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

- Yakın Komşu Arama Uygulamaları
  - Kümelerde Jaccard benzerliği
  - Dokümanlarda benzerlik
  - İşbirlikçi filtreleme
- Dokümanların Parçalar Halinde Gösterimi
  - k-shingles
  - Shingle boyutunun belirlenmesi
- Kümelerin Benzerliğini Koruyan Özetlerin Elde Edilmesi
  - Kümelerin matris gösterimi
  - Minhashing
  - Minhashing ve Jaccard benzerliği
  - Minhash imzaları
  - Minhash imzalarının hesaplanması
- Locality-Sensitive Hashing
  - Locality-Sensitive Hashing ile minhash imzaları
  - Band oluşturma yönteminin analizi

Yakın Komşu Arama Uygulamaları

- Bir küme içerisinde **benzer elemanların belirlenmesi**, **temel bir veri madenciliği problemidir.**
- Örneğin Web sayfalarının veya dokümanların benzerliği veya intihal (plagiarism) hesaplanabilir.
- **Kümelerin benzerliği, kesişim kümesinin büyüklüğüne bakılarak belirlenebilir.**
- **Kesişim kümesindeki** eleman sayısının **birleşim kümesindeki eleman sayısına oranına Jaccard benzerliği** denir.
- Jaccard benzerliği, **dokümanlardaki metin benzerliğini** bulmak için, **işbirlikçi filtrelemede** ise **benzer müşteri** veya **benzer ürünleri** bulmak için kullanılır.
- Dokümanlarda metin benzerliğini bulmak için **küçük parçalara ayırma (shingling)** tekniği kullanılabilir.

3

Konular

- Yakın Komşu Arama Uygulamaları
  - Kümelerde Jaccard benzerliği
  - Dokümanlarda benzerlik
  - İşbirlikçi filtreleme
- Dokümanların Parçalar Halinde Gösterimi
  - k-shingles
  - Shingle boyutunun belirlenmesi
- Kümelerin Benzerliğini Koruyan Özetlerin Elde Edilmesi
  - Kümelerin matris gösterimi
  - Minhashing
  - Minhashing ve Jaccard benzerliği
  - Minhash imzaları
  - Minhash imzalarının hesaplanması
- Locality-Sensitive Hashing
  - Locality-Sensitive Hashing ile minhash imzaları
  - Band oluşturma yönteminin analizi

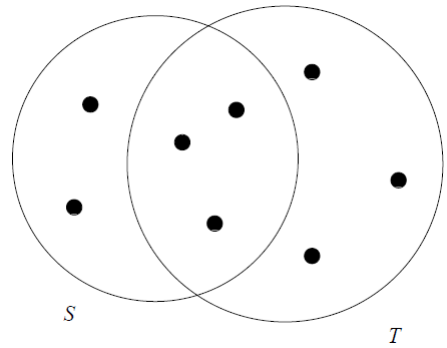
4

Kümelerde Jaccard benzerliği

- S ve T kümeleri için **Jaccard benzerliği** aşağıdaki gibi hesaplanabilir.

$$\text{SIM}(S,T) = |S \cap T|/|S \cup T|$$

- Aşağıdaki şekil için SIM(S, T) = 3/8 'dir.



5

Konular

- Yakın Komşu Arama Uygulamaları
  - Kümelerde Jaccard benzerliği
  - Dokümanlarda benzerlik
  - İşbirlikçi filtreleme
- Dokümanların Parçalar Halinde Gösterimi
  - k-shingles
  - Shingle boyutunun belirlenmesi
- Kümelerin Benzerliğini Koruyan Özetlerin Elde Edilmesi
  - Kümelerin matris gösterimi
  - Minhashing
  - Minhashing ve Jaccard benzerliği
  - Minhash imzaları
  - Minhash imzalarının hesaplanması
- Locality-Sensitive Hashing
  - Locality-Sensitive Hashing ile minhash imzaları
  - Band oluşturma yönteminin analizi

6

Dokümanlarda benzerlik

- Büyük boyutlu dokümanlarda (Web, haber metinleri) metin benzerliğini bulmak önemli bir problem türüdür.
- **Jaccard benzerliği metin benzerliği problemlerinde oldukça başarılıdır.**
- **Metin benzerliği problemlerinde karakter düzeyinde benzerlik önemlidir**, anlam benzerliğine (semantic similarity) bakılmaz.
- Çoğu uygulamada dokümanlar aynı değildir, ancak metin içerisinde aynı olan büyük kısımlar vardır.

7

Dokümanlarda benzerlik

Plagiarism

- İntihal yapılan dokümanların bulunması metin benzerliği problemidir.
- Orijinal dokümana göre bazı kelimeler değiştirilmiş veya cümlelerin yeri değiştirilmiştir.
- Dokümanın %50'den fazlası orijinal dokümanı içerebilir.
- Karakter karakter karşılaştırmak karmaşık intihal dokümanlarını bulamaz.

8

Dokümanlarda benzerlik

Mirror pages

- **Önemli ve popüler Web siteleri yük dağılımı yapmak için çok sayıda host üzerinde bulunabilir.**
- Bu sitelerdeki sayfalar çok benzerdir, ancak tümüyle aynı olmayabilir.
- Bazı sitelerde yıllara göre değişiklikler yapılmış olabilir (ders isimleri, ders sunumları, vb.).
- **Bu tür sayfaların tespit edilmesi arama motorları için önemlidir.**
- Bu şekilde benzer Web sayfalarının arama motorunun ilk sonuç sayfasında birlikte görüntülenmesi istenmez.

