# Bilgi Erişimi ve Öneri Sistemleri

## Yapay Zekâ ve Veri Dersleri Rehberi · Kısım IV: Veri

Dizinleme | Sorgu ve sıralama | İşbirlikçi süzme | Değerlendirme | Yoğun vektör araması | Komşu dersler

Sercan KÜLCÜ · Giresun Üniversitesi · Bilgisayar Mühendisliği

© Sercan Külcü. Tüm hakları saklıdır. · Sunumun etkileşimli HTML sürümü:
https://sercankulcu.github.io/files/ai_courses/sunumlar/Bolum_28_Bilgi_Erisimi_ve_Oneri_Sistemleri.html

GİRİŞ

# Ders Hangi Soruyu Sorar?

> **Kullanıcıya en ilgili belge ya da ürün milyonlarca aday içinden nasıl getirilir?**

Bilgi erişimi, bir sorguya karşılık milyonlarca belge arasından en ilgili olanları bulup **sıralayan** sistemleri inceler. Öneri sistemleri aynı işi sorgu olmadan yapar: kullanıcının geçmişinden neyi beğeneceğini kestirir. İkisinin de çıktısı bir yanıt değil, sıralı bir listedir.

## Bir bakışta

| YAKLAŞIM | GİRDİ | ÇIKTI |
| --- | --- | --- |
| eşleme ve sıralama | sorgu, belgeler, kullanıcı geçmişi | sıralı sonuç listesi |

| BAŞARI ÖLÇÜTÜ | MATEMATİK YÜKÜ | VERİ İHTİYACI |
| --- | --- | --- |
| kesinlik, duyarlılık, NDCG | orta: vektör uzayı | çok etkileşim verisi |

GİRİŞ

# Kısa Tarihçe

**1972**
Spärck Jones ters belge sıklığını (IDF) önerir: nadir terim daha çok şey söyler.

**1975**
Salton ve arkadaşları vektör uzayı modelini yayımlar: belge ve sorgu birer terim vektörüdür.

**1992**
Xerox PARC'ın Tapestry sistemi "işbirlikçi süzme" adını kullanır.

**1998**
Brin ve Page, PageRank'e dayanan Google arama motorunu tanıtır.

**2006**
Netflix Prize başlar; kazanan çözümlerde matris ayrıştırma öne çıkar.

**2020**
Yoğun vektörlerle erişim ve erişimle desteklenen üretim (RAG) yaygınlaşmaya başlar.

DERS

# Temel Kavramlar

**Ters dizin**
Her terimden onu içeren belgelere giden liste.

**TF-IDF ve BM25**
Terim sıklığı ve nadirliğiyle belge puanlamak.

**Vektör uzayı modeli**
Benzerliği kosinüs açısıyla ölçmek.

**Kesinlik, duyarlılık**
Getirilenlerin ne kadarı ilgili, ilgililerin ne kadarı geldi.

**Sıralama ölçütleri**
MAP, MRR, NDCG: sıradaki yer de önemlidir.

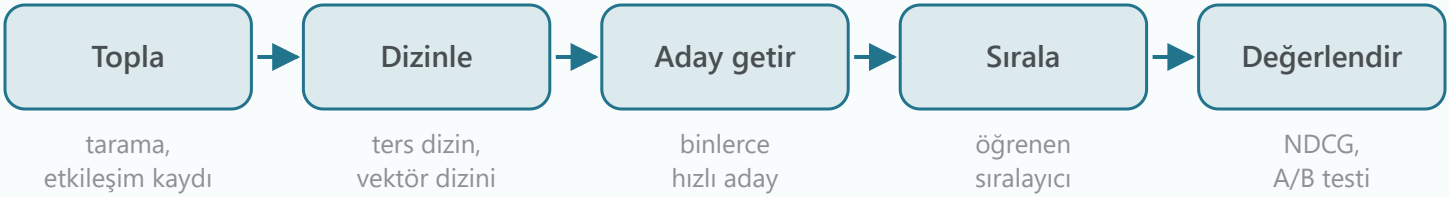
**İşbirlikçi süzme**
Benzer kullanıcıların beğenisinden öneri yapmak.

