

PROBLEMI CON GLI ALGORITMI APRIORI-UPS

- 1) Troppi set di candidati. Ci vuole tempo.
- 2) Trope scans del DB
- 3) Long Sequential Pattern complessi da gestire.

FPSS-SPAN

È una estensione del FP-GROWTH per analizzare i sequential pattern.

3) F-LIST

Prendono tutti gli item singolarmente e ordinareli per ordine decrescente di supporto. Considerando solo quelli con $\text{support} \geq \text{min-sup}$. Questo è f-list.

Tutte le sequenze possono essere divise in sottoinsiemi senza overlapping.

Esempio

Sequence Database SDB

```
< (bd) c b (ac) >  
< (bf) (ce) b (fg) >  
< (ah) (bf) a b f >  
< (be) (ce) d >  
< a (bd) b c b (ade) >
```

f_list: b:5, c:4, a:3, d:3, e:3, f:2

All seq. pat. can be divided into 6 subsets:

- Seq. pat. containing item *f*
- Those containing *e* but no *f*
- Those containing *d* but no *e* nor *f*
- Those containing *a* but no *d*, *e* or *f*
- Those containing *c* but no *a*, *d*, *e* or *f*
- Those containing only item *b*

f-list = < b:5, c:4, a:3, d:3, e:3, f:2 >

PROIEZIONI

Un set completo di sequenziali pattern che contiene l'item i ma non gli item che seguono i nel f-list è detto **i -projection** del DB.

Esempio:

Prendiamo la transazione $\langle (a\ b) (b\ f) a\ b\ f \rangle$ La f -projection è formata da tutti gli item della transazione che stanno prima di f nel f-list:

$$\langle a (b\ f) a\ b\ f \rangle$$

a-projection: $\langle a\ b\ a\ b \rangle$.

PROIEZIONI PARASSIA

Scannerizzando una volta il DB sequenziale, deriviamo tutte le proiezioni i-projection.

Supponiamo che in media le transazioni del DB contengano L item frequenti. Una transazione è proiettata su $L-1$ proiezioni diverse del DB. Queste proiezioni contengono $1+2+\dots+(L-1) = \frac{L(L-1)}{2}$ items.

Quindi la dimensione di una i-projection è $\frac{L(L-1)}{2}$ volte quella del DB. TROPPO GROSSA.

PARTIZIONE DELLE PROIEZIONI

Invece di proiettare tutto il DB in uno scan, ci concentriamo nel proiettare la sequenza dell'ultimo oggetto frequente della f-list. (Nel nostro esempio vuol dire concentrarsi su f)

Questo lo facciamo per ogni transazione.

Dopo lo scan abbiamo che l'intero DB è stato proiettato. **Non abbiamo replicazione**.

Una volta ottenuta la proiezione, la usiamo per le successive analisi.

Esempio

$\langle bf \rangle$	$\langle ce \rangle$	$\langle b fg \rangle$	b: 2	c: 1
$\langle ah \rangle$	$\langle bf \rangle$	$\langle b f \rangle$	f: 2	e: 1
$\langle d a \rangle$			a: 2	

Iniziamo analizzando le proiezioni dell'Item meno frequente (Ma sempre sopra min-supp).

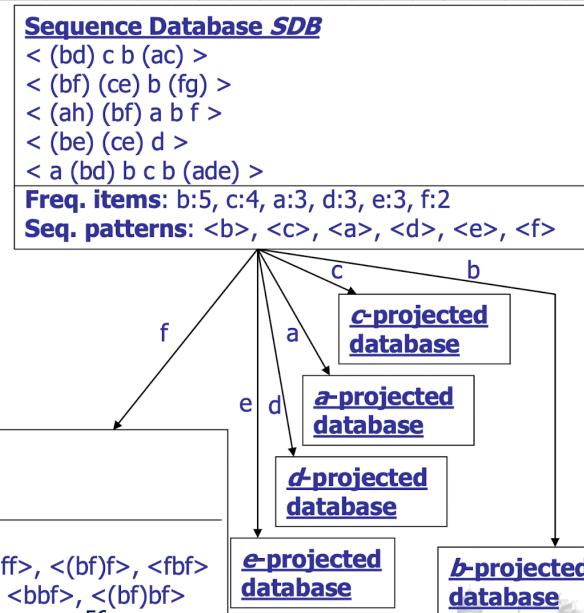
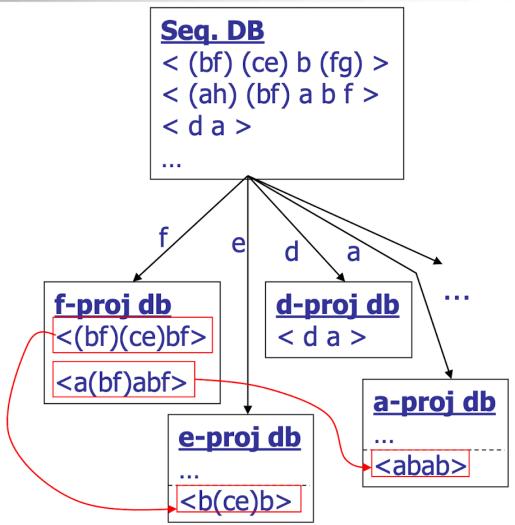
f-projection:

$$\begin{array}{l} \langle bf \rangle \langle ce \rangle \langle b f \rangle \\ \langle a(bf) \rangle \langle abf \rangle \end{array}$$

D'ora in poi usiamo la f-projection per ricavare la alta. RISPARMIAMO COSÌ MEMORIA. Nella nostra

$$f\text{-list} = \langle b:5, c:4, a:3, d:3, e:3, f:2 \rangle$$

Ora proiettiamo la e-projection usando le transazioni nelle f-projection: $\langle b (ce) b \rangle$ da



f-projection:

$\langle (\underline{bf})(\underline{ce}) \underline{b} \underline{f} \rangle$
 $\langle a (\underline{bf}) a \underline{b} \underline{f} \rangle$

Nella f-projection gli unici item frequenti sono:

b e f.

Da qui ricaviamo i possibili frequent pattern:

$\underline{\langle bf \rangle}$, $\underline{\langle fb \rangle}$, $\underline{\langle (bf) \rangle}$, $\underline{\langle ff \rangle}$, $\underline{\langle (bf)f \rangle}$, $\underline{\langle fbf \rangle}$

NOTA: Non posso avere come frequent pattern $\langle bb \rangle$ perché qui stiamo analizzando le sequenze temporali perciò in questo caso fra b e b abbiamo f.

ONE WORD SCAN

Per considerare $\{b, b, f\}$:

$\langle (bf) b \rangle, \langle \underline{bbf} \rangle, \langle (bf) bf \rangle$
?

?

.

RENDERE IL TUTTO PIÙ VELOCIS

Iniziare a generare due frequent pattern e poi da qui copiare quelli alti creare in modo diatto, che abbiano un'alta probabilità di essere frequenti

FREQUENT ITEM MATRIX

Matrice triangolare, che considera le relazioni fra items. Ogni elemento della matrice è formato da tre valori $\langle A, B, C \rangle$:

- A: numero di volte in cui i occurre dopo f : $\langle if \rangle$
- B: numero di volte in cui f occurre dopo i : $\langle fi \rangle$
- C: numero di volte in cui f occurre insieme a i : $\langle (if) \rangle$

Lungo la diagonale abbiamo il numero di volte in cui occorre la sequenza $\langle ii \rangle$.

Considerando la sequenza $\langle (bd) c b (ac) \rangle$ possiamo aumentare $F[b, c] + = 1$ per A e B ma non per C perché la sequenza (b, c) non c'è.

$F[b, c] = (4, 3, \emptyset)$					
b	c	a	d	e	f
4	1				
b	c	a	d	e	f
(4, 3, 0)					
a (3, 2, 0)	(2, 1, 1) 2				
d (2, 2, 2)	(2, 2, 0) (1, 2, 1) 1				
e (3, 1, 1)	(1, 1, 2) (1, 0, 1) (1, 1, 1) 1				
f (2, 2, 2)	(1, 1, 0) (1, 1, 0) (0, 0, 0) (1, 1, 0) 2				

Sequence Database SDB

- < (bd) c b (ac) >
- < (bf) (ce) b (fg) >
- < (ah) (bf) a b f >
- < (be) (ce) d >
- < a (bd) b c b (ade) >

Come usiamo questa matrice?