9

Dokümanlarda benzerlik

Aynı kaynaktan alınan makaleler

- **Bir haber makalesi çok sayıda gazeteye dağıtılır** veya Web sitesinde yayınlanabilir.
- **Her gazete haber metninde bazı değişiklikler yapabilir.**
- Bazı paragraflar çıkarılabilir veya ekleme yapılabilir.
- Bazıları, reklam, logo ekleyebilir veya kendi sitelerindeki başka makalelere link verebilir.
- Haber toplama uygulamaları (Google News) bu şekilde benzer makalelerin tümünü tek haber metni olarak göstermek ister.
- **Bu Web sayfaları metin benzerliğine sahiptir**, ancak **tümüyle aynı değildir.**

10

Konular

- Yakın Komşu Arama Uygulamaları
  - Kümelerde Jaccard benzerliği
  - Dokümanlarda benzerlik
  - İşbirlikçi filtreleme
- Dokümanların Parçalar Halinde Gösterimi
  - k-shingles
  - Shingle boyutunun belirlenmesi
- Kümelerin Benzerliğini Koruyan Özetlerin Elde Edilmesi
  - Kümelerin matris gösterimi
  - Minhashing
  - Minhashing ve Jaccard benzerliği
  - Minhash imzaları
  - Minhash imzalarının hesaplanması
- Locality-Sensitive Hashing
  - Locality-Sensitive Hashing ile minhash imzaları
  - Band oluşturma yönteminin analizi

11

İşbirlikçi filtreleme

- Bazı uygulamalarda kullanıcılara öneri yapılması gerekebilir.
- **Öneri listesi benzer kullanıcıların davranışına göre oluşturulabilir.**
- Kümelerin bu tür benzerliğini kullanan uygulamalar işbirlikçi filtreleme (collaborative filtering) uygulamaları olarak adlandırılır.

12

İşbirlikçi filtreleme

On-line alışveriş

- On-line alışveriş sitelerinde milyonlarca müşteri ve ürün vardır.
- **Müşterilerin satın aldıkları ürünlere göre Jaccard benzerliği hesaplanabilir.**
- Aynı türde (bilim kurgu, komedi) fakat farklı kitapları alan müşterilerin aynı grupta yer alması gereklidir.

13

İşbirlikçi filtreleme

Film puanlama

- NetFlix, hangi filmin müşteriler tarafından kiralandığını saklar ve müşterilerin film için puanlamasını alır.
- **Aynı müşteriler tarafından kiralanan** ve **yüksek puanlanan filmler benzer alınabilir.**
- **Aynı filmleri kiralayan** ve **yüksek puanlayan müşteriler benzer alınabilir.**
- Binary veri (aldı/almadı, beğendi/beğenmedi) yerine, dereceli puanlama verisi varsa (1,2,3,4,5) farklı yöntemler uygulanabilir:
  - **Düşük değerli müşteri-film ikilileri gözardı edilebilir.**
  - **Yüksek puanlı olanlar için beğendi, düşük puanlı olanlar için beğenmedi atanır.** Ardından, Jaccard benzerliği hesaplanabilir.
  - **Bir film puan değerine göre müşteri kümesinde tekrarlı alınabilir.** Ardından, **Jaccard bag benzerliği** hesaplanabilir.

14

İşbirlikçi filtreleme

- Jaccard bag benzerliği için, kesişim kümesindeki eleman sayısının kümelerdeki tüm elemanların toplam sayısına (kümelerin boyutu) oranı hesaplanır.

Örnek:

{a, **a, a, b**} ve {**a, a, b**, b, c} için **Jaccard bag benzerliği 3/9 = 1/3** olur.

Jaccard bag benzerliği kesişim kümesi **{a, a, b}** olup; boyutu **3**'tür. Toplam eleman sayısı ise **9**'dur.

Jaccard benzerliğinde ise kesişim **{a, a, b}**, birleşim {a, **a**, **a**, **b**, b, c} olur (3/6).

15

Konular

- Yakın Komşu Arama Uygulamaları
  - Kümelerde Jaccard benzerliği
  - Dokümanlarda benzerlik
  - İşbirlikçi filtreleme
- Dokümanların Parçalar Halinde Gösterimi
  - k-shingles
  - Shingle boyutunun belirlenmesi
- Kümelerin Benzerliğini Koruyan Özetlerin Elde Edilmesi
  - Kümelerin matris gösterimi
  - Minhashing
  - Minhashing ve Jaccard benzerliği
  - Minhash imzaları
  - Minhash imzalarının hesaplanması
- Locality-Sensitive Hashing
  - Locality-Sensitive Hashing ile minhash imzaları
  - Band oluşturma yönteminin analizi

16

Dokümanların Parçalar Halinde Gösterimi

- **Dokümanların kümeler halinde gösterimi sözlüksel benzerliğin (lexical similarity) bulunması amacıyla kullanılan en etkin yöntemdir.**
- Kümeler, **doküman içerisindeki kısa string'ler kullanılarak oluşturulur.**
- Böylece, kısa string'lerin yerleri farklı olsa bile dokümanlar çok sayıda ortak elemana sahip olacaktır.
- **Shingling, dokümanı küçük string'lere ayırmak için kullanılan basit ve yaygın kullanılan yöntemdir.**