**Matris ayrıştırma**
Kullanıcı-ürün tablosunu gizli etkenlere bölmek.

**Soğuk başlangıç**
Geçmişi olmayan yeni kullanıcı ya da ürün sorunu.

DERS

# Tipik İş Akışı




**Örnek:** Bir video platformu izleme kayıtlarını toplar, videoları vektörlerle dizinler, milyonlarca videodan birkaç yüz aday getirir, bunları izlenme olasılığına göre sıralar ve yeni sıralayıcıyı A/B testiyle eskisiyle karşılaştırır.

DERS

# Derste Neler İşlenir?

1. Boole erişimi ve ters dizin
2. Dizin kurma, sıkıştırma ve sözlük yapıları
3. Metin ön işleme: belirteçleme, gövdeleme
4. Puanlama: TF-IDF ve vektör uzayı modeli
5. Olasılıksal erişim ve BM25
6. Erişimi değerlendirme: kesinlik, MAP, NDCG
7. Bağlantı analizi: PageRank ve HITS
8. Öğrenerek sıralama (learning to rank)
9. Yoğun vektörle erişim, yaklaşık komşu arama
10. İçerik tabanlı öneri
11. İşbirlikçi süzme ve matris ayrıştırma
12. Öneriyi değerlendirme: çevrim dışı ölçüm, A/B testi

## Kazandırdığı beceriler

- Elasticsearch ile arama dizini kurmak
- Pyserini ile BM25 deneyi
- FAISS ile vektör araması
- Surprise ile öneri modeli eğitmek
- NDCG ve MAP ile sıralama ölçmek

Tipik bir dönemin akışıdır; sıra ve kapsam üniversiteden üniversiteye değişir.

DENE

# Yardım Merkezinde Arama

Bir alışveriş sitesinin yardım merkezinde dokuz makale var; müşteri sorusunu kendi sözcükleriyle yazıyor. Hazır bir soru seçin ya da kendiniz yazın; makaleler TF-IDF ağırlıkları ve kosinüs benzerliğiyle sıralanır.

| Makale | kosinüs benzerliği |
|---|---|
| **1. Hasarlı ürün geldi** | 0,53 |
| 2. Siparişim nerede? | 0,10 |
| 3. Siparişimi nasıl iptal ederim? | 0,09 |
| 4. Kargo ücreti ne kadar? | 0,05 |
| 5. Fatura adresi değişir mi? | 0,05 |
| 6. İade nasıl yapılır? | 0,00 |
| 7. Para iadem ne zaman yatar? | 0,00 |
| 8. Kartla taksitli ödeme | 0,00 |
| 9. Şifremi unuttum | 0,00 |

siparişim hasarlı geldi

**hasarlı geldi** | kargo ücreti | para iadesi | şifre

**TF-IDF** | Yalnız sayım (TF)

Terim kökleri ve IDF: sipa **0,81**, hasa **2,20**, geld **2,20**
Eşleşen makale: **5** / 9 · 1. ile 2. arası fark: **0,43**

En iyi sonuç: **Hasarlı** ürünün fotoğrafını çekip 48 saat içinde yazın; değişim ya da iade yapılır.

IDF, neredeyse her makalede geçen “sipariş”, “kargo” gibi sözcüklerin ağırlığını düşürür; ayırt edici “hasarlı” öne çıkar. TF düğmesiyle farkı görün. Kökü ilk dört harfle almak Türkçe ekleri basitçe yutar.