Lo usiamo per generare i 2-Itemset frequenti e il set di proiezioni del nostro DB, che saranno usati per creare i 3-Itemset.

Per farlo mettiamo da una parte due tipi di annotazioni:

- 1) item-repeated patterns
- 2) projection DB.

GENERAZIONE DI 2-LENGTH SEQUENCES

Per ogni contatore $\langle A, B, C \rangle$ se il suo valore non è sotto la threshold min-supp allora lo in output il sequential pattern corrispondente:

$F[b, a] = \langle 3, 2, \emptyset \rangle$ se min-supp=2					
b	c	a	d	e	f
4	1				
b	c	a	d	e	f
(4, 3, 0)					
a (3, 2, 0)	(2, 1, 1) 2				
d (2, 2, 2)	(2, 2, 0) (1, 2, 1) 1				
e (3, 1, 1)	(1, 1, 2) (1, 0, 1) (1, 1, 1) 1				
f (2, 2, 2)	(1, 1, 0) (1, 1, 0) (0, 0, 0) (1, 1, 0) 2				

\Rightarrow allora l'output è:

$\langle bac \rangle: 3$ e $\langle cab \rangle: 2$

NOTAZIONI DELLE ANNOTAZIONI

Prima di capire come usare la matrice dei frequent pattern, definiamo le varie notazioni che utilizzeremo:

- 1) $\{ \dots \}$: sequenza in qualsiasi ordine
- 2) $\langle \dots \rangle$: sequenza nell'ordine specificato
- 3) a^+ : + per indicare che ci sono più occorrenze di a se non presente (+) allora " a " appare una volta

COME GENERARE LE ANNOTAZIONI

Tre modi diversi:

1) Generare length-2 sequential pattern

Semplicemente per ogni counter (Intacta della matrice) e sopra la soglia di min-sup, generare i corrispondenti frequent pattern:

$$F[b,a] = \langle 3, 2, 0 \rangle \implies \langle ba \rangle : 3, \langle ab \rangle : 2$$

2) Generare annotazioni per item-repeating pattern per la riga j :

Guardando la diagonale:

SE $F[i, j] \geq \text{min-sup}$ ALLORA generare $\langle ii^+ \rangle$

Per le colonne $i \neq j$

SE $F[i, i] \geq \text{min-sup}$ ALLORA bisogna aggiungere all'annotazione i^+

SE $F[i, j] \geq \text{min-sup}$ ALLORA aggiungere j^+

Se solo uno dei contatori di $F[i, \cdot]$ supera min-sup
Allora allora aggiungere la sequenza a cui fa riferimento nelle annotazioni

Esempio

	Sequence Database SDB					
	b	c	a	d	e	f
b	4					
c	(4, 3, 0)	1				
a	(3, 2, 0)	(2, 1, 1)	2			
d	(2, 2, 2)	(2, 2, 0)	(1, 2, 1)	1		
e	(3, 1, 1)	(1, 1, 2)	(1, 0, 1)	(1, 1, 1)	1	
f	(2, 2, 2)	(1, 1, 0)	(1, 1, 0)	(0, 0, 0)	(1, 1, 0)	2
	b	c	a	d	e	f

Prendiamo $F[f, f] = 2$, con min-sup=2 dobbiamo generare l'annotazione $\{f^+ f^+\}$

Guardando le colonne diverse da f, che la riga f contiene, solo "b" e "a" hanno $F[i, i] \geq \text{min-sup}$

$$\{b^+ a^+\}$$

Anche $F[f, f] \geq \text{min-sup}$, possiamo aggiungere f^+

$$\{b^+ a^+ f^+\}$$

Ora però analizzando il contenuto di $F[f, b]$ e $F[f, a]$ solo la prima ha elementi sopra min-sup quindi dovremo NON CONSIDERARE a^+ . Da qui, dell'analisi delle sole righe f essa fuori l'annotazione

$$\{b^+ f^+\}$$

Annotiamo $\{b^+ f^+\}$ perché $F[f, b] = (2, 2, 2)$ quindi

sia la sequenza $\langle b f \rangle$ che $\langle f b \rangle$ soddisfa il min-supp, perciò con $\{ \dots \}$ facciamo riferimento a entrambi gli ordinamenti.

