VERİ MADENCİLİĞİ

Fırat İsmailoğlu, PhD

Birliktelik Kuralları
(Association Rules)

Birliktelik Kuralları Giriş

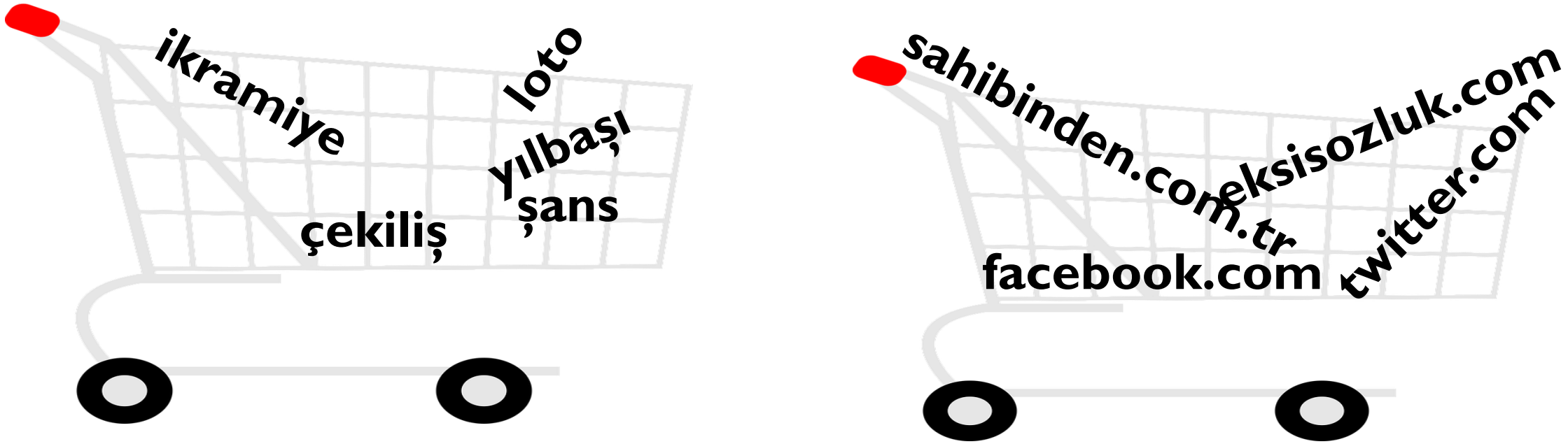
Birliktelik kuralları yada birliktelik analizi nesneler arası bağlantıların ortaya çıkarılması ve bir kural olarak ortaya konulması için vardır.

Birliktelik kuralları en yaygın olarak **pazar sepeti analizi** (market basket analysis) için kullanılır. Kişilerin alışveriş sepetleri incelenerek birlikte alınan ürünler tespit edilir. Sonuç olarak bu ürünü alan şu ürünü de aldı gibi tespitler yapılır. Dahası bu analizlere göre market rafları daha iyi bir şekilde düzenlenir, kişiye özel kampanyalar yapılabilir.

Birliktelik Kuralları Giriş

Market basket analizine ilaveten birliktelik kurallarının diğer onemli kullanim alani metin madenciliğidir. Birliktelik kuralları analizi ile bir metindeki sıklıklıa beraber kullanılan kelimleri tespit edebiliriz. Yada bir metinde en çok tekrar eden kelimleri (anahtar kelimleri) bulabiliriz.

Ayrıca web madenciliğinide (web mining) de sıkça birlikte kuralları analizi kullanılır. Böylece birlikte ziyaret edilen siteler açığa çıkarılmış olur.



Birliktelik Kuralları Giriş

Birliktelik kurallarındaki en meşhur örnek ekmek, süt, bebek bezi, yumurta, kola ve biradan oluşan pazar sepeti incelemesidir. Buarada 5 tane işlem (transaction) olmuştur. Her bir işlemin kendine ait bir işlem no'su vardır. Her bir küme birlikte alınan ürünleri gösterir.

İşlem No	Ürünler (Items)
1	{Ekmek, Süt}
2	{Ekmek, Çocuk bezi, Bira, Yumurta}
3	{Süt, Çocuk bezi, Bira, Kola}
4	{Ekmek, Süt, Çocuk bezi, Bira}
5	{Ekmek, Süt, Çocuk bezi, Kola}

Buradan çıkartacağımız bir birliktelik kuralı:

Çocuk bezi → Bira

Buradan çıkaracağımız sonuç çocuk bezi alanlar bira da alırlar.

