



Eriřim Kuralları

Yařar Tonta

Hacettepe Üniversitesi

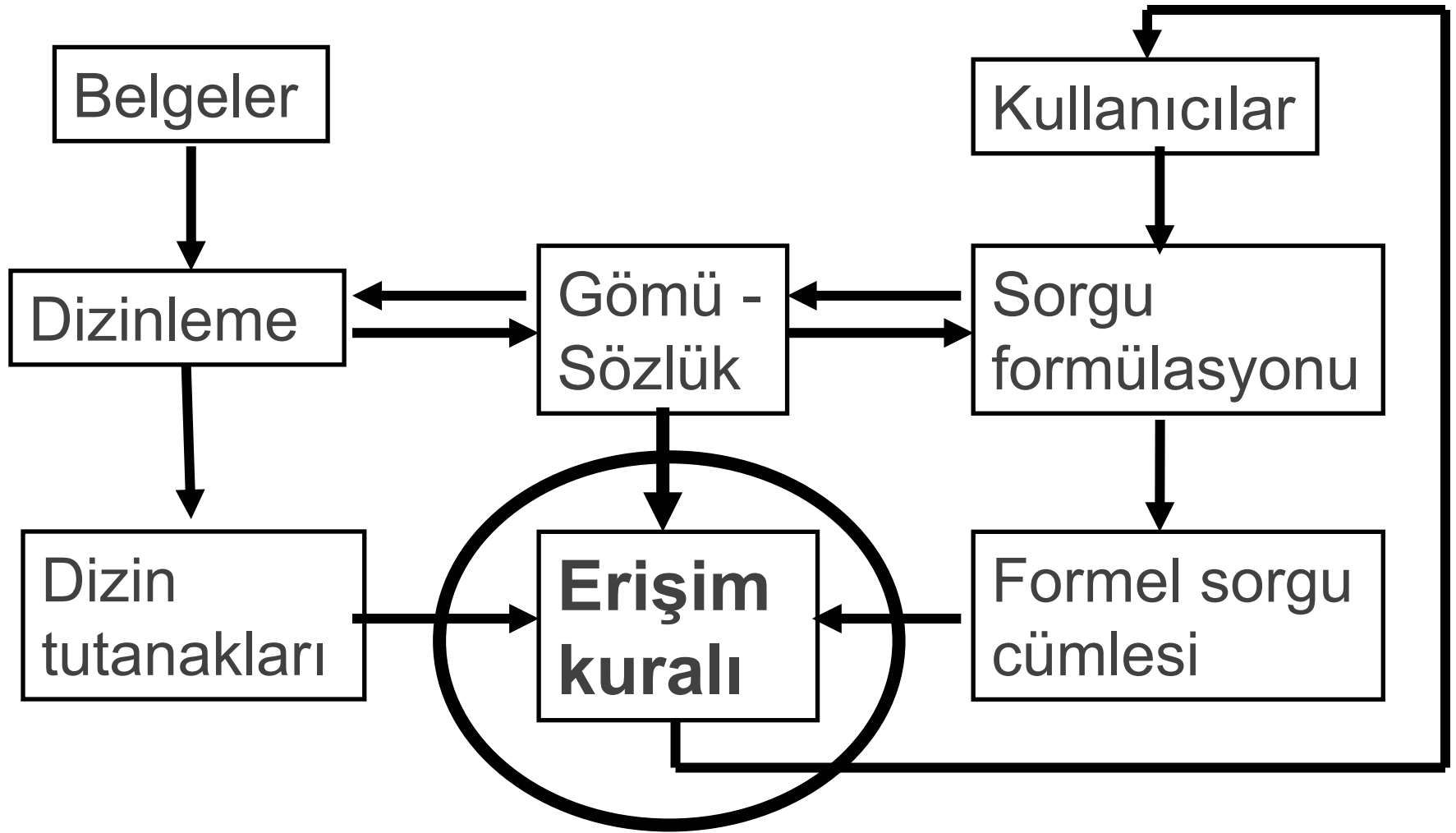
tonta@hacettepe.edu.tr

yunus.hacettepe.edu.tr/~tonta/

DOK324/BBY220 Bilgi Eriřim İlkeleri



Belge Erişim Sisteminin Mantıksal Düzenlemesi



Kaynak: Maron, 1984

Table 2.1 Summary of Retrieval Rules
Source: Compiled from Blair (1990), Chapter II.

