Bilgi Erişim Performans Ölçüleri

Yaşar Tonta

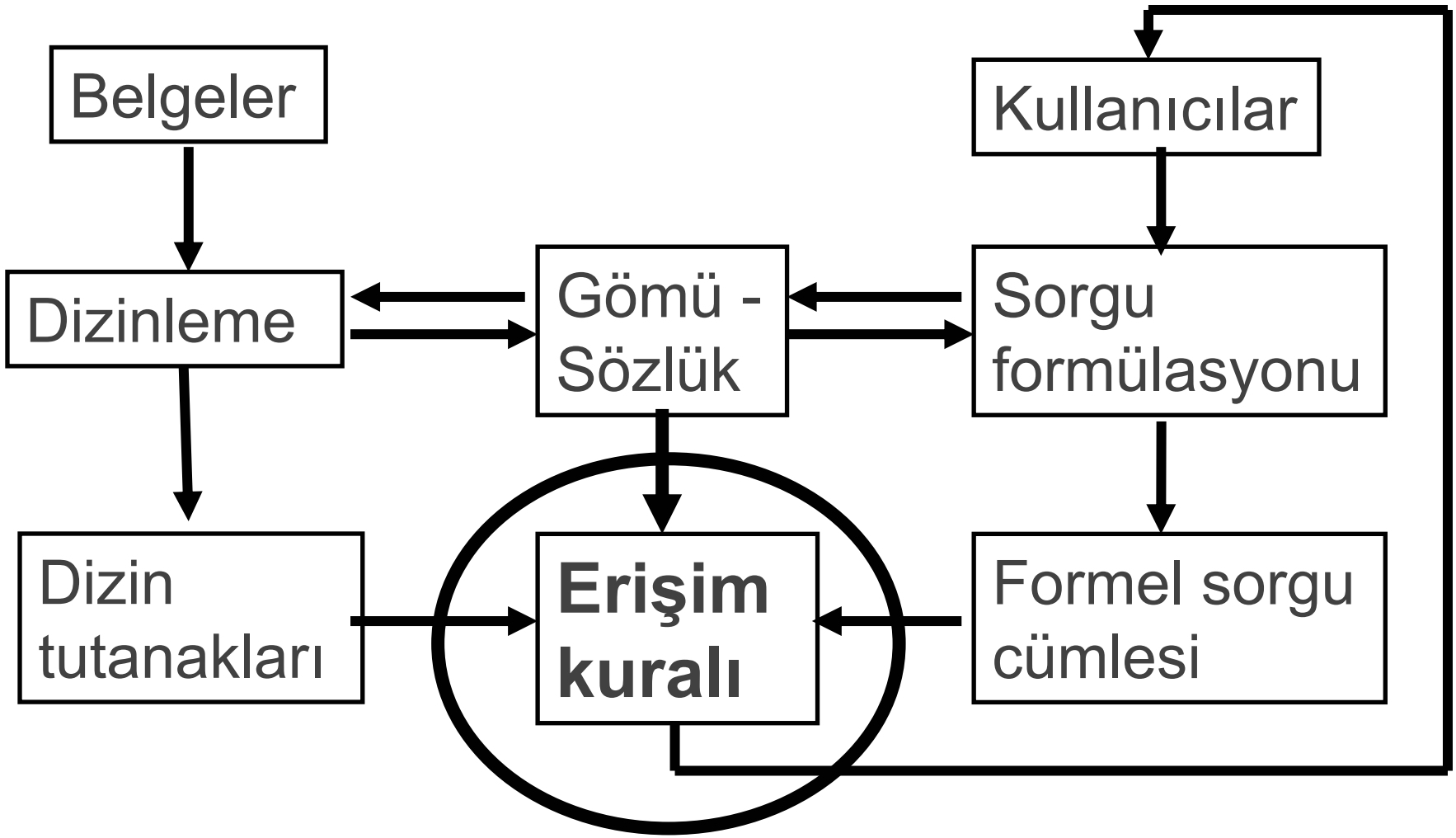
Hacettepe Üniversitesi

tonta@hacettepe.edu.tr

yunus.hacettepe.edu.tr/~tonta/

DOK324/BBY220 Bilgi Erişim İlkeleri

Belge Erişim Sisteminin Mantıksal Düzenlemesi



Kaynak: Maron, 1984

İdeal Bilgi Erişim Sistemi

- İlgili belgelerin tümüne ve salt ilgili belgelere erişim sağlamalı
- "İlgililik" kavramı
  - Nesnel ilgililik
  - Öznel ilgililik
- Birbirine benzeyen bilgileri bir araya getirmek, benzemeyenleri ayırmak

İlglilik Derecesi

- İkili ilgililik (İlgili/İlgisiz)
- 0-1 ölçeğinde ilgililik (veya 0-1000 ölçeğinde)

Hakkındalık

- Bir belgenin X konusunda olduğuna nasıl karar veririz?
- Dizin terimleri/konu başlıkları bir belgenin hangi konu(lar) hakkında olduğunu belirtir
- Dizin terimleri vermek genellikle ikili bir karardır

Ağırlıklandırma

- Belgelerde geçen terimler
- Arama sorularında geçen terimler
- Terimlere negatif ağırlık verilebilir mi?

