Bilgi Erişim Sistemleri

1

1

Başlıklar

- Bilgi Erişim Sistemi(BES) Nedir? Amaçları Nelerdir?
- BES'in Geçmişi
- BES'in Genel Mimarisi
- BES'te Performans Değerlendirmesi
- Bilgi Erişim Modelleri
- İdeal BES ve Sorunları
- BES Geliştirme Aşamasında Karşılaşılan Güçlükler
- BES'te Araştırma Konuları
- Sonuç

2

2

Bilgi Erişim Sistemi(BES) Nedir? Amaçları Nelerdir?

- Bilgi Erişim bir bilgi ihtiyacını karşılamak amacıyla büyük koleksiyonlar içindeki düzenli yapıda olmayan materyalleri bulmaktır. (C.D. Manning, P. Raghavan,H. Schütze, 2009)
- BES'in amacı kullanıcılara aradıkları bilgiyi içeren materyalleri ilgililik sırasına göre en kısa zamanda sunmaktır.

3

3

Bilgi Artışı

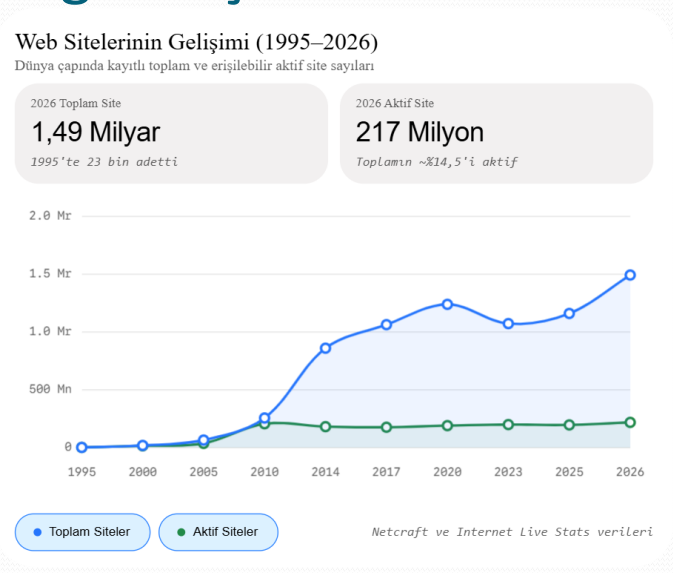
Dünya'da internet kullanımı İstatistikleri- 31 Aralık, 2011 (World Stats, 2012)				
Dünya Nüfusu (2011)	İnternet Kull. Ara. 31, 2000	İnternet Kull. Sayısı, 2011	% İnternet Kull. 2011	Büyüme 2000-2011
6,930,055,154	360,985,492	2,267,233,742	32.7 %	528.1 %

- Peter Lyman ve Hal Varian'ın 26 Mart 2001 sayılı The Economist'te yayınlanan araştırmasına göre:
- 5,4 Exabyte veri internet üzerinde yayınlanmış durumda (54 milyar *Economist* dergisinin içeriğine eşit)
- ABD'de her yıl 80 milyar fotoğraf çekiliyor
- 2 milyar röntgen filmi çekiliyor
- Günde 610 milyar e-posta mesajı gönderiliyor
- Her yıl 15 trilyon sayfa yazıcılardan çıktı alınıyor

4

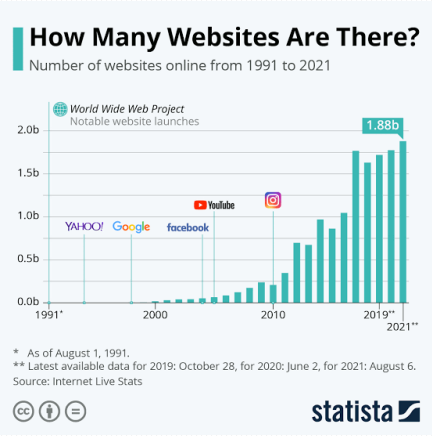
4

Bilgi Artışı



5

Bilgi Artışı



Verilerdeki Önemli Kırılma Noktaları

- Park Edilmiş Alan Adları ve Alt Alanlar: 2026 itibarıyla internette yanıt veren yaklaşık 1,5 milyar web sitesinin sadece ~%15'i (yaklaşık 217 milyon) özgün içerik barındıran aktif sitelerden oluşur. Geriye kalan %85'lik dilim park edilmiş alan adları, yönlendirmeler ve otomatik açılmış pasif alt alan adlarıdır.

Kayıtlı Alan Adı (TLD) ile Hostname Farkı: Verisign ve alan adı otoriteleri dünya genelinde kayıtlı toplam alan adını yaklaşık 401 milyon olarak raporlamaktadır. Netcraft'ın 1,5 milyara ulaşmasının sebebi, tek bir alan adı altında barındırılan yüz binlerce subdomain/hostname yapısını da anket kapsamında ayrı web sitesi olarak taramasıdır.

2020 Sonrası Temizlik ve Dalgalanmalar: Netcraft verilerinde 2017–2020 arasında görülen tepe noktalarından sonraki düşüşler, tarama algoritmalarında spam, bot veya otomatik oluşturulmuş sahte şablon sitelerin filtrelenmesinden kaynaklanmıştır. 2024–2026 döneminde ise bulut servis sağlayıcılarının (özellikle Cloudflare ve tersine proxy yapılandırmalarının) yaygınlaşmasıyla toplam tespit edilen host sayısı yeniden 1,5 milyar bandına yükselmiştir.

