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

- PageRank
  - Term spam
  - PageRank tanımı
  - Web'in yapısı
  - Dead end sayfalar
  - Spider traps
  - Arama motorunda PageRank kullanımı
- Link Spam
  - Link spam yapısı
  - Spam farm analizi
  - TrustRank
  - Spam mass
- Hub ve Otorite Sayfalar
  - HITS algoritması
  - Hub ve otorite tanımı

PageRank

- **Google, spammer'ları elimine edebilen ilk arama motorudur.**
- Spammer, arama motoru sonuçlarını kullanışsız hale getirir.
- Google, **Web sayfalarının önemini değerlendiren** PageRank algoritmasını geliştirmiştir.
- **Spammer'lar** ise PageRank algoritmasını manipüle etmek için **link spam** yöntemini kullanmaya başlamıştır.
- **Google**, spammer'ların saldırılarını engellemek için TrustRank **yöntemini geliştirmiştir.**

3

Konular

- PageRank
  - Term spam
  - PageRank tanımı
  - Web'in yapısı
  - Dead end sayfalar
  - Spider traps
  - Arama motorunda PageRank kullanımı
- Link Spam
  - Link spam yapısı
  - Spam farm analizi
  - TrustRank
  - Spam mass
- Hub ve Otorite Sayfalar
  - HITS algoritması
  - Hub ve otorite tanımı

4

Term spam

- Google'dan **önceki arama motorları** crawl ettikleri Web sayfasında yer alan **terimleri** (boşluk hariç kelime veya string) inverted index **kullanarak listelemekteydi.**
- Bir search query alındığında, **inverted index'ten ilgili sayfalar alınarak** hesaplanan rank **değerine göre sunulmaktaydı.**
- **Bir terimin sayfa başlığında olması** veya **sayfada sık geçmesi** sorguya **ilgililik düzeyini artırmaktaydı.**
- Etik olmayan yöntemlerle **ilk arama motorlarını kolaylıkla yanıltmak mümkündü.**
- Örneğin, sık arama yapılan kelimeler background rengiyle çok sayıda yazılarak rank değeri yükseltilebilmekteydi.
- Bir sayfanın herhangi bir konuyla ilgili olduğuna yönelik **arama motorlarının yanıltılması** için kullanılan tekniklere term spam denir.

5

Term spam

- **PageRank**, term spam ile mücadele için **iki yöntem geliştirmiştir:**
- **PageRank Web kullanıcılarını simüle eder.**
  - Rastgele bir sayfadan başlayıp; **outlink'leri rastgele seçen kullanıcıların (random surfer) hangi sayfalara gideceğini iterative bir şekilde belirler.**
  - **Çok gezilen sayfaları** diğerlerine göre **daha önemli kabul eder.**
  - **Google** gelen bir sorgu için cevap oluştururken **önemli sayfaları tercih eder.**
- **Bir sayfa içeriğine sadece o sayfada yer alan terimlere göre karar vermez.**
  - **O sayfaya link veren sayfalarda** linkin içerisinde veya yakınında **bulunan terimlere göre önemine karar verilir.**
  - **Spammer** kendi sayfasında term spam yapabilir, ancak kendi sayfasına link veren **diğer sayfalarda kolaylıkla term spam yapamaz.**

6

Term spam

- **Google,** bir Web sayfasının kendisi için ifade ettiğini değil, **diğer sayfaların onun için ifade ettiğini dikkate almaktadır.**
- **Spammer kendi sayfasına çok sayıda link veren sayfa oluşturabilir.** Ancak, **PageRank algoritmasında bu sayfaların da önemi düşük olacaktır.**
- **Her kullanıcı Web üzerinde** gezinirken sayfalardaki linkleri seçerek bir **oylama yapar.**
- **Web sayfasına faydalı olduğu düşünülen sayfaların linkleri konulur.**
- Faydalı olmayacağı düşünülen linkler genellikle yer almaz.
- **Kullanıcılar faydalı sayfaları** faydasız sayfalara göre **daha çok ziyaret ederler.**

7

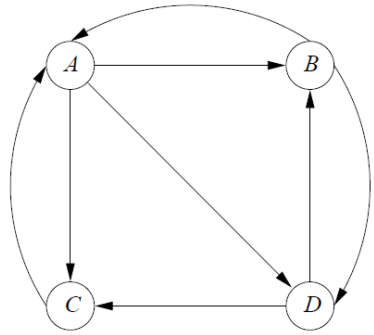
Konular

- PageRank
  - Term spam
  - PageRank tanımı
  - Web'in yapısı
  - Dead end sayfalar
  - Spider traps
  - Arama motorunda PageRank kullanımı
- Link Spam
  - Link spam yapısı
  - Spam farm analizi
  - TrustRank
  - Spam mass
- Hub ve Otorite Sayfalar
  - HITS algoritması
  - Hub ve otorite tanımı

8

PageRank tanımı

- **PageRank**, Web'teki **her sayfaya reel sayı atayan bir fonksiyondur.**
- Atanan **değeri yüksek olan sayfa** PageRank için **daha önemlidir.**
- **Web bir graf olarak düşünülebilir.** Sayfalar node, linkler kenardır.
- Bir random surfer (rastgele seçim yapan kullanıcı) A düğümünde ise; B, C ve D düğümlerini seçme olasılığı $1/3$'tür. A'da kalma olasılığı 0'dır.