Erişim Kuralları

Boole mantığı

Set kuramına dayanıyor. Boole işleçleri –VE, VEYA, DEĞİL- kullanılıyor

Vektör uzayı modeli

$$\sigma(D,Q) = \Sigma(t_k x q_k) \,/\, \sqrt{\Sigma(t_k)^2} \times \sqrt{\Sigma(q_k)^2}$$

t_k = k teriminin belgedeki değeri
q_k = k teriminin sorgudaki değeri

Olasılık modeli

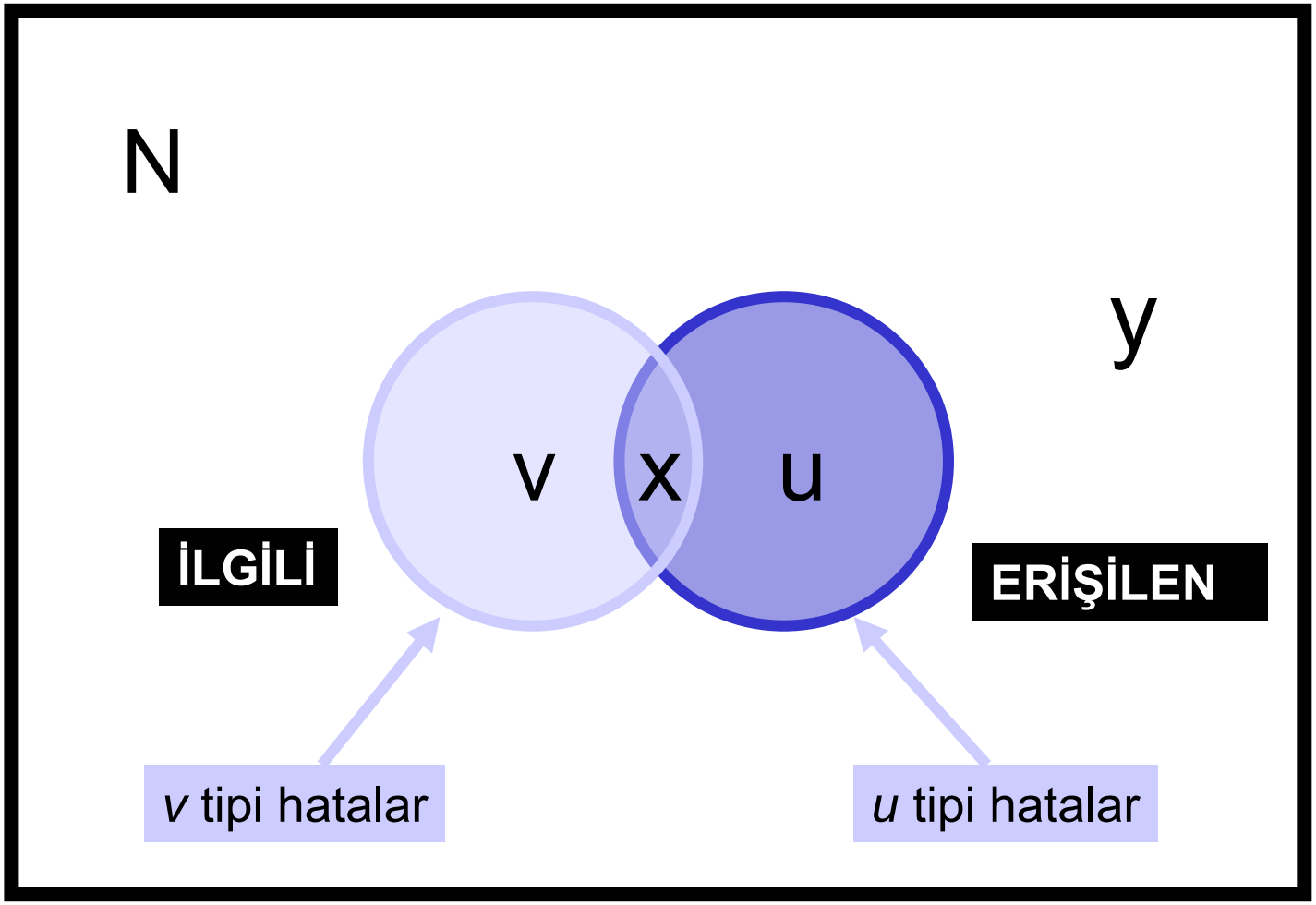
$$P(ilgili) = n / N$$
$$P(\neg ilgili) = 1 - P(ilgili) = N - n / N$$

n = ilgili belge sayısı
N = toplam belge sayısı

İstatistiksel ağırlıklandırma (tf*idf)

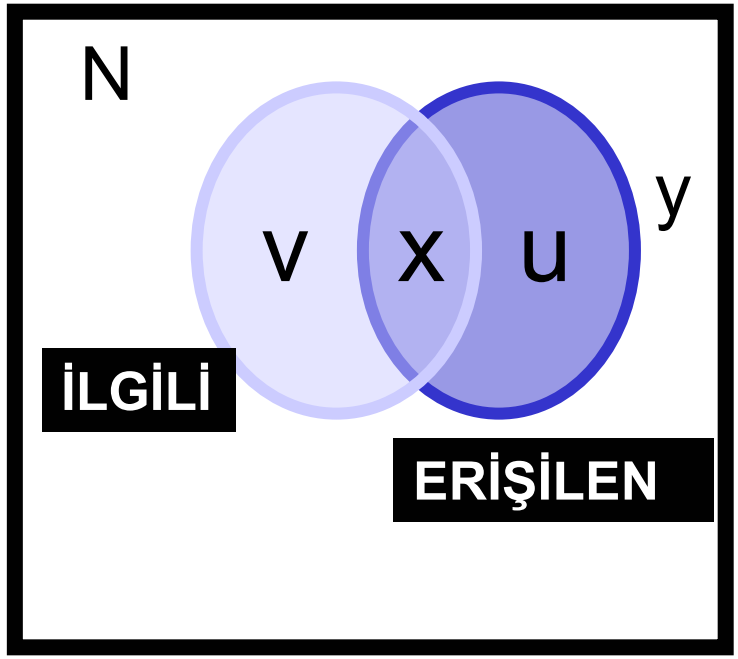
Ağırlıklandırma ilkesi: İlgili belgelerde sık **AMA** derlemin tamamında seyrek geçen terimleri daha yüksek ağırlıklandır

Bilgi Erişim Sistemleri Mükemmel Değil!



