Büyük Veri Analitiği
(Big Data Analytics)

M. Ali Akcayol
Gazi Üniversitesi
Bilgisayar Mühendisliği Bölümü

Bu dersin sunumları, "Mining of Massive Datasets, Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman, Stanford University, 2011." kitabı kullanılarak hazırlanmıştır.

Konular

- On-line Reklam
- On-line Algoritmalar
  - On-line ve off-line algoritmalar
  - Greedy algoritmalar
- Eşleştirme Problemi
  - Mükemmel eşleştirme
  - Greedy algoritması ile maximal eşleştirme
- Adwords Problemi
  - Arama reklamcılığı
  - Adwords probleminin tanımı
  - Adwords problemi için greedy yaklaşımı
  - Balance algoritması

On-line Reklam

- **Web uygulamalarının çoğu** üyelik yerine **reklamcılık ile kendilerini desteklemektedir.**
- En karlı on-line reklamcılık Web üzerinde arama uygulamalarında yapılır.
- **En etkili arama reklamcılığı** ise **adwords modeliyle sağlanmaktadır.**
- **Adwords** modelinde **arama sorguları ile reklamlar arasında eşleştirme** yapılır.
- Sorgular ile reklamlar arasındaki **eşleştirmenin optimizasyonu için greedy** veya **on-line algoritmalar kullanılır.**
- **Diğer on-line reklamcılık problemi** ise, **reklam yapılacak item'ların belirlenmesidir.**
- Benzer müşteriler belirlenerek reklam yapılacak elemanlar belirlenir.
- Reklam yapılacak elemanlar **işbirlikçi filtreleme** (collaborative filtering) ile belirlenebilir.

3

On-line Reklam

- Web çok farklı yollarla reklamcıların müşterilerine ulaşmasını sağlar.
- **Bazı siteler reklamcılara doğrudan reklamlarını yayınlama olanağı sunar** (eBay, e-ticaret).
  - Bu siteler, ücretsiz, ücretli veya komisyon karşılığı bu servisi sağlarlar.
- **Reklamlar çok farklı Web sitelerinde yer alır.**
  - Reklamcılar sayfaların görüntülenmesi ve download edilmesi halinde ücret öderler.
- Bazı **on-line mağazalar** üreticilerden **herhangi bir ücret almaksızın reklam yapabilirler** (Amazon).
  - Bu siteler kendi müşterilerinin ilgisini çekeceğini düşündükleri ürünleri seçerler.
- **Ürün reklamları arama sonuçları arasına yerleştirilir.**
  - Reklamcılar kendi reklamlarının tıklanması halinde ücret öderler.
  - Reklamcılar sorgu kelimeleri için teklif verirler ve tıklanması halinde ücret öderler.

4

On-line Reklam

Reklamların doğrudan yayınlanması

- Reklamlar bir Web sitesinde doğrudan yayınlandığında, **sorgu kelimeleriyle eşleştirme gereklidir.**
- **Inverted indeks** oluşturulmalı ve sorgu **kelimelerinin tümü reklam içerisinde bulunmalıdır.**
- Alternatif olarak, **reklamcı kendi reklamı için parametreler belirleyebilir.**
- İkinci el oto reklamı için, marka, model, renk, … olabilir.
- Web link analizinde yapıldığı gibi **reklamların önemi belirlenemez.**
- En **güncel olanlar öncelikli** gösterilebilir.
- Diğer bir yöntemde **en çok ilgi gören öncelikli** gösterilebilir.
- **Başlangıçta sık tıklananların şansı her zaman yüksek olur.**

5

On-line Reklam

Reklamların doğrudan yayınlanması

- Reklamların değerlendirilmesi için farklı etkenler vardır.
- Bir reklamın sunulan **listedeki konumu tıklanma olasılığı için çok önemlidir.**
- Bir reklamın **ilgi çekici olması sorgu terimlerine bağlıdır.**
- Herhangi bir tıklanma olmadan önce tüm reklamların gösterilme şansı olmalıdır.
- **Web sayfasında reklamların doğrudan gösterilmesi klasik medya ile reklam yapmaya benzer.**
- **Çok kişi reklamı görür**, ancak çok **az sayıda kişi yapılan reklamla ilgilidir.**
- Daha çok **ilgili kişilere** reklamın **ulaşması için** konuya **özel yayınlarda gösterilmesi** gereklidir (bilişim dergileri, sağlık dergileri, …).