Birliktelik Kuralları Problemleri

Birliktelik kuralları analizinde genelde iki büyük problemle karşılaşırız.

1. Bulunan bazı kurallar yapay olabilir; yani birlikteliğe konu olan ürünler şanstan yanyana gelmiş olabilir.

   Örnek olarak bir önceki veri setinden

   $$\text{Bira} \rightarrow \text{Yumurta}$$

   elde edilen bir kural gerçekçi olmaz. Bira ve yumurta yalnız bir kez yanyana gelmiştir. Genelleme yapamayız.
2. İşlem sayısının çok olduğu büyük veri setlerinden birliktelik kuralları elde etmek hesaplama açısından zor bir olaydır.

Sonuç olarak birliktelik analizi yapılırken tesadufi olarak bir araya gelmiş ürünler elenmeli, çıkartilabilecek tüm kuralları buulabilmek için brute-force harici akıllıca bir uöntem kullanılmalıdır.

Birliktelik Kuralları ve Nedensellik

Birliktelik kuralları bize nedensellik (causality) sağlamaz. Yani örneğin

$$\text{Çocuk bezi} \rightarrow \text{Bira}$$

kuralında çocuk bezinin alınması bira alınmasına neden olmaz.

Bu kuralın bize verdiği bilgi çocuk bezi ile bira arasında güçlü bir ilişkinin olduğu, çocuk bezi alanlarin bira almaya eğilimli olduğudur.

Birliktelik Kuralları Genel Form

Birliktelik kurallarının genel formu :

$$X \rightarrow Y$$

dir. Burada X ve Y ürünler kümesinin (itemset) ayrık birer altkümesidir. $X \cap Y = \emptyset$.

X ve Y'yi ayrık kabul ettiğimizden

$$\text{Çocuk bezi} \rightarrow \text{Bira, Çocuk bezi}$$

gibi bir kural elde etmeyiz.

Destek ve Güven (Support and Confidence)

Bir birliktelik kuralının kalitesi iki kriter ile ölçülür: destek ve güven.

Destek

Destek, bir kuralı oluşturan ürünlerin işlem listesindeki görülme oranıdır. Bir anlamda kuralın yaygınlığının ölçüsüdür.

Matematiksel gösterim için:

X ürünler kümesinin bir alt kümesi olmak üzere, $\sigma(X)$, N tane işlemdeki görülme sayısı olsun. Bu durumda $X \rightarrow Y$ formundaki bir kuralın desteği:

$$destek(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N}$$

- $\sigma(X \cup Y)$ → X ve Y'nin birlikte görüldüğü işlemler sayısı
- N → Toplam işlem sayısı

ile bulunur.

Desteği yüksek olan bir kuralda ürünlerin bir arada görülmesi şans ile açıklanamaz, bu ürünler arasında bir ilişki olduğu açık olur.

Destek

İşlem No	Ürünler (Items)
1	{Ekmek, Süt}
2	{Ekmek, Çocuk bezi, Bira, Yumurta}
3	{Süt, Çocuk bezi, Bira, Kola}
4	{Ekmek, Süt, Çocuk bezi, Bira}
5	{Ekmek, Süt, Çocuk bezi, Kola}

Yukarıdaki veri setinde toplam 5 işlemin 3'unde çocuk bezi ve bira beraber satın alınmıştır. O halde Çocuk bezi → Bira kuralının desteği:

$$destek(\text{Çocuk bezi} \quad \text{Bira}) = \frac{3}{5} = 0.6$$

Süt, Çocuk bezi → Bira kuralı için

$$destek(\text{Süt, Çocuk bezi} \rightarrow \text{Bira}) = \frac{2}{5} = 0.4$$

Güven

Güven, bir kurala ne kadar güvendiğimizin ölçüsüdür. Eğer $X \rightarrow Y$ gibi bir kuralın güveni yüksekse, bu X'i gördüğümüzde Y'yi de görme ihtimalimizin fazla olduğu anlamına gelir.