Bilgi Erişim Performansı

	İLGİLİ	İLGİSİZ	
ERİŞİLEN	x	u	n_1
ERİŞİLE-MEYEN	v	y	
	n_2		



Duyarlık = x / n_1 Erişilen ilgili belgelerin erişilen tüm belgelere oranı

Anma = x / n2 Erişilen ilgili belgelerin tüm ilgili belgelere oranı

Posa = u / u + y Erişilen ilgisiz belgelerin tüm ilgisiz belgelere oranı

Genellik = n2 / N Tüm dermedeki ilgili belgelerin oranı

Diğer Performans Ölçümleri

- **Kapsama Oranı: $|R_k| / U$**
  - Gerçekte erişilen ilgili belgelerin kullanıcının ilgili olduğunu önceden bildiği belgelere oranı
- **Yenilik Oranı: $|R_u| / |R_u| + |R_k|$**
  - Gerçekte erişilen ilgili belgelerin kullanıcının ilgili olduğunu önceden bilmediği belgelere oranı

U: kullanıcının ilgili olduğunu önceden bildiği belgeler seti

R_k: Erişilen ve kullanıcının önceden ilgili olduğunu bildiği belgelerin sayısı

R_u: Erişilen ve kullanıcının önceden ilgili olduğunu bilmediği belgelerin sayısı

Normalleştirilmiş Sıralama

Sıralama	1	2	3	4	5	6	7	8	9
Sıra1	+	+	+	+	+	-	-	-	-
Sıra2	-	-	-	-	+	+	+	+	+
Sıra3	+	+	+	-	-	-	+	-	+

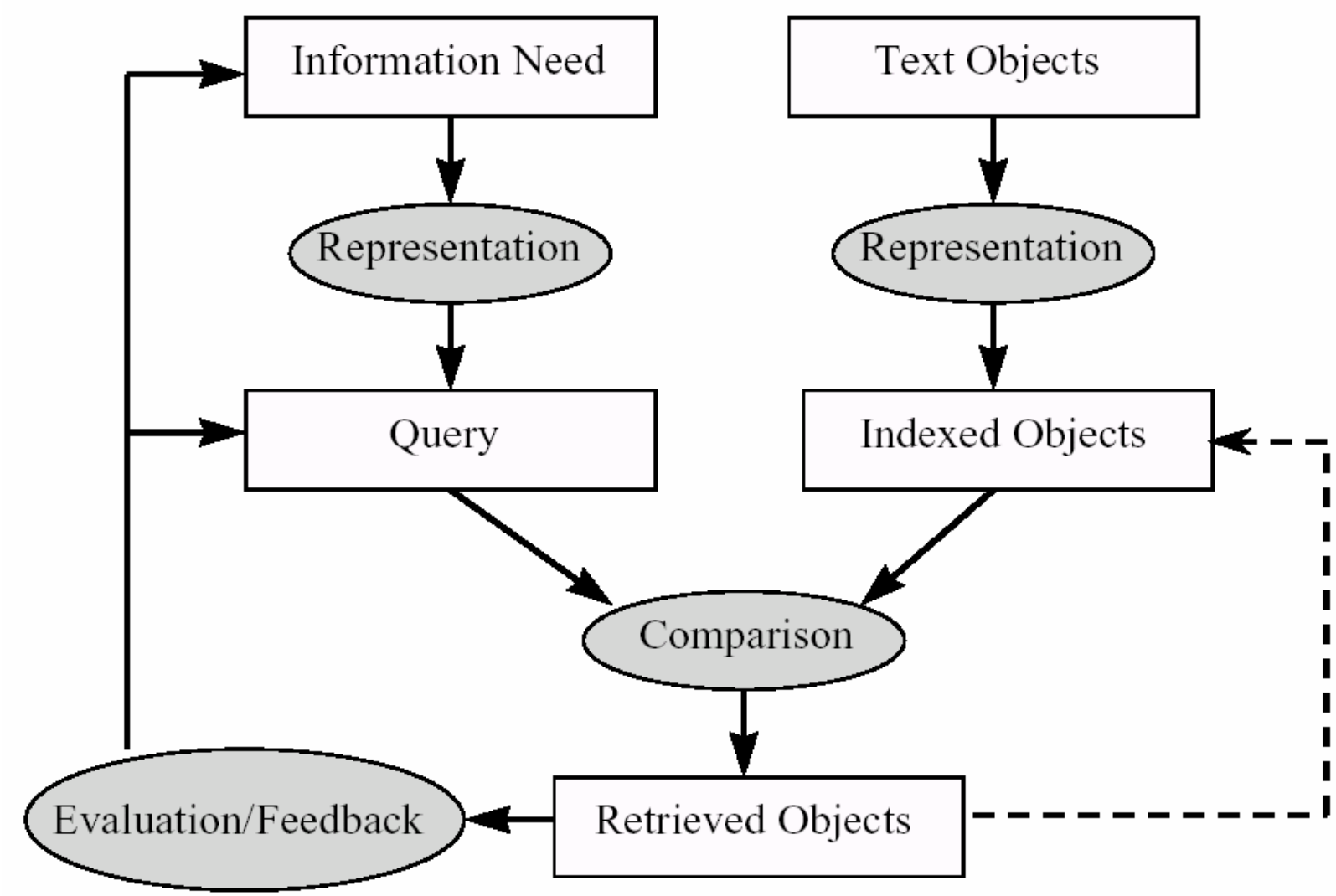
Duyarlık üç arama için de 5/9

Hangisini tercih edersiniz?