6

On-line Reklam

Reklamların doğrudan yayınlanması

- Web reklamcılığı basılı reklamcılığa göre bir çok avantaja sahiptir.
- Kullanıcı bilgilerine göre hangi reklamın gösterileceğine karar verilebilir.
- Bir kişinin bir konuya **ilgi düzeyi farklı parametrelerle belirlenebilir:**
  - **Sosyal medyada** ilgili alandaki bir **gruba üye olabilir.**
  - Konuyla **ilgili kelimeleri e-postalarında sık kullanabilir.**
  - Konuyla **ilgili arama sonuç sayfasında uzun zaman harcayabilir.**
  - Konuyla **ilgili kelimelerle çok sık arama yapabilir.**
  - Konuyla **ilgili ders, kurs gibi sayfaları bookmark yapmış olabilir.**

7

Konular

- On-line Reklam
- On-line Algoritmalar
  - On-line ve off-line algoritmalar
  - Greedy algoritmalar
- Eşleştirme Problemi
  - Mükemmel eşleştirme
  - Greedy algoritması ile maximal eşleştirme
- Adwords Problemi
  - Arama reklamcılığı
  - Adwords probleminin tanımı
  - Adwords problemi için greedy yaklaşımı
  - Balance algoritması

8

On-line Algoritmalar

- Arama sorgusundaki **kelimelere** göre **reklamların eşleştirilmesi** gereklidir.
- Bu eşleştirmeyi yapan algoritmalar **on-line algoritmalar** olarak ifade edilir.
- **On-line algoritmalar greedy yaklaşımını içerir.**
- Bu algoritmaların maximal matching yapması istenir.

9

Konular

- On-line Reklam
- On-line Algoritmalar
  - On-line ve off-line algoritmalar
  - Greedy algoritmalar
- Eşleştirme Problemi
  - Mükemmel eşleştirme
  - Greedy algoritması ile maximal eşleştirme
- Adwords Problemi
  - Arama reklamcılığı
  - Adwords probleminin tanımı
  - Adwords problemi için greedy yaklaşımı
  - Balance algoritması

10

On-line ve off-line algoritmalar

- Off-line algoritmaların özellikleri:
  - Tipik olarak **algoritmanın kullanacağı tüm veri başlangıçta sağlanır.**
  - Algoritma **veriye istediği sırada ve sayıda erişir.**
  - **En sonunda** algoritma **bir cevap üretir.**
- On-line algoritmaların özellikleri:
  - **Bazı durumlarda** algoritma **tüm veriyi görmeden karar vermek zorundadır.**
  - Stream'den alınan **sınırlı veri ile tüm stream veriyi içerecek** şekilde **cevap oluşturulabilir.**
  - Bazen stream'den **bir eleman geldiğinde** bile **algoritmanın karar vermesi gerekebilir.**

11

On-line ve off-line algoritmalar

Örnek

- **A** firması "**chesterfield**" kelimesi için **10 krş** teklif vermiş olsun.
- **B** firması "**chesterfield**" ve "**sofa**" kelimeleri için **20 krş** teklif vermiş olsun. Her iki firmada **aylık 100 TL bütçeye** sahip olsun.
- Her sorgu için en fazla bir reklam gösterilebilsin.
- "chesterfield" kelimesi için bir sorgu geldiğinde A veya B firmalarından birisinin seçilip gösterilmesi gerekir.
- **B daha yüksek teklif verdiği için B'nin reklamı gösterilir.**
- Ancak, çok sayıda "sofa" sorgusu olduğunu, az sayıda "chesterfield" sorgusu olduğunu varsayarsak, **A hiçbir zaman 100 TL bütçesini harcayamaz**, **B bütçesinin tamamını harcar.**
- Tüm "chesterfield" sorguları A için ve "sofa" sorguları B için kullanılırsa **kazanç maksimum yapılmış olur.**

12

Konular

- On-line Reklam
- On-line Algoritmalar
  - On-line ve off-line algoritmalar
  - Greedy algoritmalar
- Eşleştirme Problemi
  - Mükemmel eşleştirme
  - Greedy algoritması ile maximal eşleştirme
- Adwords Problemi
  - Arama reklamcılığı
  - Adwords probleminin tanımı
  - Adwords problemi için greedy yaklaşımı
  - Balance algoritması

13

Greedy algoritmalar

- **On-line algoritmaların çoğu greedy yaklaşımını kullanır.**
- Greedy algoritmalar her bir eleman girişi ve önceki girişler için bir fonksiyonu maksimize edecek şekilde karar oluştururlar.
- A ve B firmaları için ¨chesterfield¨ ve ¨sofa¨ kelimeleri örneğini ele alalım.
- **İlk gelen 500 sorgu ¨chesterfield¨ sonraki 500 sorgu ise ¨sofa¨ olsun.**
- **İlk 500 sorgu ile B tüm bütçesini harcar**, ardından gelen 500 sorgu için A'nın reklamı gösterilemez.
- Arama motorunun **toplam kazancı 100 TL olur.**
- Eğer ilk gelen 500 sorgu A için, ikinci gelen 500 sorgu B için kullanılabilirse **toplam kazanç 150 TL olur.**
- **Off-line algoritmalar** tüm veriyi kullanarak **optimizasyon yapabilir.**