6

BES'in Geçmişi

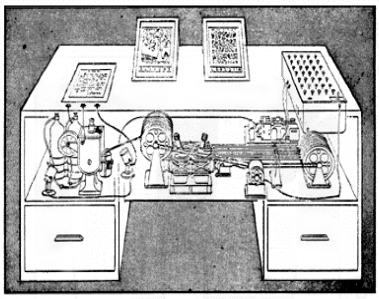
- **İçindekiler, kitap indeksi**
- **Dewey Onlu Sınıflama - 1876:** Melvil Dewey tarafından geliştirilen ve kütüphanelerdeki materyalleri konularına göre 10 temel sayısal kategoriye ve alt dallara ayıran dünyaca ünlü sınıflama sistemi.
- **Kart Katalogları – 18/19. yy:** Dijital veri tabanlarından önce kütüphanelerde kitapların yazar adı, eser adı veya konusuna göre kartlar üzerine yazılıp alfabetik olarak dizildiği geleneksel arama/kataloglama sistemidir.
- **George Boole 1847 :** Kataloglar ve kütüphane bilgi erişim sistemlerinde Boole mantığı, aranan materyallere (kitap, makale, tez, dergi vb.) kesin, filtrelenmiş ve mantıksal sınırlar çizerek ulaşmanın temel mekanizmasıdır.Geleneksel kart kataloglarından OPAC (Online Public Access Catalog) gibi modern dijital kütüphane sistemlerine geçildiğinde, araştırmacılar bu üç ana mantıksal bağlaçla sorgularını yapılandırır.
- **Claude Shannon 1937:** Sayısal Devrelerin Boole ile ifade edilmesi ve bilgisayarın doğuşu (20.yy'ın en önemli master tezi)

7

7

BES'in Geçmişi

- Vannevar Bush, 1946 : Memex, ilk Bilgi Erişim Sistemi tasarımı



Memex in the form of a desk would instantly bring files and material on any subject to the operator's fingertips. Slanting translucent viewing screens magnify supermicrofilm filed by code numbers. At left is a mechanism which automatically photographs longhand notes, pictures and letters, then files them in the desk for future reference (*LIFE* 19(11), p. 123).

Memex Nedir? "Memory" ve "Index" (veya "Extender") kelimelerinin birleşiminden türetilen Memex; bireylerin tüm kitaplarını, kayıtlarını, makalelerini ve iletişimlerini mikrofilm tabanlı bir masa/cihaz içinde saklayıp hızla taramasına imkan tanıyan kuramsal bir elektromekanik masaüstü cihazıdır.

8

8

BES'in Geçmişi-2

Coordinate Indexing Nedir?

1950'lerin başında **Morten Taube** tarafından geliştirilen (*Uniterm* sistemi olarak da bilinir), belgeleri önceden tanımlanmış katı ve hiyerarşik konu başlıklarına (Dewey veya Kongre Kütüphanesi sınıflaması gibi) hapsetmek yerine, **bağımsız anahtar terimlerin (tanımlayıcıların) serbestçe bir araya getirilmesiyle** arama yapmayı sağlayan indeksleme yöntemidir.

1. Çalışma Mantığı

İndeksleme Aşaması: Bir belge sisteme kaydedilirken hiyerarşik tek bir sınıfa sokulmaz; belgeyi niteleyen bağımsız kavramlar (anahtar kelimeler/koordinatlar) atanır (örn. *Bilgisayar*, *Tıp*, *Görüntü İşleme*).

Arama Aşaması: Araştırmacı arama yaparken bu bağımsız terimleri Boole mantığı (**VE / AND**) ile birleştirir (koordine eder).

2. Önceden Koordineli (Pre-coordinate) vs. Sonradan Koordineli (Post-coordinate)

Pre-coordinate (Geleneksel): Kavramların ilişkisi indeksleyici tarafından en başta, kataloglama anında kurulur (örn. kütüphane kartındaki sabit başlık: *"Tıbbi Bilişim - Görüntüleme Yöntemleri"*). Kullanıcı bu kalıba bağımlıdır.

Post-coordinate (Coordinate Indexing): Terimler katalogda ayrı ayrı saklanır; ilişkilendirme (koordinasyon) **arama anında kullanıcı tarafından** kurulur.

3. Bilişim ve Bilgi Erişim Açısından Önemi

Ters Çevrilmiş Dizin (Inverted Index): Her bir anahtar terimin altına o terimin geçtiği belge numaralarının yazıldığı kart sistemi (*Uniterm kartları*), modern arama motorlarının temelini oluşturan Ters dizin (inverted index) mimarisinin doğrudan öncüsüdür.

Delikli Kartlar ve Peek-a-boo (Batten) Sistemi: Belgeleri bulmak için ilgili terimlerin delikli kartları üst üste konup ışığa tutulurdu; ışığın geçtiği ortak delikler, her iki terimi de içeren belge numaralarını (yani Boole AND kesişimini) optik/mekanik olarak anında gösterirdi.

Modern arama motorlarındaki çoklu anahtar kelime sorgularının ve veritabanı filtreleme mantığının teorik kökeni bu yönteme dayanır.

9

BES'in Geçmişi-2

- Taube 1950: "coordinate indexing"

EMPRESYONISM				15897					
10	511	22	13	54*	65	126	67	18	39
80	701	52	163	114	345	816	357	58	769
550	961	62*	983	364*			627	88*	
630		92		654				148	
				794					

PISSARO				53971					
80	141	12	93	54*	25	16	47	38	19
110	191	62*	143	74		36	417	88*	239
710			253	364*			537		569
				524					

10

BES'in Geçmişi-3

W.E. Batten-1948

Çalışma Prensibi

Terim Başına Kart: Geleneksel sistemlerde her belge için bir kart açılırken, Batten sisteminde her bir **kavram/anahtar terim** için ayrı bir delikli kart (grid şablonu) oluşturulmuştur.

Koordinat ve Delikler: Kartın üzerindeki her bir koordinat pozisyonu (örneğin 100x100'lük bir ızgarada 1'den 10.000'e kadar olan numaralar) belirli bir belgeyi/patenti temsil eder. Bir belge o terimle ilişkiliyse, kartın ilgili koordinatına bir delik açılır.

Optik Kesişim (Boole AND İşlemi): Birden fazla kavramı aynı anda içeren belgeleri bulmak için ilgili terim kartları üst üste konur ve bir ışık kaynağına tutulur.

Işığın doğrudan geçtiği delikler, **her iki (veya daha fazla) kartta da ortak olan** pozisyonları gösterir.

