Türkçe Arama Motorlarında Performans Değerlendirme

Yaşar Tonta
Hacettepe Üniversitesi

Yıltan Bitirim
Doğu Akdeniz Üniversitesi

Hayri Sever
Massachusetts Üniversitesi

TOTAL Bilişim Ltd. Şti.
Ankara

Türkçe Arama Motorlarında Performans Değerlendirme

Türkçe Arama Motorlarında Performans Değerlendirme

Yaşar Tonta

Hacettepe Üniversitesi

Yıltan Bitirim

Doğu Akdeniz Üniversitesi

Hayri Sever

Massachusetts Üniversitesi

Total Bilişim Ltd. Şti.

Ankara

Dizgi ve baskı: Damla Matbaası, Ankara, tel. 0312 312 75 16

Tonta, Yaşar
Türkçe arama motorlarında performans değerlendirme / Yaşar Tonta, Yıltan Bitirim ve Hayri Sever. Ankara: Total Bilişim Ltd. Şti., 2002.
Xiv, 154s.; 23cm.
Kaynakça: 137-148; dizin: 149-152.
ISBN 975-92923-0-0
1. Web arama motorları 2. Web arama motorları – Bilgi erşim I. Bitirim, Yıltan. II. Sever, Hayri. III. Başlık
ZA4226 T616 025.04

Prof. Dr. İlhan Kum'un aziz anısına saygıyla...

ÖNSÖZ

Günümüzde çok hızlı bir elektronik bilgi artışıyla karşı karşıyayız. Basılı bilgi miktarı yaklaşık 14 yılda bir ikiye katlanırken, Internet aracılığıyla erişilen bilgiler her yıl 2-3 kat artmaktadır. Dünyadaki en zengin kütüphanelerden birisi olan Amerikan Kongre Kütüphanesi'nde yaklaşık 170 milyon belge bulunmaktadır. World Wide Web'de ise herkesin erişimine açık birkaç milyar belge bulunmaktadır. Web'e bağlı ancak doğrudan erişilemeyen intranetler üzerindeki belgeleri de bu rakama eklediğimizde dünya üzerindeki her bireye yaklaşık 90 belge düşmektedir! Yaklaşık yarım milyar civarındaki Internet kullanıcısı her gün milyarlarca belge arasından istediklerini bulmaya çalışmaktadırlar. Hızlı bilgi artışıyla başa çıkmaya çalışan Internet kullanıcılarının durumu "yangın hortumundan su içmeye çalışan" kimselere benzetilmektedir.

Internet kullanıcıları istedikleri bilgilere erişmek için çoğu zaman "arama motorları"nı kullanmaktadırlar. Milyarlarca Web sayfası arasından kullanıcıların işine yarayacak belgeleri bulmaya çalışan AltaVista, Google, Yahoo! gibi arama motorları Web üzerinde bulunan bilgilerin ancak küçük bir kısmını dizinleyebilmektedirler. Dahası, yapılan araştırmalarda söz konusu arama motorlarının bilgi erişim performanslarının pek yüksek olmadığı ortaya çıkmaktadır. Arama motorlarının kullanıcıların bilgi gereksinimlerini daha iyi karşılayabilmesi için neler yapılması gerektiği konusunda performans değerlendirme araştırmaları yapılmaktadır.

Bu araştırmada ülkemizde yaygın olarak kullanılan dört Türkçe arama motorunun (Arabul, Arama, Netbul ve Superonline) bilgi erişim performansları çeşitli ölçütlere göre değerlendirilmektedir. Arama motorlarına yöneltilen farklı türdeki sorulara karşılık erişilen "ilgili" ve "ilgisiz" belgelere dayanarak yapılan değerlendirmede her arama motoru için duyarlık, normalize sıralama, kapsama, yenilik ve ölü bağlantı oranları bulunmuş, sorularda Türkçe karakter kullanılmasının erişim sonuçlarına etkileri araştırılmış ve arama motorlarının belgeleri dizinlemek amacıyla HTML üst veri belirteçlerinden yararlanıp yararlanmadıkları test edilmiştir.

Arama motorları konusunda araştırma yapma düşüncesi üniversitede bilgi erişim sistemleri konusunda verdiğimiz dersler sırasında doğdu. Çok daha mütevazı bir girişim olarak başlayan bu araştırma yaklaşık altı ay sürdü. Araştırmanın altıncı bölümünün taslağını okuyarak görüşlerini bildiren Sayın Dr. Aydın Erar'a, kitabın kapak tasarımını yapan Sayın Erol Olcay'a, çalışmanın yayınlanmasını sağlayan Total Bilişim Teknolojisi Sanayi ve Ticaret Ltd. Şti. Genel Müdürü Sayın Yüksel Çetinkaya'ya ve kitabı basan Damla Matbaacılık Ltd. Şti. yöneticisi Sayın Kayhan Kaya ve takımına içtenlikle teşekkür ederiz.

YT, YB, HS

İÇİNDEKİLER

TABLOLAR LİSTESİ

ŞEKİLLER LİSTESİ

ÖZET

Bu çalışmada Türkçe arama motorlarının bilgi erişim performansları çeşitli ölçütlere göre değerlendirilmiştir. Ülkemizde yaygın olarak kullanılan Arabul, Arama, Netbul ve Superonline arama motorları üzerinde çeşitli türde 17 farklı soru için arama yapılmış ve bu sorulara karşılık erişilen "ilgili" ve "ilgisiz" belgelere dayanarak söz konusu dört arama motorunun çeşitli kesme noktalarındaki duyarlık ve normalize sıralama değerleri hesaplanmıştır. Arama motorlarının dizinlenen belgeleri ne kadar sıklıkla ziyaret ettikleri ve güncelleştirdikleri erişim çıktılarında yer alan "ölü" (yani erişilemeyen) adreslerin sayısına bakılarak saptanmıştır. Türkçe arama motorlarında en sık aranan beş sözcük ("mp3", "oyun", "sex", "erotik" ve "porno") dört arama motorunda aranmış ve her arama motorunun kapsama ve yenilik oranları bulunmuştur. Arabul, Arama, Netbul ve Superonline'ın belgeleri dizinlemek amacıyla "anahtar sözcük", "tanım" gibi HTML üst veri (metadata) alanlarından yararlanıp yararlanmadıkları iki küçük deneyle sınanmıştır. Kruskal-Wallis ve Mann-Whitney istatistikleri kullanılarak arama motorlarının güncellik, duyarlık, normalize sıralama, kapsama ve yenilik oranlarının birbirinden farklı olup olmadığı test edilmiştir.

Araştırmadan elde edilen belli başlı bulgular şunlardır: Arabul, Arama, Netbul ve Superonline'ın eriştiği ortalama her altı belgeden birisi ölü bağlantı içermektedir. Netbul'un ölü bağlantı oranı diğer arama motorlarından daha düşüktür. Arama motorları bazı sorular için hiç bir belgeye ya da hiç bir ilgili belgeye erişememiştir. Erişilen ortalama her altı belgeden beşi ilgisizdir. Arama motorlarının ortalama duyarlık oranları %11 (Netbul) ile %28 (Arama) arasında değişmektedir (Superonline %20, Arabul %15). Arama, ilk 5 belgede Arabul ve Netbul'dan daha fazla sayıda ilgili belgeye erişmiştir. Arama motorları erişilen ilgili belgeleri erişim çıktılarının ilk sıralarında gösterme konusunda yeterince çaba sarfetmemektedirler. Arama motorlarının ortalama normalize sıralama değerleri %20 (Arabul) ile %54 (Arama) arasında değişmektedir (Superonline %37, Netbul %30). Arama, erişim çıktılarında ilgili belgeleri Arabul'dan ve Netbul'dan daha üst sıralarda göstermektedir. Duyarlık ile normalize sıralama değerleri arasında gözlenen güçlü pozitif ilişki, değerlendirilen belge sayısı arttıkça giderek zayıflamaktadır. Arama motorları, Web'de yaygın olarak kullanılan terimlerin geçtiği spesifik arama sorularında nispeten daha az başarı göstermişlerdir. Tek sözcükten oluşan ya da "VEYA" işleci kullanılan sorularda, erişilen ilgisiz belge sayısı yüksek olmasına rağmen, arama motorları nispeten daha başarılı olmuştur. "VE" işlecinin kullanıldığı sorularda ise başarı oranı daha düşüktür. Arama motorları soruları daha iyi analiz etmek ve performansı artırmak için gövdeleme algoritmalarından yararlanmamaktadırlar. Türkçe arama motorlarında Türkçe karakter sorunu henüz çözülememiştir. Arama motorları Türkçe karakterler kullanılarak yapılan aramalarda farklı sonuçlar vermektedir. En sık aranan "mp3", "oyun", "sex", "erotik" ve "porno" soruları için Superonline'ın kapsama oranları daha yüksektir. Arama dışında diğer Türkçe arama motorlarının Türkiye adresli belgeleri/siteleri pek dizinlemedikleri ortaya çıkmıştır. Türkiye adresli belgeleri kapsamada Arama tartışmasız bir üstünlüğe sahiptir. En sık aranan sorularda hemen hemen tüm arama motorlarının yenilik oranları yüksektir. Aynı sorulara karşılık farklı arama motorları farklı ilgili belgelere erişmektedirler. HTML belgelerinde yer alan "anahtar sözcük" ve "tanım" üst veri (metadata) alanlarında geçen terimlerin bazı arama motorları (Netbul ve Superonline) tarafından dizinlendiği ve erişim amacıyla bu terimlerden yararlanılmadığı ortaya çıkmıştır.

Çalışmanın sonunda Türkçe arama motorlarının bilgi erişim performanslarını geliştirmek için bazı önerilere yer verilmektedir.

SUMMARY

Evaluation of Information Retrieval Performance of Turkish Search Engines

This is an investigation on the information retrieval performances of search engines based on various measures. We searched 17 queries of differing types on four Turkish search engines, namely Arabul, Arama, Netbul and Superonline. We classified each document/Web site contained in the retrieval results as being "relevant" or "non-relevant". Based on this classification, we calculated the precision and normalized ranking ratios in various cut-off points for each query run on each search engine. We checked the "dead" or "broken" links among the retrieval results to determine how often the crawlers of search engines visit the sites they index and how often they update their indexes, if needed. We found out the coverage and novelty ratios of each search engine by searching five keywords that have been the most frequently submitted queries to the Turkish search engines. Those keywords are "mp3", "oyun" (game), "sex", "erotik" (erotica) and "porno" (porn). By means of two modest experiments, we tested to see if Turkish search engines make use of index terms that are assigned by the authors of Web pages and included under the "keywords" and "description" meta tags of HTML documents. Using Kruskal-Wallis and Mann-Whitney statistics, we tested if up-to-dateness, precision, normalized ranking, coverage and novelty ratios of each search engine differ significantly from each other.

Major findings of our research are as follows: On the average, one in six documents retrieved by search engines was not available due to dead or broken links. Netbul retrieved fewer documents with dead or broken links than other search engines did. Some search engines retrieved no documents (so called "zero retrievals") or no relevant documents for some queries. On the average, five in six documents retrieved were not relevant. Average precision ratios of search engines ranged between 11% (Netbul) and 28% (Arama) (Superonline being 20% and Arabul 15%). Arama retrieved more relevant documents than that of Arabul and Netbul in the first five documents retrieved. Search engines do not seem to make every efforts to retrieve and display the relevant documents in higher ranks of retrieval results. Average normalized ranking ratios of search engines ranged between 20% (Arabul) and 54% (Arama) (Superonline being 37% and Netbul 30%). Arama retrieved the relevant documents in higher ranks than that of Arabul and Netbul. The strong positive correlation between the precision and normalized ranking ratios got weakened as the number of documents that we evaluated increased. Search engines were less successful in finding relevant documents for specific queries or queries that contained broad terms. Although non-relevant documents were higher in number, search engines were more successful in single-term queries or queries with Boolean "OR" operator. The success rate was lower for queries with Boolean "AND" operator. Search engines seemingly do not use stemming algorithms to better analyze queries and to increase retrieval performance. The use of Turkish characters such as "ç", "ö", and "ş" in queries still creates problems for Turkish search engines as retrieval results differed for such queries. Superonline's coverage rate was much higher than that of other search engines for the most frequently searched queries on the Turkish search engines. Except Arama, search engines index fewer documents/sites with domain names ending with ".tr". Arama is the indisputable leader in covering documents with Turkish addresses. Almost all search engines scored high in novelty ratios for the most frequently searched queries. Different search engines tend to retrieve different relevant documents for the same queries. For retrieval purposes, Netbul and Superonline seem to index and make use of metadata fields that are contained in HTML documents under "keywords" and "description" meta tags.

The research report concludes with some recommendations to improve the information retrieval performances of Turkish search engines.

1 GİRİŞ

Internet, IP protokolünü kullanarak bilgisayar ağlarını birbirine bağlayan dünya çapında bir bilgisayar ağı olarak tanımlanabilir. Bilgisayarların küresel olarak birbirine bağlanması temelinde şekillenen Internet fikri, 1962 yılında J.C.R. Licklider tarafından savunma amaçlı bir proje (DARPA: Defense Advanced Research Projects Agency) olarak başlatılmıştır. O zamanlar ARPANET olarak adlandırılan Internet, ilk defa 1969 yılında ABD'nin güneybatı bölgesindeki dört ana bilgisayarı (Kaliforniya Üniversitesinin Los Angeles ve Santa Barbara yerleşkeleri, Utah Üniversitesi, ve Stanford Araştırma Enstitüsü) çevrimiçi (online) olarak birleştirmiştir (Howe, 2001). Internet, Web belgeleri[1] içerisinde depolanmış bilgileri bir bilgisayardan başka bir bilgisayara taşıyan bir araç görevini görmektedir. Bilgiler Internet üzerinde değil Internet'e bağlı olan bilgisayarlar üzerinde bulunmaktadır. Internet sadece bilginin bir bilgisayardan başka bir bilgisayara aktarılmasını sağlamaktadır.

Amerikan Devleti tarafından desteklendiği için Internet başlangıçta sadece eğitim, araştırma ve devlet kullanımı ile sınırlandırılmıştı. Bu amaçlara hizmet etmeyen ticari kullanım '90'ların başlarına kadar yasaklanmıştı. Geçen zaman içinde ticari ağların büyümesi sonucu veri trafiğinin Amerikan Ulusal Bilim Vakfı Ağı (NSFNet: National Science Foundation Net) omurgası olmadan da ülke boyunca akması ancak mümkün olabilmişti. Internet'in ticari amaçlar da dahil tam olarak kullanılması ise 1995 yılının ortasına rastlamaktadır. Delphi ile başlayan Internet çıkışı ve hizmetleri daha sonraları AOL (American On-Line), Prodigy ve CompuServe ile devam etmiştir. Bu gelişmelerle orantılı olarak Amerikan Ulusal Bilim Vakfı'nın Internet gelişimindeki rolü ağ omurgasının desteklenmesi ve yüksek eğitim kurumlarının erişimlerinin sağlanmasının ötesine geçerek, ana okulu-ilkokul (K-12) ve yerel halk kütüphanelerinin erişimlerinin oluşturulmasına ve çok yüksek hacimli bağlantılar üzerine yapılan teknolojik araştırmaların desteklenmesine yönelmiştir. Daha önce de belirtildiği gibi, ABD'de başlangıçta yalnızca askeri alandaki bilgileri transfer etmek amacıyla geliştirilmiş olan Internet, günümüzde hemen hemen tüm dünyada kullanılan ve ticaret, eğitim, eğlence, spor, bilim, alış veriş gibi çok çeşitli konulardaki bilgiyi bünyesinde barındıran büyük bir bilgi sistemine dönüşmüştür. Dünyanın birçok yerinde bulunan her çeşit bilgisayarın, doğrudan ve saydam bir biçimde birbiriyle iletişim kurmalarını ve sunulan hizmetlerden yararlanmalarını sağlayan küresel bir ağ halini almıştır (Internet Society, 2000).

---

[1] Bu çalışmada Web belgesi, HTML (Hypertext Markup Language) veya XML (Extended Markup Language) dili ile tanımlanmış ve URI (Universal Resource Indicator) adresine sahip Internet kaynağı olarak dar anlamıyla tanımlanmıştır.

Internet üzerinden sağlanan uygulamalardan üçünü, elektronik postayı (e-posta), dosya transfer protokolünü (file transfer protocol) ve uzaktan bağlanmayı (remote login veya telnet) temel hizmetler bölümünde sınıflamak en azından tarihsel olarak yanlış bir yaklaşım olmayacaktır. Dahası, e-posta uygulamasını bilgi toplumuna giden zaman yolculuğunun başlangıç noktası olarak niteleyebiliriz.[2] E-posta insanların birbirleriyle iletişimine, etkileşimine ve yardımlaşmasına yeni bir model getirmiştir. Dosya transfer protokolü günümüzde de çok sık olarak kullanılmakta ve esas gücünü uzaktan bağlanma uygulamasından almaktadır. Söz konusu iki uygulama bilgisayar ağı aracılığıyla uzaktan araştırmanın ilk çekirdeğini oluşturmuştur.

Internet kavramının oluşturulmasına temel olan USENET ve BITNET (Because It's Time NETwork) uygulamalarından da kısaca söz etmekte yarar görüyoruz. Dünya çapında gönüllü üyeliğe dayalı bir ağ olan ve UUCP (Unix-to-Unix Copy Protocol) protokolü üzerine temellendirilen USENET, Unix işletim sistemini kullanan bilgisayarlar arasında e-posta ve e-postaya dayalı elektronik tartışma listesi hizmetleri için kullanılmaktaydı. Öte yandan, IBM bilgisayarları arasında verilen e-posta hizmetleri için ise sakla-ilet (store-and-forward) protokolüne göre çalışan BITNET kullanılmaktaydı (Bollmann-Sdorra ve Raghavan, 1993). BITNET ve USENET, Internet teknolojisinin parçaları olmamalarına rağmen, bu ağlar aracılığıyla oluşturulan tartışma/haber grupları ve kapalı listeler bugünkü çağdaş bilgi toplumunun oluşmasına önemli katkılarda bulunmuştur.

Internet üzerindeki bilgi kaynaklarının dizinlenmesinin ilk örneğini Archie oluşturur (Frank, 1996). Archie hizmeti orijinal olarak Internet üzerindeki kamuya açık (anonim) FTP arşivlerinde bulunan dosya adlarının taranabilir bir veri tabanı olarak başladı (Tennant, Ober ve Lipow, 1996). Archie yazılımı FTP sitelerini periyodik olarak dolaşarak var olan dosyaları isimleri üzerinden dizinleyerek aranabilir (ya da taranabilir) hale getirmişti.[3] Kullanıcılar archie sunucularına telnet ile bağlanıp (veya bu sunuculara e-posta gönderip) aradıkları dosya ya da program adlarını girerek ilgili dosya ya da programın kamuya açık onbinlerce bilgisayardan hangisi/hangileri üzerinde olduğunu kolayca saptayabilme ve ilgili dosyayı FTP protokolü kullanarak kendi bilgisayarlarına kopyalayabilme olanağına kavuştular (Deutsch, 1992). Archie, aradıkları dosyanın adını bilen kullanıcılar için kamuya açık FTP arşivlerini taramada kullanılan yararlı bir yazılımdı. Ancak dizinlenen dosya adları bazen içerik hakkında çok fazla bilgi içermeyebiliyordu. Dahası, hemen hemen her FTP sitesinde

[2] E-postayla ilgili RFC (Request for Comment) 1969'da yayımlanmıştır.

[3] Archie, Unix işletim sisteminde satır komutu olarak kullanılmaktaydı. Archie yazılımına olan yatırımın durması (sunucu desteğinin olmaması ve istemci üzerinde koşullandırılmaması) nedeniyle günümüzde Archie'in kullanımı artık pratik olarak mümkün değildir.

rastlanabilen yazılımlar ya da yaygın olarak kullanılmasından dolayı çok fazla anlam taşımayan dosya adları (örneğin, "readme.txt") için arama yapıldığında aramalar uzun zaman alabiliyordu.

Daha sonra mönü tabanlı bir sistem olan "gopher" ortaya çıktı. Gopher, Minnesota Üniversitesi Bilgi İşlem Birimi tarafından yerleşke bilgi sistemi (campus-wide information system) hedeflenerek geliştirildi. Gopher'i popüler yapan özellikleri onun mönü tabanlı olması değil, sunucu-istemci mimarisinde geliştirilmesi ve işletim sisteminden ve platformdan bağımsız olarak konuşlandırılmasıdır. Her bir gopher mönü tabanlı bir Internet istemcisidir. Gopher uzayını birbirleri ile döngüsel veya döngüsüz bağlantılı metin ve grafik türündeki bilgi kaynakları oluşturur. Gopher uzayının giderek genişlemesi bu uzayda yer alan bilgi kaynaklarının dizinlenmesi sorununu da beraberinde getirdi. Bu sorunun adreslenmesi VERONICA (Very Easy Rodent-Oriented Net-wide Index to Computerized Archives)[4] ile olmuştur. Nevada Üniversitesi tarafından geliştirilen VERONICA, dünyaya yayılmış binlerce Gopher mönüsünde geçen anahtar sözcükleri içeren bir veri tabanıdır. Gopher kullanıcıları gopher mönülerinde geçen anahtar sözcükleri VERONICA veri tabanından belirli bir sorgu kullanarak arayabilirler. VERONICA, ilgili anahtar sözcük ya da sözcüklerin hangi gopher sunucularında geçtiğini bularak kullanıcıların bilgi ihtiyacını karşılamayı amaçlayan bir sistemdir (Tennant et al., 1996). Bir başka deyişle, kullanıcılar Archie ile sadece dosya adlarını kullanarak kamuya açık FTP arşivlerinde arama yapabilirken, VERONICA ile gopher mönülerinde geçen herhangi bir sözcük ile arama yapabilmektedirler. Mönü seçenekleri genellikle birden fazla sözcük içerdiğinden kullanıcıların aradıkları bilgiye erişme olasılıkları daha fazladır.

1989 yılında geliştirilen WAIS (Wide Area Information Server), metin dosyalarını içerik olarak dizinleyip bunlar üzerinden sorgulamaya imkân veren bir sunucu-istemci sistemidir (Frank, 1996). İstemcilerin arama isteklerini alan WAIS sunucuları veri tabanlarında arama yapar ve sonuçları gönderirler. WAIS'in Archie ve VERONICA'dan farklı birkaç önemli özelliği bulunmaktadır. WAIS, bir belgede geçen tüm sözcükleri dizinlemekte, hem Boole işleçleri hem de doğal dille arama yapılmasına olanak sağlamakta, arama sonuçlarını belirli ölçütlere göre sıralayabilmekte ve ilgililik geribildirimi (relevance feedback) özelliği sayesinde kullanıcı tarafından ilgili bulunan bir belgeye benzeyen diğer belgeleri bulabilmektedir (Tennant et al., 1996).

[4] FTP için Archie ne ise Gopher için Veronica odur. 'Archive' sesini veren Archie aslında ülkemizde de yayımlanan bir çizgi romanın komik karakteridir ve Veronica da onun kız arkadaşıdır. Archie yaratıcılarına gönderme yapmak için Veronica isminin seçildiği bilinmektedir.

Archie, VERONICA ve WAIS'in günümüzde kullanımı kısıtlı olmasına rağmen, bu uygulamalar, sayısı hızla artan Internet kaynaklarına erişim sorununu ilk olarak gündeme getiren uygulamalardır. Kısacası, Archie, VERONICA ve WAIS etrafında oluşturulan çalışmalar günümüz arama motorlarına giden serüvenin Internet üzerindeki ilk çalışmalarıdır.

Günümüzde e-postadan sonra en sık kullanılan Internet aracı olan WWW (World Wide Web) (Berners-Lee, Cailliau, Groff ve Pollermann, 1992) ise, 1989 yılında Cenevre'deki Avrupa Parçacık Fiziği Laboratuvarı'nda (CERN) geliştirilmeye başlanmıştır. WWW, 1992 yılında Internet üzerinde kullanılmaya başlandığı dönemlerde Internet tarihinde bir devrim olarak nitelendirilmiştir (Kredel, Meuer, Schumacher ve Strohmaier, 2000). WWW'nin en önemli işlevi, Web'e bir standart getirmiş olması ve daha önce geliştirilen protokolleri (telnet, ftp, gopher, vd.) tanımasıdır. WWW'i kaba hatlarıyla, HTTP'yi (Hyper-Text Transfer Protocol) kullanan Internet üzerindeki bütün kaynaklar ve kullanıcılar olarak tanımlayabiliriz. WWW'i geliştiren ve W3C'nin (World Wide Web Consortium) kurucularından birisi olan Tim Berners-Lee, Internet'i ağ aracılığıyla erişilebilir (network-accessible) bilgi uzayı olarak nitelendirmiştir (Berners-Lee, Cailliau, Luotonen, Nielsen ve Arthur Secret, 1994). Bu bakış acısından yola çıkacak olursak, artık Internet ile eş anlamlı hale gelen WWW, res sistemi (Uniform Resource Locator (URL)), ağ protokolü (HTTP) ve hiper-metin işaretleme dilinden (Hyper-Text Markup Language (HTML)) oluşan bir yapıdır diye tanımlanabilir.

WWW kolay kullanılan arayüzü ve çoklu ortam özellikleri sayesinde çok sayıda kullanıcının ilgi odağı olmuş ve bu sayede çok geniş dağıtık bir bilgi kaynağı durumuna gelerek kişisel Web sayfalarını, çevrimiçi (online) sayısal kütüphanelerini, sanal müzelerini, ürün ve servis kataloglarını, halka açık hükümet bilgilerini, araştırma yayınlarını içerecek şekilde ve aynı zamanda FTP, Gopher, ve e-posta gibi farklı Internet hizmetlerine olanak sağlayarak çok hızlı bir şekilde büyümüştür (Gudivada, Raghavan, Grosky ve Kasanagottu, 1997). Web ve Internet'in büyümesi üç boyutta incelenebilir: Kullanıcı sayısı, Internet'e bağlı ağ (host site) sayısı ve adreslenebilir Web sayfası sayısıdır. Web'in uluslararası kullanımı hakkındaki veriler NUA Internet araştırma sayfasında yayınlanmaktadır (http://www.nua.com/surveys/). Buna göre Internet kullanıcı sayısı en azından 419 milyon civarındadır. Internet'teki host sayısı ise, netsizer şirketinin elde ettiği istatistiğe göre şu an 120 milyon civarındadır (http://www.netsizer.com/index.html).[5] Inktomi Corp. ve NEC

[5] Internet'in büyümesi üzerine verilen rakamlar kaynaklar arasında farklılık göstermesine rağmen, "host", sayfa ve kullanıcı sayılarındaki ikinin katları şeklindeki üssel (exponential) büyüme oranı hemen hemen hepsi tarafından doğrulanmaktadır (Kobayashi ve Takeda, 2000). Host sayısındaki derlemeyi bunun ışığı altında incelediğimizde, iki kaynak arasında yapılan varsayımlar cinsinden önemli farklılıklar gözükmemektedir.

Araştırma Enstitüsünün 2000 Ocak ayında yapmış olduğu açıklamada Web üzerinde 1 milyar üzerinde belge (sayfa) bulunduğu duyurulmuştur (Inktomi Corp., 2000).[6] İlgili rakamlar ve onların yıllara dağılımı) çeşitli kaynaklarca farklı olarak belirtilse bile, host/kullanıcı/sayfa büyüme oranları ölçümünde uygunluk olduğu gözlenmiştir: host ve Web sayfa sayıları her yıl ikiye katlanmaktadır (Kobayashi ve Takeda, 2000). Daha ilginç olanı ise Web üzerindeki bilgi hacminin 31 Ağustos 1998 tarihi itibarıyla 3 katrilyon sekizli (tera byte) olduğu[7] ve büyüme oranının ise her sekiz ayda bir ikiye katlandığıdır.

Yukarıda verilen tablo, WWW üzerindeki bilgilere ulaşmak için arama motorlarına olan ihtiyacı açıkça kanıtlamaktadır. Bugün, bilgiyi arayabilmek Internet yaşamının önemli bir parçası olduğundan dolayı yeni ve daha güçlü arama motorları her gün geliştirilmektedir (Jansen, 1996; Adalı, Bufi ve Temtanapat, 1997). Dünya genelinde çok geniş kullanım alanı olan AltaVista, Yahoo, Google, Excite, Lycos, HotBot, Northern Light, MSN Search (PC Computing, 1996) vb. gibi arama motorlarını değerlendirmek için yöntemler önermek ve arama motorlarının performanslarını incelemek üzere birtakım çalışmalar yapılmıştır (Lawrence ve Giles, 1998; Sullivan, 2000: 11). Ülkemizde de son zamanlarda özellikle popüler yayınlarda arama motorlarıyla ilgili bazı tanıtıcı yazılara rastlanmaktadır. Ancak akademik yönden arama motorlarının araştırmacılarımızın ilgi alanına girmesi nispeten daha yenidir. AltaVista, Excite, HotBot, Infoseek ve Northern Light adlı arama motorlarının performanslarının değerlendirildiği çalışma bu alanda ülkemizde yapılan ilk çalışmalardan birisidir (Soydal, 2000). Benzer çalışmaların son yıllarda büyük gelişme gösteren Türkçe arama motorları hakkında da yapılması gerektiği açıktır. Nitekim bu yönde bazı çabalar gösterilmektedir (Aslantürk, 2000). Bu çalışmada, ülkemizde yaygınlıkla kullanılan belli başlı Türkçe arama motorlarından Arabul, Arama, Netbul ve Superonline incelenmiş ve bu motorların bilgi erişim performansları çeşitli ölçütlere göre test edilip değerlendirilmiştir. Araştırma raporunun düzeni aşağıda kısaca tanıtılmaktadır.

Çalışmanın ilk bölümünde Internet ve World Wide Web'in gelişmesi hakkında kısa bilgiler verilmiştir.

---

[6] Web kaynaklarını birbirini dışlayan iki kategoride, derin ve yüzey Web, sınıflayalım. Derin Web, Web üzerinde bulunan ve arama motorlarının dizinlerinde yer almayan belgelerin bulunduğu kısım; yüzey Web ise, Web üzerinde bulunan ve arama motorlarının dizinlerinde yer alan belgelerin bulunduğu kısım olsun. 2000 Temmuz'da BrightPlanet şirketi tarafından yapılan inceleme sonucunda oluşturulan yayında, derin Web üzerindeki belge miktarının, yüzey Web üzerindeki belge miktarından 500 kat daha fazla olduğu açıklanmıştır (Bergman, 2001). Ayrıca BrightPlanet şirketinin incelemelerinde yer alan bir nokta da, her gün yüzey Web'deki belge sayısının 1.5 milyon arttığıdır (Bergman, 2001). Bu incelemeler göz önünde bulundurularak, 2001 yılının başlarında yüzey Web üzerinde bulunan belge sayısının 1.5 milyarın üzerinde, derin Web üzerinde bulunan belge sayısının da 750 milyarın üzerinde olduğu söylenebilir.

[7] Bu varsayım, Kobayashi ve Takeda (2000) tarafından "Alexa Internet" (http://www.alexa.com/) kaynağına dayanılarak verilmiştir.

İkinci bölümde bilgi erişim sistemlerinin temel bileşenleri (dizin terimleri, belgeler, sorgular ve erişim fonksiyonları) ve belli başlı bilgi erişim performans değerlendirme ölçütleri ("anma", "duyarlık", "normalize sıralama", "kapsama" ve "yenilik" oranları) gözden geçirilmiştir.[8]

Çalışmanın üçüncü bölümünde arama motorlarının mimari yapıları, dizinleme ve belgeleri gösterme özellikleri, erişim için kullandıkları fonksiyonlar ile arama motorlarında performans değerlendirme konusunda yapılan belli başlı çalışmalar incelenmiştir.

Dördüncü bölümde araştırmamızın tasarımı ve yöntemi açıklanmıştır. Arama motorları hakkında yanıtlamaya çalıştığımız araştırma soruları, deney için kullanılan arama motorları ve bu motorların özellikleri, arama motorlarına yöneltilen sorular, aramaların yapılması, arama motorlarının performanslarının ölçümleri ve verilerin analiziyle ilgili bilgiler bu bölümde verilmiştir.

Beşinci bölümde araştırmanın sonuçları ayrıntılı olarak verilmiştir. Bu bölümde, Arabul, Arama, Netbul ve Superonline'ın:

a) eriştikleri belgelerdeki "ölü" bağlantı oranları;
b) 17 farklı türdeki soru için çeşitli kesme noktalarında kaydettikleri duyarlık ve normalize sıralama oranları;
c) Türkçe arama motorlarında en sık aranan sözcüklerle ilgili belgeleri kapsama oranları ve bu sözcüklere karşılık eriştikleri belgelerin yenilik oranları; ve
d) belgeleri dizinlemek amacıyla "anahtar sözcük", "tanım" gibi HTML üst veri (metadata) alanlarından yararlanıp yararlanmadıkları ile ilgili iki küçük deneyin sonuçları

ile ilgili bulgular verilmiş ve dört arama motorunun performansları birbiriyle karşılaştırılmıştır.

Altıncı ve son bölümde araştırmamızın sonuçları kısaca özetlenmiş ve arama motorlarının performanslarının artırılmasıyla ilgili çeşitli önerilere yer verilmiştir.

Çalışmada yararlanılan kaynaklar Kaynakça'da listelenmiştir.

[8] "Anma (recall) değeri erişilen ilgili belge sayısının derlemdeki toplam (hem erişilen hem erişilemeyen) ilgili belge sayısına oranıdır." "Duyarlık (precision) erişilen ilgili belge sayısının erişilen toplam belge sayısına oranıdır"(Van Rijsbergen, 1979). Bu terimler Türkçede ilk kez bildiğimiz kadarıyla Aydın Köksal (1979, 1987) tarafından kullanılmıştır. Kütüphanecilik literatüründe "duyarlık" için "kesin isabet", "anma" için "erişim isabeti" terimleri de kullanılmaktadır (Tonta, 1995). Anma ve duyarlık değerleriyle ilgili daha ayrıntılı bilgi aşağıda (2.5) verilmektedir.

2 BİLGİ ERİŞİM SİSTEMLERİ

Bir bilgi erişim sisteminin temel işlevi, kullanıcıların bilgi ihtiyaçlarını karşılaması muhtemel derlemdeki ilgili (relevant) belgelerin tümüne erişmek, ilgili olmayanları da ayıklamaktır. Bir bilgi erişim sisteminin bazı belgelere erişim sağlayabilmesi için iki koşul yerine getirilmelidir. İlki, derleme eklenen her belgenin temel özellikleri geleneksel veya otomatik olarak gerçekleştirilen dizinleme işlemleri sırasında belirlenmeli ve her belge için ilgili içerik belirteçleri (dizin terimleri) oluşturulmalıdır. Bir belge için oluşturulan söz konusu içerik belirteçleri bilgi erişim sırasında belgenin tamamını temsil etmek üzere (surrogates) kullanılır. İkincisi, kullanıcılar belgelere verilen bu içerik belirteçlerini doğru olarak tahmin edip sorgu cümlelerini ona göre oluşturmalıdırlar. Bir başka deyişle, kullanıcının bilgi ihtiyacını ifade etmek için kullandığı terimlerle belgeyi temsil eden içerik belirteçleri birbiriyle karşılaştırılır ve çakışan belgelere erişilir (Tonta, 1995, 1992). Çakışma "Erişim Kuralı" (Retrieval Rule) olarak adlandırılan kuralı izler. Maron (1984, s.155) bu kuralı şöyle açıklamaktadır: "Herhangi bir resmi (formel) sorgu [cümlesi] için bu arama sorgusunda belirlenen tutanakların (records) alt setinde yer alan dizin tutanaklarının tümüne ve salt bu dizin tutanaklarına erişim sağla." Böylece, bir bilgi erişim sisteminin temel bileşenlerinin: (1) bir belge derlemi (ya da bu belgeleri temsil eden içerik belirteçlerini içeren tutanaklar), (2) kullanıcıların sorgu cümleleri, ve (3) kullanıcıların sorgu cümlelerinde yer alan terimlerle derlemdeki belgelere verilen terimleri karşılaştırarak ilgili belgeleri belirlemek için kullanılan bir erişim kuralından oluştuğu ortaya çıkmaktadır.

Şekil 1'deki işlevsel mimaride de görüleceği üzere, sistemi oluşturan temel bilgi erişim süreçlerini üçer tane ön yüz (front-end) ve arka yüz (back-end) kavramları çerçevesinde tanımlamak mümkündür. Bu şekilde kavramlar dikdörtgen, temel süreçler oval, seçenekli süreçler ise kesikli oval şekillerle gösterilmiştir. Ön yüz kavramları sistemin dış dünyaya yansıyan görünüşünü oluşturmaktadır. Benzer şekilde arka yüz kavramları kullanıcıya saydam olup bilgi erişim süreçleri arasındaki iletişimde kullanılır. Bilgi ihtiyacı, metin nesneleri ve erişim çıktısı ön yüz, sorgular, belgeler ve içerik belirteçleri arka yüz kavramlarını oluşturur.

Bilgi ihtiyacı bir düz metinle (doğal dille) ifade edilebileceği gibi dizin terimleri ve aralarındaki ilişkiler ("ve", "veya", "ve-değil", "ise/eğer", vb.) çerçevesinde de tanımlanabilir. Metin nesneleri arka planda işleyen otomatik dizinleme sürecine giriş oluşturur ve sonuçta belgeler ters dizin kütüğü (inverted file) düzenlemesi içinde içerik belirteçleri ile öznel (subjektif) olarak gösterilirler. Buradaki öznellik metin nesnelerinin içerik belirteçleri ile

gösteriminin ileride de göreceğimiz üzere çeşitlilik göstermesidir.[1] Bunun aksini ise metin yazarı, adı, yayıncı bilgisi, yayın tarihi, türü, gibi nesnel (objektif nitelikler) oluşturur.[2] Erişim çıktısı eldeki sorgu ifadesinin belgeler (ve/veya onların öznel/nesnel nitelikleri) ile eşleştirilmesiyle oluşturulurlar; yani sistemin, belge derlemi (koleksiyonu) içinde sunulan sorgu ifadesi ile ilgili olduğunu "düşündüğü" belgeleri topladığı havuza (formel anlamıyla "küme"ye) erişim çıktısı adını vermekteyiz. Erişim çıktısındaki belgeler kullanıcı bilgi ihtiyacına yakınlık derecesine göre azalan sırada sıralanırlar.[3]

Arka yüz kavramları aslında üç temel sonlu nesne küme notasyonuna karşılık gelirler. Bunlar sırasıyla *belgeler*, *içerik belirteçleri* (*anahtar sözcükler*, *dizin terimleri*[4]*)* ve *sorgulardır*. Kullanılan model ne olursa olsun, sorgular mutlaka belgeler (ya da belgeleri temsil eden içerik belirteçleri) ile eşleştirilmelidir -ki bu eşleştirmeye erişim kuralı (ya da erişim işlevi) denir. Şekil 1'de kümeleme (clustering) süreci bir anlamda aşırı yüklenmiştir. Sorguları, belgeleri ve içerik belirteçlerini tek tek özyineli (recursive) olarak temel alan kümeleme süreçleri, aynı ad ile anılmalarına rağmen amaçları ve/veya uygulanan teknikler açısından birbirlerinden farklılık gösterebilirler.[5] Şöyle ki, içerik belirteçleri eş anlamlılık temelinde kümelendirildiklerinde amaç sorgu genişletebilme ve yerden kazanç sağlama (metin nesnelerinin daha az sayıdaki belirteçler ile gösterilmesi) olmasına rağmen, belgelerin kümelendirilmesinde amaç eşleştirme sürecinin hızlandırılmasıdır. Sorguların kümelendirilmesinde ise, zaman açısından pahalı bir süreç olan geribildirim sürecine olan ihtiyacı azaltma ya da geribildirim sürecini kısa zamanda sonuçlandırma kaygısı olabileceği gibi (Mettrop ve Nieuwenhuysen, 2001)[6], performans etkinliği daha yüksek olan bilgi erişim sistemleri gerçekleştirme hedefi de güdülebilir (Lee, 1995; Belkin, Kantor, Fox ve Shaw,

---

[1] Ne şekilde dizinleme yapılırsa yapılsın, ilgili süreç sonucunda elde edilen gösterim (içerik belirteçleri kümesi) özneldir. Başka bir deyişle, bir belgenin birden fazla (ve doğru) gösterim şekli olabilir. Dizinleme işleminin elle ya da otomatik olarak yapılması bu gerçeği değiştirmez.

[2] Kütüphanecilikte bir bilgi kaynağıyla ilgili nesnel niteliklerin (yazar adı, başlık vs.) belirlenmesine "tanımlayıcı kataloglama", kaynağın hangi konu ya da konular hakkında olduğunun belirlenmesine ise "konu kataloglaması" adı verilmektedir.

[3] Başka bir deyişle, erişim çıktısı erişim fonksiyonunun değişimini oluşturan sıralı belgeler kümesidir.

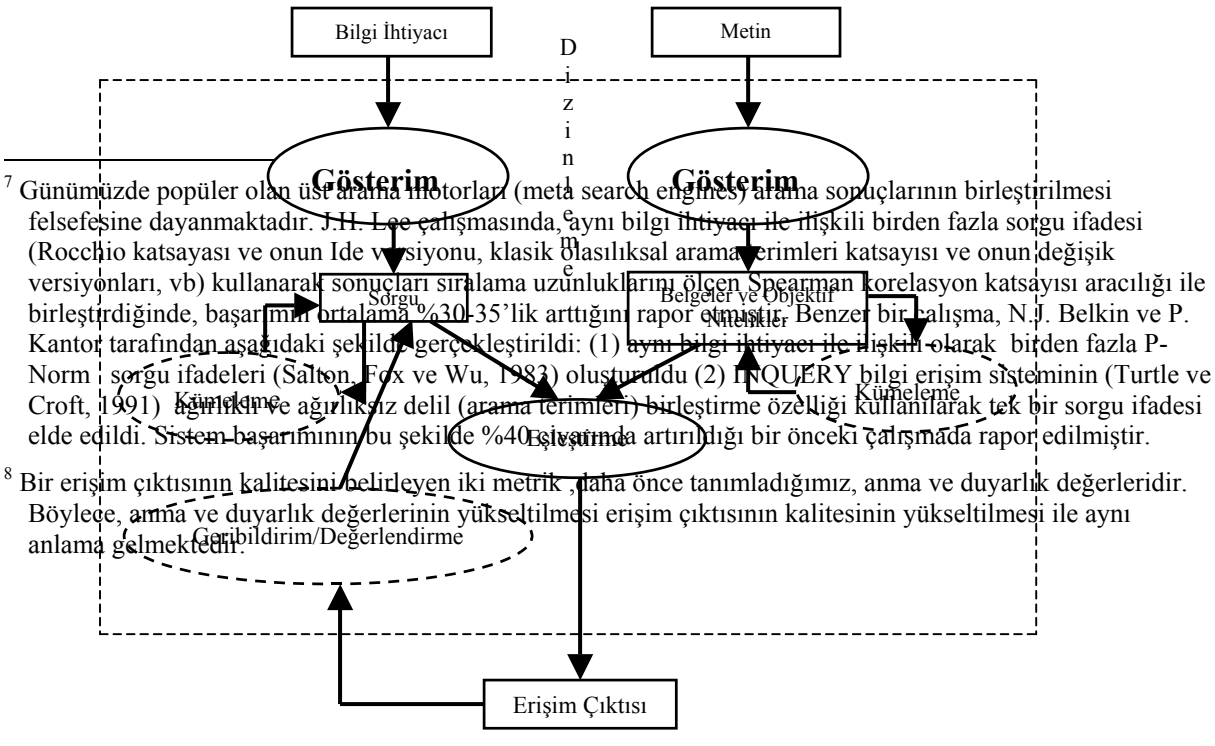
[4] "İçerik belirteçleri", "dizin terimleri" ve "anahtar sözcükler" makale boyunca eş anlamlı olarak kullanılmaktadır.

