

VERİ MADENCİLİĞİ İlişkilendirme Kuralları: Farklı Yöntemler

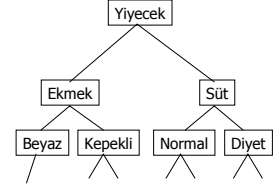
Yrd. Doç. Dr. Şule Gündüz Öğüdücü

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

1

Çok Katmanlı İlişkilendirme Kuralları

- Öğeler bir hiyerarşi oluşturabilir.
 - kavram hiyerarşisi
- Daha alt katmanlarda yer alan ürünlerin destek değeri daha düşük
- Alt seviyelerde oluşan kurallar özel durumları tanımlıyor
 - Beyaz ekmek => normal süt
 - Kepekli ekmek => normal süt
 - Kepekli ekmek => diyet süt



<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

2

Çok Katmanlı İlişkilendirme Kuralları

- Destek ve güven değerlerinin katmanlar arasında değişimi
 - $X1$ ve $X2$ öğelerinin üst düğümünde yer alan öğe X
 $\sigma(X) \geq \sigma(X1) + \sigma(X2)$
 - $\sigma(X1 \cup Y1) \geq \text{minsup}$, $X1$ öğesinin üst düğümünde yer alan öğe X , $Y1$ öğesinin üst düğümünde yer alan öğe Y
 $\sigma(X \cup Y1) \geq \text{minsup}$, $\sigma(X1 \cup Y) \geq \text{minsup}$
 $\sigma(X \cup Y) \geq \text{minsup}$
 - $\text{conf}(X1 \Rightarrow Y1) \geq \text{minconf}$ ise $\text{conf}(X1 \Rightarrow Y) \geq \text{minconf}$

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

3

Çok Katmanlı İlişkilendirme Kuralları

- 1. Yaklaşım:
Her hareket t , üst katmanlardaki öğeler cinsinden yazılarak genişletilmiş harekete t' dönüştürülür
 $t = \{\text{beyaz ekmek, normal süt}\}$
 $t' = \{\text{beyaz ekmek, normal süt, ekmek, süt, yiyecek}\}$
 t' hareketleri için ilişkilendirme kuralları bulunur
- Problemler:
 - Üst düğümlerde yer alan öğelerin destek değerleri daha büyük
 - Farklı seviyelerde farklı destek değerleri
 - Veri kümesinin boyutu artıyor

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

4

Çok Katmanlı İlişkilendirme Kuralları

- Alt katmanlarda destek değerini küçültme
 - Her katmanda bağımsız hesaplama (Level-by-level independent)
 - Üst düğümleri yaygın öğe olan öğelerin destek değerini hesaplama
 - Tek öğeye göre budama (Level-cross filtering by single item)
 - Öğeler kümesine göre budama (Level-cross filtering by k-itemset)

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

5

Çok Katmanlı İlişkilendirme Kuralları

- 2. Yaklaşım
İlk olarak en üst katman için yaygın öğeler bulunur
Her aşamada bir alt katmandaki öğeler arasındaki ilişkiler bulunur
- Problemler:
 - Kullanışlı değil: her katman için veri kümesi araştırılıyor
 - Farklı katmanlar arasındaki ilişkiler bulunamıyor

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

6

Nicel İlişkilendirme Kuralları

Kayıt No	Yaş	Medeni Durum	Araba Sayısı
1	23	Bekar	1
2	25	Evli	1
3	29	Bekar	0
4	34	Evli	2
5	38	Evli	2
...	...	...	...

- Örnek ilişkilendirme kuralları:
yaş:30-39, medeni durum: evli => araba sayısı:2
(destek: %40, güven: %100)
araba sayısı: 0-1 => medeni durum: bekar
(destek: %40, güven: %66,70)

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

7

Nicel İlişkilendirme Kuralları

- İkili değişkenlerden oluşan ilişkilendirme kurallarına dönüştürülüyor:
 - Kategorik olmayan nitelik değerleri için bölmeleme
 - yaş: 20-29, 30-39
 - ikili değişken oluşturma
 - kategorik niteliklerin her değeri için bir ikili değişken
 - kategorik olmayan niteliklerin her bölümü için bir ikili değişken