$X \rightarrow Y$ formundaki bir kuralın güveni:

$$güven(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

- $\sigma(X \cup Y)$ → X ve Y'nin birlikte görüldüğü işlemler sayısı
- $\sigma(X)$ → Yalnzca X'in görüldüğü işlem sayısı

Bir önceki veri setinde işlemlerin 4'unde çocuk bezi alinmistir. Bu 4 işlemin 3'ünde çocuk bezi ile beraber bira da alınmıştır. O halde Çocuk bezi → Bira kuralının güveni:

$$güven(\text{Çocuk bezi} \rightarrow \text{Bira}) = \frac{3}{4} = 0.75$$

Süt, Çocuk bezi → Bira kuralı için

$$güven(\text{Süt, Çocuk bezi} \rightarrow \text{Bira}) = \frac{2}{3} = 0.66$$

Güven

Not: $güven(X \rightarrow Y)$ ile $güven(Y \rightarrow X)$ aynı şey değildir. $güven(X \rightarrow Y)$; X ürünü alındığında Y'nin de alınmasının ne kadar olası olduğudur. Öte yandan $güven(Y \rightarrow X)$; Y ürünü alındığında X'in de alınmasının ne kadar olası olduğudur.

Örnek olarak $güven(\text{Çocuk bezi} \rightarrow \text{Bira}) = 0.75$ iken $güven(\text{Bira} \rightarrow \text{Çocuk bezi}) = 1$ olur. Çünkü bira alınan her işlemde çocuk bezi de alınır.

Birliktelik Kurallarının Üretilmesi

Amacımız, desteği daha önceden belirlediğimiz bir minimum destek değerinden (min_destek) yüksek ve güveni bir minimum güven değerinden $(min\_güven)$ yüksek tüm kuralları bulmaktır.

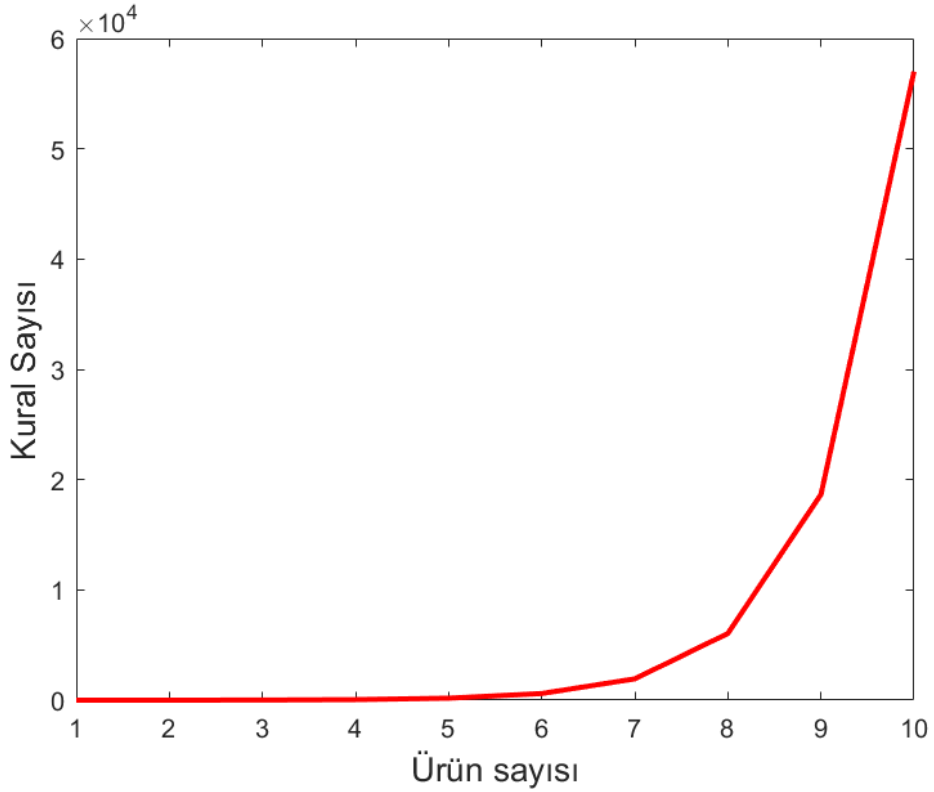
Bunun için yapabilecegimiz ilk şey önce kaba kuvvet (brute-force) yöntemiyle olabilecek bütün kuralları üretmek; daha sonra bunların içinden min_destek'ten yüksek desteği olan ve $min\_güven$'den yüksek güveni olan kuralları bulmaktır.

Birliktelik Kurallarının Üretilmesi

Teorem: d tane farklı üründen oluşan bir işlemler listesinden elde edilebilecek bütün kuralların sayısı

$$3^d - 2^{d+1} + 1$$

dir. Bu, ürün sayısı artıkça olabilecek kuralların sayısının eksponensiyel olarak arttığı anlamına gelir.