<i>Model</i>	<i>Search Request</i>	<i>Documents</i>	<i>Retrieval Rule</i>
1	Single query terms	Documents are assigned one or more index terms	If the term in the request is a member of the terms assigned to a document, then the document is retrieved
2	A set of query terms	A set of index terms	Document is retrieved if <i>all</i> the terms in the request are in the index record of the document
3	A set of query terms plus a "cut-off" value	A set of one or more index terms	Document is retrieved if it shares a number of terms with the request that <i>exceeds</i> the cutoff value
4	Same as 3	Same as 3	Documents showing with the request <i>more</i> than the specified number of terms are ranked in order of decreasing overlap
5 Weighted Requests	Set of query terms each of which has a positive number associated with it	Same as 3	Documents are ranked in decreasing order of the sum of the weights of terms common to the request and the index record



- Erişim kuralı: erişim algoritması
- Erişim statü değeri: çıktının sınıflanması ya da sıralanması
- En basit erişim kuralı: ERİŞİLDİ / ERİŞİLMEDİ (çıktı sıralı değil)
- Çıktının sıralanması: YÜKSEK STATÜYLE ERİŞİLDİ, DÜŞÜK STATÜYLE ERİŞİLDİ, ERİŞİLMEDİ (Örnek, Medline sistemi)
- Gruplar statü değerlerine göre sıralı ama kendi içinde sıralama yok (ya da zayıf sıralama var)
- Konu kataloglarındaki sıralama nasıl?



- Blair'in kitabında anılmayan daha basit erişim kuralları da vardır.
- Basit kontrol sistemlerinde her dizin kaydı tek terimden oluşur ve iki belgeye aynı terim verilemez (yer dizini)
- Her sorgu cümlesi tek terimden oluşur; sorgu terimleri dizin terimleri listesinden alınır
- Erişim statüsü değeri: ERİŞİLDİ/ERİŞİLMEDİ
- Erişim Kuralı: Sorgu terimiyle çakışan belgelere ERİŞİLDİ değeri ver, geri kalanına ERİŞİLMEDİ değeri ver
- Yazar adı/kitap adı kataloğunda yapılan aramalar bu sınıfa girer
- Fiziksel erişim için yararlı



- Blair'de söz edilmeyen bir diđer model tek eriřim noktalı sistemlerdir.
- Bu modelde her dizin kaydı tek terimden oluşur ama birden fazla belgeye aynı terim verilebilir
- Her sorgu cümlesi tek terimden oluşur; sorgu terimleri dizin terimleri listesinden alınır
- Eriřim statüsü deęeri: ERİŐİLDİ/ERİŐİLMEDİ
- Eriřim Kuralı: Sorgu terimiyle çakışan tüm belgelere ERİŐİLDİ deęeri ver, geri kalanına ERİŐİLMEDİ deęeri ver
- Örnekler: Dosyalama sistemleri, konu katalogları, video kiralama dükkanlarındaki sınıflamalar
- Her belgeye sadece bir eriřim noktasından eriřilebilir



- Her dizin kaydı bir veya daha fazla terimden oluşur
- Her sorgu cümlesi tek terimden oluşur
- Eriřim statüsü deęeri: ERİŐİLDİ/ERİŐİLMEDİ
- Eriřim Kuralı: Sadece sorgu terimiyle çakışan belgelere ERİŐİLDİ deęeri ver, geri kalanına ERİŐİLMEDİ deęeri ver
- Örnekler: yer dizini, yazar adı-eser adı-konu adı kart kataloęu
- Kitap sayısı arttıkça bu modelin etkinlięi azalır



- Boole AND işleciyle çalışır (Conjunctive)
- Her dizin kaydı bir veya daha fazla terimden oluşur
- Her sorgu cümlesi bir veya daha fazla terimden oluşur
- Erişim statüsü değeri: ERİŞİLDİ/ERİŞİLMEDİ
- Erişim Kuralı: Sadece tüm sorgu terimleriyle çakışan belgelere ERİŞİLDİ değeri ver, geri kalanına ERİŞİLMEDİ değeri ver
- Örnekler: konu katalogları; “peek-a-boo kart sistemi”