Kayıt No	Yaş	Medeni Durum	Araba Sayısı
1	23	Bekar	1
5	38	Evli	2

Kayıt No	Yaş 20-29	Yaş 30-39	M.Durum Evli	M.Durum Bekar	Araba 0	Araba 1	Araba 2
1	1	0	0	1	0	1	0
5	0	1	1	0	0	0	1

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

8

Nicel İlişkilendirme Kuralları

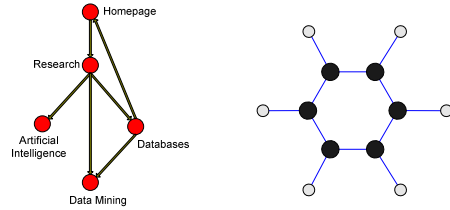
- Dönüştürmeden kaynaklanan problemler
 - bölme sayısı az: Bilgi kaybolması
 - Destek değeri küçük: fazla sayıda kural

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

9

Yaygın Altçizge Madenciliği

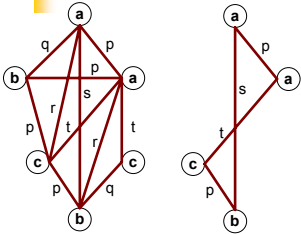
- Frequent Subgraph Mining
 - İlişkilendirme kuralları madenciliği yöntemleri uygulanarak yaygın altçizgeler bulunur
- Uygulama alanları
 - Web madenciliği
 - Biyoenformatik
 - :



<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

10

Çizge Tanımları



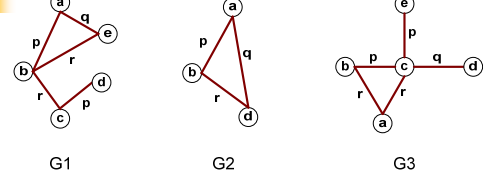
(a) Etiketli Çizge (b) Altçizge

- $V = \{x_i\}$ düğümler kümesi
 - $E = \{w_{ij}\}$ x_i ve x_j düğümleri arasındaki ağırlık
 - Altçizge: $G(V_i, E_i)$
- $$\bigcup_{i=1}^k V_i = V$$
- $$E_i = \{\{u, v\} \in E \mid u, v \in V_i\}$$

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

11

Çizge-Hareket Dönüşümü



	(a,b,p)	(a,b,q)	(a,b,r)	(b,c,p)	(b,c,q)	(b,c,r)	...	(d,e,r)
G1	1	0	0	0	0	1	...	0
G2	1	0	0	0	0	0	...	0
G3	0	0	1	1	0	0	...	0
G3	...	...	...	...	...	...	...	...

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

12

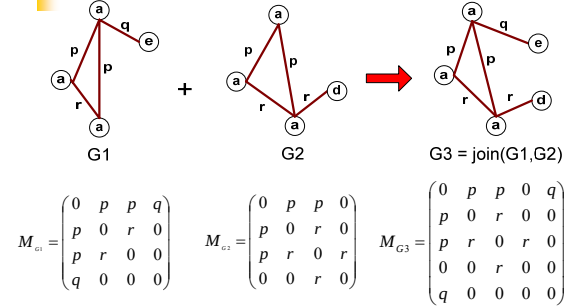
Apriori Temelli Yaklaşım

- Bir altçizgenin güven değeri:
 - Altçizgenin bulunduğu çizgelerin toplam çizgelere oranı
- Apriori temelli yaklaşım
 - yaygın (k+1)-altçizgeleri bulmak için k-altçizgelerin kullanılması
 - düğüm ekleme: k düğüm sayısı
 - ayırt ekleme: k ayırt sayısı

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

13

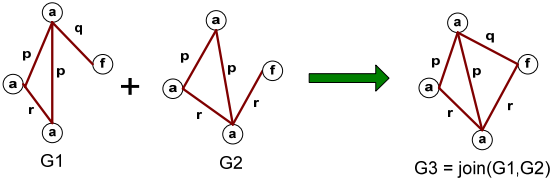
Düğüm Ekleme



<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

14

Ayırt Ekleme



<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

15

Apriori Temelli Algoritma

- Yaygın 1-altçizgelerin belirlenmesi
- Yeni yaygın altçizge oluşmayana kadar tekrarla
 - Aday oluşturma: Yaygın (k-1) altçizgeleri kullanarak k-altçizgeler oluşturma
 - Aday azaltma: (k-1)-altçizgesi yaygın olmayan adayların silinmesi
 - Destek hesaplaması: Kalan adaylar için destek değerinin hesaplanması
 - Aday eleme: min_sup değerini sağlamayan adayların silinmesi