Birliktelik Kurallarının Üretilmesi

Örnekte gördüğümüz işlemler listesinde 6 tane ürün vardı. O halde üretilebilecek toplam kural sayısı $3^6 - 2^{6+1} + 1 = 602$ olur.

Fakat örneğin $min\_destek = 0.2$ ve $min\_güven = 0.5$ alındığında 602 kuralın $\%80$'nin destek ve güven değerleri minimum değerlerin altında kalır.

Şu halde kaba kuvvet ile üretilen olabilecek tüm kuralların birçoğu gereksizdir; bunları elemek gerekir.

Birliktelik kurallarının üretilmesi için genel yaklaşım iki aşamalıdır:

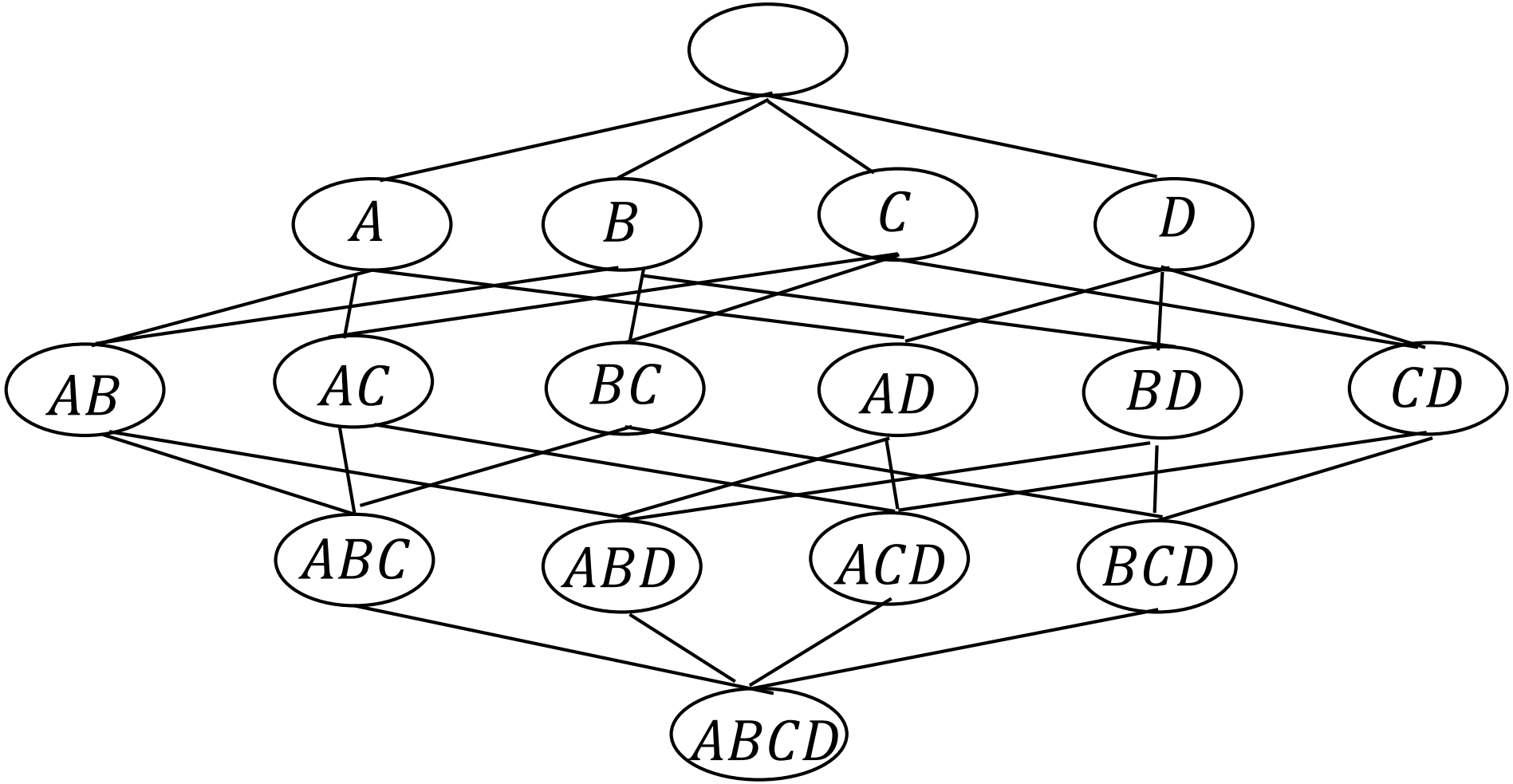
1. Destek değeri min_destek'ten daha büyük tüm ürün kümeleri (itemset) üretilir.
2. Birinci aşamada üretilen ürün kümeleri içinden güven değeri $min\_güven$'den daha yüksek olan kurallar üretilir.

