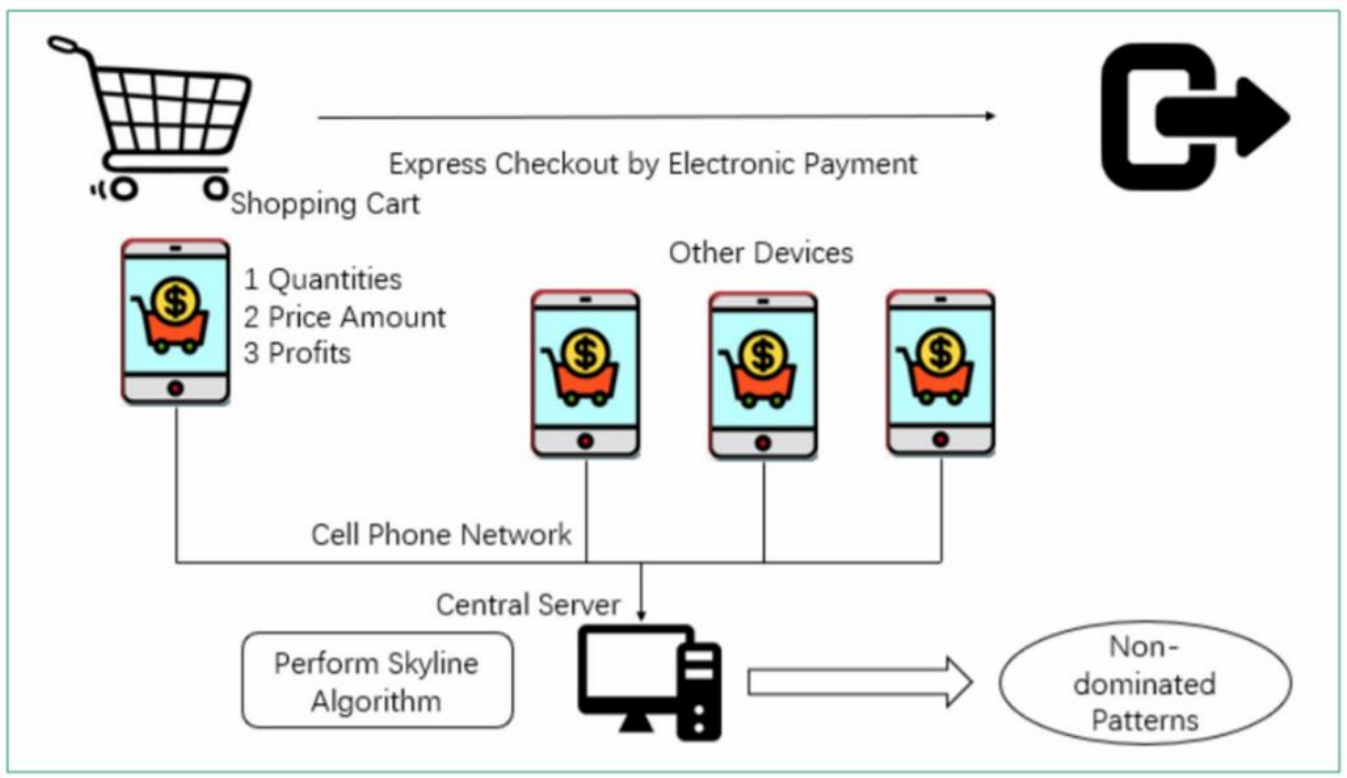


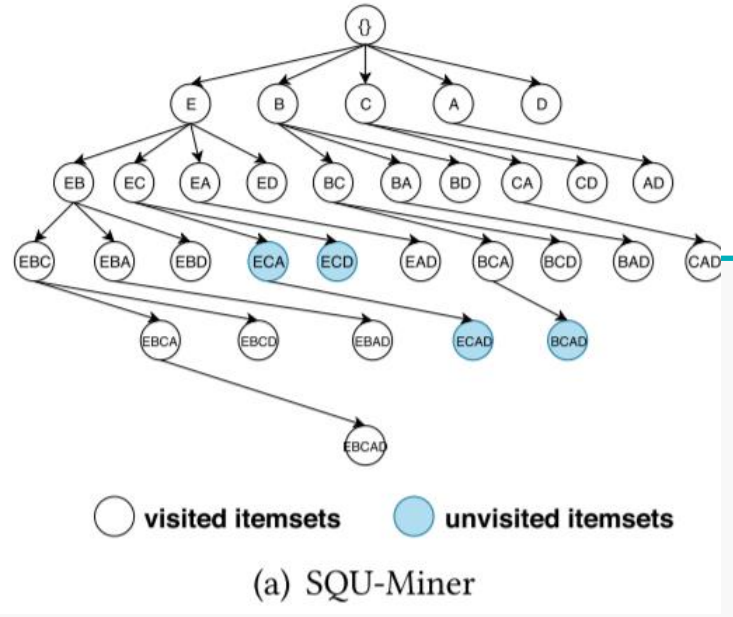
Which one is better?

Set of skyline solutions

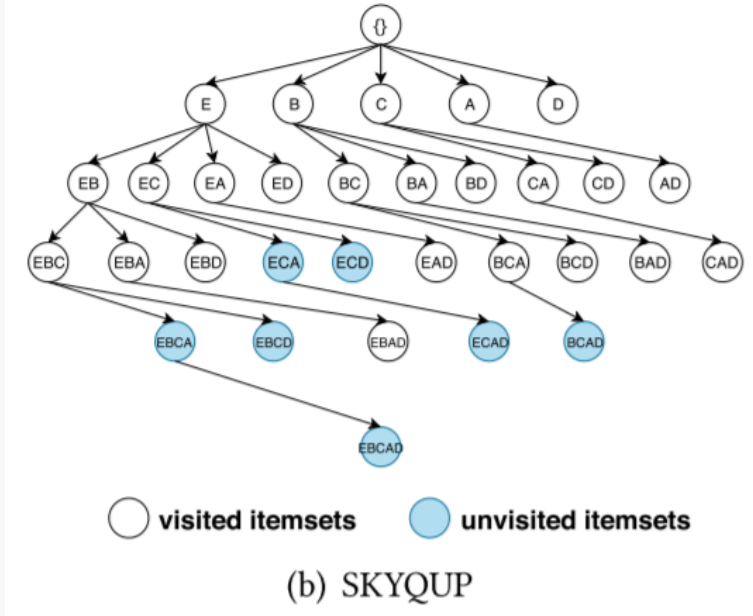
Techniques

- Utility and quantity are both considered





(a) SQU-Miner



(b) SKYQUP

Constraint-based level (3) – occupation

1. Wensheng Gan, **Jerry Chun-Wei Lin***, Jiexiong Zhang, and Philip S. Yu, “Utility Mining Across Multi-Sequences with Individualized Thresholds,” **ACM Transactions on Data Science**, Vol. 1(2), Article No 8, 2020

2. Wensheng Gan, **Jerry Chun-Wei Lin***, Philippe Fournier-Viger, Han-Chieh Chao, and Philip S. Yu, “Beyond Frequency: Utility Mining with Varied Item-Specific Minimum Utility,” **ACM Transactions on Internet Technology**, Vol. 21(1), Article No. 3, 2021 (SCI, JCR:Q1, IF:3.135)

3. Jimmy Ming-Tai Wu, Qian Teng, Gautam Srivastava, Matin Pirouz, and **Jerry Chun-Wei Lin***, “The Efficient Mining of Skyline Patterns from a Volunteer Computing Network,” **ACM Transactions on Internet Technology**, Vol. 21(4), Article No.:89, 2021, (SCI, JCR:Q1, IF:3.135)

4. Wensheng Gan, **Jerry Chun-Wei Lin***, Philippe Fournier-Viger, Han-Chieh Chao, and Philip S. Yu, “HUOPM: High Utility Occupancy Pattern Mining,” **IEEE Transactions on Cybernetics**, Vol. 50(3), pp. 1195–1208, 2019 (SCI, JCR:Q1, IF:11.448)

Occupation

- Occupation
  - On which mobile Apps do the customers spend most of their data traffic
- Utility occupation
  - In a purchase behavior (i.e., a transaction), which item does the customer spend most of the cost?
  - **u(X, Tq)/tu(Tq) in a range of [0, 1]**
  - Instead of the only high-utility measure

Which categorical app you spend the most money

Techniques

- Based on the occupation, several strategies are designed
- UO-list + FU-table
- 3 pruning strategies

UO-list

(e)		
tid	*uo*	*ruo*
5	0.1837	0.8163
6	0.6207	0.3793
8	0.5902	0.4098
10	0.2143	0.7857

(a)

FU-table

(e)		
sup	*uo*	*ruo*
4	0.4022	0.5978

$uo(e) = (0.1837 + 0.6207 + 0.5902 + 0.2143)/4 = 0.4022;$
$ruo(e) = (0.8163 + 0.3793 + 0.4098 + 0.7857)/4 = 0.5978$

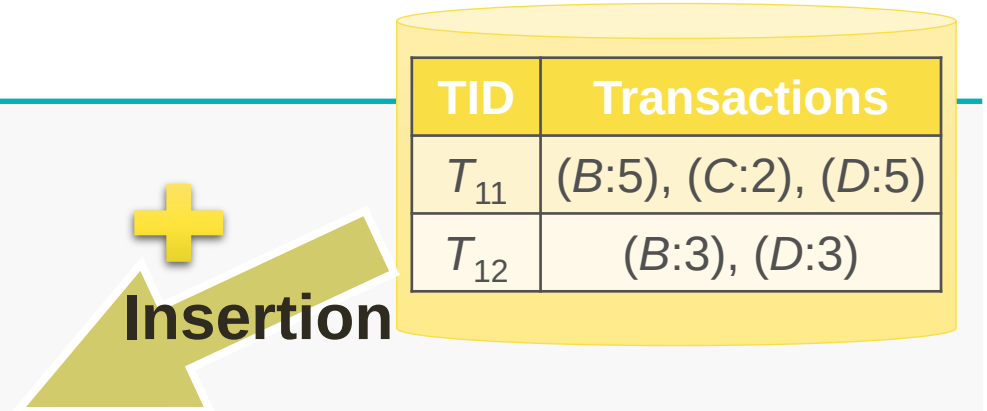
(b)

TABLE II
DERIVED HUOPS

Pattern	*sup*	*uo*	Pattern	*sup*	*uo*
(c)	8	0.6468	(bd)	4	0.3620
(e)	4	0.4022	(cd)	5	0.6881
(ab)	3	0.4334	(ce)	3	0.8776
(ac)	4	0.8273	(abd)	3	0.4959
(ad)	5	0.3609	(acd)	4	0.8972
(bc)	3	0.6554			

Big and Dynamic Models

- Data is dynamic changed
  - Insert, delete or modify
- **Data is a streaming type**
- **Data is distributed**
- **Large-scale data**



TID	Transactions
T_{11}	(*B*:5), (*C*:2), (*D*:5)
T_{12}	(*B*:3), (*D*:3)

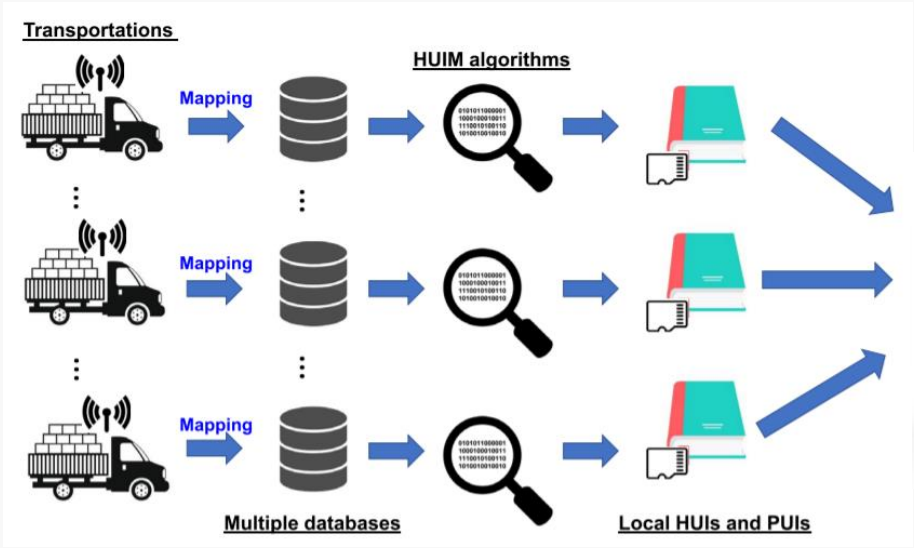
TID	Transactions
T_1	(*A*:6)
T_2	(*B*:1), (*D*:5)
…	…….
T_9	(*A*:2), (*C*:2), (*E*:3)
T_{10}	(*B*:4)

A	*B*	*C*	*D*	*E*
$6	$2	$15	$7	$10

3 different proposed models

- Big and dynamic data models
  - Streaming data
  - Large-scale data
  - Data fusion