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

16

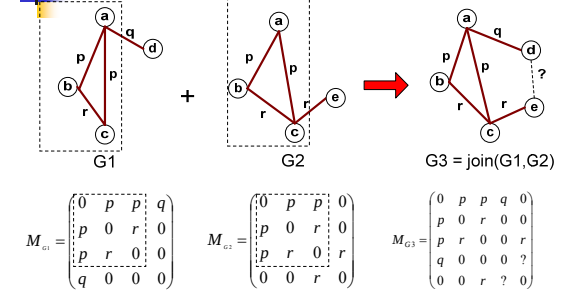
Aday Oluşturma

- Apriori yöntemi:
 - İki (k-1)-yaygın öğelerin birleştirilmesi sonucu sadece bir k-yaygın öğeler oluşur.
- Altçizge madenciliği
 - İki (k-1)-yaygın altçizgenin birleştirilmesi sonucu birden fazla k-yaygın altçizge oluşur
 - iki (k-1)-yaygın altçizgenin birleştirilmesi için (k-2)-altçizgelerinin aynı olması gerekir (çekirdek)

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

17

Düğüm Ekleme: Farklı Aday Oluşması

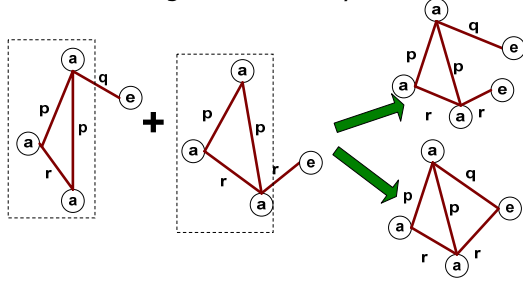


<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

18

Ayrıt Ekleme: Farklı Aday Oluşması

- 1. durum: düğüm etiketleri aynı

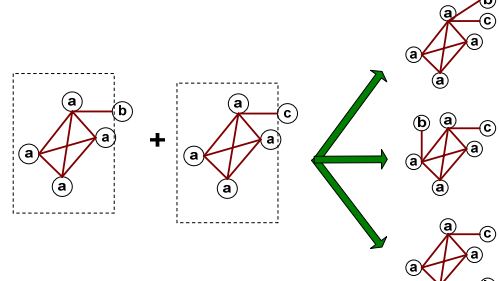


<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

19

Ayrıt Ekleme: Farklı Aday Oluşması

- 2. durum: çekirdekteki düğüm etiketleri aynı



<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

20

Çizge Madenciliği

- Jun Huan, Wei Wang, Jan Prins, and Jiong Yang, *SPIN: Mining maximal frequent subgraphs from graph databases*. Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD), 2004.
- X. Yan, X. J. Zhou, and J. Han, *Mining Closed Relational Graphs with Connectivity Constraints*, Proc. 2005 Int. Conf. on Knowledge Discovery and Data Mining (KDD'05), Chicago, IL, Aug. 2005.

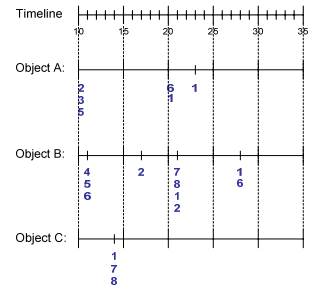
<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

21

Sıralı Dizi Verisi

Object	Timestamp	Events
A	10	2, 3, 5
A	20	6, 1
A	23	1
B	11	4, 5, 6
B	17	2
B	21	7, 8, 1, 2
B	28	1, 6
C	14	1, 8, 7

- Örnek sıralı dizi verisi:
 - müşteri hareketleri
 - tıbbi veriler
 - web verileri
 - telefon görüşme kayıtları
 - DNA dizileri



<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

22

Sıralı Dizi Tanımı

Sıralı dizi veritabanı

SID	sıralı dizi
10	<a(abc)(ac)d(cf)>
20	<(ad)c(bc)(ae)>
30	<(ef)(ab)(df)cb>
40	<eg(af)cbc>

