

Web Madenciliđi (Web Mining)

Hazırlayan: M. Ali Akcayol
Gazi Üniversitesi
Bilgisayar Mühendisliđi Bölümü

Bu dersin sunumları, "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, Bing Liu, Springer, 2011." kitabı kullanılarak hazırlanmıştır.

Konular

- Giriş
- Metin Ön İşlemleri
- Web Sayfası Ön İşlemleri
- Web Arama
- Meta-Arama ve Sonuçların Birleştirilmesi

Giriş

- Günümüzdeki arama algoritmaları kullanıcı sorgusu ile ilgililik düzeyini hesaplamak için farklı yöntemler kullanır.
- **Bir arama motorunun en önemli bileşeni, kullanıcı sorgusuyla ilgili olduğu belirlenen sayfaların rank değerlerinin belirlenmesidir.**
- Web aramanın temeli, bilgi erişim (information retrieval - IR) yöntemlerine dayanmaktadır.
- Web arama motorları crawler yazılımları ile orijinal sunuculardaki Web sayfalarını belirli aralıklarla güncellemesi gereklidir.
- Crawl yapılan sayfalar arama işlemlerinde etkinliği artırmak için indekslenmiş bir şekilde saklanır.

3

Giriş

- Web arama, en önemli IR uygulamalarının başında yer alır.
- **Web arama, IR yöntemlerinin yanı sıra çok sayıda kendisine özgü yöntemi de birlikte kullanır.**
- **Web aramada etkinlik ve yüksek performans en önemli gereksinimlerdir, çünkü Web üzerindeki doküman sayısı çok fazladır.**
- **Web kullanıcıları arama sonuçlarına etkin ve çok hızlı yanıt almak istemektedir.**

4

Giriş

- **Web sayfaları çok sayıda farklı yapıya sahip bloklardan oluşur.**
- Web sayfaları hyperlinklere ve anchor metinlere sahiptir, klasik dokümanlarda linkler yer almaz.
- **Hyperlinkler**, Web arama algoritmalarında ve elde edilen sonuçların sıralanması için kullanılan **ranking algoritmalarında kullanılan en önemli bileşenlerdir.**
- Hyperlinkler ile ilişkilendirilen anchor metinler, link verilen sayfaya ait bilgiyi içerdiğinden çok önemlidir.
- **Web sayfaları yarı yapılandırılmıştır.** Bir Web sayfasında title, metadata, body gibi alanlar bulunmaktadır.
- **Bazı alanlardaki bilgiler diğerlerinden daha fazla öneme sahiptir.**

5

Giriş

- Bu bloklardan **bazıları çok önemlidir** (menü alanları, başlık bilgileri), **bazıları ise önemli değildir** (reklamlar, copyright bilgileri, gizlilik bildirimleri) ve Web sayfasından çıkartılmalıdır.
- Bu blokların çok iyi analiz edilmesi ve faydalı olanların seçilmesi, faydalı olmayanların silinmesi gereklidir.
- **Web dokümanlarında spamming çok önemlidir.** Ancak **klasik IR dokümanlarında spamming yapılmaz.**
- Spamming ile Web sayfasının arama sonuçlarında elde edilen listede daha üst sıralarda yer alınması sağlanabilir.
- Bunun sonucunda kullanıcı sorgusuyla daha ilgili Web sayfaları alt sıralarda olduğundan kullanıcı tarafından görülemez.

6

Konular

- Giriş
- **Metin Ön İşlemleri**
- Web Sayfası Ön İşlemleri
- Web Arama
- Meta-Arama ve Sonuçların Birleştirilmesi

7

Metin Ön İşlemleri

- Bir doküman erişim için kullanılmadan önce bazı ön işlemler yapılır.
- Klasik dokümanlar için (HTML hariç) yapılan ön işlemler, **stopword silme, stemming, rakamların silinmesi, tire işaretlerinin silinmesi, noktalama işaretlerinin silinmesi ve büyük/küçük harf ayırımının giderilmesi** işleridir.
- **Stopword silme** ile dokümanın anlamına katkı sağlamayan dilin yapısal özelliklerinden dolayı eklenen kelimeler silinir (a, about, an, by, for, who, ben, ki, de/da, için, ...).
- **Stemming** ile bir kelimenin çoğul yapılması, zaman ekleri gibi öneklerinin ve soneklerinin silinmesi yapılır ("walks", "walking" ve "walker" kelimeleri "walk" yapılır).