		TID	Transaction	TU
W_1	B_1	1	(A, 2) (C, 1) (E, 1) (F, 1)	20
		2	(B, 3) (C, 1)	12
W_2	B_2	3	(C, 1) (D, 3) (F, 2)	19
		4	(A, 2) (B, 2) (D, 1)	15
	B_3	5	(A, 2) (D, 1) (E, 1) (F, 3)	21
		6	(B, 3) (E, 3)	18
	B_4	7	(A, 3) (D, 1) (F, 1)	17
		8	(B, 1) (C, 3) (E, 1)	24



Big and Dynamic Models (1) – dynamic data

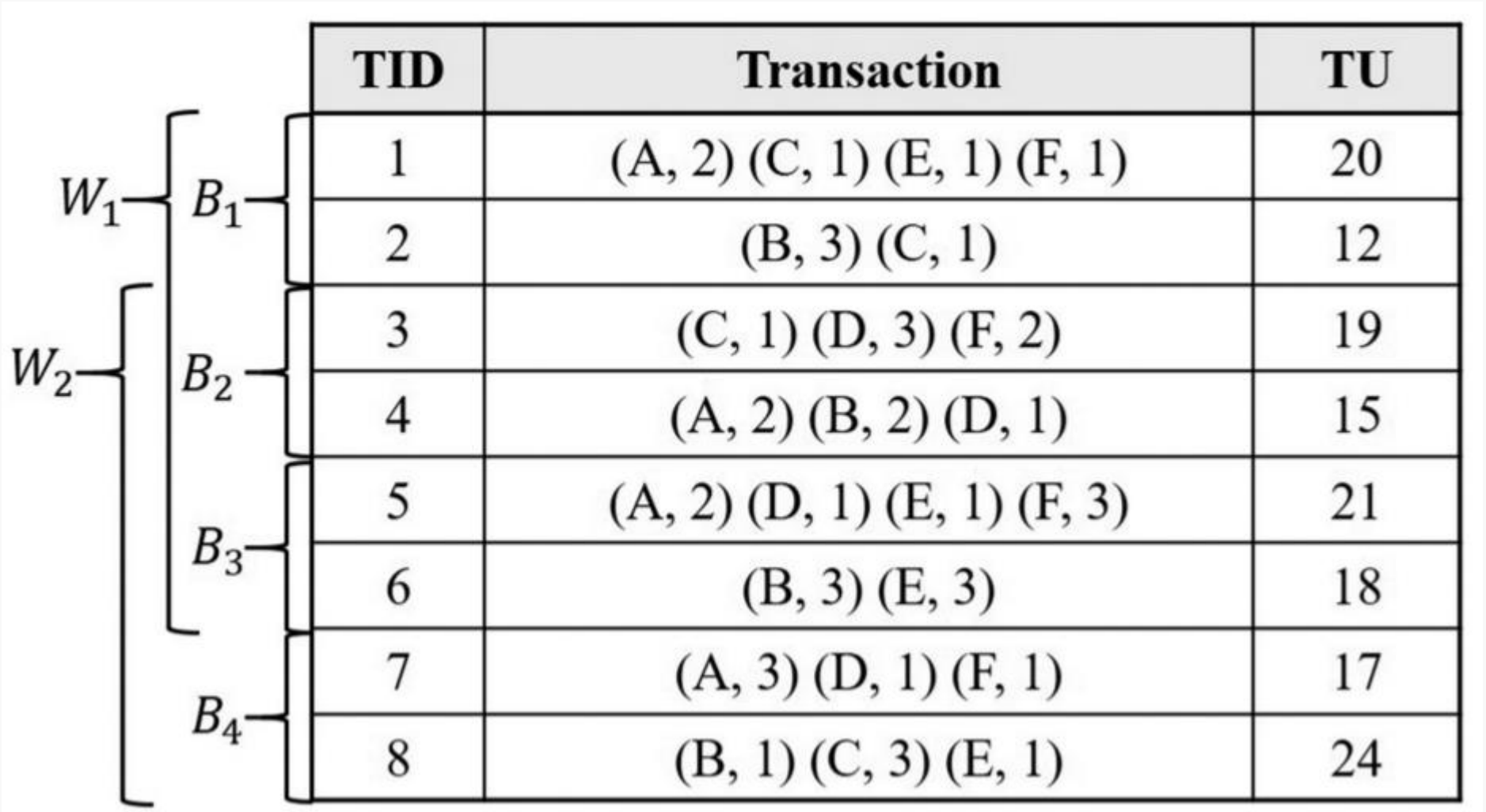
1. Yoonji Baek, Unil Yum, Heonho Kim, Hyoju Nam, Hyunsoo Kim, **Jerry Chun-Wei Lin**, Bay Vo, and Witold Pedrycz, “RHUPs: Mining Recent High Utility Patterns with Sliding Window based Arrival Time Control over Data Streams,” **ACM Transactions on Intelligent Systems and Technology**, Vol. 12(2), Article No. 16, 2021 (SCI, JCR:Q1, IF:4.654)

2. Gautam Srivastava, **Jerry Chun-Wei Lin***, Xuyun Zhang, and Yuanfa Li, “Large-Scale High-Utility Sequential Pattern Analytics in Internet of Things,” **IEEE Internet of Things Journal**, Vol. 8(16), pp. 12669–12678, 2021 (SCI, JCR:Q1, IF:9.471)

3. **Jerry Chun-Wei Lin***, Youcef Djenouri, Gautam Srivastava, Yuanfa Li, and Philp S. Yu, “Scalable Mining of High-Utility Sequential Patterns with Three-Tier MapReduce Model,” **ACM Transactions on Knowledge Discovery from Data**, 2021 (SCI, JCR:Q2, IF:2.713)

4. Gautam Srivastava, **Jerry Chun-Wei Lin***, Matin Pirouz, Yuanfa Li, and Until Yun, “A Pre-large Weighted-Fusion System of Sensed High-Utility Patterns,” **IEEE Sensors Journal**, Vol. 21(14), pp. 15626–15634, 2021 (SCI, JCR:Q2, IF:3.301)

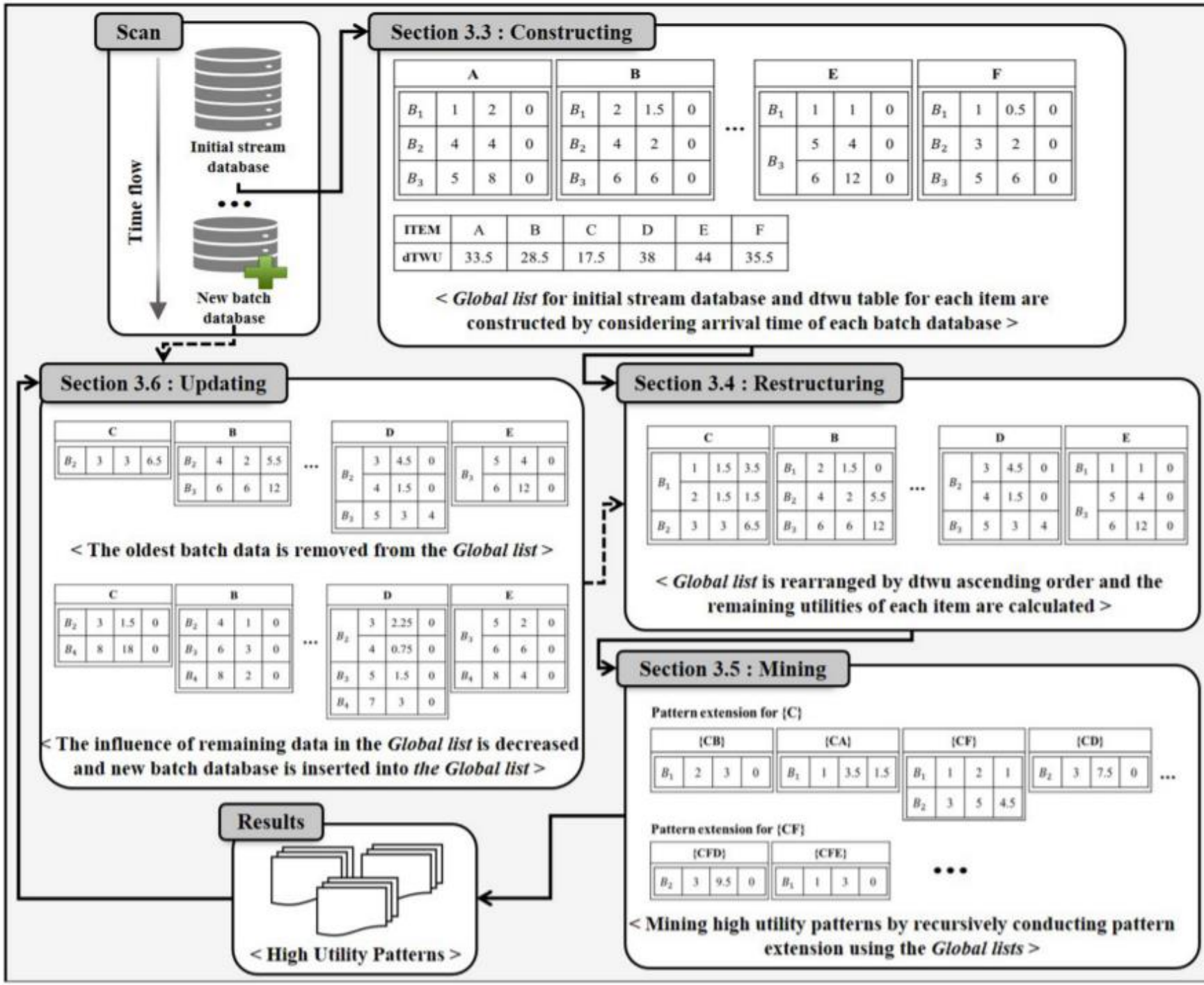
Data stream



TID	Transaction	TU
1	(A, 2) (C, 1) (E, 1) (F, 1)	20
2	(B, 3) (C, 1)	12
3	(C, 1) (D, 3) (F, 2)	19
4	(A, 2) (B, 2) (D, 1)	15
5	(A, 2) (D, 1) (E, 1) (F, 3)	21
6	(B, 3) (E, 3)	18
7	(A, 3) (D, 1) (F, 1)	17
8	(B, 1) (C, 3) (E, 1)	24

Techniques

- Damped factor for old and new slide window
- Update the built structure accordingly



Big and Dynamic Models (2) – large-scale data

1. Yoonji Baek, Unil Yum, Heonho Kim, Hyoju Nam, Hyunsoo Kim, **Jerry Chun-Wei Lin**, Bay Vo, and Witold Pedrycz, “RHUPs: Mining Recent High Utility Patterns with Sliding Window based Arrival Time Control over Data Streams,” **ACM Transactions on Intelligent Systems and Technology**, Vol. 12(2), Article No. 16, 2021 (SCI, JCR:Q1, IF:4.654)

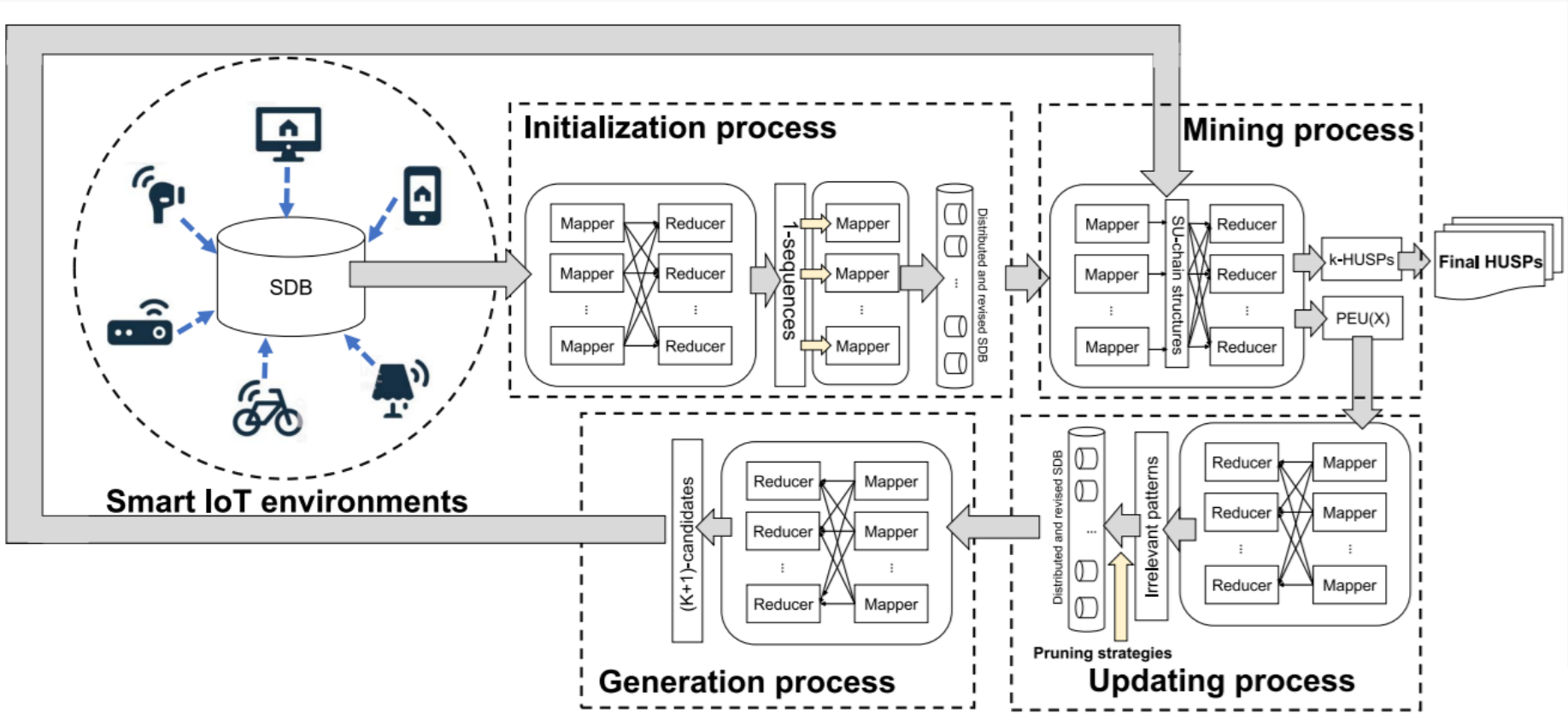
2. Gautam Srivastava, **Jerry Chun-Wei Lin***, Xuyun Zhang, and Yuanfa Li, “Large-Scale High-Utility Sequential Pattern Analytics in Internet of Things,” **IEEE Internet of Things Journal**, Vol. 8(16), pp. 12669–12678, 2021 (SCI, JCR:Q1, IF:9.471)