17

Konular

- Yakın Komşu Arama Uygulamaları
  - Kümelerde Jaccard benzerliği
  - Dokümanlarda benzerlik
  - İşbirlikçi filtreleme
- Dokümanların Parçalar Halinde Gösterimi
  - k-shingles
  - Shingle boyutunun belirlenmesi
- Kümelerin Benzerliğini Koruyan Özetlerin Elde Edilmesi
  - Kümelerin matris gösterimi
  - Minhashing
  - Minhashing ve Jaccard benzerliği
  - Minhash imzaları
  - Minhash imzalarının hesaplanması
- Locality-Sensitive Hashing
  - Locality-Sensitive Hashing ile minhash imzaları
  - Band oluşturma yönteminin analizi

18

k-shingles

- Bir doküman karakterlerden oluşan bir string'tir.
- Bir doküman için bir **k-shingle** k uzunluğundaki herhangi bir **substring**'dir.
- Her doküman kendi içerisinde bir veya daha fazla sayıda bulunan k-shingle kümesi ile ilişkilendirilebilir.

Örnek:

D = abcdabd dokümanı için **2-shingle kümesi = {ab, bc, cd, da, bd}** olur.

Dokümanda ab iki kez vardır ancak 2-shingle kümesinde bir kez alınır.

Shingle-bag olarak oluşturulursa **tüm tekrarlar da** shingle kümesinde **yer alır.**

19

k-shingles

- White space (blank, tab, newline, vb.) için farklı yöntemler kullanılabilir.
- **Birden fazla sıralı white space için bir tane blank konulabilir.**
- Böylelikle birden fazla kelime bulunan shingle'lar ayırt edilebilir.

Örnek:

''The plane was ready for touch down'' ile *''The quarterback scored a touchdown''* cümlelerinin **benzerliğini 9-shingle kullanarak bulalım.**

Boşluklar kaldırılmazsa, ilk cümle için *''touch dow''* ve *''ouch down''* 9-shingle'ları yer alır. İkinci cümle için *''touchdown''* 9-shingle yer alır.

Boşluklar kaldırılırsa her ikisinde de *''touchdown''* yer alır.

20

Konular

- Yakın Komşu Arama Uygulamaları
  - Kümelerde Jaccard benzerliği
  - Dokümanlarda benzerlik
  - İşbirlikçi filtreleme
- Dokümanların Parçalar Halinde Gösterimi
  - k-shingles
  - Shingle boyutunun belirlenmesi
- Kümelerin Benzerliğini Koruyan Özetlerin Elde Edilmesi
  - Kümelerin matris gösterimi
  - Minhashing
  - Minhashing ve Jaccard benzerliği
  - Minhash imzaları
  - Minhash imzalarının hesaplanması
- Locality-Sensitive Hashing
  - Locality-Sensitive Hashing ile minhash imzaları
  - Band oluşturma yönteminin analizi

21

Shingle boyutunun belirlenmesi

- **Shingle boyutu çok küçük olursa** hemen hemen **tüm dokümanlarda yer alabilir.**
- Dokümanların **Jaccard benzerliği yüksek olur.** Ancak, kelimeler veya cümleler ortak değildir.
- k = 1 alınırsa, tüm Web sayfalarında benzerlik yüksek çıkar.
- **k değerinin yeterli büyüklükte alınması gereklidir.**
- Herhangi bir k-shingle'ın **diğer herhangi bir dokümanda olma olasılığı düşük olmalıdır.**
- ASCII tablosundan 27 karakter kullanılan bir metin için k = 5 alınırsa 27^5 = 14.348.907 farklı shingle olasılığı vardır.
- Yapılan çalışmalar, **e-posta için k=5**, **büyük dokümanlar için k=9** alındığında benzerliğin iyi belirlendiğini göstermiştir.

22

Shingle boyutunun belirlenmesi

Hashing shingles

- Substring'leri shingle olarak kullanmak yerine, **hash fonksiyonu ile eşleştirilen sayılar kullanılabilir.**
- **k-shingle string'ler yerine**, daha **küçük boyuta sahip sayılar** shingle yerine kullanılabilir.
- Örneğin, 9-shingle ile dokümanı göstermek yerine, hash fonksiyonu ile 0 - (2^{32}-1) arasındaki sayılarla eşleştirme yapılabilir.
- Bu durumda, **bir shingle 9 byte yerine 4 byte ile gösterilir.**
- Bir shingle, 4-shingle ile gösterimi kadar yer kaplar, ancak 9-shingle ile ifade edilmiş olur ve **dokümanlar daha iyi ayırt edilir.**

23

Shingle boyutunun belirlenmesi

Kelimelerden shingle oluşturulması

- Haber makaleleri için **shingle'lar makaledeki kelimelerle oluşturulabilir.**
- Haber makalelerinde **çok sayıda stop word (and, the, to, vb.) bulunur.**
- Çoğu uygulamada stop word'ler göz ardı edilir.
- Yapılan çalışmalarda, **haber makalelerinde stop word** ve **ardındaki iki kelimenin shingle olarak alınmasının** başarılı sonuç verdiği görülmüştür.
- Böylelikle haber metni etrafındaki **kısa metinli reklamlar** ve **diğer eklentiler elimine edilebilmektedir.**
- **Aynı makaleye sahip**, ancak **etrafındaki metinler (reklam, öneri) farklı olan sayfaların Jaccard benzerliği yüksek olur.**
- **Farklı makaleye** sahip **etrafındaki metinler aynı olanların Jaccard benzerliği düşük olur.**

24

Shingle boyutunun belirlenmesi

Örnek

Haber makalesinde,
‘*A* spokesperson *for the* Sudzo Corporation revealed today *that* studies *have* shown *it is* good *for* people *to* buy Sudzo products.’’
olsun.