Bilgi Erişim Sorunları

Belge nedir ve boyu nasıl hesaplanır?
Bu belge ne hakkındadır?
Bu sorgu ne hakkındadır?
Bu sorgu ve belge aynı şey hakkında mıdır?
Bu belge verilen sorgu ile ilgili midir?
Bu belge sisteme sunulan bilgi ihtiyacı ile ilgili midir?
Bu belge ne kadar ilgilidir?
Bu veritabanı verilen sorgu ile ilgili midir?
Bu resim ne hakkındadır?

Bilgi Erişime İşlevsel Bakış



Belge İşleme ve Gösterimi

- Ön işlem: Noktalama işaretlerinin kaldırılması ve daha sonra durma listesinde bulunan kelimelerin belgeden ayıklanması.

- Gövdeleme: bir kelimeden yapım eklerinin korunup çekim eklerinin atılması.

- Belge Gösterimi için içerik terimleri ve onların göreceli ağırlıkları. Bir terimin ağırlığı onun belge içindeki sıklığı ile doğru, fakat derlem sıklığı ile ters orantılıdır.

Dizinler

- Dizin ne içermelidir?

Veritabanı sistemi asıl ve ikincil anahtarları dizinler.

- BE Problemi: anahtarları kestirebilmek?
- Çözüm: İçerik terimleri.

- Zipf Kanunu: Terimlerin dağılımı ve sıraları arasındaki ilişki sabit bir değere yakınsar.
- İçerik terimlerin göreceliği ağırlığı ne olmalıdır?

- Sıklık Modeli: Terim sıklığı? Belge sıklığı?
- Ayrımsama Modeli: belge uzayının yoğunluğunu azaltan terim iyi bir terimdir.
- Dil modeli: Belgenin söz konusu terimi üretme olasılığı ile derlemin üretme olasılığı arasındaki doğrusal ilişki ağırlığı belirler.

Zipf Kanunu

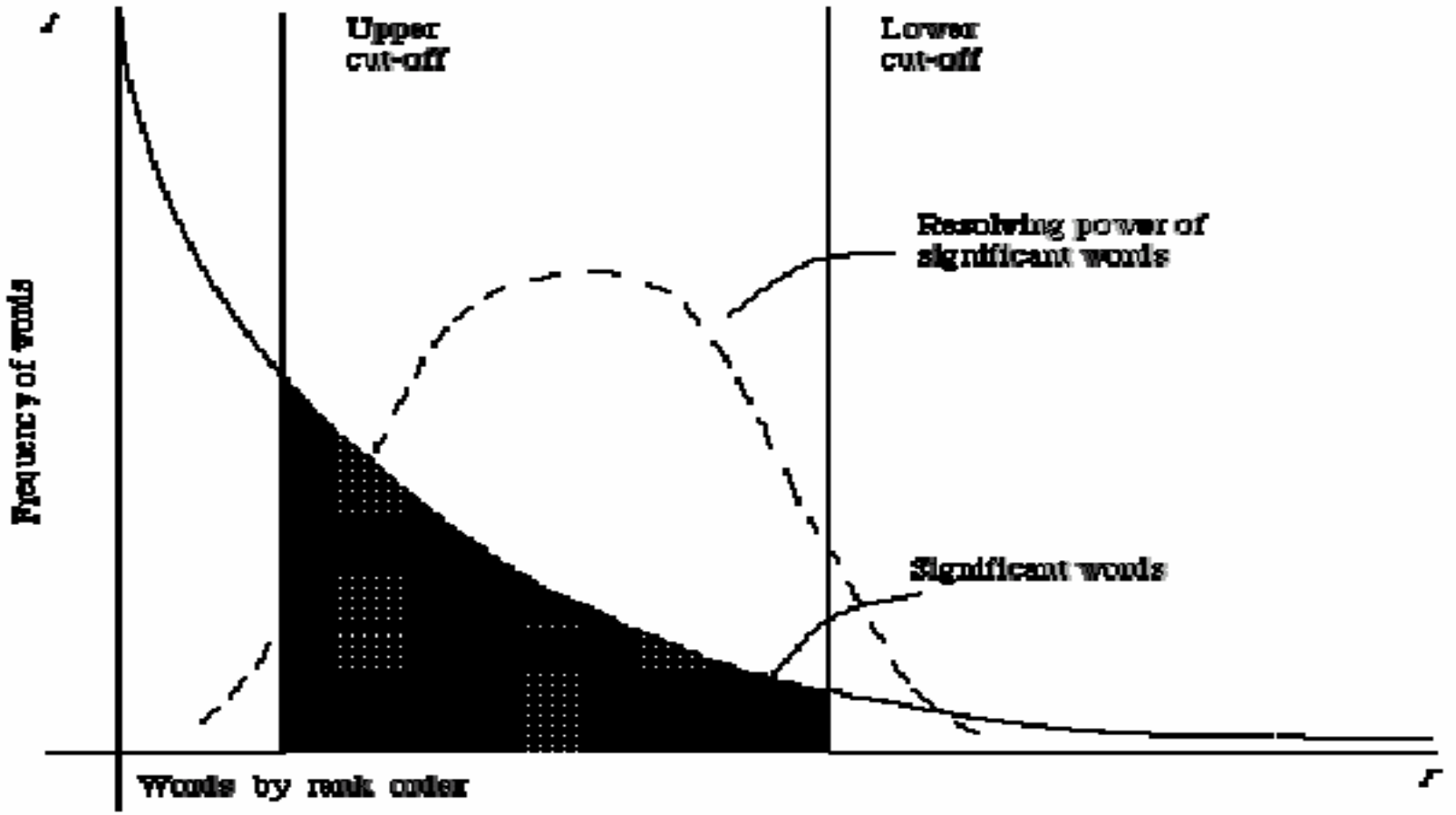
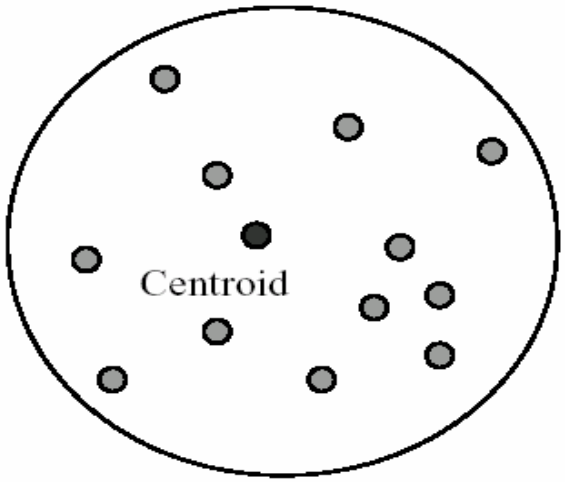
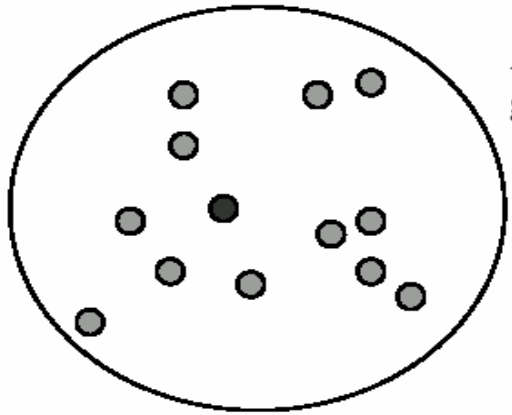