KONUM

# Ön Koşullar

Bu dersten önce bilinmesi yararlı olan dersler ve neden gerektikleri.

| Ders | Neden gerekir |
|---|---|
| Doğrusal Cebir | vektör uzayı modeli, matris ayrıştırma |
| Veri Yapıları | karma tablosu, ağaçlar ve öncelik kuyruğu: dizin ve ilk k sonuç |
| Algoritmalar | sıralama, birleştirme ve karmaşıklık: büyük dizinlerde hız |

KONUM

# Bu Ders Sana Göre mi?

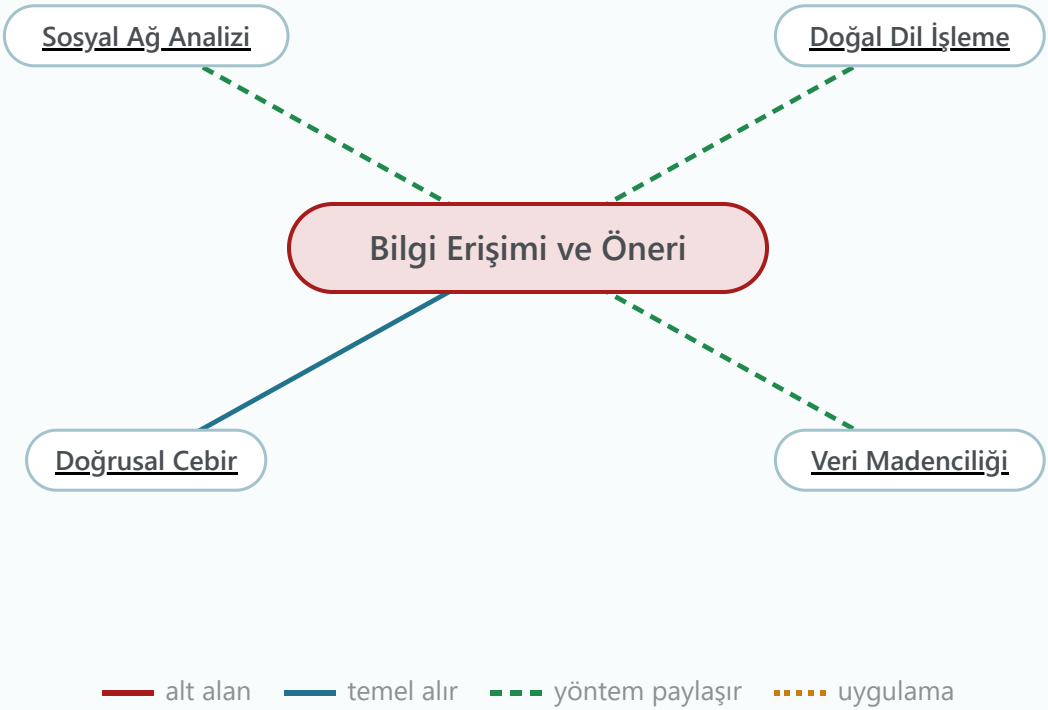
Kendine sor; "hayır" dediklerin için sağdaki sütun nereden başlayacağını gösterir.

| Kendine sor | Hayırsa önce |
|---|---|
| İki vektörün kosinüs benzerliğini hesaplayıp yorumlayabiliyor musun? | Doğrusal Cebir |
| Karma tablosunda aramanın neden sabit zaman aldığını açıklayabiliyor musun? | Veri Yapıları dersi: karma tabloları ve ağaçlar |
| Kesinlik ve duyarlılığı bir karışıklık matrisinden hesaplayabiliyor musun? | Makine Öğrenmesi |
| Bir metni sözcüklere bölüp sözcük sıklıklarını sayabiliyor musun? | Doğal Dil İşleme ve Metin Madenciliği |

Çoğuna "evet" diyebiliyorsan derse hazırsın; bir ya da iki "hayır" dönem içinde kapanabilir.