Metnin etrafında ‘’Buy Sudzo.’’ şeklinde reklam olsun.

- Eğik ve altı çizgili kelimeler stop word olarak alınabilir.
- Haber metni için ilk üç shingle aşağıdaki gibi olabilir:

  A spokesperson for
  for the Sudzo
  the Sudzo Corporation
- Cümlede toplam 9 shingle oluşturulur (cümlede 9 stop word var) ve hiçbirisi reklam metnini içermez.

25

Konular

- Yakın Komşu Arama Uygulamaları
  - Kümelerde Jaccard benzerliği
  - Dokümanlarda benzerlik
  - İşbirlikçi filtreleme
- Dokümanların Parçalar Halinde Gösterimi
  - k-shingles
  - Shingle boyutunun belirlenmesi
- Kümelerin Benzerliğini Koruyan Özetlerin Elde Edilmesi
  - Kümelerin matris gösterimi
  - Minhashing
  - Minhashing ve Jaccard benzerliği
  - Minhash imzaları
  - Minhash imzalarının hesaplanması
- Locality-Sensitive Hashing
  - Locality-Sensitive Hashing ile minhash imzaları
  - Band oluşturma yönteminin analizi

26

Kümelerin Benzerliğini Koruyan Özetlerin Elde Edilmesi

- **Shingle kümeleri çok büyük boyuttadır.**
- Bir doküm için oluşturulan **4 byte uzunluktaki shingle'lar** bile **orijinal dokümanın 4 katı alan kaplar.**
- Milyonlarca dokümanın **shingle kümelerinin hafızada saklanması mümkün değildir.**
- Bir depolama biriminde saklansa bile **doküman çiftlerinin benzerliklerini bulmak çok uzun zaman alır.**
- Büyük boyuttaki kümelerin **küçük boyuttaki imzaları (signatures)** ile saklanması daha uygundur.
- **İki kümenin sadece imzası kullanılarak Jaccard benzerliğinin tahmin edilmesi amaçlanmaktadır.**
- **İmzalar ile** kesin benzerliğin bulunması mümkün değildir, **yaklaşık bir sonuç bulunur.**

27

Konular

- Yakın Komşu Arama Uygulamaları
  - Kümelerde Jaccard benzerliği
  - Dokümanlarda benzerlik
  - İşbirlikçi filtreleme
- Dokümanların Parçalar Halinde Gösterimi
  - k-shingles
  - Shingle boyutunun belirlenmesi
- Kümelerin Benzerliğini Koruyan Özetlerin Elde Edilmesi
  - Kümelerin matris gösterimi
  - Minhashing
  - Minhashing ve Jaccard benzerliği
  - Minhash imzaları
  - Minhash imzalarının hesaplanması
- Locality-Sensitive Hashing
  - Locality-Sensitive Hashing ile minhash imzaları
  - Band oluşturma yönteminin analizi

28

Kümelerin matris gösterimi

- Aşağıdaki matriste (karakteristik matris) ilk sütun evrensel küme elemanlarını, diğer sütunlar kümeleri göstermektedir.

Element	S_1	S_2	S_3	S_4
a	1	0	0	1
b	0	0	1	0
c	0	1	0	1
d	1	0	1	1
e	0	0	1	0

- 1 değeri ilgili elemanın kümede olduğunu, 0 değeri olmadığını gösterir.
- Evrensel Küme = $\{a, b, c, d, e\}$
  $S_1 = \{a, d\}$
  $S_2 = \{c\}$
  $S_3 = \{b, d, e\}$
  $S_4 = \{a, c, d\}$

29

Kümelerin matris gösterimi

- Kümeler genellikle matris şeklinde saklanmaz, matris görünümü sunum amacıyla kullanılır.
- Matrislerin hemen hemen tamamı sparse matris şeklindedir.
- **Matrislerdeki 1'lerin pozisyonu saklanır** ve toplamda daha küçük alan kaplar.
- Karakteristik matriste, satırlar ürünleri sütunlar ise müşterilerin ürünleri alıp almadığını gösterebilir.

30

Konular

- Yakın Komşu Arama Uygulamaları
  - Kümelerde Jaccard benzerliği
  - Dokümanlarda benzerlik
  - İşbirlikçi filtreleme
- Dokümanların Parçalar Halinde Gösterimi
  - k-shingles
  - Shingle boyutunun belirlenmesi
- Kümelerin Benzerliğini Koruyan Özetlerin Elde Edilmesi
  - Kümelerin matris gösterimi
  - Minhashing
  - Minhashing ve Jaccard benzerliği
  - Minhash imzaları
  - Minhash imzalarının hesaplanması
- Locality-Sensitive Hashing
  - Locality-Sensitive Hashing ile minhash imzaları
  - Band oluşturma yönteminin analizi

31

Minhashing

- Kümeler için oluşturulan **imzalar (minhash)** çok sayıda hesaplamanın sonucunda elde edilir.
- Matristeki bir sütunun **minhash gösterimi için satırların bir permütasyonu alınır.**
- **Bir sütun (küme) için minhash değeri** o sütundaki **ilk 1 olan satırın numarasıdır (veya elemanıdır).**
- Her kümenin yukarıdan itibaren ilk 1 değerinin farklı elemanla başladığı permütasyon bulunursa, kümeler daha iyi ayırt edilebilir.