Figure 2.1. A plot of the hyperbolic curve relating f, the frequency of occurrence and r, the rank order (Adapted from Schultz[44] page 120)

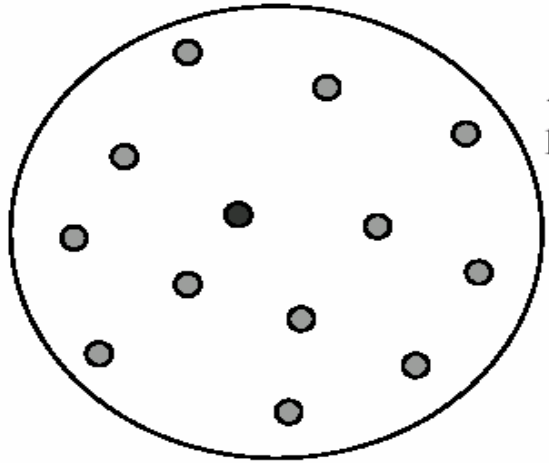
Ayrımsama Modeli



Document space with all terms



After removal of a good discriminator



After removal of a poor discriminator

Sorgu İfadesi

- 2 temel sorgu dili türü
  - Boole, yapılı
  - Serbest metin
- Birçok sistem birisini ya da her ikisini birden desteklemektedir.
- Sorgu ifadesinin oluşturulmasında kullanıcı arayüzü önemlidir.
- Sorgu ifadesinin oluşturulması için araçlar
  - Sorgu işleme ve ağırlıklandırma
  - Sorgu genişletme
  - Sözlükler ve eşanlamlı sözlük
  - İlgililik geribildirimi

Sorgu İşleme

➢ Sorgu işleme adımları otomatik belge dizinlemeninkilere çok benzemektedir.
- Durma Kelime Listesi farklı olabilir
- Metin daha az gramatik ve kısa olabilir

➢ Kullanıcı etkileşimi mümkün ve istenebilir
➢ Sorgu-tabanlı gövdeleme ve durma kelimeleri

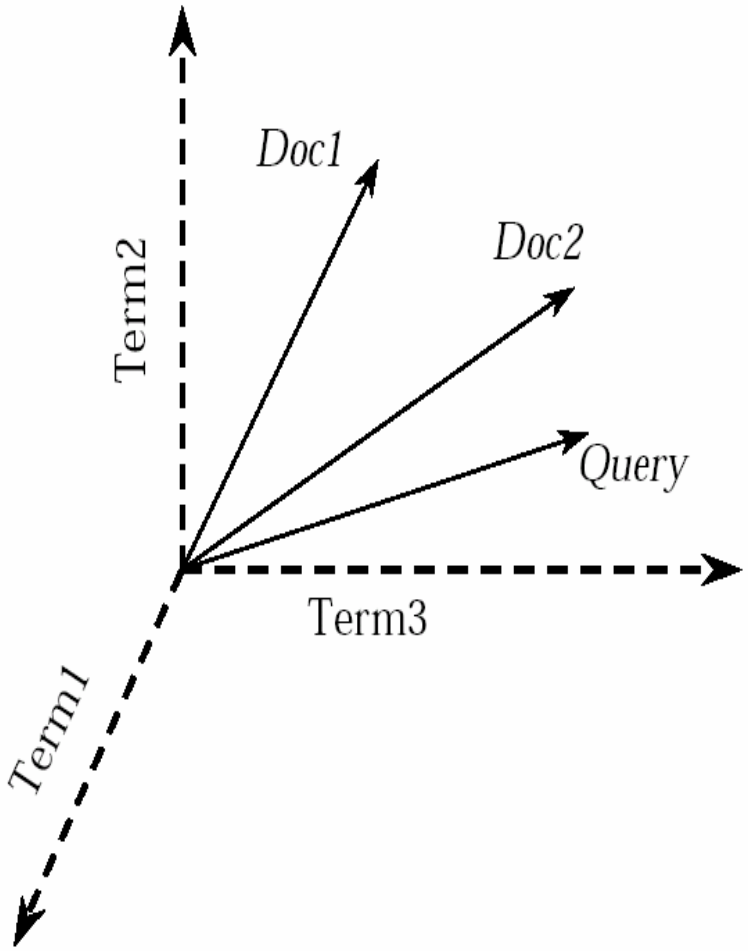
➢ Diğer olası adımlar
- Tamlamaların tanınması
- Negatiflerin tanınması
- İlgili kelimelerle sorguların genişletmesi

Erişim Modelleri

Boole model kesin eşleştirme yaklaşımına dayanmaktadır.