- Her sorgu cümlesi birden fazla terimden oluşur (sıralı ya da sırasız)
- Dizin kayıtlarında sorgu cümlesindeki terimlerle en fazla çakışan terim bulunan belgelere en yüksek erişim statüsü değeri atanır, yani belgeler dizin kayıtlarıyla sorgu cümlelerindeki terimlerin çakışma oranına göre sıralanır
- Erişim statüsü değeri: 0 (Çakışma yok) / ERİŞİLMEDİ
- Erişim statüsü değeri: 1 (Yani sadece 1 terim çakıştı) -> En düşük düzeyde ERİŞİLDİ
- Erişim Kuralı: (i) Her dizin kaydı bir veya daha fazla terimden oluşur (ii) Her sorgu cümlesi bir veya daha fazla terimden oluşur; terimler dizin kaydındaki terim listesinden gelir (iii) erişim statüsü değeri: 0,1,2,3 gibi pozitif değerler (iv) Kural: dizin kaydıyla sorgu cümlesinde çakışan terim sayısına göre erişim değeri ver (0, 1, 2, 3...)
- Erişim çıktısı sıralı

Model 5: Ağırlıklı Sorgu Sistemleri



- Sorgu cümlesinde geçen terimlere ağırlık verilir ama dizin kaydında geçenlere verilmez
- Sorgu oluşturma sırasında kullanıcı terimleri ağırlıklandırır
- Eşik değeri belirlenebilir
- Sıralama çakışma oranına göre belirlenir
- Computers 5, Library 3, Information centers 3, Archives 0 (aslında tüm diğer terimlere 0 verilmiş olur)
- Her üç terim de geçen belgelere en üst sırada erişilir (11, 8, 8, 6, 5, 3, 3, 0)
- Terimlere negatif ağırlık verilmesi ne anlama gelir?
- Computers 5, Library 3, Information centers 3, Archives -5
- Dört terim de geçen belgeler “cezalandırılmış” ve sıralamada aşağıya inmiş olur (6, 11, 3, 3, 1)
- “Her Boole isteğine karşılık gelen eşik değeri olan ağırlıklı sorgu sistemi vardır” (İspatlayınız)

Yardıma için soru yazın			
6	Set of query terms	Set of index terms each of which has a positive number assigned to it	Documents are ranked in decreasing order of the sum of the weights of terms common to the request and the index record
Weighted			
Indexing			
7	Same as 5	Same as 6	Documents are ranked by the sum of products each of which results from the multiplication of the weight of the term in the request by the weight of the same term in the index record
Weighted			
Requests and			
Indexing			
8	Same as 5	Same as 6	The weights of the terms common to the request and an indexing record are treated as vectors. The value of a retrieved document is the cosine of the angle between the vectors
Cosine Rule			
9	Requests are any Boolean combination of query terms with AND, OR, and NOT	A set of one or more index terms	<p>i) AND: Retrieve only documents that match all terms in the request</p> <p>ii) OR: Retrieve only documents that match any term in the request</p> <p>iii) NOT: retrieve all documents that do not match any term in the request</p>
Boolean			
Requests			



- Dizin kaydında geçen terimlere ağırlık verilir ama sorgu cümlesinde geçenlere verilmez
- Dizinleme sırasında katalogcu terimleri ağırlıklandırır
- Sıralama çakışma oranına göre belirlenir
- Dizin terimlerine negatif ağırlık verilmesi ne anlama gelir?



- Dizin kaydında ve sorgu cümlesinde geçen terimlere ağırlık verilir
- Dizinleme sırasında katalogcu terimleri ağırlıklandırır
- Sorgulama sırasında kullanıcı terimleri ağırlıklandırır
- Sıralama çakışma oranına göre belirlenir, bir terimin sorgu değeriyle dizin değerinin çarpımı esas alınır



- Model 7'ye benzer (hem sorgu hem dizin terimleri ağırlıklandırılmış ve erişilen kayıtlar sıralanmış)
- Arama ve dizin terimleri çok boyutlu bir uzayda vektörler olarak tanımlanır
- Çakışmanın gücü (yani sıralama) arama ve dizin terimleri vektörleri arasındaki açının kosinüsüne göre hesaplanır
- Vektör uzayı modeline dayanan belge erişim sistemleri 1960'lardan beri kullanımdadır (ör., SMART)



- Boole AND, OR, NOT işleçleriyle çalışır
- Her dizin kaydı bir veya daha fazla terimden oluşur
- Her sorgu cümlesi bir veya daha fazla terimden oluşur
- Erişim statüsü değeri: ERİŞİLDİ/ERİŞİLMEDİ
- Erişim Kuralı:
 - Sorgu A VE B biçimindeyse, dizin kaydında her iki terimin (A, B) de geçtiği belgelere eriş
 - Sorgu A VEYA B biçimindeyse, dizin kaydında A, B veya hem A hem de B'nin geçtiği belgelere eriş
 - Sorgu A DEĞİL B biçimindeyse dizin kaydında A geçen ama B geçmeyen belgelere eriş
- Örnekler: Boole mantığıyla çalışan çevrimiçi kataloglar ve diğer sistemler”