32

Minhashing

- Örnekteki elemanlar için $\{b, e, a, d, c\}$ permütasyonu alınsın.

Element	S_1	S_2	S_3	S_4
b	0	0	1	0
e	0	0	1	0
a	1	0	0	1
d	1	0	1	1
c	0	1	0	1

$h(S_1) = a$
$h(S_2) = c$
$h(S_3) = b$
$h(S_4) = a$

33

Konular

- Yakın Komşu Arama Uygulamaları
  - Kümelerde Jaccard benzerliği
  - Dokümanlarda benzerlik
  - İşbirlikçi filtreleme
- Dokümanların Parçalar Halinde Gösterimi
  - k-shingles
  - Shingle boyutunun belirlenmesi
- Kümelerin Benzerliğini Koruyan Özetlerin Elde Edilmesi
  - Kümelerin matris gösterimi
  - Minhashing
  - Minhashing ve Jaccard benzerliği
  - Minhash imzaları
  - Minhash imzalarının hesaplanması
- Locality-Sensitive Hashing
  - Locality-Sensitive Hashing ile minhash imzaları
  - Band oluşturma yönteminin analizi

34

Minhashing ve Jaccard benzerliği

- **Rastgele oluşturulan permütasyonda,** iki küme için **minhash** fonksiyonu ile elde edilen **değerlerin aynı olma olasılığı, Jaccard benzerliğine eşittir.**

Element	S_1	S_2	S_3	S_4
b	0	0	1	0
e	0	0	1	0
a	1	0	0	1
d	1	0	1	1
c	0	1	0	1

- Herhangi iki sütun için (S_1 ve S_2) üç durum ortaya çıkar:
  - İkisinde de 1 vardır.
  - İkisinden birisinde 1, diğerinde 0 vardır.
  - İkisinde de 0 vardır.

35

Minhashing ve Jaccard benzerliği

- S_1 ve S_2 için **sparse matris elde edilir.** İkisinde de 0 daha fazladır.

Element	S_1	S_2	S_3	S_4
b	0	0	1	0
e	0	0	1	0
a	1	0	0	1
d	1	0	1	1
c	0	1	0	1

- İkisinin de 1 olduğu durum $S_1 \cap S_2$, birisinin 1 olduğu durum $S_1 \cup S_2$ dir.
- **$(S_1 \cap S_2)$ / $(S_1 \cup S_2)$ oranı** iki kümenin **minhash değerlerinin ($h(S_1)$, $h(S_2)$)** (Jaccard benzerliği) **eşit olma olasılığını gösterir.**
- Üstten itibaren **ilk 1 olan satırda her ikisi de 1 ise $h(S_1) = h(S_2)$** alınır.
- **İlk 1 gelen satırda diğer küme 0 ise, $h(S_1) \neq h(S_2)$ alınır.** 1 olan satır o küme için minhash değeridir.
- **$h(S_1) = h(S_2)$ olasılığı, $(S_1 \cap S_2)$ / $(S_1 \cup S_2)$ oranına bağlıdır.**

36

Konular

- Yakın Komşu Arama Uygulamaları
  - Kümelerde Jaccard benzerliği
  - Dokümanlarda benzerlik
  - İşbirlikçi filtreleme
- Dokümanların Parçalar Halinde Gösterimi
  - k-shingles
  - Shingle boyutunun belirlenmesi
- Kümelerin Benzerliğini Koruyan Özetlerin Elde Edilmesi
  - Kümelerin matris gösterimi
  - Minhashing
  - Minhashing ve Jaccard benzerliği
  - Minhash imzaları
  - Minhash imzalarının hesaplanması
- Locality-Sensitive Hashing
  - Locality-Sensitive Hashing ile minhash imzaları
  - Band oluşturma yönteminin analizi

37

Minhash imzaları

- Bir küme topluluğunu gösteren **karakteristik matris M için**, **rastgele** seçilen n sayısı ($n \leq N!$, N evrensel küme eleman sayısı) kadar **permütasyon oluşturulabilir.**
- Herhangi **bir S kümesi için minhash fonksiyonu n kez çağrılarak** S kümesi için **n farklı minhash değeri belirlenir (h_1, h_2,… h_n).**
- Elde edilen **$[h_1(S), h_2(S), \ldots, h_n(S)]$ vektörü, S kümesi için minhash signature** olarak adlandırılır.
- **M karakteristik matrisinden signature matrisi elde edilir.**
- **Signature matrisinin sütun sayısı,** M matrisinin sütun sayısı ile aynıdır, ancak **satır sayısı farklı olabilir** (hash fonksiyonu sayısı kadar).