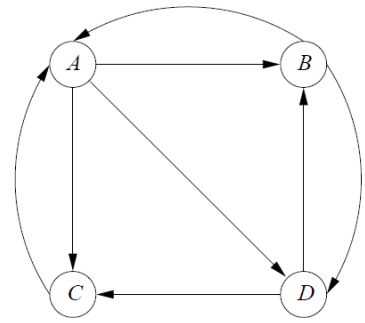


PageRank tanımı

- Random surfer'ın **bir sonraki geçişi için** transition matrix (M) **tanımlanabilir.**

sonraki → satırlar; mevcut → sütunlar

$$M = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix} \end{array} \qquad \mathbf{v}_0 = \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}$$



- $m_{ij} = 1/k$ ise, ***j*. sayfada *k* tane outlink vardır** ve **bir tanesi *i*.sayfaya verilmiştir.** $m_{ij} = 0$ ise link verilmemiştir.
- M matrisinin her sütunundaki değerlerin toplamı 1'e eşittir (stokastik matris).
- Herhangi **bir sütundaki olasılık dağılımı PageRank fonksiyonudur.**

PageRank tanımı

- Bir random surfer **başlangıçta Web'teki *n* sayfadan birisinde başlasın.**
- Başlangıç vektörü $\mathbf{v}_0$ **her eleman için** $1/n$ **değerine sahiptir.**
- $\boldsymbol{M}$, Web'teki **transition matrix** (geçiş matrisi) **olsun.**
- **Bir adım sonraki olasılık dağılımı** $\boldsymbol{M}\,\mathbf{v}_0$ **olur.**
- İki adım sonra $M(M\mathbf{v}_0)=M^2\mathbf{v}_0$ olur.
- Random surfer'ın sonraki adımda ***i* node'una geçme** olasılığı $\mathbf{x}_i$ aşağıdaki gibi hesaplanır:

$$\mathrm{x}_i = \sum_j m_{ij}\mathrm{v}_j$$

- v_j , random surfer'ın **önceki adımda *j* node'unda olma** olasılığıdır.
- m_{ij} , random surfer'ın ***j* node'unda iken *i* node'una geçme** olasılığıdır.

11

PageRank tanımı

- Aşağıdaki şartlar altında $\mathbf{v}$ dağılımının limit değeri $\mathbf{v} = \boldsymbol{M}\,\mathbf{v}$ eşitliğini sağlar:
  - **Graf strongly connected** yapıdadır (herhangi bir node'a herhangi bir node'dan ulaşılabilir.).
  - **Dead end yoktur** (outlink olmayan sayfa yoktur.).
- $\mathbf{v}$ vektörüne M matrisinin principle eigenvector'ü denir.
- $\mathbf{v}$ vektörü **random surfer'ın** uzun bir süre sonunda **hangi sayfada olacağını gösterir.**
- Başlangıç vektörü $\mathbf{v}_0$ kullanılarak $\mathbf{v}$ **vektörünün değeri belirli bir iterasyon sonrası için hesaplanır.**
- Çok küçük değişim oluncaya kadar iterasyon devam ettirilir.
- **Web için 50-75 arasında iterasyon yeterlidir.**

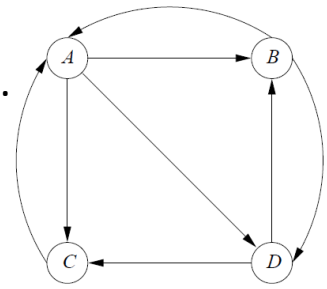
12

PageRank tanımı

Örnek

- M **transition matrisi** ve $\mathbf{v}_0$ **başlangıç vektörü.**

$$M = \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix} \qquad \mathbf{v}_0 = \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}$$



- Eigenvector değerlerinin iterasyonlarla değişimi aşağıdaki gibi olur.

$$\begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}, \begin{bmatrix} 9/24 \\ 5/24 \\ 5/24 \\ 5/24 \end{bmatrix}, \begin{bmatrix} 15/48 \\ 11/48 \\ 11/48 \\ 11/48 \end{bmatrix}, \begin{bmatrix} 11/32 \\ 7/32 \\ 7/32 \\ 7/32 \end{bmatrix}, \ldots, \begin{bmatrix} 3/9 \\ 2/9 \\ 2/9 \\ 2/9 \end{bmatrix}$$

- İlk satır değeri A **düğümüne aittir** ve **diğerlerinin** $3/2$ **katı çıkmıştır.**

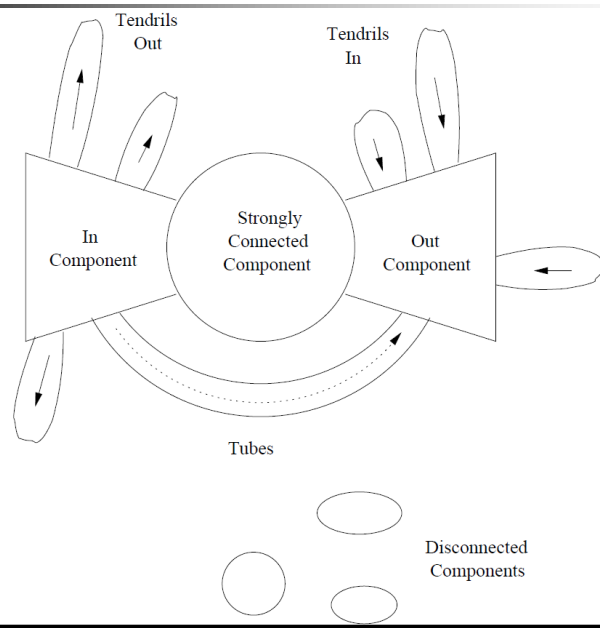
13

Konular

- PageRank
  - Term spam
  - PageRank tanımı
  - Web'in yapısı
  - Dead end sayfalar
  - Spider traps
  - Arama motorunda PageRank kullanımı
- Link Spam
  - Link spam yapısı
  - Spam farm analizi
  - TrustRank
  - Spam mass
- Hub ve Otorite Sayfalar
  - HITS algoritması
  - Hub ve otorite tanımı

14

Web'in yapısı



15

Web'in yapısı

- Web'te çok sayıda strongly connected component (SCC) vardır (dead end sayfa yoktur).
- SCC olmayan büyük bir kısım vardır.
- In-component: **Linkler takip edilerek SCC'ye ulaşan sayfaları içerir.** Ancak, **SCC'lerden bu elemanlara ulaşılamaz.**
- Out-component: **Linkler takip edilerek SCC'den ulaşılabilen sayfaları içerir.** Ancak, **bunlardan SCC'ye ulaşılamaz.**
- Tendrils out: **In-component'lerden ulaşılan sayfalardır.** Bunlardan in-component'lere ulaşılamaz.
- Tendrils in: **Out-component'lere ulaşılan sayfalardır.** Out-component'lerden bunlara ulaşılamaz.