[5] İleride de belirtileceği üzere erişim çıktısındaki belgelerin kümelendirilmesi arama motorlarında kullanıcı arayüzü tasarımının bir parçası olarak önem kazanmıştır (Leuski, 2001). Belgeler tek tek ilgililik derecesine göre kullanıcıya sunulmaz, bunun yerine genellikle iki veya daha fazla belgeden oluşan öbekler halinde kullanıcıya sunulur. Google arama motoru (www.google.com) olaya benzer bir perspektiften bakarak içerik olarak aynı olan fakat farklı site adreslerine sahip belgeleri eleme amacıyla erişim çıktısını arka planda kümeleme tekniği uygulamaktadır.

[6] Sorguları kümelendirmede kullanılan ve ikili tercih ilişkisi tabanlı basamak inme algoritması (steepest descent algorithm) bilgi süzgeçleme alanına da başarıyla uygulanmıştır (Mettrop ve Nieuwenhuysen, 2001).

1995).[7] İçerik belirteçlerinin kümelendirilmesinde LSA (Latent Semantic Analysis) tekniği (Deerwester, Dumais, Furnas, Landauer ve Harshman, 1990; Foltz, 1996), belirteçlerin ayrım gücünü temel alan sıradüzensel (hiyerarşik) (Van Rijsbergen, 1979) veya düz kümeleme teknikleri kullanılabilir (Salton, 1989; Salton, Wong ve Yu, 1976; Sezer, 1999). Oysaki, sorguların kümelenmesinde sorgularla ilgili belgelerin kesişim derecesi temel alınılır.

Şekil 1'de görüldüğü üzere, tipik bir bilgi erişim sistemi geribildirim özelliğine sahiptir. Sistem tarafından döndürülen belge çıktısının kullanıcının bilgi ihtiyacını karşılamaktan uzak olduğu durumlarda, kullanıcı geribildirim sürecini başlatarak daha kaliteli bir belge çıktısı[8] elde etmek isteyebilir. İleride değinileceği üzere, tipik bir geribildirim sürecinde, hata (herhangi bir belgenin eldeki bilgi ihtiyacı ile ilgili olması bağlamında, sistem kararının kullanıcı görüşü ile örtüşmemesi) oranının tekrarlı ve etkileşimli bir süreç boyunca kullanıcının tatmin olabileceği bir düzeye indirgenmesi hedeflenir (Salton ve Buckley, 1990).



[7] Günümüzde popüler olan üst arama motorları (meta search engines) arama sonuçlarının birleştirilmesi felsefesine dayanmaktadır. J.H. Lee çalışmasında, aynı bilgi ihtiyacı ile ilişkili birden fazla sorgu ifadesi (Rocchio katsayası ve onun Ide versiyonu, klasik olasılıksal arama terimleri katsayısı ve onun değişik versiyonları, vb) kullanarak sonuçları sıralama uzunluklarını ölçen Spearman korelasyon katsayısı aracılığı ile birleştirdiğinde, başarımın ortalama %30-35'lik arttığını rapor etmiştir. Benzer bir çalışma, N.J. Belkin ve P. Kantor tarafından aşağıdaki şekilde gerçekleştirildi: (1) aynı bilgi ihtiyacı ile ilişkili olarak birden fazla P-Norm sorgu ifadeleri (Salton, Fox ve Wu, 1983) oluşturuldu (2) INQUERY bilgi erişim sisteminin (Turtle ve Croft, 1991) ağırlıklı ve ağırlıksız delil (arama terimleri) birleştirme özelliği kullanılarak tek bir sorgu ifadesi elde edildi. Sistem başarımının bu şekilde %40 civarında artırıldığı bir önceki çalışmada rapor edilmiştir.

[8] Bir erişim çıktısının kalitesini belirleyen iki metrik, daha önce tanımladığımız, anma ve duyarlık değerleridir. Böylece, anma ve duyarlık değerlerinin yükseltilmesi erişim çıktısının kalitesinin yükseltilmesi ile aynı anlama gelmektedir.

Şekil 1. Bir bilgi erişim sisteminin işlevsel mimarisi

Bu bölumde arka yüz kavramlarını teşkil eden belgeler, içerik belirteçleri ve sorgular üç alt başlık halinde incelenmekte, eşleştirme süreci (ya da daha bilinen adıyla erişim fonksiyonları) ve etkinlik ölçümleri tartışılmaktadır.

2.1 İçerik Belirteçleri

İçerik belirteci bir belgenin veya bilgi ihtiyacının gösterimi (temsil edilmesi) için kullanılır. Rasgele işlenen metinler üzerinden aynı alan (domain) içinde olsalar bile ortak yapı kalıpları elde edebilmek çoğunlukla mümkün değildir. Zaten işlenen veri (örneğin, metin) üzerinde bir yapı empoze edilebiliyorsa, bu süreç, doğasına göre veri tabanı veya uzman modeller aracılığı ile daha etkin olarak herhangi bir bilgi erişim modeline gerek kalmaksızın modellenebilir. Ele alınan bir metin, (bütünlük arz eden) bir bilgi taşıdığı için, buradaki kritik soru bu metnin veya belgenin içerik açısından nasıl temsil edileceğidir. Çünkü gerek duyulduğunda bu belgeye erişilebilmelidir. Başka bir deyişle, belge işleme sürecine yakından bakıldığında, belgeleri çoğu zaman sisteme sunuldukları halleri ile değil, belgelerin içeriğini yansıtan belirteç kümesi (surrogate record) halinde kullanma zorunluluğu görülür. Bu içerik belirteçlerine anahtar sözcük, üst veri (metadata), dizin terimi, tanımlayıcı, veya kısaca terim gibi adlar verilir.

1950'lerin sonunda bir metnin konusunu belirten sözcükleri (metindeki geçiş sıklıklarına dayanarak) belirlemeye yarayan bir program geliştiren Hans Peter Luhn, anahtar sözcüklerle dizinleme ve arama yapmanın modern "kaşif"i olarak bilinmektedir. Luhn ilk kez bir makale adında geçen sözcükleri, her bir sözcüğün basılı dizinlerde "giriş" (entry) olarak yer almasını algoritmik olarak sağlayan bilgisayarla dizinlemeyi geliştirmiştir. KWIC (Key-Word-in-Context) olarak bilinen bu dizinleme türü bibliyografik dizinlerin hazırlanmasında halen kullanılmaktadır (Svenonius, 2000, s. 28, 44, 190).

Her bir dizin terimi belgelerin içeriğini çoğu zaman bütünüyle değil, ancak bir yönüyle ifade eder ve bir belge için bir çok dizin terimi seçilir. Verilen bir belge için dizin terimlerinin seçilmesi sürecine *dizinleme* adı verilir. Dizinleme süreci kontrollu veya kontrol edilmeyen bir terim sözlüğü (vocabulary) üzerinden elle (manual) ya da otomatik olarak gerçekleştirilebilir. Kontrollü dizinlemede bir belgeyi temsil edecek terimlerin seçimi belli bir konu sözlüğü temel alınarak konu uzmanlarınca yapılır. Bu tarz ile yüksek bir biçimdeşlik (uniformity) ve kalite elde etmek mümkündür; fakat dizinlemenin yavaş ve maliyetli olması ve en önemlisi kullanıcının sorgu ifade etmede kullanacağı kelime dağarcığının kontrollü sözlükle çakışması gereksinimi kontrollü dizinlemenin dezavantajlarından birisidir.[9] Kullanıcılar konu uzmanları tarafından kullanılan kontrollü sözlüklerle (örneğin, Kongre Kütüphanesi Konu Başlıkları Listesi) aşina değildirler (Tonta, 1990). Dahası, konu uzmanları tarafından aynı kontrollü sözlüğün kullanıldığı durumlarda bile dizinleme tutarlılığı (indexing consistency) son derece düşük olmaktadır (Tonta, 1991). Yapılan deneysel araştırmalar otomatik dizinlemenin kontrollü dizinleme ile elde edilen performansı yakaladığını göstermiştir (Van Rijsbergen, 1979; Salton, 1989).

Metin yazarının sözcük dağarcığı ile bir bilgi erişim sistemi kullanıcısının sözcük dağarcığı arasındaki farka *sözcük dağarcığı farkı* diyelim. Kullanımı daha pratik ve hızlı olan otomatik dizinleme sözcük dağarcığı farkının açılmasına yol açar. Sözcük dağarcıkları arasındaki fark, belge derlemindeki bir belgenin verilen bir sorgu ifadesi ile ilgili olmasına (ya da daha da kısıtlayacak olursak her ikisinin de aynı kavrama karşılık geldiğini

[9] Aslına bakılırsa, bu sorun, belgelerin tam metinlerinde geçen her sözcüğün dizinlendiği, kontrol edilmeyen bir terim sözlüğü kullanıldığı zaman da tam olarak ortadan kalkmamaktadır (Tonta, 1995). Kullanıcıların bilgi ihtiyaçlarını ifade etmek için kullandıkları terimlerle ilgili belgelerin tam metinlerinde geçen terimler arasındaki çakışma sorgu cümlelerinde geçen terim sayısı arttıkça hızla düşmektedir. Kullanıcılar, hakkında bilgi bulmak istedikleri konuları sorgu cümlelerinde iyi tanımlayamamaktadırlar. Zaten tam olarak ne aradıklarını tanımlayabilselerdi belki de ilgili bilgi erişim sistemini kullanmalarına gerek kalmayacaktı. Bilgi erişimin temel paradoksu "hakkında bilgi bulmak için bilmediğin bir şeyi tanımlama gereği"dir. Bu paradoks, bir bakıma, "sözlük" sözcüğünün anlamını bilmeyen (ve yakınında sorabileceği birisi olmayan) bir kimsenin çaresizliğine benzetilebilir (Blair ve Maron, 1985).

varsaymamıza) rağmen, söz konusu belgenin çoğunlukla elimizdeki sorgu ifadesi ile eşleşmemesine (ya da eşleştirme derecesinin düşük olmasına) yol açar.

Bilgi erişim sistemlerinde dağarcık farkını kapatmak için kullanılan araçlardan birisi de *gömü*lerdir (thesauri). Tipik bir bilgi erişim sistemi için gömü, terimlerin belli bir ilişkiye göre düzenlenmesidir (Srinivassan, 1992). Gömü, dizinleme ve erişim hizmetlerinde terimlerin kullanımına rehberlik eder. Bilgi erişim sisteminin sorgu işleme sürecinde yardımcı yapı olarak, satırda belgeleri ve sütunda (dizin) terimleri tuttuğunu varsayalım (ki bu uygulama oldukça sık kullanılan ve "ters dizin" kütüğü olarak adlandırılan bir yapıdır). O zaman, eş anlamlılar ilişkisinin karşılıklı dışlayan kümeler olduğunu varsayarak, gömünun dizinlemede kullanımı sütunların belirli bir ad altında birleştirilmesinden ibarettir. Erişimde ise, eş anlamlılar ilişkisinin kesişen kümeler olduğunu varsayarak, gömü sorgu genişletmeye karşılık gelir. Eş anlamlı olan her bir terim için verilen bir sorgu belge uzayına karşı eşleştirilir.

Gömüler, elle ve otomatik olmak üzere iki türlü üretilirler. Elle gömü üretimi insan emeği ile terimler arasında önceden ilişki (eş/zıt anlamlı ilişkiler, dar/geniş terimler, vd.) kurulmasına ve bu ilişkilerin gömü oluşturmak için kullanılmasına dayanır. Gömü sıradüzensel (hiyerarşik) ilişkiler şeklinde de oluşturulabilir. Örneğin, '*A*' ve '*B*' herhangi iki eş anlamlı küme olsun. '*A*' terimi '*B*' teriminden daha dar (narrower) anlamlıdır, ya da '*B*' terimi '*A*' teriminden daha geniş (broader) anlamlıdır demek, matematiksel olarak '*A*'nın '*B*'nin alt kümesi (ya da tersi) olduğunu belirtmektir. Bu tür tek yönlü ilişkiler birbirleriyle aşağıdan (dar anlamlıdan) yukarı (geniş anlamlılar) doğru bağlandıklarında sıradüzensel ilişki oluşturulur. Sıradüzensel ilişkiler gömü işlevinin yanı sıra, sorgu sonuçlarını süzmede de kullanılırlar. Bu tür ilişkilerle oluşturulan bilgi erişim sistemleri kavram tabanlı sistemler olarak da adlandırılmaktadır (McCune, Tong, Dean ve Shapiro, 1985).[10]

Otomatik gömü üretimi teknikleri, terimlerin herhangi bir belgede birlikte geçme olasılıklarını temel alır. Herhangi iki terimin aynı gömü alt kümesi içerisinde yer alması onların anlamsal olarak eş anlamlı olduğu anlamına gelmez; ilgili iki terimin aynı küme içerisinde yer alması demek, yalnızca ve yalnızca sistemin verilen derlem bazında bu iki terimi istatistiksel olarak biribirinden ayırt edememesi demektir. Gömü yapısının bir bilgi erişim sisteminin etkinliğine (effectiveness) olan katkısı üzerinde yapılan çalışmalarda gömünün üretildiği derleme benzer derlemlerde kullanılması şartıyla anma değerinde

[10] Ali Alsaffar ve ötekilerinin çalışmaları kavram tabanlı sistemlerin bir sürekli (persistent) sıradüzensel ilişki tutmadan Boole (Alsaffar, Deogun, Raghavan ve Sever, 1999) veya vektör (Alsaffar, Deogun, Raghavan ve Sever, 2000) tabanlı sistemlerin üstüne etkin olarak nasıl kurulabileceği açısından ilginçtir.

%20'lere yaklaşan artışlar elde edilebildiği görülmüştür (Salton, 1989; Crouch ve Yang, 1992; Chen ve Lynch, 1992). Türkçede ise, alanı çok farklı konuları içeren küçük bir derlemde, çeşitli parametrelere göre üretilen gömüler için yapılan performans araştırmasında, gömü kullanımının erişilen ilgili belge sayısını artırmazken, ilgili belgeleri erişim çıktısında üst sıralara yerleştirdiği görülmüştür (Sezer, 1999).

2.2 Belgeler

Tipik bir bilgi erişim sisteminde belgeler terimler ile gösterilir. Verilen bir derlem bağlamında terim sözlüğü geleneksel olarak aşağıdaki gibi gerçekleştirilebilir: (1) harf olmayan karakterler boşluklarla yer değiştirilir; (2) tek harfli sözcükler silinir; (3) bütün karakterler küçük harfli yapılır; (4) durma listesinde adı geçen sözcükler silinir; (5) sözcükler gövdelenir (stemming); (6) tek karakterli gövdeler atılır. Son adım olarak, istenirse, (6). adımın sonunda elde edilen listedeki yüksek sıklıklı[11] sözcükler terim sözlüğünden çıkarılarak derleme duyarlı ikinci bir durma listesi oluşturulur. Ya da, yüksek sıklıklı sözcükler, otomatik eş anlamlı sözlük oluşturmanın bir parçası olarak, orta sıklıklı sözcüklerle birleştirilerek tamlama (phrase) oluştururlar.[12] Türkçe gibi sondan eklemeli (agglunative) dillerde gövdelemenin (bir sözcükten çekim eklerinin atılıp, yapım eklerinin korunması) bilgi erişim sistemi içindeki önemi yadsınamaz. Nitekim GÖVDEBUL algoritması (Duran, 1999) kullanılarak yapılan deneylerde anma ve duyarlık değerlerinde gövdeleme yapmaksızın yapılan sorgulara göre sırasıyla ortalama %20 ve %25 artış gözlenmiştir (Sezer, 1999). Bu deneylerde Türkçeye yerelleştirilen SMART sistemi kullanılmıştır (http://ata.cs.hun.edu.tr/~km/arsiv.html).

Otomatik olarak elde edilen gövdelenmiş sözcüklere "terim" denir. Daha önce de belirtildiği gibi terimler hem belgeleri göstermede hem de sorguları ifade etmede kullanılırlar. Bu ikisi arasında bir ayrım yapmak istediğimizde öncekine belge terimleri ve diğerine de sorgu terimleri adını vereceğiz. Bir belge teriminin ağırlığı terim belge içinde yer alıyorsa bir, aksi takdirde sıfırdır (ikili ağırlık). Bu yaklaşıma Boole modeli adı verilir. Diğer bir popüler yaklaşım ise vektör tabanlı[13] modellerde kullanılan *tf*idf* değerleridir. Burada, (**tf***)* terimin

[11] Bir sözcüğün sıklığı, ilgili sözcüğü taşıyan belgelerin derlem içindeki sayılarına eşittir.

[12] Tamlamaya katılan orta sıklıklı sözcük kendi başına terim sözlüğünde de yer alır.

[13] Vektör uzayı modelinde sorgular ve belgeler terim vektörleri biçiminde ele alınır. '*t*' tane ayrık terimin olduğu bir derlemde, *i.* belge,

$D_i = (a_{i1}, a_{i2}, \ldots a_{it})$,

ilgili belgede geçme sıklığı, yani *terim sıklığı*dır (term frequency). Terimin derlemde geçtiği belge sayısına ise *belge sıklığı* (document frequency) **(df)** denir. Terim sıklığı yüksek olan bir terim aynı zamanda derlem içindeki diğer belgelerde de sık geçiyorsa, ilgili terimin ayırt edici özelliği veya belge içindeki diğer terimlere göre göreceli değeri düşük olmalıdır. Bir terimin terim sıklığı (yani ilgili bir belgede geçme sıklığı) yüksek ve derlemdeki diğer belgelerde geçme sıklığı düşükse, o terimin göreceli ağırlığı yüksek olmalıdır. Bu kıstası sağlamak için *devrik belge sıklığı (inverse document frequency)* **(idf)** kullanılmıştır. "*idf*" parametresi terimin belge sıklığı arttıkça azalan özelliktedir. Tipik bir *idf* parametresi *log(N/df$_j$)*'dir. Burada *N*, derlemdeki toplam belge sayısı; *df$_j$*, *j*. terimin belge sıklığıdır. t_j teriminin D_i belgesi için ağırlığı w_{ij} ile gösterilirse, w_{ij}

$$wij=tf_{ij}*log(N/df_j) \qquad (1)$$

formülü ile hesaplanır. Yukarıda df_j , t_j teriminin belge sıklığı; tf_{ij}, t_j *teriminin* D_i belgesinde geçme sıklığı (terim sıklığı) ve *N* derlemdeki toplam belge sayısıdır.[14]

Terimler birbirleri ile belirli bir ilişki altında kümelendiği gibi belgeler de kümelere (clusters) bölünebilirler. Buradaki ideal amaç ise, belge arama uzayını, anma değerini sabit tutarak, küçültmektir. Belgeleri kümeleme süreci, belgeler birbiri ile karşılaştırılıp benzer bulunanların kümelenmesi ile en alt düzeyde başlar. Daha sonra kümeler birbiri ile karşılaştırılarak bir üst seviyede kümelenir. Bu işlem, tek bir küme kalana dek sürer. Oluşan yapıda sorgu en üst düzeyden başlayarak kümelerle karşılaştırılmaya başlanır ve en ilgili bulunan küme yönünde ilerlenir. Literatürde bu işleme *sıradüzensel kümeleme* (hierarchical clustering) denir (Van Rijsbergen, 1979). Bu yaklaşım arama motorlarının büyük bir çoğunluğunca 'directory search' (rehber arama) adı altında sağlanmaktadır. Kavram tabanlı bir arama motoru olan Excite'da (http://www.excite.com) ise, rehber aramaya ek olarak, geniş (broad) arama sonuçları düz (flat) olarak kümelendirilerek kullanıcıya sunulmaktadır. Böylece

---

ve *j*. sorgu ,

$Q_j=(q_{j1},q_{j2},...,q_{jt})$

biçiminde gösterilir. Burada a_{ik} ve q_{jk} sırasıyla, *k* teriminin D_i belgesi ve Q_j sorgusu içindeki göreceli ağırlıklarıdır.

[14] *"tf*idf"* metodunda terimlerin göreceli ağırlıkları önem taşır. *tf*idf* metodu ile birlikte diğer terim ağırlıklarını tartışan ve bu terimlerin karşılaştırmalı etkinliğini gösteren çalışmalara da rastlanmaktadır (Salton ve Buckley, 1988).

kullanıcı sorgusunu daraltmada ya da 'rafine' etmede hazır bloklardan biri veya birkaçıyla işleme devam ederek bilgi ihtiyacını istenilen düzeyde tatmin edebilmektedir.[15]

2.3 Sorgular

Bir sorgu, kullanıcının bilgi ihtiyacının resmi (formal) olarak belirtilmesidir. Kullanıcı çok değişik biçimlerde bir sorguyu ifade edebilir.

Arama terimleri (ya da sözcükleri) Boole işleçleri ile bağlanır (Salton, 1989; Van Rijsbergen, 1979). Boole işleçleri *ve (and), ya da (or)* ve *değil (and not)'*dir. 'Ve' işleci ile bağlanan terimlerin hepsini içeren belgeler, 'ya da' işleci ile bağlanan terimlerden en az birini içeren belgeler, 'değil' işleci ile bağlanan terimi içermeyen belgeler erişim çıktısında yer alabilirler.

Kullanıcı doğal dil ile sorgu ihtiyacını belirleyebilir. İlgili sorgu metni, Bölüm 2.2'de adımları verilen tipik bir dizinleme sürecinde olduğu gibi, arama terimleri sorgu vektörüne çevrilir. Sorgu vektörü ağırlıklandırılmış arama terimlerini (örneğin, *tf*idf* kullanılarak) içerebileceği gibi, ağırlıkların değişimini basit bir şekilde ikili değerler kümesi ile sınırlandırabilir (bir arama terimi ilgili sorgu vektöründe ya vardır ya da yoktur, fakat her ikisi olamaz). Doğal dilde girilen sorgularda ise terimlerin tamamının aynı belgede bulunma şartı yoktur. Belgenin, kullanıcının bilgi ihtiyacı ile ilgili olma derecesi, sorgu terimlerinin ne kadarını içerdiği ile doğru orantılıdır. Dolayısıyla sorguda geçen terimlerin tamamını içeren bir belge bu açıdan en iyi belgedir. Ancak bir belgenin erişim çıktısında yer alması için sorgu cümlesinde geçen tüm terimleri içermesi gerekmeyebilir. Kullanıcı tarafından verilen bir eşik değerini (threshold) aşan belgeler de erişim çıktısında yer alabilir. Başka bir deyişle, örneğin, kullanıcı bilgi ihtiyacına %80 veya daha fazla benzerlik gösteren belgeleri görmek isteyebilir.

Olasılık modeli arama terimlerini, geribildirim aracılığı ile ilgili belgelerde bulunabilme olasılıklarını temel alarak ağırlıklandırır; belge terimleri ise ikili ağırlığa sahiptirler (Robertson ve Jones, 1976; Crestani, Lalmas, Van Rijsbergen ve Campbell, 1998). Bu modelde, sorgu başlangıçta arama sözcüklerinin bir listesi olarak ya da doğal dilde ifade edilir. Sistem tarafından döndürülen belge çıktısının kullanıcının bilgi ihtiyacını

[15] Bu tür kullanıcı arayüzleri ile ilgilenen okuyucuya 'Light House' (http://www.lighthouse.org) aracını salık verebiliriz (Leuski, 2001). Bu araç, bir arama motoru tarafından döndürülen belgeleri iki boyutlu kümelendirerek (başka bir deyişle sınıflandırarak veya gruplandırarak) kullanıcıya grup etiketleri ile birlikte sunmaktadır.

karşılamaktan uzak olduğu durumlarda, kullanıcı geribildirim sürecini başlatarak daha kaliteli bir belge çıktısı elde etmek isteyebilir. Bu sürece *geribildirim* süreci denir (Salton ve Buckley, 1990). Geribildirim sürecinde, kullanıcı erişim çıktısındaki belgeleri çeşitli ilgililik düzeylerine göre sınıflandırır. Bu sınıflandırma temel alınarak, yapılan sınıflandırma hatası düzeltilmeye (daha doğrusu azaltılmaya) çalışılır. En basit ve en çok kullanılan sınıflandırma düzeyi, ilgili ve ilgisiz olmak üzere ikilidir (çok düzeyli geribildirim için bkz. Wong, Ziarko, Raghavan ve Wong (1989); Bollmann-Sdorra, Raghavan ve Sever (1999)). Hangi teknik uygulanırsa uygulansın, sınıflandırıcılar (classifiers), pozitif ve negatif örnekleri içeren belirli bir sıralı belge kümesi (erişim çıktısı) üzerinden eğitilirler (tümevarım süreci). Anma ve duyarlık değerleri açısından daha kaliteli olacağı varsayılan yeni bir erişim çıktısı ise arama sözcüklerinin yeniden ağırlıklandırılmasıyla elde edilir (tümdengelim süreci) (Wong ve Yao, 1990). [16] Eğitim aşamasında kullanıcı tarafından sisteme sunulan bilgiler kullanılarak, sorgu ifadesi içinde yer alan bir arama terimi eldeki belgede yer alıyorsa, belgenin ilgili olabilme olasılığı Bayes modeli (Duda ve Hart, 1973) üzerinde birtakim varsayımlar[17] yapılarak hesaplanır. Bu olasılık değeri arama teriminin yeni ağırlığını oluşturur.

Kavram tabanlı modeller ise kullanıcının bilgi ihtiyacını kurallar biçiminde ifade eder (Alsaffar et al., 2000, 1999; McCune et al., 1985). Ana kavramın alt kavramları bir üst kavramı oluştururken birbirleri ile 've' işleci ile bağlanabileceği gibi 'veya' işleci ile de bağlanabilir (örneğin, eğer belge (<kavram_1> ve <kavram_2>) veya <kavram_3>) içeriyorsa o zaman <ana kavram> belgede geçiyor demektir). Bir alt kavram, diğer bir üst kavramı belirli bir inanç derecesiyle belirleyebilir (Alsaffar et al., 2000). Bu yönüyle arama terimleri, yani belgede yazılı (literal) olarak yer alması istenen somut kavramlar) kullanıcı tarafından ağırlıklandırılabilir. Kavram, vektör, ve Boole tabanlı modeller arasındaki köprü P-Norm cümlecikleri ile kurulabilir (Alsaffar et al., 2000; Salton et al., 1983; Akal, 2000). Ayrıca vektör modeli içinde Boole modeli sorgu dilinin kullanılması konusundaki ilginç bir yaklaşım için okuyucu (Wong et al., 1989) no'lu analitik çalışmayı gözden geçirebilir.

2.4 Erişim Fonksiyonları

[16] Göz önünden kaçırılmaması gereken husus, geri bildirim sürecinin erişim modelinden bağımsız olup herhangi birine takılabılır (plug-in) olmasıdır.

[17] İkili bağımsız modeli içinde tanımlı bu varsayımlar aşağıdaki gibidir: (1) terimlerin ilgili belgelerdeki ve ilgisiz belgelerdeki dağılımı birbirinden bağımsızdır (2) belge terimleri ikili değere sahiptirler (Salton, 1989; Van Rijsbergen, 1979; Crestani et al., 1998).

Sorgu cümlesindeki terimlerle dizin terimleri arasında eşleşme olup olmadığı çeşitli erişim fonksiyonları kullanılarak belirlenebilir. Blair (1990) 12 değişik erişim fonksiyonunu ayrıntılı olarak incelemektedir.[18] Bu fonksiyonlar kabaca üç grup olarak sınıflandırılabilir:

1) Sorgu ve dizin terimlerinin *n*-boyutlu bir uzaydaki vektörler olarak işlem gördüğü ve ağırlıklandırıldığı vektör uzayı erişim fonksiyonu;
2) Sorgu ve dizin terimleri arasında kesin eşleşme (exact match) gerektiren erişim fonksiyonları/Boole erişim fonksiyonları; ve
3) Sorgu ve dizin terimlerinin olasılık kuramına göre ağırlıklandırılmasına dayalı erişim fonksiyonları.

Aşağıda söz konusu üç gruptaki erişim fonksiyonlarının resmi tanımları verilmektedir.

Daha önce bir bilgi erişim sisteminde üç ana nesne kümesi olduğunu söylemiştik. Bunlar sırasıyla, içerik belirteçleri (veya kısaca terimler), belgeler ve sorgulardır. Terimler hem sorguları hem de belgeleri göstermede kullanıldığı için, vektör uzayı modelinde pratik olarak sorgular ve belgeler terim uzayında bir nokta olarak görülebilir (ve bu varsayım sıkça yapılır).[19] Bu yaklaşımda her iki noktadan geçen ayrık (distinct) iki vektör (belge vektörü ve sorgu vektörü) düşünülür. Bu iki vektörün vektörel çarpımı -ki iki vektör arasındaki açının kosinüsüne eşit olduğundan kosinüs katsayısı olarak da bilinir- ya da skalar çarpımı -iç çarpım katsayısı olarak da bilinir- sorgu-belge noktaları arasındaki benzerliğin derecesini verebilir. Bu katsayılar aşağıda verilmiştir:

$$\text{İç Çarpımı } (D_r, Q_s) = \sum^t a_{ri} * q_{si} \qquad (2)$$

$$\text{Vektör Çarpımı } (D_r, Q_s) = (\sum^t a_{ri} * q_{si}) / (\sum^t (a_{ri})^2 * \sum^t (q_{si})^2)^{1/2} \qquad (3)$$

Formüllerde D_r belge vektörünü, Q_s sorgu vektörünü, a_{ri} ve q_{si} ise i. öğenin, sırasıyla, belge vektörü D_r ve sorgu vektörü Q_s'teki ağırlıklarını temsil etmektedir.

Boole modelinde bir belge veya sorgu, terimler kümesinin bir alt kümesi olarak düşünebilir. Bu durumda, iki küme (sorgu-belge) arasındaki eşleştirmelerin derecesi erişim fonksiyonunun değerini oluşturur. Örneğin, Jaccard katsayısı eldeki iki küme ($D_r=\{d_{r1}, d_{r2}, \ldots, d_{rt}\}$ ve $Q_s=\{q_{s1}, q_{s2}, \ldots, q_{st}\}$) arasındaki kesişimin oranını verir. Diğer yandan Dice katsayısı ise D_r ve Q_s kümeleri arasındaki kesişimi onların ortalama büyüklükleriyle ilişkilendirir. Aşağıda her iki katsayının resmi tanımları verilmiştir:

---

[18] Blair'in kapsamlı olarak incelediği erişim fonksiyonlarının kısa bir özeti için bkz. (Tonta, 1995).

[19] Terim uzayı kullanılarak yapılan modellemede [belgelerin ve sorguların gösterimi, karşılıklı (sorgu-belge) ve kendi içlerindeki (belge-belge, sorgu-sorgu) ilişkiler] olası paradoks durumlar Bollmann-Sdorra ve Raghavan'ın (1993) ilginç analitik çalışmasında daha ayrıntılı olarak incelenmektedir.

Jaccard Katsayısı $(D_r, Q_s) = |(D_r \times Q_s)| / |(D_r + Q_s)|$ *(4)*

Dice Katsayısı $(D_r, Q_s) = 2*|(D_r \times Q_s)| / (|D_r| + |Q_s|)$ *(5)*

Olasılık modelinde ise, daha önce de belirtildiği üzere, sorgu terimleri, geribildirim aracılığı ile ilgili belgelerde bulunabilme olasılıkları temel alınarak ağırlıklandırılır; belge terimleri ise genellikle ikili ağırlıklandırılır. Terimlerin ilgili belgelerde ve ilgisiz belgelerde dağılımının birbirinden bağımsız olduğunu varsayalım.[20] Daha ileri giderek, herhangi bir t_i belge terim değişkeni için aşağıdaki koşullu öncel (a priori) olasılıkları göz önünde bulunduralım:

$p_{ri}=(a_{ri}=1:$ *ilgili*$(Q_s))$ ve

$q_{ri}=(a_{ri}=0:$ *ilgisiz*$(Q_s))$.

Burada *ilgili*(Q_s) ve *ilgisiz*(Q_s) verilen bir Q_s sorgu ifadesi için sırasıyla ilgili ve ilgisiz belgeleri döndüren fonksiyonlar olsun. O zaman, kolayca görüleceği gibi, p_i eldeki belgenin ilgili olması halinde t_i'nin 1 olma olasılığını ve q_i eldeki belgenin ilgisiz olması durumunda t_i'nin 0 olma olasılığını verir. Aşağıdaki olasılık erişim fonksiyonu (eldeki Q_s sorgusuna göre derlem içindeki D_s belgesinin erişim değeri) kullanıldığında, sistemin hata yapma olasılığının en aza indirgendiği ve bu anlamda optimal olduğu ispatlanmıştır (Robertson ve Jones, 1976; Crestani et al., 1998):

Olasılık Erişim Fonksiyonu $(D_r:Q_s)$: $\sum t_i \log((p_i*(1-q_i))/(q_i*(1-p_i)))$. *(6)*

Yukarıdaki p_i ve q_i değerleri Q_s sorgusu için döndürülen erişim çıktısı üzerindeki kullanıcı değerlendirmeleri kullanılarak tahmin edilir. Ancak geribildirim üzerinden öncel olasılık değerlerini (p_i ve q_i) tahmin etmek pratik değildir.[21]

---

[20] İkil bağımsız erişim modelinde (IBEM) (Robertson ve Jones, 1976) göz önünde bulundurulan terimlerin (ilgili ve ilgisiz) belgeler içindeki dağılımının birbirlerinden bağımsız olduğu varsayımı, gerçeği yansıtmayan bir varsayım olduğu gerekçesi ile devamlı şekilde eleştirilmiştir. Bununla birlikte, Cooper (1995) yukarda verilen varsayıma aslında IBEM'de ihtiyaç duyulmadığını ve onun daha güçsüz versiyonu olan 'sıralı bağımlılık' varsayımının yeterli olacağına işaret etmiştir. Sıralı bağımlılık (linked dependence) kısaca aşağıdaki gibi açıklanabilir: bir belgenin ilgili ve ilgisiz sınıfllarda olma olasılıklarının oranı onu oluşturan terimlerin ilgili ve ilgisiz sınıflarda olma olasılık oranlarının tek tek çarpımına eşittir.

[21] Tahmin için kullanılan diğer yöntemler hakkında Yu ve Lee'nin (1986) çalışmasına; belge terimlerinin ikil değerler taşıması yerine kesikli değerler taşıması durumunda olasılık erişim fonksiyonu oluşturmadaki yaklaşım için sırasıyla Yu ve Lee'nin (1986) ve Bollmann-Sdorra ve diğerlerinin (1999) çalışmalarına bakılabilir.

Son olarak, erişim fonksiyonlarının her bir döndürülen belgeyi kesikli değerlerle ilişkilendirmesinin avantajlarını da sıralamakta yarar görüyoruz:

- Çıktıda döndürülen belgeler en benzer belge en üstte olacak şekilde sıralanabilir;
- En benzer belgeler ilk dönen belgeler olduğu için kullanıcıya en iyi '*n*' belge döndürülerek duyarlılık değeri artırılabilir;
- Erişimde en iyi dönen belge kullanıcıya danışılmaksızın direkt geribildirim olarak kullanılabilir.

2.5 Etkinlik

Bilgi erişim sistemlerinin etkinliği tipik olarak *anma*, *duyarlık* ve *posa* (ya da yanlış alarm) ölçütleri ile ölçülür. Bu ölçütlerin hesaplanmasında Tablo 1'de gösterilen ikili sınıflama tablosu kullanılır. Bu tablo her bir sorgu için oluşturulur. İlgili tablonun başlığında 'ikili sınıflama' tamlamasının olmasının nedeni, sistemin bilgi erişim sürecindeki tipik davranışının bir ikili sınıflama örneği göstermesidir (eldeki sorgu ile eşleştirilen belge ya ilgilidir ya da ilgisizdir). İkili sınıflama tablosunda her bir hücre ilgili satır ve sütunun kesişimini gösterir. Örneğin, '*a*' sistem tarafından erişilen ve kullanıcının ilgili (relevant) bulduğu belge sayısını, '*b*' sistem tarafından erişilen ancak kullanıcının ilgisiz bulduğu ("false drops") belge sayısını, '*a*+*b*' ilgili ya da ilgisiz erişilen toplam belge sayısını, '*a*+*c*' ise bir sorguya karşılık erişilen ya da erişilemeyen derlemdeki toplam ilgili belge sayısını verir. Çeşitli ölçütlere veya hedeflere göre farklı etkinlik ölçütleri bu tabloya dayanılarak çıkarılabilir. Burada çok iyi bilinen anma, duyarlık ve posa değerlerine yer verilecektir. *Anma*, kimi zaman *hedefi vurma oranı* olarak da adlandırılır, sistem tarafından erişilen ilgili belgelerin (*a*) derlemdeki toplam ilgili belgelere (*a*+*c*) oranını verir.[22] *Duyarlık*, sistem tarafından erişilen ilgili belgelerin (*a*) erişim çıktısında yer alan (ilgili ve ilgisiz) toplam belgelere (*a*+*b*) oranını verir.[23] Anma ve duyarlık değerleri 0 ile 1 arasında değişmektedir. Anma ve duyarlık değerleri ne kadar yüksek olursa bir bilgi erişim sisteminin etkinliğinin de o kadar yüksek olduğu kabul edilmektedir (Salton, 1989). Posa ise, sistem tarafından ilgili olduğu varsayılıp erişilen (*b*) fakat gerçekte ilgisiz olan belgelerin toplam ilgisiz belgelere (*b*+*d*) oranını verir.[24] Bu oran "bir sistemin ilgisiz belgeleri ne derece sağlıklı olarak reddettiğini ölçer" (Blair, 1990, s. 116).

---

[22] Döndürülen/erişilen belgenin ilgili olduğu verildiğinde erişim çıktısına dahil edilmesinin olasılığı, Pr(P→R), anma değeri ile tahmin edilir.

[23] Erişilen belgenin erişim çıktısına dahil edildiği bilgisi verildiğinde, belgenin ilgili olma olasılığı, Pr(R→P), duyarlık değeri ile tahmin edilir.

[24] Erişilen belgenin ilgisiz olduğu bilgisi verildiğinde, belgenin erişim çıktısına dahil edilmesi olasılığı, Pr(¬P→R), posa değeri ile tahmin edilir. Arama motorlarında (ya da genelde derlemdeki belge sayısının

Tablo 1. İkili Sınıflama tablosu

	İlgili (P)	İlgisiz (¬P)	
Erişilen (R)	a	b	a + b
Erişilemeyen (¬R)	c	d	c + d
	a + c	b + d	a + b + c+ d

Bir sistemin etkinliği çoğunlukla anma ve duyarlık değerleri ile ifade edilir.[25] Tabi bu değerler her bir sorgu bazında kesin değerler olabileceği gibi, belirli sayıdaki sorgular üzerinden mikro ya da makro ortalamalar alınarak da hesaplanabilir. Mikro ortalamada sayıların, makro ortalamada ise oranların aritmetik ortalaması alınır. Örneğin, bir arama motoruna iki soru yönelttiğimizi varsayalım. İlkinde, erişilen beş belgeden ikisi ilgili bulunsun, ikincisinde ise erişilen 10 belgeden birisi ilgili bulunsun. Bu iki soru için mikro ortalama yöntemi kullanılırsa ortalama duyarlık değeri %20 ((2+1)/(5+10)=3/15=0,2), makro ortalama yöntemi kullanılırsa %25 ((2/5)+(1/10)/2)=(0,4+0,1)/2=0,5/2=0,25) olarak bulunur. Mikro ortalama yöntemi belgelere, makro ortalama yöntemi sorgulara ağırlık verir. Bir başka deyişle, makro ortalama, sistemin tipik bir kullanıcı için tahmini değerini temsil ederken, mikro ortalama derlemde çok sayıda ilgili belge bulunan sorgulara gereğinden fazla ağırlık verir (Rocchio, 1971).

Blair'in (1990, s. 73-74) de vurguladığı gibi, bilgi erişim temelde bir deneme-yanılma süreci olduğundan, bilgi erişim sistemlerindeki belgelere erişmek için yapılan hemen hemen her aramada ilgili belgelerin yanı sıra değişen oranlarda ilgisiz belgelere de erişilmektedir. Ancak ideal bir bilgi erişim sistemi ilgili belgelerin tümüne ve salt ilgili belgelere erişim sağlar. Yukarıda açıklandığı üzere, duyarlık hesaplamasında, erişim çıktısında yer alan ilgili ve ilgisiz belge sayıları kullanılır; fakat kimi zaman sistemin aynı duyarlık değerine sahip erişim çıktıları arasından ilgili ve/veya önemli[26] olan belgeleri en iyi ön plana çıkaran erişim çıktısını seçmesi istenebilir (Kobayashi ve Takeda, 2000). Bu durumu aşağıdaki örnek (Tablo 2) ile açıklayalım.

---

yüksek olduğu bilgi erişim sistemlerinde) posa değerinin ölçüldüğü araştırmalara rastlanmamıştır. Çünkü yüz milyonlarca belge üzerinde arama yapılan Web ortamında posa değeri hemen hemen hep sıfır çıkacaktır.

[25] Anma, duyarlık ve yanlış alarm değerleri arasındaki ilişkiler için bkz. (Van Rijsbergen, 1979).

[26] Popüler olan belgelere bağlantı veren 'hub' sayfalara veya kendileri popüler olan sayfalara (authoritative) kısaca önemli sayfalar adını vermekteyiz.