- Sorgular belge özelliklerini işlenenler olarak kabul eden mantık ifadeleridir.
- Geri getirilen belgeler genelde sıralanmaz.
- Acemi/Tecrübesiz kullanıcılara Boole sorgu ifadesi zor gelebilir.
- Boole geri-erişim modeli ile Boole sorguları birbirlerinden ayırma gereksinimi
- Saf Boole işleçleri: VE, VEYA, VE DEĞİL
- Bir çok sistem uzaklılık işleçlerine sahiptir
- Bir çok sistem basit düzenli ifadeleri desteklemektedir

Vektör Uzayı Bilgi Geri Erişim Modeli



- Belge, terimlerin bir vektörü olarak gösterilir.
- Sorgu, serbest metin veya terimlerin bir vektörü olarak gösterilir.
- İki vektör arasındaki açı benzerlik ile ters orantılıdır.
- Belgeleri sorguya benzerliklerine göre sıralar.

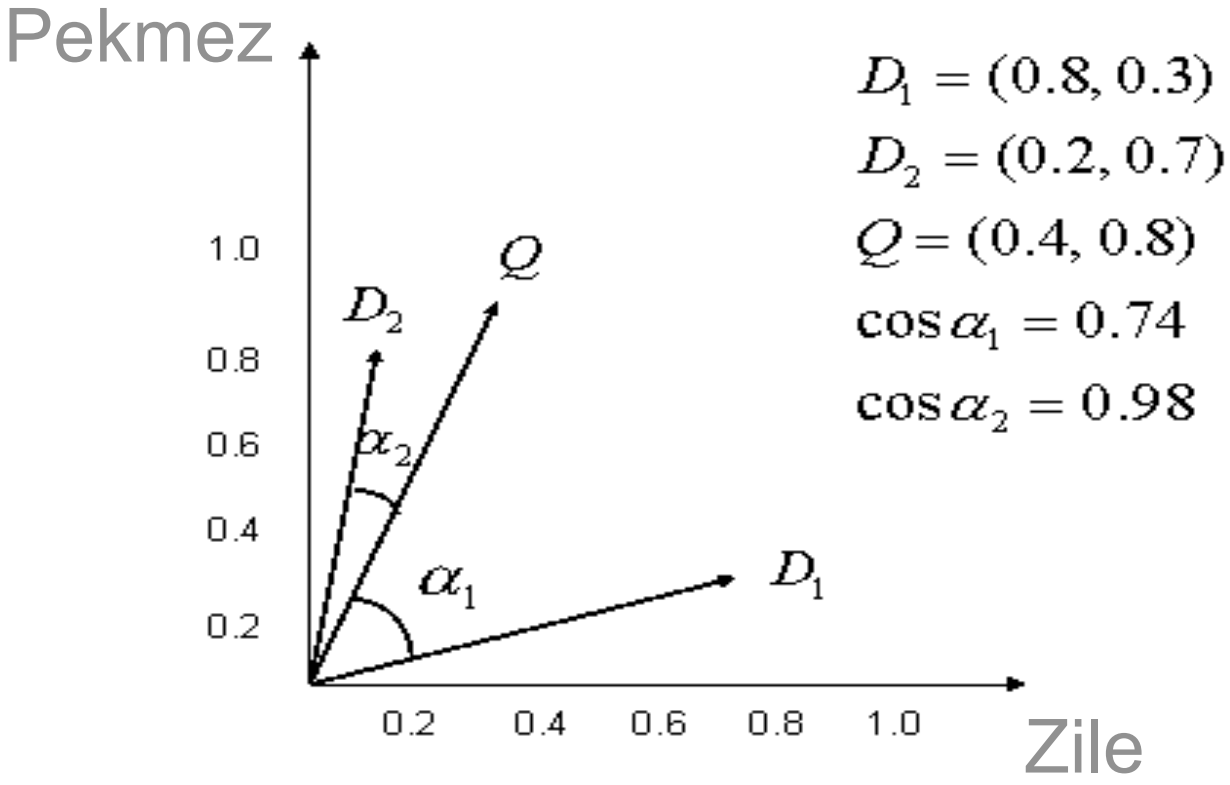
Benzerlik ölçüleri

- Bir bilgi erişim sistemi arama sorusuna benzeyen belgelere erişmeli benzemeyenleri reddetmelidir
- Bir dermedeki hangi belgelerin arama sorusunda istenenlere benzediğini (yani ilgili olduğunu) belirlemeye yarayan çeşitli benzerlik ölçüleri vardır

Vektör Uzayında Benzerlik: Ortak Ölçümler

Sim(X,Y)	Binary Term Vectors	Weighted Term Vectors
Inner product	$\lvert X \cap Y \rvert$	$\sum x_i . y_i$
Dice coefficient	$\frac{2 \lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert}$	$\frac{2 \sum x_i . y_i}{\sum x_i^2 + \sum y_i^2}$
Cosine coefficient	$\frac{\lvert X \cap Y \rvert}{\sqrt{\lvert X \rvert} \sqrt{\lvert Y \rvert}}$	$\frac{\sum x_i . y_i}{\sqrt{\sum x_i^2 . \sum y_i^2}}$
Jaccard coefficient	$\frac{\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert - \lvert X \cap Y \rvert}$	$\frac{\sum x_i . y_i}{\sum x_i^2 + \sum y_i^2 - \sum x_i . y_i}$