3. **Jerry Chun-Wei Lin***, Youcef Djenouri, Gautam Srivastava, Yuanfa Li, and Philp S. Yu, “Scalable Mining of High-Utility Sequential Patterns with Three-Tier MapReduce Model,” **ACM Transactions on Knowledge Discovery from Data**, 2021 (SCI, JCR:Q2, IF:2.713)

4. Gautam Srivastava, **Jerry Chun-Wei Lin***, Matin Pirouz, Yuanfa Li, and Until Yun, “A Pre-large Weighted-Fusion System of Sensed High-Utility Patterns,” **IEEE Sensors Journal**, Vol. 21(14), pp. 15626–15634, 2021 (SCI, JCR:Q2, IF:3.301)

Techniques (1)

- MapReduce structure to handle the IoT data
  - Apriori-like MapReduce model



Techniques (2)

- 3-tier structure for the high-utility sequential pattern mining
  - List-based model for performance improvement

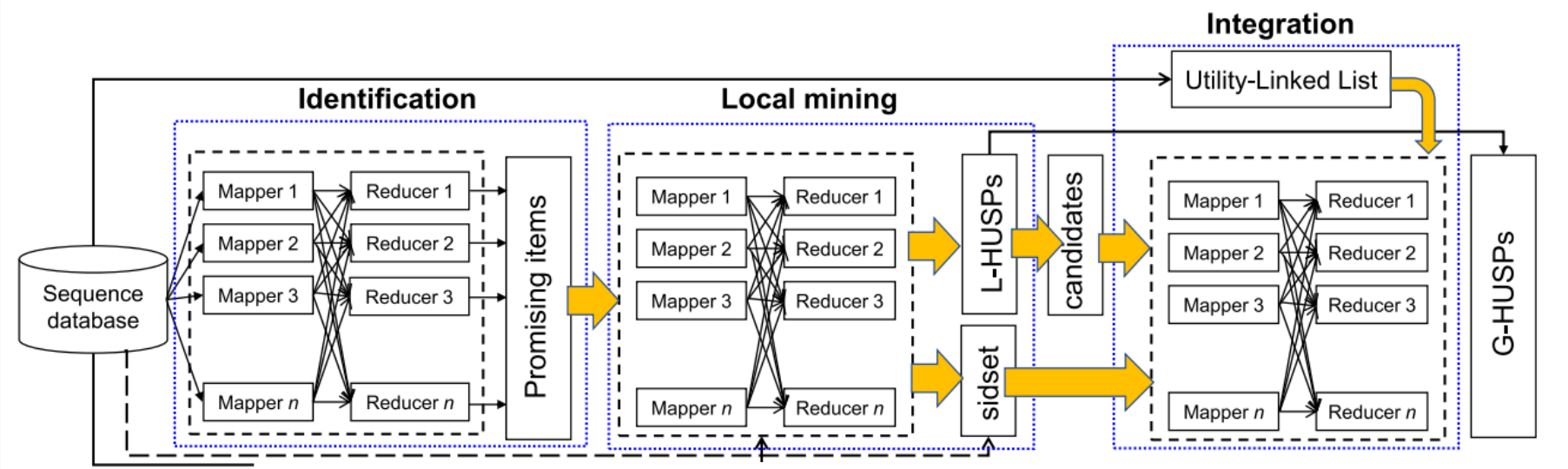


Table 5. The Utility-linked List of s_1

UP Information	<[(*a*, 10, 3) (*c*, 12, 5)], [(*a*, 15, 6) (*b*, 3, 7) (*c*, 8, -)], [(*a*, 20, -) (*b*, 15, -) (*d*, 8, -)], [*e*, 3, -]>
Header Table	(*a*, 1) (*b*, 4) (*c*, 2) (*d*, 8) (*e*, 9)

Big and Dynamic Models (3) – data fusion

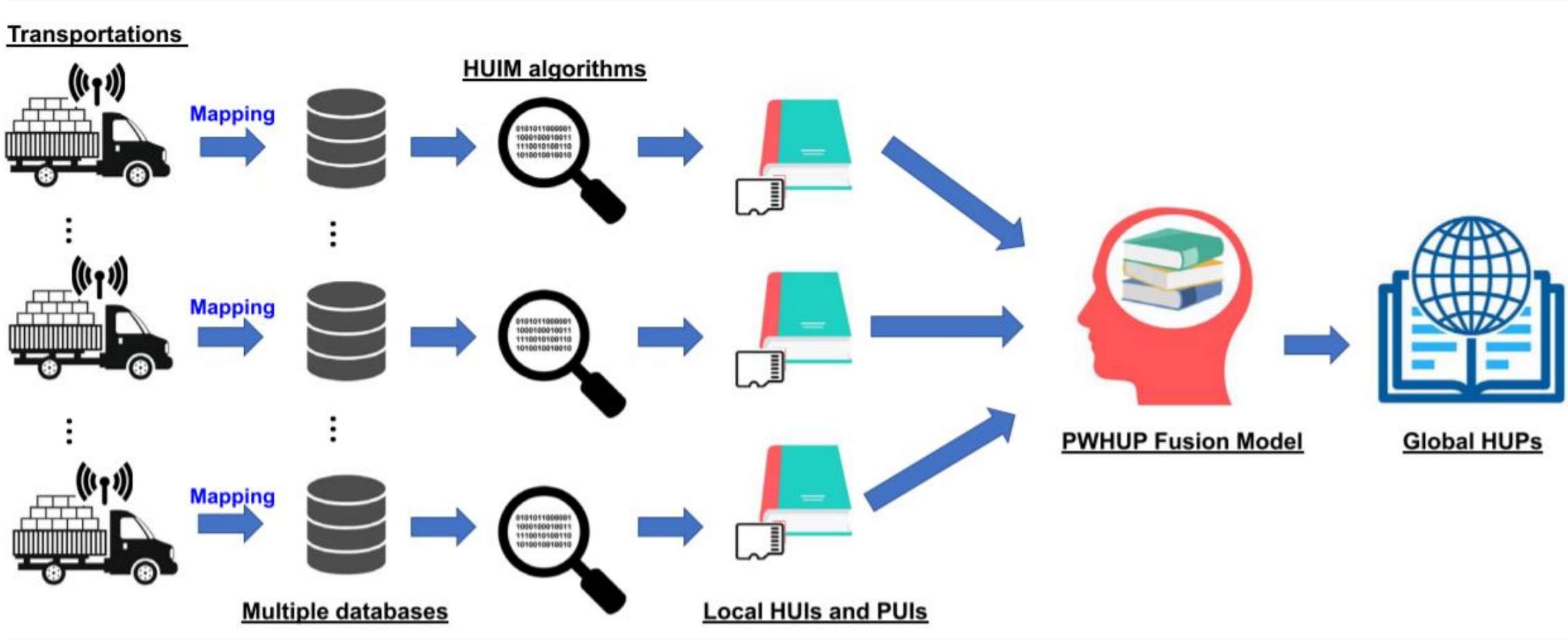
1. Yoonji Baek, Unil Yum, Heonho Kim, Hyoju Nam, Hyunsoo Kim, **Jerry Chun-Wei Lin**, Bay Vo, and Witold Pedrycz, “RHUPs: Mining Recent High Utility Patterns with Sliding Window based Arrival Time Control over Data Streams,” **ACM Transactions on Intelligent Systems and Technology**, Vol. 12(2), Article No. 16, 2021 (SCI, JCR:Q1, IF:4.654)

2. Gautam Srivastava, **Jerry Chun-Wei Lin***, Xuyun Zhang, and Yuanfa Li, “Large-Scale High-Utility Sequential Pattern Analytics in Internet of Things,” **IEEE Internet of Things Journal**, Vol. 8(16), pp. 12669–12678, 2021 (SCI, JCR:Q1, IF:9.471)

3. **Jerry Chun-Wei Lin***, Youcef Djenouri, Gautam Srivastava, Yuanfa Li, and Philp S. Yu, “Scalable Mining of High-Utility Sequential Patterns with Three-Tier MapReduce Model,” **ACM Transactions on Knowledge Discovery from Data**, 2021 (SCI, JCR:Q2, IF:2.713)

4. Gautam Srivastava, **Jerry Chun-Wei Lin***, Matin Pirouz, Yuanfa Li, and Until Yun, “A Pre-large Weighted-Fusion System of Sensed High-Utility Patterns,” **IEEE Sensors Journal**, Vol. 21(14), pp. 15626–15634, 2021 (SCI, JCR:Q2, IF:3.301)

Distributed environments

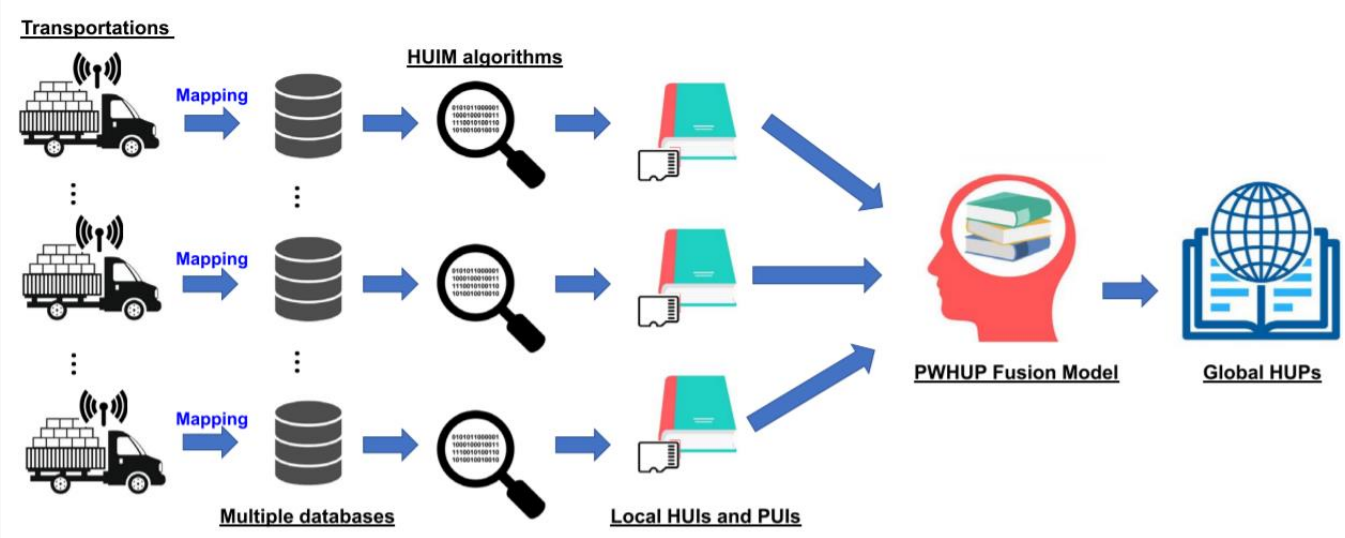


Techniques

- Information loss for the straightforward integration
  - Based on a threshold value
- Utilize the weighted models

$$w_{R_t} = \frac{Num(R_t)}{\sum_{j=1}^{n} Num(R_j)}$$

$$w_{B_k} = \frac{\sum_{R_t \in GP \cap R_t \in B_k} Num(R_t) \times w_{R_t}}{\sum_{i=1}^{n} \sum_{R_h \in GP \cap R_h \in B_i} Num(R_h) \times w_{R_h}}$$



- Apply the pre-large concept to keep the potential HUIs
  - Reduce the rescanning step of DB

