

TD2 : Sequential Pattern Analysis

Exercice 1 :

Clients	Caddies
C1	a,b, c
C2	b,c
C3	a,c,d
C2	d,a,c
C1	b,c,d
C3	a,b,c
C3	a
C1	b,c

- 1) Changer la base de données vers une base séquentielle.
- 2) En utilisant l'algorithme GSP, retrouver les candidats ainsi que les motifs fréquents, sachant que le $\text{supp_min} = 3$.

Solution :

1)

C1	<(a,b,c),(b,c,d),(b,c)>
C2	<(b,c),(d,a,c)>
C3	<(a,c,d),(a,b,c),a>

2)

<a>	3
	3
<c>	3
<d>	3

	a	b	c	d
A	<a,a> :1	<a,b> :2	<a,c> :2	<a,d> :1
B	<b,a> :2	<b,b> :1	<b,c> :2	<b,d> :2
C	<c,a> :2	<c,b> :2	<c,c> :3	<c, d> :2
D	<d,a> :1	<d,b> :2	<d,c> :2	<d,d> :0

	a	b	c	d
A		<(a,b)> :2	<(a,c)> :3	<(a,d)> :2
B			<(b,c)> :3	<(b,d)> :1
C				<(c,d)> :3
D				

Les motifs fréquents de taille 2:<(a,c)>, <(a,b)>,<(b,c)>,<(d,c)>

Les motifs fréquents sont :<a>,,<c>,<d>, <c,c>, <(a,c)>,<(b,c)>,<(d,c)>

Exercice2 :

Sid	Sequence
1	< a(ac)(adc) >
2	< (ba)(fb)a >
3	< (ab)bfb(ae) >
4	< a(af)d >
5	< d(fac) >
6	< (adf)(ae) >

- 1) En utilisant l'algorithme GSP, retrouver les candidats ainsi que les motifs fréquents, sachant que le $\text{supp_min} = 3$.

<a>	6
	2
<c>	2
<d>	4
<e>	2
<f>	5

	a	d	f
A	<a,a> :5	<a,d> :1	<a,f> :3
D	<d,a> :2	<d,d> :0	<d,f> :1
F	<f,a> :3	<f,d> :1	<f,f> :0

	a	d	f
A		<(a,d)> :2	<(a,f)> :3
D			<(d,f)> :1
F			

Les motifs fréquents de taille 2 : <a,a>, <f,a>,<a,f> et <(a,f)>

<aaa>	1
<aaf>	0
<faa>	0
<afa>	2
<faf>	0
<(af)a>	1
<a(af)>	1

Les motifs séquentiels fréquents : <a>, <f>, <d>, <aa>, <f,a>, <a,f>, <(a,f)>

2) Appliquer l'algorithme PrefixSpan sur la même base et avec le même supp_min.

Sid	Sequence
1	< a(ac)(adc) >
2	< (ba)(fb)a >
3	< (ab)bfb(ae) >
4	< a(af)d >
5	< d(fac) >
6	< (adf)(ae) >

<a>	6
	2
<c>	2
<d>	4
<e>	2
<f>	5

Sid	Sequence
1	< a(a)(ad) >
2	< (a)(f)a >
3	< (a)f(a) >
4	< a(af)d >
5	< d(af) >
6	< (adf)(a) >

Projection selon le <a> prefix

<a>_Prefx :

<a(ad)>

<fa>

<fa>

<(af)d>

<(_f)>

<(_df)a>

	<a>	<d>	<(_d)>	<f>	<_f>
supp	5	2	2	3	3

Donc les sous séquences fréquentes de taille 2 commençant par **a** sont : <aa>, <af>, <(af)>

Donc on continue en projetant sur chacune de ces sous séquences fréquentes

<aa>_Prefix :

<a>

<>

<>

<(_f)>

<>

<>

Rq : à partir de cette projection nous n'avons pas de sous séquence fréquentes de taille 3 qui commence par la sous séquence <aa>.

Projection en utilisant le préfixe <af>

<af>_Prefix :

<>

<a>

<a>

<>

<>

<>

	<a>
supp	2

Donc, il n'y a pas de sous séquences de taille 3 commençant par <af>

Projection en utilisant le préfixe <(af)>

<(af)>_Pref :

<>

<>

<>

<>

<a>

Donc il n'y a pas de sous séquence de taille 3 qui commence par <(af)>

Projection selon le <d> prefix

<d>Prefix :

<>

<>

<>

<>

<af>

<(_f)a>

Donc , il n'y a pas de sous-séquence de taille 2 qui commence par **d**

Projection selon le <f> prefix

<>

<a>
 <a>
 <d>
 <>
 <a>

	<a>	<d>
supp	3	1

Donc la sous séquence de taille 2 commençant par <f> est <fa>

Projection selon <fa>

<>
 <>
 <>
 <>
 <>
 <>

Conclusion : Les motifs fréquents sont :

<a>, <d>, <f>, <aa>, <fa>, <af>, <(af)>