Item	Output length-2 sequential patterns	Ann. on repeating items
f	$\langle bf \rangle : 2, \langle fb \rangle : 2, \langle (bf) \rangle : 2,$	$\{b^+ f^+\}$
e	$\langle be \rangle : 3, \langle (ce) \rangle : 2$	$\langle b^+ e \rangle$
d	$\langle bd \rangle : 2, \langle db \rangle : 2, \langle (bd) \rangle : 2, \langle cd \rangle : 2, \langle dc \rangle : 2, \langle da \rangle : 2$	$\{b^+ d\}, \langle da^+ \rangle$
a	$\langle ba \rangle : 3, \langle ab \rangle : 2, \langle ca \rangle : 2, \langle aa \rangle : 2$	$\langle aa^+ \rangle, \{a^+ b^+\}, \langle ca^+ \rangle$
c	$\langle bc \rangle : 4, \langle cb \rangle : 3,$	$\{b^+ c\}$
b	$\langle bb \rangle : 4$	$\langle bb^+ \rangle$

3) Generazione di annotazioni sulle proiezioni del DB per la riga j :

Per ogni $i < j$, scegliere un $K < i$ e prendere

$$F[i, j] - F[K, j] - F[i, K]$$

Se guardando queste tre F , si può formare un pattern triple ALLORA dovranno aggiungere K alla proiezione di i (i -projection)

In parole povere si sta cercando di costruire un pattern con 3 item sulla base delle informazioni contenute nelle varie $F[\ast, \ast]$.

Esempio

Se prendiamo $F[e, f] = (1, g, \emptyset)$, $F[a, f] = (3, \emptyset, \emptyset)$ e $F[a, e] = (1, 1, 3)$, non possono generare pattern con tre item perché NON SODDISFA il vincolo che $\langle f e \rangle$, $\langle a f \rangle$ e $\langle (ae) \rangle$ possono esistere contemporaneamente

nello stesso 3-sequence.

Se combiniamo $F[a, e] = (3, 1, 1)$ ora possiamo generare un 3-pattern valido:

$$\left. \begin{array}{l} F[e, f] = (1, \underline{g}, \emptyset) \\ F[a, f] = (\underline{3}, \emptyset, 1) \\ F[a, e] = (\underline{3}, 1, 1) \end{array} \right\} \Rightarrow \langle \underline{a} \underline{f} \underline{e} \rangle$$

Esempio matrice di prime

Considerando la riga f non ci sono combinazioni di appropriati $F[\ast, \ast]$ in grado di generare 3-pattern

Se consideriamo $f = e$, prendendo $c = b$ e $K = C$:

$$\left. \begin{array}{l} F[b, c] = (4, 3, \emptyset) \\ F[b, e] = (3, 3, 1) \\ F[c, e] = (1, 1, 2) \end{array} \right\} \Rightarrow (ce)b \mid b(ce)$$

Possiamo annotare $\langle ce \rangle : b$ ovvero le ce-projection che possiamo creare con la sola aggiunta di "b".

Risultato totale:

Item	Output length-2 sequential patterns	Ann. on repeating items	Ann. on Projected DBs
f	$\langle bf \rangle : 2, \langle fb \rangle : 2, \langle (bf) \rangle : 2,$	$\{b^+ f^+\}$	\emptyset
e	$\langle be \rangle : 3, \langle (ce) \rangle : 2$	$\langle b^+ e \rangle$	$\langle (ce) \rangle : \{b\}$
d	$\langle bd \rangle : 2, \langle db \rangle : 2, \langle (bd) \rangle : 2, \langle cd \rangle : 2, \langle dc \rangle : 2, \langle da \rangle : 2$	$\{b^+ d\}, \langle d a^+ \rangle$	$\langle da \rangle : \{b, c\}, \langle cd \rangle : \{b\}$
a	$\langle ba \rangle : 3, \langle ab \rangle : 2, \langle ca \rangle : 2, \langle aa \rangle : 2$	$\langle a a^+ \rangle, \langle a^+ b^+ \rangle, \langle c a^+ \rangle$	$\langle ca \rangle : \{b\}$
c	$\langle bc \rangle : 4, \langle cb \rangle : 3,$	$\{b^+ c\}$	\emptyset
b	$\langle bb \rangle : 4$	$\langle b b^+ \rangle$	\emptyset