Ürün Kümelerinin Üretilmesi

d tane üründen elde edilebilecek tüm ürün kümelerinin sayısı 2^d dir.

Örnek olarak diyelim ki elimizde 4 tane ürün olsun: A, B, C, D.

Bu 4 üründen elde edilebelicek tüm ürün kümelerinin sayısı (alt küme sayısı) $2^4 = 16$ dir.



Ürün Kümelerinin Üretilmesi

Görüldüğü gibi d tane üründen elde edilebilecek ürün seti çok fazladır (2^d). Dolayısıya bu alt kümelerin her birini uretip bunlarin destek değerlerini olçmek çok masraflidir. Bunu engellemk icin apriori prensibi göze alınır.

Apriori Prensibi

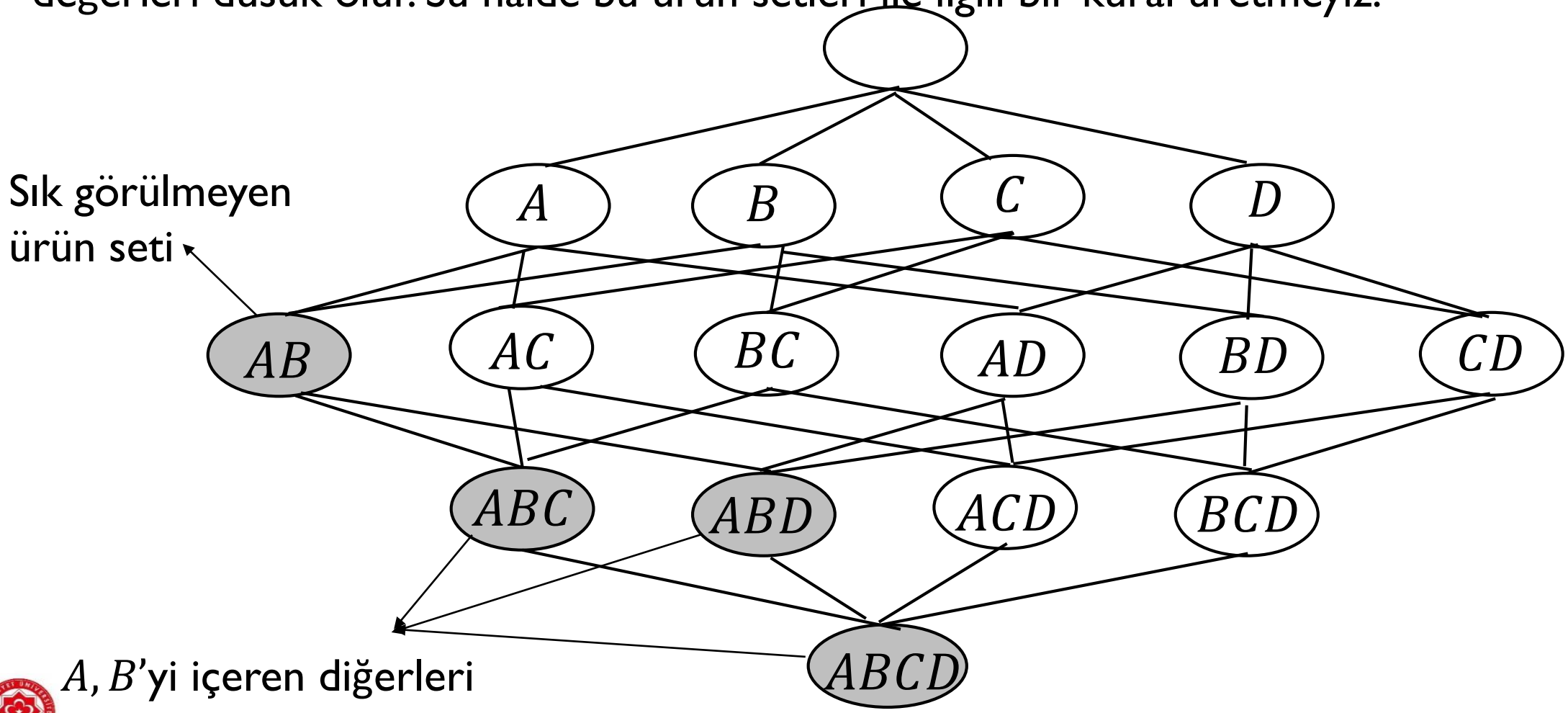
Eğer bir ürün kümesi sık görülmüyorsa, bu ürün kümesin üst kümeleri de sık görülmez (yani bu ürün kümesini içeren kümeler de sık görülmez).