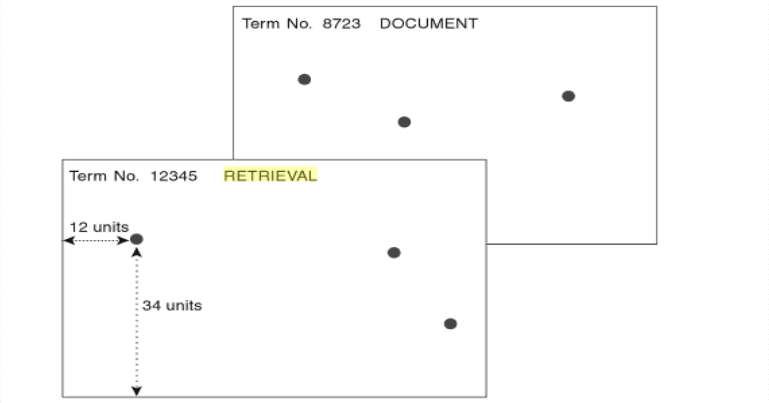
Böylece bilgisayarların henüz yaygın olmadığı bir dönemde, Boole **VE (AND)** kesişim kümesi saniyeler içinde optik olarak tespit edilmiştir.

11

11

BES'in Geçmişi-3

- W.E. Batten 1948 :



12

12

BES'in Geçmişi-4

Calvin Mooers ve Zatocoding Sistemi

Amerikalı bilgisayar bilimci **Calvin Mooers**, literatüre bizzat "**Information Retrieval**" **(Bilgi Erişim)** terimini kazandıran (1950) ve bu alandaki ilk mekanik arama mimarilerini tasarlayan öncüdür. 1940'ların sonu ile 1951 civarında kurduğu Zator Company bünyesinde geliştirdiği **Zatocoding** (Zatocode), bilgi erişim tarihinde devrim niteliğinde bir indeksleme ve sıkıştırma yöntemidir.

Çalışma Mantığı:

Üst Üste Binen Kodlama (Superimposed Coding)

Kenardan Delikli Kartlar (Edge-Notched Cards): Klasik kartların aksine, belgelerin özetleri veya referansları kartın ortasında yer alır; kartın kenarlarında ise numaralandırılmış delikler dizisi bulunur.

Rastgele / Karma Kodlama (Random Superimposed Hash): Her bir konu veya anahtar terime, tek bir delik yerine rastgele seçilmiş sabit sayıda bir delik kombinasyonu (örneğin 40 delikli bir kenarda rastgele 4 delik konumu) atanır.

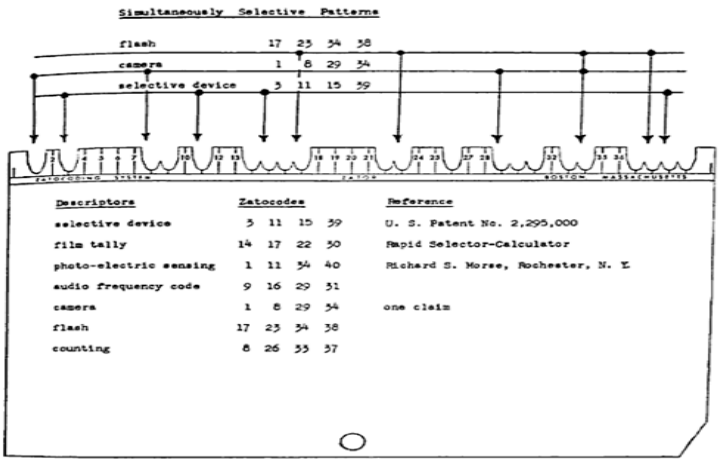
Üst Üste Kod Çentikleme: Bir belge birden fazla konuyu içeriyorsa, bu terimlerin kodları aynı kartın kenarına üst üste çentiklenir (açılır). Farklı konuların kodları aynı kartta örtüşür.

13

13

BES'in Geçmişi-4

- Calvin Mooers 1951: Zatocode



14

14

BES'in Geçmişi-5

The Western Reserve Rapid Searching Selector

1950'li yılların ortalarında (özellikle 1956 civarında) **Western Reserve University** (günümüzdeki adıyla *Case Western Reserve University*) bünyesindeki **Center for Documentation and Communication Research (CDCR)** merkezinde **James W. Perry** ve **Allen Kent** tarafından tasarlanmış ve inşa edilmiş öncü bir bilgi arama makinesidir.

Çalışma Mantığı ve Mimarisi

Delikli Kağıt Şeritler: Belgelerin bibliyografik verileri ve anlamsal kodlamaları delikli kağıt teypler (şeritler) üzerinde sıralı olarak kaydedilirdi.

Elektronik Röle ve Tarama: Sistem, yüksek hızda kayan şeritleri okumak için elektro-mekanik röleler ve vakum tüplü elektronik devreler kullanırdı.

Özel Mantıksal Kodlama (Telegraphic Abstracts): Belgeler sadece serbest anahtar kelimelerle değil, Perry ve Kent'in geliştirdiği "telgrafik özet" (telegraphic abstract) adı verilen özel bir semantik/dilbilimsel kodlama yapısıyla indekslenirdi. Bu sayede makine, kavramlar arasındaki rolleri ve mantıksal ilişkileri ayırt edebiliyordu.

15

15

BES'in Geçmişi-5

Bilgi Erişim Tarihindeki Yeri

Karmaşık Boole Sorgularını Donanımla Çözme: Kullanıcıların girdiği çok adımlı Boole sorgularını (VE, VEYA, DEĞİL ve parantezli kombinasyonlar) delikli şeritler üzerinde donanımsal düzeyde arayabilen ilk tam işlevsel sistemlerden biridir.

İlk Bilgisayarlı Bilgi Erişim Ağı: Bu makineyle edinilen tecrübe, daha sonra Amerikan Metal Derneği (ASM - American Society for Metals) için dünyanın ilk otomatik literatür tarama servisinin ve ardından GE-225 gibi genel amaçlı ana bilgisayarlar üzerinde çalışan bilgi erişim sistemlerinin kurulmasını sağlamıştır.