Tablo 2. Normalize sıralama

Sıralama	1	2	3	4	5	6	7	8	9
EÇ1	+	+	+	+	+	-	-	-	-
EÇ2	-	-	-	-	+	+	+	+	+
EÇ3	+	+	+	-	-	-	+	-	+

Yukardaki tabloda '+' ve '-' sırasıyla ilgili ve ilgisiz belgeleri; EÇ1, EÇ2 ve EÇ3 aynı bilgi ihtiyacı için ifade edilen üç ayrı sorgu ifadesi ile ilişkili döndürülen erişim çıktıları olsunlar. Duyarlığı 'DK' ile gösterelim. O zaman, $DK_{EÇ1}=DK_{EÇ2}=DK_{EÇ3}=5/9$'dur; fakat sıralamalara göz attığımızda her üçünün farklı çıktılar olduğunu farkederiz (her üç erişim çıktısı erişim çıktı boyutunun sabitlendiği durumlarda tipik olarak ortaya çıkabilir).

Yukarıdaki tartışmanın önemli görüş noktalarından birisini, duyarlık değerleri aynı olmasına karşın kullanıcıların, ilgili belgelerin erişim çıktısında olabildiğince üst sıralarda yer aldığı arama sonuçlarını tercih etmeleri oluşturmaktadır. Çünkü kullanıcılar daha az çaba sarfederek ilgili belgelere eriştikleri arama sonuçlarının daha değerli olduğunu düşünmektedirler. Öte yandan, bir erişim çıktısında ilgisiz belgelerin en üst sıralarda yer aldığı, buna karşılık ilgili belgelerin çıktıda ya hiç yer almadığı ya da çıktının en sonunda listelendiği arama sonuçları kullanıcıların sabrını zorlayıp onları arama yapmaktan vazgeçirebilir. Bu metrik gözetilerek oluşturulan ölçüte "normalize sıralama" adı verilmektedir. Sıralama elde edilen erişim çıktısında en ilgili olduğu varsayılan belgenin ilk sırada, ilgililik derecelerine göre diğer belgelerin de izleyen sıralarda yer alması demektir. Normalize sıralama (S_{norm}) elde edilen erişim çıktılarındaki sıralamaya bağlı olarak bir bilgi erişim sisteminin etkinliğini ölçmektedir (Yao, 1995). Normalize sıralama değerinin hesaplanması için kullanılan formül aşağıda verilmektedir.

S_{norm}: $$S_{norm}(\Delta) = \frac{1}{2}\left(1 + \frac{S^+ - S^-}{S^+_{\max}}\right) \qquad (7)$$

Bu formülde:

Δ : erişim çıktısı sıralaması;

S^+ : erişim çıktısında ilgili belgelerin ilgisiz belgelerin önünde yer aldığı belge çiftleri sayısı;

S^- : erişim çıktısında ilgisiz belgelerin ilgili belgelerin önünde yer aldığı belge çiftleri sayısı; ve

S^+_{max} : mümkün olan en fazla S^+ sayısıdır.

Yukarıdaki örneğimize (S^+_{max} değerini 20 kabul ederek) devam edecek olursak:

S_{norm}(EÇ1)=1/2(1+(20-0)/20) = 1;

S_{norm}(EÇ2)=1/2(1+(0-20)/20) = 0; ve

S_{norm}(EÇ3)=1/2(1+(13-9)/20) = 0.6

değerlerini elde ederiz. . Bir başka deyişle, kullanıcının, duyarlık değerleri aynı olmasına karşın, normalize sıralama değerlerine bakarak bu üç arama sonucundan ilkini diğerlerine tercih edeceği kolayca söylenebilir.

Elde edilen değerlere dikkatle bakıldığında, normalize sıralama değerinin ilgisiz belgeleri başarılı bir şekilde reddetmeyen (yani "yanlış alarm" veren) bilgi erişim sistemlerini cezalandırdığı görülecektir. Normalize sıralama değerinin, bir bakıma, tüm ilgili belgelerin ve salt ilgili belgelerin erişim çıktısında yer aldığı "ideal erişim etkinliği" ile derlemdeki tüm ilgisiz belgelerin çıktının başında, ilgili belgelerin de çıktının en sonunda yer aldığı "en kötü erişim etkinliği"[27] arasındaki değerlere belirli bir anlam yüklemeye yaradığı söylenebilir.

Birkaç ilgili belgeye hızla erişim sağlamak isteyen kullanıcılar açısından normalize sıralama değeri önemli olabilir. Öte yandan, kapsamlı arama yapan kullanıcılar (örneğin, belli bir konuda yayımlanmış tüm belgelere erişmek isteyen kullanıcılar) ya da belli bir konuda daha önce herhangi bir belge yayımlanmadığını bilgi erişim sistemi aracılığıyla doğrulatmak isteyen kullanıcılar (örneğin, patent aramaları) normalize sıralama değerlerine itibar etmeyebilirler. Normalize sıralama değeri bir bilgi erişim sisteminin etkinliğini ölçmede tek başına bir ölçüt olarak sıklıkla kullanılmasa da, ilgili belgelere sürekli ilk sıralarda erişen bilgi erişim sistemlerinin diğerlerine göre performans yönünden daha etkin sistemler olduğunu kabul etmek gerekmektedir.

Bilgi erişim sistemlerinin etkinliğini ölçmede kullanılan "kapsama" ve "yenilik" oranlarından da kısaca söz etmekte yarar vardır. Kapsama oranı (coverage ratio), erişilen ve kullanıcının daha önceden ilgili olduğunu bildiği belge sayısının, ilgili olduğu bilinen toplam belge sayısına oranıdır. Yenilik oranı (novelty ratio) erişilen ve kullanıcının daha önce görmediği ilgili belgelerin erişilen ilgili belgelere oranıdır (Korfhage, 1997, s. 198). Kapsama ve yenilik oranlarını hesaplamak için aşağıdaki formüller kullanılır:

[27] Aslına bakılırsa, bilgi erişim sistemleri bir sorgu karşılığında derlemdeki tüm belgelere erişim sağlanmasına genellikle izin vermez.

$$\textit{Kapsama oranı} = |R_k| / U \qquad (8)$$

$$\textit{Yenilik oranı} = |R_u| / |R_u| + |R_k| \qquad (9)$$

Formüllerde *U,* kullanıcının daha önceden bildiği ilgili belgelerin setini, $|R_k|$, erişilen ve kullanıcının ilgili olduğunu önceden bildiği belge sayısını, $|R_u|$ ise erişilen ve kullanıcının daha önceden görmediği ilgili belgelerin sayısını ifade etmektedir.[28] Örneğin, kullanıcının, aradığı konuda toplam 15 ilgili belge (*U)* olduğunu bildiğini varsayalım. Sistem, kullanıcının sorusuna karşılık toplam 10 ilgili belgeye erişir ve bunlardan 4'ü ($|R_k|$), kullanıcının daha önceden bildiği belgeler olursa kapsama oranı 4/15 olur ($|R_k| / U$). Aynı örneği kullanacak olursak, erişilen ilgili belgeler arasında kullanıcının daha önceden görmediği 6 belge bulunmaktadır ($|R_u|$). Dolayısıyla yenilik oranı 6/10 olur (Korfhage, 1997, s. 198). Yüksek kapsama oranı sistemin, kullanıcının görmek istediği belgelerin çoğuna eriştiği, yüksek yenilik oranı ise sistemin, kullanıcının daha önceden bilmediği yeni belgelere eriştiği anlamına gelmektedir.

Kuşkusuz kullanıcı, gerçekte daha önceden bildiği belgelerle ilgili değildir. Kullanıcı açısından yüksek yenilik oranı tercih edilir (Korfhage, 1997, s. 198).

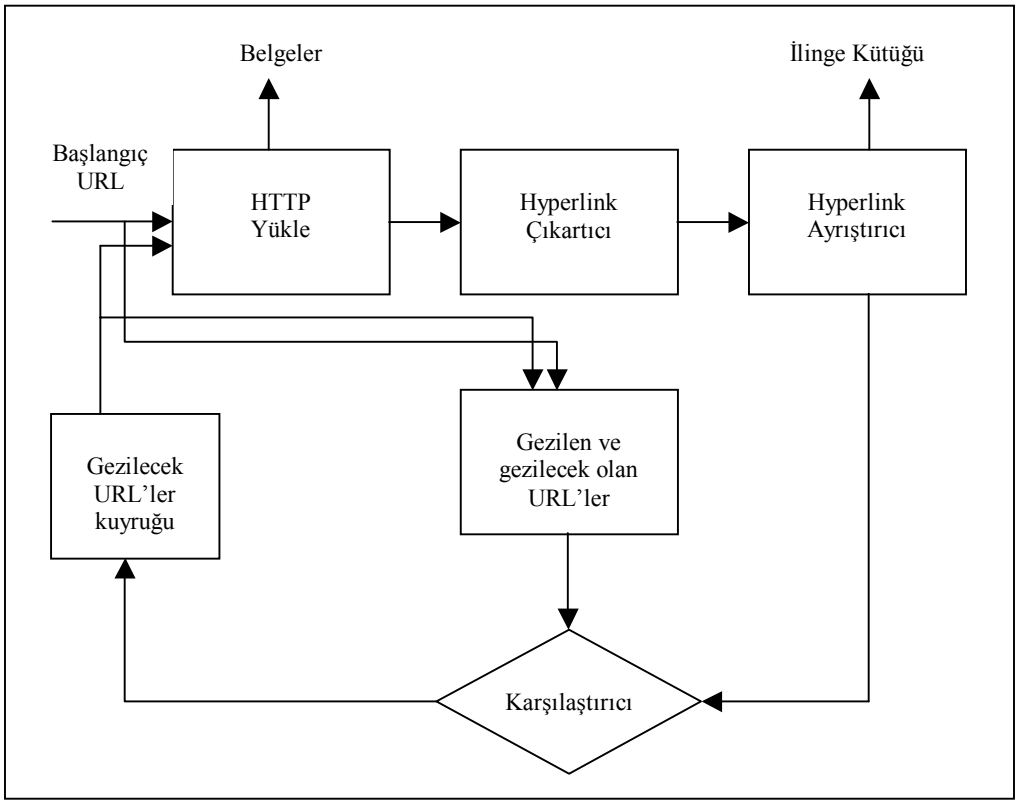
[28] Formül için bkz. http://home.himolde.no/~molka~/in350/c4y00.html.

3 ARAMA MOTORLARI

Arama motorları, son elli yıldır geliştirilmekte olan bilgi erişim sistemlerini temel almaktadırlar. Bununla birlikte, arama motorları gerek mimari açıdan gerekse işlevsel özellikleri açısından bilgi erişim sistemlerinden farklılıklar gösterir. Bu bölümde arama motorları hakkında hem mimari hem de işlevsel açıdan tanımlayıcı bilgiler verilmekte ve geleneksel bilgi erişim sistemleriyle arama motorları arasındaki farklı yönlere dikkat çekilerek, konuyla ilgili literatür kısaca incelenmektedir.

3.1 Mimari Yapı

Arama motorlarının esas bileşenlerden birisi, Web üzerindeki herhangi bir sitenin yerel diske indirilmesini sağlayan ağ sörfçüsü (*network surfer*) işlevini gören bir robottur (web crawler, spider). Tipik bir robotun genel görünümü Şekil 2'de verilmiştir.



Şekil 2. Robotun işlevsel görünümü

Robot, hafif bağlı üç alt modülden oluşmaktadır: *yükleyici* (*downloader*), *çıkartıcı* (*extractor*) ve *ayrıştırıcı* (*parser*). Yükleyici, bir başlangıç adresi (*seed node, root node*) ile çalıştırılır.[1] Yükleyici gezilecek URL'ler *(*Uniform Resource Locator*)* kuyruğundaki adresleri önce-enlemesine (breadth-first) dolaşmaya başlar. Yükleyici dolaştığı belgeleri HTTP protokolünü kullanarak getirir ve getirdiği belgeleri hiper-bağlantı çıkartıcıya iletir. Yanlış ya da geçersiz bağlantıların tespit edilmesi, bu bağlantıların ziyaret edilmek istenmesi durumunda, Web sunucusundan gelen hata iletileri yardımıyla yapılır. Çıkartıcı, gelen belge içerisinde <*a href*="..."> ya da <*frame src="..."*> biçimindeki HTML (Graham, 1997) takılarını araştırır. Ayrıştırıcı, göreli URL'leri tam URL adreslerine çevirir. Belge içindeki tam URL adresleri için herhangi bir işlem yapılmaz. Ayrıştırıcıdan gelen URL adresi

[1] Web alanı düzeyinde yeni çekirdek (seed) adreslerin nasıl elde edildiği bilgisi literatür taraması esnasında elde edilememiştir. Fakat bu konudaki güvenilir çözüm, herhangi bir sitedeki Web sayfaları işlenirken elde edilen site dışı hiper-bağlantıların bir veri tabanında sonradan işlenmek üzere ayrık (distinct) olarak saklanmasıdır. Arama motorları Internet'teki yeni sayfaları robotların sörfü esnasında tespit edebildiği gibi, bu sayfaların yaratıcıları arama motoruyla bağlantı kurup, arama motorunu sayfadan haberdar edebilmektedirler (Gordon ve Pathak, 1999). Arama motorları genelde eldeki siteleri düzenli aralıklarla ziyaret edip her site ile ilgili değişiklikleri tespit eder (Sullivan, 2000).

gezilecek URL'ler kuyruğuna atılmadan önce bir döngüye girilip kilitlenme durumu oluşmasın diye o ana kadar gezilen URL'ler ile karşılaştırılır. Bu aşamada ayrıca gezilecek URL değerlerinin başlangıç olarak verilen URL ile aynı alan adını (domain name) taşıyıp taşımadığı da kontrol edilir. Aynı alan adını taşımayan URL'ler ziyaret edilmez. Böylece robotun sadece istenen bir sitedeki belgeleri getirmesi sağlanmış olur. İşletimin sonunda, getirilen belgeler yerel olarak depolanır. Ayıca, verilen başlangıç URL değeri için dolaşılan URL'lerin listesi ve onların ilinge (topology) bilgisi çıktı olarak verilmektedir. İlinge bilgisi, getirilen belgede referans verilen tam URL adreslerinin listesini içermektedir. Bu tür ilingesel bilgiler bir elektronik katalogda dizinlenecek sayfaların tespitinde doğrudan kullanılmaktadır (Deogun, Sever ve Raghavan, 1998).

Robotun bulduğu her şey, arama motorlarının ikinci bileşeni olan "veri tabanı"na kaydedilir. Arama motorunun diğer bir bileşeni ise "ajan" olarak adlandırılan arama motoru yazılımıdır. Bu yazılım, dizinde kayıtlı olan milyonlarca sayfa içinden en ilgili olduğunu "düşündüğü" siteleri eleyerek bunları (genelde) ilgililik derecelerine göre sıralar (Sullivan, 2001).

Web robotları basit programlar olmasına rağmen Web üzerinde bulunan milyonlarca dokümanı kullanıcıların hizmetine sunmak ve aranan bilgiye kolay ve doğru bir şekilde erişilmesini sağlamak amacıyla çalışmaktadırlar. Hatta zaman zaman site sahibinin saklı tuttuğu materyalleri de otomatik ve hızlı bir şekilde keşfedebilmektedirler. Bu yüzden birçok robot gayri resmi "robotları dışlama protokolü"ne (robots exclusion protocol) göre belirlenmiş kurallar kümesi dahilinde hareket etmek zorunda kalmaktadır.

İsimleri "AbachoBOT"'dan "ZyBorg"'a kadar değişen bu robotlar tüm popüler arama motorları tarafından kullanılmaktadırlar. Örneğin; Inktomi "Slurp", AltaVista "Scooter", Google ise "Googlebot" robotlarını kullanmaktadır. Bazı arama motorları değişik amaçlar için birden fazla robot da kullanmaktadır (örneğin, yeni sayfaları bulmak için bir robot, sayfa bağlantılarını kontrol etmek için başka bir robot şeklinde). Ama bu robotların tümü arama motorları için çalışmamaktadır. Kimi robotlar sayfa bağlantılarının canlı olup olmadığını kontrol etmekte (link checker), kimisi sayfa değişimini denetlemekte (page change monitors), kimisi sayfanın HTML kodunun doğruluğunu ve standartlara uyumluluğunu kontrol etmekte (validators), kimisi FTP istemcisi (FTP client) olarak indirilecek olan dosyaların yönetiminde, kimisi de sayfa ziyaretlerinde (web browser) kullanılmaktadır.

3.2 Dizinleme

Dizinlemede ve ilgili belgeleri saklamada arama motorlarının karşılaştıkları tipik sorunlar ile bilgi erişim sistemlerinin çözmesi gereken sorunlar birbirinden farklıdır. Bu tür sorunlar, ki bu alt bölümün gerisinde işlenecektir, çoğunlukla değişik ve kendine özgü çözümleri gerekli kılar.

Bilgi erişim sistemlerinde dizinlenecek belgeler durağandır (statik). Başka bir deyişle, bir belge bir defa dizinlendikten sonra bir daha dizinleme işlemine tabi tutulmaz. Halbuki Web kaynakları tahmini olarak ortalama 75 gün değişmeden kalmaktadırlar (Brake, 2001). Kahle, yapılan tahminlere göre Web kaynaklarının %40'ının her ay değiştiğini (Kahle, 1996), Internet ortamındaki bir bağlantının (link) ortalama ömrünün 44 gün olduğunu belirtmektedir (Kahle, 1997).[2] Web'deki belgelerden örneklem seçilerek yapılan ve 120 hafta süren "uzunlamasına" bir araştırmada bir Web sayfasının ya da bir Web sitesinin "yarı ömrü"nün (half-file) iki yıl civarında olduğu bulunmuş ve Web sayfası/sitesi içeriğinin bir yıllık bir sürede değiştiği saptanmıştır (Koehler, 1999). Bu tür bilgiler, arama motorlarının mimarisinin bilgi erişim sistemlerininkine göre elbette farklı olmasını gerektirmektedir. Örneğin, daha önce dizinlenmiş kaynaklarda belirli aralıklarla günleme yapılabilir (HTTP protokolü bu tür kararları destekleyen sorgulara olanak vermektedir). Günleme yapılacaksa ilgili kaynağı yeniden dizinleyen ve eskisinin yerine yerleştiren bir robot modülün belirli aralıklarla, mevcut veri tabanının belirli bir kısmını rastgele denetleyerek işletilebilir. Aslında, Internet'in üssel olarak büyümesi (her sekiz ayda bir bilgi hacminin ikiye katlanma eğilimi göstermesi) ve Internet kaynaklarının sık sık değiştirilmesi mimariyi daha karmaşık hale getirdiği gibi bilinen arama motorlarının toplam Web kaynaklarının ne kadarını dizinleyebildiklerini ve bunların kesişim oranlarını tahmin etmeyi de güçleştirmektedir. Bu konuda göstergeler ümit verici olmaktan uzaktır: Internet'in küçük bir yüzdesi dizinlenebilmekte ve bu yarış her geçen gün arama motorları aleyhine işlemektedir (Lawrence ve Giles, 1998; Bergman, 2001; Kobayashi ve Takeda, 2000).

Dizinlenebilen Web sayfalarının azlığı problemi, dizinlenecek sayfaların *kalite*sinin göz önünde bulundurulmasını gündeme getirmiştir. Kimi çalışmalarda kaliteli olma hiper-metnin cebrik çizge özelliklerinden yola çıkılarak gündeme getirilmiştir (Deogun et al., 1998; Furner, Ellis ve Willet, 1996; Doorenbos, Etzioni ve Weld, 1996; Etzioni ve Weld, 1994). Örneğin, Deogun ve diğerlerinin (1998) çalışmasında, yazarlar çevrimiçi bir katalogda detaylı ürün

[2] Sadece bu istatistiki bilgi bile Türkiye'de bir süre önce kanunlaştırılmaya çalışılan (ancak veto edilen) Internet'i "zapt-u rapt" altına almaya yönelik girişimlerin ne kadar "naïf" olduğunu göstermeye yeterlidir kanısındayız.

bilgisinin bulunduğu sayfaları (hiper-metinde düğüm olarak da adlandırılır) referans sınıfında, benzer ürünlere ait bilgilerin tablolar ve/veya listeler kullanılarak topluca verildiği sayfaları da 'özel' sınıfında toplamışlardır. İlgili sınıflar gerek cebrik özellikleri (giriş veya çıkış bağlantıları istatistiği) gerekse kullanılan HTML yapılarının türlerine bakılarak tanınmıştır. Büyük hacimli dört Web kataloğunda (toplam 122 MB) SMART sistemi kullanılarak yapılan çalışmada, referans sayfalarında ve 'özel' sayfalardaki liste ve tablo yapılarının içeriklerini dizinlemekle tüm katalogları dizinlemenin performans açısından birbirlerinden bir farkı olmadığı rapor edilmiştir. Bu da, sonuç olarak, gerçekleştirilen deneydeki toplam 122 MB'lık dizinleme uzayını 17 MB'a indirgemiştir. HTML yapılarının Web üzerinden bilgi keşfedilmesinde kullanılması aslında yeni bir olay değildir. Doorenbos ve diğerleri, bilgi keşfetme şemsiyesi altında bir alış veriş aracı (shopping agent) geliştirmişlerdir (Doorenbos et al., 1996). Geliştirilen bu araçta, verilen bir ürün ismi için en ucuz fiyatları veren alış veriş siteleri (veya katalogları) taranırken, HTML yapılarının ve gösterim biçimlerinin (kalın, italik, boşluk, vb) uyumlu ve sürekli kullanıldığı varsayılmıştır.

Geleneksel yaklaşımda, örneğin, belgelerin yazım kalitesi (ya da metin değeri) oldukça yüksektir. Halbuki, Web sayfalarında yapılan yazım hataları bir istisna olmanın çok ötesindedir. Örneğin, yapılan bir doktora çalışmasında her sitede sık olarak kullanılan kelimelerden ortalama 200 tanesinin ve her üç yabancı soyadından birisinin yanlış hecelenmiş olduğu görülmüştür (Badino, 2001). Bunun arama motorlarına getirdiği ek yük sadece gövdelemeyle sınırlı değildir; arama motorlarının aynı zamanda düzeltme yapabilme yeteneklerinin de olması gerekmektedir.[3] Bir Internet sayfasının kaliteli olması, kimi zaman da, ne kadar sayfanın kendisine referans verdiği (authoritative) veya ne kadar çok authoritative sayfalara referans verdiği (hub) ile ölçülebilir olmuştur (Kleinberg, 1998; Lynch, 1997).

Başka bir sorun ise ikilenen (duplicate) sayfaların yüzdesinin giderek artmasıdır. Bir araştırmaya göre Web sayfalarının %30'u tekrarlardan oluşmaktadır (Kirsch, 1998). Tekrarlı sayfaların tanınması ve yalnızca bir kez dizinlenmesi birçok araştırmaya konu olmuştur (Kobayashi ve Takeda, 2000; Kirsch, 1998). Tipik olarak herhangi bir arama motorunun değerlendirilmesinde de tekrarlı Internet kaynaklarının erişim çıktısında yer alıp almadığı ölü bağlantılar ile birlikte sık sık anılan bir kriter olmuştur.

---

[3] Bilindiği üzere, Türkçe gövdeleme (Duran, 1999) bir dil üyesini *tanıma* işlemi; halbuki düzeltme ise dil üyelerini *üretme* problemidir. Verilen bir kelimeye ortalama 1.65 gövde karşılık gelmektedir. Bu da, kelime düzeltmenin Türkçede sezgisel yöntemlerle çözümlenebileceğini göstermektedir.

3.3 Belgelerin Gösterimi

Arama motorları dizinlemeyi azaltmak için, geleneksel bir bilgi erişim sisteminin aksine, verilen bir belgeyi olduğu gibi dizinlemez (Kobayashi ve Takeda, 2000; Laursen, 1998). Tipik olarak, bir Web sayfasının[4] başlık kısmı, üst veri belirteçlerinin (metadata tags) içerikleri, tam metnin ilk bir-iki paragrafı dizinlenir. Web sayfalarının insan gözüne hitap eden bir şekilde hazırlanması, ama öte yandan bu sayfaların arama motorları tarafından kolayca bulunmasının beklenmesi arama etkinliğini (örneğin duyarlık) olumsuz etkilemektedir (Olgun ve Sever, 2000; Küçük, Olgun ve Sever, 2000).

Web sayfalarının arama motorlarına hitap eden kısmıyla ilgili ilk adım, HTML 3.2 standardında belirteçlerinin tanımlanmasıyla atılmıştır.[5] HTML kodunun basında bulunan ve <head> … </head> alanı ile sınırlanan, üst veri belirteçleri görüntülenebilir olmayıp tamamen robotlara hitap etmektedir. Arama motorları açısından ilginç olabilecek iki belirteç ismi "tanım" (description) ve "anahtar sözcük"tür (keyword). Aşağıda Türk Kütüphaneciler Derneği'nin (TKD) Web sitesinden (http://www.kutuphaneci.org.tr/turk/) alınan bir örnekte "tanım" ve "anahtar sözcük" üst veri belirteçleri görülmektedir (Şekil 3).

Şekil 3. Türk Kütüphaneciler Derneği Web sitesi üst veri alanları

[4] Burada dolaylı olarak ilgili Web sayfasının kalite açısından robot trafından indirilip dizinlemeye değer bulunduğunu varsayıyoruz.

[5] HTML 4.1 için bakınız: http://www.w3.org/TR/REC-html40/struct/global.html#h-7.4.4

Bir Web sitesinde yer alan üst veri belirteçlerinin listesi (author, description, keyword, vs.) <head> etiketi içinde yer alan bir profil niteliğindeki biricik URI adresi ile kontrol edilebilir. Ancak bu, zorunlu değildir. Üst veri içeriklerini belirli bir sözcük haznesi ve kodlama kuralları ile kontrol etmek mümkün değildir. Bu durum arama motorları açısından ciddi bir sorun yaratmazken, duyarlık ve anma değerlerinin yüksek olması gereken veri tabanı uygulamaları için yeterli olmaktan çok uzaktır. Örneğin, yazar adı alanında isim ve soyad olarak mı yoksa soyad ve isim olarak mı kodlama yapılmıştır? Ya da birbirinden farklı iki ayrı tanım alanı içinde yer alan "bilgisayar ürünlerinin fiyat listesi" ile "bilgi teknolojisi malları ve ücretleri" anlamsal olarak sayfaları birbirlerine ne derece yaklaştırmaktadır? Bu sorunun cevabı veri tabanı sistemlerinde, bilgi erişim sistemlerindekinden farklı olarak, kesin olmak zorundadır. Bu amaçla yönlü çizge tabanlı bir veri tabanı modeli olan RDF (Resource Description Framework) (W3C, 1999) ve RDF'nin serileştirilmesi[6] için kullanılan XML (W3C, 1997) dili tanımlanmıştır. Internet kaynakları arasında ilişki kurabilen ve genişletilebilir olan RDF üstüne kütüphanecilik uygulamaları için kullanılmak üzere 15 elemandan oluşan Dublin Core (DC) standardı tanımlanmıştır (Dublin Core, 1998).[7] Başka bir deyişle, üst veri, Web kaynağının içeriğini makinenin anlayabileceği dilde tanımlamak amacı ile kullanılmaktadır.

Üst verinin bir Web kaynağına yerleştirilmesi kolay olmasına karşın mevcut Web sayfalarında kullanımı düşüktür. 1998'de yapılan bir araştırmada polimer kimya konulu Web sayfalarının yaklaşık %25'inde HTML üst veri belirteçleri kullanıldığı ortaya çıkmıştır (Qin ve Wesley, 1998). 1999'da yapılan bir başka araştırmada ise bu oran %34 olarak bulunmuştur (Lawrence ve Giles, 1999). Ancak Web sayfalarında Dublin Core üst veri belirteçlerinin kullanımı ise çok daha düşüktür. 1998'de yapılan bir araştırmada örnek olarak seçilen 1024 ev sayfasının sadece yedisinin Dublin Core üst veri belirteçleri içerdiği görülmüştür (O'Neill, Lavoie ve McClain, 1998). Bir başka çalışmada bu oran binde üç olarak bulunmuştur (Lawrence ve Giles, 1999). 2001 yılında yapılan bir çalışmada Web sayfalarında üst veri kullanmayanların %50'sinin üst veri hakkında herhangi bir bilgileri olmadığı ortaya çıkmıştır (Klarin, Pavelić ve Pigac, 2001). Dolayısıyla üst veri hakkında Web editörlerinin yeterince bilgi sahibi olmadıkları görülmektedir.

---

[6] Bir kodlama dili aracılığı ile metin türü bilgilerin bilgisayarın işleyebileceği hale çevrilmesi işlemine "serileştirme" adı verilmektedir.

[7] Türkçe RDF/DC editörü ve bu konuda genel bir tartışma için bkz. (Olgun ve Sever, 2000; Küçük et al., 2000).

Üst verilerle ilgili bir başka nokta "spam"dır.[8] Web sayfalarının dizinlenmesine çözüm olarak düşünülen üst veri sistemi kısa bir süre sonra kötüye kullanılmaya başlanmış, Web sitelerinin arama motorlarında üst sıralarda yer almasını sağlayabilecek "spam" teknikleri geliştirilmiştir (Henshaw, 2001). Böylece Web kaynağının üst verisine, kaynak ile ilgili olmayan ve arama motorlarında arama için kullanılan en güncel, en genel ve en popüler sözcükleri yerleştirerek erişilen sonuç listelerindeki sıralamalarda üst sıralara çıkmak amaçlanmaktadır. Kuşkusuz erişim açısından önemli dizinleme bilgileri içermesi gereken üst veri belirteçlerinin "spam" ile kirletilmesi erişim etkinliğini azaltmaktadır. Arama motoru servisleri "spam"ı tanıyabilecek ve önlem alabilecek algoritmalar geliştirmeye çalışmaktadırlar (Notess, 2001). Ancak kişilerin bilgiye erişimi engelleme pahasına da olsa kendi popülarite veya ticari kazançlarını ön planda tutmaları bu çalışmaların henüz tam anlamıyla başarı kazanmasını engellemektedir. Bundan dolayı, AltaVista, HotBot, Infoseek ve WebCrawler gibi arama motorları HTML üst veri belirteçlerini belgelerin gösteriminde sınırlı olarak kullanmalarına karşılık, Excite ve Lycos gibi bazı arama motorları üst veri etiketlerinden yararlanmamaktadır (Laursen, 1998). Onüç arama motoru üzerinde yapılan bir başka araştırmada ise tüm motorların "başlık" belirtecini (title tag), AltaVista, HotBot ve Infoseek'in anahtar sözcük ve tanım belirteçlerini, HotBot'ın "yazar" belirtecini (author tag), AltaVista ve Lycos'un şekil, resim ve görüntülerle ilgili başlık ya da resim altı (caption) gibi alternatif metin bilgisi veren "alt" belirtecini (alternative tag)[9] dizinledikleri gözlenmiştir (Mettrop ve Nieuwenhuysen, 2001).

3.4 Erişim Fonksiyonu

[8] "Spam" kelime olarak, 'genellikle öğleleri sade veya sandviç içinde tüketilen pembe renginde bir konserve et' anlamına gelmektedir. Spam, Amerika Birleşik Devletleri'nde göreceli olarak popüler olan ama birçok kimse tarafından da hiç bir estetik ve beslenme değeri olmayan yiyecek türü olarak değerlendirilmektedir. Kelimenin bilişim jargonuna, "aksi takdirde istenmeyecek veya sorulmayacak olan aynı mesajın/e-postanın birçok e-posta hesabına ve/veya Usenet haber grubuna gönderilmesi" anlamıyla girmiştir. (Spam karşıtı bir portal adresi için bkz.: http://spam.abuse.net/). Bu mesaj bombardımanı çoğunlukla bir ticari avantaj sağlamak için kullanılmaktadır. Bu çalışmada söz ettiğimiz 'spam' ise 'SEP' (Search Engine Persuasion) veya 'Web Spam' olarak adlandırılmaktadır. Burada söz konusu olan, bir arama motorunun erişim fonksiyonun nasıl çalıştığını ve bir belgenin nasıl dizinlendiğini doğruya yakın kestirebilmek ve bu bilgiyi bir avantaj (veya kişisel tatmin için) sağlamak üzere kullanmaktır. Bir başka deyişle spam, arama motorlarına bir belgeyi o belgenin HTML koduyla oynayarak gerçek içeriğinin ötesinde başka birşeyle ilgiliymiş gibi "yutturmak"tır (örnekler için bkz: (Laursen, 1998)).

[9] Alt belirteci şekil, resim ve görüntülerin yüklenmediği ya da kullanıcının bu özelliği kullanmak istemediği durumlarda sayfayla ilgili alternatif metin bilgisi sunması açısından yararlıdır. Bu belirtecin Web madenleme araçları (Web mining tools) tarafından sayfalar arasındaki ilişkilerin ortaya çıkarılmasında ya da bağlantıların anlamsal olarak sınıflandırılmasında da kullanıldığı görülmektedir.

İkinci bölümde genel bilgi erişim sistemleri için verilen erişim fonksiyonları arama motorları için de geçerlidir. AltaVista, Yahoo! gibi nispeten büyük arama motorları hem ticari sır olması açısından hem de "spam"a yol açmamak için başvurdukları erişim fonksiyonlarını ve dizinleme tekniklerini açıklamamaktadır. Bununla birlikte, söz konusu arama motorlarının çoğunun daha önce akademik ortamda geliştirildikleri bilindiği için, kullandıkları erişim fonksiyonları şu veya bu şekilde tahmin edilebilmektedir. Örneğin, Infoseek arama makinesi Massachusetts Üniversitesi tarafından geliştirilen INQUERY[10] bilgi erişim sisteminin ticari sürümüdür ve ilgililik (relevance) hesaplamasının belge istatistiği (*tf*idf*), kısmen sayfanın başka sayfalar tarafından ne kadar sıklıkla referans verildiğine (popülaritesine) ve bu sayfadan bağlantı verilen sayfaların popülaritesine dayanmaktadır (Kirsch, 1998). Google arama motoru yalnızca belge istatistiğini değil, sayfanın 'hub' ve 'authoritative' bağlantılarını da dikkate almaktadır (Kleinberg, 1998; Kobayashi ve Takeda, 2000). AltaVista ise belge sıklığına dayalı ağırlıklı Boole araması (weighted Boolean search) yapmaktadır (Silverstein, Henziger, Marais ve Moricz, 1999). Excite kavram tabanlı arama yapan, Boole sorgu dilini kullanan ve gövdeleme tekniğinden yararlanmayan bir arama motorudur (Jansen, Spink, Bateman ve Saracevic, 1998).[11] Kavramlar, terimlerin kümelendirilmesine (çevrimiçi eşanlamlı sözlük) dayanır. Excite aramada ise 'latent semantic' analiz metodunun (Deerwester et al., 1990; Foltz, 1996) hesaplama-zaman etkinliği açısından basitleştirilmiş şeklini kullanmaktadır.[12]

Erişim fonksiyonunda bir sorgu ile belge arasındaki benzerlik hesaplamasında basit olarak her ikisinde de geçen ortak terimler temel alınabileceği gibi, bir belgeyi kendisini oluşturan yapısal bileşenlerin (başlık, anahtar sözcükler, özet, tam metin, vb. gibi) bir bütünü gibi görüp, belgenin çeşitli bileşenlerinde geçen arama terimlerine farklı ağırlıklar verilebilir. Örneğin, erişim fonksiyonu çeşitli belge bileşenlerinin sorgu ile benzerliklerinin toplamı olan bir polinom şeklinde düşünüldüğünde, başlık bileşeninin sorgu ile benzerliği belgenin tam

---

[10] INQUERY çıkarsama-ağı tabanlı (inference network-based) bir bilgi erişim sistemidir (Turtle ve Croft, 1991).

[11] Kanımızca Excite arama motorunun ilginç yanlarından birisini oluşturan 'More Like This' (Buna benzer diğer sayfaları bul) özelliği bu çalışmanın kaleme alındığı sırada doğrulanamadı. Büyük bir olasılıkla kaldırılmış olan bu özellik 'ilgililik geribildirimi" (relevance feedback) tekniğine' başvuruyordu ve bu yönüyle arama motorları arasında biricik (unique) bir perspektif sağlıyordu. Klasik bilgi erişim sistemlerinde kullanılan tekniğe bağlı olarak, bilinen küçük veri derlemlerinde %28-%46 arasında (Salton ve Buckley, 1990), büyük veri derlemlerinde (TREC D1 ve D2 gibi) %14-%21 arasında (Lee, 1995) performans artırımı sağlayan ilgililik geribildirimi tekniğinin arama motorları arasında aynı öneme sahip olmaması, arama motorlarında üzerinde araştırma yapılan sorunların klasik bilgi erişim sistemlerinin sorunlarından farklı olduğunun önemli bir göstergesidir.

[12] "Intelligent Concept Extraction" adı altında Excite tarafından patenti alınmıştır (bkz. http://www.excite.com/ice/tech.html).

metniyle benzerliği ile aynı kefeye konmayabilir. Bir başka deyişle, örneğin, belge başlığında geçen bir terim, belgenin konusunu belirlemede daha ağırlıklı olarak değerlendirilebilir. Deneysel olarak bilinen bu gerçek, bir anlamda eldeki belgenin ilgililik derecesini tayin etmede farklı kaynaklardan gelen kanıtların birleştirilmesi şeklinde düşünülebilir. Nitekim '90'ların ortalarında ortak bir veri tabanı (ya da belge derlemi) üzerinde farklı erişim modelleri çalıştırılarak eldeki sorgular değerlendirildiğinde, farklı erişim fonksiyonlarına göre erişilen sonuçların birleştirilmesinin erişim performansını büyük ölçüde (tek bir işlemeye, sorguya ya da alt modele göre göreceli olarak) artırdığı gözlenmiştir (Lee, 1997, 1995).[13]

Erişim fonksiyonunun belgenin çeşitli bileşenlerini eldeki sorguyla eşleştirirken farklı ağırlıklar kullanabileceğini daha önce belirtmiştik. Bu özellik aşağıda verilen örnekle daha ayrıntılı olarak açıklanmaktadır (Yuwono ve Lee, 1996).

Boole modelindeki erişim fonksiyonunun ikil (binary) mantıkla çalıştığı bilinen bir gerçektir. Zaten bu yüzden Boole modelinde erişim çıktısındaki belgelerde sıralama yoktur (Salton, 1989). Bir başka deyişle, erişim çıktısının en başında yer alan bir belge ile en sonunda yer alan belge aynı erişim değerlerine sahiptir. Fakat ufak bir trük ile -ki çoğu arama motorlarında bu yapılmaktadır- Boole erişim fonksiyonu kullanılarak sıralama yapmak mümkün olabilmektedir:

$$\text{İç Çarpımı } (D_r, Q_s) = \sum^{t} a_{ri} * q_{si.} \qquad (10)$$

Burada D_r'yi r ile gösterilen URL adresine sahip bir Web belgesi ve Q_s'yi s ile gösterilen bir numaraya sahip bir sorgu ifadesi olarak düşünebiliriz. Daha da ileri giderek D_r ve Q_s'yi sırasıyla belge terimlerinden, a_{ri}, ve arama terimlerinden, q_{si}, oluşan listeler olarak yorumlayalım ($1 <= i <= t$). Bu iç çarpım bize ortak terimler için eşit bir tamsayı değeri döndürecektir. Yukardaki iç çarpım kolayca görülebileceği gibi Boole 'VE' işlecine karşılık

[13] Bu tür aramaya iç üst arama (internal metasearch) adı verilir. Başka bir deyişle, kendi basına işletimde olmayan fakat bir ana makineye bağlı olarak çalışan alt bileşenlerin erişim çıktıları seçilen bir birleştirme (combination/fusion) algoritması çerçevesinde tek bir erişim çıktısı haline getirilir. İç üst arama tekniklerini daha popüler olan dış arama motorları, örneğin, profusion (http://www.profusion.com) ya da metacrawler (http://www.metacrawler.com) ile karıştırmamak gerekir. Burada söz konusu olan, bu çalışmanın ana temasını oluşturan ve kendi başına işletimde olan arama motorlarına bir yazılım aracı (meta search engine agent) tarafından ilgili sorgu ifadesinin yönlendirilip sonuçların tek bir erişim çıktısı altında birleştirilmesidir. Dış üst arama motorunun, ilgili bağımsız çalışan arama motorlarına müdahale imkanı olmayıp, nasıl bir erişim fonksiyonu kullanıldığı da bilinmeyebilir. Hatta kullanılan veri tabanları (dizinlenen Web sayfaları) ortak olmak zorunda değildir. Burada kritik nokta döndürülen sonuçların normalize edilmesi (belgelerin sıralama değerlerinin modellenmesi) (Montague ve Aslam, 2001) ve birleştirilmesidir (Belkin et al., 1995). Birleştirmedeki espri farklı erişim stratejilerinin (Boole, bulanık mantık, vektör, olasılık, vb. gibi) benzer ilgili belgeler ve farklı ilgisiz belgeler döndürmeleridir.

gelir. Doğal olarak 'VEYA' ve 'DEĞİL' işleçleri nasıl yorumlanacak diye sorulabilir. Her bir Boole ifadesi anlamı değişmeksizin DNF (Disjunctive Normal Form) formuna çevrilebilir. DNF formuna çevrilen bir sorgu, birbirlerine 'VEYA' ile bağlanmış bağımsız cümleciklerden oluşur -ki her bir cümlecikteki terimler de birbirlerine 'VE' ile bağlanmıştır. Bu bağımsız cümlecikler kendi başına sorgu olarak düşünülüp yukardaki iç çarpım işlemi gerçekleştirilir. Sonuç listeler aşağıdaki gibi birleştirilebilir: Bir belgenin toplam erişim değeri ilgili sonuç listelerindeki erişim değerlerinin toplamıdır. 'DEĞİL' işleci DNF çevriminin sonucunda terimlere tümleyen olarak yansıtılır (belgenin ilgili terimi içermemesi anlamına gelmektedir). Bu da erişim fonksiyonuna sonradan budama işlemi (post-pruning technique) yapma fırsatını verir: birleştirilen sıralı erişim çıktısı üzerinden bir geçiş yapılarak ilgili terimi içeren belgeler sonuç listeden çıkarılır.

Şimdi de Boole modelinde referansların nasıl işleneceğini tartışalım.