16

Web'in yapısı

- Tubes: **In-component'lerden ulaşılan** ve **out-component'lere ulaşan sayfalardır.** SCC'den bunlara veya bunlardan SCC'ye ulaşılamaz.
- Isolated components: **Kendisine ulaşılamayan ve kendisinden diğerlerine ulaşılamayan elemanlardır.**
- Aşağıdaki iki sorundan kaçınmak gerekir:
  - **Dead end sayfalar:** Dead end sayfalara ulaşan surfer başka sayfaya geçemez. Bu sayfalara ulaşan sayfalar için PageRank değeri elde edilemez.
  - **Spider traps:** İçerisinde outlink bulunan ancak **başka sayfalara linke sahip olmayan bir grup sayfadır.**
- Bu iki problemin çözümü için taxation metodu kullanılabilir.
- Taxation metodunda, **random surfer'ın bir sayfadan ayrılma olasılığı sonludur** ve **yeni bir surfer herhangi bir sayfadan başlayabilir.**

17

Konular

- PageRank
  - Term spam
  - PageRank tanımı
  - Web'in yapısı
  - Dead end sayfalar
  - Spider traps
  - Arama motorunda PageRank kullanımı
- Link Spam
  - Link spam yapısı
  - Spam farm analizi
  - TrustRank
  - Spam mass
- Hub ve Otorite Sayfalar
  - HITS algoritması
  - Hub ve otorite tanımı

18

Dead end sayfalar

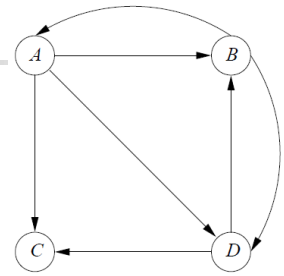
- İçerisinde hiç **outlink olmayan** sayfalar **dead end sayfalar** olarak adlandırılır.
- M **matrisi stokastik olmaz** (bazı sütunların toplamı 0'a eşittir.).
- Bir sütunun değerleri toplamı en çok 1 olan matris substokastik olarak adlandırılır.
- M^i **v artan üsler için hesaplandığında** elde edilen **vector'ün bazı değerleri veya tüm değerleri 0'a yaklaşır.**
- Web sayfalarının **göreceli önemine yönelik bilgi elde edilemez.**

19

Dead end sayfalar

Örnek

$$M = \begin{bmatrix} 0 & 1/2 & 0 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix} \quad v_0 = \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}$$



- C **(dead end)** sayfasına ulaşan random surfer sonraki adımda görünmez.
- M substokastik bir matristir (Burada C sütunu toplamı 0'dır.).
- Başlangıçtan itibaren hesaplanan **vektör değerleri 0'a doğru gider.**

$$\begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}, \begin{bmatrix} 3/24 \\ 5/24 \\ 5/24 \\ 5/24 \end{bmatrix}, \begin{bmatrix} 5/48 \\ 7/48 \\ 7/48 \\ 7/48 \end{bmatrix}, \begin{bmatrix} 21/288 \\ 31/288 \\ 31/288 \\ 31/288 \end{bmatrix}, \ldots, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

- Random surfer'ın iterasyon arttıkça herhangi bir yerde olma olasılığı 0 olur.

20

Dead end sayfalar

- Dead end sayfa probleminin çözümünde iki yaklaşım kullanılabilir.
- **Graftan dead end sayfalar atılır.**
  - Bunun sonucunda çok sayıda **yeni dead end sayfa oluşabilir.**
  - Strongly connected component'lere ulaşıncaya kadar **recursive olarak hepsi atılır.**
  - Recursive silme işleminin sonucunda **out-component'lerden**, **tendril'lerden** ve **tubes sayfalardan** bir kısmı da silinebilir.
  - SCC, in-component ve isolated component'ler kalır.
  - **Grafta yer almayan sayfaların** PageRank değeri **öncüllerinin değerlerinin toplamı ile hesaplanır.**
  - Graftaki **öncül sayfaların değeri, silinen sayfaların adedine bölünür.**
- **Random surfer'ın izlediği süreç değiştirilir.**
  - **Random surfer'ın** her durumda **Web üzerinde hareket ettiği (bir sayfadan her durumda ayrıldığı) varsayılır** (taxation metodu).