16

16

BES'in Geçmişi-6

- 1957 yılında Bracken ve Tillit: Dokümanların bilgileri indekslenmiş bir şekilde bilgisayarda tutuluyor, sorgular bir manyetik teybe kaydedildikten sonra ve indekslenip sıralanıyor. Dokümanlar ve sorgular bool işlemlerinden geçirilerek karşılaştırılıyordu.

- 1958'de Luhn ve Bar Hillel arama işlemlerinin daha isabetli yapılabilmesi için basit kelime karşılaştırması dışında kelimelerin frekanslarını ve kelimelerin tüm dokümanlardaki kullanılma sıklığı parametrelerini kullanmaya başladılar. (tf, idf)

17

17

BES'in Geçmişi-7

- 1961 yılında Chemical Abstracts Service adlı kuruluş 600 önemli dergiyi kapsayan "Chemical Titles" adını verdiği konu indeksini bilgisayar vasıtasıyla üretti ve bastı.
- Bilgisayarların hızlanması ve rastgele erişimli dosya yapısı sayesinde 1963 yılında MIT'de MAC projesi ile interaktif bilgi erişim sistemi hayata geçti.
- 1972'de LEXIS/NEXIS hukuk ve gazete erişim sisteminin temelleri atıldı. Böylece ilk defa bir erişim sistemi ile kaynakların tam metinleri üzerinde arama yapılmaya başlandı.

18

18

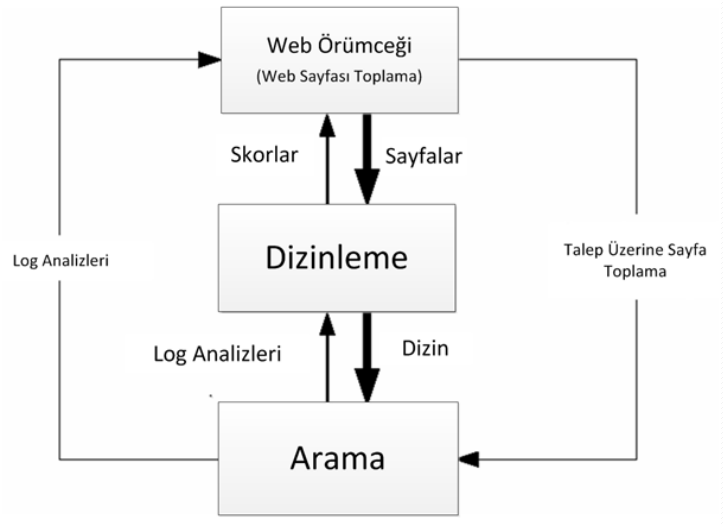
BES'in Geçmişi-8

- 1969'da DARPA, **ARPANET** adlı 4 sunucunun bağlı olduğu ilk ağ yapısını kurmuştur.
- R.E. Kahn 1972'de paket veri paylaşımı için **TCP**(Transmission Control Program) adlı basit bir yapı tasarladı. Bu yapı günümüzdeki internetin ilk haliydi.
- Cenevre'de Cern araştırma merkezinde fizikçi Tim Berners- Lee 1991'de hiper metin modelinin kullanarak **WWW**(World Wide Web)'i yarattı.

19

19

BES'in Genel Mimarisi



20

20

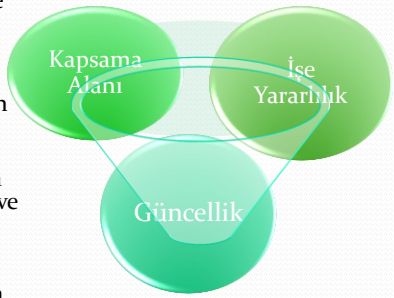
Web Örümceği

İnternetteki web sayfalarını keşfetmeyi ve toplamayı amaçlar.

Temel Görevi ve Amacı

Web örümceğinin birincil görevi, **arama motorları** (Google, Bing, Yandex vb.) için internetin güncel haritasını çıkarmak ve içerikleri toplamaktır.

Toplanan bu sayfalar, arama motorlarının dev veri merkezlerinde işlenir; kelimeler ve anlamsal yapılar çözümlenerek **ters çevrilmiş dizin (inverted index)** yapılarına kaydedilir. Bir kullanıcı arama motorunda arama yaptığında aslında o an tüm interneti değil, örümceklerin önceden toplayıp indekslediği bu veritabanını sorgular.

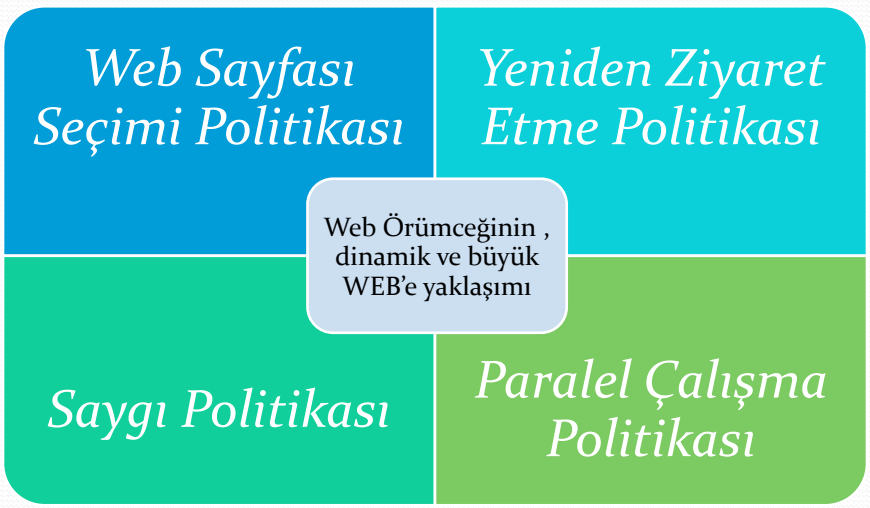


Toplanan Sayfaların Kalitesi

21

21

Web Örümceği-2



22

22

Dizinleme

- Bir Bilgi Erişim Sisteminde derlemdeki belgeler gerektiğinde hızlı bir biçimde erişilebilmesi için dizinlenerek (indeksleme) saklanır.
- Bir sorgu sözcüğü bütün derlemde baştan sona aranmaz, dizine bakılır ve hangi dokümanlarda hangi ağırlıkta geçtiği bilgisine kolayca erişilerek aramanın makul sürelerde gerçekleşmesi mümkün olur.