8

Metin Ön İşlemleri

- **Rakamların silinmesi** ile klasik IR sistemlerinde doküman içerisindeki rakamlar silinir (Web arama motorları rakamları da indeksler).
- **Tire işaretlerinin silinmesi** ile ifadeler arasındaki tutarsızlıklar giderilir (“state-of-the-art” ifadesi “state of the art” yapılır).
- **Noktalama işaretlerinin silinmesi** ile tire işaretlerinde olduğu gibi tutarsızlıklar giderilir.
- **Büyük/küçük harf ayırımının giderilmesi** ile doküman ya tümü küçük harf ya da tümü büyük harf yapılır.

9

Konular

- Giriş
- Metin Ön İşlemleri
- **Web Sayfası Ön İşlemleri**
- Web Arama
- Meta-Arama ve Sonuçların Birleştirilmesi

10

Web Sayfası Ön İşlemleri

- **HTML dokümanında farklı text alanları vardır.** Her text alanı farklı öneme sahiptir. Bu farklı text alanlarının ayırt edilmesi gereklidir.
- **Anchor text'ler** arama motorları tarafından yüksek öneme sahiptir. Anchor text'lerin doğru tanımlanması gereklidir.
- **HTML etiketlerinin** indeks yapılmadan önce silinmesi gereklidir.
- **Temel içerik bloklarının** belirlenmesi gereklidir.

11

Konular

- Giriş
- Metin Ön İşlemleri
- Web Sayfası Ön İşlemleri
- **Web Arama**
- Meta-Arama ve Sonuçların Birleştirilmesi

12

Web Arama

- Günümüzdeki arama algoritmaları vector space model ve term matching yöntemlerini kullanır.
- Bir arama motoru Web sayfalarını crawl yaparak başlar.
- Crawl edilen sayfalar **parse edilir, indekslenir ve saklanır**.
- Sorgu sırasında, bilgiye erişim için indeks etkin bir şekilde kullanılır.

Parsing

- **Parser** bir HTML dokümanını giriş olarak alır, indekslenmek üzere **token veya terim kümesi üretir**.
- Bir parser YACC (Yet Another Compiler Compiler) veya Flex (Fast Lexical Analyzer) gibi açık kaynak kodlu lexical analyzer kullanılarak yapılandırılabilir.

13

Web Arama

Indexing

- **Arama motorları arama etkinliğini artırmak için birden fazla indeks oluşturabilir.**
- Örneğin title ve anchor text sayfa hakkında daha doğru bilgiyi sağlayabilir. Bunlar için ayrıca indeks oluşturulabilir.
- Sayfa metinlerinin tümünü içeren full indeks te ayrıca oluşturulur.
- Arama algoritması öncelikle title ve anchor text'leri içeren indeks'te arama yapar ardından full indeks'te arama yapar.
- **Yeterli sayıda sayfa ilk aramada elde edilirse full indeks üzerinde arama yapılmaz.**

14

Searching and Ranking

- Girilen bir kullanıcı sorgusuna göre arama aşağıdaki adımlardan oluşur:
 - Sorgu terimleri üzerinde **önişlemler**
 - İndeks üzerinde **sorgudaki tüm terimleri bulunduran sayfaların bulunması**
 - Kullanıcıya sunulmak üzere sayfaların **rank değerlerinin belirlenmesi**
- **Bir arama motorunun en önemli kısmı ranking algoritmasıdır.**
- Ticari arama motorlarında ranking algoritmalarıyla ilgili çok az bilgi bulunmaktadır.