- Elemanların sıralı listesi (<a(abc)(ac)d(cf)>)
- Elemanlar öğeler kümesi, öğeler sıralı değildir, genelde alfabetik sırada sıralanırlar (abc)
- Elemanlar zaman veya yer ile nitelendirilir
- Sıralı dizinin uzunluğu |s|: Sıralı diziyi oluşturan öğeler sayısı
- k-sıralı dizi: k öğesi olan sıralı dizi

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

23

Sıralı Dizi Örnekleri

- Web verileri:

< {Homepage} {Electronics} {Digital Cameras} {Canon Digital Camera} {Shopping Cart} {Order Confirmation} {Return to Shopping} >

- Kütüphaneden alınan kitaplar:

< {Fellowship of the Ring} {The Two Towers} {Return of the King} >

- DNA dizileri

ACAAGATGCCATTGTCCCCCGCCTCTGTCTGTCTCTCCGGGGCCACGGCC

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

24

Aldiz Tanımı

- $\langle b_1 b_2 \dots b_m \rangle$ sıralı dizisi içinde $\langle a_1 a_2 \dots a_n \rangle$ sıralı dizinin bulunabilmesi için ($m \geq n$):
 - $i_1 < i_2 < \dots < i_n$ koşulunda $a_1 \subseteq b_{i_1}, a_2 \subseteq b_{i_2}, \dots, a_n \subseteq b_{i_n}$
 - $\langle a(abc)(ac)d(cf) \rangle$ sıralı dizinin bir altdizisi: $\langle a(bc)dc \rangle$
 - $\langle ad \rangle, \langle (ad)c(bc)(ae) \rangle$ sıralı dizinin altdizisi değil
- Bir altdizinin destek değeri: Altdizinin bulunduğu sıralı dizilerin toplam sıralı dizi sayısına oranı
- Sıralı dizi örüntüsü: Destek değeri en küçük destek değerine (minsup) eşit ya da büyük olan altdiziler

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

25

Sıralı Dizi Madenciliği

- Sıralı dizi veritabanında en az *minsup* destek değerine sahip altdizilerin bulunması (sıralı dizi örüntülerinin bulunması)

sıralı dizi veritabanı

SID	sıralı dizi
10	$\langle a(abc)(ac)d(cf) \rangle$
20	$\langle (ad)c(bc)(ae) \rangle$
30	$\langle (ef)(ab)(df)cb \rangle$
40	$\langle eg(af)cbc \rangle$

minsup = 2 için $\langle (ab)c \rangle$ bir sıralı dizi örüntüsü

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

26

Zorluklar

- Veri kümesi içinde çok büyük sayıda sıralı dizi örüntüsü bulunabilir
 - Sıralı dizi madenciliği algoritmasının bütün sıralı dizi örüntülerini bulması gerekir
 - Veri kümesini az sayıda tarayarak sonuç üretmesi gerekir
 - ölçeklenebilir
 - Değişik veri tipleri üzerinde çalışabilmesi gerekir

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

27

Sıralı Dizi Örüntülerinin Temel Özelliği

- Apriori:
 - Sıralı dizi s , *minsup* değerini sağlamıyorsa, s 'in hiçbir üstkümesi de *minsup* değerini sağlamaz.
 - Örnek: $\langle hb \rangle$ *minsup* değerini sağlamıyor $\rightarrow \langle hab \rangle$ ve $\langle (ah)b \rangle$ sıralı dizileri de *minsup* değerini sağlamaz.

SID	Sıralı dizi
10	$\langle (bd)cb(ac) \rangle$
20	$\langle (bf)(ce)b(fg) \rangle$
30	$\langle (ah)(bf)abf \rangle$
40	$\langle (be)(ce)d \rangle$
50	$\langle a(bd)bcb(ade) \rangle$

minsup = 2

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

28

GSP (A Generalized Sequential Pattern Mining Algorithm)

- Ramakrishnan Srikant, Rakesh Agrawal, *Mining Sequential Patterns: Generalizations And Performance Improvements*, Proc. 5th Int. Conf. Extending Database Technology, EDBT'96
- 1. Aşama:
 - Veritabanını bir kez tarayarak 1 öğeden oluşan sıralı dizi örüntülerini bulunur
- 2. Aşama: Yeni bir sıralı dizi örüntüsü bulunmayınca kadar tekrarla
 - Aday oluşturma: $(k-1)$. adımda bulunan sıralı dizi örüntüleri birleştirilerek k uzunluklu sıralı dizi örüntü adayları bulunur
 - Aday azaltma: $(k-1)$ altdizileri *minsup* değerini sağlamayan adaylar elenir
 - Destek değeri hesaplama: Kalan sıralı dizi örüntü adayları için veritabanı taranarak destek değeri hesaplanır
 - Aday eleme: *minsup* değeri sağlamayan sıralı dizi örüntü adayları elenir.