21

Dead end sayfalar

Örnek

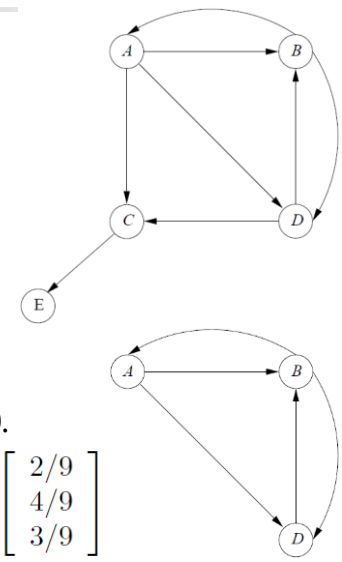
- E sayfası C sayfasının ardılıdır ve dead end'tir.
- ***E* sayfası silindiğinde *C* sayfası dead end olur.**
- Graf A, B ve D sayfalarından oluşur.
- Geçiş matrisi aşağıdaki gibidir.

$$M = \begin{bmatrix} 0 & 1/2 & 0 \\ 1/2 & 0 & 1 \\ 1/2 & 1/2 & 0 \end{bmatrix} \qquad v_0 = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

- PageRank değerleri A = 2/9, B = 4/9 ve D = 3/9.

$$\begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}, \begin{bmatrix} 1/6 \\ 3/6 \\ 2/6 \end{bmatrix}, \begin{bmatrix} 3/12 \\ 5/12 \\ 4/12 \end{bmatrix}, \begin{bmatrix} 5/24 \\ 11/24 \\ 8/24 \end{bmatrix}, \ldots, \begin{bmatrix} 2/9 \\ 4/9 \\ 3/9 \end{bmatrix}$$

- C = 1/3*A+1/2*D =1/3*2/9+1/2*3/9 = 13/54 ve E = C = 13/54 olur.



22

Konular

- PageRank
  - Term spam
  - PageRank tanımı
  - Web'in yapısı
  - Dead end sayfalar
  - Spider traps
  - Arama motorunda PageRank kullanımı
- Link Spam
  - Link spam yapısı
  - Spam farm analizi
  - TrustRank
  - Spam mass
- Hub ve Otorite Sayfalar
  - HITS algoritması
  - Hub ve otorite tanımı

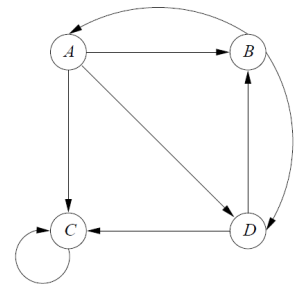
23

Spider traps

- **Bir spider trap bir grup node'tur.** Dead end değillerdir, ancak **kendilerinden başka sayfalara giden bağlantıları yoktur.**
- İterasyonun sonunda PageRank değerinin tamamını kendilerine alırlar.

Örnek

$$M = \begin{bmatrix} 0 & 1/2 & 0 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 1 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix} \quad v_0 = \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}$$



- C bir node'dan oluşan **spider trap'tir.**

$$\begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}, \begin{bmatrix} 3/24 \\ 5/24 \\ 11/24 \\ 5/24 \end{bmatrix}, \begin{bmatrix} 5/48 \\ 7/48 \\ 29/48 \\ 7/48 \end{bmatrix}, \begin{bmatrix} 21/288 \\ 31/288 \\ 205/288 \\ 31/288 \end{bmatrix}, \ldots, \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

- **Random surfer C sayfasından hiçbir zaman ayrılamaz.**

24

Spider traps

- Random surfer'ın **outlink kullanmadan rastgele bir sayfaya geçişi için küçük bir olasılık tanımlanır.**
- Yeni PageRank değeri aşağıdaki gibi hesaplanır.

$$\mathbf{v}' = \beta M\mathbf{v} + (1-\beta)\mathbf{e}/n$$

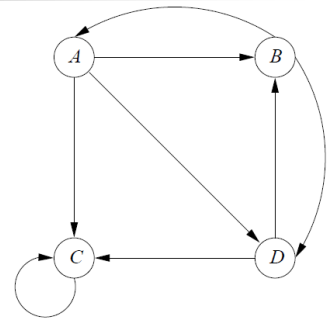
- β 0.8-0.9 aralığında bir sabit, **e** tüm değerleri 1 olan vektör, n ise Web'teki node sayısıdır.
- $\beta M\,\mathbf{v}$, random surfer'ın **out-link'ler kullanılarak β olasılığında geçişini belirler.**
- $(1-\beta)\mathbf{e}/n$ ise, **$(1-\beta)$ olasılığında surfer'ın rastgele bir sayfaya geçişini belirler.**

25

Spider traps

Örnek

$$M = \begin{bmatrix} 0 & 1/2 & 0 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 1 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix} \qquad \mathbf{v}_0 = \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}$$



$$\mathbf{v}' = \beta M\mathbf{v} + (1-\beta)\mathbf{e}/n$$

$\beta = 0.8$, $n = 4$

$$\mathbf{v}' = \begin{bmatrix} 0 & 2/5 & 0 & 0 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 0 & 4/5 & 2/5 \\ 4/15 & 2/5 & 0 & 0 \end{bmatrix} \mathbf{v} + \begin{bmatrix} 1/20 \\ 1/20 \\ 1/20 \\ 1/20 \end{bmatrix}$$

$$\begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}, \begin{bmatrix} 9/60 \\ 13/60 \\ 25/60 \\ 13/60 \end{bmatrix}, \begin{bmatrix} 41/300 \\ 53/300 \\ 153/300 \\ 53/300 \end{bmatrix}, \begin{bmatrix} 543/4500 \\ 707/4500 \\ 2543/4500 \\ 707/4500 \end{bmatrix}, \ldots, \begin{bmatrix} 15/148 \\ 19/148 \\ 95/148 \\ 19/148 \end{bmatrix}$$

26

Konular

- PageRank
  - Term spam
  - PageRank tanımı
  - Web'in yapısı
  - Dead end sayfalar
  - Spider traps
  - Arama motorunda PageRank kullanımı
- Link Spam
  - Link spam yapısı
  - Spam farm analizi
  - TrustRank
  - Spam mass
- Hub ve Otorite Sayfalar
  - HITS algoritması
  - Hub ve otorite tanımı