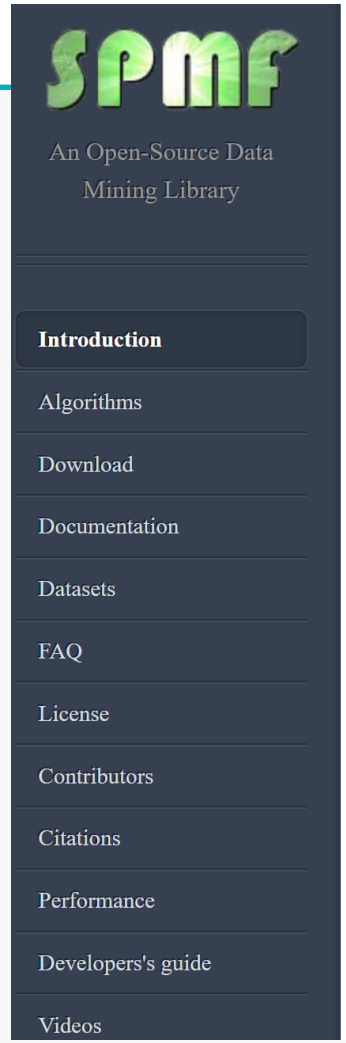
Other selected papers in utility-oriented mining

- Usman Ahmed, **Jerry Chun-Wei Lin***, Rizwan Yasin, Youcef Djenouri, and Gautam Srivastava, “An Evolutionary Model to Mine High Expected Utility Patterns from Uncertain Databases,” **IEEE Transactions on Emerging Topics in Computational Intelligence**, Vol. 5(1), pp. 19–28, 2021
- Jimmy Ming-Tai Wu, **Jerry Chun-Wei Lin***, and Ashish Tamrakar, “High-Utility Itemset Mining with Effective Pruning Strategies,” **ACM Transactions on Knowledge Discovery from Data**, Vol. 13(6), Article No. 58, 2019 (SCI, JCR:Q2, IF:2.713)
- Heonho Kim, Unil Yun, Bay Vo, **Jerry Chun-wei Lin**, and Witold Pedrycz, “Periodicity-Oriented Data Analytics on Time Series Data for Intelligence System,” **IEEE Systems Journal**, 2020 (SCI, JCR:Q2, IF:3.921)
- Wensheng Gan, **Jerry Chun-Wei Lin***, Jiexiong Zhang, Hongzhi Yin, Philippe Fournier-Viger, Han-Chieh Chao, and Philip S. Yu, “Mining Across Multi- Dimensional Sequences,” **ACM Transactions on Knowledge Discovery from Data**, Vol. 15(5), Article No.:82, 2021 (SCI, JCR:Q2, IF:2.713)
- Heonho Kim, Unil Yun, Bay Vo, **Jerry Chun-Wei Lin**, and Witold Pedrycz, “Periodicity-Oriented Data Analytics on Time Series Data for Intelligence System,” **IEEE Systems Journal**, 2020 (SCI, JCR:Q2, IF:3.921)
- Wensheng Gan, **Jerry Chun-Wei Lin*,** Jiexiong Zhang, Philippe Fournier-Viger, Han-Chieh Chao, and Philip S. Yu, “Fast Utility Mining on Sequence Data,” **IEEE Transactions on Cybernetics**, Vol. 51(2), pp. 487–500, 2021 (SCI, JCR:Q1, IF:11.448)

Pattern mining survey works

- Wensheng Gan, **Jerry Chun-Wei Lin***, Philippe Fournier-Viger, Han-Chieh Chao, and Philip S. Yu, “A Survey of Parallel Sequential Pattern Mining,” **ACM Transactions on Knowledge Discovery in Data**, Vol. 13(3), Article No. 25, 2019 (SCI, JCR:Q2, IF:2.713)
- Wensheng Gan, **Jerry Chun-Wei Lin*,** Philippe Fournier-Viger, Han-Chieh Chao, Vincent S. Tseng, and Philip S. Yu, “A Survey of Utility-Oriented Pattern Mining,” **IEEE Transactions on Knowledge and Data Engineering**, Vol. 33(4), pp. 1306–1327, 2021 (SCI, JCR:Q1, IF:6.977)
- Asma Belhadi, Youcef Djenouri, **Jerry Chun-Wei Lin**, and Alberto Cano, “Trajectory Outlier Detection: Algorithms, Taxonomies, Evaluation and Open Challenges,” **ACM Transactions on Management Information Systems**, Vol. 11(3), Article No.:16, 2020

SPMF project



Introduction

SPMF is an **open-source software** and **data mining mining library** written in **Java**, specialized in **pattern mining** (the discovery of patterns in data)
.

It is distributed under the **GPL v3 license**.

It offers implementations of **196 data mining algorithms** for:

- **association rule mining,**
- **itemset mining,**
- **sequential pattern**
- **sequential rule mining,**
- **sequence prediction,**
- **periodic pattern mining,**
- **episode mining**
- **high-utility pattern mining,**
- **time-series mining.**
- **clustering** and **classification**,

SPMF can be used as a **standalone program** with a simple user interface or from the **command line**.

Moreover, the **source code** of each algorithm can be easily integrated in other **Java** software.

Besides, some unofficial wrappers are available for other languages such as Python, R and Weka.

SPMF is fast and lightweight (no dependencies to other libraries).

The current **version** is **v2.42b** and was released the **11th March 2020.**

PKDD 2017, JMLR 2016

PPSF project

PRIVACY-PRESERVING AND SECURITY FRAMEWORK

AN OPEN-SOURCE OF PRIVACY-PRESERVING AND SECURITY LIBRARY

Introduction Algorithms Download Contributors Citations License Other resources Documentation Performance FAQ

Introduction

PPSF is an **open-source privacy-perserving library** written in **Java**.

It offers implementations of 13 privacy preserving algorithms for:

- K-anonymity
- Privacy preserving data ming
- Privacy preserving utility ming

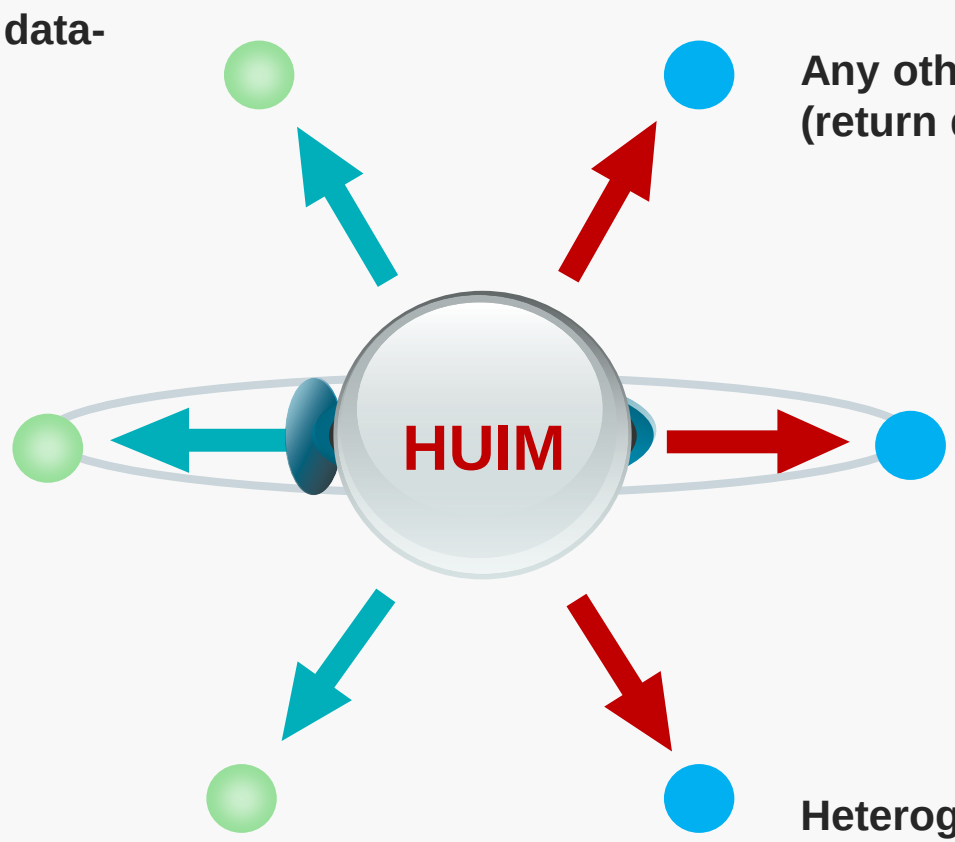
These algorithms were integred in a standalone software that can be easily implemented in different systems.

IEEE ICDM 2018

Conclusion and future work

Conclusions:

Future works:



Pattern mining tasks

- How to obtain the new knowledge for decision making
- How to design the effective structure to keep the information
- How to reduce the search space of the candidates

Instead of data mining, how about ML/DL?

Other research works (1) - Fintech

- IEEE TFS
- IEEE IOTJ
- …etc.

IEEE TRANSACTIONS ON FUZZY SYSTEMS 1

Effective Fuzzy System for Qualifying the Characteristics of Stocks by Random Trading

Mu-En Wu, Jia-Hao Syu, Jerry Chun-Wei Lin*, and Jan-Ming Ho

***Abstract*—Trading strategies can be divided into two categories; i.e., those with momentum characteristic and those that appear contrarian. The characteristics of trading strategies have been widely studied; however, there has been relatively little work on the characteristics of stocks. Furthermore, there is no standard approach to the classification of stocks in terms of momentum and contrarian. This paper presents a fuzzy momentun contrarian uncertain characteristic system for the classification and quantification of stock characteristics. Random trading, stop-loss, and take-profit mechanisms are first used to identify characteristics, and then a novel profitability index with type-2 fuzzy-set module is used to quantify them. In the experiments, 41 stocks on the Taiwan 50 index were deemed suitable for momentum strategies, whereas 9 stocks were deemed suitable for contrarian strategies. An uphill relationship between profitability index and trading performance is observed, which produced correlation coefficients of 0.148–0.539, and classification accuracy of 52.0%–60.0%. However, the proposed system greatly improved classification performance, resulting in correlation coefficients of 0.572–0.722 with accuracy of 63.6%–84.5%. In the real-world application, the proposed system outperforms the benchmark among all datasets, and increases the profitability by 1.5 times on Taiwan 50 dataset. These results clearly demonstrate the efficiency of the proposed system in the quantification and classification of stocks suited to momentum- and contrarian-type trading strategies and also in the real-world applications.**

***Index Terms*—Profitability index, random trading, momentum, contrarian**

I. Introduction

THE allocation of financial assets to companies or commodities in expectation of gaining a profit (i.e., investment) is crucial to economic growth [1], and trading strategies are crucial to investment performance. Overall, trading strategies can be divided into two categories; i.e., those with momentum characteristic and those that appear contrarian [2]. Momentum-type strategies are based on the belief that the price will follow recent trends [3]. Contrarian-type strategies are based on the belief that prices will move against recent trends [4]. These two types of strategy also tend to generate opposing trading signals.

Mu-En Wu is with the Department of Information and Finance Management, National Taipei University of Technology, Taiwan. Email: mnwu@ntut.edu.tw

Jia-Hao Syu is with the Department of Computer Science and Information Engineering, National Taiwan University, Taiwan. Email: f08922011@ntu.edu.tw

Jerry Chun-Wei Lin is with Department of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, Bergen, Norway. Email: jerrylin@ieee.org (*Corresponding author)

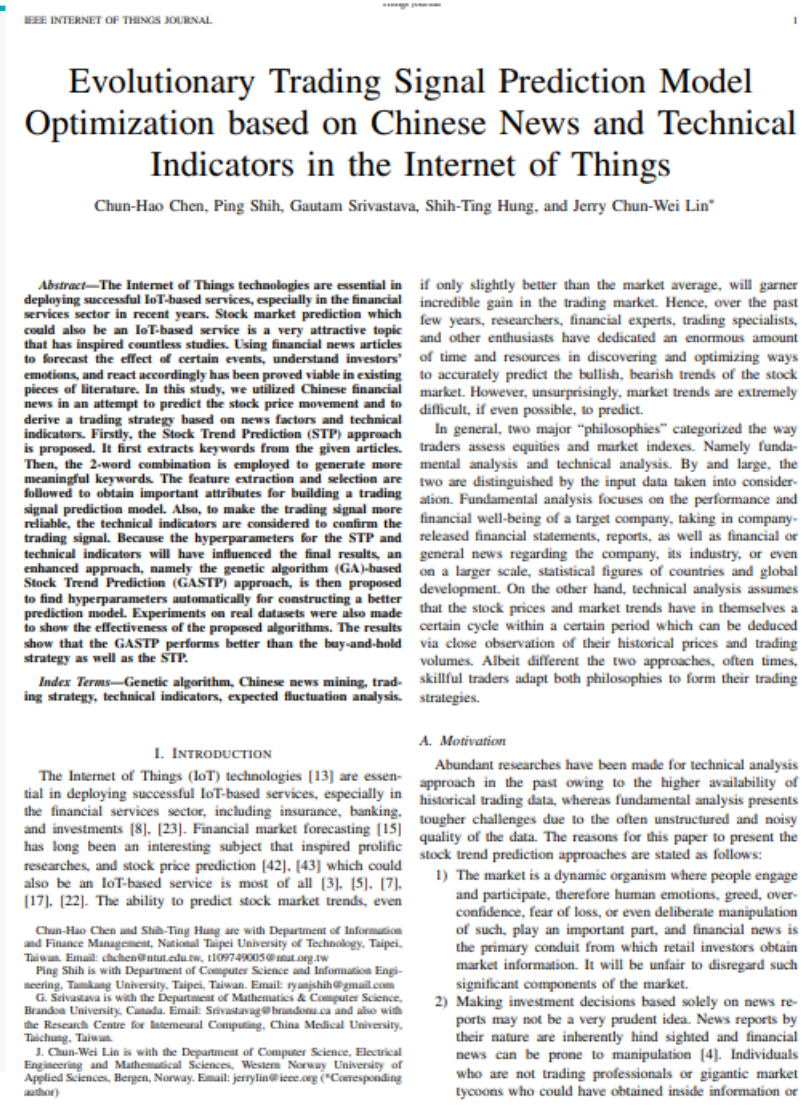
Jan-Ming Ho is with the Institute of Information Science, Academia Sinica, Taiwan. Email: hoho@iis.sinica.edu.tw

The characteristics of trading strategies have been widely studied [5]; however, there has been relatively little work on the characteristics of stocks. Furthermore, there is no standard approach to the classification of stocks as momentum-type or contrarian-type. Searching among the thousands of existing trading strategies is time-consuming and largely ineffectual. The adoption of an erroneous trading strategy or misidentifying the characteristics of the target stock can result in enormous losses. Investors need a system to facilitate the classification and quantification of stocks to inform their decisions with regard to trade strategies. In this paper, a system based on fuzzy analysis methods is presented, referred to as the Fuzzy mOmentun Contrarian Uncertain characteristic System (FOCUS). This paper makes the following contributions:

1) Random trading algorithms are designed to analyze the characteristics of stocks.
2) A profitability index, which uses a type-2 fuzzy-set is developed to quantify those characteristics.
3) An uncertainty factor in the system is devised to filter out stocks that resist classification.
4) The proposed system helps to elucidate the characteristics of stocks and thereby eliminates the time wasted assessing unsuitable trading strategies.