Internet yapısal açıdan yorumlanacaksa yönlü çizge (ya da hiper-metin veri tabanı) olarak düşünülebilir. Bu bağlamda bir Web belgesinin (daha genel olarak Internet kaynağının) bir uzaklıktaki komşuluk kümesini ilgili belgeye bağlantı veren ya da ilgili belgenin bağlantı verdiği belgelerin kümesi olarak tanımlayalım. Bu kavram *yakın komşu* (D_r) ile gösterilsin. Referans ilişkisinin Internet ortamında yakın komşuluk ilişkisi ile özdeş olduğunu düşünelim.[14] Bu durumda, yukarda verilen iç çarpımdaki belge terimi ağırlığı aşağıdaki gibi düşünülebilir:

Sorgu terimi belge terimleri içinde ise $a_{ri} = c_1$; sorgu terimi yakın komşuluk içindeki belgelerin herhangi birisinde geçiyorsa c_2; aksi tadirde *0*. c_1 ve c_2 sabitleri tasarımcının ilgili yapısal benzerlikleri nasıl ağırlıklandıracağına bağlı olarak değişir. Örneğin c_2 değeri tayin edilirken referans edilen/eden referans sayısı (Google motoru tarafından tutulmaktadır) veya referans eden/edilen belgenin kalitesi hesaba katılabilir.

3.5 Arama Motorlarında Performans Değerlendirmeyle İlgili Çalışmalar

[14] Dikkatli okuyucu referans ile yakın komşuluk ilişkilerinin özdeşliğinin gerçek hayattaki durumu yansıtmaktan uzak olduğunu düşünebilir. Kaba bir sınıflama ile her bir bağlantı ya organizasyon türü (bir sonraki, bir önceki, üstteki, ev, vb.) ya da çeşitli anlamsal ilişkileri içine alan referans türü (genelleştirme/özelleştirme veya alt/üst bileşen içinde düşünülebilir (Frei ve Stieger, 1995). Bu açıdan düşünüldüğünde, her bir bağlantının referans etme/edilme anlamına gelmeyeceği bir gerçektir; fakat her bir soyutlamanın kendi içinde yanlışlık içerebileceği düşünülerek basitlik uğruna yukarıdaki özdeşliğin geçerli olduğu varsayılabilir.

Bundan önceki alt bölümlerde arama motorlarının çeşitli yönleriyle ilgili araştırmalara yeri geldikçe değinildi. Bu alt bölümde arama motorlarında bilgi erişim performansının değerlendirilmesiyle doğrudan ilgili çalışmalar kısaca özetlenmektedir.

Geleneksel bilgi erişim sistemlerinin performans değerlendirmesinde kullanılan anma ve duyarlık gibi ölçümler arama motorlarının performans değerlendirmesinde de genellikle kullanılmaktadır. Fakat, aşağıda da açıklandığı gibi, arama motorlarının kendine özgü özelliklerinden dolayı anma ve duyarlık ölçümlerinde bazı değişiklikler yapılması gerekmektedir. Bunun yanı sıra, yapılan araştırmalarda arama motorlarının kapsam, güncellik ya da kırık bağlantılar (broken links), yanıt süresi, insan faktörleri ve kullanıcı arayüzü gibi ölçütler yönünden de incelendiği görülmektedir (Oppenheim, Morris ve McKnight, 2000).

Anma, bilindiği gibi, erişilen ilgili belgelerin derlemdeki toplam ilgili belgelere oranını Arama motorları tipik bilgi erişim sistemleriyle karşılaştırılamayacak kadar büyük hacimli belge derlemleri üzerinde aramalar gerçekleştirdiklerinden, belirli bir soru için derlemdeki toplam ilgili belge sayısını bulmak hemen hemen olanaksızdır. Buna benzer bir sorunla daha önce yüz yüze gelen TREC (Text REtrieval Conference) konferansları (http://trec.nist.gov/), sorunu "havuzlama" yöntemi ile çözmeye çalışmışlardır.[15] Bu yönteme göre, bir bilgi ihtiyacı ile ilişkili her bir işlemenin[16] (run) sonucunda dönen 1000 belgeden oluşan erişim çıktısının

[15] Bilgi erişim sistemlerinin değerlendirilmesinde yöntemler ve kalite testleri (benchmark collections) yönünden geçmişten gelen oldukça zengin bir birikim vardır (Sparck Jones, 1971; Salton, 1971). Bilinen test derlemleri CACM, CISI, Cranfield ve NPL olup, tam bilgi verirler; yani, sorgular ve belgeler terim vektörleri cinsinden tanımlı olup, her bir sorgu için ilgili belgeler liste halinde tutulur (bkz: ftp://ftp.cs.cornell.edu/pub/smart/). Bu testler bilgi erişim alanında karşılaşılan meydan okuyucu sorunların çözümü doğrultusunda oluşturulan yeni modellerin test edilmesinde ve, daha önemlisi, ortak bazda karşılaştırılmasında zamanla yetersiz kalmışlardır. Bü nedenle, 1990'da Amerikan İleri Savunma Araştırma Projeleri Ajansı'nın (DARPA) TIPSTER metin projesi (http://www.nist.gov/itl/div894/894.02/related_projects/tipster/) çerçevesinde, Ulusal Standartlar ve Teknoloji Enstitüsü'nün (NIST: National Institute of Standards and Technology) bilgi erişim teknolojilerini değerlendirmede kullanılmak üzere çok geniş bir metin (ya da genel olarak belge) derlemi oluşturması istendi (Voorhees ve Harman, 1999). İlk TREC konferansı 1992 yılında ticari kuruluşların ve çoğu DARPA veya NIST tarafından desteklenen akademik çevrelerin katılımıyla gerçekleştiğinde, eldeki derlem 2GB büyüklüğündeki yaklaşık bir milyon belgeden oluşuyordu (1998'e kadar süren TIPSTER programı 4 ciltlik Tipster CD'leri ile anılmaktadır). Ticari ve akademik bilgi erişim sistemlerinin test yatağı (test bed) olarak hizmet veren TREC, ulusal kimlikten sıyrılarak zamanla uluslararası bir yarış arenası haline dönüşmüştür. (2000 yılının Kasım ayında yapılan 9. TREC konferansına 17 ülkeden 69 akademik veya ticari grup katılmıştır). Yeni modellerin ya da tekniklerin denendiği bu konferanslar birkaç ana görev (task) ve kimisi sonradan ana görev olan bir çok izlerden (tracks) oluşmaktadır. İşte bu görevlerden birisi olan 'ad hoc' (bilgi ihtiyaçlarından oluşturulan sorgular aracılığı ile belgeler derlemini araştıran ve ilgili olduğuna inanılan belgelerin bir belge erişim çıktısı içersinde düzenlenerek geri getirilmesi sürecini yöneten sistemlerin başarılarının incelenmesi) TREC-8'den sonra yerini Web erişim izine bıraktığında Web için oluşturulan derlemin büyüklüğü 100 GB büyüklüğünde 18.5 milyon sayfadan oluşuyordu.

[16] Bir bilgi erişim sistemi (ya da arama motoru) bir göreve ya da ize birden fazla katılabilir. Örneğin, 'ad hoc' görevinde bir bilgi ihtiyacı (TREC terminolojisinde "konu" olarak adlandırılır) başlık, tanım, açıklama (narrative) ve kavramlar (TREC-2'den sonra "kavramlar"dan vazgeçildi) yapılarından oluşan bir mizanpajla ifade ediliyordu. Bir sistem yalnızca başlığı ya da tüm kısımları otomatik olarak ya da elle (orijinal ya da genişletilmiş Boole ya da geribildirim teknikleri ile sorguların genişletilmesi yolu) işleyerek sorguları oluşturabilir. Herhangi bir kombinasyon bir "işleme" olarak anılır.

ilk 100 belgesi bir havuzda toplanır. Bir değerlendirici, ki çoğunlukla bilgi ihtiyacını oluşturan kişidir, havuzda toplanan tekil (unique) belgelerin (ki konu başına ortalama 1500-2000 civarındadır) üzerinden geçerek ilgili belgeleri saptar. Eldeki derlemde bunlar dışında ilgili belge olmadığı kabul edilir ve bununla birlikte 1000'lik erişim çıktısı kullanılarak her bir işlemenin ilgili konuya göre anma ve duyarlık değerleri hesaplanır. Buradaki espri iki temel varsayıma dayanmaktadır: (1) İlgili belgeler büyük bir olasılıkla üst sıralara (örneğin, erişim çıktısının %10'luk kesimi) yerleşecektir (Voorhees ve Harman, 2000); ve (2) Kullanılan birbirinden oldukça farklı arama stratejileri sonucu farklı belgelere erişim sağlanacaktır (Lee, 1997; 1995; Belkin et al., 1995). Bu iki varsayım zaman içinde çeşitli deneylerle doğrulanmıştır.

Havuzlama yöntemine benzer bir başka yöntem de gerçek hayatta işletimde olan arama motorlarının ortalama anma değerlerinin hesaplanmasında kullanılmak üzere Clarke ve Willet (1997) tarafından önerilen "göreli anma" (relative recall) değeridir. Bu yöntem bir arama motoru tarafından bulunan ilgili belgelerin diğer arama motorlarının bulduğu ilk belgeler arasında yer alıp almadığının kontrol edilmesine dayanmaktadır.

Arama motorlarında duyarlık değerlerinin ölçülmesi geleneksel bilgi erişim sistemlerinden biraz farklılık göstermektedir. Geleneksel sistemlerde çoğu zaman erişilen tüm belgelere bakarak duyarlık değeri hesaplanırken, arama motorlarında ise erişilen belge sayısının çok yüksek olması ve bu belgelerin hepsinin tek tek değerlendirilememesi nedeniyle belirli kesme (cut-off) noktalarında duyarlık değerlerinin hesaplanması yoluna gidilmektedir. Bir başka deyişle, belirli bir soru için erişim çıktısında yer alan tüm belgeler üzerinden duyarlık değerini hesaplamak yerine, belirli sayıda (5, 10, 15, 20... gibi) belge görüldükten sonra her aşamada duyarlık değerlerinin nasıl değiştiği hesaplanmaktadır. Buradaki varsayım, çoğu arama motoru kullanıcılarının erişim çıktısında yer alan belgelerin çok azını (bir ya da iki ekran dolusu) görmek istemeleridir. Nitekim, yapılan araştırmalarda bu varsayımın geçerliliği kanıtlanmış, kullanıcıların gözden geçirdikleri ekran sayısı ortalama 1,39 (standart sapma 3,74) olarak bulunmuştur (Silverstein et al. 1999). Konuyla ilgili bir başka çalışmada (Jansen et al., 1998) ise kullanıcıların ilk ve ikinci ekranları görme oranı sırasıyla %58 ve %19 olarak bulunmuştur.

Geleneksel bilgi erişim sistemleriyle arama motorları arasındaki önemli farklardan birisi de sorgu cümlelerinde kullanılan ortalama sözcük sayısıdır. Tipik bir bilgi erişim sisteminde sorgu ifadelerinde ortalama 7,9 ile 14,95 sözcük yer almasına (Jansen et al., 1998) rağmen, arama motorlarına girilen sorgularda bu rakam ortalama 2,3 civarındadır (Silverstein et al., 1999; Kirsch, 1998; Jansen et al., 1998). Bu durumu Infoseek şirketinin başkanı S. Kirsch,

“Web kullanıcıları bir-iki kelimelik sorgularıyla bizden mucizeler yaratmamızı bekliyorlar” diye alaycı bir şekilde özetlemiştir (Kirsch, 1998). Gerçekten de arama motorlarının işlem kütükleri kullanılarak yapılan araştırmalarda en popüler sorguların tek sözcükten oluşan sorgular olduğu görülmektedir. Örneğin, aralarında "sex", "Playboy", "Penthouse", "chat", "nude", "porn", "erotica", "games" gibi sözcüklerin de bulunduğu toplam 15 sözcük Infoseek'te yapılan bütün aramaların %12'sini oluşturmaktadır (Kirsch, 1998). AltaVista'da yapılan yaklaşık bir milyar arama sorusunun incelenmesinden de benzer sonuçlar elde edilmiş, sırasıyla "sex", "applet", "porno", "mp3" ve "chat" gibi tek sözcükten oluşan sorular en sık aranan sözcükler olmuştur (Silverstein et al., 1999). Arama motorları, tek sözcükle arama yapma konusundaki bu meydan okumayı, Web kullanıcılarının tipik olarak anmadan çok duyarlık ile ilgilendiği ilkesini de göz önünde bulundurarak, çok referans alan sayfalara öncelik verme yolunu seçerek karşılamaya çalışmaktadır.

Arama motorlarında performans değerlendirmesi konusunda bu zamana dek yapılan araştırmalar birkaç çalışmada topluca özetlenmiştir (Oppenheim et al., 2000, s. 14, 23; Soydal, 2000).

Konuyla ilgili olarak yapılan ilk çalışmalardan birisinde Gudivada ve diğerleri (1997) iki soruyu (“latex software” ve “multiagent system architecture”) 13 farklı arama motoru üzerinde Boole işleçlerini kullanarak ve tamlama olarak ayrı ayrı aramışlar ve elde ettikleri sonuçları erişilen belge sayıları açısından karşılaştırmışlardır. Erişim çıktılarında ilgili belgelerin ilgisiz belgeler arasında dağıldığı görülmüş, bu nedenle kullanıcıların salt sıralamada başta gelen belgelere bakmalarının yeterli olmayacağı sonucuna varılmıştır. Arama motorlarının, kapsamları birbirinden farklı dizinler üzerinde arama yapmaları nedeniyle bu çalışmada performans değerlendirme ölçümleri kullanılmamıştır.

Chu ve Rosenthal’ın (1996) çalışması geleneksel performans değerlendirme ölçümlerinden duyarlığın kullanıldığı ilk araştırmalardan birisidir. Araştırmacılar AltaVista, Excite ve Lycos üzerinde gerçekleştirilen 10 arama sorgusu için duyarlık oranlarını sırasıyla %78, %55 ve %45 bulmuşlardır. Benzer bir çalışmada Leighton ve Srivastava (1999) 15 soru için erişilen ilk 20 Web sitesi üzerinden AltaVista, Excite, HotBot, Infoseek ve Lycos’un duyarlık değerlerini hesaplamışlardır. AltaVista, Excite ve Infoseek’in daha iyi performans gösterdikleri (%50’nin üzerinde), Lycos’un kısa ve yapılanmamış sorularda, HotBot’un ise yapılanmış sorularda daha başarılı olduğu görülmüştür.

AltaVista, Yahoo! gibi popüler arama motorlarının günümüzde yüz milyonlarca Web sayfasını dizinledikleri bilinmektedir. Bu tür büyük derlemlerde kesin anma (absolute recall) değerini hesaplamak için gerekli olan derlemdeki toplam ilgili belge sayısını bulmak hemen

hemen olanaksız olduğundan, yapılan ilk çalışmalarda anma değerlerinin ölçülmesi yoluna gidilmediği görülmektedir. Her arama motorunun farklı Web sayfalarını dizinlemesi, farklı arama motorları için elde edilen performans değerlerini karşılaştırmayı da güçleştirmektedir. Clarke ve Willet (1997) göreli anma (relative recall) değerini kullanarak AltaVista, Excite ve Lycos üzerinde 30 soruya dayanan bir araştırma gerçekleştirmişlerdir. Bu araştırmada söz konusu arama motorları için bulunan ortalama anma değerlerinin (yaklaşık %60), geleneksel bilgi erişim sistemlerinde genelde elde edilen sonuçların aksine, ortalama duyarlık değerlerinden (%35) daha yüksek olduğu görülmüştür. Anma değerleri açısından söz konusu arama motorları arasında istatistiksel açıdan anlamlı bir faklılık yoktur. Duyarlık açısından ise AltaVista (%46) ile Lycos (%25) arasındaki performans değerleri istatistiksel açıdan anlamlı bulunmuştur.

Göreli anma değerlerinin kullanıldığı bir başka araştırma Gordon ve Pathak (1999) tarafından gerçekleştirilmiştir. Araştırmacılar gerçek bilgi gereksinimlerinden kaynaklanan toplam 33 soruyu sekiz farklı arama motoru üzerinde deneyerek, bilgiye gereksinim duyan deneklerin yaptığı ilgililik değerlendirmelerine göre çeşitli kesme (cut-off) noktalarında anma ve duyarlık değerlerini hesaplamışlardır.[17] Buna göre çeşitli arama motorlarında erişilen ilk 10 belgede duyarlık değerleri %41 (AltaVista) ile %18 (Yahoo!), anma değerleri ise (erişilen ilk 15-25 belgede) %16 (AltaVista) ile %6 (Yahoo!) arasında değişmektedir.

Soydal (2000) AltaVista, Excite, HotBot, Infoseek ve Northern Light üzerinde gerçekleştirdiği bir araştırmada erişilen ilk 10 ve ilk 20 belge üzerinden ortalama (göreli) anma ve duyarlık değerlerini hesaplamıştır. Adı geçen arama motorları arasında ortalama duyarlık değerleri (yaklaşık %50) açısından anlamlı bir farklılık olmadığı görülmüştür. Ortalama anma değerleri ise %14 (Infoseek) ile %31 (Northern Light) arasında değişmektedir. Infoseek ile Northern Light arasındaki anma değerleri istatistiksel açıdan anlamlı bulunmuştur.

Yukarıda (3.3) Web sayfalarının hazırlanmasında yazar, anahtar sözcük, tanım vb. gibi HTML üst veri belirteçlerinin (meta tags) belgelerin içeriğini tanımlamada kullanıldığından söz etmiş ve arama motorlarının erişim amacıyla bu alanlardan yeterince yararlanmadığını vurgulamıştık. Web belgelerinin hazırlanmasında HTML üst veri belirteçleri kullanımının arama motorlarında erişim etkinliğini artırıp artırmadığı çeşitli araştırmalara konu olmuştur. Turner ve Brackbill (1998) AltaVista ve Infoseek üzerinde yaptıkları kontrollü araştırmada anahtar sözcük (keyword) üst veri belirtecinin kullanıldığı belgelerde üst veri belirteci

[17] TREC derlemiyle çalışan Web erişim grubundaki araştırmacılar da duyarlık değerlerini kesme noktası kullanarak hesaplamışlardır (bkz. Hawking, Craswell, Thislewaite ve Harman, 1999).

kullanılmayanlara oranla erişilebilirliğinin önemli ölçüde arttığını saptamışlardır. Ancak, popüler arama motorları kullanılarak yapılan bir başka kontrollü araştırmada üst veri belirteçlerinin kullanımının erişim sonuçlarını pek etkilemediği ortaya çıkmıştır. Elektronik bir dergi olan *First Monday*'de (http://www.firstmonday.dk) yayımlanan ve üst veri belirteçleri boş olan makalelere arama motorları kullanılarak erişim sağlanmışır. Daha sonra ise bu makalelere üst veri belirteçleri eklenmiş ve aramalar tekrarlanarak söz konusu makalelerin erişim çıktısında daha üst sıralarda yer alıp almadıkları test edilmiştir. Yapılan testlerde üst veri belirteçlerinin kullanımının erişim sıralamasını tek başına etkilemediği görülmüştür (Henshaw ve Valauskas, 2001). Anlaşıldığı kadarıyla, Web sayfalarının hazırlanmasında üst veri belirteçlerinin kullanımı açısından henüz bir standartlaşmaya gidilmediğinden, çoğu arama motorları üst veri belirteçlerini erişim sırasında dikkate almamaktadırlar.

Çeşitli araştırmacılar arama motorlarında çeşitli erişim ve sıralama algoritmalarının performanslarını değerlendirmişlerdir. Savoy ve Picard (2001) basit anahtar sözcüğe dayalı dizinleme stratejilerinin terim sıklığına dayanan dizinleme stratejilerinden daha başarılı olduğunu, sorgu cümlesinde daha fazla anahtar sözcük kullanmanın ortalama duyarlığı artırdığını, dur listesi kullanmanın erişim etkinliğini artırdığını, TREC 8'de kullanılan bilgi erişim modellerinin yaklaşık 2 GB'lık Web derlemi üzerinde de yüksek performans sergilediğini, Web sayfası başlığında yer alan terimleri ağırlıklandırmanın ortalama duyarlık üzerinde önemli bir etkisi olmadığını, sadece başlıkta yer alan terimlerin dizinlenmesinin erişim etkinliğini zayıflattığını, gövdeleme kullanılmadığında çoğu arama stratejilerinde ortalama duyarlığın önemli ölçüde düştüğünü bulmuşlardır. Yuwono ve Lee'nin (1996) araştırmasında ise vektör uzayı modeline dayalı erişim algoritmalarının daha başarılı sonuçlar verdiği, sadece üst veri alanlarında yer alan bilgilere dayanan algortimaların, sezgisel olmalarına rağmen, pek başarılı olmadığı ortaya çıkmıştır.

Arama motorları tarafından erişilen ilgili belgeler arasındaki çakışma oranı (overlap) çeşitli araştırmalara konu olmuştur. Yukarıda anılan Gordon ve Pathak'ın (1999) çalışmasında yedi arama motoru arasındaki çakışma oranı sadece %7 olarak bulunmuştur. Soydal'ın (2000) çalışmasında da beş arama motoru için benzer bir sonuç (%11) elde edilmiştir. Bharat ve Broder (1998) ise dört arama motoru (AltaVista, HotBot, Excite ve Infoseek) arasındaki çakışma oranının %1'den az olduğunu bulmuştur. 1997 yılında söz konusu dört arama motoru tarafından dizinlenen toplam 200 milyon civarındaki Web sayfasından sadece yaklaşık iki milyonu dört arama motoru tarafından da dizinlenmiştir. Bir başka deyişle, bu bulgular farklı arama motorlarının Web uzayında farklı ilgili belgelere erişim sağladığını ortaya çıkmaktadır.

4 YÖNTEM VE TASARIM

Bu bölümde Türkçe arama motorları üzerinde gerçekleştirdiğimiz bilgi erişim performans değerlendirmesi deneyiyle ilgili olarak; araştırma soruları, arama motorlarının ve soruların seçimi, soruların formüle edilmesi, verilerin toplanması ve analizi, ilgililik (relevance) değerlendirmesi, duyarlık, normalize sıralama, kapsama ve yenilik oranlarının hesaplanması, verilerin analizi ve sonuçların değerlendirilmesiyle ilgili ayrıntılı bilgiler verilmektedir.

4.1 Araştırma Soruları

Araştırmamızda aşağıdaki sorulara yanıt aranmaktadır:

- Türkçe arama motorları sorulan sorularla ilgili belgelere erişmede ne kadar başarılıdır? Arama motorlarının duyarlık performansları arasında fark var mıdır?
- Türkçe arama motorları ilgili belgelere mümkün olduğunca erişim çıktısının ilk sıralarında erişmede ne kadar başarılıdır? Arama motorlarının normalize sıralama performansları arasında fark var mıdır?
- Türkçe arama motorlarının eriştikleri belgelerin ne kadarı "canlı"dır? Başka bir deyişle, çeşitli nedenlerle erişilemeyen bağlantıların erişilen belgelere oranı nedir? Arama motorları arasında bağlantıların güncelliği açısından fark var mıdır?
- Türkçe arama motorları en sık aranan sözcüklerin ne kadarını kapsamaktadır? Arama motorlarının kapsama oranları arasında fark var mıdır?
- Türkçe arama motorları HTML belgelerinde gömülü "anahtar sözcük", "tanımlama" gibi dizinleme bilgisi içeren üst veri (metadata) belirteçlerinden ne ölçüde yararlanmaktadır?
- Türkçe arama motorlarında "ç", "ş", "ü" gibi Türkçeye özgü karakterler kullanılarak yapılan aramalarda sorun var mıdır?
- Türkçe arama motorlarında en sık aranan sorular ("mp3", "oyun", "sex", "erotik" ve "porno") için dört arama motorunun ilgili belgeleri kapsama oranları birbirinden farklı mıdır?
- Türkçe arama motorlarında en sık aranan sorular için dört arama motorunun Türkiye adresli belgeleri kapsama oranları birbirinden farklı mıdır?
- Türkçe arama motorlarında en sık aranan sorular için dört arama motoru birbirinden farklı ("yeni") ilgili belgelere erişmekte midir? Arama motorlarının yenilik oranları birbirinden farklı mıdır?
- Türkçe arama motorlarında en sık aranan sorular için dört arama motoru birbirinden farklı ("yeni") Türkiye adresli ilgili belgelere erişmekte midir? Arama motorlarının Türkiye adresli ilgili belge bulmadaki başarıları (yenilik oranları) birbirinden farklı mıdır?

- Türkçe arama motorlarının sorgu ifadelerinde yer alan terimler arasındaki belirli cebrik ilişkileri sonuç erişim çıktılarında koruma bakımından Türkçe arama motorları birbirinden farklı mıdır?

- Türkçe arama motorlarının dar kapsamlı sorular için ilgili belgelere erişilebilme kapasiteleri nedir? Arama motorları arasında bu bakımdan fark var mıdır?
- Türkçe arama motorlarının geniş kapsamlı sorular için ilgili belgelere erişebilme ve ilgisizleri ayırt edebilme kapasiteleri nedir? Arama motorları arasında bu bakımdan fark var mıdır?
- Türkçe arama motorlarında Türkçe gövdeleme algoritması kullanılmakta mıdır?
- Türkçe arama motorlarının özellikleri (yardım, görüntüleme, ileri düzey arama komutlarını ve Boole işleçlerini yorumlayabilme, vd.) nelerdir?

4.2 Türkçe Arama Motorları Listesi

Türkçe arama motorlarında performans değerlendirmesi yapmak amacıyla ülkemizde kullanılan popüler arama motorlarından Arabul, Arama, Netbul ve Superonline seçilmiştir. Aşağıda bu arama motorlarıyla ilgili bilgiler verilmektedir.

- *Arabul (http://www.arabul.com):* 1996 Kasımında aktif hale gelmiş Türkiye ve Türklerle ilgili Web sitelerini dil, içerik, kapsam ayırt etmeden düzenli bir kategorik yapı içinde sunan ve Türkçe karakterler kullanarak arama yapma olanağı sağlayan bir arama motorudur.
- *Arama (http://www.arama.com):* Tüm Web'de ya da Arama'nın kategorileri üzerinde ve Türkçe karakter kullanarak arama yapma olanağı sağlayan bir arama motorudur.
- *Netbul (http://www.netbul.com):* Arama motoru olarak HotBot'u (www.hotbot.com) kullanmakta olan Netbul, kategorilerinde, Internet Rehberinde ve Internet üzerinde arama olanağı sunan ve ayrıca resim arama özelliği de bulunan bir arama motorudur.
- *Superonline (http://www.superonline.com):* Arama motoru olarak AltaVista'yı (www.altavista.com) kullanmakta olan Superonline, isteğe bağlı olarak site içerisinde, sadece Türkçe siteler içerisinde ve/veya tüm Web'de Türkçe karakterler kullanarak arama yapma olanağı vermektedir. Superonline resim, mp3/ses ve video aramalarına da imkan vermektedir.

Arama motorlarına ilişkin çeşitli özellikler aşağıda beş başlık altında incelenmekte ve araştırma için seçtiğimiz dört arama motoru söz konusu özellikler yönünden karşılaştırılmaktadır.

4.2.1 Düzenli İfadeler

Bu grupta tekil artı (‘+’) ve tekil eksi (‘-’) işleçlerinin yanı sıra, herhangi bir terimle eşleştir (Match Any Term) ve bütün terimlerle eşleştir (Match All Terms) seçenekleri ve isim tamlamaları (phrase) (“<tamlama>” ile gösterilir) ile ilgili özellikler açıklanmakta ve dört arama motoru söz konusu özellikler yönünden Tablo 3'te karşılaştırılmaktadır.

+<terim> : Terimi içeren belgelerin arama motoru tarafından döndürülmesini belirtir. Örneğin, +kıbrıs şeklindeki sorgu ifadesi Kıbrıs terimini içeren tüm sayfaların erişim çıktısına ilave edilmesine yol açar -ki bu anlamda bir Boole ‘ve’ işleci işlevi görür.

-<terim>: Terimi içeren belgelerin arama motoru tarafından döndürülmemesini sağlayan matematiksel komut. Örneğin, ‘çiçek –kıbrıs’ şeklindeki sorgu ifadesi “çiçek” terimini içeren fakat “kıbrıs” terimini içermeyen belgelerin istendiğini belirtir.

“<tamlama>“ : Arama motorunun tırnak işaretleri arasında bulunan tamlamayı içeren sayfaları sonuç olarak geri döndürmesini sağlayan komut. Örnek: “bilgi erişim sistemleri”.

Herhangi bir terim ile eşleştirme (Match Any Term): Arama motorunun sorgu ifadesinde belirtilen arama terimlerinden herhangi birisini içeren sayfaları sonuç olarak geri döndürmesi.

Bütün terimlerle eşleştirme (Match All Terms): Arama motorunun sorgu ifadesinde belirtilen tüm arama terimlerini içeren sayfaları sonuç olarak geri döndürmesi.

Tablo 3. Matematiksel komutlar

Komut	Kullanım şekli	Arabul	Arama	Netbul	Superonline
Terimi ekle	+	✓	✓	✓	✓
Terimi çıkar	-	✓	✓	✓	✓
Tamlama	“ “	✓	✓	✓	✓
Sorguda geçen herhangi bir terimle eşleştir	Otomatik		✓	✓	✓
	Diğer	✓			
Sorguda geçen tüm terimlerle eşleştir	Otomatik	✓			
	Diğer		✓	✓	✓

✓ : Özelliğin bulunduğunu belirtir.

4.2.2 İleri Düzey Arama Komutları

Bir belgeyi başlık, site ve URL adresiyle sorgulamada ileri düzey arama komutları kullanılır. İlgili arama ifadesinde joker karakterleri (*, ?) de kullanılabilir. Aşağıda ileri düzey arama

komutları açıklanmakta ve dört arama motoru bu komutlar açısından Tablo 4'te karşılaştırılmaktadır.

Başlık Araması: Arama motorunun sorguda geçen kelime veya kelimeleri sayfa başlıkları içerisinde aramasıdır. Örneğin, *title:"Ev Sayfası"*; sayfa başlığı içerisinde "Ev Sayfası" kelimesi geçen sayfaları arar.

Site Araması: Arama motorunun sorguyu belirli bir bilgisayar üzerinde bulunan site içerisinde araması. Örneğin, *+host:cmpe.emu.edu.tr +personel*; "cmpe.emu.edu.tr" sitesi içerisinde 'personel' kelimesini arar.

URL Araması: Arama motorunun sorguda geçen kelime veya kelimeleri belirtilen URL içerisinde aramasıdır. Örneğin, *+url:.gov +turkey*; URL'i içerisinde ".gov" olan tüm sayfalarda "turkey" sözcüğünü arar.

Bağlantı Araması: Arama motorunun sorguda geçen URL'e referans veren sayfaları aramasıdır. Örneğin, *link:cmpe.emu.edu.tr*; "cmpe.emu.edu.tr"a bağı olan sayfaları arar.

'':* Sorgu içerisinde bir veya birden fazla bilinmeyen harfi temsil eder. Örneğin, *+portakallı +meyve**; 'portakallı' teriminin ve 'meyve' ile başlayan terimlerin bulunduğu sayfaları bulur.

'?': Sorgu içerisinde tek bir bilinmeyen harfi temsil eder. Örnek: çiçek??; yedi harften oluşan ve ilk beş harfi "çiçek" olan terimin geçtiği sayfaları bulur.

Tablo 4. İleri düzey komutları

Komut	**Kullanım şekli**	**Arabul**	**Arama**	**Netbul**	**Superonline**
Başlık araması	title:	--	--	✓	✓
	Diğer	✓ mönüden seçim	--	--	--
Site araması	Domain:	--	--	✓	--
	Diğer	✓ mönüden seçim	--	--	✓ host:
URL araması	url:	--	--	--	✓
	Diğer	✓ mönüden seçim	--	--	--
Bağlantı araması	link:	--	--	--	✓
	diğer:	--	--	✓ Linkdomain:	--
Joker	*	--	✓	✓	✓
	?	--	--	--	--

✓ : Özelliğin bulunduğunu belirtir. -- : Özelliğin bulunmadığını belirtir.

Arabul arama motorunda, belirtilen alanda (domain) arama yapılabilme özelliğinde bir aksaklık olduğu gözlenmiştir. Örneğin, detaylı arama kullanıldığında ve "Sadece aşağıdaki site ya da domain'deki sayfalardan araştır" kısmına "com.tr" yazılıp "bitirim" sözcüğü aratıldığı zaman "http://members.nbci.com/bitirimteam" veya "http://www.bitirimteam.da.ru" gibi sonu "com.tr" ile bitmeyen bağlantı adreslerine de erişim çıktısında yer verildiği görülmüştür.

4.2.3 Arama Yardımı Özellikleri

Arama yardımı özellikleri kapsamında arama motorlarında kullanılan 'İlgili Arama' (Related Search), 'Kümeleme' (Clustering), 'Benzer Bulma' (Find Similar), 'Gövdeleme' (Stemming), 'Tarih Sınırlaması' (Date Range), 'İçinde Arama' (Search Within), 'Büyük/Küçük Karaktere Duyarlılık' (Case Sensitivity), 'Tek Sonuç/Popülarite Sıralaması' (Direct Hit/Popularity Ranking), 'Ticari İsim Bağlantısı' (RealNames Link), 'Kaynak Türü Seçimi' (Source Type Selection), 'Web/Türkçe Siteler/Kılavuz içinde Arama' (Search Web/Turkish Sites/Categories), 'Puanlama', ve 'Sınıflandırma' özellikleri açıklanmakta ve dört arama motoru bu özellikler açısından Tablo 5'te karşılaştırılmaktadır.

İlgili Arama: Sonuç olarak erişilen belgenin URL'sine bağı (link) olan sayfaları arama özelliğidir. O belge için "link:" komutunun çalıştırılmasını otomatik olarak yapar.

Kümeleme: Kümeleme yaparak her siteye ait sadece tek bir belgenin sonuç sayfasında görülmesini sağlayan özelliktir. "site-adı.com" ve "www.site-adı.com" iki farklı küme olarak değerlendirilir.

Benzeyenleri Bulma: Döndürülen belgenin içerik olarak benzeri olan diğer belgeleri arama özelliğidir.

Gövdeleme: Arama motorunun arama yapılacak olan kelimenin kökünü alıp, kök kısmını kullanarak kelimenin değişik biçimleriyle arama yapmasıdır. Örneğin, sorgudaki kelime "compute" ise "computing" kelimesinin geçtiği belgeler de bulunur.

Tarih Sınırlaması: Belirtilen iki tarih arasında yayımlanan sayfaları bulabilme özelliğidir.

İçinde Arama: Arama motoru tarafından erişilen belgeler arasında arama yapabilme özelliğidir.

Büyük/Küçük Karaktere Duyarlılık: Küçük ve büyük harflere karşı arama motorunun duyarlı olabilme özelliğidir. Örneğin, sorguya yazılan "mısır" kelimesi için arama motoru "MISIR", "MısıR", "mıSır" gibi tüm kelimeleri arar. "MISIR" yazıldığında ise arama motoru sadece içinde büyük harfle "MISIR" kelimesi geçen belgeleri arar.

Tek Sonuç/Popülarite Sıralaması: Arama motorunun sitelerin kaç kere ziyaret edildiği ve ziyaret süreleriyle ilgili bilgileri değerlendirip bu bilgileri kullanarak en popüler siteleri gösterebilme özelliğidir.

Ticari İsim Bağlantısı: Sorguda aranan kelime eğer bir firma tarafından kendi adına kayıtlı ise arama motorunun kayıtlı firmaya direkt olarak bağ (link) vermesidir. Örneğin, "nike" kelimesi girildiği zaman arama motoru Nike firmasının direkt Web adresini sonuç sayfasında verir.

Kaynak Türü Seçimi: Aranacak olan kaynağın türünün (mp3/video/resim vb. gibi) belirlenmesini sağlayan bir özelliktir.

Web/Türkçe Siteler/Kılavuz içinde Arama: Arama motorunun tüm Web'i veya sadece Türkçe siteleri veya kendi içerisinde bulunan kategorileri arayabilme özelliğidir.

Puanlama: Arama motorunun bulduğu belgelere ilgililik puanı verebilme özelliğidir. Erişilen belgenin ilgililik oranı genellikle yüzde ya da 1 ile 1000 arasında değişen bir sayıyla belirtilmektedir.

Sınıflandırma: Arama motorunun bulduğu sonuçları sınıflandırabilme özelliğidir.

Tablo 5. Arama yardımı özellikleri

Özellik		Arabul	Arama	Netbul	Superonline
İlgili aramalar		--	--	--	--
Öbekleme		--	--	--	✓
Benzeyenleri bulma		--	--	--	--
Gövdeleme		--	--	--	--
Tarih değişimi		--	--	--	✓
İçinde arama		--	--	--	--
Küçük harf/büyük harf duyarlılığı		--	✓	--	✓
Kesin sonuç/Popülarite sıralaması		--	--	--	--
Ticari isim bağlantısı		--	--	--	--
Kaynak türü seçimi		--	--	✓ Sadece resim	✓
Arama	tüm Web	✓	✓	✓	✓
	Türkçe siteler	--	--	✓	✓
	Kategoriler	✓	✓	✓	--
Puanlama		--	--	--	✓ Web'de arama dışında
Sınıflandırma		✓	--	--	--

✓ : Özelliğin bulunduğunu belirtir. -- : Özelliğin bulunmadığını belirtir.

Arama ve Netbul arama motorlarında Boole "VEYA" işleci yardım sayfalarında belirtildiği şekilde çalışmamaktadır. Arama'nın "yardım" sayfasında "İki kelimeden bir tanesinin geçtiği sayfaları bulmak için, iki kelimeyi yan yana bir boşluk bırakarak veya iki kelimenin arasına "|" (veya) işareti koyarak yapabilirsiniz" denmektedir. Dolayısıyla, örneğin, "<sözcük1> <sözcük2>" veya "<sözcük1> | <sözcük2>" sorguları arama motorunda aratıldığında, "<sözcük1>" sözcüğü aratıldığında erişilen liste ile "<sözcük2>" sözcüğü aratıldığında erişilen listenin birleşiminin erişim çıktısı olarak döndürülmesi beklenir. Fakat sistem sadece "<sözcük1>" sözcüğü aratılmış gibi bir erişim çıktısı vermektedir. Benzeri bir biçimde Netbul'un "yardım" sayfasında da "Birden fazla kelimenin işaretsiz olarak yan yana yazılması, VEYA anlamına gelir" denmektedir. Fakat Netbul'da "<sözcük1> <sözcük2>" gibi bir sorgu çalıştırıldığında hiçbir belgeye erişilememektedir.

4.2.4 Erişim Çıktısı Görüntüleme Özellikleri

Varsayılan (default) sayıda erişilen belgenin gösterilmesi, gösterilen belge sayısının artırılması, sonuçların belirli ölçütlere (başlığa göre alfabetik, belgenin yaratılış tarihine göre, vd.) göre sıralanması, Belgelerin belirli kısımlarının (örneğin, sadece başlıklar) gösterilmesi, belge büyüklüğü, erişilen toplam belge sayısı, son güncelleme tarihi vs. Arama motorlarının erişim çıktılarını görüntülemede kullandığı ölçütlerden bazılarıdır. Tablo 6'da ilgili özellikler açısından dört arama motoru karşılaştırılmaktadır.

Tablo 6. Görüntüleme özellikleri

Özellik	Arabul	Arama	Netbul	Superonline
Varsayılan (default) erişilen belge sayısı	10	15	Netbul Kategorileri için - 10 Internet Index için - 40	10
Sonuçların Sayısını Artır	✓	--	--	--
20 Sonucu Gör	✓	--	--	--
Sırala	--	✓	--	✓
Tarihe Göre Sırala	--	--	--	--
URL Adresi	✓	✓	✓	✓
Sayfanın Başlığı	✓	✓	✓	✓
Sadece başlıkları göster	--	--	--	--
Sayfadan Alıntı	✓	✓	✓	✓
Belge Büyüklüğü	--	✓	--	✓
Sayfaya Bağlantı	✓	✓	✓	✓
Toplam Kayıt Sayısı	✓	✓	--	✓
Kaydın Son Güncelleme Tarihi	--	--	--	✓

✓ : Özelliğin bulunduğunu belirtir. -- : Özelliğin bulunmadığını belirtir.

4.2.5 Boole Komutları

Bu kısımda ise arama motorlarında kullanılan Boole mantıksal işleçleri (‘OR'/VEYA, ‘AND'/VE, ‘NOT/AND NOT’/DEĞİL/VE DEĞİL, ‘( )’, ve ‘NEAR’/YAKIN) açıklanmaktadır. Tablo 7'de dört arama motoru, ilgili Boole işleçleri açısından karşılaştırılmaktadır.

> *OR/VEYA*: Bu işleç kullanılarak birden fazla kelimeden oluşan sorgu cümlelerinde kelimelerden en az bir tanesinin geçtiği belgelere erişilir. Örneğin, *papatya VEYA nilüfer* sorgusu için içinde "papatya" ya da "nilüfer" veya her ikisi de geçen tüm belgeler bulunur.

AND/VE: Bu işleç kullanılarak girilen kelimeler "ve" mantık kuralına uygun olarak işlem görür ve sorgu cümlesinde yer alan bütün kelimelerin geçtiği belgeler bulunur. Örneğin, *nilüfer VE çiçek* sorgusunda içinde hem "nilüfer" hem de "çiçek" geçen belgelere erişilir.

NOT/AND NOT/DEĞİL/VE DEĞİL: Bu işleç içinde istemediğimiz kelimeler geçen belgeleri dışlamak için kullanılır. Örneğin, *nilüfer DEĞİL çiçek* sorgusunda "nilüfer" ile ilgili olsa bile içinde "çiçek" kelimesi geçen belgelere erişilmez.

( ): Bu işleç kullanılarak içinde birden fazla kelime geçen sorgu cümlelerinde hangi kelimelerin öncelikli olarak aranacağı belirlenir. Örneğin, *çiçek VE (papatya VEYA nilüfer)* sorgusunda önce içinde "papatya" ya da "nilüfer" kelimelerinden en az biri geçen belgeler belirlenir, daha sonra bu belgelerde aynı zamanda "çiçek" kelimesinin geçip geçmediğine bakılır ve geçenlere erişilir.

NEAR/YAKIN: Bu işleç kullanıldığında sorgu cümlesinde yer alan kelimelerin ilgili belgelerde en az kaç kelime arayla geçmesi gerektiği tanımlanır. Örneğin, *shakespeare YAKIN komedi* sorgusu ile içinde "Shakespeare muhteşem komedi yazılarında..." ibaresi geçen bir belgeye erişilir.