14

Greedy algoritmalar

Competitive ratio

- **On-line algoritmalarda elde edilen sonuç** hiçbir zaman **off-line algoritmalardan elde edilen sonuç kadar iyi olamaz.**
- Bir on-line algoritmanın sonucu, off-line algoritmanın sonucunun en çok c katı kadar iyi olabilir ($c < 1$).
- c sabit sayısına competitive ratio denir.
- A ve B firmaları için ¨chesterfield¨ ve ¨sofa¨ örneğinde en iyi sonuç 100 TL/150 TL = **2/3** oranında olur.
- Competitive ratio değeri 2/3 olur.

15

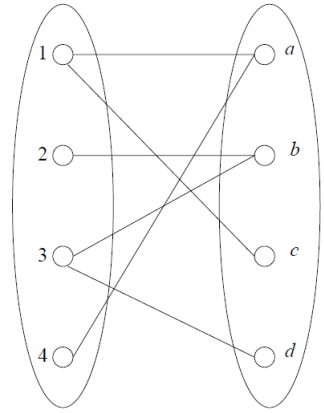
Konular

- On-line Reklam
- On-line Algoritmalar
  - On-line ve off-line algoritmalar
  - Greedy algoritmalar
- Eşleştirme Problemi
  - Mükemmel eşleştirme
  - Greedy algoritması ile maximal eşleştirme
- Adwords Problemi
  - Arama reklamcılığı
  - Adwords probleminin tanımı
  - Adwords problemi için greedy yaklaşımı
  - Balance algoritması

16

Eşleştirme Problemi

- **Reklamların kullanıcı sorgularıyla eşleştirilmesi**, maximal matching problemi olarak adlandırılır.
- Maximal matching problemi **bipartite (iki parçalı) graf** içerir.
- İki tür düğümden birisi reklamları diğeri ise sorguları gösterir.



17

Konular

- On-line Reklam
- On-line Algoritmalar
  - On-line ve off-line algoritmalar
  - Greedy algoritmalar
- Eşleştirme Problemi
  - Mükemmel eşleştirme
  - Greedy algoritması ile maximal eşleştirme
- Adwords Problemi
  - Arama reklamcılığı
  - Adwords probleminin tanımı
  - Adwords problemi için greedy yaklaşımı
  - Balance algoritması

18

Mükemmel eşleştirme

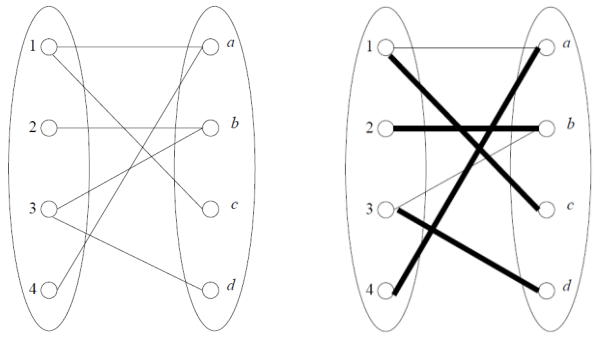
- Bir bipartite grafta, **eğer hiçbir node iki veya daha fazla kenarın sonu değilse, buna eşleştirme (matching) denir.**
- Eğer **tüm node'lar eşleştirmede yer alıyorsa,** buna **mükemmel eşleştirme (perfect matching) denir.**
- Perfect matching için sol ve sağdaki düğüm sayılarının eşit olması gerekir.
- Bir grafta elde edilen **en büyük eşleştirme** ise maximal matching olarak adlandırılır.

19

Mükemmel eşleştirme

Örnek

- $\{(1, a), (2, b), (3, d)\}$ kenar kümesi bir eşleştirmedir.
- $\{(1, c), (2, b), (3, d), (4, a)\}$ kenar kümesi **mükemmel bir eşleştirmedir.**
- Mükemmel eşleştirmede her node kesinlikle bir kez yer alır.