15

Searching and Ranking

- Klasik IR sistemlerinde dokümanların rank değerlerinin belirlenmesi için **cosine similarity** gibi yöntemler kullanılır.
- **Cosine similarity gibi yöntemler içerik tabanlıdır ve Web aramada yeterli değildir.**
- Kullanıcı sorgusu için çok sayıda ilgili doküman elde edilebilir.
- **Örneğin, Google arama motorunda “data mining” şeklinde sorguya 41.900.000 adet sonuç bulunmaktadır.**
- Bu sayfaların kullanıcıya hangi sırada sunulduğu ve üstte hangilerinin yer alacağı önemlidir.
- **Web üzerindeki sayfaların kalitesi ve güvenilir olması ranking için çok önemlidir.**

16

Web Arama

Searching and Ranking

- Bir Web sayfasının kaliteli olup olmadığı içeriğe göre belirlenemez.
- **Hyperlink'ler sayfaların kalitesinin belirlenmesinde en önemli bileşenlerdir.**
- X sayfasında Y sayfasına link verildiğinde, **X sayfasının yazarı Y sayfasının içeriğine güvendiğini dolaylı olarak gösterir.**
- **Bir sayfa ne kadar çok link alırsa (in-links) o kadar kaliteli olduğu varsayılır.**
- PageRank algoritmasının temelinde in-links sayısı yer almaktadır.
- **Web sayfaları kendi içeriklerine ve aldıkları linklere göre değerlendirilebilir.**

17

Web Arama

Searching and Ranking

- **İçerik tabanlı (content based) değerlendirme** aşağıdaki bilgilere dayanır:
 - Occurence type
 - Count
 - Position

Occurence type

- Sorgu terimlerinin bulunduğu yerler:
 - **Title:** Sayfa başlığında yer alabilir
 - **Anchor text:** Linke ait metinde yer alabilir
 - **URL:** <http://www.domain.edu/Webmining>
 - **Body:** Sayfa içerisinde herhangi bir yer olabilir. (bold, font size gibi değerlerine göre değerlendirilir.)

18

Web Arama

Searching and Ranking

Count

- Her terimin sayfa içerisinde bulunduğu **tekrar sayısıdır**.

Position

- Her terimin **sayfada bulunduğu pozisyonu gösterir**.
- Birden fazla terime sahip proximity sorgularda terimlerin birbirine yakınlığını belirlemek için kullanılır.
- **İçerik tabanlı skor hesabında terimlerin bulunduğu yerlere göre ağırlıklandırma yapılır (title, URL, body, ...).**

19

Konular

- Giriş
- Metin Ön İşlemleri
- Web Sayfası Ön İşlemleri
- Web Arama
- **Meta-Arama ve Sonuçların Birleştirilmesi**

20

Meta-Arama ve Sonuçların Birleştirilmesi

- **Kullanıcıların %75'i** sorgu sonucunda gelen **ilk sayfadaki sonuçlara bakmaktadır**. Sadece %25'i ikinci sayfaya geçmektedir.
- **Kullanıcıların %70'i** doğrudan **organik sonuçları seçmektedir** ve reklam içerikli sonuçlara bakmamaktadır.
- **Şirketlerin %81'i** kendileri için **yazılan blog bilgilerini dikkate almaktadırlar**.
- Google arama motorunun sunduğu organik listedeki sonuçlardan **1.sıradakini seçen kullanıcı oranı %18, 2.sıradakini seçen %10 ve 3.sıradakini seçen %7** oranındadır.
- Bing arama motoru organik sonuçlarından **1.sıradakini seçen kullanıcı oranı %9.7, ikinci sıradakini seçen %5.5 ve 3.sıradakini seçen %2.7** oranındadır.