27

Arama motorunda PageRank kullanımı

- Her arama motoru kendine özgü ve **gizlenmiş bir formül ile sayfa sıralamasını yapmaktadır.**
- **Google** Web sayfalarının sıralaması için **250'den fazla farklı özelliği kullandığını belirtmektedir.**
- Bir sayfanın **sonuç listesinde yer alabilmesi için sorgudaki bir veya daha fazla kelimeyi içermesi zorunludur.**
- Genellikle **tüm kelimeleri içermeyen sayfaların ilk 10 sırada görülme şansı çok küçüktür.**
- Sorgudaki aranan **kelimelerin sayfaların başlığında** veya **linklerde** (kendi kendisine linkler hariç) **olması daha önemlidir.**

28

Konular

- PageRank
  - Term spam
  - PageRank tanımı
  - Web'in yapısı
  - Dead end sayfalar
  - Spider traps
  - Arama motorunda PageRank kullanımı
- Link Spam
  - Link spam yapısı
  - Spam farm analizi
  - TrustRank
  - Spam mass
- Hub ve Otorite Sayfalar
  - HITS algoritması
  - Hub ve otorite tanımı

29

Link Spam

- **PageRank** algoritmasının **geliştirilmesiyle term spam** yöntemleri **etkisini kaybetmiştir.**
- **Spammer'lar** ise PageRank algoritmasını yanıltmaya yönelik **yeni yöntemler geliştirmiştir.**
- PageRank algoritması için **bir sayfanın öneminin yapay bir şekilde artırılması** amacıyla kullanılan yöntemlere link spam denir.
- **Link spam** yöntemlerinin **etkisiz olması için** TrustRank ve spam mass ölçümü gibi yöntemler geliştirilmiştir.

30

Konular

- PageRank
  - Term spam
  - PageRank tanımı
  - Web'in yapısı
  - Dead end sayfalar
  - Spider traps
  - Arama motorunda PageRank kullanımı
- Link Spam
  - Link spam yapısı
  - Spam farm analizi
  - TrustRank
  - Spam mass
- Hub ve Otorite Sayfalar
  - HITS algoritması
  - Hub ve otorite tanımı

31

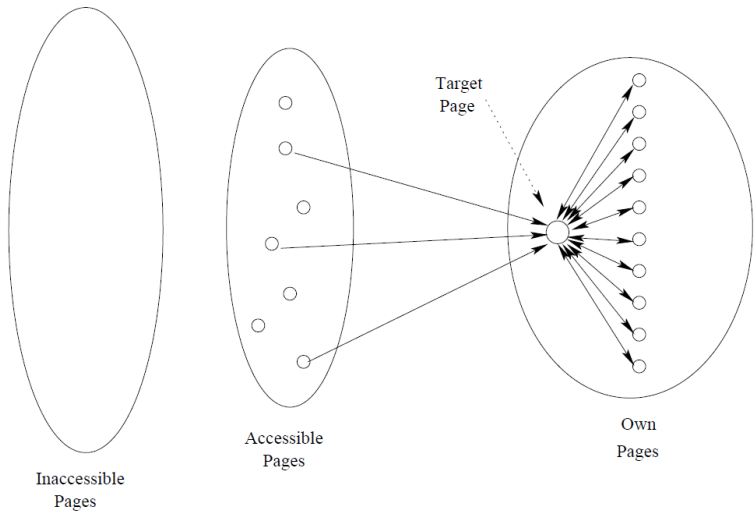
Link spam yapısı

- **Bir sayfanın** veya bir grup sayfanın **PageRank değerini artırmak için oluşturulan sayfa topluluğuna** spam farm denir.
- Spammer açısından Web üç kısma ayrılır:
  - **Erişilemez sayfalar:** Spammer bu sayfaları etkileyemez. Web'in büyük bölümü bu kısımdadır.
  - **Erişilebilir sayfalar:** Spammer tarafından doğrudan kontrol edilmeyen, ancak spammer'ın etkileyebildiği sayfalardır.
  - **Sahip olduğu sayfalar:** Spammer'ın sahip olduğu ve kontrol ettiği sayfalardır.
- **Spam farm spammer'ın sahip olduğu sayfalardan oluşur.**
- Arama motorları tarafından **crawl yapılsa bile spam farm sayfalar kullanıcı için faydasızdır.**

32

Link spam yapısı

- Birtakım yöntemlerle **erişilebilir sayfalardan spam farm sayfalara link verilir.**



33

Link spam yapısı

- Dışarıdan **kendisine link verilmemesi halinde** spam farm içindeki **Web sayfalarını arama motorları crawl yapamaz.**
- Günümüzde çok sayıda **blog** ve **haber sitesi** İnternet kullanıcılarını **yorum post** etmeleri için davet etmektedir.
- **Spammer'lar** bu tür sitelere **spam farm sayfaların linkini içeren çok sayıda yorum mesajı post etmektedirler** ("I agree. Please see my article at www.mySpamFarm.com").
- **Spammer** PageRank değerini yükseltmek istediği sayfadan **spam farm sayfalara link verir.**
- **Spam farm sayfaların tamamı** da sadece PageRank değeri yükseltilmek **istenen sayfaya link verir.**

34

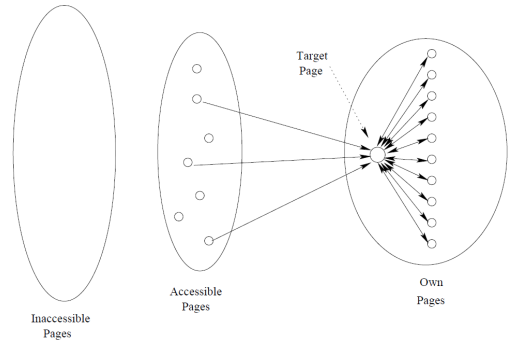
Konular

- PageRank
  - Term spam
  - PageRank tanımı
  - Web'in yapısı
  - Dead end sayfalar
  - Spider traps
  - Arama motorunda PageRank kullanımı
- Link Spam
  - Link spam yapısı
  - Spam farm analizi
  - TrustRank
  - Spam mass
- Hub ve Otorite Sayfalar
  - HITS algoritması
  - Hub ve otorite tanımı