20

Konular

- On-line Reklam
- On-line Algoritmalar
  - On-line ve off-line algoritmalar
  - Greedy algoritmalar
- Eşleştirme Problemi
  - Mükemmel eşleştirme
  - Greedy algoritması ile maximal eşleştirme
- Adwords Problemi
  - Arama reklamcılığı
  - Adwords probleminin tanımı
  - Adwords problemi için greedy yaklaşımı
  - Balance algoritması

21

Greedy algoritması ile maximal eşleştirme

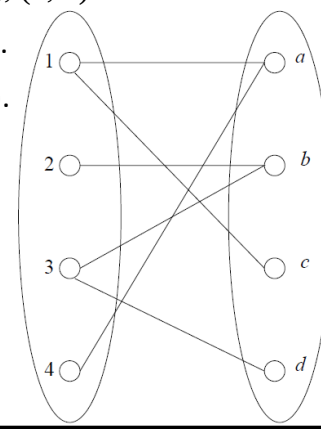
- Off-line algoritmalar ile maximal eşleştirme n tane node'a sahip graf için $O(n^2)$ ile elde edilebilmektedir.
- On-line greedy algoritmalar ile maximal matching yapılabilmektedir.
- Greedy yaklaşımında ***x* ve *y* node'ları hiçbir kenarın ucunda değilse, *x* ve *y* arasında kenar çizilir.**
- x ve y node'larından birisi **bir kenarın ucunda** ise ***x* ve *y* node'ları atlanır.**

22

Greedy algoritması ile maximal eşleştirme

Örnek

- Tüm node'lar **lexicographically sıralansın.**
- Bu sıralamada soldaki node'un bağlı olduğu sağdaki node'lar da kendi içinde sıralanır $(1, a)$, $(1, c)$, $(2, b)$, $(3, b)$, $(3, d)$, $(4, a)$.
- $(1, a)$ seçilir. $(1, c)$ seçilemez (1 eşleştirilmiştir).
- $(2, b)$ seçilir. $(3, b)$ seçilemez (b eşleştirilmiştir).
- $(3, d)$ seçilir. $(4, a)$ seçilemez.
- $(1, a)$, $(2, b)$, $(3, d)$ kenarları seçilir.
- **Elde edilen eşleştirme maximal değildir.**



23

Konular

- On-line Reklam
- On-line Algoritmalar
  - On-line ve off-line algoritmalar
  - Greedy algoritmalar
- Eşleştirme Problemi
  - Mükemmel eşleştirme
  - Greedy algoritması ile maximal eşleştirme
- Adwords Problemi
  - Arama reklamcılığı
  - Adwords probleminin tanımı
  - Adwords problemi için greedy yaklaşımı
  - Balance algoritması

24

Adwords Problemi

- Adwords problemi ile **ilk defa Google Adwords sisteminde karşılaşmıştır.**
- 2000'li yıllarda **Overture isimli firma** yeni bir arama **önermiştir.**
- **Arama kelimelerine reklamcılar** tarafından **teklif** verilmektedir.
- İlgili kelime arama sorgusunda varsa **yüksek tekliften başlanarak reklamlar gösterilmektedir.**
- Reklamcının listede sunulan **linki tıklanırsa reklam veren ücret ödemektedir.**

25

Konular

- On-line Reklam
- On-line Algoritmalar
  - On-line ve off-line algoritmalar
  - Greedy algoritmalar
- Eşleştirme Problemi
  - Mükemmel eşleştirme
  - Greedy algoritması ile maximal eşleştirme
- Adwords Problemi
  - Arama reklamcılığı
  - Adwords probleminin tanımı
  - Adwords problemi için greedy yaklaşımı
  - Balance algoritması

26

Arama reklamcılığı

- Google tarafından Overture firmasının önerdiği arama reklamcılığı değiştirilerek kullanılmıştır.
- Önerilmesinden birkaç yıl sonra **Google Adwords sistemine aşağıdaki özelliklerle birlikte adapte etmiştir:**
  - Google her sorgu için **sınırlı sayıda reklam göstermektedir.**
  - Adwords kullanıcıları **aylık tüm tıklanmalar için bütçeye sahiptir**.
  - Google sadece toplam teklif değerine göre sıralama yapmaz, **geçmişteki tıklanma oranını da (click-through rate) gözönüne alır.**

27

Konular

- On-line Reklam
- On-line Algoritmalar
  - On-line ve off-line algoritmalar
  - Greedy algoritmalar
- Eşleştirme Problemi
  - Mükemmel eşleştirme
  - Greedy algoritması ile maximal eşleştirme
- Adwords Problemi
  - Arama reklamcılığı
  - Adwords probleminin tanımı
  - Adwords problemi için greedy yaklaşımı
  - Balance algoritması