38

Konular

- Yakın Komşu Arama Uygulamaları
  - Kümelerde Jaccard benzerliği
  - Dokümanlarda benzerlik
  - İşbirlikçi filtreleme
- Dokümanların Parçalar Halinde Gösterimi
  - k-shingles
  - Shingle boyutunun belirlenmesi
- Kümelerin Benzerliğini Koruyan Özetlerin Elde Edilmesi
  - Kümelerin matris gösterimi
  - Minhashing
  - Minhashing ve Jaccard benzerliği
  - Minhash imzaları
  - Minhash imzalarının hesaplanması
- Locality-Sensitive Hashing
  - Locality-Sensitive Hashing ile minhash imzaları
  - Band oluşturma yönteminin analizi

39

Minhash imzalarının hesaplanması

- **Büyük** bir **karakteristik matrisi** doğrudan **permütasyonu alarak kullanmak hem süre hem de iş yükü açısından uygun değildir.**
- Milyonlarca veya milyarlarca satırın permütasyonu **zaman alıcıdır** ve **satırların** yeniden **sıralanması gerekir.**
- **Karakteristik matrisin permütasyon etkisini** rastgele **hash fonksiyonu kullanarak simüle etmek mümkündür.**
- Satırların ***n*** **tane rastgele permütasyonu yerine**, satırlar için ***n*** **tane rastgele hash fonksiyonu seçilir.**
- **SIG(*i*, *c*),** signature matriste **bir elemandır** ve ***i*.hash fonksiyonun *c* sütunundaki değerini gösterir.**
- **Başlangıçta** tüm i ve c 'ler için **SIG(*i*, *c*) =** ∞ olarak atanır.

40

Minhash imzalarının hesaplanması

- Herhangi bir satır r aşağıdaki şekilde elde edilir:
  - $h_1(r), h_2(r), \ldots, h_n(r)$ **hesaplanır.**
  - c **sütununda 0 değeri varsa**, herhangi bir **işlem yapılmaz.**
  - Eğer c **sütununda 1 değeri varsa**, $i = 1, 2, \ldots, n$ için **SIG(*i*, *c*) ve $h_i(r)$ den küçük olan atanır.** $\min(\text{SIG}(i, c), h_i(r))$

Row	S_1	S_2	S_3	S_4	$x+1 \mod 5$	$3x+1 \mod 5$
0	1	0	0	1	1	1
1	0	0	1	0	2	4
2	0	1	0	1	3	2
3	1	0	1	1	4	0
4	0	0	1	0	0	3

41

Minhash imzalarının hesaplanması

Örnek:

- Aşağıdaki karakteristik matris için satırlardaki harfler yerine satır numaraları (0, 1, 2, 3, 4) yazılmıştır.

Element	S_1	S_2	S_3	S_4
a	1	0	0	1
b	0	0	1	0
c	0	1	0	1
d	1	0	1	1
e	0	0	1	0

Row	S_1	S_2	S_3	S_4
0	1	0	0	1
1	0	0	1	0
2	0	1	0	1
3	1	0	1	1
4	0	0	1	0

- $h_1(x) = x+1 \mod 5$ ve $h_2(x) = 3x + 1 \mod 5$ seçilmiştir.
- Hash fonksiyonları **satır numaralarını giriş olarak alır.**

42

Minhash imzalarının hesaplanması

Row	S_1	S_2	S_3	S_4
0	1	0	0	1
1	0	0	1	0
2	0	1	0	1
3	1	0	1	1
4	0	0	1	0

Örnek - devam:

- Başlangıç imza matrisi aşağıdadır.

$h_1(x) = x+1 \mod 5$

$h_2(x) = 3x+1 \mod 5$

	S_1	S_2	S_3	S_4
h_1	∞	∞	∞	∞
h_2	∞	∞	∞	∞

- $h_1(0) = 1$, $h_2(0) = 1$ ($1 < \infty$) olur. **Satır 0 sadece S_1 ve S_4 'te 1 değerine sahiptir.** S_1 ve S_4 sütunlarına 1 atanır. Yeni tahmin matrisi aşağıdadır.

	S_1	S_2	S_3	S_4
h_1	1	∞	∞	1
h_2	1	∞	∞	1

- $h_1(1) = 2$, $h_2(1) = 4$ olur. **Satır 1 sadece S_3 'te 1 değerine sahiptir.** S_3 sütununa 2 ve 4 atanır. Yeni tahmin matrisi aşağıdadır.

	S_1	S_2	S_3	S_4
h_1	1	∞	2	1
h_2	1	∞	4	1

43

Minhash imzalarının hesaplanması

Row	S_1	S_2	S_3	S_4
0	1	0	0	1
1	0	0	1	0
2	0	1	0	1
3	1	0	1	1
4	0	0	1	0

Örnek - devam:

$h_1(x) = x+1 \mod 5$

$h_2(x) = 3x+1 \mod 5$

- $h_1(2) = 3$, $h_2(2) = 2$ olur. **Satır 2 sadece S_2 ve S_4 'te 1 değerine sahiptir.** S_2 sütununa 3 ve 2 atanır. S_4 sütunu 1 (1 < 3) ve 1 (1 < 2) olduğundan değişmez.

	S_1	S_2	S_3	S_4
h_1	1	3	2	1
h_2	1	2	4	1

- $h_1(3) = 4$, $h_2(3) = 0$ olur. **Satır 3, S_1, S_3 ve S_4 'te 1 değerine sahiptir.** h_1 satırındaki tüm değerler 4 ten küçüktür. h_2 satırındaki tüm değerler 0 dan büyüktür.