35

Spam farm analizi

- PageRank **taxation parametresi** $\beta = 0.85$ olarak belirlenmiş olsun.
- β **değeri** sayfanın **bir sonraki iterasyonda** ardından gelecek sayfalar için hesaplanan **PageRank değerini etkiler.**
- **Spam farm içerisinde** m **adet destekleyici sayfa** olsun.
- **Web'te toplam** n **sayfa** olsun.
- t **ise bir adet hedef sayfa** olsun.
- x **değeri,** p **adet erişilebilir sayfanın** t **sayfasına link vererek sağlayacağı** toplam PageRank değeri olsun.
- y **değeri,** t **sayfasının hesaplanan PageRank** değeri olsun.



36

Spam farm analizi

- Her bir **destekleyici sayfa için PageRank değeri** aşağıdaki gibi hesaplanır.

$$\beta y/m + (1-\beta)/n$$

- İlk terim ***t*** **sayfasının diğer sayfalara sağlayacağı değeri** göstermektedir.
- *t* sayfasından kendisinin outlinki olan diğer sayfalara βy dağıtılır.
- $\beta y/m$ **ile sonraki** m **adet sayfaya eşit olarak dağıtılır.**
- İkinci terimde ise PageRank'in **(1-**β**) oranı Web'teki tüm sayfalara dağıtılır.**
- Hedef ***t*** **sayfasının** PageRank **değeri** y, üç farklı kaynaktan elde edilir:
  - x, dışarıdan ***t*** **sayfasına** linki olan sayfalardan **gelen PageRank değeridir.**
  - **Destek sayfalarından** ***t*** **sayfasına gelen PageRank değeri.** $\beta(\beta y/m + (1-\beta)/n)$
  - **Web'in tamamında (1-**β**)/**n **ile** ***t*** **sayfasına düşen PageRank değeridir.** Çok küçük bir değerdir, analizi kolaylaştırmak için ihmal edilebilir.

37

Link spam analizi

- **İlk iki kaynaktan** ***t*** **sayfasına gelen toplam PageRank değeri** aşağıdaki gibi yazılabilir:

$$y = x + \beta m\left(\frac{\beta y}{m} + \frac{1-\beta}{n}\right) = x + \beta^2 y + \beta(1-\beta)\frac{m}{n}$$

$$y = \frac{x}{1-\beta^2} + c\frac{m}{n}$$

$$c = \beta(1-\beta)/(1-\beta^2) = \beta/(1+\beta)$$

Örnek

- $\beta = 0{,}85$ olursa $\mathbf{1/(1-\beta^2) = 3{,}6}$ olur.
- $\mathbf{c = \beta/(1+\beta) = 0{,}46}$ olur.
- Spam farm **dışarıdan gelen** x **PageRank değerini 3,6 kat** yükseltmiştir.
- **Web'in içindeki oranına göre** (m/n) (destekleyici sayfa sayısının tüm sayfa sayısına oranı) **PageRank değeri %46 elde edilir.**

38

Link spam analizi

- **Arama motorlarının link spam'i algılayıp elimine etmesi gereklidir.**
- Arama motorları tarafından bir sayfanın çok sayfaya link verdiği ve bu sayfaların da sadece kendisine link verdiği spam farm aranır.
- **Bu yapıya uygun sayfalar indeksten çıkartılır.**
- Spammer'lar farklı yapılar geliştirerek PageRank üzerinde aynı etkiyi elde etmeye çalışmışlardır.
- **Spam farm** sayfaların algılanıp elimine edilmesi **için iki farklı yöntem geliştirilmiştir:**
  - **TrustRank:** Spam sayfaların skorunu azaltır.
  - **Spam mass:** Spam sayfaları tanımlayacak **bir hesaplama yapar** ve tümüyle elimine eder veya **PageRank değerini** önemli oranda **azaltır.**

39

Konular

- PageRank
  - Term spam
  - PageRank tanımı
  - Web'in yapısı
  - Dead end sayfalar
  - Spider traps
  - Arama motorunda PageRank kullanımı
- Link Spam
  - Link spam yapısı
  - Spam farm analizi
  - TrustRank
  - Spam mass
- Hub ve Otorite Sayfalar
  - HITS algoritması
  - Hub ve otorite tanımı

40

TrustRank

- **TrustRank spam olmadığı düşünülen sayfa kümesine (topic) sahiptir.**
- Bir **spam sayfa güvenilir sayfaya kolaylıkla link verebilir**, **ancak güvenilir sayfa spam sayfaya link vermez.**
- Spammer'ların link yerleştirebildiği siteler blog siteleri ve diğer benzeri sitelerdir.
- Bu durumda, **güvenilirliği yüksek olan bir blog sitesi** veya **kullanıcılarından yorum alan saygın bir haber sitesi de güvenilir kabul edilmez!!!**
- Spammer'lar yorum olarak gönderdikleri metin içeriğine kendi sayfalarının linklerini yerleştirir.