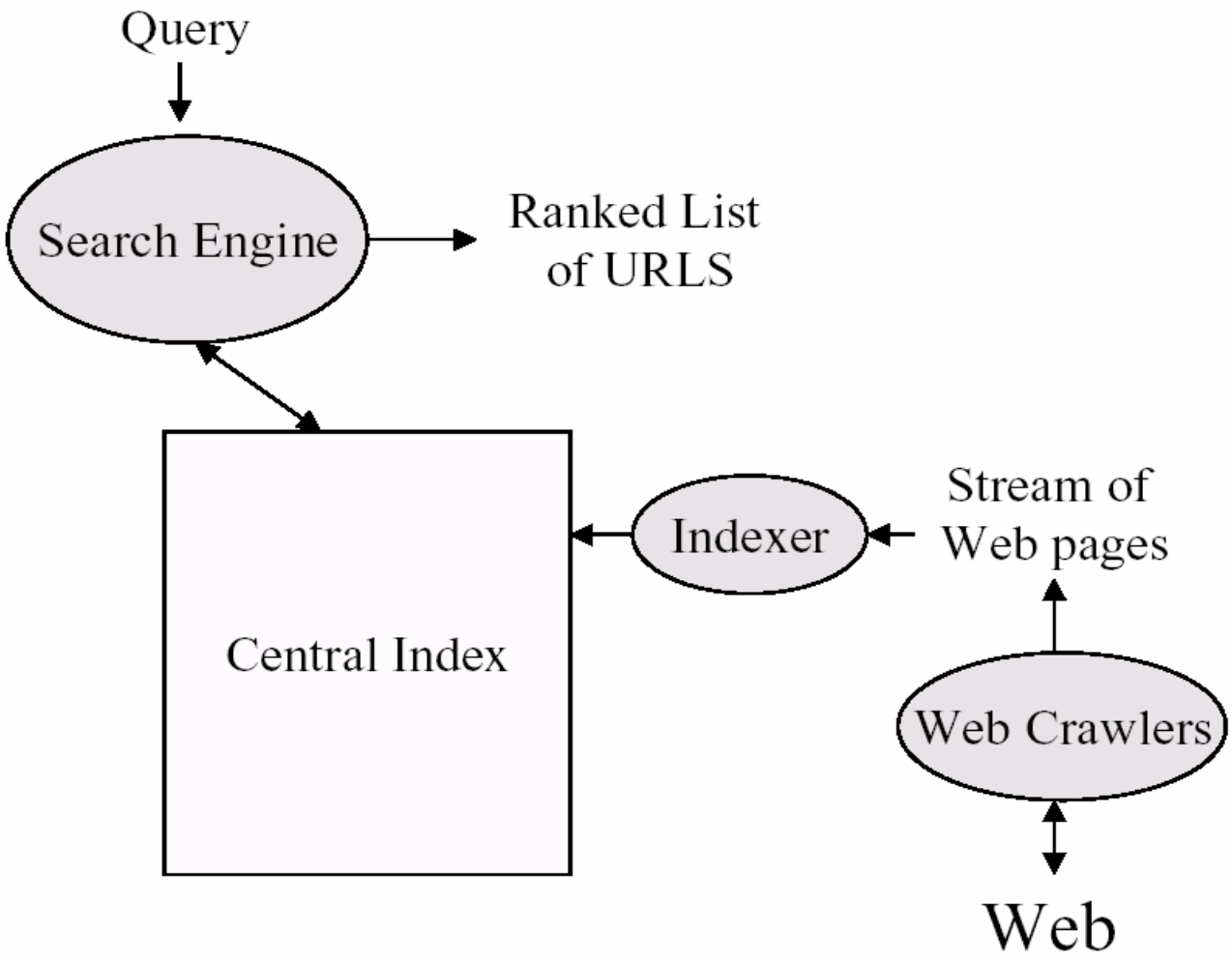
Benzerlik Skorunun Hesaplanması: Kosinüs Katsayısı



$$D_1 = (0.8, 0.3)$$
$$D_2 = (0.2, 0.7)$$
$$Q = (0.4, 0.8)$$
$$\cos\alpha_1 = 0.74$$
$$\cos\alpha_2 = 0.98$$

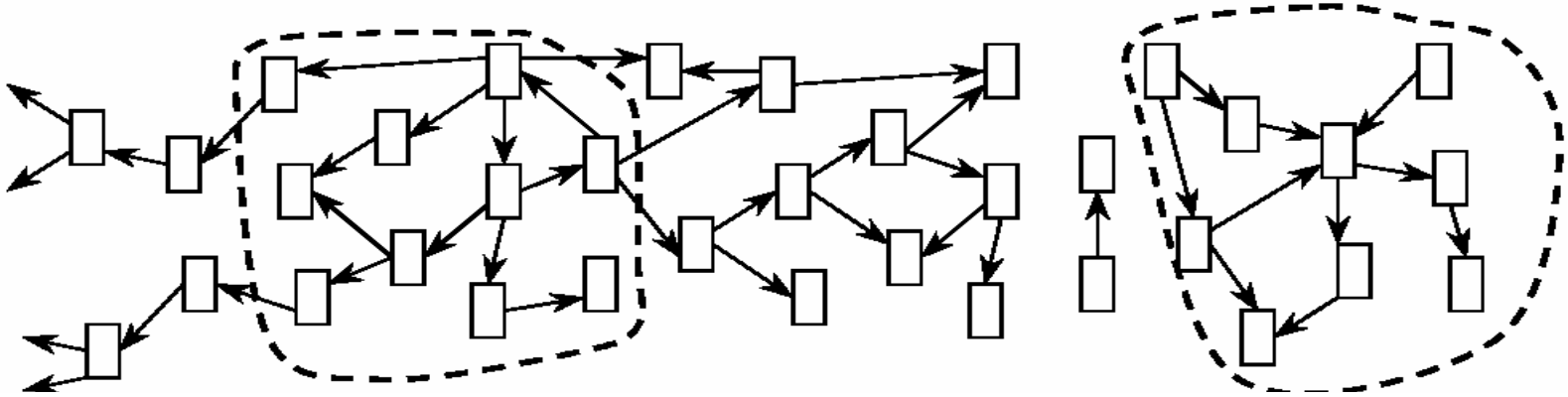
http://www.sims.berkeley.edu/courses/is296a-3/f98/lectures/ir-background/sld038.htm