Dizinleme-2

- Dizin yapısı iki tablodan oluşur. Birinci tabloda derlemdeki tüm sözcükler ve bu sözcüklerin kaç dokümanda ,toplam kaç defa kullanıldığı bilgisi saklanır.
- İkinci tabloda ise sözcüğün hangi dokümanlarda kaç defa bulunduğu bilgisi ilk tablo ile ilişkili biçimde saklanır.

Arama

- İhtiyaç duyulan bilgiyi içeren kaynaklara ulaşmak için belli bir yöntem izlenerek sorgu ve kaynakların değerlendirilmesi ve uygun kaynakların kullanıcıya sunulması sürecidir.

25

25

Performans Değerlendirmesi

- **Anma (Recall)** sorgu sonucu erişilen ilgili dokümanların koleksiyondaki tüm ilgili dokümanlara oranıdır. (Saracevic, 1995)
- Anma tamlığı, bütünlüğü (completeness) gösterir. (Fensel, 2007)
- Erişilen dokümanlar kadar erişilemeyenlerle de ilgilidir.

$$Anma\ (Recall) = \frac{Erişilen\ ilgili\ dokümanların\ sayısı}{koleksiyondaki\ tüm\ ilgili\ dokümanların\ sayısı}$$

26

26

Performans Değerlendirmesi -2

- **Duyarlılık(Precision)** sorgu sonucu erişilen ilgili dokümanların , erişilen tüm dokümanlara oranıdır.
- Erişilen dokümanların kaç tanesinin sorgu ile ilişkili olduğu bilgisini verir. Sistemin "Doğruluğunu" ifade eder. (Fensel, 2007)

$$Duyarlılık\ (Precision) = \frac{Erişilen\ ilgili\ dokümanların\ sayısı}{Erişilen\ tüm\ dokümanların\ sayısı}$$

27

27

Performans Değerlendirmesi -3

- **Kullanışlılık** – Sorgu ve sonuç gösterme sayfası kullanıcı dostu olmalıdır.
- **Hız** – Arama sonuçlarının hızlı gelmesi özellikle bazı alanlarda zaruri olabilir. (sağlık, güvenlik vs...)

28

28

Bilgi Erişim Modelleri

- Vektör Uzayı Modeli
- Boole Modeli

29

29

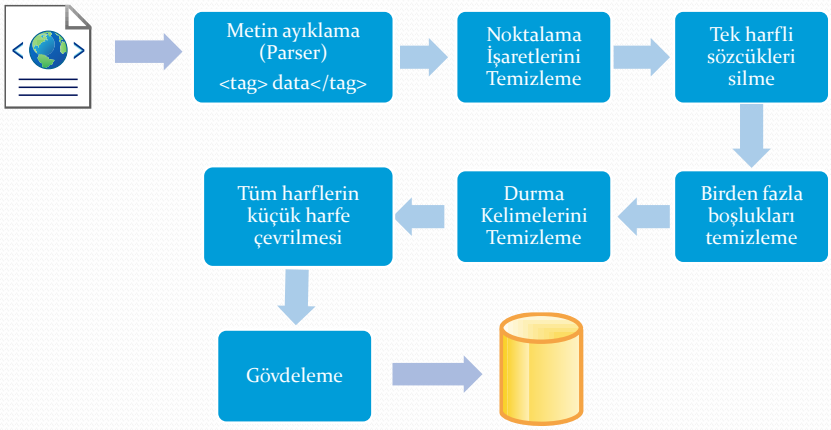
Vektör Uzayı Modeli (VSM)

- Bir doküman setindeki dokümanların ortak bir vektör uzayında vektörel olarak ifade edilmesine Vektör Uzayı Modeli denir.
- Sorgu ile dokümanların benzerliklerinin bulunabilmesi için, dokümanlara ait vektörler ile sorgu vektörü, vektör iç çarpımı kullanılarak benzerlik hesaplaması yapılır. (C.D. Manning, P. Raghavan,H. Schütze, 2009)

30

30

VSM – Belge İçeriğinin Arındırılması



31

31

Türkçe, İngilizce Durma Kelimeleri

ve	bir	bu	da	eğer
de	için	ile	olarak	kabul
daha	çok	en	gibi	artık
sonra	olan	ama	kadar	tam
ne	ise	iki	var	çünkü
büyük	yeni	her	o	bütün
ilk	son	ancak	değil	iş
dedi	diye	olduğunu	ki	mı
önce	yüzde	olduğu	göre	diğer
içinde	bin	yok	iyi	bunu
milyon	zaman	karşı	söyledi	milyar
arasında	ya	ilgili	gün	nasıl
oldu	tarafından	yer	kendi	başka
önemli	hem	bile	mi	eden
aynı	dün	ortaya	gelen	geçen
tek	böyle	biz	ben	yine
kez	etti	sadece	siyasi	şimdi
bunun	birlikte	özel	konusunda	şu
nedeniyle	hiç	şöyle	üzerine	eski
yaptığı	bugün	yapılan	devam	bazı
tüm	pek	eğer		

I	in		
a	is		
about	it		who
an	of		will
are	on		with
as	or		the
at	that		www
be	the		
by	this		
com	to		
for	was		
from	what		
how	when		
	where		