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

29

GSP: Aday Oluşturma

- İlk durum ($k=2$): Bir uzunluklu sıralı dizilerden iki uzunluklu sıralı diziler oluşturma
 - $s_{11} = \langle i_1 \rangle$ ve $s_{12} = \langle i_2 \rangle$ sıralı dizilerin birleştirilmesi sonucunda $s_{21} = \langle i_1 i_1 \rangle, s_{22} = \langle i_1 i_2 \rangle, s_{23} = \langle i_2 i_2 \rangle, s_{24} = \langle i_2 i_1 \rangle$ ve $s_{25} = \langle i_1 i_2 \rangle$ sıralı dizileri oluşur
- Genel durum ($k > 2$):
 - k uzunluklu s_{k1} ve s_{k2} sıralı dizileri şu şartlar altında birleştir:
 - s_{k1} sıralı dizinin ilk öğesi ve s_{k2} sıralı dizinin son öğesi silindiğinde oluşan altdizinin aynı olması
 - $(k+1)$ uzunluklu sıralı dizi örüntü adayı s_{k3} sıralı diziye s_{k2} sıralı dizinin son öğesinin eklenmesi ile bulunur
 - s_{k1} sıralı dizinin son iki öğesi aynı elemana ait: s_{k3} sıralı dizinin son öğesi s_{k1} sıralı dizisinin son elemanının parçası olur
 - diğer durumda s_{k1} sıralı dizisinin son öğesi s_{k2} sıralı dizisinin son öğesi ile aynı olur.

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

30

Örnek

- $s_1 = \langle \{a\} \{b\} \{c\} \{d\} \rangle$ ve $s_2 = \langle \{b\} \{c\} \{d\} \{e\} \rangle$ sıralı dizileri birleştirilirse
 - aday sıralı dizi $\langle \{a\} \{b\} \{c\} \{d\} \{e\} \rangle$
 - s_2 sıralı dizideki son iki öge (d ve e) aynı elemana ait
- $s_1 = \langle \{a\} \{b\} \{c\} \{d\} \rangle$ ve $s_2 = \langle \{b\} \{c\} \{d\} \{e\} \rangle$ sıralı dizileri birleştirilirse
 - aday sıralı dizi $\langle \{a\} \{b\} \{c\} \{d\} \{e\} \rangle$
 - s_2 sıralı dizideki son iki öge (d ve e) aynı elemana ait değil
- $s_1 = \langle \{a\} \{b\} \{f\} \{d\} \rangle$ ve $s_2 = \langle \{a\} \{b\} \{d\} \{e\} \rangle$ sıralı dizileri $\langle \{a\} \{b\} \{d\} \{e\} \rangle$ sıralı dizi örüntü adayını oluşturmak üzere birleştirilemez
 - $s_2 = \langle \{b\} \{f\} \{d\} \{x\} \rangle$ veya $s_2 = \langle \{b\} \{f\} \{d\} \{x\} \rangle$ şeklinde olmadığı için

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

31

Örnek

- Bir öğeden oluşan dizilerin bulunması
- Adaylar
 - $\langle a \rangle, \langle b \rangle, \langle c \rangle, \langle d \rangle, \langle e \rangle, \langle f \rangle, \langle g \rangle, \langle h \rangle$
- Veritabanı taranarak destek sayıları hesaplanır

$min_sup = 2$

SID	Sıralı Dizi
10	$\langle (bd)cb(ac) \rangle$
20	$\langle (bf)(ce)b(fg) \rangle$
30	$\langle (ah)(bf)abf \rangle$
40	$\langle (be)(ce)d \rangle$
50	$\langle a(bd)bcb(ade) \rangle$