Tablo 7. Boole komutları

Komut	Kullanım şekli	Arabul	Arama	Netbul	Superonline
OR	OR	✓ OR/VEYA/^	✓ |/BOŞLUK	✓ BOŞLUK	✓ |/VEYA
AND	AND	✓ AND/VE/ BOSLUK	✓ &	--	✓ VE/&
NOT	NOT	✓ DEĞIL	--	--	✓ DEĞİL
	AND NOT	--	--	--	✓ VE DEĞİL/!
Nesting	( )	✓	--	--	✓
NEAR	NEAR	--	--	--	✓ YAKIN/~

✓ : Özelliğin bulunduğunu belirtir. -- : Özelliğin bulunmadığını belirtir.

4.3 Sorular

Arabul, Arama, Netbul ve Superonline arama motorlarının bilgi erişim performanslarını çeşitli açılardan değerlendirmek için toplam 17 soru seçilmiştir. Bu soruların bir kısmı tarafımızdan oluşturulmuş, bir kısmı da daha önce yabancı arama motorlarının performanslarını değerlendirmek üzere kullanılan sorular arasından seçilmiştir. Yabancı arama motorlarında denenen bazı soruların Türkçe arama motorlarında da denenerek ilgili konularda Türkçe belgelere erişilip erişilemediği test edilmiştir.

Soruların listesi, hangi bilgi ihtiyaçlarını karşılamak üzere oluşturuldukları ve erişilen ilgili belgelerde hangi özelliklerin arandığı Şekil 4'te verilmektedir. Şekilde her bir bilgi ihtiyacı için belirlenen başlık (koyu renkte), daha ayrıntılı tanım ve erişilen listede ilgili belgelere karar verilirken kullanılacak olan ölçütler yer almaktadır. Soruları oluşturma amaçları aşağıda açıklanmaktadır.

Birinci ("Internet ve etik") ve 2. soruda ("Barok müzik") belli bir konuya odaklanan bilgi ihtiyaçları göz önünde bulundurulmuştur. Üçüncü soruda "Prozac" adlı ilaçla ilgili bilgi edinmek amaçlanmıştır. Aynı adı taşıyan rock grubu ile ilgili bilgiler ilgisiz sayılmış, böylece arama motorlarının belli sözcükleri içeren belgeleri ayıklama kapasitesinin olup olmadığı sınanmıştır. Bu çalışmanın ana temasını oluşturan Türkçe arama motorlarının özellikleri, kullanım ve etkinlinlik değerlendirilmeleri ile ilgili belgeler 4. soruyu oluşturmuştur. Beşinci ve altıncı sorularda ("Baris Manco şarkılarına ait mp3'ler" ve "Barış Manço şarkılarına ait mp3'ler") arama motorlarının "ş", "ı" ve "ç" gibi Türkçe karakterleri nasıl algıladığı; bunların en yakın İngilizce karakterlerle aranmasının sonuçlarda ne gibi değişiklikler yarattığı gözlenmek istenmiştir. Yedinci soruda ("DPT" nedir?) Devlet Planlama Teşkilatının ev sayfasına yönlendirilmesi beklenmektedir. Sekizinci soruda "uzaylı" hakkında genel bir bilgi edinilmek istenmiş ve kullanıcının konuyu özellikle genel tutup, cevaplardan yola çıkarak bilgi ihtiyacını daraltmak (refine) isteme olasılığı göz önünde bulundurulmuştur. Benzer amaçlı bir diğer soru da dokuzuncu sorudur ("uzaylılar"). Amaç "uzay" (13. soru), "uzaylı" ve "uzaylılar" sorgularının sonuçlarından elde edilen bilgilere dayanarak arama motorlarının gövdeleme (stemming) yapıp yapmadığını belirlemektir. Onuncu ("Demirel ve Sezer"), 11. ("Demirel veya Sezer") ve 12. ("Demirel veya Sezer ve TEMA") sorular Boole işleçleri kullanılarak yapılan ve kişilerle ilgili bilgi edinmeyi amaçlayan aramalardır. Ayrıca, 10. ve 12. sorulara karşılık erişilen belgelerin 11. soru için erişilenlerin bir alt kümesi olması gerektiği düşünülmüştür. Onüçüncü, 14. ve 15. sorular belirli bir konuda yapılan oldukça genel konu aramalarıdır. Sırasıyla, "Uzay", "Evren" ve "Uzay veya Evren" hakkında bilgi bulmak amaçlanmıştır. Bu sorularda arama motorlarının kapsamlı konu aramalarında ilgili belgeleri ilgisiz belgelerden ayırt etmek, yanlış düşmeleri (false drops) azaltmak için çaba sarfedip sarfetmediklerini görmek amaçlanmıştır. Onbeşinci soruda iki geniş kapsamlı konu aramasının "VEYA" Boole işleciyle birleştirilmesi sonucu erişilen belgelerin nasıl sıralandıklarını görmek amaçlanmıştır. Onaltıncı soruda ("Atatürk ve Fikriye Hanım") arama motorlarının tarihsel araştırmalar için yararlı olup olamayacaklarını test etmek amacıyla kullanılmıştır. Onyedinci soruda ("TBMM Başkanı Ömer İzgi hakkında bilgi") ise arama

motorlarının güncel konularda bilgi edinmek amacıyla kullanılıp kullanılamayacağı test edilmiştir.

1. **Internet ve etik**. Internet ile ilgili etik değerler. Internet kullanımı ve yayıncılıkla ilişkili etik veya ahlaki değerler. Etik değerlerle ilişkili üst veri belirteçleri (metatags) kullanımı üzerine tartışma ilgili bilgi ihtiyacını tatmin edici olarak kabul edilmiştir. Üst veri belirteçlerinin işlenmesi, spam e-posta ve/veya Web sayfalarını süzgeçleme (filtering) yazılımları veya araçları hakkında bilgiler. Internet ve etik değerler hakkında, ikisi arasında ilişki kurmaksızın, bilgi veren kaynaklar ilgisiz sayılmıştır.
2. **Barok müzik ve özellikleri**. Genel olarak Barok müzik üzerine bir tartışma veya özellikleri ile ilgili ayrıntılı bilgi, Barok çağı sanatçıları hakkındaki bilgiler ilgili bilgi ihtiyacını karşılamada yeterli bulunmuştur.
3. **"Prozac" hakkında bilgi (bir rock grubu olan "Prozac" hakkındaki bilgiler hariç)**. Prozac adlı ilacın özellikleri, kullanım yerleri, yan etkileri ve/veya elektronik satın alma kaynakları hakkındaki belgeler konu ile ilgili sayılmıştır. Aynı adı taşıyan rock grubu hakkındaki bilgiler ilgisiz sayılmıştır.
4. **Internet'te Türkçe arama motorlarının değerlendirmesiyle ilgili çalışmalar**. Türkçe arama motorları ile ilgili başarım (performans) ölçümleri, karşılaştırmalar, ve/veya istatistiksel çalışmalar ilgili sayılmıştır.
5. **Baris Manco'nun şarkılarına ait mp3'ler**. Belgenin ilgili olabilmesi için Barış Manço'nun şarkılarına ait mp3'lerin Web sayfası içerisinde indirilebilir olması gerekmektedir.
6. **Barış Manço'nun şarkılarına ait mp3'ler**. 5. soruyla aynı (yazım farkı)
7. **"DPT" nedir?** "DPT" kısaltmasının açılımını veren, genelde "Devlet Planlama Teşkilatı" hakkındaki belgeler ilgili sayılmıştır.
8. **"uzaylı" hakkında genel bilgi**. Uzaylılar hakkında genel bir bilgi edinilmek istenmektedir. Sadece uzay konusu hakkında bilgi içeren kayıtlar ilgisiz olarak kabul edilmiştir. Uzaylılar ile ilgili (belgenin görselliğini zenginleştirmek amacı ile Web tasarımcısı tarafından çizilmiş olmayan) resimler içeren belgeler uzaylıların varlığı hakkında görsel bir bilgi verdiği için ilgili sayılmıştır. Ayrıca uzaylılar ile ilgili yazılar, haberler ve/veya kişisel yorumlar içeren belgeler de ilgili belge olarak kabul edilmiştir.
9. **"uzaylılar" hakkında genel bilgi**. 8. soruyla aynı (tekil-çoğul özelliği).
10. **Türkiye Cumhuriyeti Cumhurbaşkanlarından Süleyman Demirel ve Ahmet Necdet Sezer'i konu alan belgeler**. Süleyman Demirel ve Ahmet Necdet Sezer hakkında yayınlanan haberleri; özgeçmişlerini, konuşmalarını, basın bildirilerini ve/veya Cumhurbaşkanlığı görevinde bulunma tarihlerini içeren belgeler ilgili sayılmıştır.
11. **Türkiye Cumhuriyeti Cumhurbaşkanlarından Süleyman Demirel veya Ahmet Necdet Sezer'i konu alan belgeler**. Süleyman Demirel **veya** Ahmet Necdet Sezer hakkında yayınlanan haberleri; özgeçmişlerini, konuşmalarını, basın bildirilerini ve/veya Cumhurbaşkanlığı görevinde bulunma tarihlerini içeren belgeler ilgili sayılmıştır.
12. **Türkiye Cumhuriyeti Cumhurbaşkanlarından Süleyman Demirel veya Ahmet Necdet Sezer'in TEMA konusundaki yaklaşımları**. Süleyman Demirel veya Ahmet Necdet Sezer'in TEMA konusundaki düşüncelerini, yorumlarını ve/veya açıklamalarını içeren belgeler ilgili sayılmıştır.
13. **Uzay hakkında bilgi**. Uzaydaki gezegenlerden bahseden, uzayda canlı olup olmadığını tartışan ve/veya uzay ile ilişkili çalışmalar hakkında bilgi içeren bilimsel veya güncel yazı, haber, veya kişisel yorum içeren belgeler ilgili sayılmıştır.
14. **Evren hakkında bilgi**. Evrendeki gezegenlerden bahseden, evrende canlı olup olmadığını tartışan ve/veya evren ile ilişkili çalışmalar hakkında bilgi içeren bilimsel veya güncel yazı, haber, veya kişisel yorum içeren belgeler ilgili kategorisinde yer alacaktır.
15. **Uzay veya Evren hakkında bilgi**. Uzaydaki/evrendeki gezegenlerden bahseden, uzayda/evrende canlı olup olmadığını tartışan ve/veya uzay/evren ile ilişkili çalışmalar hakkında bilgi içeren bilimsel veya güncel yazı, haber, veya kişisel yorum içeren belgeler ilgili sayılmıştır.
16. **Atatürk ve Fikriye Hanım**. Türkiye Cumhuriyeti'nin kurucusu ve ilk Cumhurbaşkanı olan Mustafa Kemal Atatürk ve Fikriye Hanım arasındaki ilişki hakkında bilgi edinmek amaçlanmaktadır. Fikriye Hanım'ın ve Atatürk'ün adlarının geçtiği, aralarındaki ilişkiden söz eden belgeler ilgili sayılmıştır.
17. **TBMM Başkanı Ömer İzgi hakkında bilgi**. Bu soruda, arama motorlarının güncel konularla ilgili bilgi bulma özellikleri test edilmek istenmektedir. Şimdiki TBMM Başkanı Ömer İzgi hakkında bilgi veren

Şekil 4. Soru listesi

Kısaca özetlenecek olursa; yukarıda verilen 17 soruyla arama motorlarının etkinlikleri aşağıda verilen beklentiler açısından test edilmiştir:

- Farklı türdeki sorular için arama motorlarının erişim etkinliği;
- Boole işleçleri kullanılarak ifade edilen sorgularda erişim etkinliği;
- Dar kapsamlı sorular için ilgili belgelere erişilebilmesi;
- Geniş kapsamlı sorular için ilgili belgelere erişilebilmesi ve ilgisizlerin ayırt edilebilmesi;
- Bir veya iki sözcükle ifade edilen bilgi ihtiyaçlarının karşılanabilmesi;
- Gövdeleme algoritması kullanılması;
- Türkçe karakter kodlamadan kaynaklı sorunların en aza indirilmesi; ve
- Dizinlenen Internet kaynakları daha sonra belirli aralıklarla güncellenmesi ve ölü bağlantıların ayıklanması.

4.4 Soruların Formülasyonu

Yukarıda verilen sorular için, seçilen dört arama motoru üzerinde aramalar gerçekleştirilmiştir (14-28 Kasım 2001). Doğal dil ile belirtilen (bkz. Şekil 4) bilgi gereksinimlerinin arama motorlarında aranabilmesi için, söz konusu bilgi gereksinimlerinin arama motorlarının "anlayabileceği" şekle sokulması gerekmektedir. Bir başka deyişle, bu sorular her arama motorunda kullanılan kural ya da ifadelerle gerçekleştirilmelidir. Bu bağlamda, aynı bilgi gereksinimini karşılamak üzere farklı arama motorlarına yönlendirilen sorular sözdizimi (sentaks) ve kullanılan işleç ya da işaretler açısından farklı olabilmektedir. Şekil 5'te erişim etkinliğini ölçmek üzere kullanılan 17 sorunun seçilen dört arama motoru üzerinde nasıl çalıştırıldığı gösterilmektedir. Soruların formüle edilmesinde daha önce verilen (bkz. 4.2) her arama motorunun özellikleriyle ilgili bilgilerden yararlanılmıştır.

Sorguların çalıştırılmasında aşağıdaki noktalara dikkat edilmiştir:

- Tutarlılığı sağlamak için tüm sorular bütün arama motorlarında aynı araştırmacı (YB) tarafından gerçekleştirilmiş ve sonuçlar yine aynı kişi tarafından değerlendirilmiştir.
- Bir arama motoruna ait farklı sorgulama biçimleri varsa, en basit olanı kullanılmıştır.
- Bazı arama motorları (Arabul, Netbul, Superonline) kategoriler üzerinde de arama yapmakta ve erişim çıktısında kategoriler ve belgeler ayrı ayrı gösterilmektedir. Kuşkusuz herhangi bir soru karşılığı entellektüel dizinleme işlemi sonucu oluşturulan kategorilere isabet eden aramalarda duyarlık daha

yüksek olabilmektedir. Netbul ve Superonline'da kategoriler üzerinde arama seçeneği dikkate alınmamış, sırasıyla "Internet'te" ve "Web" seçenekleri ile aramalar gerçekleştirilmiştir. Arabul'da ise böyle bir ayrım yapmak mümkün değildir. Erişim çıktısında önce erişilen kategoriler daha sonra ise bireysel belgeler gösterilmektedir. Farklı arama motorlarından elde edilen sonuçları birebir karşılaştırabilmek amacıyla Arabul'da erişilen kategoriler dikkate alınmamış, daha sonra erişilen bireysel belgeler değerlendirilmiştir.

- İdeal olarak bir sorgunun bütün arama motorlarında aynı anda çalıştırılması gerekmektedir. Böylece sürekli çalışan dizinleme yazılımlarının iki arama motorunun denenmesi sırasında geçen zaman zarfında yeni adresleri dizinlemesi ve daha sonra denenen motorun bu nedenle daha başarılı bulunması olasılığı ortadan kalkmaktadır. Araştırmamızda aynı sorgular farklı arama motorlarında mümkün olduğu kadar kısa aralıklarla çalıştırılmış ve bütün sorguların araştırılması yaklaşık iki haftada bitirilmiştir.
- Sorgular çalıştırılıp erişim çıktıları alındıktan sonra, erişilen belgelerin en kısa zamanda incelenmesi gerekmektedir. Böylece arama motorunun dizinlediği ve belirli bir sorguya karşılık eriştiği belgelerin değerlendirme yapılana dek geçen sürede silinme ya da değiştirilme olasılığı ortadan kalkmaktadır. Araştırmamızda bu durum dikkate alınmış ve erişim çıktıları üzerinde değerlendirmeler erişimden hemen sonra bir hafta içinde (Aralık 2001) gerçekleştirilmiştir.

Soru	Arabul	Arama	Netbul	Superonline
1	internet ve etik	internet & etik	+internet +etik	internet ve etik
2	”barok müzik”	”barok müzik”	”barok müzik”	”barok müzik”
3	+prozac –rock	+prozac -rock	+prozac -rock	+prozac –rock
4	“arama motoru” ve değerlendirme	“arama motoru” & değerlendirme	+“arama motoru” +değerlendirme	“arama motoru” ve değerlendirme
5	”baris manco” ve mp3	”baris manco” & mp3	+”baris manco” +mp3	”baris manco” ve mp3
6	”barış manço” ve mp3	”barış manço” & mp3	+”barış manço” +mp3	”barış manço” ve mp3
7	DPT	DPT	DPT	DPT
8	uzaylı	Uzaylı	uzaylı	Uzaylı
9	uzaylılar	uzaylılar	uzaylılar	uzaylılar
10	“Süleyman Demirel” ve “Ahmet Necdet Sezer”	“Süleyman Demirel” & “Ahmet Necdet Sezer”	+“Süleyman Demirel” +“Ahmet Necdet Sezer”	“Süleyman Demirel” ve “Ahmet Necdet Sezer”
11	“Süleyman Demirel” veya “Ahmet Necdet Sezer”	“Süleyman Demirel” | “Ahmet Necdet Sezer”	“Süleyman Demirel” “Ahmet Necdet Sezer”	“Süleyman Demirel” veya “Ahmet Necdet Sezer”
12	“Süleyman Demirel” veya “Ahmet Necdet Sezer” +tema	“Süleyman Demirel” | “Ahmet Necdet Sezer” +tema	“Süleyman Demirel” “Ahmet Necdet Sezer” +tema	“Süleyman Demirel” veya “Ahmet Necdet Sezer” +tema
13	uzay	Uzay	uzay	Uzay
14	evren	Evren	evren	Evren
15	Uzay veya evren	uzay | evren	uzay evren	uzay veya evren
16	atatürk ve fikriye	atatürk & fikriye	+atatürk +fikriye	atatürk ve fikriye
17	“Ömer İzgi”	“Ömer İzgi”	“Ömer İzgi”	“Ömer İzgi”

Şekil 5. Arama sorularının formülasyonu

Tüm sorular için dört arama motoru tarafından erişilen belgelerle ilgili ham veriler işlem kütüklerine kaydedilmiştir. Bu verilere Web aracılığıyla erişilebilmektedir.[1]

4.5 İlgililik Değerlendirmeleri

Arama motorlarında her soru ayrı ayrı çalıştırılmış ve erişim çıktıları üzerinde ilgililik (relevance) değerlendirmeleri gerçekleştirilmiştir. Belirli bir soruya karşılık erişilen belgelerden hangilerinin ilgili, hangilerinin ilgisiz olduğuna aramayı gerçekleştiren kişi tarafından karar verilmiştir. Söz konusu karar verilirken erişilen her belgenin aranan sorunun konusu hakkında (aboutness) olup olmadığına bakılmıştır. Araştırmacının erişilen belgeleri daha önce görmediği varsayılmıştır. Her ilgili belge erişilen diğer ilgili belgelerden bağımsız olarak değerlendirilmiştir.

İlgililik değerlendirmesi yapılırken aşağıdaki noktalara dikkat edilmiştir (Soydal, 2000, s. 46):

- Erişilen belgeler teker teker incelenip "ilgili" ya da "ilgisiz" olarak sınıflandırılmıştır.
- Web üzerinde bulunan ve adresleri farklı olan her belge farklı bir bilgi kaynağı olarak değerlendirilmiştir.
- Aynı bilgiyi içeren ve fakat farklı adresi olan belgeler (mirror pages), farklı belgeler olarak değerlendirilmiştir.
- Aynı bilgiyi içeren ve adresleri de aynı olan belgelerden ilki değerlendirilmiş, diğer(ler)i kullanıcının bu adreslere bakmayacağı düşünülerek "ilgisiz" kabul edilmiştir. Benzer bir biçimde, büyük ve küçük harf kullanımı nedeniyle URL adresleri farklı gözüken ama aynı bilgileri içeren belgelerden ilki değerlendirilmiş, sonraki(ler) "ilgisiz" kabul edilmiştir.
- Erişilen sayfalarda bilginin kendisi değil, fakat bu bilginin yer aldığı başka bir sayfaya bağlantı (link) bulunuyorsa, bu sayfalar da "ilgili" olarak değerlendirilmiştir.
- Hata veren, taramanın yapıldığı tarihte çalışmayan, ilgili görünen sayfalara bağlantıların bulunduğu ve fakat bu bağlantıların çalışmadığı sayfalar ile taşınmış sayfalar (bağlantı olmasına rağmen bunların çalışmadığı durumlarda) "ilgisiz" olarak kabul edilmiştir.
- İngilizce ya da Türkçe dışında bir dilde hazırlanmış sayfalar "ilgisiz" olarak değerlendirilmiştir.

[1] Bkz. http://cmpe.emu.edu.tr/bitirim/home/.

Her soru için erişilen belgeler üzerinde yapılan ilgililik değerlendirmeleri ikinci bir araştırmacı tarafından da gözden geçirilmiştir. İşlem kütükleri üzerinden yapılan söz konusu gözden geçirmede iki araştırmacının ilgililik değerlendirmelerinde büyük ölçüde aynı görüşte oldukları ortaya çıkmıştır.

4.6 Performans Ölçümleri

Arama motorlarının erişim etkinliğini belirleyen en önemli etmenlerden birisi kullanılan erişim yöntemleridir. Fakat kullanıcı açısından bakıldığında, bir arama aracının kullandığı yöntem önem taşımamaktadır; kullanıcıyı sadece arama motorunun belirli bir soru için eriştiği belgeler ilgilendirmektedir. Kısacası, kullanıcı için önemli olan arama motorunun performansıdır. Bir önceki kesimde sözü edilen ilgililik değerlendirmeleri bu anlayışla gerçekleştirilmiştir.

Daha önceki kesimlerde (2.5 ve 3.5) bilgi erişim sistemlerinin etkinliğini ölçmek için kullanılan anma, duyarlık ve posa gibi belli başlı değerlere ve ilgili değerleri kullanılarak yapılan araştırmalara değinilmişti.

Bu araştırmada önce Clarke ve Willet (1997) tarafından önerilen ortalama anma değerlerinin hesaplanması kararlaştırılmış ve ön değerler elde edilmiştir. Ancak her soru için erişilen ilgili belge sayısının çok düşük olmasından dolayı anma değerlerinin hesaplanmasından vazgeçilmiştir. Çünkü ilgili belge sayısının sığ olduğu bir ortamda anma değerleri rahatlıkla sorgulanabilir.[2]

Duyarlık değerleri ise her soru için çeşitli kesme noktaları (5, 10, 15 ve 20) kullanılarak hesaplanmıştır. Duyarlık değerlerinin hesaplanması için kullanılan formül aşağıdadır:

$$D_k = \text{Erişilen ilk (k) belge arasındaki ilgili belgelerin sayısı / Erişilen ya da gösterilen toplam belge sayısı (k)} \qquad (11)$$

Bu formülde *k* katsayısı çeşitli kesme noktalarını ifade etmektedir. Örneğin, kesme noktasının 10 olarak alındığı bir erişimde 10 belge içindeki toplam ilgili belge sayısı 5 ise duyarlık değeri 0,5 olarak gerçekleşir ($D_{10} = 5 / 10 = 0{,}5$). Erişilen toplam belge sayısı kesme noktası olarak belirlenen sayıdan daha az ise erişilen toplam belge sayısı üzerinden duyarlık

[2] Bunun tersi de mümkündür; yani milyarlarca belgenin söz konusu olduğu ve bunlardan çok azının dizinlenebildiği bir ortamda anmanın değeri kendiliğinden azalmaktadır (Hawking et al., 1999).

değeri hesaplanır. Erişilen toplam belge sayısının 20'den fazla olduğu durumlarda ise ilk 20 belgeden sonra gelen belgeler duyarlık hesaplamalarında dikkate alınmamıştır.

Tüm sorular için bu hesaplamalar farklı arama motorlarında ayrı ayrı yapıldıktan sonra, her soru ve her arama motoru için makro ortalama yöntemi kullanılarak ortalama duyarlık değerleri bulunmuştur. Bilindiği gibi, makro ortalamada her soru için erişilen ilgili belge sayısı ve erişilen toplam (ilgili ve ilgisiz) belge sayısı bulunur, ilgili belge sayısı erişilen toplam belge sayısına bölünerek duyarlık değeri bulunur. Ortalama duyarlık değerini bulmak için ise her soru için hesaplanan duyarlık değerleri toplanarak toplam soru sayısına bölünür. Benzeri bir biçimde, belirli bir soru için dört arama motorundan elde edilen ortalama duyarlık değerini bulmak için arama motorlarının o soru için hesaplanan duyarlık değerleri toplanır ve dörde bölünür.

Bu çalışmada, her arama sorusu için elde edilen ilgili belge sayısının genelde düşük olması nedeniyle, arama motorlarının performans değerlendirmesi için "normalize sıralama" değerleri de bulunmuştur. Daha önce de değinildiği gibi, kullanıcılar arama motorlarının belirli bir soruya karşılık olarak eriştikleri belgelerin çok azını görme eğilimindedirler. Bu bakımdan, erişilen ilgili belgeleri tutarlı bir biçimde erişim çıktısının başlarında listeleyen arama motorlarının kullanıcılar tarafından daha çok tercih edileceği kolayca öne sürülebilir.

Bu çalışmada her soru için dört arama motorunda gerçekleştirilen arama sonuçları üzerinde toplam belge sayısının 5, 10, 15 ve 20 olduğu durumlarda (kesme noktalarında) normalize sıralama değerleri hesaplanmıştır. Normalize sıralama değerlerinin hesaplanmasında daha önce kesim 2.5'te verilen S_{norm} formülü kullanılmıştır.

Erişilen toplam belge sayısının kesme noktası olarak belirlenen sayıdan daha az olduğu durumlarda kullanıcının erişilemeyen belgeler karşısında nötr olduğu varsayılmıştır. Çünkü kullanıcı, erişilemeyen belgelerin ilgili olduğu halde sistem tarafından kaçırıldığını (miss) bilmediği gibi, ilgisiz olduğu halde erişilen (false drop) belgeleri de bilmeyebilir. Bu gibi durumlarda, kullanıcının nötr düzeyi göz önünde bulundurularak erişim çıktısının boyu kesme noktası olarak alınan sayıya genişletilmiştir. Erişilen toplam belge sayısının 20'den fazla olduğu durumlarda ise ilk 20 belgeden sonra gelen belgeler normalize sıralama değerlerinin hesaplanmasında dikkate alınmamıştır.

Normalize sıralama değerlerinin erişilen toplam belge sayısının kesme noktası olarak kullanılan sayıdan (örneğin, 5) az olduğu durumlarda nasıl hesaplandığı aşağıda çeşitli örneklerle gösterilmektedir. Örneklerde "+" ilgili, "-" ilgisiz, "*n*" ise nötr belgeyi temsil etmektedir.

1. + - - + *n* : S_{norm} = ½ (1+(4-4)/8)=0,5 (S_{max}: + + *n* - -);
2. - + + + - : S_{norm} = ½ (1+(3-3)/6)=0,5 (S_{max}: + + + - -);
3. - - - - - : S_{norm} = 0 (S_{max} tanımlı değil, fakat erişim çıktısı - - - - - + biçiminde düşünülerek S_{norm}= ½ (1+(5-5)/5)=0 olarak verilebilir. Bu durumda hedeflenen erişim çıktısı kuşkusuz S_{max}: + - - - - - olacaktır.);
4. - - *n n n* : S_{norm} = ½ (1+(0-6)/6)=0 (S_{max}: *n n n* - -);
5. + *n n n n* : S_{norm} = 1 (S_{max}: + *n n n n*).

Tüm sorular için normalize sıralama değerleri farklı arama motorlarında ayrı ayrı hesaplandıktan sonra, her soru ve her arama motoru için makro ortalama yöntemi kullanılarak ortalama normalize sıralama değerleri bulunmuştur.

Türkçe arama motorlarının kapsama ve yenilik oranları, 1-14 Haziran 2001 tarihleri arasında Arabul'da gerçekleştirilen aramalarda en sık aranan ve tek sözcükten oluşan beş soru ("mp3", "oyun", "sex", "erotik" ve "porno") ile ölçülmüştür. Seçilen sorular yabancı arama motorlarında en sık aranan sözcüklerle de uyumluluk göstermektedir.

Kapsama ve yenilik oranlarını sağlıklı bir biçimde ölçmek için her soruya karşılık erişilen toplam belge sayısının en az 50 olması gerekmektedir. Kapsama ve yenilik oranlarının hesaplanmasında izlenen yol şöyle özetlenebilir:

Belirlenen anahtar sözcükler ("mp3", "oyun", "sex", "erotik" ve "porno") dört arama motorunda, Internet'te arama yapılacak şekilde, çalıştırılarak erişilen ilk 1000'er belge havuzda toplanmıştır. Arama motoru 1000'den az belgeye eriştiyse erişilen tüm belgeler havuza atılmıştır. Daha sonra her anahtar sözcük için erişilen belgeler 50'lik öbekler (ilk 50, ilk 100, ilk 150, ..., ilk 1000) halinde listelenme sıralarına göre alınarak her arama motoru için kapsama ve yenilik oranları aşağıdaki formüllere göre hesaplanmıştır.

$$\textit{Kapsama oranı} = TBS({}_{a}S^{b}_{c1}) / TBS({}_{a}S^{b}_{c1} \cup {}_{a}S^{b}_{c2} \cup {}_{a}S^{b}_{c3} \cup {}_{a}S^{b}_{c4}) * 100 \qquad (12)$$

Formülde:

a	: öbek sayısını (ilk 50, ilk 100, ilk 150,…, ilk 1000);
b	: sorguyu ("mp3", "oyun", "sex", "erotik", ve "porno");
c	: arama motorunu (Arabul, Arama, Netbul, Superonline);
${}_{a}S^{b}_{c}$	: *b* sorgusu *c* arama motorunda çalıştırıldığında erişilen ilk *a* kadar belge kümesini;
TBS(*S*)	: *S* kümesindeki tekil belgelerin sayısını (birden fazla ve aynı olan belgeler bir belge kabul edilir)

temsil etmektedir. Formül, sorgu *b* çalıştırıldığında *c* arama motorunun ilk *a* belge öbeği için kapsama oranını vermektedir.

Aynı notasyon kullanılarak yenilik oranının formülü de verilebilir:

$$\textit{Yenilik oranı} = TBS\left(\left({}_{a}S^{b}_{c1} - \left({}_{a}S^{b}_{c2} \cup {}_{a}S^{b}_{c3} \cup {}_{a}S^{b}_{c4}\right)\right) / a\right) * 100 \qquad (13)$$

Formül, sorgu *b* çalıştırıldığında *c* arama motorunun ilk *a* belge öbeği için yenilik oranını vermektedir.

Kapsama ve yenilik oranlarının hesaplanması aşağıda bir örnekle açıklanmaktadır: "mp3" anahtar sözcüğü için Arabul arama motorunda erişilen ilk 50 belge alınmış, birden fazla ve aynı olan belgeler bir belge sayılarak erişilen toplam belge sayısı bulunmuştur. Bu sayı, dört arama motoru tarafından erişilen ilk 50'şer belge kümesinin birleşiminin toplam sayısına bölünmüş (toplamda da birden fazla ve aynı olan belgeler bir belge sayılmıştır) ve çıkan sonuç 100 ile çarpılarak kapsama oranı bulunmuştur. Bu işlem ilk 50, ilk 100, ilk 150, ilk 200 ,…, ve ilk 1000 belge için yinelenmiş ve Arabul arama motoru için her 50'lik öbekteki kapsama oranları hesaplanmıştır. Her bir üst öbek bir alt öbeği de içermektedir. Örneğin, "ilk 100" öbek kullanılırken "ilk 50" öbek için alınan belgeler, "ilk 150" öbek kullanılırken ise "ilk 100" öbek için alınan belgeler de kullanılmaktadır. Özetlemek gerekirse, kapsama oranı bir arama motoru tarafından erişilen tekil ilgili belgelerin dört arama motoru tarafından erişilen tekil ilgili belgelere oranıdır.

Yenilik oranı ise şöyle hesaplanmıştır: Arabul'da "mp3" anahtar sözcüğü arandığında erişilen ilk 50 belgelik kümeden, aynı soru için diğer arama motorları (Arama, Netbul, Superonline) tarafından erişilen ilk 50'şer belge kümelerinin birleşimi çıkarılıp, geriye kalan tekil belge sayısının toplamı 50'ye bölünmüş (toplamda, birden fazla ve aynı olan belgeler tek belge sayılmıştır), çıkan sonuç 100 ile çarpılarak yenilik katsayısı hesaplanmıştır. Özetlemek gerekirse, yenilik oranı bir arama motoru tarafından erişilen ve fakat diğer üç arama motoru tarafından erişilemeyen tekil ilgili belgelerin o arama motoru tarafından erişilen belgelere oranıdır.

Bu yöntem sorgular ve arama motorları bazında her 50'lik öbek için yinelenmiş ve kapsama ve yenilik oranları hesaplanmıştır. Daha sonra, arama motorlarının Türkiye adresli (sonu ".tr" ile biten) belgeler açısından kapsama ve yenilik oranları hesaplanmıştır. Bu amaçla, bir önceki adımda her sorgu için havuzda toplanan belgeler, belge sıralarına göre kontrol edilip site adresleri ".tr" ile bitmeyenler havuzdan çıkarılmıştır. Havuzda kalan ve

sonu ".tr" ile biten adresler yeniden sıralanmıştır. Örneğin; "mp3" anahtar sözcüğü Arabul arama motorunda arandığında erişilen ilk 50 belge sırasıyla (1) www.aaa.com, (2) www.bbb.com.tr, (3) www.ccc.com, (4) www.ddd.com.tr, (5) www.eee.edu.tr,… olsun. Sonu ".tr" ile bitmeyen site adresler havuzdan çıkarıldığında sıra şöyle değişecektir: (1) www.bbb.com.tr, (2) www.ddd.com.tr, (3) www.eee.edu.tr,… Havuzdaki ayıklama işlemleri bittikten sonra, ".tr" adresli belgeler için kapsama ve yenilik oranları daha önceki formüller kullanılarak hesaplanmıştır.

Türkçe arama motorlarında günleme sıklığını belirlemek için, bilgi erişim sisteminin tanımını içeren bir paragraflık bir belge hazırlanmış ve aynı belge farklı Internet adreslerine sahip iki Web sunucusuna (http://cmpe.emu.edu.tr/bitirim/vartanbitirim, http://www.geocities.com/vartanbitirim) yerleştirilmiştir. Daha sonra bu adresler Arabul, Arama, Netbul ve Superonline arama motorlarının site kayıt formları doldurularak 18 Ekim 2001 tarihinde kaydedilmiştir. Arabul, eklenen adresin 1 ile 4 hafta arasında dizinleneceğini belirtmesine rağmen, 5 Şubat 2002 tarihinde yapılan kontrolde ilgili belgenin henüz dizinlenmediği gözlenmiştir. Arama, eklenen adreslerin 6 saat içerisinde dizinleneceğini belirtmiştir. Ancak adresler eklendikten yaklaşık bir dakika sonra yapılan aramada her iki adresin de dizinlendiği görülmüştür. Netbul ve Superonline arama motorlarında eklenen adreslerin ne kadar sürede dizinleneceği belirtilmemiştir. 5 Şubat 2002'de yapılan kontrolde her iki adresin de Netbul ve Superonline'da dizinlenmediği görülmüştür.

Günleme sıklığıyla ilgili testin diğer bir aşaması ise arama motorlarının dizinlerinde yer alan bir Web sayfası üzerinde yapılan günlemenin ne kadar sürede dizinlere yansıdığını ölçmektir. Bir başka deyişle, aradan geçen süre arama motorlarının dizinleme yazılımlarının (örümceklerinin) aynı adresi hangi sıklıkta ziyaret ettiğini göstermektedir. Ancak kaydedilen adresler bir arama motoru (Arama) dışında arama motorları tarafından henüz dizinlenmediklerinden bu aşamaya geçilememiştir. Yine de, 17 soru için erişilen belgelerdeki çalışmayan ya da ölü bağlantıların sayısı örümceklerin aynı adresleri hangi sıklıkta ziyaret ettikleri konusunda kabaca da olsa bir fikir vermektedir kanısındayız (bkz. Şekil 6).

Arama motorlarının Web belgelerinde yer alan üst veri (metadata) öğelerinden erişim amacıyla yararlanıp yararlanmadıkları iki küçük uygulamayla test edilmiştir. Bu amaçla önce TKD ev sayfasında (bkz. Şekil 3) yer alan üst veriler kullanılarak dört arama motoru üzerinde arama yapılmış ve TKD'nin sayfasına üst veriler aracılığıyla erişilip erişilemediği test edilmiştir. İkinci testte her arama motoru tarafından dizinlendiği kesin olarak bilinen ve üst veri alanları dolu olan birer Web sayfası seçilmiştir. Daha sonra, her belgedeki üst veri alanlarında yer alan anahtar sözcükler kullanılarak sorgular oluşturulmuş ve bu sorgulara

karşılık arama motoru tarafından dizinlendiği kesin olarak bilinen sayfalara erişilip erişilemediği gözlenmiştir.

4.7 Verilerin Analizi

Araştırmada elde ettiğimiz bulgular çeşitli yöntemlerle analiz edilmiştir. Arama motorlarının çeşitli kesme noktalarındaki duyarlık ve normalize sıralama oranları ile güncellik, kapsama ve yenilik oranları tablo ve şekillerle verilmiş ve belli başlı bulgular özetlenmiştir. Arama motorlarının dizinlemede üst veri belirteçlerinden yararlanıp yararlanmadıklarını test etmek amacıyla yapılan denemelerin sonuçları da tablo ve şekiller halinde verilmiştir.

Arama motorlarının güncellik, duyarlık ve normalize sıralama performansları arasında fark olup olmadığı parametrik olmayan (nonparametric) Kruskal-Wallis ve Mann-Whitney U testleri uygulanarak sınanmıştır. Duyarlık ve normalize sıralama değerleri arasında ilişki (korelasyon) olup olmadığı Pearson korelasyon katsayısı (r) ile test edilmiştir. Bu testlerle ilgili kısa bilgi aşağıda verilmektedir.

Bilindiği gibi, ikiden fazla örneklemden elde edilen ortalamalar arasında fark olup olmadığını test etmek için varyans analizi (F-testi) yaygınlıkla kullanılır. Fark varsa bu farkın hangilerinden kaynaklandığını bulmak için t testi uygulanır. Ancak parametrik bir test olan F-testini (ve t- testini) uygulayabilmek için verilerin normal dağılım göstermesi ve birbirine benzemesi (homojenlik) gerekmektedir. Arama motorlarının toplam 17 soru için kaydettikleri duyarlık, normalize sıralama ve güncellik değerleri normal dağılım göstermemektedir. Yapılan testlerde dört arama motorunun duyarlık değerlerinin varyanslarının homojen olmadığı görülmüştür.

Bu nedenle, arama motorlarının duyarlık değerleri arasında istatistiksel açıdan anlamlı bir fark olup olmadığını ölçmek için parametrik olmayan Kruskal-Wallis testi uygulanmıştır. Kruskal-Wallis testi parametrik F-testi ile yapılan varyans analizine alternatif olarak bilinmektedir (Kartal, 1998, s. 211). Kruskal-Wallis testinde ikiden fazla (k) arama motoru tarafından kaydedilen tüm (N) gözlem değerlerine (örneğin, duyarlık değerleri) sıra numarası verilmekte ve bu numaralar gerçek değerlerin yerine yazılmaktadır. Daha sonra her arama motoruna ait sıra numaraları (n_j) toplanarak sütun toplamları bulunmakta (T_j) ve Kruskal-Wallis test istatistiği (H ile gösterilmektedir) hesaplanmaktadır.

$$H = \left[12(N/N+1)\sum_{j=1}^{k}(T_j^2/n_j) \right] - 3(N+1) \qquad (14)$$

Hesaplanan *H* istatistiği belirlenen güven düzeyinde tablodan bulunan kritik değerden büyükse H_0 hipotezi (örneğin, arama motorları arasında duyarlık değerleri açısından fark yoktur) reddedilmektedir.[3] Bir başka deyişle, en az iki arama motorunun duyarlık değerleri arasında istatistiksel açıdan anlamlı bir fark olduğu ortaya çıkmaktadır.

Değerler arasında fark olması durumunda bu farkın hangi arama motorundan/motorlarından kaynaklandığını bulmak için Mann-Whitney U- testi yapılmıştır. Mann-Whitney U- testi bağımsız örneklem *t*- testinin parametrik olmayan karşılığıdır. İki farklı arama motoruna ait değerlerin sıralamaları karşılaştırılarak *U*- istatistiği hesaplanır. İki farklı arama motoruna ait sıralamaların ortalamaları arasındaki farkın istatistiksel açıdan anlamlı olup olmadığı *U*- istatistiği ile test edilir.

U- istatistiğini hesaplamak için önce iki arama motorunun değerleri birleştirilerek sıra numaraları verilir. İki arama motorundan herhangi birisi için mümkün olan en büyük sıralar toplamı ile mevcut sıralar toplamı arasındaki fark *U*- değerini verir. Bu hesaplama işlemi için aşağıdaki formüller kullanılır:

n_1 büyüklüğündeki örneklem için,

$$U_1 = n_1 n_2 + ((n_1 (n_1 + 1)) / 2) - T_1 \qquad (15)$$

n_2 büyüklüğündeki örneklem için,

$$U_2 = n_1 n_2 + ((n_2 (n_2 + 1)) / 2) - T_2 \qquad (16)$$

Formülde T_1 *ve* T_2 sırasıyla birinci ve ikinci arama motorlarının sıralar toplamını vermektedir. Bu şekilde hesaplanan U_1 ve U_2 değerlerinden küçüğü *U* istatistiği olarak alınır. Hesaplanan *U* istatistiği örnek büyüklükleri (araştırmamızda 17) ve önem düzeyine göre hazırlanmış *U*- testi kritik değerler tablosundaki kritik değerle karşılaştırılır. *U* istatistiği tablo değerinden küçükse H_0 hipotezi (örneğin, Arama ile Arabul arama motorlarının duyarlık değerleri arasında fark yoktur) reddedilir. Bir başka deyişle, belirlenen güven düzeyinde iki

[3] Her arama motoruna ait gözlem sayısı 5 ya da daha az ise kritik değer Kruskal-Wallis kritik değerler tablosundan, gözlem sayısı 5'ten fazlaysa kritik değer χ^2 (ki- kare) tablosundan elde edilir. Çünkü gözlem sayılarının yeterince büyük olması durumunda ($n_j > 5$) Kruskal-Wallis istatistiği serbestlik derecesi (SD) = *k-1*'lik bir dağılım gösterir. Araştırmamızda her arama motoruna ait gözlem sayısı 17 olduğundan kritik değerler için χ^2 tablosu kullanılmıştır. Kruskal-Wallis istatistiği hakkında ayrıntılı bilgi için bkz. Kartal (1998, s. 211 ve devamı) ve Ünver ve Gamgam (1999, s. 373 ve devamı).

arama motorunun duyarlık değerleri arasında istatistiksel açıdan anlamlı bir fark olduğuna karar verilir.[4]

Çeşitli kesme noktalarındaki ortalama duyarlık değerleriyle ortalama normalize sıralama değerleri arasındaki ilişki Pearson korelasyon katsayısı (*r)* ile test edilmiştir. Pearson korelasyon katsayısı (*r)* -1 ile 1 arasında değerler almakta ve iki örneğe ait değerler arasında fark olup olmadığını, fark varsa bu farkın yönünü (negatif ya da pozitif) ve gücünü gösterir. İyi bilinen ve hipotez testlerinde yaygın olarak kullanılan Pearson korelasyon katsayısının hesaplanmasıyla ilgili ayrıntılı bilgi çeşitli istatistik kitaplarından edinilebilir.