32

32

Gövdeleme

- Gövdeleme kelimeleri çekim eklerinden ayırma, yalın hallerine döndürme işlemidir.
- Kelimelerin ek alarak farklılaşması sorgular ve belgelerin kesişme olasılığını etkilemektedir.
- Ör1. kalitesindeki -> kalite
- kaliteden->kalite
- Ör2. buldular -> bul
- bulacağı->bul

33

33

Vektör Oluşturma

- **Temel Vektör:** *t1, t2, t3…tn*
  - n adet kelimeden oluşur,
  - Derlemdeki tüm kelimeleri kapsamalıdır.
- **Doküman/Sorgu Vektörleri:** *f1, f2, f3…fn*
  - sözlükte bulunan n adet tekil kelimenin doküman/sorgu içerisindeki sıklık değeri(frekansı) vektör katsayısı olarak kullanılır (“terim ağırlıklı gösterim”)

(C.D. Manning, P. Raghavan,H. Schütze, 2009)

34

34

VSM - Örnek

- **D_1**: *"Küçük alanda domates yetiştirme teknikleri" bloğu yeri sınırlı olanlar için kullanışlı bir blog. Domatesten verim almak için bu blogdaki dikim ve bakım tekniklerinin dikkatle uygulanması gerekiyor.*
- **D_2**: *Domates yetiştirme teknikleri domatesin türüne göre farklılık göstermektedir. Salkım domatesler saksı gibi küçük alanlarda bile çok az sulama ile yetiştirilirken salçalık büyük domatesler derin ve geniş toprağa ihtiyaç duyuyor.*
- **D_3**: *Saksıda salkım domates yetiştirdim, bütün yaz çok az su ile inanılmaz verim aldım.*
- **D_4**: *Salçalık domates mart ayında zam şampiyonu oldu, fiyatı %30 arttı. Domates yetiştiricileri durumdan memnun.*
- **D_5**: *Pazarda satılan sözde organik domateste yüksek oranda zirai ilaç tespit edildi. Uzmanlar sebzelerin mevsiminde tüketilmesinde fayda olduğunu söyledi.*
- **D_6**: *Yunanistan'da halk başbakanı domates atarak protesto etti.*
- **Sorgu**: *Küçük alanda domates yetiştirme teknikleri*

Sözlük	D1	D2	D3	D4	D5	Sorgu
Küçük	1	1	0	0	0	1
alan	1	1	0	0	0	1
yetiştir	1	2	0	1	0	1
teknik	2	1	0	0	0	1
domates	2	3	1	1	1	1
yer	1	0	0	0	0	0
sınır	1	0	0	0	0	0
olan	1	0	0	0	0	0
kullanışlı	1	0	0	0	0	0
blog	3	0	0	0	0	0
verim	1	0	1	0	0	0
almak	1	0	1	0	0	0
dikim	1	0	0	0	0	0
bakım	1	0	0	0	0	0
uygula	1	0	0	0	0	0
gerek	1	0	0	0	0	0
tür	0	1	0	0	0	0
farklılık	0	1	0	0	0	0
göstermek	0	1	0	0	0	0
saksı	0	1	1	0	0	0
derin	0	1	0	0	0	0
geniş	0	1	0	0	0	0
toprak	0	1	0	0	0	0
ihtiyaç	0	1	0	0	0	0
duy	0	1	0	0	0	0
bütün	0	0	1	0	0	0
yaz	0	0	1	0	0	0
su	0	1	1	0	0	0
inanılmaz	0	0	1	0	0	0
salçalık	0	1	0	1	0	0
mart	0	0	0	1	0	0
ay	0	0	0	1	0	0
zam	0	0	0	1	0	0
şampiyon	0	0	0	1	0	0
ol	0	0	0	1	1	0
fiyat	0	0	0	1	0	0
art	0	0	0	1	0	0
durum	0	0	0	1	0	0
memnun	0	0	0	1	0	0
Pazar	0	0	0	0	1	0
sat	0	0	0	0	1	0
söz	0	0	0	0	1	0
organik	0	0	0	0	1	0
yüksek	0	0	0	0	1	0
oran	0	0	0	0	1	0
zirai	0	0	0	0	1	0

Benzerlik hesaplaması

- Sorgu ve haberin benzerliği, vektörlerinin birbirleriyle yaptıkları açıyla ters orantılıdır.
- İki vektör arasındaki açı ne kadar küçükse iki vektör o kadar benzerdir.

$$Cos(\Theta) = \frac{\vec{A}.\vec{B}}{\|A\|.\|B\|} \Rightarrow sim(a,b) = \frac{\sum_{w=1}^{n} tf_a(w).tf_b(w).idf(w)}{\sqrt{\sum_{w=1}^{n} tf_a^2 . \sum_{w=1}^{n} tf_b^2}}$$

Benzerlik hesaplaması - tf

$$sim(a,b) = \frac{\sum_{i=1}^{n} tf_a(i).tf_b(i).idf(i)}{\sqrt{\sum_{i=1}^{n} tf_a^2 . \sum_{i=1}^{n} tf_b^2}}$$

- $Tf_a(i)$ *(term frequency):* temel vektörde i. sıradaki kelimenin a dokümanı içindeki frekansı/kullanım sıklığını gösterir.
- Bir kelime bir dokümanda sık geçiyorsa kelimenin o belgenin konusunu ve içeriğini ifade etmede önemli olma ihtimali yüksektir.

37

Benzerlik hesaplaması - idf

- ***Idf* (i)**(inverse document frequency) : *temel vektörde i. sıradaki kelimenin* tüm dokümanlar içindeki frekansını ifade eder.
- Bir kelime derlemdeki dokümanların çoğunluğunda çok fazla bulunuyorsa belgeleri birbirinden ayırt etme olasılığı o kadar düşük demektir.

$$Idf(i) = \log_{10}\left\{\frac{N}{df(i)}\right\}$$

  - **N:** toplam doküman sayısı
  - **df(i):** *temel vektörde i. sıradaki kelimenin* tüm dokümanlarda bulunma sıklığı

38

Eşik Değeri (Threshold)

- Erişim işlemleri sonucunda bulunan belgelerin benzerlik değerlerinden belli bir skor değerinin altında kalanlar ilgisiz kabul edilir.
- İlgili / ilgisiz ayrımının yapılmasını sağlayan bu skor değeri **eşik değer** olarak kabul edilir.
- Eşik değerleri seçilirken küçük bir test doküman seti üzerinde deneme yapılır.
- Bu deneme sonucunda anma ve duyarlık değerlerinin maksimum olduğu noktadaki skor değeri Eşik Değeri olarak alınır.