KONUM

# Dersin Komşuları



ÖRTÜŞME

# Örtüştüğü Dersler

| İlişki | Ders | Ortak konu |
|---|---|---|
| temel alır | Bu ders Doğrusal Cebir dersini temel alır | vektör uzayı modeli, matris ayrıştırma |
| yöntem paylaşır | Doğal Dil İşleme ve Metin Madenciliği ile yöntem paylaşır | metin temsili, TF-IDF, anlamsal arama |
| yöntem paylaşır | Veri Madenciliği ile yöntem paylaşır | birliktelik, işbirlikçi süzme |
| yöntem paylaşır | Sosyal Ağ Analizi ile yöntem paylaşır | PageRank, bağlantı çözümlemesi |

# Ayrıştığı Noktalar (1/2)

| Ölçüt | Bilgi Erişimi ve Öneri | Veri Madenciliği | Doğal Dil İşleme |
|---|---|---|---|
| **Temel soru** | Kullanıcıya en ilgili belge ya da ürün milyonlarca aday içinden nasıl getirilir? | Büyük veri yığınlarındaki gizli ve işe yarar örüntüler nasıl bulunur? | Bilgisayar insan dilini nasıl çözümler, anlar ve üretir? |
| **Yaklaşım** | eşleme ve sıralama | keşfe dayalı çözümleme | dilbilim, istatistik ve derin öğrenme |
| **Girdi** | sorgu, belgeler, kullanıcı geçmişi | büyük tablo ya da işlem verisi | metin |
| **Çıktı** | sıralı sonuç listesi | kural, küme, aykırı değer | etiket, çeviri, özet, yanıt |
| **Başarı ölçütü** | kesinlik, duyarlılık, NDCG | ilginçlik, destek, güven | doğruluk, BLEU, insan değerlendirmesi |
| **Matematik yükü** | orta: vektör uzayı | orta | orta |
| **Veri ihtiyacı** | çok etkileşim verisi | çok | çok metin (derlem) |

- **Veri Madenciliği:** Veri madenciliği veride örüntü keşfeder; bu ders keşfi bir kullanıcının o anki ihtiyacına göre sıralı listeye çevirir.
- **Doğal Dil İşleme:** NLP dilin yapısını ve anlamını çözümler; bu ders metni yalnızca ilgililik puanı için temsil eder ve sıralar.

# Ayrıştığı Noktalar (2/2)

| Ölçüt | Bilgi Erişimi ve Öneri | Makine Öğrenmesi | Sosyal Ağ Analizi |
|---|---|---|---|
| **Temel soru** | Kullanıcıya en ilgili belge ya da ürün milyonlarca aday içinden nasıl getirilir? | Bir program örneklerden nasıl genelleme yapar? | Kişiler ve ilişkilerden oluşan ağlar nasıl çözümlenir? |
| **Yaklaşım** | eşleme ve sıralama | istatistiksel öğrenme | çizge kuramı ve istatistik |
| **Girdi** | sorgu, belgeler, kullanıcı geçmişi | etiketli ya da etiketsiz örnekler | düğümler ve bağlantılar |
| **Çıktı** | sıralı sonuç listesi | tahmin yapan model | merkezilik, topluluk, yayılım |
| **Başarı ölçütü** | kesinlik, duyarlılık, NDCG | görülmemiş veride hata | modülerlik, yorumlanabilirlik |
| **Matematik yükü** | orta: vektör uzayı | yüksek: doğrusal cebir, olasılık, optimizasyon | orta: çizge, matris |
| **Veri ihtiyacı** | çok etkileşim verisi | orta ve çok | ağ verisi |

- **Makine Öğrenmesi:** MÖ tek tek örnekleri doğru tahmin etmeye çalışır; bu ders listenin sırasını, özellikle ilk birkaç sonucu önemser.
- **Sosyal Ağ Analizi:** Sosyal ağ analizi ilişkilerin yapısını inceler; bu ders o yapıyı (ör. PageRank) yalnız sıralama sinyali olarak kullanır.