Araştırmamızda tüm istatistik testler için %95 güven düzeyi (iki yönlü) kullanılmıştır. İstatistik testlerin hesaplanmasında SPSS for Windows (sürüm 9.05) yazılımından yararlanılmıştır.

[4] Mann-Whitney U testiyle ilgili bilgiler için Kartal'ın eserinden (1998, s. 189 ve devamı) yararlanılmıştır. Ayrıntılı bilgi için bkz. Kartal (1998) ve Ünver ve Gamgam (1999).

5 BULGULAR VE YORUM

Bu bölümde araştırmadan elde edilen bulgular verilmekte ve söz konusu bulgular analiz edilerek yorumlanmaktadır. Önce 17 soru için dört arama motorundan elde ettiğimiz erişim sonuçları güncellik ile duyarlık ve normalize sıralama değerleri açısından incelenmiş ve birbiriyle karşılaştırılmıştır. Daha sonra Türkçe arama motorlarında en sık aranan beş sözcüğe dayanarak ölçülen arama motorlarının kapsama ve yenilik oranlarıyla ilgili bulgular verilmiştir. Son olarak, arama motorlarında erişim amacıyla üst veri belirteçlerinden yararlanılıp yararlanılmadığı konusunda yaptığımız iki küçük testin sonuçları kısaca özetlenmiştir.

Toplam 17 soru için Arabul, Arama, Netbul ve Superonline arama motorlarında erişilen belgeler üzerinde yapılan ilgililik değerlendirmeleri Şekil 6'da verilmektedir. Değerlendirmelerde (varsa) erişilen ilk 20 belge dikkate alınmıştır. Erişilen her belge "ilgili" veya "ilgisiz" olarak sınıflandırılmış, aramanın gerçekleştirildiği anda çalışmayan (ölü) bağlantılar da not edilmiştir. Şekil 6 esas olarak güncellik, duyarlık ve normalize sıralama değerlerini analiz ederken kullandığımız ham verilerin birçoğunu içermektedir. Örneğin, her arama motorundan elde edilen çalışmayan bağlantı sayısına bakarak arama motorlarının güncelliği Şekil 6'dan yararlanılarak oluşturulmuştur. Benzeri bir biçimde, arama motorlarının çeşitli kesme noktalarındaki duyarlık ve normalize sıralama değerleri ile ilgili tablolar Şekil 6'dan yararlanılarak hesaplanmıştır.

Sorgular	Arabul	Arama	Netbul	Superonline
1.	0 - - - - - - -	+ + + - - - - - - - 0 0 - 0 - - - 0 0 -	- - - - - - - - - - - - 0 - - - - - - - - -	- 0 - - - - - - 0 - - - - - 0 - - - - -
2.	+	0 0 - - 0 - - + + - - 0 0 0	- -	0 - + - - 0 + 0 - 0 0 -
3.	0	+ - - - - + - 0 0 - + - 0 - 0 0 0 - - -	+ - - - - - + + - 0 - + - - 0 - 0 - - -	+ - + - + + - - + + + + - 0 + - - - - +
4.		- - - - - - - 0 - - - 0 0 - -	- - - - 0 - - 0 - - - - - - - - - - - -	- 0 0 0
5.	- 0 -	+ + + + - - - 0 - -	+ + - - - - - - - - + - - - - - - - - -	0 - - - - 0 - - - - - - - - - 0 - + - -
6.	- 0	+ + + 0 - - - - - - - - 0 - - - - 0 0 0	- - 0 - - - - - - - - - 0 - 0 - - - - -	- - 0 - - - - 0 - 0 - - - - - - - - - -
7.	+ 0 - +	+ 0 + + + + + - + 0 + - + + + - - + - + -	- - + - - - - - - - - - - - - - - - - -	- + 0 - - - - - - - - - - - - - 0 - - -
8.	- + + + 0 - - - + 0 - + - - - - + - 0 -	- - + + + 0 - + - 0 - 0 0 0 0 - - - - -	- - - + - + + - - 0 - - -	- + 0 - - - - 0 - - + + - - - - 0 - + -
9.	+ 0 - + 0 - - - - - + + 0 + + 0 + + + 0	0 + + + - - - - 0 - 0 + - 0 - - 0 - 0 0	- - + - - + - - - -	+ - - + - - - - - - - - - 0 + - - - -
10.		- + + 0 0 0 + 0 0 0 + 0 - + 0 0 0 0 0 0	- - + - + - - - - - - - - - - - - - -	+ 0 + + + - 0 + + + + - 0 0 - - - - 0 0
11.		+ + + 0 0 0 + 0 0 0 + 0 - + 0 0 0 0 0 0	- - + - + + + - - - - - - - - - - - - -	+ 0 + + + + 0 + + + + + 0 0 - - - - 0 0
12.		- -	- - - - - - - - - - - - - - - - - - - -	- 0 - - - - 0 - - - - - - - - 0 0 - 0 -
13.	0 - - - - - - + + 0 + + - + 0 - - - - -	+ - - - 0 0 - - 0 0 + - 0 - 0 - 0 + 0 -	- - - - + - - - - - - - + - - - - - -	+ - - - - - - - - - - - - - - - - + - -
14.	- - 0 - - - - - - - - - - - - - 0 0 - -	- 0 - - - - - - 0 - 0 - 0 - - 0 0 - 0 0	- - - - - 0 - - - - - - - - - - - - -	- - - - - - - - - - - - - 0 - - - 0 + -
15.	- 0 - - + - 0 - - - - 0 + 0 - - - 0 - -	+ + - + + - - - + - + - + - 0 - 0 - + - 0	- 0 - - - + - + - - + - - + - - - + - -	- + - + 0 + - - - + - - - - + - + - - -
16.		- + + + - + - - + - - + + - - -	- - - - - - -	0 + - + - - - - + + - + - 0 + - - 0 - -
17.			-	- - 0 - - -

Not: “+” ilgili, “-” ilgisiz belgeyi, “0” ise çalışmayan ya da ölü bağlantıları göstermektedir. Belirli bir gözde herhangi bir işaretin olmaması ilgili arama motorunun o soru için ilgili ya da ilgisiz hiç bir belgeye erişemediğini göstermektedir.

Şekil 6. İlgililik değerlendirmeleri

5.1 Arama Motorlarının Güncelliği

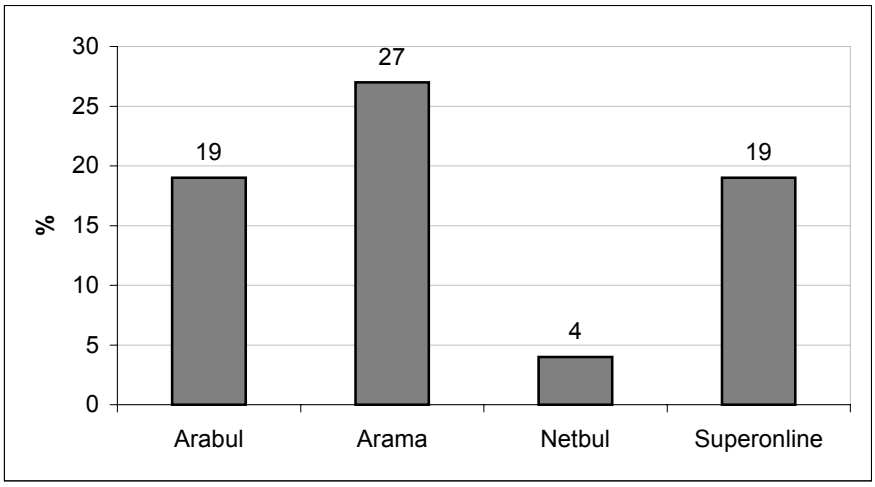
Bir önceki bölümde (4.6) değinildiği gibi, arama motorlarının güncelliğini ölçmek için tasarladığımız deney, ilgili arama motorlarından bazılarının gönderilen sayfaları uzun süre dizinlememeleri nedeniyle gerçekleştirilememiştir. Örneğin; iki farklı Web sunucusuna (http://cmpe.emu.edu.tr/bitirim/vartanbitirim, http://www.geocities.com/vartanbitirim) yerleştirilmiş olan aynı belge dört arama motoruna da, (Arama, Arabul, Netbul ve Superonline) 18 Ekim 2001 tarihinde ayrı ayrı kaydedilmiş, fakat 5 Şubat 2002 tarihine kadar yapılan denetlemeler sonucunda söz konusu belgenin Arama dışındaki diğer arama motorları tarafından dizinlenmediği görülmüştür. Ancak arama motorlarının güncelliği belirli bir soru karşılığında erişilen belgelerin "canlı" olup olmamasıyla da ölçülmektedir. Bir başka deyişle, erişilen belgelerin URL adreslerine tıklandığında belgenin Web'den kaldırılması, adının değişmesi, bir başka sunucu ya da dizin altında listelenmesi, vb. gibi nedenlerle "adres bulunamadı" mesajıyla karşılaşılıyorsa bu adres "ölü" demektir. Bu da, arama motoru yazılımlarının dizinlenen adresleri hangi sıklıkla ziyaret ederek ölü olup olmadıklarının kontrol ettiklerinin bir göstergesidir.

Arabul, Arama, Netbul ve Superonline arama motorlarının güncelliğini ölçebilmek için her soru için erişilen belgeler arasındaki "ölü" bağlantılar sayılarak makro ortalama yöntemiyle arama motorlarının ölü bağlantı yüzdeleri hesaplanmıştır. Tablo 8'de arama motorlarına göre her soru için erişilen/değerlendirilen belge sayıları ve ölü bağlantı sayıları verilmektedir. Şekil 7'de ise arama motorlarının ortalama ölü bağlantı oranları verilmektedir.

Tablo 8. Arama motorlarının ölü bağlantı oranları

Sorgular	Arabul			Arama			Netbul			Superonline			Ort.
	ÖBS	ETBS	%	ÖBS	ETBS	%	ÖBS	ETBS	%	ÖBS	ETBS	%	(%)
1	1	8	13	5	20	25	1	20	5	3	20	15	**14**
2	0	1	0	6	14	43	0	2	0	5	12	42	**21**
3	1	1	100	6	20	30	3	20	15	1	20	5	**38**
4	0	0	0	3	15	20	2	20	10	3	4	75	**26**
5	1	3	33	1	10	10	0	20	0	3	20	15	**15**
6	1	2	50	5	20	25	3	20	15	3	20	15	**26**
7	1	4	25	2	20	10	0	20	0	2	20	10	**11**
8	3	20	15	6	20	30	1	13	8	3	20	15	**17**
9	5	20	25	7	20	35	0	10	0	1	20	5	**16**
10	0	0	0	13	20	65	0	20	0	6	20	30	**24**
11	0	0	0	13	20	65	0	20	0	6	20	30	**24**
12	0	0	0	0	2	0	0	20	0	5	20	25	**6**
13	3	20	15	8	20	40	0	20	0	0	20	0	**14**
14	3	20	15	8	20	40	1	20	5	2	20	10	**18**
15	5	20	25	3	20	15	1	20	5	1	20	5	**13**
16	0	0	0	0	16	0	0	7	0	3	20	15	**4**
17	0	0	0	0	0	0	0	1	0	1	6	17	**4**
Ortalama Ölü Bağlantı Sayısı - Oranı	**1,4**		**%19**	**5,1**		**%27**	**0,7**		**%4**	**2,8**		**%19**	**%17**

Not: Erişilen ilk 20 belge değerlendirilmiş ve ortalama ölü bağlantı yüzdesi makro ortalama yöntemine göre hesaplanmıştır. ÖBS: Ölü Bağlantı Sayısı; ETBS: Erişilen/Değerlendirilen Toplam Belge Sayısı



Şekil 7. Arama motorlarının ortalama ölü bağlantı oranları

Tablo 8 ve Şekil 7'den de görülebileceği gibi, Arama, %27 ile ölü bağlantı oranı en yüksek arama motorudur. Bir başka deyişle Arama'nın eriştiği her dört belgeden yaklaşık birisi ölü bağlantı içermektedir. Arama'yı %19 ile Arabul ve Superonline izlemektedir. Ölü bağlantı oranı en düşük arama motoru ise Netbul'dur (%4). Arama'nın soru başına ortalama ölü bağlantı sayısı 5,1, Superonline'ın 2,8, Arabul'un 1,4, Netbul'un ise 0,7'dir. Arama motorlarının eriştikleri ortalama her 6 belgeden birisi (%17) ölü bağlantı içermektedir.

Arama motorlarının ölü bağlantı oranları arasındaki farkın istatistiksel açıdan anlamlı olup olmadığını saptamak için Kruskal-Wallis testi uygulanmış ve fark anlamlı bulunmuştur (H = 17,49, $s.d$ = 3, $p < .001$).[1] Ölü bağlantı oranı en düşük olan Netbul'un değerleri çıkarılarak Kruskal-Wallis katsayısı yeniden hesaplanmıştır. Netbul'un dışındaki diğer üç arama motorunun (Arabul, Arama ve Superonline) ölü bağlantı oranları arasında anlamlı bir fark yoktur (H = 3,02, $s.d.$ = 2, $p < .220$). Bir başka deyişle, Netbul'un bağlantılarının diğer arama motorlarının bağlantılarından daha sık aralıklarla güncelleştirildiği ve ölü bağlantıların daha hızlı ayıklandığı söylenebilir.

Arama motorlarının sorulara göre ölü bağlantı sayıları karşılaştırıldığında (Tablo 8, son sütun), soru türleriyle ölü bağlantı sayısı arasında bir ilişki olmadığı gözlenmektedir. Tüm arama motorları için ortalama ölü bağlantı oranı en yüksek olan soru 3. sorudur. Ancak bunun nedeni Arabul'un 3. soru için toplam 1 belgeye erişmesi ve bu belgenin de ölü olmasından kaynaklanmaktadır. Ölü bağlantı oranları 4. ve 6. sorular için %26'dır. Superonline'ın 4. soru için ölü bağlantı oranının yüksek olması bu soru için ortalamayı yükseltmiştir. Benzeri bir biçimde 10. ve 11. sorularda ortalama ölü bağlantı oranı (%24), Arama'nın oranları nedeniyle yüksek çıkmıştır. Ölü bağlantı oranları en düşük olan sorular 12., 16, ve 17. sorulardır (%4). Bu sorular için arama motorlarının genellikle ya ilgili belgelere erişemedikleri, ya da erişilen toplam belge sayısının düşük olduğu görülmektedir.

Arama motorlarının hangi sorular için sıfır sonuç verdiği Tablo 8'den izlenebilir. Arabul 17 sorudan 6'sı (4, 10, 11, 12, 16 ve 17. sorular) için hiç bir belgeye erişememiştir. Bu rakam, Arabul'un soruların %35'i için sıfır sonuç verdiği anlamına gelmektedir. Arama, bir soruya (17. soru) karşılık sıfır sonuç vermiştir. Netbul ve Superonline ise tüm sorular için en az bir belgeye erişmiştir.

[1] Hesaplanan Kruskal-Wallis (H) değeri, χ^2 (3, 0,05) = 7,81 tablo değeriyle karşılaştırılmıştır.

5.2 Arama Motorlarının Duyarlık ve Normalize Sıralama Performansları

Yöntem ve Tasarım bölümünde (4. bölüm) de değinildiği gibi, Türkçe arama motorlarının performans düzeylerini belirlemek için duyarlık ve normalize sıralama değerleri kullanılmıştır. Toplam 17 sorgu (bkz. 4.3) Arabul, Arama, Netbul ve Superonline arama motorlarında tek tek aranmış, erişilen belgeler belirlenen ölçütlere göre (bkz. 4.5) "ilgili" ve "ilgisiz" olarak sınıflandırılmış ve daha önce verilen formüller kullanılarak (bkz. 4.6, formül 6 ve formül 8) arama motorlarının çeşitli kesme noktalarındaki duyarlık ve normalize sıralama değerleri hesaplanmıştır.[2] Erişilen belge sayısı sıfır olan sorgularda normalize sıralama değeri sıfır olarak alınmıştır. Tüm sorular için ortalama duyarlık ve ortalama normalize sıralama değerleri makro ortalama yöntemi kullanılarak bulunmuştur.

5.2.1 Bireysel Değerlendirme

5.2.1.1 Arabul

Arabul için 5, 10, 15 ve 20 belge değerlendirildikten sonra kaydedilen duyarlık ve normalize sıralama değerleri Tablo 9'da verilmektedir. Tablo 9'da dikkati çeken ilk noktalardan birisi, Arabul'un 17 sorudan 6'sı için hiç bir belgeye erişememiş olmasıdır.[3] Geri kalan 11 soru için Arabul, toplam 119 belgeye erişmiş,[4] bunlardan 25'i ilgili bulunmuştur (soru başına yaklaşık 2 ilgili belge). Dikkati çeken bir başka nokta ise Arabul'un toplam 5 soru için[5] herhangi bir ilgili belgeye erişememesidir (yani bu sorular için duyarlık değeri 0'dır). Bir başka deyişle, sıfır sonuç veren sorular da eklendiğinde, Arabul, toplam 17 sorudan 11'inde ilgili belgelere erişememiştir. Arabul'un ilgili belgelere eriştiği sorularda[6] erişilen ilk 20 belge için duyarlık değerleri %10 ile %100, normalize sıralama değerleri ise %33 ile %100 arasında değişmektedir.

Arabul'un ortalama duyarlık değerleri çeşitli kesme noktalarında pek büyük bir değişim göstermemektedir (%16). Ortalama normalize sıralama değerleri ise erişilen ilk 5 belgede %16 iken ilk 10 belgede önemli bir artış kaydederek %22'ye yükselmiş, ancak erişilen belge

---

[2] Arama motorlarının her sorgu için eriştikleri belgelerin listeleri http://cmpe.emu.edu.tr/bitirim/home adresinden çevrimiçi olarak görülebilir.

[3] Dördüncü, 10., 11., 12., 16. ve 17. sorular.

[4] Toplam erişilen belge sayısı ilgili (İİ) ve ilgisiz (İZ) belgelerin toplamına eşittir. Gerçekte erişilen toplam belge sayıları kuşkusuz daha yüksektir. Ancak araştırmamızda ilk 20 belge değerlendirmeye alınmıştır.

[5] Birinci, 3., 5., 6. ve 14. sorular.

[6] İkinci, 7., 8., 9., 13. ve 15. sorular.

sayısı 15 ve 20'ye yükseldiğinde aynı artış sürdürülememiştir (sırasıyla %19 ve %21). Arabul, 17 sorudan 11'inde ilgili belgeye erişemediğinden tüm kesme noktalarında hem duyarlık hem de normalize sıralama değerlerinin ortancaları sıfırdır.

Tablo 9. Arabul'un çeşitli kesme noktalarında duyarlık ve normalize sıralama değerleri

	Arabul															
	Kesme noktası 5				Kesme noktası 10				Kesme noktası 15				Kesme noktası 20			
Sorgular	İİ	İZ	D	NS	İİ	İZ	D	NS	İİ	İZ	D	NS	İİ	İZ	D	NS
1	0	5	0	0	0	8	0	0	0	8	0	0	0	8	0	0
2	1	0	100	100	1	0	100	100	1	0	100	100	1	0	100	100
3	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	3	0	0	0	3	0	0	0	3	0	0	0	3	0	0
6	0	2	0	0	0	2	0	0	0	2	0	0	0	2	0	0
7	2	2	50	50	2	2	50	50	2	2	50	50	2	2	50	50
8	3	2	60	50	4	6	40	67	5	10	33	70	6	14	30	69
9	2	3	40	67	2	8	20	88	6	9	40	33	9	11	45	33
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	5	0	0	2	8	20	13	5	10	33	22	5	15	25	48
14	0	5	0	0	0	10	0	0	0	15	0	0	0	20	0	0
15	1	4	20	0	1	9	10	56	2	13	13	42	2	18	10	58
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Toplam	**9**	**32**			**12**	**57**			**21**	**73**			**25**	**94**		
İİ (ortanca)	**0**				**0**				**0**				**0**			
D (ort.)			**16**				**14**				**16**				**15**	
D (ortanca)			**0**				**0**				**0**				**0**	
NS (ort.)				**16**				**22**				**19**				**21**
NS (ortanca)				**0**				**0**				**0**				**0**

İİ: İlgili belge sayısı; **İZ**: İlgisiz belge sayısı; **D**: Duyarlık; **NS**: Normalize sıralama; **ort**.: Aritmetik ortalama

5.2.1.2 Arama

Arama için 5, 10, 15 ve 20 belge değerlendirildikten sonra kaydedilen duyarlık ve normalize sıralama değerleri Tablo 10'da verilmektedir. Arama 17 sorudan sadece 1'inde (17. soru) hiç bir belgeye erişememiştir. Geri kalan sorular için Arama, toplam 277 belgeye erişmiş, bunlardan 64'ü ilgili bulunmuştur (soru başına yaklaşık 4 ilgili belge). Üç soru[7] için ise erişilen belgeler arasında ilgili belgelere rastlanmamıştır (duyarlık değeri 0). Arama'nın erişilen ilk 20 belge için duyarlık değerleri %14 ile %60 (ort. = %21, ortanca = %15), normalize sıralama

[7] Dördüncü, 12. ve 14. sorular.

değerleri ise %0 ile %100 (ort. = %54, ortanca = %65), arasında değişmektedir. Soru başına erişilen ilgili belgelerin ortancası tüm kesme noktalarında 3'tür.

Tablo 10. Arama'nın çeşitli kesme noktalarında duyarlık ve normalize sıralama değerleri

	Arama															
	Kesme noktası 5				Kesme noktası 10				Kesme noktası 15				Kesme noktası 20			
Sorgular	İİ	İZ	D	NS	İİ	İZ	D	NS	İİ	İZ	D	NS	İİ	İZ	D	NS
1	3	2	60	100	3	7	30	100	3	12	20	100	3	17	15	100
2	0	5	0	0	2	8	20	13	2	12	14	32	2	12	14	20
3	1	4	20	100	2	8	20	75	3	12	20	67	3	17	15	76
4	0	5	0	0	0	10	0	0	0	15	0	0	0	15	0	0
5	4	1	80	100	4	6	40	100	4	6	40	59	4	6	40	52
6	3	2	60	100	3	7	30	100	3	12	20	100	3	17	15	100
7	4	1	80	25	7	3	70	57	10	5	67	58	12	8	60	65
8	3	2	60	0	4	6	40	58	4	11	27	77	4	16	20	84
9	3	2	60	50	3	7	30	86	4	11	27	75	4	16	20	83
10	2	3	40	67	3	7	30	71	5	10	33	56	5	15	25	71
11	3	2	60	100	4	6	40	88	6	9	40	69	6	14	30	80
12	0	2	0	0	0	2	0	0	0	2	0	0	0	2	0	0
13	1	4	20	100	1	9	10	100	2	13	13	65	3	17	15	53
14	0	5	0	0	0	10	0	0	0	15	0	0	0	20	0	0
15	4	1	80	50	5	5	50	76	7	8	47	70	8	12	40	72
16	3	2	60	50	5	5	50	64	7	8	47	63	7	9	44	55
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Toplam	**34**	**43**			**46**	**106**			**60**	**161**			**64**	**213**		
İİ (ortanca)	**3**				**3**				**3**				**3**			
D (ort.)			**40**				**27**				**24**				**21**	
D (ortanca)			**60**				**30**				**20**				**15**	
NS (ort.)				**50**				**58**				**52**				**54**
NS (ortanca)				**50**				**71**				**63**				**65**

İİ: İlgili belge sayısı; **İZ**: İlgisiz belge sayısı; **D**: Duyarlık; **NS**: Normalize sıralama; **ort**.: Aritmetik ortalama

Arama'nın ortalama duyarlık değerleri çeşitli kesme noktalarında önemli değişiklikler göstermektedir. Erişilen ilk 5 belgede %40 olan ortalama duyarlık değeri kesme noktası arttıkça giderek düşmüş ve erişilen ilk 20 belgede %21'e inmiştir. Duyarlık değerlerinin ortancaları ise ortalamalardan daha hızlı bir düşüş göstermiştir (ilk 5 belgede %60, ilk 20'de %15). Ortalama normalize sıralama değerleri ise erişilen ilk 5 belgede %50 iken ilk 10 belgede önemli bir artış kaydederek %58'e yükselmiş, ancak erişilen belge sayısı 15 ve 20'ye yükseldiğinde aynı artış sürdürülememiştir (sırasıyla %52 ve %54). Normalize sıralama değerlerinin ortancaları ortalamaya benzer bir değişim göstermiştir.

5.2.1.3 Netbul

Netbul için 5, 10, 15 ve 20 belge değerlendirildikten sonra kaydedilen duyarlık ve normalize sıralama değerleri Tablo 11'de verilmektedir. Netbul bütün sorular için en az bir belgeye erişmiştir. Bir başka deyişle, Netbul'da sıfır sonuç veren soru olmamıştır. Ancak Netbul, toplam 17 sorudan 8'inde hiç bir ilgili belgeye erişememiştir (duyarlık değeri 0).[8] Geri kalan sorular için Netbul, toplam 273 belgeye erişmiş, bunlardan 26'sı ilgili bulunmuştur (soru başına yaklaşık 3 ilgili belge). Netbul'un erişilen ilk 20 belge için duyarlık değerleri %5 ile %25 (ort. = %9, ortanca = %5), normalize sıralama değerleri ise %0 ile %89 (ort. = %34, ortanca = %26), arasında değişmektedir. Soru başına erişilen ilgili belgelerin ortancası kesme noktası 5 iken 0, diğer kesme noktalarında 1'dir.

Tablo 11. Netbul'un çeşitli kesme noktalarında duyarlık ve normalize sıralama değerleri

	Netbul															
	Kesme noktası 5				Kesme noktası 10				Kesme noktası 15				Kesme noktası 20			
Sorgular	İİ	İZ	D	NS	İİ	İZ	D	NS	İİ	İZ	D	NS	İİ	İZ	D	NS
1	0	5	0	0	0	10	0	0	0	15	0	0	0	20	0	0
2	0	2	0	0	0	2	0	0	0	2	0	0	0	2	0	0
3	1	4	20	100	3	7	30	52	4	11	27	59	4	16	20	72
4	0	5	0	0	0	10	0	0	0	15	0	0	0	20	0	0
5	2	3	40	100	2	8	20	100	3	12	20	78	3	17	15	84
6	0	5	0	0	0	10	0	0	0	15	0	0	0	20	0	0
7	1	4	20	50	1	9	10	78	1	14	7	86	1	19	5	89
8	1	4	20	25	3	7	30	48	3	10	23	45	3	10	23	33
9	1	4	20	50	2	8	20	63	2	8	20	30	2	8	20	26
10	2	3	40	17	2	8	20	69	2	13	13	81	2	18	10	86
11	2	3	40	17	4	6	40	54	4	11	27	75	4	16	20	83
12	0	5	0	0	0	10	0	0	0	15	0	0	0	20	0	0
13	1	4	20	0	1	9	10	56	2	13	13	42	2	18	10	58
14	0	5	0	0	0	10	0	0	0	15	0	0	0	20	0	0
15	0	5	0	0	2	8	20	31	4	11	27	34	5	15	25	45
16	0	5	0	0	0	7	0	0	0	7	0	0	0	7	0	0
17	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0
Toplam	**11**	**67**			**20**	**130**			**25**	**188**			**26**	**247**		
İİ (ortanca)	**0**				**1**				**1**				**1**			
D (ort.)			**13**				**12**				**10**				**9**	
D (ortanca)			**0**				**10**				**7**				**5**	
NS (ort.)				**21**				**32**				**31**				**34**
NS (ortanca)				**0**				**31**				**30**				**26**

İİ: İlgili belge sayısı; **İZ**: İlgisiz belge sayısı; **D**: Duyarlık; **NS**: Normalize sıralama; **ort**.: Aritmetik ortalama

[8] Birinci, 2., 4., 6., 12., 14., 16. ve 17. sorular.

Netbul'un erişilen ilk 5 belgede %13 olan ortalama duyarlık değeri ilk 20 belgede %9'a düşmüştür. Çeşitli kesme noktalarındaki duyarlık değerlerinin ortancaları ise büyük değişiklikler göstermektedir. Erişilen ilk 5 belgede duyarlık değerlerinin ortancası %0 iken, ilk 10 belgede %10'a yükselmiş, ilk 15 belgede %7'ye, ilk 20 belgede ise %5'e düşmüştür.

Netbul'un ortalama normalize sıralama değerlerinin genelde kesme noktası yükseldikçe arttığı gözlenmiştir. Erişilen ilk 5 belgede %21 olan ortalama duyarlık değeri, ilk 10 belgede %32'ye yükselmiş, ilk 15 belgede biraz düşmüş (%31), ilk 20 belgede ise yeniden %34'e yükselmiştir. Netbul'un ortalama normalize sıralama değerlerinin ortancaları da ortalamalara benzer bir değişim göstermiştir (erişilen ilk 5 belgede 0, ilk 10'da %31, ilk 15'te %30, ilk 20'de %26).

Netbul'un ortalama duyarlık ve normalize sıralama değerlerindeki ve bu değerlerin ortancalarındaki değişim, Netbul'un toplam 8 soru için hiç bir ilgili belgeye erişememesi ve diğer sorular için de erişilen ilk 5 belgede genellikle ilgili belgelere rastlanmaması ile açıklanabilir.

5.2.1.4 Superonline

Superonline için 5, 10, 15 ve 20 belge değerlendirildikten sonra kaydedilen duyarlık ve normalize sıralama değerleri Tablo 12'de verilmektedir. Superonline da bütün sorular için en az bir belgeye erişmiştir. Yani, Superonline'da sıfır sonuç veren soru olmamıştır. Superonline toplam 17 sorudan 4'ünde[9] hiç bir ilgili belgeye erişememiştir (duyarlık değeri 0). Superonline toplam 302 belgeye erişmiş, bunlardan 54'ü ilgili bulunmuştur (soru başına yaklaşık 3 ilgili belge). Superonline'ın erişilen ilk 20 belge için duyarlık değerleri %0 ile %50 (ort. = %16, ortanca = %10), normalize sıralama değerleri ise %0 ile %84 (ort. = %39, ortanca = %47), arasında değişmektedir. Soru başına erişilen ilgili belgelerin ortancası erişilen belge sayısı 5, 10 ve 15 iken 1'dir. Fakat ortanca, erişilen ilk 20 belgede 2'ye yükselmiştir. Superonline'da gösterilen belge sayısı arttıkça ilgili belgelere erişme olasılığının da arttığı söylenebilir.

Superonline'ın erişilen ilk 5 belgede %25 olan ortalama duyarlık değeri ilk 20 belgede %16'ya düşmüştür. Duyarlık değerlerinin ortancaları da ilk 5 belgede %20 iken ilk 20 belgede %10'a düşmüştür.

Tablo 12. Superonline'ın çeşitli kesme noktalarında duyarlık ve normalize sıralama değerleri

[9] Birinci, 4., 16. ve 17. sorular.

	Superonline															
	Kesme noktası 5				Kesme noktası 10				Kesme noktası 15				Kesme noktası 20			
Sorgular	İİ	İZ	D	NS	İİ	İZ	D	NS	İİ	İZ	D	NS	İİ	İZ	D	NS
1	0	5	0	0	0	10	0	0	0	15	0	0	0	20	0	0
2	1	4	20	50	2	8	20	56	2	10	17	34	2	10	17	25
3	3	2	60	50	6	4	60	46	9	6	60	50	10	10	50	63
4	0	4	0	0	0	4	0	0	0	4	0	0	0	4	0	0
5	0	5	0	0	0	10	0	0	0	15	0	0	1	19	5	11
6	0	5	0	0	0	10	0	0	0	15	0	0	0	20	0	0
7	1	4	20	75	1	9	10	89	1	14	7	93	1	19	5	95
8	1	4	20	75	1	9	10	89	3	12	20	47	4	16	20	47
9	2	3	40	67	2	8	20	88	2	13	13	92	3	17	15	71
10	4	1	80	25	7	3	70	43	8	7	53	73	8	12	40	84
11	4	1	80	25	8	2	80	38	10	5	67	72	10	10	50	86
12	0	5	0	0	0	10	0	0	0	15	0	0	0	20	0	0
13	1	4	20	100	1	9	10	100	1	14	7	100	2	18	10	56
14	0	5	0	0	0	10	0	0	0	15	0	0	1	19	5	5
15	2	3	40	50	4	6	40	50	5	10	33	56	6	14	30	61
16	2	3	40	50	4	6	40	38	6	9	40	43	6	14	30	63
17	0	5	0	0	0	6	0	0	0	6	0	0	0	6	0	0
Toplam	**21**	**63**			**36**	**124**			**47**	**185**			**54**	**248**		
İİ (ortanca)	**1**				**1**				**1**				**2**			
D (ort.)			**25**				**21**				**19**				**16**	
D (ortanca)			**20**				**10**				**7**				**10**	
NS (ort.)				**33**				**37**				**39**				**39**
NS (ortanca)				**25**				**38**				**43**				**47**

İİ: İlgili belge sayısı; **İZ**: İlgisiz belge sayısı; **D**: Duyarlık; **NS**: Normalize sıralama; **ort**.: Aritmetik ortalama

Superonline'da erişilen belge sayısı arttıkça ortalama normalize sıralama değerlerinin de arttığı gözlenmiştir (erişilen ilk 5 belgede %33, ilk 20 belgede %39). Benzeri bir biçimde normalize sıralama değerlerinin ortancalarında da sürekli bir artış gözlenmiştir (erişilen ilk 5 belgede %25, ilk 20 belgede %47).

5.2.2 Toplu Değerlendirme

5.2.2.1 Arama Motorlarının Eriştikleri İlgili Belge Sayıları

Arama motorlarının performansları, sıfır sonuç veren (bir soruya karşılık hiç bir belgeye erişilememesi) veya duyarlık değeri sıfır olan (bir soruya karşılık hiç bir ilgili belgeye erişilememesi) soru sayısı açısından değerlendirilebilir. Arabul toplam 17 sorudan 6'sında, Arama ise 1'inde hiç bir belgeye erişememiştir. Netbul ve Superonline ise her soru için en az 1 belgeye erişmiştir. Netbul toplam 17 sorudan 8'inde hiç bir ilgili belgeye erişememiştir. Netbul'u 5 soruyla Superonline, 4 soruyla Arabul ve 3 soruyla Arama izlemektedir. Sıfır sonuç veren ve duyarlık değeri sıfır olan soru sayısı birlikte değerlendirildiğinde, Arabul 17

sorudan 11 (%65), Netbul 8 (%47), Superonline 5 (%29), Arama ise 4 (%24) soruda ilgili belgelere erişememiştir.

Sorulara göre arama motorlarının ilgili belgelere erişip erişemedikleri ve erişilen ilk 20 belgedeki ilgili belge sayıları Tablo 13'te verilmektedir. Hiç bir arama motoru 4., 12. ve 17. sorular için ilgili belgeye erişememiştir. Birinci, 6. ve 14. sorular için üç arama motoru, 16. soru için iki, 2., 3., 5., 10. ve 11. sorular için ise en az bir arama motoru ilgili belgelere erişememiştir. Yedinci, 8., 9., 13. ve 15. sorular için ise tüm arama motorları en az bir ilgili belgeye erişmiştir.

Tablo 13. Sorulara göre erişilen ilgili belge sayısı

Soru	Arabul	Arama	Netbul	Superonline
1	0	3	0	0
2	1	2	0	2
3	0	3	4	10
4	0	0	0	0
5	0	4	3	1
6	0	3	0	0
7	2	12	1	1
8	6	4	3	4
9	9	4	2	3
10	0	5	2	8
11	0	6	4	10
12	0	0	0	0
13	5	3	2	2
14	0	0	0	1
15	2	8	5	6
16	0	7	0	6
17	0	0	0	0
Toplam	**25** (119)	**64** (277)	**26** (273)	**54** (302)
Ortalama	**1,5**	**3,8**	**1,5**	**3,2**

Not: Gri renkli gözler ilgili arama motorunun o soru için ya hiç bir belgeye erişemediğini ya da erişilen belgeler arasında hiç bir ilgili belge bulunmadığını ifade etmektedir. Toplam satırındaki ilk değer tüm sorular için arama motorunun eriştiği ilgili belge sayısını, parantez içindeki değer ise erişilen toplam belge sayısını vermektedir.

Toplam 17 soru için en fazla ilgili belgeye erişen arama motoru Arama'dır (64). Arama'yı Superonline (54), Netbul (26) ve Arabul (25) izlemiştir. Bu açıdan bakıldığında, Arama'nın soru başına erişilen ortalama ilgili belge sayısının (3,8) diğer arama motorlarından yüksek olması kolayca açıklanabilir (Superonline 3,2; Arabul ve Netbul 1,5). Arabul ve Netbul'un sırasıyla toplam 11 ve 8 soru için hiç bir ilgili belgeye erişememeleri bu arama motorlarının soru başına düşen ortalama ilgili belge sayısına da yansımıştır.

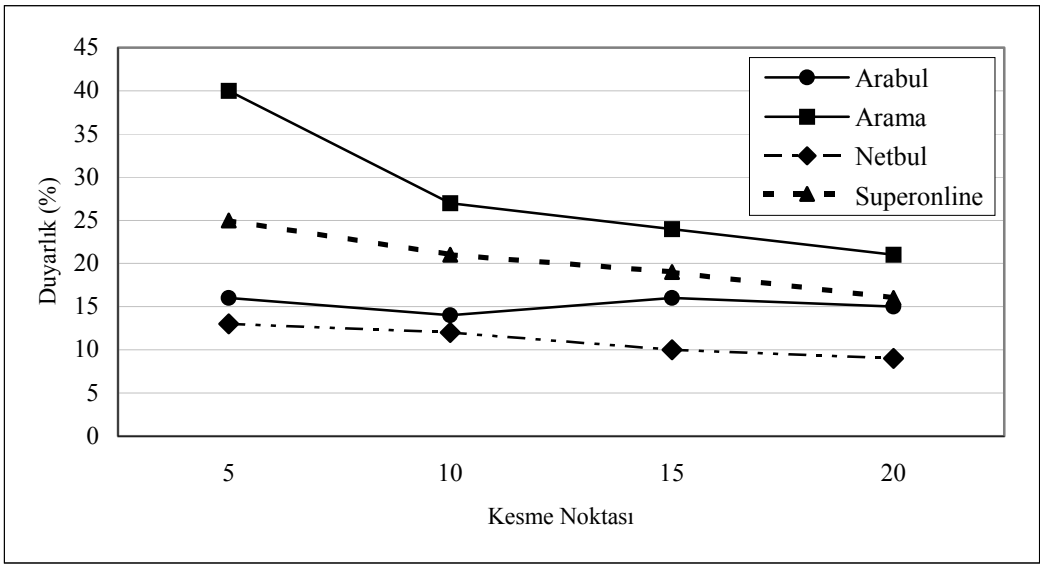
Netbul 17 soru için 273 belgeye erişmesine rağmen, bunlardan sadece 26'sı ilgili bulunmuştur. Arabul'da ise sıfır sonuç vermeyen toplam 11 soruya karşılık 119 belgeye erişilmiş ve bunlardan 25'i ilgili bulunmuştur.

Dört arama motorunun eriştiği toplam ilgili belge sayısı 169'dur. Bir an için her arama motoru tarafından erişilen belgelerin tekil olduğunu varsayacak olursak, bu rakam, soru başına dört arama motoru tarafından ortalama 10 ilgili belgeye erişildiği anlamına gelmektedir.

Dört arama motoru tarafından erişilen toplam belge sayısı 971'dir. Bunlardan 169'u ilgili bulunmuştur. Arama motorlarının eriştikleri yaklaşık her 6 belgeden 5'i ilgisizdir.

5.2.2.2 Arama Motorlarının Ortalama Duyarlık Değerleri

Arama motorlarının çeşitli kesme noktalarında (erişilen ilk 5, 10, 15, ve 20 belge için) kaydettikleri ortalama duyarlık değerleri Şekil 8'de verilmektedir.



Şekil 8. Ortalama duyarlık değerleri

Şekil 8'den de görülebileceği gibi, çeşitli kesme noktalarında en yüksek duyarlık değerini Arama kaydetmiştir (ort. %28). Arama'yı Superonline (%20) ve Arabul (%15) izlemektedir. Superonline ve Arabul'un ortalama duyarlık değerleri kesme noktası 20'ye yükseldiğinde hemen hemen birbirine eşit hale gelmiştir. En düşük duyarlık değeriyse Netbul'a aittir (ort. %11). Daha önce de değinildiği gibi, Arama 17 soru için en fazla ilgili belgeye erişen arama

motoru olmuştur. Bu açıdan bakıldığında, Arama'nın ortalama duyarlık değerinin diğerlerinden yüksek olması kolayca açıklanabilir. Arabul'un toplam 17 sorudan 6'sı için sıfır sonuç verdiğini belirtmekte yarar vardır. Ortalama duyarlık değerleri açısından Arama, Superonline'a oranla %40, Arabul'a oranla %87, Netbul'a oranla ise yaklaşık %250 daha iyi bir performans göstermiştir.

Kesme noktası yükseldikçe ortalama duyarlık değerlerinin de düştüğü görülmektedir. Bu düşüş Arama'da çok belirgindir. Arama'da ortalama duyarlık değeri kesme noktası 5 iken %40, 10 iken %27, 15 iken %24, 20 iken %21'dir. Bir başka deyişle, kesme noktası artırıldığında Arama'nın ortalama duyarlık değeri yaklaşık yarı yarıya düşmüştür. Superonline ve Netbul'da ise ortalama duyarlık değerlerindeki düşüş nispeten daha yavaştır (Superonline'da %25'ten %16'ya, Netbul'da %13'ten %9'a). Arabul'da ise çeşitli kesme noktalarında ortalama duyarlık değerleri hemen hemen aynı kalmıştır.