Örneğin aşağıdaki veri setinde yumurta yanlızca bir kez görülmüştür. O halde yumurtayı içeren hicbir ürün kümesi sık görülmez, yani destek değerleri çok düşük olur.

İşlem No	Ürünler (Items)
1	{Ekmek, Süt}
2	{Ekmek, Çocuk bezi, Bira, Yumurta}
3	{Süt, Çocuk bezi, Bira, Kola}
4	{Ekmek, Süt, Çocuk bezi, Bira}
5	{Ekmek, Süt, Çocuk bezi, Kola}

Ürün Kümelerinin Üretilmesi

Diyelimki A, B, C, D ürünlerinden oluşan veri setinde A ve B çok sık bir arada görülmesin. Bu durumda A, B'yi iceren ABC, ABD ve $ABCD$ ürün setleri de pek sık görülmez, destek değerleri dusuk olur. Su halde bu urun setleri ile ilgili bir kural uretmeyiz.



A, B'yi içeren diğerleri kümelerde sık görülmez!

ör.

İşlem No	Ürünler (Items)
1	{Ekmek, Süt}
2	{Ekmek, Çocuk bezi, Bira, Yumurta}
3	{Süt, Çocuk bezi, Bira, Kola}
4	{Ekmek, Süt, Çocuk bezi, Bira}
5	{Ekmek, Süt, Çocuk bezi, Kola}

Veri seti için her bir urunun alınma adedi tablosu:

Ürün	Adet
Bira	3
Ekmek	4
Kola	2
Çocuk Bezi	4
Süt	4
Yumurta	1

Diyelimki $min\_destek = \%60$ olsun. Bu, bir ürün setinin sık görülmesi için tüm işlemlerin en az $\%60$'inda yer almasi demektir. Örnekte 5 işlem olduğu için, bir urunun yada urunlerin sık görüldüğünü söylemek için 5 islemin %60'i olan 3 işlemde görulmelidir.

Kola 2 ve yumurta yalnizca 1 islemde görüldüğünden bu urunleri içeren birliktelik kurali olusturamayiz. Bu urunleri cikartirsak iki urun içeren ürün kümeleri şöyle olur.

Ürün	Adet
{Bira,Ekmek}	2
{Bira,Çocuk Bezi}	3
{Bira,Süt}	2
{Ekmek ,Çocuk Bezi}	3
{Ekmek, Süt}	3
{Çocuk Bezi,Süt}	3

{Bira,Ekmek} ve {Bira,Süt} ikilileri 2 defa birlikte görüldüğünden bu urun kumelerini çıkartırız.
Su halde diğer ikililerden iki uzunluğunda aday kurallar uretebiliriliriz.

Örnegin Bira →Çocuk Bezi, Çocuk Bezi →Bira, Ekmek →Çocuk Bezi, Çocuk → Bezi Ekmek,… vb.
Üretilen bu aday kuralların güven değerleri ölçulmelidir; eğer güvenleri min _*güven*'den yüksekse birliktelik kurali olarak düşünebiliriz.

Kola ve yumurtayi en basta elemistik. Kalan urunler olan Ekmek, Cocuk bezi, Bira ve Sütü bir onceki asamada buldugumuz ikililere ekleyerek uc urun iceren urun kumelerini inceliyelim.

{Bira,Çocuk Bezi, Ekmek}
{Bira,Çocuk Bezi, Süt}
{Ekmek ,Çocuk Bezi,Bira}
{Ekmek ,Çocuk Bezi,Süt}
{Ekmek, Süt,Bira}
{Ekmek, Süt,Çocuk Bezi}
{Çocuk Bezi,Süt, Bira}
{Çocuk Bezi,Süt, Ekmek}

Yanda koyu renkle gösterilenler bir önceki asamada buldugumuz sık olamayan {Bira, Ekmek} ve {Bira, Süt} ikililerinden birini içerirler. O halde bu ürün setleri sık görülemez.

Açık renkte görülen ürün setleri ise aslında tamamen ayni ürünleri (Ekmek, Cocuk Bezi ve Süt) içeren ürün setidir; birbirinin aynısıdır.
Bu urun setinin bütün alt kümeli sık görüldüğünden kendininde sık görülme olasılığı vardır.