UYGULAMA

# Uygulama Alanları

**Web arama**
Sorguya en ilgili sayfaları sıralamak.

**E-ticaret önerisi**
Bunu alanlar şunları da aldı.

**Video ve müzik**
Kişiye özel akış ve çalma listeleri.

**Kurumsal arama**
Şirket belgeleri ve e-postalarda arama.

**Haber akışı**
Gönderileri ilgiye göre sıralamak.

**RAG**
Dil modeline doğru belgeleri getirmek.

UYGULAMA

# Riskler ve Etik

**Filtre balonu**

Hep beğendiğinin benzerini göstermek görüş çeşitliliğini daraltır; haber ve video önerilerinde kutuplaşma etkisi tartışılır.

**Popülerlik yanlılığı**

Çok tıklanan ürün daha çok önerilir, böylece daha çok tıklanır; yeni ve küçük üreticiler görünmez kalır.

**Sorgu kayıtları ele verir**

AOL'ün 2006'da anonimleştirip yayımladığı arama kayıtlarından gazeteciler bazı kullanıcıların kimliğini buldu.

**Sıralamayı manipüle etmek**

Bağlantı çiftlikleri, sahte yorumlar ve anahtar sözcük doldurma sıralamayı bozar; sistem saldırgana karşı tasarlanmalıdır.

Ayrıntılı tartışma: Açıklanabilir Yapay Zekâ ve Yapay Zekâ Etiği.

UYGULAMA

# Araçlar ve Örnek Veriler

## Araçlar ve kütüphaneler

Apache Lucene | Elasticsearch | FAISS | Pyserini | Surprise | scikit-learn

## Tanınmış veri setleri

MovieLens | Netflix Prize | MS MARCO | TREC koleksiyonları | Cranfield koleksiyonu | Amazon ürün yorumları

UYGULAMA

# Sık Karıştırılanlar

| Yaygın düşünce | Doğrusu |
|---|---|
| Arama motoru sorgudaki sözcükleri bulmaktan ibarettir. | Asıl iş eşleşen binlerce belgeyi **sıralamaktır**; kullanıcı çoğu zaman yalnız ilk sayfaya bakar. |
| En yüksek puanlı ürünleri önermek yeterlidir. | Herkese aynı listeyi vermek kişiselleştirme değildir; çeşitlilik ve yenilik de ölçülür. |
| Doğruluk yüksekse öneri iyidir. | Puan tahmini hatası düşük bir model kötü sıralama yapabilir; NDCG gibi sıra ölçütleri gerekir. |
| Yoğun vektörler klasik aramayı tamamen bitirdi. | BM25 hâlâ güçlü bir taban çizgisidir; birçok sistem ikisini birlikte kullanır. |

SONRASI

# Bu Dersten Sonra

## Bu dersi temel alan ya da genişleten dersler

Bu ders çoğunlukla uç noktadır; öteki dersleri temel alır.

## İlgili meslekler

Arama mühendisi | Öneri sistemi mühendisi | Veri bilimci | Sıralama (ranking) araştırmacısı

Bu dersten sonra doğal dil işlemeyle anlamsal aramaya, sosyal ağ analiziyle bağlantı temelli sıralamaya, büyük veri sistemleriyle milyarlarca belgeyi dizinlemeye geçilebilir. Erişimle desteklenen üretim, bu dersi büyük dil modelleriyle doğrudan buluşturur.

SONRASI

# Kaynaklar


- Manning, Raghavan ve Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2008.
- Baeza-Yates ve Ribeiro-Neto, *Modern Information Retrieval*, 2. baskı, Addison-Wesley, 2011.
- Aggarwal, *Recommender Systems: The Textbook*, Springer, 2016.
- Koren, Bell ve Volinsky, "Matrix Factorization Techniques for Recommender Systems", *IEEE Computer* 42(8), 2009.

# Son

## Bölüm 28: Bilgi Erişimi ve Öneri Sistemleri

← Önceki bölüm · Alanın haritası · Sonraki bölüm →