Kesme noktası yükseldikçe, yani erişim çıktılarında aşağı doğru inildikçe, ortalama duyarlık değerlerinin düştüğüne diğer araştırmalarda da rastlanmıştır. Örneğin, Soydal'ın (2000, s. 59) araştırmasında AltaVista, Excite, HotBot, Infoseek ve Northern Light'ta erişilen ilk 10 belge ile ilk 20 belge arasında ortalama duyarlık değerleri %7 ile %12 oranında bir düşüş göstermiştir.

Türkçe arama motorlarında erişilen ilk 20 belge için elde ettiğimiz ortalama duyarlık değerleri yurt dışında yapılan benzeri bir çalışmanın bulgularıyla karşılaştırılabilir. Leighton ve Srivastava 1999 yılında yaptıkları araştırmada erişilen ilk 20 belge için AltaVista, Excite, HotBot ve Infoseek'in ortalama duyarlık değerlerini sırasıyla %53, %58, %35 ve %52 olarak bulmuşlardır. Dört Türkçe arama motorunda 17 soru için erişilen ilk 20 belgede elde ettiğimiz ortalama duyarlık değerleri yabancı arama motorlarında elde edilen değerlerin üçte biri civarındadır (Arabul %15, Arama %21, Netbul %9, Superonline %16). Araştırmamızın amacı her ne kadar Türkçe arama motorlarının performanslarını yabancı arama motorlarınınkilerle karşılaştırmak değilse de, böyle bir karşılaştırmanın ilginç olabileceği kanısındayız.

Arama'da ilk 5 sonuç ile ilk 10 sonuç arasındaki ortalama duyarlık değeri diğer arama motorlarına oranla çok daha hızlı bir düşüş göstermiştir. Bu durum, Arama'da kullanılan erişim algoritmasının ilgili belgeleri ilk sıralarda göstermek üzere programlanmış olabileceğini akla getirmekte ise de ne yazık ki elimizde bu savı destekleyen ampirik veriler bulunmamaktadır.

Arama motorlarının çeşitli kesme noktalarındaki duyarlık değerleri arasında istatistiksel açıdan anlamlı bir fark olup olmadığı çeşitli testlerle sınanmıştır. Tablo 9, 10, 11 ve 12'deki veriler incelendiğinde arama motorlarının duyarlık değerlerinin normal dağılım göstermediği

ortaya çıkmaktadır. Yapılan testlerde dört arama motorunun duyarlık değerlerinin varyanslarının homojen olmadığı görülmüştür. Bu nedenle, arama motorlarının duyarlık değerleri arasında istatistiksel açıdan anlamlı bir fark olup olmadığını ölçmek için parametrik olmayan Kruskal-Wallis testi uygulanmıştır. Fark varsa bu farkın hangi arama motorundan/motorlarından kaynaklandığını bulmak için ise Mann-Whitney *U*- testi uygulanmıştır. Bu testler hakkında daha geniş bilgi bir önceki bölümde (bkz. bölüm 4.7) verilmiştir.

Arama motorlarının kesme noktası 10, 15 ve 20 iken kaydettikleri duyarlık değerleri arasında %95 güven düzeyinde[10] istatistiksel yönden anlamlı bir fark gözlenmemiştir (kesme noktası 10 için $H = 6{,}89$, *s.d.* = 3, $p = .076$; kesme noktası 15 için $H = 5{,}37$, *s.d.* = 3, $p = .147$; kesme noktası 20 için $H = 5{,}22$, *s.d.* = 3, $p = .156$). Üç kesme noktası için hesaplanan Kruskal-Wallis (H) değerleri tablodaki kritik değerden (χ^2 (3, 0,05) = 7,81) küçüktür. Ancak kesme noktası 5 iken arama motorlarının duyarlık değerleri arasındaki fark istatistiksel yönden anlamlıdır ($H = 8{,}77$, *s.d.* = 3, $p = .033$). Bu farkın hangi arama motorlarından kaynaklandığını saptamak için Mann-Whitney *U*- testi uygulanmış ve Arama'nın duyarlık değerinin Arabul'dan ($U(34)=79{,}5$) ve Netbul'dan ($U(34)=75{,}0$) farklı olduğu ortaya çıkmıştır.[11] Bu fark, Şekil 8'de de açıkça görülmektedir. Kesme noktası 5 iken Arama'nın ortalama duyarlık değeri %40, Arabul'un %16, Netbul'un ise %13'tür.

Özetle, ilk 5 belgede Arama, Arabul ve Netbul'dan daha fazla sayıda ilgili belgeye erişmektedir. Arama motorlarının daha yüksek kesme noktalarında eriştikleri ilgili belge sayıları ise birbirine benzemektedir.

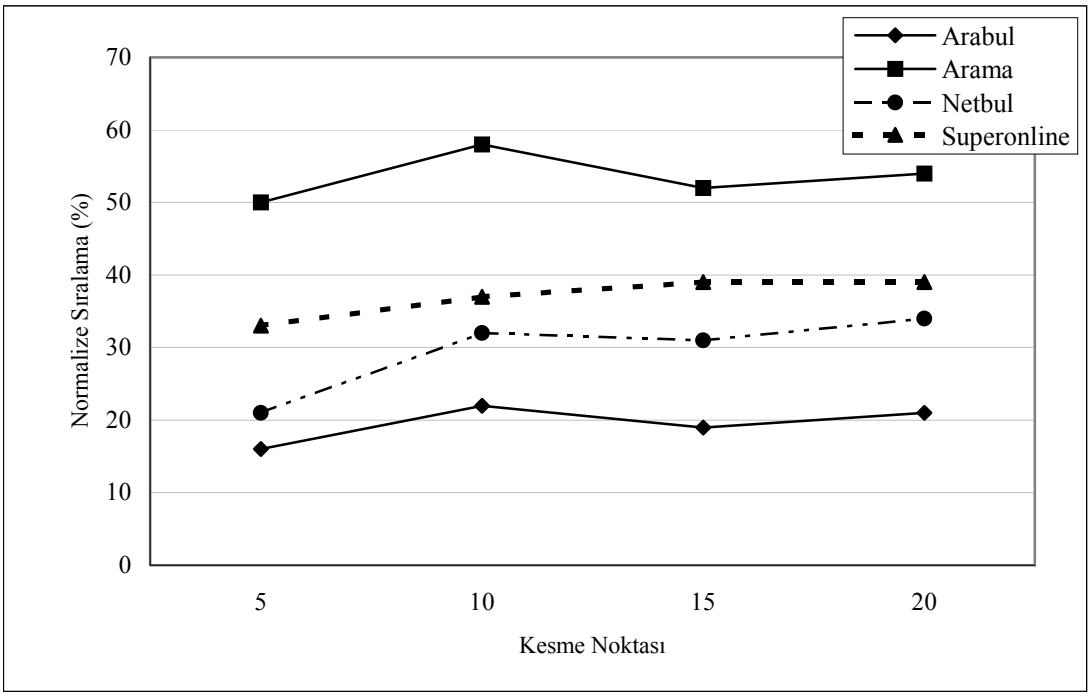
5.2.2.3 Arama Motorlarının Ortalama Normalize Sıralama Değerleri

Arama motorlarının çeşitli kesme noktalarında (erişilen ilk 5, 10, 15, ve 20 belge için) kaydettikleri ortalama normalize sıralama değerleri Şekil 9'da verilmektedir. Daha önce de değinildiği gibi (bkz. 4.6), normalize sıralama değeri arama motorlarının erişilen ilgili belgeleri erişim çıktısının ilk sıralarında gösterip göstermediklerini ölçmektedir. Bu açıdan

[10] Daha önce de belirttiğimiz gibi (bkz. 4.7), araştırmamızda tüm istatistik testler için %95 güven düzeyi (iki yönlü) kullanılmıştır.

[11] $n_1 = 17$, $n_2 = 17$, $\alpha = 0.05$ ve çift yönlü test için *U*- testi kritik değeri 87'dir.

bakıldığında, erişilen ilgili belgeleri sürekli ilk sıralarda gösteren arama motorlarının normalize sıralama değerlerinin de daha yüksek olması beklenir.



Şekil 9. Ortalama normalize sıralama değerleri

Şekil 9'dan da görülebileceği gibi, Arama, çeşitli kesme noktalarında en yüksek normalize sıralama değerini kaydetmiştir (ort. %54). Arama'yı %37 ile Superonline, %30 ile Netbul izlemektedir. En düşük ortalama normalize sıralama değeri ise Arabul'a aittir (%20). Ortalama normalize sıralama değerleri açısından Arama, Superonline'a oranla %46, Netbul'a oranla %80, Arabul'a oranla ise %270 daha iyi bir performans göstermiştir. Hatırlanacağı gibi, sıfır sonuç veren sorular için normalize sıralama değeri sıfır olarak alınmıştır. Bu açıdan bakıldığında, Arabul'un ortalama normalize sıralama değerinin düşük olması bu arama motorunun toplam 17 sorudan 6'sı için sıfır sonuç vermesiyle açıklanabilir.

Çeşitli kesme noktalarında arama motorlarının ortalama normalize sıralama değerlerinde büyük dalgalanmalar gözlenmemektedir. İlginçtir, dört arama motorunda da erişilen ilk 10 belge için ortalama normalize sıralama değerleri ilk 5 belgedekinden daha yüksek çıkmıştır. Örneğin, Netbul'da ortalama normalize sıralama değeri erişilen ilk 5 belge için %21 iken bu oran ilk 10 belgede %32'ye yükselmiştir (%52 artış). Bu oranlar Arabul'da %16'dan %22'ye (%38 artış), Arama'da %50'den %58'e (%16 artış), Superonline'da ise %33'ten %37'ye (%12 artış) yükselmiştir. Ancak kesme noktası artarak önce 15'e, daha sonra da 20'ye çıktığında

Arama'nın ve Arabul'un ortalama normalize sıralama değerleri hafifçe düşerken, Netbul ve Superonline'ınkiler biraz yükselmiştir.

Arama motorlarının çeşitli kesme noktalarındaki normalize sıralama değerleri arasında istatistiksel açıdan anlamlı bir fark olup olmadığını görmek için Kruskal-Wallis testi uygulanmıştır. Kesme noktası 20 iken arama motorlarının normalize sıralama değerleri arasında anlamlı bir fark gözlenmemiştir (H = 7,42, *s.d.* = 3, p = .060).[12] Ancak diğer kesme noktalarında arama motorlarının normalize sıralama değerleri arasında gözlenen farklar istatistiksel yönden anlamlıdır (kesme noktası 5 için H = 7,97, *s.d.* = 3, p = .047; kesme noktası 10 için H = 8,51, *s.d.* = 3, p = .037; kesme noktası 15 için H = 7,81, *s.d.* = 3, p = .050).[13]

Kesme noktası 5, 10 ve 15 iken hangi arama motorlarının normalize sıralama değerlerinin birbirinden farklı olduğunu görmek için Mann-Whitney *U*- testi uygulanmıştır. Kesme noktası 5 iken Arama ile Arabul'un (U(34) = 80,5), kesme noktası 10 iken Arama ile Arabul (U(34) = 72,0) ve Netbul'un (U(34) = 82,0), kesme noktası 15 iken Arama ile Arabul'un (U(34) = 69,5) normalize sıralama değerleri arasındaki farklar istatistiksel açıdan anlamlıdır.[14]

Özet olarak, Arama'nın normalize sıralama değerleri üç kesme noktasında da Arabul'unkilerden daha yüksektir. Arama'nın performansı, kesme noktası 10 iken Netbul'dan da yüksek çıkmıştır. Diğer arama motorlarının normalize sıralama performansları arasındaki farklar istatistiksel yönden önemli değildir.

Bazı arama motorlarının ortalama normalize sıralama değerleri istatistiksel yönden birbirinden farklı olmasına rağmen, bu değerlerin çeşitli kesme noktalarındaki değişimi birbirine benzemektedir. Bir başka deyişle, arama motorları erişilen ilgili belgeleri erişim çıktısının en üst sıralarında gösterme konusunda birbirlerinden pek farklı gözükmemektedir. Çünkü ilgili belgelerin erişim çıktısındaki dağılımlarında arama motorlarına göre farklı bir örüntü (pattern) göze çarpmamaktadır. Hatta, tüm arama motorlarında ortalama normalize sıralama değerlerinin erişilen ilk 5 belge için ilk 10 belgeden daha düşük olması ve erişilen ilk 15 ve ilk 20 belge için normalize sıralama değerlerinin benzer dağılımlar göstermesi, normalize sıralama ölçümünün düşük kesme noktalarında arama motorları arasındaki performans farklarını yeterince güçlü bir biçimde ortaya çıkaramadığını düşündürmektedir.

---

[12] Hesaplanan Kruskal-Wallis (H) değerleri, daha önce duyarlık değerlerinde de yapıldığı gibi, χ^2 (3, 0.05) = 7,81 tablo değeriyle karşılaştırılmıştır.

[13] Yüzde 95 güven düzeyinde gözlenen bu farklar sınır değerlere çok yakındır. Nitekim %99 güven düzeyinde bu farklar ortadan kalkmaktadır (%99 güven düzey için kritik değer χ^2 (3, 0.01) = 11,3 4'tür).

[14] *U*- testi kritik değeri 87'dir (n_1 = 17, n_2 = 17, α = 0.05, çift yönlü).

Araştırmamızda her soru için erişilen ilk 20 belge değerlendirmeye tabi tutulduğundan, kesme noktasının daha yüksek (ilk 100, ilk 200 gibi) tutulduğu durumlarda normalize sıralama değerinin arama motorları arasındaki performans farklarını daha iyi ortaya çıkarıp çıkarmadığı test edilememiştir.

Normalize sıralama değerinin arama motorları arasındaki performans farklarını yeterince güçlü bir biçimde ortaya çıkaramamasının bir başka nedeni, araştırmamızda soru başına düşen ortalama ilgili belge sayısının düşük olması ve bazı sorular için erişilen toplam belge sayısının 20'den daha az olmasıdır. İlgili belge sayısının genelde düşük olması duyarlık değerlerini etkilediği gibi normalize sıralama değerlerini de etkilemektedir. İlgili belgeler erişim çıktısında seyrek dağıldığından normalize sıralama değerleri düşük çıkmakta ve arama motorları arasında belirgin bir fark göze çarpmamaktadır. Öte yandan, herhangi br soru için erişilen toplam belge sayısı düşükse normalize sıralama değeri belirli bir noktadan sonra değişmemektedir. Örneğin, belirli bir soruya karşılık toplam 1 belgeye erişildiyse (örneğin, Arabul, soru 2) ve bu belge de ilgiliyse normalize sıralama değeri bu soru için tüm kesme noktalarında 1 çıkmaktadır. Oysa, örneğin, erişilen 20 belgeden 12'sinin ilgili olduğu bir soru için (Arama, soru 7) kesme noktası 5, 10, 15 ve 20 iken normalize sıralama değerleri sırasıyla %25, %57, %58 ve %65 değerlerini vermektedir. Superonline 3. ve 11. soruları için eşit sayıda (10) ilgili belgeye erişmesine karşın, bu sorular için çeşitli kesme noktalarındaki normalize sıralama değerleri birbirinden oldukça farklı gözükmektedir.

5.2.2.4 Ortalama Duyarlık ve Normalize Sıralama Değerleri Arasındaki İlişki

Tüm kesme noktalarında kaydedilen ortalama duyarlık değerleriyle ortalama normalize sıralama değerleri arasında istatistiksel açıdan anlamlı bir ilişki gözlenmiştir (Pearson's r = .61, $p < .05$). Bir başka deyişle, ortalama duyarlık değerinin yüksek olduğu durumlarda genellikle ortalama normalize sıralama değeri de yüksek olmaktadır. Ancak kesme noktası yükseldikçe ortalama duyarlık değeriyle ortalama normalize sıralama değeri arasındaki ilişkinin giderek zayıfladığı görülmüştür (kesme noktası 5 iken Pearson's r = .97, 10 iken r = .89, 15 iken r = .70, ve 20 iken r = .61'dir). Bu durumun, erişim çıktılarında listelerin sonuna doğru gidildikçe ilgili belge sayısının azalmasından kaynaklandığı söylenebilir. Dolayısıyla ortalama duyarlık ile ortalama normalize sıralama değerleri arasındaki düşük kesme noktalarında güçlü olan ilişki kesme noktası yükseldikçe giderek zayıflamaktadır. Nitekim, ortalama duyarlık değerleri açısından arama motorları Arama (%28), Superonline (%20), Arabul (%15) ve Netbul (%11) biçiminde sıralanmış, ortalama normalize sıralama değerleri

açısından ise bu sıralama Arama (%54), Superonline (%37), Netbul (%30) ve Arabul (%20) şeklini almıştır. Görüldüğü gibi, ortalama duyarlık ile ortalama normalize sıralama değerleri arasındaki ilişkinin güçlü olduğu kesme noktalarında (5 ve 10) arama motorlarının sıralaması değişmezken (Arabul ve Superonline), ilişkinin zayıfladığı kesme noktalarında (15 ve 20) Netbul ve Arabul yer değiştirmiştir.

Öte yandan, tek tek arama motorlarının çeşitli kesme noktalarında kaydettikleri ortalama duyarlık değerleriyle ortalama normalize sıralama değerleri arasında güçlü bir negatif ilişki olduğu gözlenmiştir. Bir başka deyişle, ortalama duyarlık değerleri kesme noktası yükseldikçe tüm arama motorlarında düşerken, ortalama normalize sıralama değerleri tüm arama motorlarında yükselmiştir (Arabul, Arama, Netbul ve Superonline için ilgili *r* katsayısı sırasıyla -.86, -.51, -.79 ve -.94'tür).

5.2.2.5 Arama Motorlarının Sorulara Göre Ortalama Duyarlık ve Normalize Sıralama Değerleri

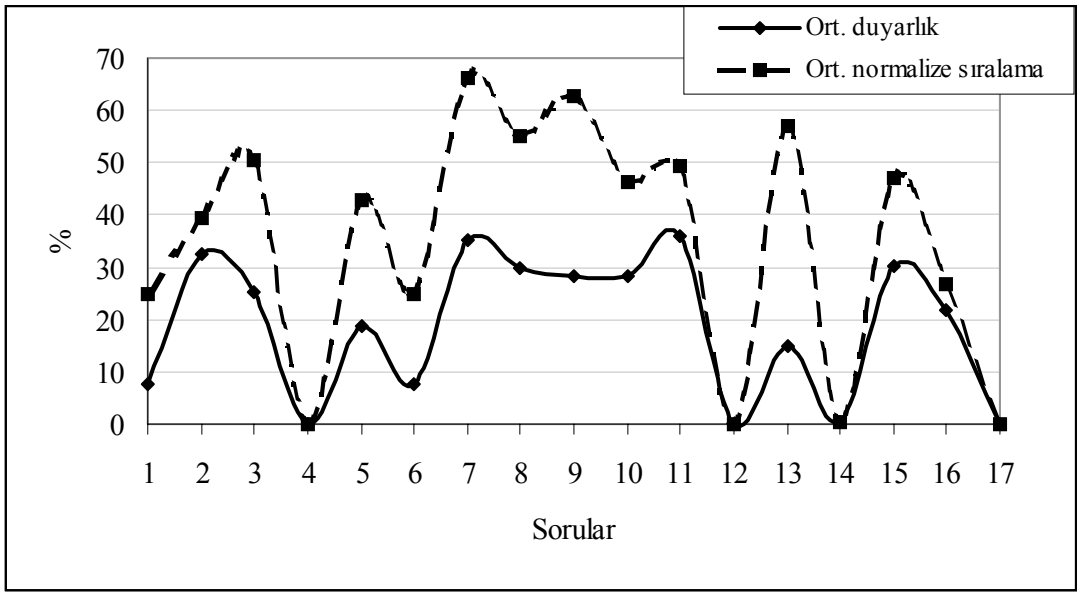
Buraya dek arama motorlarının tüm sorular için çeşitli kesme noktalarındaki duyarlık ve normalize sıralama değerlerini ve bu değerler açısından performanslarının birbirlerinden farklı olup olmadıklarını inceledik. Şimdi de dört arama motorunun sorulara göre gösterdikleri performansları gözden geçireceğiz. Ancak her arama motorunun farklı sorular için gösterdiği performansın diğerlerinden çok farklı olduğu gözlenmiştir. Bu bakımdan, her soru için dört arama motorunun kaydettikleri duyarlık ve normalize sıralama değerlerini gösteren şekillerde belirli bir yönelim göze çarpmamaktadır. Bu nedenle, aşağıda dört arama motorunun her soru için kaydettikleri duyarlık ve normalize sıralama değerleri topluca değerlendirilmektedir.

Tablo 14'te (Tablo 13'te de yer alan arama motorlarının her soru için eriştikleri ilgili belge sayılarına ek olarak) ve Şekil 10'da dört arama motorunun her soru için kaydettiği ortalama duyarlık ve ortalama normalize sıralama değerleri verilmektedir. Ortalama değerler arama motorlarının tüm kesme noktalarında kaydettikleri değerlerin ortalamalarıdır. Söz konusu ortalama duyarlık ve ortalama normalize sıralama değerleri arama motorlarının hangi sorularda nispeten daha başarılı olduklarını ortaya çıkarmaktadır.

Tablo 14. Sorulara göre arama motorlarının ortalama duyarlık ve ortalama normalize sıralama değerleri

						Ortalama	
Soru no.	**Soru**	**Arabul**	**Arama**	**Netbul**	**Superonline**	**Duyarlık (%)**	**Normalize Sıralama (%)**

1	internet ve etik	0	3	0	0	8	25
2	barok müzik	1	2	0	2	33	39
3	Prozac	0	3	4	10	25	51
4	arama motorları	0	0	0	0	0	0
5	baris manco'nun mp3'leri	0	4	3	1	19	43
6	barış manço'nun mp3'leri	0	3	0	0	8	25
7	dpt	2	12	1	1	35	66
8	uzaylı	6	4	3	4	30	55
9	uzaylılar	9	4	2	3	28	63
10	demirel ve sezer	0	5	2	8	28	46
11	demirel veya sezer	0	6	4	10	36	49
12	demirel veya sezer ve tema	0	0	0	0	0	0
13	uzay	5	3	2	2	15	57
14	evren	0	0	0	1	0	0
15	uzay veya evren	2	8	5	6	30	47
16	atatürk ve fikriye hanım	0	7	0	6	22	27
17	ömer izgi	0	0	0	0	0	0
Ort.						**19**	**35**



Şekil 10. Sorulara göre arama motorlarının ortalama duyarlık ve ortalama normalize sıralama değerleri

Arama motorlarının tüm sorular için kaydettikleri ortalama duyarlık değerleri % 0 ile %36 arasında değişmektedir (ortalama duyarlık değerlerinin ortalaması %19, ortancası %22'dir). Daha önce de değinildiği gibi, hiç bir arama motoru 4., 12., ve 17. sorular için ilgili belgeye erişememiştir. Dolayısıyla bu sorular için arama motorlarının ortalama duyarlık ve ortalama normalize sıralama değerleri sıfırdır. Ondördüncü soru için ortalama duyarlık değeri %1'den az, 1. ve 6. sorular için %8, 13. soru için %15, 5. soru için ise %19'dur.

Arama motorlarının ortalama duyarlık değeri açısından en başarılı oldukları sorular 11., 7. ve 2. sorulardır (sırasıyla %36, %35, %33). Sekizinci ve 15. sorular için ortalama duyarlık değeri %30'dur. Onuncu, 9., 3. ve 16. sorular için ortalama duyarlık değeri %28'le %22 arasında değişmektedir.

Arama motorlarının tüm sorular için kaydettikleri ortalama normalize sıralama değerleri % 0 ile %66 arasında değişmektedir (ortalama normalize sıralama değerlerinin ortalaması %35, ortancası %43'tür). Hiç bir arama motorunun ilgili belgeye erişemediği 4., 12. ve 17 soruların normalize sıralama değeri doğal olarak sıfırdır. Ondördüncü soru için ortalama normalize sıralama değeri de %1'in altındadır. Ortalama duyarlık derecesi açısından %8 ile alt sıralarda yer alan 1. ve 6. soruların ortalama normalize sıralama değerleri de düşüktür (%25).

Arama motorlarının ortalama normalize sıralama değeri açısından en başarılı oldukları sorular 7., 9. ve 13. sorulardır (sırasıyla %66, %63, %57). Sekizinci ve 3. sorular için ortalama normalize sıralama değeri %50'nin üstündedir (sırasıyla %55 ve %51). Dört soru

için (11., 15., 10. ve 5. sorular) ortalama normalize sıralama değerleri %40'lardadır. İkinci soru için ortalama normalize sıralama değeri %39, 16. soru için ise %27'dir.

Arama motorlarının hem ortalama duyarlık hem de ortalama normalize sıralama değerleri açısından en başarılı oldukları sorular ise 7., 9., 8., ve 11. sorulardır.

Şekil 10'dan da kolayca görülebileceği gibi, arama motorlarının kaydettikleri ortalama duyarlık ile ortalama normalize sıralama değerleri arasında güçlü bir pozitif ilişki vardır (r = .87, p < .01).

5.2.3 Niteliksel Değerlendirme

Aşağıda arama motorlarının tüm sorular için kaydettikleri ortalama duyarlık ve ortalama normalize sıralama değerleri niteliksel yönden değerlendirilmektedir. Bu alt bölümde yer alan ve araştırma sonuçlarına dayanan kimi çıkarsamalarımız bir miktar önyargı içeriyormuş gibi yorumlanabilir. Ancak her çıkarsamanın kaçınılmaz olarak belli bir oranda taraflılık içerdiği, aksi takdirde çıkarsamanın ilginç olmayacağı bilinen bir gerçektir (Mitchell, 1997). Bu bakımdan aşağıdaki çıkarsamalar nihai yargılar olarak görülmemelidir. Aksine, araştırmamızda elde edilen önemli bulgulara dayanan bu çıkarsamalar yapılacak yeni çalışmalarla test edilmelidir kanısındayız. Aşağıdaki niteliksel değerlendirmeler daha önce Bölüm 4.1 ve 4.3'te ayrıntılı olarak verilen sorulara dayanmaktadır.

Belli bir konuya odaklanan bilgi ihtiyaçları göz önünde bulundurulduğunda arama motorlarının davranışının ne yönde olacağı 1., 2., 4. ve 12. sorular nezdinde incelenmiştir. İlk soruda Internet ile ilgili etik değerler araştırılmıştır. Arama arama motoru dışındaki diğer üç arama motoru bu soruda tamamen başarısız olmuşlardır. Arama ise ilgili sayfaları ilgisiz sayfaların önüne yerleştirmede oldukça başarılı olmuştur. İkinci soruda ansiklopedik bir bilgi ("Barok müzik") araştırılmıştır. Netbul bu bilgi ihtiyacını karşılayamamıştır. Bu soruya karşılık Arama ve Superonline ikişer ilgili belgeye erişmiştir. Bununla birlikte, ilk 10'luk öbek için Superonline'ın normalize sıralama değeri Arama'ya göre daha iyidir (sırasıyla %13 ve %56). Arabul ise yalnızca bir belgeye erişmiş, bu belge de Barok müzikle ilgili bulunmuştur. Bu nedenle, Arabul bu soruda gerek duyarlık ve gerekse de normalize sıralama değerleri açısından mükemmel skoru elde etmiştir. Başka bir deyişle, Arabul ansiklopedik bilgilere erişmede en başarılı arama motoru olmuştur. Dördüncü soruda ise bu araştırmanın konusunu oluşturan Türkçe arama motorlarının değerlendirilmesi ile ilgili çalışmalara erişilmek amaçlanmıştır. Anlaşıldığı kadarıyla, 4. soruda yer alan ve arama motorlarında sık

rastlanan “Internet”, “arama” gibi terimler başarıyı etkilemiştir. Bunu arama motorlarının eriştiği birçok ilgisiz belgeyle açıklamak mümkündür. Dahası, Internet'te Türkçe arama motorlarının değerlendirmesiyle ilgili olarak yapılmış olan bir çalışma (Aslantürk, 2000: http://ata.cs.hun.edu.tr/~aslantur/Akademik) en az altı aydır Web’de yer almasına karşılık arama motorlarınca dizinlenmemiş olabilir. Kısacası, spesifik ve akademik esaslı bir bilgi ihtiyacına ulaşmada tüm arama motorları başarısız olmuştur. Onikinci soruda Cumhurbaşkanlarımızdan Süleyman Demirel veya Ahmet Necdet Sezer’in TEMA (Türkiye Erozyonla Mücadele ve Ağaçlandırma ve Doğal Varlıkları Koruma) Vakfı ile igili düşünceleri hakkında belgelere erişmek amaçlanmıştır. Bu konuyla ilgili çok sayıda Internet kaynağı elde edeceğimiz beklentisi maalesef gerçekleşmemiştir.[15] Bu soruya tatmin edici düzeyde yanıt vermekte söz konusu dört arama motoru tamamen başarız olmuştur. İlginçtir, bu soruyla ilgili olarak yabancı bir arama motoru (google.yahoo.com) çok daha başarılı olmuştur. Google'ın kapsamının ve/veya erişim algoritmasının performansının bu başarıda ne derece etkili olduğu ilginç bir araştırma konusu oluşturmaktadır. Öte yandan, web’de Demirel, Sezer ve TEMA ile ilgili belgeler olmasına karşın (bkz. 10. ve 11. sorular), bu soruda geçen kavramlar çok yaygın olduğundan ve aramada Boole işleçlerinin kullanımı gerektiğinden, arama motorlarının erişim algoritmaları ilgili belgeleri ilk sıralarda gösterememiş olabilir.

Birinci ve 6. sorular için de arama motorlarının ortalama duyarlık değerleri oldukça düşüktür (%8). Bu sorular için Arama dışında ilgili belgelere erişen arama motoru olmamıştır. Yaygın olarak web’de yer alan “internet” ve “etik” kavramları ilk sorunun duyarlık değerini düşürmüş olabilir. Ancak Barış Manço’nun şarkılarına ait mp3’lerle ilgili üç belgeye sadece Arama tarafından erişilmiştir.

Altıncı soruyla ilgili ilginç bir durum göze çarpmaktadır. Beşinci soru da Barış Manço’nun şarkılarına ait mp3’lerle ilgilidir. Ancak 5. soruda sanatçının adında geçen Türkçe harfler (“ı”, “ş” ve “ç”) kullanılmadan arama yapılmıştır. Bu soru için Arabul dışındaki diğer

---

[15] On ikinci soru ile ilgili Internet kaynaklarına ulaşmak için, 16.02.02 tarihli aşağıdaki iki sorgu Google’a sunulmuştur: (1) http://google.yahoo.com/bin/query?p=tema+demirel&hc=0&hs=0 veya (2) http://google.yahoo.com/bin/query?p=tema+sezer&hc=0&hs=0. Dönenler arasından rastgele seçilen ilgili kaynaklar aşağıda verilmiştir: (a) TEMA tarafından Cumhurbaşkanımız Ahmet Necdet Sezer’e gönderilen 16.02.2001 tarihli açık mektup, http://www.tema.org.tr/tema/kampanya/tema_haber/end_bol_kanuntasari_cumh.html , (b) TEMA Vakfı başkanı Hayrettin Karaca’nın “Türkiye Hızla Açlığa, Yoksulluğa ve Çölleşmeye Doğru Gidiyor” adlı tespit makalesi -ki içinde Cumhurbaşkanımız Süleyman Demirel’in TEMA’ya katkıları konu edilmektedir, http://www.elegans.com.tr/44/html/karaca.html (c) TEMA ve Cumhurbaşkanı Süleyman Demirel’in konu alındığı yazı, http://www.tema.org.tr/english/mission/public.html (d) Saint-Joseph Okulu 8B sınıfı öğrencisi Levent Gürel’in TEMA’nın okul gezileri çerçevesinde düzenlenen ve Süleyman Demirel’in onurlandırdığı etkinlikle ilgili anısı, http://www.sj.k12.tr/html/kardelen/05/tema.html, ve (e) Süleyman Demirel’in 1994 yılında TEMA tarafından düzenlenen Erozyonla Mücadele Haftası açılış konuşmasını konu alan haber, http://www.byegm.gov.tr/yayinlarimiz/TURKHABER/94/T20.htm.

arama motorları Arama 4, Netbul 3, Superonline 1 olmak üzere toplam 8 ilgili belgeye erişmiştir. Anlaşıldığı kadarıyla, Netbul ve Superonline, Türkçe karakter kullanılmadan yapılan aramalarda ilgili belgelere erişmiştir. İşin ilginç yanı, Arama, Türkçe harf kullanılarak ve kullanılmadan yapılan bu iki aramada farklı sayıda ilgili belgeye erişmiştir.

Dört arama motorunun 5. ve 6. sorular için eriştikleri ilgili ve ilgisiz belgelerin sayıları Tablo 15'te verilmektedir. Tabloda dikkati çeken bir nokta, Türkçe harfler kullanılarak ve kullanılmadan yapılan aramalarda arama motorları farklı sayıda belgeye erişmektedirler. Arama, Türkçe karakter kullanılmadan yapılan aramada daha fazla (10), Arabul ise daha az (3) belgeye erişmiştir. Netbul ve Superonline, Türkçe harf kullanılmadan yapılan aramalarda sırasıyla 3 ve 1 ilgili belgeye erişmişler, Türkçe harf kullanıldığında ise hiç bir ilgili belgeye erişememişlerdir. Bir başka deyişle, bu aramalarda farklı belgelere erişilmektedir.

Tablo 15. Arama motorlarında Türkçe karakter kullanımı

Arama Motoru	**Sorgu**	**Erişilen İlgili Belge Sayısı**	**Erişilen Toplam Belge Sayısı**
Arabul	"baris manco" ve mp3	0	3
	"barış manço" ve mp3	0	2
Arama	"baris manco" ve mp3	4	10
	"barış manço" ve mp3	3	20
Netbul	"baris manco" ve mp3	3	20
	"barış manço" ve mp3	0	20
Superonline	"baris manco" ve mp3	1	20
	"barış manço" ve mp3	0	20
Toplam		**11**	**115**

Not: Değerlendirmede erişilen ilk 20 belge dikkate alınmıştır.

Türkçe arama motorlarında Türkçe karakter sorununun henüz çözülemediği anlaşılmaktadır. Aramaların önemli bir kısmının Türkçe karakterler kullanılarak yapılacağı göz önüne alınacak olursa, sorunun ivediliği daha belirgin olarak ortaya çıkmaktadır. Yakın zamana kadar Web'de Türkçe karakterler yaygın olarak kullanılmadığından, arama motorları Web sayfalarını her iki şekilde de dizinlemiş olabilirler. Ancak Türkçe karakterler farklı arama motorlarında farklı kurallara göre işlem gördüğünden ve yapılan aramalarda farklı belgelere erişildiğinden kullanıcılar şaşırmaktadır. Türkçe karakter sorunu kullanıcıya yansıtılmadan çözülebilir. Bunun için oluşturulacak bir dönüştürüm tablosu ve yakınsamalı (genel olarak yaklaşık) arama algoritması (Badino, 2001) yararlı olabilir. Arama motorları bu

zamana dek işlem kütükleri aracılığıyla topladıkları istatistiklere dayanarak bu dönüştürüm tablosuna temel olacak kuralları belirleyebilirler.

Yedinci soruda, "DPT nedir?" diye sorduğumuzda, beklentimiz DPT ev sayfasının ilk sıraya yerleştirilmesi idi. Bu konuda, Arabul ve Arama başarılı olurken Netbul ve Superonline başarısız kalmışlardır. Özellikle, Netbul ve Superonline dikkat çekecek ölçüde birçok ilgisiz belgeye erişmiştir.

Sekizinci soruda "uzaylı" hakkında genel bir bilgi edinilmek istenmiş ve kullanıcının konuyu özellikle genel tutup, cevaplardan yola çıkarak bilgi ihtiyacını daraltmak (refine) isteme olasılığı göz önünde bulundurulmuştur. Benzer amaçlı bir diğer soru da dokuzuncu sorudur ("uzaylılar"). Buradaki asıl amaç "uzay" (13. soru), "uzaylı" ve "uzaylılar" sorgularının sonuçlarından elde edilen bilgilere dayanarak arama motorlarının gövdeleme (stemming) yapıp yapmadığını belirlemektir. Dizinlemede ve sorgu işlemede gövdelemeye başvurulduğunda, "uzaylı" ve "uzaylılar" sorgularının gövdeleri aynı olduğundan özdeş erişim çıktıları döndürmesi gerekmektedir -ki hiç bir arama motoru bunu başaramamıştır. Sonuç olarak; Türkçe tabanlı arama motorları gövdeleme yapmamaktadır. Diğer yandan, gövdelemeye başvurmayan arama motorlarının alt dizgi (substring) aramaya başvurup başvurmadıklarını denetlemek için, 13. sorgunun erişim çıktısınının 8. veya 9. sorgularının erişim çıktılarını içerip içermediğine bakmak yeterli olacaktır. Ancak bu nokta erişim çıktılarında erişilen ilk 20 belgenin değerlendirilmesi nedeniyle test edilememiştir. İlginçtir, aynı isimden ("uzay") türetilen bu üç soruda, soru spesifikleştikçe dört arama motoru tarafından erişilen toplam ilgili belge sayısı artmıştır. Ortalama duyarlık 13. soruda %15, 9. soruda %28, 8. soruda ise %30'dur. Bu durum, arama motorlarının gövdeleme algoritmalarından yararlanmadıklarının diğer bir göstergesidir. Dört arama motorunun içinde "uzay" terimi geçen sorularda nispeten daha başarılı oldukları gözlemlenmiştir. Aslına bakılırsa, örnek sorularda gövdeleme algoritması kullanılmamasının kullanıcının lehine işlediği sonucuna varılabilir. Ancak "uzay" teriminden daha az yaygın olan terimler için aynı şeyi söylemek mümkün olmayabilir. Çünkü erişim çıktılarında konuyla doğrudan ilgisi olmayan birçok belge yer almıştır (adında "uzay" geçen ya da sahibinin adı "uzay" olan işyerlerinin siteleri vs.).

Ondördüncü soru için ise sadece Superonline ilgili bir belgeye erişmiştir. Bu soru için arama motorlarının ortalama duyarlık ve ortalama normalize sıralama değerleri %1'in altında olduğundan Tablo 14 ve Şekil 10'da farkedilememektedir. Bu soruya karşılık dört arama motoru da birçok belgeye erişmiştir. Ancak "evren" teriminin çok çeşitli bağlamlarda

kullanılması (firma ismi, 7. Cumhurbaşkanı, vs) arama motorlarının başarımını etkilemiş gözükmektedir.

Onyedinci soru için arama motorları tarafından hiç bir belgeye erişilememesi ilginçtir. Şimdiki Meclis Başkanı Ömer İzgi hakkında web'de belge bulunamaması arama motorlarının kişilerle ilgili güncel bilgileri yeterince dizinlemediklerini düşündürmektedir.

Dört arama motorunun da en az bir ilgili belgeye eriştiği 7. ("dpt"), 8. ("uzaylı"), 9. ("uzaylılar"), 13. ("uzay") ve 15. ("uzay" veya "evren") sorular aynı zamanda ortalama duyarlık ve normalize sıralama değerleri açısından arama motorlarının en başarılı olduğu sorulardır. Bu soruların ortak yönlerinden birisi soruların ya tek sözcükten oluşması ya da soruların "VEYA" Boole işleci içermesidir.[16] Tek sözcükten oluşan 3. soruda ("prozac") da arama motorları oldukça başarılıdır (ortalama duyarlık %25, ortalama normalize sıralama %51). Bu soruda, aynı adlı "rock" müzik grubuyla ilgili belgeler (Arabul dışında) arama motorları tarafından başarıyla ayıklanmış ve adı geçen ilaçla ("prozac") ilgili belgelere erişilmiştir. Arama motorlarının başarısız oldukları tek sözcükten oluşan tek soru 14. sorudur ("evren"). Daha önce de değindiğimiz gibi, arama teriminin hangi bağlamda kullanıldığının belirtilmemiş olması bu sorudaki başarıyı etkilemiştir. Bu sorudaki dille ilgili belirsizlik (lingusitic ambiguity) giderilmiş olsaydı, arama motorları bu soruda da tıpkı "prozac" sorusunda olduğu gibi başarılı olabilirlerdi. Bu tür sorularda başarılı olabilmek için kullanıcıların tıklama bilgilerinden yararlanılarak bir terimin daha çok hangi bağlamda arandığı saptanabilir. Kullanıcıya daha fazla yardımcı olmak için çevrimiçi kavramsal listelerden (gömü) yararlanılabilir.

Arama motorlarının Boole işleçlerinin ("VE", "VEYA" ve "DEĞİL") kullanıldığı sorularda sergiledikleri başarımı topluca değerlendirmekte yarar vardır. "VEYA" işlecinin kullanıldığı sorularda arama motorlarının genelde başarılı olduklarına yukarıda değinmiştik. Nitekim, "VEYA" işlecinin kullanıldığı 11. soruda ("demirel veya sezer") Arabul dışındaki diğer üç arama motoru %36 ile en yüksek ortalama duyarlık değerine ulaşmışlardır.

Yukarıda anılan 3. soru da, aslına bakılırsa, "DEĞİL" işlecinin kullanıldığı bir sorudur. Bütün arama motorlarında "prozac" DEĞİL "rock" biçiminde aranan bu soruda da Arabul dışındaki diğer üç arama motoru başarılı olmuştur (ortalama duyarlık %25, ortalama normalize sıralama %51).

---

[16] Arama motorlarının farklı yönlerini test etmek için sorulan bu sorulardan 4'ünde "uzay" sözcüğünün geçmiş olması tamamen rastlantıdır. Arama motorlarının "uzay" konusuna özel ilgi duyarak konuyla ilgili belgeleri daha özenli dizinlemiş olmaları çok düşük bir olasılıktır.