Arama Motorunun Merkezi Mimarisi

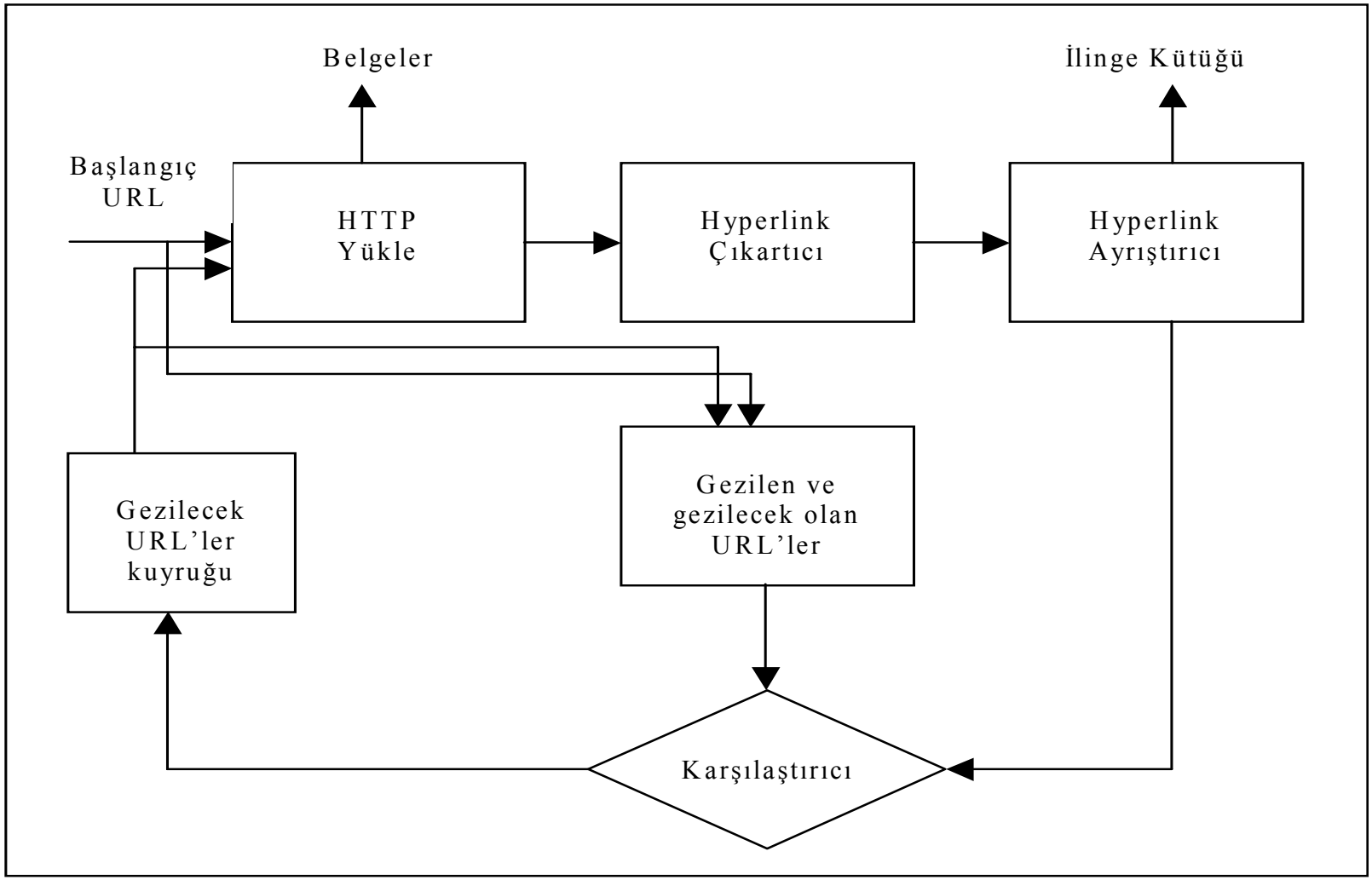


Web Örümceği ve Veri Toplama

- Hiper-bağlantılı belgeler çizgedeki düğümler olarak görülebilir.
  - – İlginç altçizgeler: alan isimleri kesişen düğümler
  - – İzole altçizgeler: Dışardan referans almayan düğümler
- Veri toplama meseleleri:
  - – Her bir düğüm nasıl bir kere ziyaret edilecek
  - – Düğümlerin temsili örneklemi nasıl elde edilir



Web Örümceği İşlevsel Mimarisi



Veri Toplama ve Sorgu Dili: Ortak Sorunlar

- Göreceli yollar: <A HREF="../../../bby220/">Yayınlar</A>
- Tekrarlı sayfalar (%30): Aynı sayfa, farklı adres.
- Javascript: Dinamik HTML
- Çok büyük sayfalar: 10 MB sayfayı gerçekten tümü ile dizinlemek istiyor musunuz?
- Dinamik içerik: Web kaynakları tahmini olarak ortalama 75 gün değişmeden kalmaktadırlar.
- Kaliteli Web sayfaları: Nasıl ölçülür?
- Meta öznitelikler: description, keywords, title, vs.
- Bir kaç kelimelik sorgular (ortalama 1.5)

Üst Arama Motorları