41

TrustRank

- **TrustRank** için **güvenilir sayfalardan oluşan bir küme** oluşturulması gereklidir.
- Bu sayfalar **manuel olarak belirlenebilir.**
  - **En yüksek PageRank değerine sahip olan sayfalar** alınabilir.
  - **Link spam bir sayfanın PageRank değerini yükseltir,** ancak **güvenilir sayfalar düzeyine yaklaştıramaz.**
- **Spammer'ların kontrol etmelerinin zor olduğu bir domain alınır** (.edu, .mil, .gov).
  - **Güvenilir sayfalar** genellikle **ABD'deki sitelerden oluşmaktadır.**
  - **Sayfaların** iyi bir dağılım için **farklı ülkelerden de seçilmesi daha uygun olur.**

42

Konular

- PageRank
  - Term spam
  - PageRank tanımı
  - Web'in yapısı
  - Dead end sayfalar
  - Spider traps
  - Arama motorunda PageRank kullanımı
- Link Spam
  - Link spam yapısı
  - Spam farm analizi
  - TrustRank
  - Spam mass
- Hub ve Otorite Sayfalar
  - HITS algoritması
  - Hub ve otorite tanımı

43

Spam mass

- Spam mass yönteminde, **her sayfanın PageRank değerinin bir kısmının spam'den geldiği kabul edilir.**
- Bir p sayfasının **PageRank değeri** r ve **TrustRank değeri** t olsun.
- **TrustRank değeri sayfanın içeriğine** ve **aldığı linklere göre hesaplanır.**
- p sayfasının **spam mass değeri aşağıdaki gibi hesaplanır.**

$$spam\_mass_p = (r - t)/r$$

- **Negatif** veya **küçük pozitif spam mass değerleri** sayfanın muhtemelen **spam olmadığını** gösterir.
- **1'e yakın spam mass değerleri,** sayfanın muhtemelen **spam olduğunu** gösterir.
- Yapılan çalışmalar, **elimine edilen linklerin büyük bölümünün spam farm olduğunu göstermiştir.**

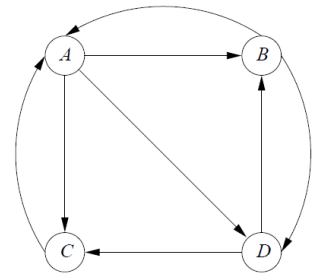
44

Spam mass

Örnek

- Graf için **PageRank, TrustRank** ve **Spam Mass** değerlerini hesaplayalım.
- Tablodaki değerler hesaplanmış olsun.

Node	PageRank	TrustRank	Spam Mass
A	$3/9$	$54/210$	0.229
B	$2/9$	$59/210$	-0.264
C	$2/9$	$38/210$	0.186
D	$2/9$	$59/210$	-0.264



- ***B*** ve ***D***, spam mass değerleri **negatif olduğundan spam değildir.**
- ***A*** ve ***C*** için PageRank değeri TrustRank değerlerinden büyük olduğundan **spam mass değerleri hesaplanır.**

$$A = (3/9 - 54/210)/(3/9) = 0{,}229$$

- Spam mass değeri **0'a yakın** olduğundan **muhtemelen spam değildir.**

45

Konular

- PageRank
  - Term spam
  - PageRank tanımı
  - Web'in yapısı
  - Dead end sayfalar
  - Spider traps
  - Arama motorunda PageRank kullanımı
- Link Spam
  - Link spam yapısı
  - Spam farm analizi
  - TrustRank
  - Spam mass
- Hub ve Otorite Sayfalar
  - HITS algoritması
  - Hub ve otorite tanımı

46

Hub ve Otorite Sayfalar

- Hub ve Otorite yaklaşımı **PageRank algoritmasından kısa süre sonra geliştirilmiştir.**
- **Hub ve otorite algoritması** kısaca HITS (Hyperlink-Induced Topic Search) olarak da adlandırılır.
- **HITS algoritması**, PageRank algoritması gibi **iteratif vektör ve matris hesaplamasını kullanır.**
- HITS algoritması, PageRank algoritmasındaki kullanıcı sorgusundan önceki önişlemleri yapmaz.
- HITS algoritması, kullanıcı sorgusu geldiğinde **sadece gelen sorgu için rank hesaplanır.**
- **Ask arama motoru HITS algoritmasını kullanmaktadır.**

47

Konular

- PageRank
  - Term spam
  - PageRank tanımı
  - Web'in yapısı
  - Dead end sayfalar
  - Spider traps
  - Arama motorunda PageRank kullanımı
- Link Spam
  - Link spam yapısı
  - Spam farm analizi
  - TrustRank
  - Spam mass
- Hub ve Otorite Sayfalar
  - HITS algoritması
  - Hub ve otorite tanımı

48

HITS algoritması

- HITS algoritması **iki tür önemli sayfa tanımlar:**
  - Belirli bir konu hakkında **bilgi sağlayan sayfa (otorite).**
  - Bir konu hakkında **bilgi sağlayan sayfaları gösteren sayfa (hub).**
- Bir bölümdeki **derslerin listesini** bulunduran sayfa **hub** sayfadır.
- Dersler hakkında **bilgi içeren sayfalar** ise **otorite** sayfalardır.
- **PageRank** algoritmasında, eğer bir sayfaya **önemli sayfalar link vermişse** o sayfa **önemlidir**.
- **HITS** algoritmasında, bir sayfa **önemli hub sayfadır** eğer **önemli otorite sayfalara link vermişse.**
- **HITS** algoritmasında, bir sayfa **önemli otorite sayfadır** eğer **önemli hub sayfalar kendisine link vermişse.**

49

Konular

- PageRank
  - Term spam
  - PageRank tanımı
  - Web'in yapısı
  - Dead end sayfalar
  - Spider traps
  - Arama motorunda PageRank kullanımı
- Link Spam
  - Link spam yapısı
  - Spam farm analizi
  - TrustRank
  - Spam mass
- Hub ve Otorite Sayfalar
  - HITS algoritması
  - Hub ve otorite tanımı