39

39

Vektör Uzayı Modelinin Avantajları

- Sorgu sonucu derecelendirilebilmektedir . Böylece dokümanlar benzerlik değerine göre sıralanarak kullanıcıya sunulabiliyor.
- Terim ağırlıklandırma, sorgu sonucu döndürülen belgelerin kalitesini artırmaktadır.

40

40

Vektör Uzayı Modelinin Dezavantajları

- Terimleri birbirlerinden bağımsız görmesi nedeniyle belgedeki kelimeler arası ilişkiler göz ardı ediliyor.
- Farklı terimlerle ifade edilmiş benzer konuları içeren dokümanları bulmak zor.

41

41

Boole Modeli

- Bu modelde de Vektör Uzayı Modeli'nde olduğu gibi dokümanlar vektörel olarak ifade edilir fakat vektör katsayıları 1 veya 0 olarak ifade edilir.
- Bir kelimenin bir doküman içerisinde birden fazla geçmesi önem taşımamaktadır.
- Pek çok belgede geçen kelimeler, diğerleri ile aynı ağırlığa sahiptir.

42

42

Boole Modeli - Örnek

- ***D1**: Gezgin iletişim sistemleri*
- **D2**: Kablosuz gezgin iletişim ağları
- **D3**: Geniş alanlı kablosuz veri ağları
- **D4**: Veritabanı yönetim sistemleri
- **D5**: Bilgi teknolojileri

Sözlük Vektörü	D1	D2	D3	D4	D5
kablosuz	0	1	1	0	0
gezgin	1	1	0	0	0
iletişim	1	1	0	0	0
sistemleri	1	0	0	1	0
veri	0	0	1	0	0
ağları	0	1	1	0	0
bilgi	0	0	0	0	0
teknolojileri	0	0	0	0	1
geniş	0	0	1	0	1
alanlı	0	0	1	0	0
veritabanı	0	0	0	1	0
yönetim	0	0	0	1	0

43

43

Boole Modeli

- Boole Modeli Küme teorisi üzerine kuruludur. Sorgu kelimeleri ile doküman terimlerinin kesişmesi mantığı üzerinde çalışır.
- Kullanıcı sorgu kelimeleri arasına **AND, OR, NOT** mantıksal operatörleri yerleştirir.
- Ör. bilgi VE erişim
- Sorgu ve doküman vektörleri kullanıcının girdiği operatörlere göre mantıksal işleme sokulur. Sonuçta TRUE çıkıyorsa doküman kullanıcıya sunulur, FALSE çıkarsa sunulmaz.
- İşlem Sırası: Parantez içi -> AND -> OR

44

44

Boole Modeli - Örnek

- ***Sorgu**: kablosuz VE gezgin VE iletişim*
  ***Sonuç**: D2*
- ***Sorgu**: (kablosuz VEYA gezgin) VE iletişim*
  ***Sonuç**: D1, D2*
- ***Sorgu**: (kablosuz VEYA gezgin) VE (iletişim VEYA ağları VEYA teknolojileri VEYA sistemleri)*
  ***Sonuç**: D1,D2, D3*

Sözlük Vektörü	D1	D2	D3	D4	D5
kablosuz	0	1	1	0	0
gezgin	1	1	0	0	0
iletişim	1	1	0	0	0
sistemleri	1	0	0	1	0
veri	0	0	1	0	0
ağları	0	1	1	0	0
bilgi	0	0	0	0	0
teknolojileri	0	0	0	0	1
geniş	0	0	1	0	1
alanlı	0	0	1	0	0
veritabanı	0	0	0	1	0
yönetim	0	0	0	1	0

45

45

Boole Modeli - Sorunlar

- **Benzerlik derecelendirme yok:** sorgu ile belgenin ne kadar ilgili olduğunu gösteren bir skor değeri üretilmiyor bunun yerine ilgili / ilgisiz (1/0) yargısı var.
- **Terim ağırlıklandırma yok:** bir belge içerisinde 1 kez geçen terimle 100 kez geçen terim aynı ağırlıkta, bu belge içeriğini ifade eden değerli terimleri seçmemizi engelliyor.
- **Sorgu oluşturmak zor:** kullanıcılar bilgi ihtiyaçlarını ifade edebilmek için karmaşık sorgular kurabilmesi gerekiyor.
- **Hata toleransı yok:** bilgi ihtiyacı tam ifade edilmek zorunda, yaklaşık sonuç alma olasılığı yok.

46

46

İdeal Bilgi Erişim Sistemi

- Kullanıcının aradığı konu ile ilişkili materyalleri döndürmelidir.
- Kullanıcının aradığı konu ile ilişkili tüm materyalleri döndürmelidir.
- Sonuçlar güncel olmalıdır.
- Sonuçları hızlı döndürmelidir.
- Ara yüzü kullanışlı olmalıdır. (M. Kobayashi, K. Takeda, 2000)

47

47

İdeal Bilgi Erişim Sistemi Neden Mümkün Değil?

- Sorgu Sorunları: ihtiyacı ifade etme sıkıntısı
- Dil Sorunları: gövdeleme zor
- Konu Belirleme Sorunları: metadata tanımlarında eksiklikler ve kötüye kullanma sorunları
- WEB'in büyüklüğü ve dinamikliği: Netcraft'ın ulaşabildiği web sitesi sayısı Mart 2012 itibariyle 646 milyona ulaşmıştır ve bu sayı her ay yaklaşık %5 oranında artmaktadır.
- Web kaynaklarındaki yazım hataları: içerik kontrolü olmaması nedeniyle kalite düşük.