10	Same as 9	Entire text of the documents is searchable (except stop words)	Same as Model 9 with adjacency operators
Full Text			
Retrieval			
11	Single terms	A set of one or more index terms	The request term is looked up in a thesaurus (online) and semantically related terms are added to the request term
Simple			
Thesaurus			
12	Single terms	A set of one or more index terms	The request term is looked up in a thesaurus (online) and semantically related terms above a given cut-off value (weight) are added (disjunctively) to the request term. The cut-off value could be given by the inquirer.
Weighted			
Thesaurus			



- Boole VE, VEYA, DEĞİL işleçleriyle çalışır
- Her sorgu cümlesi bir veya daha fazla terimden oluşur
- Her dizin kaydı belgede geçen (dur listesi dışındaki) terimlerden oluşur
- Eriřim statüsü değeri: ERİŐİLDİ/ERİŐİLMEDİ
- Eriřim Kuralı:
 - Sorgu A VE B biçimindeyse, dizin kaydında her iki terimin (A, B) de geçtiđi belgelere eriř
 - Sorgu A VEYA B biçimindeyse, dizin kaydında A, B veya hem A hem de B'nin geçtiđi belgelere eriř
 - Sorgu A DEĞİL B biçimindeyse dizin kaydında A geçen ama B geçmeyen belgelere eriř
 - Sorgu A BİTİŐİK (ADJACENT) B biçimindeyse, dizin kaydında her iki terimin arka arkaya geçtiđi belgelere eriř (örnek, information adjacent retrieval)
 - Sorgu A İLE (WITH) B biçimindeyse, dizin kaydında her iki terimin aynı paragrafta geçtiđi belgelere eriř (örnek, information with retrieval; bazen kaç kelime arayla geçmesi gerektiđi de tanımlanabilir, information (w3) retrieval)
 - Sorgu A AYNI (SAME) B biçimindeyse, dizin kaydında her iki terimin aynı cümlede geçtiđi belgelere eriř
- Örnekler: Boole mantıđıyla çalışan tam metin bilgi eriřim sistemleri



- Küme kuramına dayalı modeller
 - Boole
 - Genişletilmiş Boole
 - Bulanık mantık (fuzzy logic)
- Vektör Modelleri (cebirsal)
- Olasılıksal modeller (probabilistic)
- Diğerleri (ör., sinir ağları)



- 1) Sorgu terim(ler)iyle dizin terimlerinin Boole erişim kurallarına göre kesin çakışmasına dayanır (Model 1-4, 9-12);
 - Dizinleme ve sorgu terimleri ikili, yani bir terim belgeye ya atanır ya atanmaz (sorgu cümlesinde ya vardır ya yoktur)
 - Her terim erişim açısından eşit değerdedir.
 - Birden çok terimle yapılan aramalarda eşik değerleri konulabilir (Model 3 ve 4).
 - Arama terimleri bir gömüden (thesaurus) alınan ilgili terimler eklenerek zenginleştirilebilir (Model 11 ve 12).
 - Erişilen kayıtlar kabaca erişilenler ve erişilemeyenler olarak sıralıdır (weakly ordered) (Model 1-3, 12).
 - Veya sorgu cümlesinde ve dizin kaydında bulunan terim sayısına göre sıralanır (Model 4).
 - Terimler arasındaki ilişkiler Boole mantığına göre tanımlanabilir (ör.,dizin kaydında sadece A ve B arama terimleri birlikte geçen belgeleri getir) (Model 9 ve 10).
 - Boole modeli “bilgisayara dayalı belge erişim sistemleri için en popüler erişim tasarımıdır” (Blair, 1990, s.44).

Olasılığa Dayalı Erişim Kuralları (Model 5-7)



- Ağırlıklandırılmış sorgu terimleri (Model 5)
- Ağırlıklandırılmış dizin terimleri (Model 6)
- Ağırlıklandırılmış sorgu ve dizin terimleri (Model 7)
-
- Erişim amaçları açısından bir terimin önemi kullanıcı tarafından belirlenir
- Erişilen kayıtlar sorgu ve dizin terimleri arasındaki çakışmanın gücüne göre sıralanır