finding frequent subsequences from set of sequences.

↑

Sequential Pattern Mining (SPM)

Sequence = Set of ordered events

represented by < >

Challenge → finding all subsequences

Example:-

CID	TID	Transactions
1	100	a, b, c, d
3	111	a, f, d, e
1	122	d, e, p
3	133	b, f, s, a

Sequential data

CID	Sequences
1	< (abcd), (dep) >
3	< (afde), (bfsa) >

sid	Sequence
10	< a (abc) (ac) d (cf) >
20	< (ad) c (bc) (ae) >
30	< (ef) (ab) (df c) b >
40	< eg (af) c b c >

min supp = 2

< (ab) c > ✓

< eg > ✗

- SPM vocabulary
  - Itemset (non-empty set of items)
  - Sequence (ordered list of events) order matters
  - Event (itemset (unordered list of items))

↓

$(I_2\ I_1\ I_3)$ where I_1, I_2, I_3 belongs to set of item I.

e_1 = abc

e_2 ade

s = < (abc) (ade) >

length = 6

↓

no. of instances of items in a sequence.

1 - Sequence → Length 1

Sequence with lengh-1 is called 1-pattern.

- Subsequence (α) is a sequence which is part of another sequence (β)

Agr kuch integers exist krtay hain jisme α key elements β main order main dikhte hain, tou α is β ka subsequence and β is α supersequence

$\alpha = \langle (ab) d \rangle$
$\beta = \langle (abc) (d.e) \rangle$
α subsequence of β

$\beta = \langle (abd) \rangle$ ✗

- Sequence database
  collection of sequences and often stores as tuples

  $\langle SID, S \rangle$
  SID ↓ identifier
  S → sequence.

- Support of a sequence in a sequence database.

- Support of sequence (α) in a sequence database (S) is (no. of tuples that contains sequence (α))

calculated as no. of times sequence appears in a database.

- Frequent Sequence.
  . Sequence (α) is considered frequent if its support is greater or equal to specified minimum support threshold.

Algo:- Apriori based (GSP)
Pattern-growth method (Freespan & prefixspan)
Vertical format based mining (SPADE)
Mining Closed sequential pattern (Clospan)

Sequence pattern.

Example:- min_supp=2

length of 1 is 9 & there are 3 items in the first 3 events from 1, it contributes to supp (a) to be $\boxed{1}$.

SID	Sequence
1	<a(abc)(ac)d(cf)>
2	"
3	"
4	"

~~Subsequence & support~~

Example:-

<a(bc)df> is sub sequence of <a(abc)(ac)d(cf)>

It means ais kay events [redacted] sequence main order kay Sath aa rahay hain in

Challenges of SPM:-

- Many number of possible sequential pattern are hidden within large datasets. which makes mining complex

Notes / Uni example.

Example:- min- supp = 3

SID	Sequence
1	< (bd) cb (ac) >
2	< (bf) (ce) b (fg) >
3	< (ah) (bdf) a bf >
4	< (be) (ce) d >
5	< a (bd) bcb (ade) >

1)

Candidate	Support
a	3
b	5
c	4
d	4
e	3
f	2 X
g	1 X
h	1 X

→

candidate	Support.
a	3
b	5
c	4
d	4
e	3

~~2)~~

~~2- length Seq (k)~~	
~~ab~~	~~2~~
~~ac~~	~~2~~
~~ad~~	
~~ae~~	
~~(a,b)~~	
~~(a,c)~~	
~~(a,d)~~	
~~(a,e)~~	

2) 2 events Candidate.

	<a>	<b>	<c>	<d>	<e>
<a>	aa	ab	ac	ad	ae
<b>	ba	bb	bc	bd	be
<c>	ca	cb	cc	cd	ce
<d>	da	db	dc	dd	de
<e>	ea	eb	ec	ed	ec

Candidate	Supp	Validity	Candidate	Supp	Validity
aa	2		dd	2	
ab	2		de	1	
ac	1		ea	0	
ad	2		eb	1	
ae	1		ec	1	
ba	3	✓	ed	1	
bb	4	✓	ee.	1	
bc	4	✓			
bd	2				
be	3	✓			
ca	2				
cb	3	✓			
cc	1				
cd	2				
ce	1				
da	3	✓			
db	3	✓			
dc	2				

3) Generating 1-event Candidate.

~~a,b~~
~~a,c~~
~~a,d~~
~~a,e~~
~~a,f~~
~~a,g~~
~~a,h~~
~~b,c~~
~~b,d~~
~~b,e~~
~~b,f~~
~~b,g~~
~~b,h~~
~~c,d~~

Candidate	Supp	Validity
<(a,b)>	0	
<(a,c)>	1	
<(a,d)>	1	
<(a,e)>	1	
<(b,c)>	0	
<(b,d)>	3	✓
<(b,e)>	1	
<(c,d)>	0	
<(c,e)>	2	
<(d,e)>	1	

Frequent 2 - Sequences
<ba>
<bb>
<bc>
<be>
<cb>
<da>
<db>
<(b,d)>

→ for 4th step

→ for 5th step

4) Generating 3-event Candidate

	<a>	<b>	<c>	<d>	<e>
<ba>	baa	bab	bac	bad	bae
<bb>	bba	bbb	bbc	bbd	bbe
<bc>	bca	bcb	bcc	bcd	bce
<be>	bea	beb	bec	bed	bee
<cb>	cba	cbb	cbc	cbd	cbe
<da>	daa	dab	dac	dad	dae
<db>	dba	dbb	dbc	dbd	dbe

<baa	0		beb	1		dba	2
<bab	1		bec	0		dbd	1
<bac	0		bed	1		dbe	1
<bad	0		bee	0			
<bae	0		cba	2			
<bba	2		cbb	0			
bbb	1		cbc	1			
bbc	1		cbd	0			
bbd	1		cbe	1			
bbe	1		daa	0			
bca	2		dab	1			
bcb	3	✓	dac	0			
bcc	1		dad	0			
bcd	2		dae	0			
bce	1		dba	2			
bea	0		db[illegible]	1			

5) 2 event candidates,

		a	b	c	d	e
b,d	>	(b,d)a	(b,d) b	(b,d) e	(b,d) d	(b,d)e

(beb), (b,d)

(b,d) a	3	✓
(b,d) b	2	X
(b,d) c	2	X
(b,d) d	1	X
(b,d) e	1	X