48

48

Bilgi Erişim Sistemi Geliştirme Aşamasında Karşılaşılan Güçlükler

- Donanımsal Altyapı Gereksinimleri: Büyük veri yığınlarını saklama/erişme problemi
- İnsan Kaynağı Gereksinimi: testler ve konu belirleme işlemleri
- Bilimsel Araştırmaların Paylaşılmaması Sorunu: gövdeleme yazılımları, sözlükler...

49

49

Bilgi Erişim Sistemlerinde Araştırma Konuları - Birleştirilmiş Arama Motorları (Meta Search Engines)

- Kullanıcıyı birden fazla arama motorunda arama yapma külfetinden kurtarır.
- Kullanıcı farklı arama motorlarının birbirinden farklı sorgu dillerini öğrenmek zorunda kalmaz.
- Sistem arama motorlarının en yüksek ilişkili sonuçlarını kullanıcıya en üstte sunduğu için kullanıcının aradığını kolayca bulma olasılığı yükselmektedir. (Meng, August 2009)
- AllInOneNews, Search aggregator, Travelfox

50

50

Bilgi Erişim Sistemlerinde Araştırma Konuları - Kişiselleştirilmiş arama (personalized search)

- Kullanıcının internetteki aktivitelerinin, bilgi tüketim alışkanlıklarının izlen kullanıcı hakkında edinilen bilgilerin bilgi erişim sistemlerinde kullanılması ve kullanıcıya anlamlı sonuçlar döndürülmesi olarak tanımlanır. (J. Pitkow, H. Schütze, T. Cass, R. Cooley, D. Turnbull, A. Edmonds, E. Adar, T. Breuel, 2002)
- Özel hayatın gizliliği ve kullanıcıların zamanla ilgi alanlarının değişmesi konularına dikkat edilmelidir.

51

51

Bilgi Erişim Sistemlerinde Araştırma Konuları - Anlamsal arama (semantic search)

- Anlamsal aramada sorgu sözcüklerinin dokümanda bulunma sıklığı değil kullanıcının gerçekte neyi aradığı ve web içeriklerinin konusu /anlamı önemlidir.
- Sorgu sözcükleriyle eş anlamlı sözcükler de aramaya dahil edilir.
- Sözcüklerin tüm varyasyonları aramaya dahil edilir: improve, improved, improvement
- Sadece anahtar kelime değil konu ile ilgili diğer kelimelerin de bilinmesi gerekir.(ontological knowledge) Ör: üst solunum yolu enfeksiyonu ->(nezle, grip, sinüzit, farenjit)

52

52

Bilgi Erişim Sistemlerinde Araştırma Konuları - Anlamsal arama (semantic search) - Sorunlar

- Anlamsal arama için yapılması gereken ön çalışma (ontoloji oluşturma) çok uzun zaman almaktadır.
- Doğal dil sorgularını makinenin anlayabileceği ontolojik sorgulara dönüştürmek henüz mümkün olmamıştır.
- Web içeriğinde konunun, anlamın bulunması da çözülmeyi bekleyen sorunlar arasındadır.
- Hakia, DBpedia

53

53

Bilgi Erişim Sistemlerinde Araştırma Konuları - Soru yanıtlama sistemleri (question answering)

- Kullanıcının sorduğu soruya cevap olarak bulduğu bilgileri toparlayıp tek bir sonuç olarak döndürür.
- Günümüz soru yanıtlama sistemleri "soru sınıflama" modülü ile sorunun ve cevabın türünü belirlerler.
- Soru analiz edildikten sonra sistem metinler üzerinde karmaşık doğal dil işleme tekniklerini çalıştırır.
- Bu sırada bir filtre cevabın türüne göre paragraflar arasından cevabı bulur. Örneğin soruda "kim" sorusu soruluyorsa cevap bir özel isim olmalıdır.
- webclopedia.com, answerbus.com, answers.com

54

54

Bilgi Erişim Sistemlerinde Araştırma Konuları - Konu Tespit ve Takip Sistemleri (Topic Detection and Tracking)

- haber yayınlarının izlenerek yeni ve ilginç bir haber olduğunda ilgililerin uyarılmasını sağlayan sistemlerdir.
- Dört ana modülden oluşur.

- ***"İlk Hikaye Algılama Modülü"*** sisteme yeni bir hikaye ulaştığında bunun tartıştığı konu daha önceden tanımlanmış konularla ilgili değilse, yeni bir konu olduğunu belirler.
- "***Küme Belirleme Modülü"*** gelen hikaye bir ilk hikaye değilse ilgili konu kümesine yerleştirir, eğer ilk hikaye ise bunun için yeni bir küme oluşturur.
- ***"Haber İzleme Modülü"*** haber kaynaklarından sisteme gelen haberleri değerlendirilerek, bu haberlerin daha önceden belirlenmiş olan konularla ilgili olup olmadıklarını araştırır.
- ***"Hikaye Bağlantı Algılama"*** modülü ise sisteme ulaşan iki farklı hikayenin aynı konuyu tartışıp tartışmadıklarını anlamayı amaçlar.

55

55

Sonuç

- Mobil cihazların sabit ve dizüstü bilgisayarların yerine geçmesi internete erişimi kolaylaştıracak, içerik oluşturma hızlanacak,
- 4G ve 5G ile bant genişliklerinin 2-3 katına çıkması nedeniyle veri iletimi hızlanacak,
- Bu gelişmeler Dünya'da üretilen bilginin daha da hızla artmasına neden olacaktır.
- Bilgi Erişim Sistemleri sadece metin değil aynı oranda ses, video, resim arama araçları olacak.
- Bu nedenlerle Bilgi Erişim sistemlerinin hem işi zorlaşacak hem de ihtiyacın artmasıyla gün geçtikçe daha da önem kazanacaktır.
- Doğru kaynağa en kısa zamanda ulaşabilmek için kullanıcıyı tanıyan,arama yaparken niyetini anlayan,web içeriğinin konusunu tespit edebilen akıllı sistemlere ihtiyaç vardır.

56

56



57