	S_1	S_2	S_3	S_4
h_1	1	3	2	1
h_2	0	2	0	0

- $h_1(4) = 0$, $h_2(4) = 3$ olur. **Satır 4 sadece S_3 'te 1 değerine sahiptir.**

	S_1	S_2	S_3	S_4
h_1	1	3	0	1
h_2	0	2	0	0

44

Minhash imzalarının hesaplanması

Row	S_1	S_2	S_3	S_4
0	1	0	0	1
1	0	0	1	0
2	0	1	0	1
3	1	0	1	1
4	0	0	1	0

Örnek - devam:

	S_1	S_2	S_3	S_4
h_1	1	3	0	1
h_2	0	2	0	0

- **Jaccard benzerliği imza matrisinden tahmin edilebilir.**
- **S_1 ve S_4 aynıdır** ve $\mathrm{SIM}(S_1, S_4) = 1$ olarak tahmin edilebilir.
- S_1 ve S_4 için gerçek **Jaccard benzerliği $\mathbf{SIM}(S_1, S_4) = 2/3$ tür.**
- **S_1 ve S_3 'ün yarısı aynıdır** ve $\mathrm{SIM}(S_1, S_3) = 1/2$ olarak tahmin edilebilir.
- S_1 ve S_3 için gerçek **Jaccard benzerliği $\mathbf{SIM}(S_1, S_3) = 1/4$ tür.**
- **S_1 ve S_2 'nin benzerliği 0 dır.** $\mathrm{SIM}(S_1, S_2) = 0$ olarak tahmin edilebilir.
- S_1 ve S_2 için gerçek **Jaccard benzerliği $\mathbf{SIM}(S_1, S_2) = 0$ dır.**
- **Büyük ölçekli örnekler için tahmin edilen ve gerçek değer yakın olur.**

45

Konular

- Yakın Komşu Arama Uygulamaları
  - Kümelerde Jaccard benzerliği
  - Dokümanlarda benzerlik
  - İşbirlikçi filtreleme
- Dokümanların Parçalar Halinde Gösterimi
  - k-shingles
  - Shingle boyutunun belirlenmesi
- Kümelerin Benzerliğini Koruyan Özetlerin Elde Edilmesi
  - Kümelerin matris gösterimi
  - Minhashing
  - Minhashing ve Jaccard benzerliği
  - Minhash imzaları
  - Minhash imzalarının hesaplanması
- Locality-Sensitive Hashing
  - Locality-Sensitive Hashing ile minhash imzaları
  - Band oluşturma yönteminin analizi

46

Locality-Sensitive Hashing

- **Minhashing ile büyük dokümanlar** benzerlikleri yaklaşık korunarak **küçük imzalar şeklinde ifade edilir.**
- Ancak, **en çok benzeyen doküman çiftlerini etkin ve kısa sürede bulmak mümkün değildir.** (Doküman az bile olsa doküman çifti çok fazladır.)
- 1.000.000 doküman 1.000'er byte imza ile gösterilirse 1 GB alan gerekir.
- Toplam 0,5 trilyon çift vardır. Her çift için benzerlik 1 mikrosaniyede yapılırsa, **tüm benzerlikler yaklaşık 6 (5,787) günde hesaplanır.**

$$C(1.000.000, 2) = \binom{1.000.000}{2} = \frac{1.000.000!}{(1.000.000-2)!\,2!} = 0{,}5.10^{12}$$

- Tüm çiftlerin benzerliği hesaplanacaksa süre kısaltılamaz.
- Ancak, **genellikle en çok benzeyen çiftler bulunmaya çalışılır.**
- Arama için locality-sensitive hashing (near-neighbor search) kullanılır.

47

Konular

- Yakın Komşu Arama Uygulamaları
  - Kümelerde Jaccard benzerliği
  - Dokümanlarda benzerlik
  - İşbirlikçi filtreleme
- Dokümanların Parçalar Halinde Gösterimi
  - k-shingles
  - Shingle boyutunun belirlenmesi
- Kümelerin Benzerliğini Koruyan Özetlerin Elde Edilmesi
  - Kümelerin matris gösterimi
  - Minhashing
  - Minhashing ve Jaccard benzerliği
  - Minhash imzaları
  - Minhash imzalarının hesaplanması
- Locality-Sensitive Hashing
  - Locality-Sensitive Hashing ile minhash imzaları
  - Band oluşturma yönteminin analizi

48

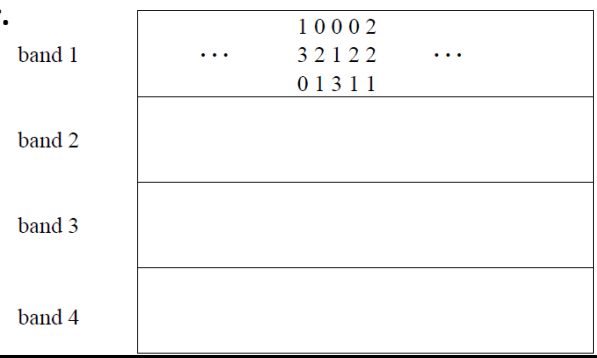
Locality-Sensitive Hashing ile minhash imzaları

- Locality-sensitive hashing ile **aynı elemana çok sayıda hashing yapılır.**
- **Benzer çiftler** farklı olan çiftlere göre daha çok **aynı sıraya/lokasyona yerleşir.**
- **Aynı hash değerine** sahip çiftler **aday çift** olarak alınır.
- **Sadece aday çiftlerin** benzer olduğu varsayılarak **benzerlikleri hesaplanır.**
- Benzer olmayıp da hash değeri aynı olan çiftlerin **false positive (FP)** az sayıda olması beklenir.
- **FP** olanlar tüm çiftler içerisinde genellikle **küçük bir kısmı oluşturur.**
- Çok küçük bir kısımda **false negative (FN)** olabilir. (Benzer olup da farklı hash değeri alanlar.)
- Büyük kısım **true positive (TP)** ve **true negative (TN)** olur.