Douglas [6] defined random trading as the poorly-planned process of making trades without the guidance of a plan based on informative data (i.e., prices or market information). Nonetheless, a random trading strategy can be used to reveal investment behaviors and the characteristics of stocks and trading strategies [7]. Among the thousands of trading strategies that have been developed in the field of finance, stop-loss [8] and take-profit [9] are two common momentum-type and contrarian-type strategies. Several studies investigated for the momentum and contrarian effect [10] through the stop-loss and take-profit mechanisms [11]. In this paper, a random trading based on stop-loss and take-profit strategies is employed to investigate the characteristics of stocks. A Profitability Index (PI) is then created to indicate the trading performance of a given stock under momentum- and contrarian-type strategies. The proposed PI aims to quantify the degree of the suitability of target stock to momentum- and contrarian-type trading.

Unfortunately, the intangibility of momentum and contrarian concepts hinders the task of quantify the degree of these characteristics. In addition, the degree of characteristics is not absolute and well-defined, but relative and uncertain. Fuzzy-set theory [12] is used to model situations in a manner that makes it easier for humans to make rational decisions in uncertainty and imprecision environments. Type-2 fuzzy-



IEEE INTERNET OF THINGS JOURNAL 1

Evolutionary Trading Signal Prediction Model Optimization based on Chinese News and Technical Indicators in the Internet of Things

Chun-Hao Chen, Ping Shih, Gautam Srivastava, Shih-Ting Hung, and Jerry Chun-Wei Lin*

***Abstract*—The Internet of Things technologies are essential in deploying successful IoT-based services, especially in the financial services sector in recent years. Stock market prediction which could also be an IoT-based service is a very attractive topic that has inspired countless studies. Using financial news articles to forecast the effect of certain events, understand investors' emotions, and react accordingly has been proved viable in existing pieces of literature. In this study, we utilized Chinese financial news in an attempt to predict the stock price movement and to derive a trading strategy based on news factors and technical indicators. Firstly, the Stock Trend Prediction (STP) approach is proposed. It first extracts keywords from the given articles. Then, the 2-word combination is employed to generate more meaningful keywords. The feature extraction and selection are followed to obtain important attributes for building a trading signal prediction model. Also, to make the trading signal more reliable, the technical indicators are considered to confirm the trading signal. Because the hyperparameters for the STP and technical indicators will have influenced the final results, an enhanced approach, namely the genetic algorithm (GA)-based Stock Trend Prediction (GASTP) approach, is then proposed to find hyperparameters automatically for constructing a better prediction model. Experiments on real datasets were also made to show the effectiveness of the proposed algorithms. The results show that the GASTP performs better than the buy-and-hold strategy as well as the STP.**

***Index Terms*—Genetic algorithm, Chinese news mining, trading strategy, technical indicators, expected fluctuation analysis.**

I. Introduction

The Internet of Things (IoT) technologies [13] are essential in deploying successful IoT-based services, especially in the financial services sector, including insurance, banking, and investments [8], [23]. Financial market forecasting [15] has long been an interesting subject that inspired prolific researches, and stock price prediction [42], [43] which could also be an IoT-based service is most of all [3], [5], [7], [17], [22]. The ability to predict stock market trends, even if only slightly better than the market average, will garner incredible gain in the trading market. Hence, over the past few years, researchers, financial experts, trading specialists, and other enthusiasts have dedicated an enormous amount of time and resources in discovering and optimizing ways to accurately predict the bullish, bearish trends of the stock market. However, unsurprisingly, market trends are extremely difficult, if even possible, to predict.

In general, two major "philosophies" categorized the way traders assess equities and market indexes. Namely fundamental analysis and technical analysis. By and large, the two are distinguished by the input data taken into consideration. Fundamental analysis focuses on the performance and financial well-being of a target company, taking in company-released financial statements, reports, as well as financial or general news regarding the company, its industry, or even on a larger scale, statistical figures of countries and global development. On the other hand, technical analysis assumes that the stock prices and market trends have in themselves a certain cycle within a certain period which can be deduced via close observation of their historical prices and trading volumes. Albeit different the two approaches, often times, skillful traders adapt both philosophies to form their trading strategies.

Chun-Hao Chen and Shih-Ting Hung are with Department of Information and Finance Management, National Taipei University of Technology, Taipei, Taiwan. Email: chchen@ntut.edu.tw, t109749005@ntut.org.tw

Ping Shih is with Department of Computer Science and Information Engineering, Tamkang University, Taipei, Taiwan. Email: ryanjshih@gmail.com

G. Srivastava is with the Department of Mathematics & Computer Science, Brandon University, Canada. Email: Srivastavag@brandonu.ca and also with the Research Centre for Interneural Computing, China Medical University, Taichung, Taiwan.

J. Chun-Wei Lin is with the Department of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, Bergen, Norway. Email: jerrylin@ieee.org (*Corresponding author)

A. Motivation

Abundant researches have been made for technical analysis approach in the past owing to the higher availability of historical trading data, whereas fundamental analysis presents tougher challenges due to the often unstructured and noisy quality of the data. The reasons for this paper to present the stock trend prediction approaches are stated as follows:

1) The market is a dynamic organism where people engage and participate, therefore human emotions, greed, overconfidence, fear of loss, or even deliberate manipulation of such, play an important part, and financial news is the primary conduit from which retail investors obtain market information. It will be unfair to disregard such significant components of the market.
2) Making investment decisions based solely on news reports may not be a very prudent idea. News reports by their nature are inherently hind sighted and financial news can be prone to manipulation [4]. Individuals who are not trading professionals or gigantic market tycoons who could have obtained inside information or

Other research works (2) – trajectory outlier analysis

- IEEE T-ITS
- IEEE IOTJ
- ACM TKDD
- …etc.

Deep Active Learning Intrusion Detection and Load Balancing in Software-Defined Vehicular Networks

Usman Ahmed, Jerry Chun-Wei Lin, *Senior Member, IEEE*, Gautam Srivastava, *Senior Member, IEEE*, Unil Yun, and Amit Kumar Singh, *Senior Member, IEEE*

***Abstract*— Software-defined vehicular networks (SDVN) can help analyze and reconfigure networks. Massive data generation in autonomous vehicles can lead to issues in network configuration, routing, network characteristics, and system load factors. Load balancing in vehicle sensors helps reduce delays and improve resource utilization. In this paper, we propose a load balancing algorithm to map sensor data, vehicles and data centers performing tasks. A dynamic convergence method is proposed to help identify vehicle system load factors and compare their termination criteria. We also propose a packet-level intrusion detection model. After all load balancing, the model can track the attack on the network. The proposed model further combines the entropy-based active learning and the attention-based model to efficiently identify the attacks. Experiments are then conducted on the standard KDD data to validate the developed models with and without an attention-based active learning mechanism. Our experimental results show that the load balancing mechanism is able to achieve more performance gains than previous techniques. Moreover, the results show that the developed model can improve the decision boundary by using a pooling strategy and an entropy uncertainty measure.**

***Index Terms*— Software-defined networking (SDN), network performance, intelligent load balancing, road traffic management, smart city application.**

I. Introduction

HETEROGENEOUS applications lead to huge amounts of data in distributed computing. The future of the Internet of Things (IoT) [1] will connect several heterogeneous devices to interact with distributed data [2]. Several objects (e.g., vehicle networks) will be connected to data using advanced technologies, i.e., 5G/6G. This will lead to data explosion with the increase of data transmission issues. Objects may include traffic flow analysis and security systems equipped with sensors connected to devices connected via vehicle networks. Analysis of real-time environments is necessary to ensure that information is transmitted securely to IoT gateways.

With the rise of autonomous vehicles, the future of intelligent transportation systems is also gaining traction. This evolution toward new and improved sensor-based devices will make travel more comfortable and secure. Such transformation will increase the need for security and improve data portability to support transportation infrastructure. Improving vehicle sensors can help improve data processing rates for each real-time collection. The collected sensor data will be processed on cloud-based servers [3]. The servers help store, process, and then update the data in real time. This helps transportation infrastructure analyze connected vehicles (CV) via vehicle-to-infrastructure (V2I) communications for traffic control and traffic flow estimation. CV trajectory analysis can help in identifying the vehicles with temporal and spatial applications leading to accurate traffic information. In addition to the benefits of CV, infrastructure connectivity also opens the door to cyber attacks. Malicious attackers can conduct targeted cyber attacks against the data-driven V2I applications. This leads to problems in the traffic operation of the vehicle, including safety and mobility. The trajectory benefits the data-driven application to analyze the cyber security.

To improve the security of CV systems, Security Credential Management Systems (SCMS) are being implemented by USDOT [4]. This type of system requires a digital certificate to be sent with each CV message. The receiver verifies the received message before extracting information. However, adaptive attackers used the legalized communication device to send fake data with a valid certification. The attacker spoofed the identity of the sender, which modified the attached message (e.g., speed and location data). SCMS then used the forged data from CV, to verify the identity and message [4]. It has been reported that identity theft has been carried out and the vehicle owner's software is exploited via the electronic control units (ECU) or the infotainment system [4]. The vehicle owner has arbitrary access, so in practice the attacks on the fake trajectories can be carried out quickly. Another problem is the analysis of the trajectories of the deviation point group. This type is typically used for financial rewards and is often used

Manuscript received October 1, 2021; revised January 12, 2022 and March 1, 2022; accepted April 6, 2022. This work was supported in part by the National Centre for Research and Development through the project Automated Guided Vehicles integrated with Collaborative Robots for Smart Industry Perspective under Contract NOR/POLNOR/CoBotAGV/0027/2019-00. The Associate Editor for this article was A. Jolfaei. *(Corresponding author: Jerry Chun-Wei Lin.)*

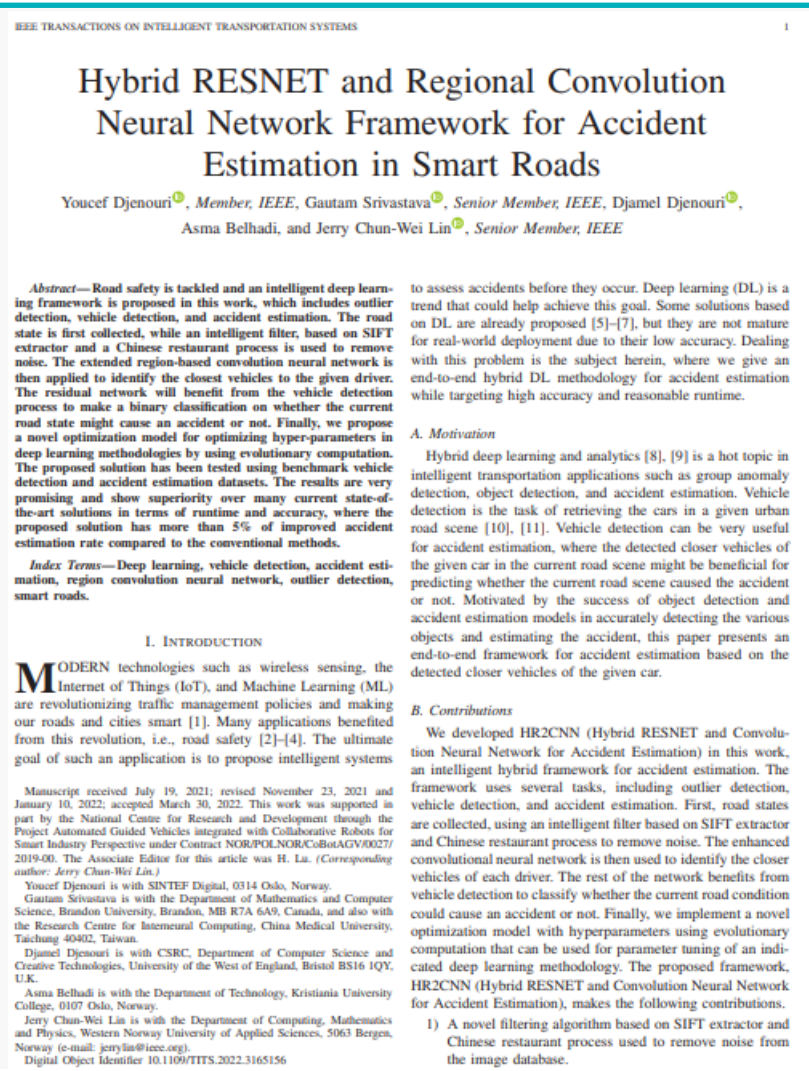
Usman Ahmed and Jerry Chun-Wei Lin are with the Department of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, 5063 Bergen, Norway (e-mail: usman.ahmed@hvl.no; jerrylin@ieee.org).