21

Meta-Arama ve Sonuçların Birleştirilmesi

- Google, Yahoo, Live ve Ask arama motorlarında yapılan araştırmaya göre,
 - **İki farklı arama motorunun ilk sayfa sonuçlarındaki çakışma oranı %8.9,**
 - **Üç farklı arama motorunun ilk sayfa sonuçları arasında çakışma oranı %2.2**
 - **Dört farklı arama motorunun ilk sayfa sonuçları arasındaki çakışma oranı ise %0.6** çıkmıştır.

	Percentage % of G-Y-L-A Total Results
Shared by all 4 engines	0.6%
Shared by any 3 engines	2.2%
Shared by any 2 engines	8.9%
Unique to 1 engine	88.3%

22

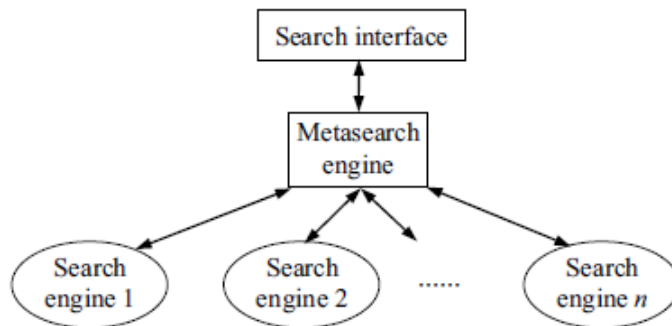
Meta-Arama ve Sonuçların Birleştirilmesi

- Kullanıcı sadece **Google** arama motorunu kullanırsa sorgusuyla ilgili Web'teki **%72.7 en iyi sonucu ilk sayfada görememektedir.**
- Kullanıcı sadece **Yahoo** arama motorunu kullanırsa **%69.2** ve sadece **Live** kullanırsa **%69.9** en iyi sonucu **ilk sayfada görememektedir.**
- Meta arama motorları, normal arama motorlarına göre Web'teki bilgilerin daha büyük bir kısmında arama yapar.
- Kullanıcıların farklı arama motorlarına sorgu girerek gelen sonuçları incelemeleri yerine tek bir sorgu ile birleştirilmiş tek bir liste halindeki sonuçları almaları sağlanır.
- Meta arama motorları kullanıcının girdiği sorguya hangi arama motorunun daha uygun olduğuna çok hızlı ve doğru bir şekilde karar verebilirler.

23

Meta-Arama ve Sonuçların Birleştirilmesi

- Meta-arama motoru, kullanıcı sorugusunu **birden fazla arama motoruna gönderir ve gelen sonuçları birleştirerek kullanır.**
- **Meta-arama motoru kendisine ait veritabanına sahip değildir.**



24

Meta-Arama ve Sonuçların Birleştirilmesi

- Arama arayüzünden alınan kullanıcı sorgusu seçilen arama motorlarına gönderilir.
- Arama motorlarından gelen sonuçlar birleştirilerek kullanıcıya sunulur.
- **Meta arama motorları Web'teki arama bölgesini artırır.**
- Web çok büyük bir bilgi kaynağıdır ve **arama motorları çok küçük bir kısmına erişebilir.**
- Eğer bir arama motoru kullanılırsa erişemediği bölgedeki ilgili dokümanlar elde edilemez.
- **Her arama motoru elde ettiği ilgili sayfalar için rank değeri hesaplar.**

25

Meta-Arama ve Sonuçların Birleştirilmesi

- **Meta aramada en önemli işlem, arama motorlarından alınan sonuçların birleştirilmesidir.**
- Farklı arama motorlarından gelen sonuçlardan aynı olanların belirlenmesi gereklidir.
- Elde edilen sonuçlardaki aynı sayfaların teke indirilmesi gereklidir.
- Meta arama motorunun elde edeceği liste için farklı arama motorlarının rank değerlerinin birleştirilmesi gereklidir.
- Elde edilen **sonuçların benzerliğine göre veya rank pozisyonlarına göre birleştirme işlemi yapılır:**
 - Borda ranking
 - Condorcet ranking
 - Reciprocal ranking