49

Locality-Sensitive Hashing ile minhash imzaları

- **Signature matrisi *b* tane banda bölünür.** Her band *r* adet satıra sahiptir.
- Her band bir hash fonksiyonuna sahiptir ve **satırdaki integer değerleri büyük bir hash tablosuna eşleştirir.**
- **Her band için ayrı bir hash tablosu kullanılır.**
- Aşağıda 2. ve 4. sütunlar (0, 2, 1), aynı hash tablosunda **aynı bucket'a eşleştirilirler.**



50

Locality-Sensitive Hashing ile minhash imzaları

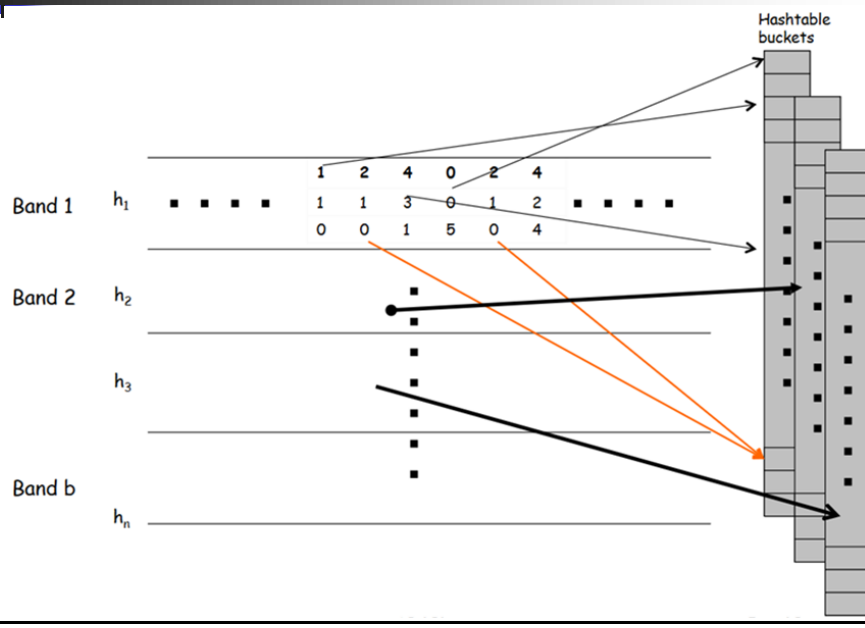
- Aşağıda 1. (1, 3, 0) ve 2. sütunların (0, 2, 1) hash tablosunda **farklı bucket'a eşleştirilmesi beklenir.**

band 1	...	1 0 0 0 2 3 2 1 2 2 0 1 3 1 1	...
band 2			
band 3			
band 4			

- Band 1 hash fonksiyonu ile **aynı bucket'a eşleştirilmeyenler**, **diğer 3 band için aynı bucket'a eşleştirilme olasılığına sahiptir.**

51

Locality-Sensitive Hashing ile minhash imzaları



52

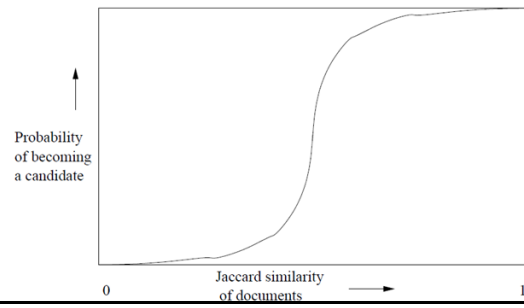
Konular

- Yakın Komşu Arama Uygulamaları
  - Kümelerde Jaccard benzerliği
  - Dokümanlarda benzerlik
  - İşbirlikçi filtreleme
- Dokümanların Parçalar Halinde Gösterimi
  - k-shingles
  - Shingle boyutunun belirlenmesi
- Kümelerin Benzerliğini Koruyan Özetlerin Elde Edilmesi
  - Kümelerin matris gösterimi
  - Minhashing
  - Minhashing ve Jaccard benzerliği
  - Minhash imzaları
  - Minhash imzalarının hesaplanması
- Locality-Sensitive Hashing
  - Locality-Sensitive Hashing ile minhash imzaları
  - Band oluşturma yönteminin analizi

53

Band oluşturma yönteminin analizi

- Toplam b tane band ve her birinde r adet satır olsun.
- İki çiftin Jaccard benzerliği s olsun. **Aday çift olma olasılığı** ile **Jaccard benzerliği** arasında aşağıdaki gibi **S-eğrisi** oluşur.
- Benzerlik için threshold değeri, b (bant sayısı)ve r (satır sayısı) nin fonksiyonu $(1/b)^{1/r}$ kullanılarak hesaplanır.
- $b = 16$, $r = 4$ için $(1/16)^{1/4} = 1/2$.
- Jaccard benzerliği $s \geq 0.5$ çiftlerin aday çift olma olasılığı yüksektir.



54

Ödev

- Locality-sensitive hashing için hash fonksiyonu seçimine yönelik bir makale ödevi hazırlayınız.