50

Hub ve otorite tanımı

- Web sayfalarının **ne kadar iyi hub sayfa olduğunu** veya **ne kadar iyi otorite sayfa olduğunu** gösteren **iki skor tanımlanır.**
- Sayfaların tamamı **h** (hub) ve **a** (otorite) vektörleri ile ifade edilebilir.
- İki vektörde ***i*.değer, *i*.sayfanın hub** veya **otorite değerini** gösterir.
- **Bir sayfanın hub değerini hesaplamak için**, **ardından gelen sayfaların otorite değerleri kullanılır.**
- **Bir sayfanın otorite değerini hesaplamak için**, **önünde olan sayfaların hub değeri kullanılır.**
- İteratif şekilde hesaplanan hub ve otorite değerleri **her adımdan sonra maksimum 1 olacak şekilde ölçeklenir.**
- **h** ve **a** vektörlerinin hesaplanması için **link matrisi oluşturulur.**

51

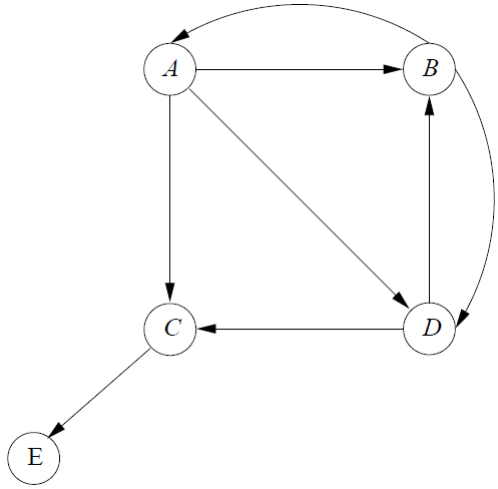
Hub ve otorite tanımı

- ***n* adet sayfa için *L* link matrisi *n*×*n*** boyutunda **kare matristir.**
- L_{ij} **= 1** ise ***i*.sayfadan *j*.sayfaya link vardır**, $L_{ij} = 0$ ise link yoktur.
- L^T ise L link matrisinin transpozudur.
- $L^T_{ij} = 1$ ise ***j*.sayfadan *i*.sayfaya link vardır**, $L^T_{ij} = 0$ ise link yoktur.
- L^T **matrisi** PageRank algoritmasındaki M **geçiş matrisine benzer**. L^T 'nin **1** olduğu yerde M geçiş matrisi **(1/outlink_sayısı)** değerine eşittir.
- **Dead end** ve **spider traps** sayfalar HITS algoritmasının **anlamlı çift vektörü bulmasına engel olmaz.**
- **Taxation** veya graf üzerinde **preprocess yapılması gerekmez.**

52

Hub ve otorite tanımı

Örnek



$$L = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$L^{\mathrm{T}} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

53

Hub ve otorite tanımı

- Bir **sayfanın hub değeri ardındaki otorite sayfaların değeriyle gösterilir.**

$$\mathbf{h} = \lambda L\mathbf{a}$$

- Burada **λ ölçeklendirme sabitidir.**
- Bir **sayfanın otorite değeri önündeki hub sayfaların değeriyle gösterilir.**

$$\mathbf{a} = \mu L^T\mathbf{h}$$

- Burada **μ ölçeklendirme sabitidir.**
- **h** ve **a** birbirinden bağımsız şekilde hesaplanabilir.

$$\mathbf{h} = \lambda\mu LL^T\mathbf{h}$$

$$\mathbf{a} = \lambda\mu L^TL\mathbf{a}$$

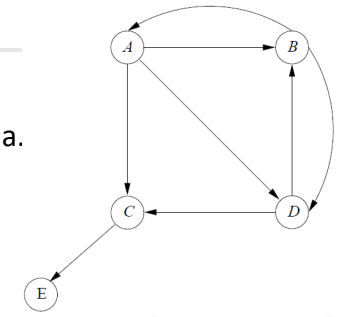
- **h** vektörünün tüm **değerleri 1 alınarak başlanır.**
- $\mathbf{a} = L^T\mathbf{h}$ ve $\mathbf{h} = L\mathbf{a}$ için hesaplama yapılır ardından ölçekleme yapılır.

54

Hub ve otorite tanımı

Örnek

- HITS algoritmasının iki iterasyonu için hesaplama.



$$L = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$L^{\mathrm{T}} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

$\mathbf{h}$	$L^{\mathrm{T}}\mathbf{h}$	$\mathbf{a}$	$L\mathbf{a}$	$\mathbf{h}$
$\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 2 \\ 2 \\ 2 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 1/2 \\ 1 \\ 1 \\ 1 \\ 1/2 \end{bmatrix}$	$\begin{bmatrix} 3 \\ 3/2 \\ 1/2 \\ 2 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 1/2 \\ 1/6 \\ 2/3 \\ 0 \end{bmatrix}$
$\begin{bmatrix} 1/2 \\ 5/3 \\ 5/3 \\ 3/2 \\ 1/6 \end{bmatrix}$	$\begin{bmatrix} 3/10 \\ 1 \\ 1 \\ 9/10 \\ 1/10 \end{bmatrix}$	$\begin{bmatrix} 29/10 \\ 6/5 \\ 1/10 \\ 2 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 12/29 \\ 1/29 \\ 20/29 \\ 0 \end{bmatrix}$	
$L^{\mathrm{T}}\mathbf{h}$	$\mathbf{a}$	$L\mathbf{a}$	$\mathbf{h}$	

A en büyük hub'tır ve değeri 1'dir.

B ve *D* önemli hub'tır.

E'nin hub değeri 0'dır.

B ve C en iyi otoritedir.

55

Ödev

- Topic-sensitive PageRank hakkında bir araştırma ödevi hazırlayınız.

56