28

Adwords probleminin tanımı

- Hangi reklamın gösterileceğinin on-line belirlenmesi gereklidir.
- **On-line algoritmaya girişler:**
  - Arama sorguları için reklamcıların **tekliflerinin kümesi**
  - Her reklamcı-sorgu çifti için **click-through oranı**
  - Her reklamcı için **aylık bütçe**
  - Her **arama sorgusu için** gösterilecek **reklam sayısı limiti**
- Her arama sorgusu için reklamcı kümesi ile **oluşturulan cevap:**
  - Her sorgu için **belirlenen reklam sayısından fazla reklam sunulamaz.**
  - Her reklamcının **arama sorgusuna teklifi vardır.**
  - Her reklamcı listede tıklanması halinde **yeterli bütçeye sahiptir.**
- Bir **reklamın değeri, teklif miktarı** ile **click-through oranının çarpımıdır.**
- On-line **algoritmanın kazancı aylık toplam getirisiyle ölçülür.**

29

Konular

- On-line Reklam
- On-line Algoritmalar
  - On-line ve off-line algoritmalar
  - Greedy algoritmalar
- Eşleştirme Problemi
  - Mükemmel eşleştirme
  - Greedy algoritması ile maximal eşleştirme
- Adwords Problemi
  - Arama reklamcılığı
  - Adwords probleminin tanımı
  - Adwords problemi için greedy yaklaşımı
  - Balance algoritması

30

Adwords problemi için greedy yaklaşımı

- **Adwords problemi için** sadece **on-line algoritmalar uygundur.**
- Aşağıdaki **varsayımlar** alınmıştır:
  - **Her sorgu için bir reklam gösterilir.**
  - **Tüm reklamcılar aynı bütçeye sahiptir.**
  - **Tüm click-through oranları eşittir.**
  - **Tüm teklifler 0 veya 1'dir.**
  - **Tüm reklamların değeri (teklif * click-through rate) eşittir.**
- **Greedy algoritması her arama sorgusu için teklif veren reklamcılardan birisini seçer.**

31

Adwords problemi için greedy yaklaşımı

Örnek

- A ve B olarak **iki reklamcı**, x ve y olarak **iki tür sorgu** olsun.
- A sadece x **için teklif** versin, B hem x hem de y **için teklif** versin.
- **Her iki reklamcının** da **bütçesi** 2 **olsun.**
- Gelen sorgu $xxyy$ şeklinde olsun.
- Greedy algoritması **ilk iki** x**'i** B**'ye atayabilir. Kalan** y**'ler için ödeme alamaz.**
- Algoritmanın **toplam getirisi** 2 **olur.**
- Optimum bir **off-line algoritma** x**'leri** A**'ya** ve y**'leri** B**'ye atayabilir.**
- Bu durumda **off-line algoritmanın getirisi** 4 **olur.**
- Greedy algoritması için **competitive ratio** $1/2$ **olur.**

32

Konular

- On-line Reklam
- On-line Algoritmalar
  - On-line ve off-line algoritmalar
  - Greedy algoritmalar
- Eşleştirme Problemi
  - Mükemmel eşleştirme
  - Greedy algoritması ile maximal eşleştirme
- Adwords Problemi
  - Arama reklamcılığı
  - Adwords probleminin tanımı
  - Adwords problemi için greedy yaklaşımı
  - Balance algoritması

33

Balance algoritması

- **Competitive ratio değerini $3/4$'e yükseltmek** için **greedy algoritmasında iyileştirme yapılır.**
- Balance algoritması, gelen bir **sorguyu teklif veren bir reklamcıya atarken**, **kalan bütçesi yüksek** olana öncelik verir.
- Böylelikle **teklif veren reklamcıların bütçeleri dengeli bir şekilde azalmış olur.**

34

Balance algoritması

Örnek

- A ve B olarak iki **reklamcı** ve x ve y olarak iki tür **sorgu** olsun.
- A sadece x **için teklif** versin, B hem x hem de y için **teklif** versin.
- **Her iki reklamcının** da **bütçesi** 2 **olsun.**
- Gelen sorgu $xxyy$ şeklinde olsun.
- Balance algoritması **ilk** x**'i** A**'ya** veya B**'ye atayabilir.**
- Çünkü, her **ikisi** x**'e teklif vermiştir** ve **kalan bütçeleri aynıdır.**
- **İkinci** x**'** i A veya B**'den diğerine atar.**
- **Birinci** y**'yi** B**'ye atar. İkinci** y **atanamaz** çünkü B**'nin bütçesi kalmamıştır.**
- Balance algoritması için toplam getiri 3 olur.
- Competitive ratio $3/4$ olur.

35