26

Meta-Arama ve Sonuçların Birleştirilmesi

Borda ranking

- Aday sayfaların oylanmasına dayanan birleştirme yapar.
- Toplam n tane aday varsa, bir oylama sonucunda **birinci sırada yer alan n puan, ikinci sırada yer alan $n-1$ puan, ...** şeklinde devam eder.
- Eğer **oylanmayan** adaylar varsa **kalan puan oylanmayan adaylar arasında eşit dağıtılır.**
- En yüksek puan alan aday kazanan adaydır.
- **Sıralama puanı yüksek olandan düşük olana doğru yapılır.**

27

Meta-Arama ve Sonuçların Birleştirilmesi

Condorcet ranking

- Her adayın **ikili olarak diğerleriyle sıralamadaki yerinin karşılaştırılmasına dayanır.**
- Bir oylamada **bir aday diğerine göre daha yüksek rank değerine sahipse kazanmış olur (win).**
- Bir oylamada bir aday diğerine göre daha düşük rank değerine sahipse kazanmış olur **(lose).**
- Bir oylamada **iki adayda oylanmadıysa birbirlerine bağlanır (tie).**

28

Meta-Arama ve Sonuçların Birleştirilmesi

Reciprocal ranking

- Her oylamada **birinci sıradaki 1 puan, ikinci sıradaki 1/2 puan, üçüncü sıradaki 1/3 puan, ...** şeklinde puan alır.
- Eğer bir aday oylanmazsa boş geçilir.
- Sonuç rank değerleri elde edilen skorların toplamı alınarak elde edilir.
- **Skor değeri yüksek olan ilk sırada yer alır.**
- Üst sıradaki adaylara Borda ranking yöntemine göre daha yüksek puan verir.

29

Meta-Arama ve Sonuçların Birleştirilmesi

Örnek

- Bir meta arama 5 arama motorundan aşağıdaki listeleri alsın.

system 1: *a, b, c, d*

system 2: *b, a, d, c*

system 3: *c, b, a, d*

system 4: *c, b, d*

system 5: *c, b*

Borda Ranking

$$\text{Score}(a) = 4 + 3 + 2 + 1 + 1.5 = 11.5$$

$$\text{Score}(b) = 3 + 4 + 3 + 3 + 3 = 16$$

$$\text{Score}(c) = 2 + 1 + 4 + 4 + 4 = 15$$

$$\text{Score}(d) = 1 + 2 + 1 + 2 + 1.5 = 7.5$$

- Sonuç ranking: **b, c, a, d**

30

Meta-Arama ve Sonuçların Birleştirilmesi

Örnek

Condorcet Ranking

	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>
<i>a</i>	-	1:4:0	2:3:0	3:1:1
<i>b</i>	4:1:0	-	2:3:0	5:0:0
<i>c</i>	3:2:0	3:2:0	-	4:1:0
<i>d</i>	1:3:1	0:5:0	1:4:0	-

	<i>win</i>	<i>lose</i>	<i>tie</i>
<i>a</i>	1	2	0
<i>b</i>	2	1	0
<i>c</i>	3	0	0
<i>d</i>	0	3	0

system 1: *a, b, c, d*
 system 2: *b, a, d, c*
 system 3: *c, b, a, d*
 system 4: *c, b, d*
 system 5: *c, b*

■ Sonuç ranking: **c, b, a, d**

31

Meta-Arama ve Sonuçların Birleştirilmesi

Örnek

Reciprocal Ranking

$$\text{Score}(a) = 1 + 1/2 + 1/3 = 1.83$$

$$\text{Score}(b) = 1/2 + 1 + 1/2 + 1/2 + 1/2 = 3$$

$$\text{Score}(c) = 1/3 + 1/4 + 1 + 1 + 1 = 3.55$$

$$\text{Score}(d) = 1/4 + 1/3 + 1/4 + 1/3 = 1.17$$

■ Sonuç ranking: **c, b, a, d**

32