Aday	Des
$\langle a \rangle$	3
$\langle b \rangle$	5
$\langle c \rangle$	4
$\langle d \rangle$	3
$\langle e \rangle$	3
$\langle f \rangle$	2
$\langle g \rangle$	1
$\langle h \rangle$	1

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

32

Örnek

- iki uzunluklu sıralı dizilerin bulunması
- 51 aday

	$\langle a \rangle$	$\langle b \rangle$	$\langle c \rangle$	$\langle d \rangle$	$\langle e \rangle$	$\langle f \rangle$
$\langle a \rangle$	$\langle aa \rangle$	$\langle ab \rangle$	$\langle ac \rangle$	$\langle ad \rangle$	$\langle ae \rangle$	$\langle af \rangle$
$\langle b \rangle$	$\langle ba \rangle$	$\langle bb \rangle$	$\langle bc \rangle$	$\langle bd \rangle$	$\langle be \rangle$	$\langle bf \rangle$
$\langle c \rangle$	$\langle ca \rangle$	$\langle cb \rangle$	$\langle cc \rangle$	$\langle cd \rangle$	$\langle ce \rangle$	$\langle cf \rangle$
$\langle d \rangle$	$\langle da \rangle$	$\langle db \rangle$	$\langle dc \rangle$	$\langle dd \rangle$	$\langle de \rangle$	$\langle df \rangle$
$\langle e \rangle$	$\langle ea \rangle$	$\langle eb \rangle$	$\langle ec \rangle$	$\langle ed \rangle$	$\langle ee \rangle$	$\langle ef \rangle$
$\langle f \rangle$	$\langle fa \rangle$	$\langle fb \rangle$	$\langle fc \rangle$	$\langle fd \rangle$	$\langle fe \rangle$	$\langle ff \rangle$

Apriori özelliği kullanılmadan aday sayısı
 $8 * 8 + 8 * 7 / 2 = 92$

Apriori aday sayısını
 %44.57 azaltıyor

	$\langle a \rangle$	$\langle b \rangle$	$\langle c \rangle$	$\langle d \rangle$	$\langle e \rangle$	$\langle f \rangle$
$\langle a \rangle$		$\langle ab \rangle$	$\langle ac \rangle$	$\langle ad \rangle$	$\langle ae \rangle$	$\langle af \rangle$
$\langle b \rangle$			$\langle bc \rangle$	$\langle bd \rangle$	$\langle be \rangle$	$\langle bf \rangle$
$\langle c \rangle$				$\langle cd \rangle$	$\langle ce \rangle$	$\langle cf \rangle$
$\langle d \rangle$					$\langle de \rangle$	$\langle df \rangle$
$\langle e \rangle$						$\langle ef \rangle$
$\langle f \rangle$						

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

33

Örnek

- 3 uzunluklu adayların bulunması
 - 2 uzunluklu sıralı dizilerden 3 uzunluklu adaylar oluşturma
 - Apriori özelliği kullanılıyor
 - $\langle ab \rangle, \langle aa \rangle$ ve $\langle ba \rangle$: 2 uzunluklu sıralı diziler
 $\rightarrow \langle aba \rangle$: 3 uzunluklu sıralı dizi
 - $\langle (bd) \rangle, \langle bb \rangle$ ve $\langle db \rangle$: 2 uzunluklu sıralı diziler
 $\rightarrow \langle (bd)b \rangle$: 3 uzunluklu sıralı dizi
 - 46 aday oluşuyor
- 3 uzunluklu sıralı dizilerin bulunması
 - Veritabanı taranarak adayların destek sayıları hesaplanır
 - 46 adaydan 19 3 uzunluklu sıralı dizi örüntüsü bulunur

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

34

GSP Örnek

1. adım: 8 aday, 6 adet 1-sıralı dizi örüntüsü.
2. adım: 51 aday, 19 adet 2-sıralı dizi örüntüsü. 10 aday DB'de değil
3. adım: 46 aday, 19 adet 3-sıralı dizi örüntüsü. 20 aday DB'de değil
4. adım: 8 aday, 6 adet 4-sıralı dizi örüntüsü
5. adım: 1 aday, 1 adet 5-sıralı dizi örüntüsü

$minsup = 2$

ID	Sıralı dizi
10	$\langle (bd)cb(ac) \rangle$
20	$\langle (bf)(ce)b(fg) \rangle$
30	$\langle (ah)(bf)abf \rangle$
40	$\langle (be)(ce)d \rangle$
50	$\langle a(bd)bcb(ade) \rangle$