Gautam Srivastava is with the Department of Mathematics and Computer Science, Brandon University, Brandon, MB R7A 6A9, Canada, and also with the Research Centre for Interneural Computing, China Medical University, Taichung 40402, Taiwan (e-mail: srivastavag@brandonu.ca).

Unil Yun is with the Department of Computer Engineering, College of Electronics and Information Engineering, Sejong University, Seoul 143-747, South Korea (e-mail: yunei@sejong.ac.kr).

Amit Kumar Singh is with the Department of Computer Science and Engineering, NIT Patna, Patna, Bihar 800005, India (e-mail: amit.singh@nitp.ac.in).

Digital Object Identifier 10.1109/TITS.2022.3166864



Hybrid RESNET and Regional Convolution Neural Network Framework for Accident Estimation in Smart Roads

Youcef Djenouri, *Member, IEEE*, Gautam Srivastava, *Senior Member, IEEE*, Djamel Djenouri, Asma Belhadi, and Jerry Chun-Wei Lin, *Senior Member, IEEE*

***Abstract*— Road safety is tackled and an intelligent deep learning framework is proposed in this work, which includes outlier detection, vehicle detection, and accident estimation. The road state is first collected, while an intelligent filter, based on SIFT extractor and a Chinese restaurant process is used to remove noise. The extended region-based convolution neural network is then applied to identify the closest vehicles to the given driver. The residual network will benefit from the vehicle detection process to make a binary classification on whether the current road state might cause an accident or not. Finally, we propose a novel optimization model for optimizing hyper-parameters in deep learning methodologies by using evolutionary computation. The proposed solution has been tested using benchmark vehicle detection and accident estimation datasets. The results are very promising and show superiority over many current state-of-the-art solutions in terms of runtime and accuracy, where the proposed solution has more than 5% of improved accident estimation rate compared to the conventional methods.**

***Index Terms*— Deep learning, vehicle detection, accident estimation, region convolution neural network, outlier detection, smart roads.**

I. Introduction

MODERN technologies such as wireless sensing, the Internet of Things (IoT), and Machine Learning (ML) are revolutionizing traffic management policies and making our roads and cities smart [1]. Many applications benefited from this revolution, i.e., road safety [2]–[4]. The ultimate goal of such an application is to propose intelligent systems to assess accidents before they occur. Deep learning (DL) is a trend that could help achieve this goal. Some solutions based on DL are already proposed [5]–[7], but they are not mature for real-world deployment due to their low accuracy. Dealing with this problem is the subject herein, where we give an end-to-end hybrid DL methodology for accident estimation while targeting high accuracy and reasonable runtime.

A. Motivation

Hybrid deep learning and analytics [8], [9] is a hot topic in intelligent transportation applications such as group anomaly detection, object detection, and accident estimation. Vehicle detection is the task of retrieving the cars in a given urban road scene [10], [11]. Vehicle detection can be very useful for accident estimation, where the detected closer vehicles of the given car in the current road scene might be beneficial for predicting whether the current road scene caused the accident or not. Motivated by the success of object detection and accident estimation models in accurately detecting the various objects and estimating the accident, this paper presents an end-to-end framework for accident estimation based on the detected closer vehicles of the given car.

B. Contributions

We developed HR2CNN (Hybrid RESNET and Convolution Neural Network for Accident Estimation) in this work, an intelligent hybrid framework for accident estimation. The framework uses several tasks, including outlier detection, vehicle detection, and accident estimation. First, road states are collected, using an intelligent filter based on SIFT extractor and Chinese restaurant process to remove noise. The enhanced convolutional neural network is then used to identify the closer vehicles of each driver. The rest of the network benefits from vehicle detection to classify whether the current road condition could cause an accident or not. Finally, we implement a novel optimization model with hyperparameters using evolutionary computation that can be used for parameter tuning of an indicated deep learning methodology. The proposed framework, HR2CNN (Hybrid RESNET and Convolution Neural Network for Accident Estimation), makes the following contributions.

1) A novel filtering algorithm based on SIFT extractor and Chinese restaurant process used to remove noise from the image database.

Manuscript received July 19, 2021; revised November 23, 2021 and January 10, 2022; accepted March 30, 2022. This work was supported in part by the National Centre for Research and Development through the Project Automated Guided Vehicles integrated with Collaborative Robots for Smart Industry Perspective under Contract NOR/POLNOR/CoBotAGV/0027/2019-00. The Associate Editor for this article was H. Lu. *(Corresponding author: Jerry Chun-Wei Lin.)*

Youcef Djenouri is with SINTEF Digital, 0314 Oslo, Norway.

Gautam Srivastava is with the Department of Mathematics and Computer Science, Brandon University, Brandon, MB R7A 6A9, Canada, and also with the Research Centre for Interneural Computing, China Medical University, Taichung 40402, Taiwan.

Djamel Djenouri is with CSRC, Department of Computer Science and Creative Technologies, University of the West of England, Bristol BS16 1QY, U.K.

Asma Belhadi is with the Department of Technology, Kristiania University College, 0107 Oslo, Norway.

Jerry Chun-Wei Lin is with the Department of Computing, Mathematics and Physics, Western Norway University of Applied Sciences, 5063 Bergen, Norway (e-mail: jerrylin@ieee.org).

Digital Object Identifier 10.1109/TITS.2022.3165156

Other research (3) – Mental health

- IEEE JBHI
- IEEE TCSS
- ACM TALLIP
- ….etc.

Hyper-Graph Attention Based Federated Learning Methods For Use in Mental Health Detection

Usman Ahmed, Jerry Chun-Wei Lin*, and Gautam Srivastava

***Abstract*—Internet-Delivered Psychological Treatment (IDPT) has become necessary in the medical field. Deep neural networks (DNNs) require large, diverse patient populations to train models that achieve clinician-level performance. However, DNN models trained on limited datasets have poor clinical performance when used in a new location with different data. Thus, increasing the availability of diverse as well as distinct training data is vital. This study proposes a structural hypergraph as well as an emotional lexicon for word representation. An embedding model based on federated learning was developed for mental health symptom detection. The model treats text data as a collection of consecutive words. The model then learns a low-dimensional continuous vector while maintaining contextual linkage. The generated models with attention-based mechanisms as well as federated learning are then tested experimentally. Our strategy is suitable for vocabulary diversification, grammatical word representation, as well as dynamic lexicon analysis. The goal is to create semantic word representations using an attention network model. Later, clinical processes are used to mark the text by embedding it. Experimental results show the encoding of emotional words using the structural hypergraph. The 0.86 ROC was achieved using the bidirectional LSTM architecture with an attention mechanism.**

***Index Terms*—text clustering, NLP, Internet-delivered interventions, word sense identification, adaptive treatments**

I. Introduction

Wearable gadgets as well as innovations on the Internet of Medical Things (IoMT) have made remote patient monitoring possible like never before. Through the process of learning patterns from provided data by devices, machine learning, as well as deep learning algorithms, are significantly helping physicians diagnose patients remotely. Classical machine learning (ML), as well as deep learning (DL) models have the disadvantage of requiring patient data to be transferred from specialized devices, sensors, as well as wearables to centralized servers so that the data can be trained with ML/DL models. Because of the nature of data in healthcare, the techniques mentioned thus far for transferring patient data to centralized servers can pose significant privacy, as well as security issues.

Recent advancement in ML/DL is federated learning, where data is not transmitted to centralized servers. However, the ML model itself is distributed to different nodes (devices) for training data [1]. Parameters of the device models are then transmitted to a centralized model for training the model globally. Federated learning can protect patient data privacy by preventing sensitive information from being exposed to intruders like hackers. COVID-19 is a global health disaster which has threatened the livelihood of millions of people. Advanced ML technologies have been used to build models to predict and diagnose diseases for combating coronavirus. Furthermore, partly because of unstable communication methods, as well as potential attackers, the huge amount of data collected during this time can pose various security and privacy issues. Privacy-preserving federated learning becomes a superior alternative to ensure the security of patient data during the transfer, as well as the training process. Therefore, we have compiled several high quality works on this topic that use state-of-the-art federated learning technologies to protect healthcare data, as well as provide valuable guidance for the current society.

According to WHO [2], *depression* is a severe problem among the most disabling diseases in the world. Depressive disorders affect approximately 264 million people worldwide. Due to a lack of interpersonal interaction, as well as trust, most cases of depression go untreated [3]. Because they do not seek treatment, the leading cause of mortality among individuals between the ages of 15, as well as 29 years is suicide or 76% - 85% within middle-income countries. In addition to mental health issues, early detection is hindered by a lack of resources, inexperienced medical personnel, social stigma, as well as rapid response [2]. People are often humbled by their inability to maintain stability with one's mental state is part due to both shy aspects, as well as nervousness. Anytime that a patient goes through any kind of thorough evaluation of psychological aspects of their current state, the problem persists [4]; as a result, people with depression may decide against future therapy to manage their existing health problems.

When dealing directly with IDPT, or Internet-delivered Psychological Treatment, we can say without a doubt that this level of treatment helps people deal with their own psychological problems with less resources. A tunnel-based approach is rigid as well as not interoperable [5]. The approach is not adaptable due to low adoption, as well as a more significant number of dropouts. User adoption must be considered overtime. With an IDPT system that assesses user behavior, as well as emotional exchanges, adaptation is possible. For example, Mukhiya *et al.* notes that culturally-based mental health issues influence individualized user behavior [6].

U. Ahmed and J. C. W Lin are with the Department of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, 5063, Bergen, Norway. Email: usman.ahmed@hvl.no, jerrylin@ieee.org. Website: http://ikelab.net. (*Corresponding author: Jerry Chun-Wei Lin)

G. Srivastava is with the Department of Mathematics & Computer Science, Brandon University, Canada as well as the Research Centre for Interneural Computing, China Medical University, Taichung, Taiwan. Email: SRIVASTAVAG@brandonu.ca

Deep Explainable Hate Speech Active Learning on Social-Media Data

Usman Ahmed and Jerry Chun-Wei Lin, *Senior Member, IEEE*

***Abstract*—Hate speech is demonstrably aimed at social tension and violence. Detection becomes increasingly difficult as overlapping emotional feelings occur. However, there are still several unresolved issues with informal and indirect targeting of negative communication, including sarcasm, misrepresentation, and praise for the target's or society's immoral behavior. In this study, we proposed a method for instance selection based on attention network visualization. The goal is to categorize, modify, and expand the number of training instances. To this end, we first used the lexicons of hate speech and online forums to train the embedding using transfer learning. Then, we used synonym expansion to the semantic vectors. The active learning approach was used to train the task using the result-label pairs. The entropy-based selection and visualization techniques help select unlabeled text for each active learning cycle. The approach is improved, and the number of training instances is increased to improve the model's accuracy. The active learning cycles are repeated until all unlabeled texts are converted to labeled text. The semantic embedding and lexicon expansion improve the model receiver operating characteristics (ROCs) from 0.89 to 0.91. The bidirectional LSTM with attention and active learning achieved 0.90 for precision–recall. The learned model can visualize the position-weighted terms to illustrate why hate speech is classified.**

***Index Terms*—Deep learning, ethnic hate, explainable machine learning (ML), hate speech detection.**

I. Introduction

Artificial intelligence in medicine (AIM) has the potential to improve medical data by enabling clinicians to understand various data more naturally and accurately [1]. Advanced data-driven technologies can help clinicians interpret complex heterogeneous medical data and make appropriate diagnoses and treatments for patients [2]. In today's pandemic environment, artificial intelligence (AI) tools help physicians and lighten their workload [3]. Machine learning (ML) is used to replicate human behavior using labeled data. The AIM is linked to the electronic health records (EHRs) of healthcare organizations [4]. Data provided by EHR systems enable more effective use of AI technologies. The EHR includes both organized data (e.g., patient demographics, diagnoses, and procedures) and unstructured data (e.g., physician notes and clinical reports).

Despite the promising research efforts at AIM, practical clinical application is limited by the problem of data interpretability. Numerous studies have shown that medical professionals lack confidence in AIM [5]. The primary sources of trust concerns include limited access to medical data, lack of experience integrating clinical processes, legislative compilation, and the technical difficulties associated with developing and operating AIM systems. In 2017, the Defense Advanced Research Projects Agency (DARPA) published a report on its Explainable AI (XAI) research program [6]. According to DARPA, emerging AI systems are limited in understanding and communicating. This is because humans cannot interpret AI systems, i.e., understand why an AI system makes a particular decision.

In addition, researchers claim that physicians easily rely on drugs such as aspirin although the underlying mechanism is unclear. When AI systems begin to explain the decision-making through the lens of performance, physicians will accept them. On the other hand, drug regulatory agencies guarantee that each drug is thoroughly tested and undergoes randomized clinical trials before it is approved for general use. In addition, regulatory agencies, such as the Food and Drug Administration in the United States, conduct postmarket surveillance. The goal was to quickly remove drugs from the market with unforeseen or catastrophic side effects. The AI systems were constructed using a small dataset and may have been compromised due to problems with generalization for the additional samples. The AI system lacks an equivalent mechanism for taking full responsibility for safety and efficacy concerns. XAI helps determine the validity of target decisions and ensures consensus among medical specialists. By acting as decision support systems, XAIs increase their confidence in AIM [7]. Explainability is critical to the use of AI in healthcare decision-making [7], [8].

In recent years, social media platforms, such as Twitter and Facebook, have gained popularity among the general public. They are packed with user-generated content, including text, social media data, photos, and videos. Typically, hate speech can be considered an expression that denigrates a person or group based on their sexual orientation, race, ethnicity, religion, gender, color, or national origin. Given a large amount of user-generated content on the Internet, especially on social media, it is becoming increasingly important to identify and potentially stop the transmission of hate speech, i.e., in the fight against racism and sexism. Given the vast amount of user-generated information on the Internet, especially on social media, it is increasingly important to identify and potentially curb the spread of hate speech, for example, in the fight against sexism and xenophobia. The term "hate speech" varies from

Manuscript received December 29, 2021; revised February 24, 2022; accepted April 2, 2022. *(Corresponding author: Jerry Chun-Wei Lin.)*

The authors are with the Department of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, 5063 Bergen, Norway (e-mail: usman.ahmed@hvl.no; jerrylin@ieee.org).

Digital Object Identifier 10.1109/TCSS.2022.3165136

Other research (4) – privacy and security

- ACM TOIT
- IEEE IOTJ
- …etc.

A Multi-Threshold Ant Colony System-based Sanitization Model in Shared Medical Environments

JIMMY MING-TAI WU, Shandong University of Science and Technology, China
GAUTAM SRIVASTAVA, Brandon University, Canada and China Medical University, Taiwan
JERRY CHUN-WEI LIN, Western Norway University of Applied Sciences, Norway
QIAN TENG, Shandong University of Science and Technology, China

During the past several years, revealing some useful knowledge or protecting individual's private information in an identifiable health dataset (i.e., within an Electronic Health Record) has become a tradeoff issue. Especially in this era of a global pandemic, security and privacy are often overlooked in lieu of usability. Privacy preserving data mining (PPDM) is definitely going to be have an important role to resolve this problem. Nevertheless, the scenario of mining information in an identifiable health dataset holds high complexity compared to traditional PPDM problems. Leaking individual private information in an identifiable health dataset has becomes a serious legal issue. In this article, the proposed Ant Colony System to Data Mining algorithm takes the multi-threshold constraint to secure and sanitize patents' records in different lengths, which is applicable in a real medical situation. The experimental results show the proposed algorithm not only has the ability to hide all sensitive information but also to keep useful knowledge for mining usage in the sanitized database.

CCS Concepts: • **Information systems → Data cleaning**; • **Theory of computation → Evolutionary algorithms**; • **Security and privacy → Data anonymization and sanitization**;

Additional Key Words and Phrases: Privacy-preserving data mining, evolutionary algorithm, sensitive itemsets, ant colony system

ACM Reference format:
Jimmy Ming-Tai Wu, Gautam Srivastava, Jerry Chun-Wei Lin, and Qian Teng. 2021. A Multi-Threshold Ant Colony System-based Sanitization Model in Shared Medical Environments. *ACM Trans. Internet Technol.* 21, 2, Article 49 (May 2021), 26 pages.
https://doi.org/10.1145/3408296

1 INTRODUCTION

Since the early 2010s, data mining techniques [15, 16] have been utilized and applied in different domains and applications that can be used to retrieve useful and meaningful information from

This article is supported by National Natural Science Foundation of China (61976126).
Authors' addresses: J. M.-T. Wu, Q. Teng, Shandong University of Science and Technology, 579 Qianwangang Rd, Qingdao, Shandong, 266590, China; emails: wmt@wmt35.idv.tw, qrape@foxmail.com; G. Srivastava, Brandon University, 270 18th St, Brandon, MB R7A 6A9, Canada and China Medical University, Taichung, 44020, Taiwan; email: srivastavag@brandonu.ca; J. C.-W. Lin (corresponding author), Western Norway University of Applied Sciences, Inndalsveien 28, Bergen, 5063, Norway; email: jerrylin@ieee.org.

1533-5399/2021/05-ART49 $15.00
https://doi.org/10.1145/3408296

Privacy-Preserving Multiobjective Sanitization Model in 6G IoT Environments

Jerry Chun-Wei Lin, *Senior Member, IEEE*, Gautam Srivastava, *Senior Member, IEEE*, Yuyu Zhang, Youcef Djenouri, and Moayad Aloqaily, *Member, IEEE*

***Abstract*—The next revolution of the smart industry relies on the emergence of the Industrial Internet of Things (IoT) and 5G/6G technology. The properties of such sophisticated communication technologies will change our perspective of information and communication by enabling seamless connectivity and bring closer entities, data, and "things." Terahertz-based 6G networks promise the best speed and reliability, but they will face new man-in-the-middle attacks. In such critical and high-sensitive environments, the security of data and privacy of information still a big challenge. Without privacy-preserving considerations, the configuration state may be attacked or modified, thus causing security problems and damage to data. In this article, motivated by the need to secure 6G IoT networks, an ant colony optimization (ACO) approach is presented by adopting multiple objectives as well as using transaction deletion to secure confidential and sensitive information. Each ant in the population is represented as a set of possible deletion transactions for hiding sensitive information. We utilize the use of a prelarge concept to assist in the reduction of multiple database scans in the evaluation progress. We then also adopt external solutions to maintain discovered Pareto solutions, thus improving effectiveness to find optimized solutions. Experiments are conducted comparing our methodology to state-of-the-art bioinspired particle swarm optimization (PSO) as well as genetic algorithm (GA). Our strong results clearly show that the designed approach achieves fewer side effects while maintaining low computational cost overall (Chen *et al.*, 2020).**

***Index Terms*—5G/6G, ant colony, decomposition, deep learning, IIoT, object detection, particle swarm optimization (PSO), smart factory.**

Manuscript received May 21, 2020; revised July 30, 2020 and September 15, 2020; accepted October 8, 2020. Date of publication October 21, 2020; date of current version March 24, 2021. *(Corresponding author: Gautam Srivastava.)*
Jerry Chun-Wei Lin is with the School of Information and Control Engineering, Qingdao University of Technology, Qingdao 266520, China, and also with the Department of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, 5063 Bergen, Norway (e-mail: jerrylin@ieee.org).
Gautam Srivastava is with the Department of Mathematics and Computer Science, Brandon University, Brandon, MB R7A 6A9, Canada, and also with the Research Centre for Interneural Computing, China Medical University, Taichung 40402, Taiwan (e-mail: srivastavag@brandonu.ca).
Yuyu Zhang is with the School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China (e-mail: zhangyuyu@stu.hit.edu.cn).
Youcef Djenouri is with the Department of Mathematics and Cybernetics, SINTEF Digital, 0314 Oslo, Norway (e-mail: youcef.djenouri@sintef.no).
Moayad Aloqaily is with the Faculty of Engineering, Al Ain University, Al Ain, UAE (e-mail: maloqaily@ieee.org).
Digital Object Identifier 10.1109/JIOT.2020.3032896

I. Introduction

IN SMART industry, sensors as well as smart devices are widely deployed to collect events and information in wireless sensor networks (WSNs) deployed in the Internet of Things (IoT). Due to characteristics, such as being large-scale, dynamic, self-organization, and constrained, the IoT environments are attractive targets for many attacks [2], [3], thus privacy and security concerns have become an emerging issue in IoT. With the current push in society to 5G and future considerations for 6G networks, the need for secure communication protocols and privacy techniques for user data is paramount. The main reason that can be mentioned here is that in the 6G environment, every object connects to each other to share or exchange the information and data efficiently. Although the data communication and connection in 6G brings high efficiency of data exchanging and sharing, the exponentially increasing networks for the IoT environment also bring data privacy and confidentiality threats, which will be the problem in the future 6G networks [4], [5]. Without strong encryption techniques for collected events as well as data, attacks on the configuration state may cause serious security problems to systems. Sensitive and confidential information shared by IoT applications need protection and can cause development and deployment issues if security is not at the forefront of application design [6]. Privacy-preserving data-mining (PPDM) [7] minimizes the probability of sensitive data disclosure. PPDM can be widely applied to IoT-based applications and domains, allowing for the sanitization of confidential events and information for security purposes.

In PPDM, the most common approach to hide confidential information is through an intricate process known as data sanitization by using deletion operations on transactions. During this process, unfortunately several side effects may be produced, such as missing cost, hiding failure, and artificial cost hereby known as the three factors or three known side effects we wish to minimize. The problem of minimization of these side effects is a known NP-hard problem [7]. In PPDM, when considering data sanitization, it is important to select the appropriate victims (transactions or items) to be deleted from the database that can minimize side effects. Agrawal and Srikant [7] first stated the problem of PPM. Lindell and Pinkas [8] then presented the ID3-based model which is used to primarily solve the PPDM problem. Several algorithms that are related are also presented to hide sensitive rules or items based on the sanitization procedure [9]–[11].

Other research (5) – neural evolving and optimization

- IEEE TNNLS
- IEEE TETCI
- …etc

Multi-Objective Neural Evolutionary Algorithm for Combinatorial Optimization Problems

Yinan Shao, Jerry Chun-Wei Lin*, Gautam Srivastava, Dongdong Guo, Hongchun Zhang, Xiaonan Meng, and Alireza Jolfaei

***Abstract*—There has been a recent surge of success in optimizing deep reinforcement learning models with neural evolutionary algorithms. This type of method is inspired by biological evolution and uses different genetic operations to evolve neural networks. Previous neural evolutionary algorithms mainly focused on single-objective optimization problems. In this paper, we present an end-to-end multi-objective neural evolutionary algorithm based on decomposition and dominance (MONEADD) for combinatorial optimization problems. The proposed MONEADD is an end-to-end algorithm that utilizes genetic operations and rewards signals to evolve neural networks for different combinatorial optimization problems without further engineering. To accelerate convergence, a set of non-dominated neural networks is maintained based on the notion of dominance and decomposition in each generation. In inference time, the trained model can be directly utilized to solve similar problems efficiently, while the conventional heuristic methods need to learn from scratch for every given test problem. To further enhance the model performance in inference time, three multi-objective search strategies are introduced in this work. Our experimental results clearly show that the proposed MONEADD has a competitive and robust performance on a bi-objective of the classic Travel Salesman problem (TSP), as well as Knapsack problem up to 200 instances. We also empirically show that the designed MONEADD has good scalability when distributed on multiple GPUs.**

***Index Terms*—Neural combinatorial optimization, neural evolutionary algorithm, deep reinforcement learning, multi-objective learning, attention mechanism.**

I. Introduction

OVER the past few years, deep reinforcement learning (DRL) has proven to be an effective tool for many practical problems, such as recommendation systems [1], [2], [3], [4], computer vision [5], [6], [7], [8] and natural language processing [9], [10], [11], [12]. One interesting application is the combinatorial optimization problem. Combinatorial optimization problems are conventionally solved by heuristic algorithms. Given a certain problem, heuristic algorithms randomly generate a set of solutions and evolve them generation by generation through genetic operators. Different from heuristic algorithms, DRL aims to build an end-to-end model for combinatorial optimization problems. There are two advantages to adopt DRL into combinatorial optimization problems:

1) DRL models can be easily generalized to several different combinatorial tasks without any hand-craft engineering, while genetic operators of heuristic algorithms are often task-specific.
2) DRL models can learn from large unsupervised data, and generate solutions within a forward pass of a neural network, while most heuristic algorithms should always search from the scratch for every test problem, even the problems are similar to each other.

Bello *et al.* [13] introduced DRL into combinatorial optimization problems. They adopt a pointer network to model the optimization problem and regard negative tour length as a reward signal applied in TSP problems. Network parameters are updated by the policy gradient method. They demonstrate that DRL can be successfully applied to different combinatorial optimization problems.

Most related DRL studies focused on single-objective optimization [14], [15], [16], [17], [18]. As for multi-objective optimization problems (MOPs), a common solution is to adopt a predefined weight vector to combine different objectives and regard this value as a reward signal. Unfortunately, the weights of different objectives are usually hard to define in real-world applications. For example, in e-commerce, a seller may focus on revenue and paid orders simultaneously. While paid orders are much sparser than revenue, it is quite hard to find a specific weight vector to combine these two objectives.