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

35

GSP Örnek

3-sekans örüntüleri

$\langle \{1\} \{2\} \{3\} \rangle$
$\langle \{1\} \{2\} \{5\} \rangle$
$\langle \{1\} \{5\} \{3\} \rangle$
$\langle \{2\} \{3\} \{4\} \rangle$
$\langle \{2\} \{5\} \{3\} \rangle$
$\langle \{3\} \{4\} \{5\} \rangle$
$\langle \{5\} \{3\} \{4\} \rangle$

Aday Oluşturma

$\langle \{1\} \{2\} \{3\} \{4\} \rangle$
$\langle \{1\} \{2\} \{5\} \{3\} \rangle$
$\langle \{1\} \{5\} \{3\} \{4\} \rangle$
$\langle \{2\} \{3\} \{4\} \{5\} \rangle$
$\langle \{2\} \{5\} \{3\} \{4\} \rangle$

Aday Eleme

$\langle \{1\} \{2\} \{5\} \{3\} \rangle$

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

36

GSP Algoritmasında Problemler

- Çok sayıda sıralı dizi örüntü adayı oluşturulabilir
 - 1000 adet 1-sıralı dizi örüntülerinden $1000 \times 1000 + \frac{1000 \times 999}{2} = 1,499,500$ adet sıralı dizi örüntü adayı oluşur.
- Veritabanı birçok kez taranır
- Uzun sıralı dizileri belirlemek çok zor
 - sıralı diziler ard arda eklenerek daha uzun sıralı diziler oluşuyor
 - 100-sıralı dizi oluşturmak için 10^{30} sıralı dizi gerekiyor.

$$\sum_{i=1}^{100} \binom{100}{i} = 2^{100} - 1 \approx 10^{30}$$

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

37

Sıralı Dizi Örüntü Adayı Oluşturma

- Parçala ve çöz yaklaşımı
 - veritabanı yinelemeli küçültülür
 - oluşan her veritabanında sıralı dizi örüntüleri bulunur
- Algoritmalar:
 - J. Han J. Pei, B. Mortazavi-Asi, Q. Chen, U. Dayal and M.C. Hsu. *FreeSpan: Frequent pattern-projected sequential pattern mining*. KDD'00
 - J. Pei, J. Han, B. Mortazavi-Asi, H. Pinto, Q. Chen, U. Dayal, and M.C. Hsu. *Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth*. In Proceedings of the 17th International Conference on Data Engineering, 2001.

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

38

FreeSpan Algoritması

- Örnek: Sıralı dizi veritabanı ve min_sup=2

Sıralı dizi Veritabanı

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

- 1. Adım: 1-sıralı dizi örüntülerinin bulunması ve destek değeri küçülen sırada sıralanmaları (f-list)

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

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

39

FreeSpan Algoritması

- 2. Adım: Arama uzayının bölünmesi. Sıralı dizi veritabanı altkümelerine bölünür
 - f* ögesi bulunan altdiziler
 - e* ögesi bulunan *f* ögesi bulunmayan
 - d* ögesi bulunan *e* ve *f* öğeleri bulunmayan
 - a* ögesi bulunan *d*, *e* ve *f* öğeleri bulunmayan
 - c* ögesi bulunan *a*, *d*, *e* ve *f* öğeleri bulunmayan
 - b* ögesi bulunan
- Her altdizi (α) için bu altdizilerin bulunduğu sıralı dizilerden oluşan veritabanları, α -yansımali veritabanları (α -projected database), oluşturulur.

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

40

FreeSpan Algoritması

- c* ögesi bulunan *a*, *d*, *e* ve *f* öğeleri bulunmayan sıralı diziler
 - $\langle c \rangle$ -yansımali veritabanı: $\langle b c b c \rangle$, $\langle b c b \rangle$, $\langle b c \rangle$, $\langle b b c b \rangle$
 - c* ögesi bulunan *a*, *d*, *e* ve *f* öğeleri bulunmayan bütün altdizilerin bulunması: $\langle bc \rangle$: 4, $\langle cb \rangle$: 3
 - Veritabanı yansıtılarak sıralı diziler bulunmaya devam edilir.

Sıralı dizi Veritabanı

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

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

41