To solve these types of multi-objective optimization problems, the concept of Pareto optimality is commonly used. The definitions of multi-objective optimization problems are then described below. Let $u, v \in R^m$, u is said to dominate $v \iff u_i \geq v_i$ for every $i \in 1, \ldots, m$ and $u_j > v_j$ for at least one index $j \in 1, \ldots, m$. A point $x^* \in \Omega$ is Pareto optimal if there is no point $x \in \Omega$ such that $F(x)$ dominates $F(x^*)$. $F(x^*)$ is then called a Pareto optimal vector. Pareto front (PF) is composed of all the Pareto optimal vectors [19]. An example of PF is shown in Fig. 1.

In this paper, we propose the Multi-Objective Neural Evolutionary Algorithm based on Decomposition and Dominance (MONEADD), which directly outputs a Pareto front for given multi-objective combinatorial optimization problems. This is the first neural evolutionary algorithm for multi-objective reinforcement learning. Different from a conventional evolutionary algorithm that can directly solve MOPs, neural evolutionary algorithms utilize genetic operators to evolve neural networks,

Yinan Shao, Dongdong Guo, Hongchun Zhang and Xiaonan Meng are with the Alibaba Inc, Hangzhou, China. (E-mail: {wusi.syn, dongdong.gdd, hongchun.zhc,xiaonan.mengxn}@alibaba-inc.com)

Jerry Chun-Wei Lin is with the Western Norway University of Applied Sciences, Bergen 5063, Norway. Email: jerrylin@ieee.org (*Corresponding author)

Gautam Srivastava is with the Department of Mathematics and Computer Science, Brandon University, Brandon, Canada as well as with the Research Center for Interneural Computing, China Medical University, Taichung 40402, Taiwan, and with the College of Information and Electrical Engineering, Asia University, Taiwan. e-mail: srivastavag@brandonu.ca

Alireza Jolfaei is with the Department of Computing, Macquarie University, Australia. e-mail: alireza.jolfaei@mq.edu.au

Security and Privacy in Shared HitLCPS Using a GA-Based Multiple-Threshold Sanitization Model

Jimmy Ming-Tai Wu, Gautam Srivastava, Alireza Jolfaei, *Senior Member, IEEE*, Matin Pirouz, and Jerry Chun-Wei Lin, *Senior Member, IEEE*

***Abstract*—In Cyber-Physical Systems (CPS), especially in human-in-the-loop situations (also known as HitLCPS), the security and privacy for keeping sensitive information private is considered an emerging topic in recent decades. Many techniques in privacy-preserving data mining (PPDM) can be applied directly to HitLCPS. However, most of them to date have focused on handling singular threshold problems for data sanitization. If a sensitive itemset includes more items, it has a higher probability of being identified due to its specificity. In this work, we propose a new concept of multiple support thresholds to assist in resolving this issue. The proposed method assigns a stricter threshold for an itemset. Furthermore, a genetic-algorithm (GA)-based model is involved in the designed algorithm to minimize side effects. In our experimental results, the GA-based PPDM approach is compared with traditional Greedy PPDM approaches. The strong experimental results clearly show that our proposed method can give similar performance to conventional algorithms while still maintaining higher-levels of security and privacy protection than previous methods.**

***Index Terms*—Privacy preservation, optimization, genetic algorithm, security, multi-threshold.**

I. Introduction

IN THE past decade, data mining techniques [1], [2] have been utilized and applied in different domains and applications that can be used to retrieve useful and meaningful information from very large datasets (big data). The fundamental algorithm to mine required patterns is called Apriori [3] which uses the minimum support threshold to first and foremost discover the known set of frequent itemsets (FI) from a given database using a level-wise approach. The next step is the use of a combinational approach to derive all of the effective association rules (ARs) that are based on a minimum confidence threshold. Since what is known as the level-wise approach makes use of a "generate-and-test" mechanism that in essence almost always requires major computational costs, the efficient frequent pattern (FP)-tree structure [4] has been developed to keep the frequent 1-itemsets in the tree. Next, a recursive FP-growth mining algorithm is developed to mine the set of frequent itemsets. Several extensions in knowledge discovery in databases (KDD) were then implemented to handle different scenarios and domains for retrieving various knowledge for decision-making, i.e., sequential pattern mining [5] and high-utility pattern mining [6]. Since utility and sequence factors are two important issues for mining the significant information for decision making, several works regarding high-utility sequential pattern mining have been introduced and discussed [7]–[9].

Although KDD techniques can be used to mine the relationship of attributes in databases, confidential/private information can also be revealed or referred from related information during the mining progress [10]. For example, purchase behaviors can refer to the visited malls and even the gender of customers that should be considered confidential information in data analytics. Thus, using what is known as privacy-preserving data mining (PPDM) techniques first arose as an important issue in recent decades. The main purpose of PPDM can be defined easily as the ability to high user-defined sensitive and private information while still maintaining the ability to discover both useful and meaningful information simultaneously from databases for efficient decision-making. Thus, the private or personal information can thus be hidden and secured. One technique of PPDM includes perturbation or sanitization approach in which the confidential information regarding any patient's record is perturbed using a random process. This process distorts sensitive data values by changing them using a process of adding, subtracting, or perturbing the data through other means. In addition, for the sanitization process to be able to hide most of the sensitive information, some of the associated rules may as a side effect be lost as well as artificial rules have a chance of appearing both as common side effects of the sanitization overall process. One of the most well-known side effects common in PPDM includes missing cost, hiding failure, as well as artificial cost. To date, there have been many algorithms proposed that claim to sanitize any original database and be able to effectively and

Manuscript received May 4, 2020; revised September 14, 2020; accepted October 1, 2020. Date of publication September 21, 2021; date of current version January 21, 2022. This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant program (RGPIN-2020-05363) held by Dr. Gautam Srivastava. *(Corresponding author: Jerry Chun-Wei Lin.)*

Jimmy Ming-Tai Wu is with the College of Computer Science, and Engineering, Shandong University of Science, and Technology, Qingdao 266590, China (e-mail: wmt@wmt35.idv.tw).

Gautam Srivastava is with the Department of Mathematics and Computer Science, Brandon University, Brandon, Manitoba R7A 6A9, Canada, with the Research Center for Interneural Computing, China Medical University, Taichung 40402, Taiwan and also with the College of Information and Electrical Engineering, Asia University, Taiwan (e-mail: srivastavag@brandonu.ca).

Alireza Jolfaei is with the Department of Computing, Macquarie University, Sydney, New South Wales 2113, Australia (e-mail: alireza.jolfaei@mq.edu.au).

Matin Pirouz is with the Department of Computer Science, California State University, Fresno 93740, California USA (e-mail: matin.pn@gmail.com).

Jerry Chun-Wei Lin is with the Department of Computer Science, Electrical Engineering, and Mathematical Sciences, Western Norway University of Applied Sciences, 5063 Bergen, Norway (e-mail: jerrylin@ieee.org).

Digital Object Identifier 10.1109/TETCI.2020.3032701

Results in CoBotAGV

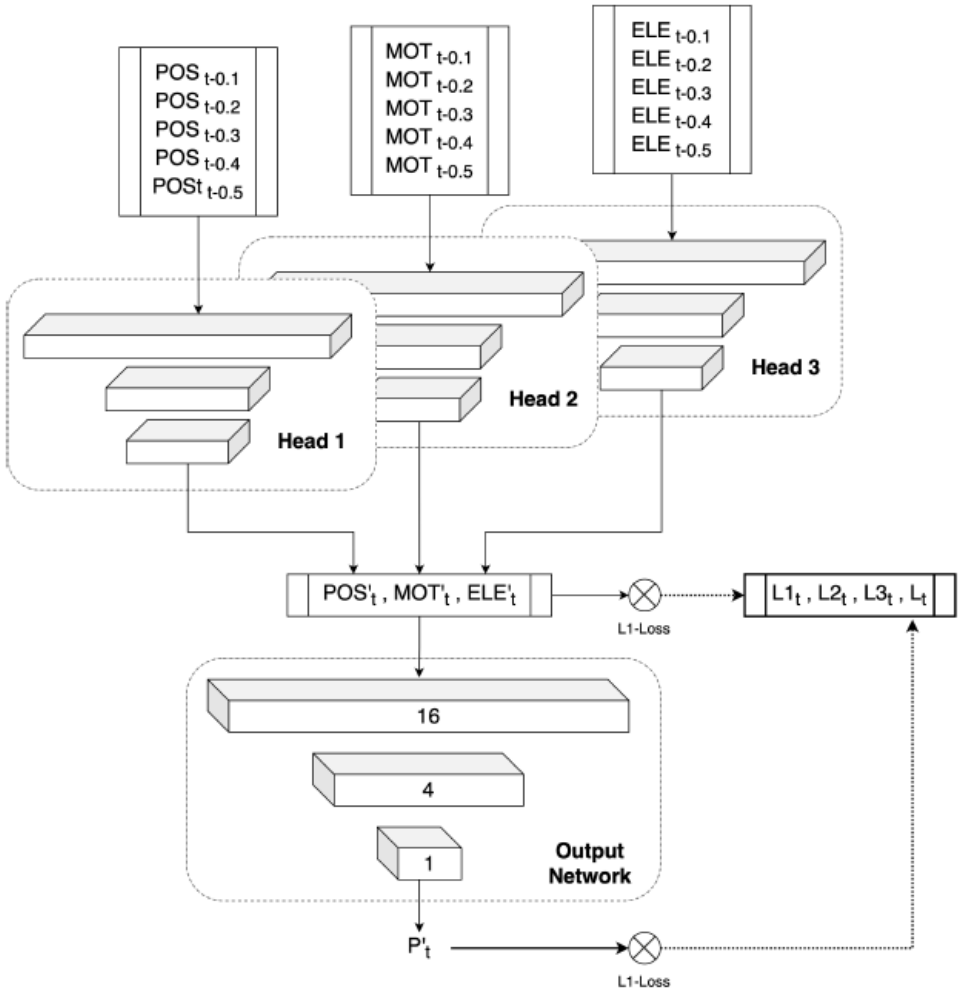


Figure 2: Architecture of the 2SMH

Multi-Head Learning Model for Power Prediction of Uncrewed Ground Vehicles

Anonymous Author(s)
Affiliation
Address
email

Abstract

Predicting the power performance of unmanned vehicles is an emerging topic, and real-time prediction models have gained widespread acceptance. However, some models focus only on limited features, while others directly incorporate all features into the prediction model. There are inevitable tradeoffs between feature diversity and interference, and between feature engineering and prediction efficiency. In this work, we propose a 2-stage multi-head learning (2SMH) that predicts future features over multiple head networks and then aggregates them as input to the subsequent network to predict power consumption. Experimental results show that the proposed 2SMH significantly outperforms the benchmark, with a maximum Mean Absolute Error (MAE) of 30.4%. The 2SMH also exhibits significant improvement in few-shot and zero-shot prediction, demonstrating its outstanding generality and robustness. Moreover, the developed transfer mechanisms MAE show a reduction of 18.2% and 13.7%, respectively. In summary, the 2SMH achieves excellent MAEs and demonstrates remarkable transferability and predictability.

1 Introduction

With the rapid and encouraging development of data mining and machine learning, power prediction is an emerging research topic that includes smart city, smart buildings, and electric vehicles. As for large-scale power prediction, it is indispensable in smart cities and can be used as a decision-making tool for energy management of smart power grids [1], including the discussions of dynamic pricing [2] and operational reserves [3]. In the smaller scale of smart buildings, techniques from smart meters, sensor systems [4], and energy storage [5] are often used for damage identification [6] and energy efficiency [7].

Due to the rapid development of electric automobiles, the power prediction of electric automobiles is also an important topic, including battery research [8], route information [9], and driving behavior analysis [10]. As for unmanned ground, air, and surface vehicles, various implementations for aerial photography, ocean exploration, and smart factories [11] will play a role. Due to limited battery capacity, real-time power prediction for unmanned vehicles is an important issue.

Research on smart cities and buildings mainly focuses on system design of IoT data collection with feature engineering. Feature preprocessing can delay real-time predictions, leading to the use of traditional and simple models. For electric vehicles, short-term power prediction models are already widely used. However, some models focus only on limited data sources (e.g., only driving behavior [10], route information [9], or battery [8]). On the other hand, some models use multiple data sources but feed all features directly into the prediction model, which can lead to noise interference between features. In summary, there are inevitable tradeoffs between feature diversity and interference, and between feature engineering and prediction efficiency.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

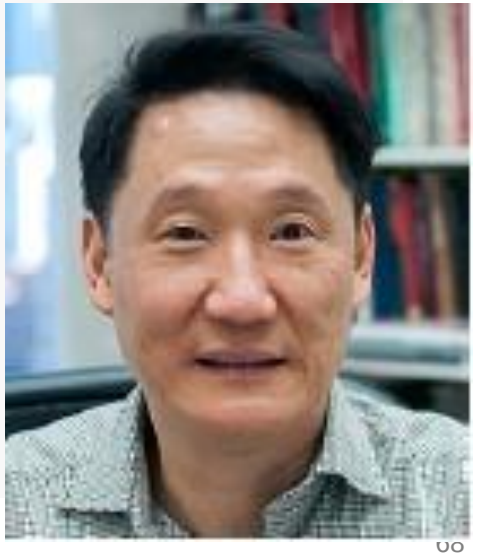
Special thanks to all the collaborators in the team











Thank you, any questions?

Jerry Chun-Wei Lin

jerrylin@ieee.org
IKE Lab: http://ikelab.net