“VE” işlecinin kullanıldığı 5. ve 6. sorularla ilgili geniş bir değerlendirme (Türkçe karakter kullanımı nedeniyle) daha önce yapılmıştı. Bu iki soru için arama motorlarının başarımı nispeten daha düşük olmuştur. Onuncu soruda (“demirel ve sezer”) Arabul dışındaki üç arama motoru oldukça başarılıdır (ortalama duyarlık %28, ortalama normalize sıralama %46). “Atatürk ve Fikriye Hanım” sorusunda (16. soru) ise Arama ve Superonline başarılı olmuş, diğer iki arama motoru ilgili belgelere erişememiştir (ortalama duyarlık %22, ortalama normalize sıralama %27). Bu durum, arama motorlarının tarihi araştırmalar için de kullanılabileceğini, ancak tarihle ilgili daha fazla belge dizinlenmesi gerektiğini göstermektedir. Arama motorlarının en düşük başarım gösterdikleri “VE” işleci içeren soru 1. sorudur (“internet ve etik”). Bu soru için sadece Arama üç ilgili belgeye eriştiğinden, ortalama duyarlık ve normalize sıralama değerleri düşüktür (sırasıyla %8 ve %25). Hem “VEYA” hem de “VE” işleci içeren 12. soruda (“demirel veya sezer ve tema”) ise arama motorları hiç bir ilgili belgeye erişemediğinden, ortalama değerler sıfır olarak gerçekleşmiştir.

Araştırmamızda arama motorlarında “VE” ve “VEYA” işleçlerini tutarlı bir biçimde kullanılıp kullanılmadığı test edilmiştir. Onuncu (“demirel ve sezer”) ve 11. sorularda (“demirel veya sezer”) Arama, Netbul ve Superonline’ın erişim sonuçlarının tutarlı olduğu görülmektedir. Beklendiği gibi, 11. soruda erişilen ilgili belge sayısı (ve dolayısıyla ortalama değerler) her üç arama motorunda da 10. soruda erişilen belge sayısından daha yüksektir. (Arabul bu iki soru için hiç bir belgeye erişememiştir.)

Onüçüncü (“uzay”) ve 15. (“uzay veya evren”) sorular da arama motorlarının tutarlılığını ölçmek için kullanılmıştır. “VEYA” işleci tutarlı çalıştığı takdirde 15. soru için erişilen ilgili belge sayısının daha yüksek olması ve 13. soru için erişilen ilgili belgelerin bir alt kümesi olması beklenir. Arama, Netbul ve Superonline’ın erişim çıktıları bu beklentiyi doğrulamıştır. Arabul’da ise ilgili belge sayısında düşme olmuştur.

Bu sonuçlara bakarak, arama motorlarının Boole işleci içeren soruları genelde tutarlı bir biçimde yorumladıkları söylenebilir.

Onüçüncü, 14. ve 15. sorular aynı zamanda nispeten genel konularla ilgili kapsamlı konu aramalarına örnek olarak seçilmiştir. Ondördüncü (“evren”) soruyla ilgili hususa daha önce değinmiştik. Bu tür genel ve tek terimden oluşan sorularla ilgili olarak yapılan aramalarda genellikle yanlış düşmelere (false drops) ve ilgisiz belgelerin ilgili belgelerden önce listelenmesine sık rastlanır. Nitekim aynı şey bizim araştırmamızda da görülmüş, her arama motoru “uzay” veya “evren” ile ilgili en az 20 belgeye erişmiş, ancak bunlardan birçoğu ilgisiz çıkmıştır. Anlaşıldığı kadarıyla, 15. soruda “uzay” ve “evren” terimlerinin aynı soruda geçmesi bu soru için duyarlık değerini tek sözcükten oluşan 13. ve 14. sorulara oranla

artırmıştır. Dilde belirsizlik hususuna 14. soru bağlamında yukarıda değinmiştik. Aynı şeyler 13. ve 15. sorular için de geçerlidir. Genel olarak arama motorlarının başarımının kapsamlı sorular için daha da iyileştirilebileceği söylenebilir.

5.3 Kapsama ve Yenilik Oranları

Arama motorlarının performansları kapsama ve yenilik oranları açısından da birbirleriyle karşılaştırılmıştır. Daha önce de değindiğimiz gibi (bkz. 4.6), kapsama oranı arama motorlarının daha önceden ilgili olduğunu bildiğimiz belgelere erişme açısından başarılarını, yenilik oranı ise ilgili olduğunu bilmediğimiz belgelere erişme açısından başarılarını ölçmek için kullanılmaktadır. Kapsama ve yenilik oranları Türkçe arama motorlarında en sık aranan beş sözcük (“mp3”, "oyun", “sex”, “erotik” ve “porno”) kullanılarak Türkiye adresli ve Türkiye adresli olmayan belgelere göre ayrı ayrı hesaplanmıştır.

Kapsama ve yenilik oranlarını hesaplamak için arama motorları tarafından erişilen ilk 1000 belgenin kullanılması planlanmıştır. Ancak belirlenen sorgular arama motorlarında çalıştırıldığında, Netbul’un eriştiği belgelerden sadece ilk 240’ını, Arama’nın ise sadece ilk 300’ünü listelediği görülmüştür. Bir başka deyişle, söz konusu iki arama motorunun kapsama ve yenilik oranlarını hesaplamak için kullandığımız tekil ilgili belge “havuz”una katkıları sırasıyla en fazla 240 ve 300 belgeyle sınırlıdır. (Arabul ve Superonline’da ise ilk 1000 belge alınmıştır.) Ayrıca Arama’da bazı sorular için gösterilen sayfalarda yer alması gereken belgelerin listelenmediği görülmüştür. Arama’nın boş sayfalarda listelenmesi gereken belgeleri toplam içinde gösterdiği saptanmıştır.

Netbul, “internet” seçeneği üzerinde arama yaptığı zaman “sex”, “erotik”, ve “porno” sorularına karşılık herhangi bir belgeye erişememiştir. Ancak Netbul’daki arama “internet rehberinde” yapıldığında ilgili belgelere erişilmiştir. Bu durumun diğer arama motorlarına karşı haksızlık olacağı düşünülerek “rehber” ya da “dizin” üzerinde yapılan aramalarda erişilen ilgili belgeler hesaplamalarda dikkate alınmamıştır. Arabul’un bazı sorular için listelemiş olduğu kendi kategorileri de aynı nedenle değerlendirmeye katılmamıştır.

Toplam beş soru için (“mp3”, "oyun", “sex”, “erotik” ve “porno”) tekil ilgili belge havuzunda toplanan belge sayısı 9944’tür. Bu belgelerin sorulara ve arama motorlarına göre dağılımı ile her arama motorunun havuza katkısı ve her sorunun havuzdaki payı Tablo 16'da verilmektedir. Havuzda toplanan belgelerin yaklaşık yarısı Superonline'a aittir. Arabul'un havuza katkısı %31,5, Arama'nın %18,6, Netbul'un ise yaklaşık %5'tir. Netbul ve Arama'nın

havuza katkıları yukarıda açıklanan nedenlerden dolayı sınırlı kalmıştır. Superonline'ın havuza katkıdaki ağırlığı nedeniyle sorulara göre havuzda toplanan belgelerin dağılımı birbirine yakındır ("mp3" ve "oyun" %17, "sex" %23, "erotik" %22 ve "porno" %21).

Tablo 16. Kapsama ve yenilik oranlarını hesaplamak için kullanılan "havuz" değerleri

Sorgu	Arabul			Arama			Netbul			Superonline			Her sorunun havuza toplam katkısı (%)
	TBS	EBS	**%**	**TBS**	EBS	**%**	**TBS**	EBS	**%**	**TBS**	EBS	**%**	
"mp3"	193	193	1,9	240	2950	2,4	240	240	2,4	1000	15,422,517	10,1	**16,8**
"oyun"	175	175	1,8	300	12360	3,0	240	240	2,4	1000	35,239	10,1	**17,2**
"sex"	1000	1431	10,1	285	2269	2,9	0	0	0,0	1000	18,181,033	10,1	**23,0**
"erotik"	886	886	8,9	300	783	3,0	0	0	0,0	1000	1,731,010	10,1	**22,0**
"porno"	877	877	8,8	208	223	2,1	0	0	0,0	1000	2,140,263	10,1	**21,0**
Toplam belge sayısı / Her arama motorunun havuza katkısı (%)	**3131**	3562	**31,5**	**1333**	18585	**13,4**	**480**	480	**4,8**	**5000**	37,510,062	**50,3**	**100,0**

TBS: Toplanan Belge Sayısı; **EBS**: Erişilen Sayısı

Not: Arama motorlarının listelediği tüm belgeler havuzda toplanmıştır. Her arama motoruna ait ilk sütun her soru için o arama motorunun havuza katkısını, ikinci sütun ise bu katkının havuzda toplanan tüm belgelere oranını vermektedir.

Sonu ".tr" ile bitmeyen adresler havuzdan ayıklandığında geriye toplam 1417 belge kalmıştır.[17] Bu belgelerin sorulara ve arama motorlarına göre dağılımı ile her arama motorunun havuza katkısı ve her sorunun havuzdaki payı Tablo 17'de verilmektedir.

Tablo 17. Kapsama ve yenilik oranlarını hesaplamak için kullanılan "havuz" değerleri (sadece alan adı ".tr" ile biten belgeler)

Sorgu	Arabul	%	Arama	%	Netbul	%	Superonline	%	Her sorunun havuza toplam katkısı (%)
"mp3"	0	0,0	218	15	0	0,0	0	0,0	15,4
"oyun"	21	1,5	300	21	22	1,6	116	8,2	32,4
"sex"	6	0,4	258	18	0	0,0	0	0,0	18,6
"erotik"	7	0,5	287	20	0	0,0	0	0,0	20,7
"porno"	4	0,3	178	13	0	0,0	0	0,0	12,8
Toplam belge sayısı / Arama motorunun havuza toplam katkısı (%)	**38**	**2,7**	**1241**	**87,6**	**22**	**1,6**	**116**	**8,2**	**100,0**

Not: Arama motorlarının listelediği belgelerden sadece alan adı ".tr" ile biten tüm belgeler havuzda toplanmıştır. Her arama motoruna ait ilk sütun her soru için o arama motorunun havuza katkısını, ikinci sütun ise bu katkının havuzda toplanan tüm belgelere oranını vermektedir.

[17] Bu belgelere http://cmpe.emu.edu.tr/bitirim/home adresinden çevrimiçi olarak erişilebilir.

Havuzda toplanan ve alan adı ".tr" ile biten belgelerin büyük bir çoğunluğuna (%88) Arama'nın katkısıyla erişilmiştir. Türkçe adresli belgelerin bulunduğu havuza katkı bakımından Arama'yı %8,2 ile Superonline, %2,7 ile Arabul, %1,6 ile Netbul izlemektedir. Superonline'ın ve Arabul'un katkıları belgeler alan adına göre ayıklandıktan sonra önemli derecede düşmüştür (Superonline %50'den %8'e, Arabul %31'den yaklaşık %3'e). Buna karşılık, Arama'nın bütün belgelerde %13 olan havuza katkısı, alan adı ".tr" ile biten belgelerde %88 olarak gerçekleşmiştir.

Arama'nın havuza katkısı büyük ölçüde Türkiye adresli belgelerden oluşmaktadır. Arama'nın havuza giren toplam 1333 belgesinden 1241'i nin (%93) alan adı ".tr" ile bitmektedir. Diğer üç arama motoru için bu oranlar görmezden gelinebilir düzeylerdedir. Superonline'ın havuza katkıda bulunduğu toplam 5000 belgeden sadece 116'sı (%2), Arabul'un 3131 belgeden sadece 38'i (%1), Netbul'un ise 480 belgeden sadece 22'si (yaklaşık %5) Türkiye adreslidir. Başka bir deyişle, Türkçe arama motorlarında en sık aranan beş soru için dört arama motoru tarafından erişilen her 7 belgeden sadece birisi Türkiye adreslidir (1418/9944). Türkçe adresli bu 7 belgeden 6'sına Arama tarafından erişilmektedir (1241/1418). Bu terimler için Superonline, Arabul ve Netbul'un büyük ölçüde yabancı adresli belgelerden yararlandıkları ortaya çıkmaktadır.

Havuzda toplanan tüm belgelerin en sık aranan beş terime göre dağılımının birbirine yakın olduğuna, oranların %17 ("mp3" ve "oyun") ile %23 ("sex") arasında değiştiğine yukarıda değinmiştik. Bu terimlerin Türkiye adresli belgeler açısından dağılımında bazı küçük farklılıklar göze çarpmaktadır. Örneğin, "oyun" terimiyle ilgili belgelerin oranı Türkiye adresli belgelerde daha yüksektir (%32). Buna karşılık, "porno" teriminin havuzdaki tüm belgeler arasındaki %21'lik payı Türkiye adresli belgelerde %13'e düşmüştür.

Aşağıda arama motorlarının kapsama ve yenilik oranlarıyla ilgili bulgular özetlenmektedir.

5.3.1 Kapsama Oranları

5.3.1.1 Arama Motorlarının Tüm Belgeleri Kapsama Oranları

Sorulara göre dört arama motorunun tüm belgeler için kapsama oranları ile arama motorlarının tüm öbeklerde kaydettikleri (makro ortalama yöntemine göre hesaplanmış) ortalama kapsama oranları Tablo 18'de verilmektedir. Bu oranlar havuzda toplanan (Türkiye

adresli ve Türkiye adresli olmayan) tüm ilgili belgelere dayanılarak hazırlanmıştır. Tablo 18'de dikkati çeken önemli noktalar aşağıda özetlenmektedir.

Tablo 18. Arama motorlarının kapsama oranları (Genel)

		Belge Öbek Sayısına Göre Kapsama Oranları (%)																				
Sorgu	**Arama Motoru**	50	100	150	200	250	300	350	400	450	500	550	600	650	700	750	800	850	900	950	1000	**Ort.**
	Arabul	26	26	27	26	23	22	21	20	19	18	17	16	16	15	15	14	14	13	13	12	19
mp3	**Arama**	26	26	26	27	28	27	25	24	23	22	21	20	20	19	18	17	17	16	16	15	22
	Netbul	26	26	27	27	28	27	25	24	23	22	21	20	20	19	18	17	17	16	16	15	22
	Superonline	26	26	27	27	29	34	37	41	44	46	49	51	53	55	57	59	60	62	63	64	45
	Arabul	27	27	27	24	20	18	17	16	16	15	14	14	13	13	12	12	12	11	11	10	16
oyun	**Arama**	27	27	27	28	29	31	30	29	27	26	25	24	23	22	22	21	20	20	19	18	25
	Netbul	27	27	26	27	27	25	23	22	21	20	20	19	18	17	17	16	16	15	15	14	21
	Superonline	27	27	27	28	29	31	35	38	41	44	46	48	50	52	54	56	57	59	60	61	44
	Arabul	33	33	33	33	33	34	35	37	38	39	39	40	41	41	42	42	43	43	43	44	38
sex	**Arama**	33	33	33	33	33	32	29	26	24	22	21	19	18	17	16	15	14	14	13	13	23
	Netbul	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Superonline	33	33	33	33	33	34	36	37	38	39	40	41	41	42	42	43	43	43	44	44	39
	Arabul	33	33	33	33	33	33	35	36	37	38	39	40	41	41	42	42	42	42	41	40	38
erotik	**Arama**	33	33	33	33	33	33	30	27	25	23	21	20	19	18	17	16	15	14	14	14	24
	Netbul	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Superonline	33	33	33	33	33	33	35	36	38	39	39	40	41	41	42	42	43	43	45	46	38
	Arabul	33	33	33	33	35	37	38	40	40	41	42	42	43	43	44	44	44	44	43	42	40
porno	**Arama**	33	33	33	33	29	26	23	20	19	17	16	15	14	13	12	11	11	10	10	10	19
	Netbul	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Superonline	34	34	34	34	36	37	39	40	41	42	42	43	44	44	44	45	45	46	47	48	41

Not: Havuzdaki tüm ilgili belgeler dikkate alınmıştır. Bazı sorular için (örneğin, "sex" ve "erotik") bazı öbeklerde dört arama motorunun toplam kapsama oranı yuvarlama hatasından dolayı %100'ün altındadır.

En sık aranan 5 sorudan 4'ü için en yüksek ortalama kapsama oranı Superonline'a aittir. Bir soru ("erotik") için ise Superonline ve Arabul'un en yüksek ortalama kapsama oranları birbirine eşittir (%38). En sık aranan sorular için erişilen her 5 ilgili belgeden yaklaşık 2'sine Superonline tarafından erişilmiştir. Superonline'ı bir soruda ("erotik") Superonline en yüksek, iki soruda ("sex" ve "porno") ikinci en yüksek ortalama kapsama oranları ile Arabul izlemiştir. Arama ise dört soru için ikinci en yüksek ortalama kapsama oranlarını kaydeden arama motoru olmuştur. Netbul ise bir soru ("mp3") dışında tüm sorularda en düşük ortalama kapsama oranına erişmiştir.

Bazı sorular için (örneğin, "mp3" ve "oyun") dört arama motorunun çeşitli öbeklerdeki kapsama oranlarının toplamının %100'ü aştığı dikkat çekmiş olabilir. Aynı durum, ortalama kapsama oranlarının toplamı için de söz konusudur (bkz. Tablo 18, son sütun). Aslına bakılırsa, bir soru için dört arama motorunun kapsama oranlarının toplamı en az 100 olmalıdır. Ama bu toplam 100 ile sınırlı değildir. Örneğin, arama motorlarından her birinin

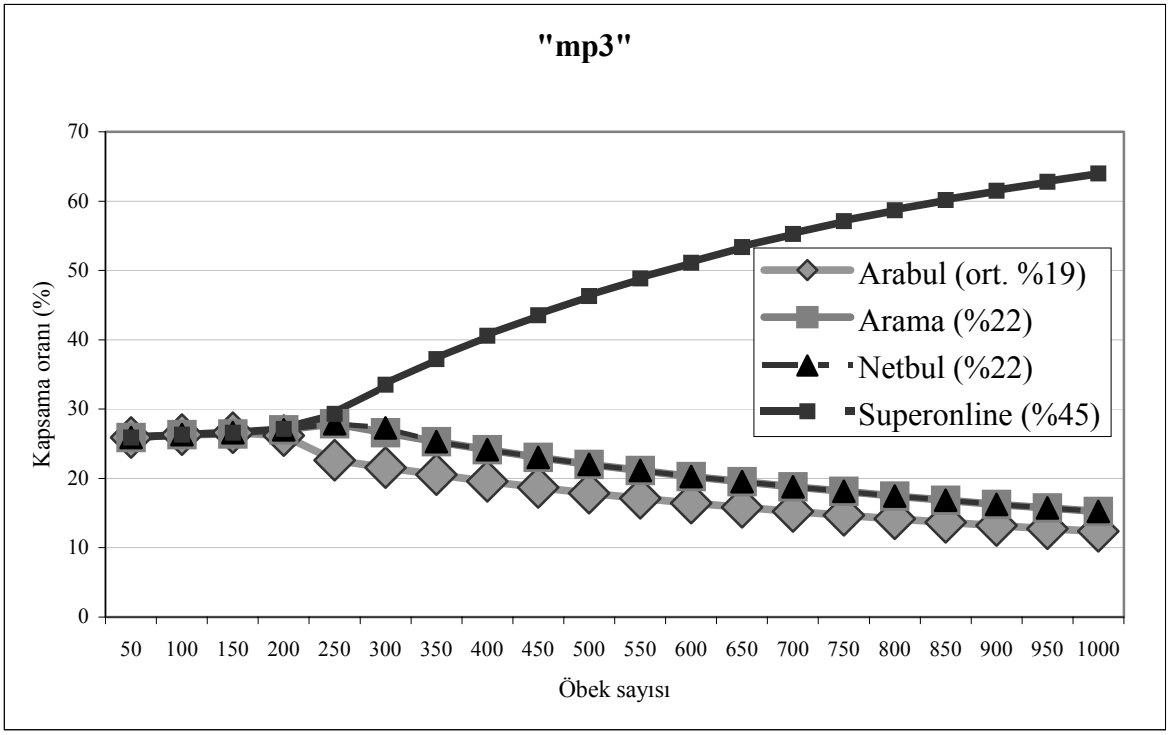
25’er farklı ilgili belgeye eriştiğini varsayalım. Bu durumda dört arama motoru tarafından erişilen toplam tekil ilgili belge sayısı 100 olur ve arama motorlarının bu soru için kapsama oranları da eşit şekilde (her biri %25) dağılır. Bunun tam tersini düşünelim: Dört arama motoru da aynı 25 tekil ilgili belgeye erişmiş olsun. Bu durumda dört arama motoru tarafından erişilen toplam tekil ilgili belge sayısı 25 olur ve her arama motoru ilgili belgelerin tamamını (%100) kapsamış olur. Böyle bir soru için dört arama motorunun ortalama kapsama oranlarının toplamı 400 olur. Başka bir deyişle, arama motorları ne kadar çok ortak belgeye erişirlerse kapsama oranı da o kadar çok %100’ü aşacaktır. Oranların toplamını %100’den çıkararak arama motorlarının eriştikleri toplam ilgili belgeler arasındaki ortak belgelerin yüzdesini bulabiliriz. Örneğin, “mp3” sorusu için bu oran %8’dir ((19 + 22 + 22 + 45) – 100)). Yani, “mp3” sorusu için dört arama motoru tarafından erişilen toplam ilgili belgelerin %8’ine birden fazla arama motoru tarafından erişilmiştir. Bu oranlar “oyun” için %6, diğer üç soru (“sex”, “erotik” ve “porno”) için %0’dır. Son üç soru için her arama motoru genelde birbirinden farklı ilgili belgelere erişmiştir.

Tablo 18’de özellikle ilk öbeklerde arama motorlarının kapsama oranlarının her soru için hemen hemen birbirine eşit olduğu gözükmektedir. İlk iki soru için her arama motorunun kapsama oranı %26-%27 civarındadır. Son üç soruda ise Netbul herhangi bir belgeye erişemediğinden geri kalan üç arama motorunun kapsama oranı %33’tür. Başka bir deyişle, arama motorlarının kapsama oranları ilk öbeklerde birbirinden farklı değildir.

Bir soru için dört arama motorunun kapsama oranları toplamının %100’ü geçip geçmemesi bir başka açıdan da yorumlanabilir. Toplamın %100’e eşit veya biraz üzerinde olması her arama motorunun farklı ilgili belgelere eriştiğini, yani her arama motorunun yenilik oranının yüksek (%100’e eşit veya yakın) olduğunu gösterir. Nitekim, yenilik oranları ile ilgili bulgular bu ilişkiyi doğrulamaktadır (bkz. 5.3.2).

Superonline ve Arabul’un havuza en fazla belge sağlayan (sırasıyla 5000 ve 3131) arama motorları olarak en yüksek kapsama oranlarına erişmeleri olağan gözükmektedir. Öte yandan, Arama ve Netbul’un ortalama kapsama oranlarının nispeten düşük olmasının nedeni bu arama motorlarının havuza katkıda bulundukları belge sayılarının (sırasıyla 300 ve 240) sınırlı olmasıdır. Dahası, Netbul, "sex", "erotik" ve "porno" soruları için "internet" üzerinde arama seçeneğinde hiç bir belgeye erişemediğinden bu sorular için ortalama kapsama oranı sıfırdır.

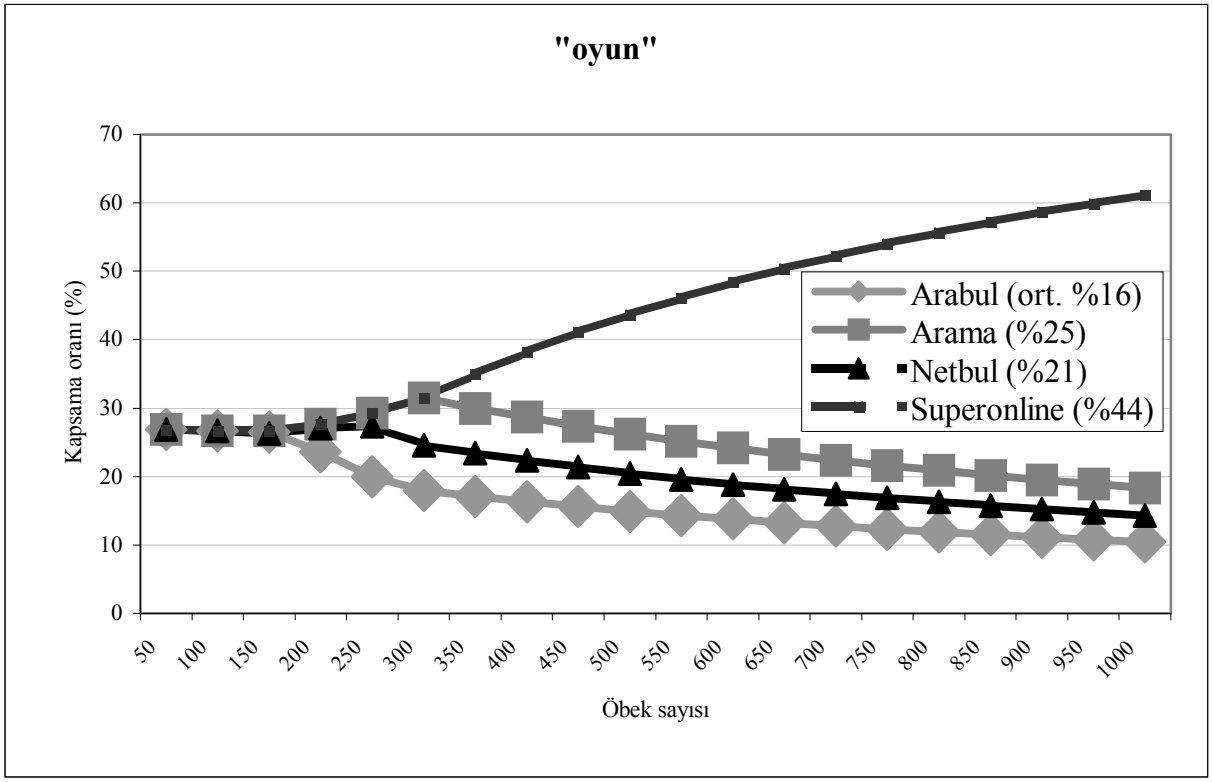
Her arama motorunun havuza katkıda bulunduğu belge sayısının, o arama motorunun öbekteki belge sayısına göre kapsama oranını da etkilediği görülmektedir. Şekil 11’de dört arama motorunun “mp3” sorusu için kapsama oranlarını göstermektedir.



Şekil 11. Arama motorlarının "mp3" için öbekteki belge sayısına göre kapsama oranları

Şekil 11'den de görülebileceği gibi, "mp3" sorusu için dört arama motorunun kapsama oranları ilk 200 belgede hemen hemen birbirine eşittir. Ancak daha sonra Superonline'ın kapsama oranının giderek yükseldiği, diğer üç arama motorunun kapsama oranlarının ise giderek düştüğü görülmektedir. Bunun temel nedeni, daha önce de vurguladığımız gibi, arama motorlarının havuza katkıda bulunduğu belge sayılarıyla ilgilidir. Arama ve Netbul'un havuza katkıda bulundukları belge sayılarına (sırasıyla 300 ve 240) daha önce değinmiştik. Nitekim, ilgili öbek sayısına ulaşıldığında bu iki arama motorunun kapsama oranlarının giderek düşmeye başladığı Şekil 11'de açıkça görülmektedir. Arabul'da böyle bir sınır olmamasına rağmen, Arabul'un "mp3" sorusu için havuza katkıda bulunduğu belge sayısı 193'tür. Bu bakımdan Arabul'un kapsama oranı da öbek sayısı 200'den itibaren düşmeye başlamıştır. Havuza en fazla belgeyle katkıda bulunan Superonline'ın kapsama oranı ise öbek sayısı 200'den itibaren yükselmeye başlamıştır. Çünkü Arabul, Netbul ve Arama'nın kapsama oranları belirli bir soru için havuza katkıda bulundukları belge sayısına (örneğin, Arabul) ya da tüm sorular için belirlenen sınıra eşit olduğunda (Netbul ve Arama) söz konusu arama motorlarının kapsama oranlarının bu noktalardan itibaren yükselmesi mümkün değildir. Nitekim, benzer bir duruma "oyun" sorusunda da rastlanmıştır (bkz. Şekil 12).

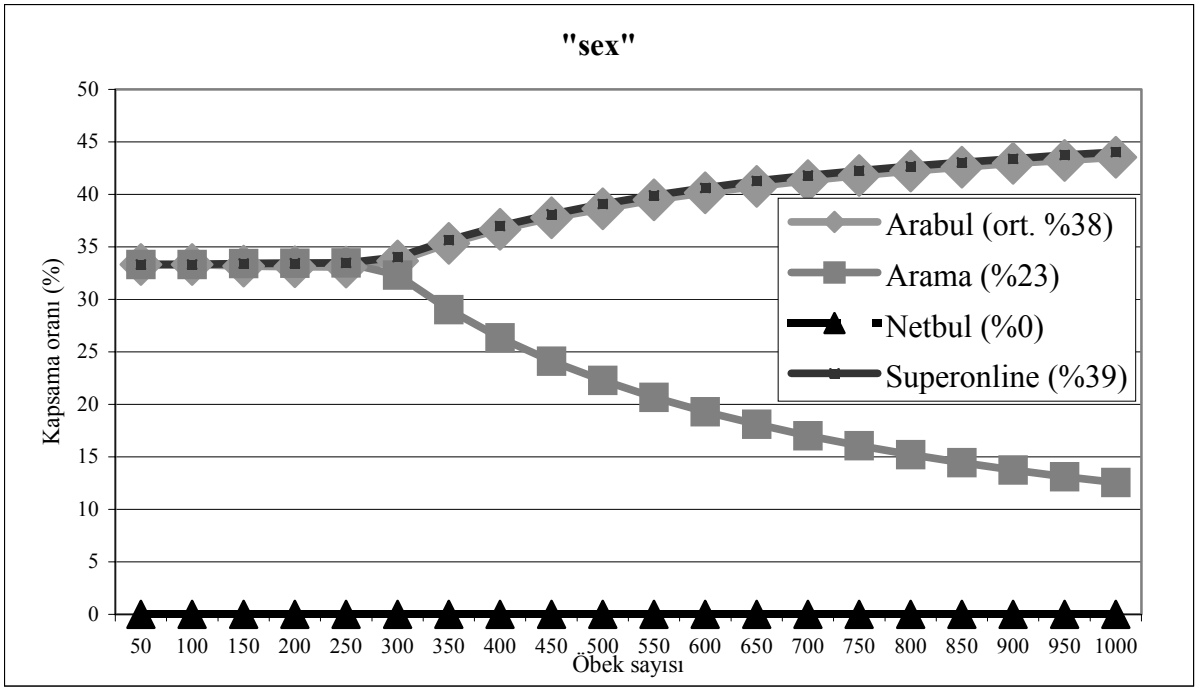
Yukarıda özetlediğimiz hususlar arama motorlarının "oyun" sorusundaki kapsama oranları için de geçerlidir.



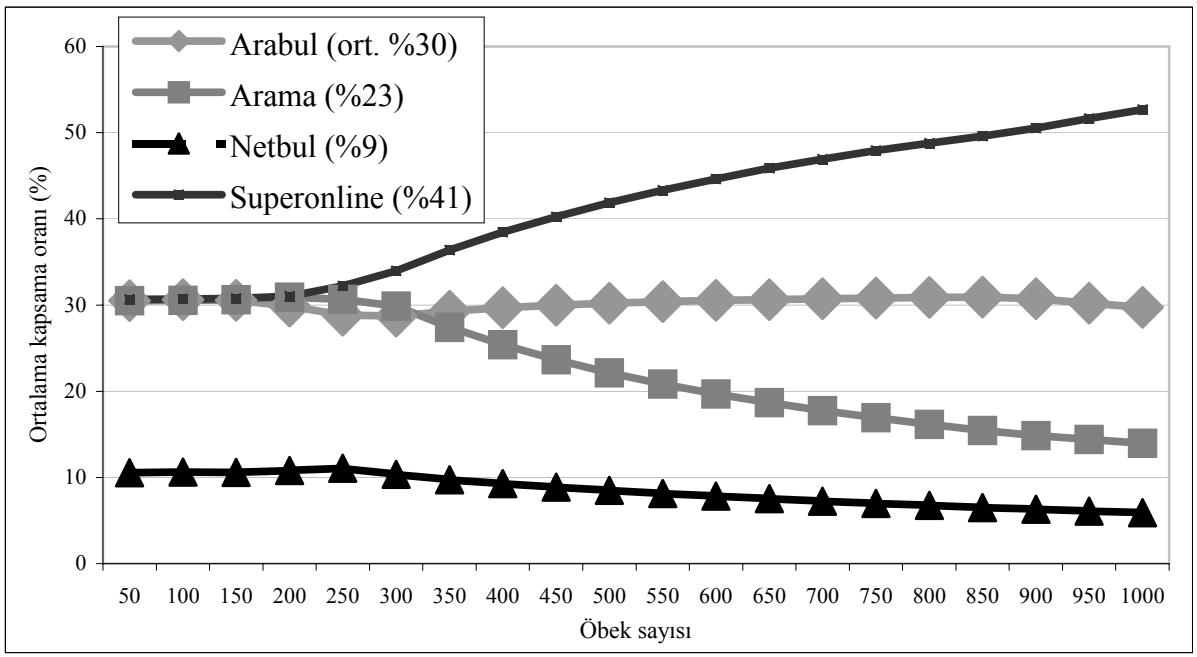
Şekil 12. Arama motorlarının "oyun" için öbekteki belge sayısına göre kapsama oranları

Diğer üç soruda ("sex", "erotik" ve "porno") arama motorlarının kapsama oranları birbirine çok benzemektedir. Bu üç soru için Netbul'un kapsama oranı sıfırdır. Arama'nın havuza katkıda bulunduğu belge sayısı ise 208 ile 300 arasında değişmektedir. Arabul ise aynı sorular için havuza 877 ile 1000 belgeyle katkıda bulunmuştur. Şekil 13'de "sex" sorusu için arama motorlarının kapsama oranları verilmektedir.[18] Dört arama motorunun en sık aranan beş soru için ortalama kapsama oranları ise Şekil 14'te verilmektedir.

[18] Diğer iki soru ("erotik" ve "porno") için de arama motorları benzer kapsama oranlarına sahip olduğundan, bu sorularla ilgili şekillere yer verilmemiştir.



Şekil 13. Arama motorlarının "sex" için öbekteki belge sayısına göre kapsama oranları



Şekil 14. Arama motorlarının en sık aranan beş soru için ortalama kapsama oranları

5.3.1.2 Arama Motorlarının Türkiye Adresli Belgeleri Kapsama Oranları

Sorulara göre dört arama motorunun Türkiye adresli belgeler[19] için kapsama oranları ile her soru için arama motorlarının kaydettikleri (makro ortalama yöntemine göre hesaplanmış) ortalama kapsama oranları Tablo 19'da verilmektedir. Bu oranlar havuzda toplanan sadece Türkiye adresli ilgili belgelere dayanılarak hazırlanmıştır. Tablo 19'da dikkati çeken önemli noktalar aşağıda özetlenmektedir.

Tablo 19. Arama motorlarının Türkiye adresli belgeleri kapsama oranları

		Belge Öbek Sayısına Göre Kapsama Oranları (%)																				
Sorgu	**Arama Motoru**	50	100	150	200	250	300	350	400	450	500	550	600	650	700	750	800	850	900	950	1000	**Ort.**
	Arabul	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
mp3	**Arama**	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	Netbul	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Superonline	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Arabul	16	9	7	6	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	6
oyun	**Arama**	38	43	51	58	63	67	67	67	67	67	67	67	67	67	67	67	67	67	67	67	63
	Netbul	16	9	7	6	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	6
	Superonline	38	43	39	33	29	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	29
	Arabul	11	6	4	3	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3
sex	**Arama**	89	94	96	97	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	97
	Netbul	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Superonline	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Arabul	12	7	4	3	3	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3
erotik	**Arama**	88	93	96	97	97	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	97
	Netbul	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Superonline	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Arabul	8	4	3	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3
porno	**Arama**	92	96	97	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	97
	Netbul	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Superonline	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Not: Sadece alan adı ".tr" ile biten ilgili belgeler dikkate alınmıştır.

Bir önceki kesimde (5.3.1.1) arama motorlarının ortalama kapsama oranları toplamının %100'ü aşmasının erişilen ortak ilgili belgelerin bir göstergesi olduğunu vurgulamıştık. Arama motorlarının beş soru için Türkiye adresli tekil belgelere erişme oranlarının toplamı sadece ikinci soru ("oyun") için %100'ü aşmaktadır (%104). Başka bir deyişle, bu soru için tüm arama motorları tarafından erişilen tekil belgelerin %4'lük bir bölümüne birden fazla arama motoru tarafından erişilmiştir. Diğer sorularda ise ortalama kapsama oranlarının

[19] "Türkiye adresli belgeler" URL adresi ".tr" uzantısıyla biten adreslerdeki belgelerdir. Bu belgelerin içeriğinin çoğunlukla Türkçe olduğu varsayılabilir. Ancak Türkçe arama motorları tarafından bulunan ve sonu ".tr" ile bitmeyen adreslerdeki belgelerin önemli bir kısmının içeriğinin de Türkçe olduğu görülmektedir. Arama motorları Türkçe içerikli belgeleri saptamada HTML'deki "Language" üst veri belirtecinden (meta tag) yararlanmaktadır. Türkçe arama motorlarının Türkçe belgeleri bulma ve kapsama oranları başka bir çalışmamızda ele alınacaktır.

toplamı %100'dür. Yani, birden fazla arama motoru tarafından erişilen tekil belgelerin oranı yüzde sıfır civarındadır. Nitekim, ilk soru için tüm ilgili belgelere bir arama motoru (Arama) tarafından erişilmiştir. Üçüncü, 4., ve 5. sorular için ise sadece iki arama motorunun (Arama ve Arabul) tekil ilgili belgelere eriştikleri görülmektedir. Bu sorular için farklı arama motorlarının eriştikleri ortak belgelerin çok az olduğunu söylemek mümkündür.

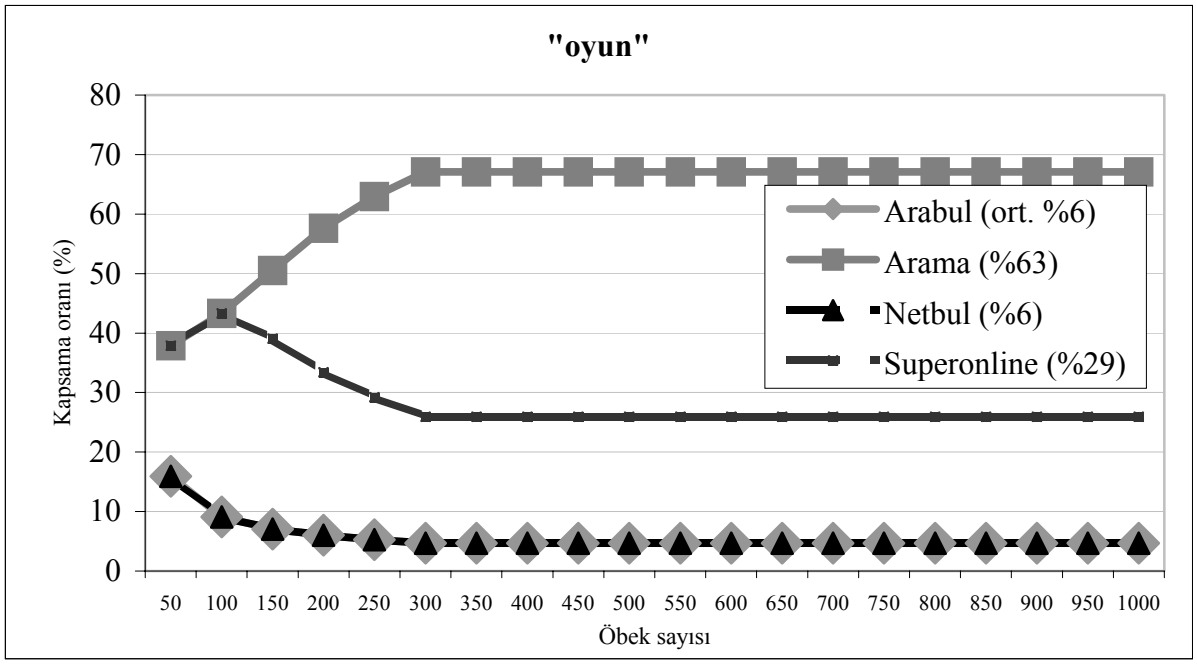
Arama, Türkçe arama motorlarında en sık aranan beş soru için ilgili belgelerin büyük bir çoğunluğunu kapsamaktadır. İlk soru ("mp3") için Arama'nın ortalama kapsama oranı %100'dür. Diğer üç arama motorunun bu soruya karşılık Türkiye adresli belgeleri dizinlemedikleri anlaşılmaktadır. "Sex" ve "erotik" soruları için de Arama'nın ortalama kapsama oranı %97'dir. Bu sorularla ilgili geri kalan (%3) Türkiye adresli belgelerin Arabul tarafından dizinlendiği görülmektedir. "Oyun" sorusu için Arama'nın ortalama kapsama oranı %63, Superonline'ın %29, Arabul ve Netbul'un ise %6'dır.

Öbekteki belge sayısı yükseldikçe Arama'nın kapsama oranı da başlangıçta yükselmiş, diğer arama motorlarınınkiler ise düşmüştür. Örneğin, "sex", "erotik" ve "porno" soruları için Arama'nın kapsama oranı ilk 50 belgede sırasıyla %89, %88 ve %92 iken, ilk 250 belgede bu oranlar %98'e yükselmiş bu oranlar 1000 belgeye dek devam etmiştir. Bunun iki nedeni vardır: İlki, yukarıda da değinildiği gibi, bu sorularla ilgili Türkiye adresli 7 belgeden 6'sı Arama tarafından kapsanmıştır. Arama dışındaki diğer arama motorları en sık aranan beş soru için ya hiç bir belgeye erişememiş, ya da eriştikleri belge sayıları çok düşük kalmıştır. Örneğin, Superonline "oyun" sorusu dışındaki diğer sorular için Türkiye adresli hiç bir belgeye erişememiştir. Superonline sadece "sex" sorusunda hatırı sayılır oranda (ort. %29) Türkiye adresli ilgili belgelere erişmiştir. Netbul, "sex", "erotik" ve "porno" soruları için "internet" üzerinde arama seçeneğinde hiç bir belgeye erişemediğinden bu değerlendirmede dezavantajlı durumdadır. Arabul ise en sık aranan 5 sorudan 4'ü için az sayıda da olsa bazı Türkiye adresli belgelere erişmiştir (4 soru için toplam 38 belge).

İkinci neden ise bu sorular için diğer arama motorlarının kapsadığı ilgili belge ya hiç olmadığından ya da sayısı düşük olduğundan, ilk 250-300 öbekte maksimum düzeye ulaşılmış ve kapsama oranları daha sonra değişmemiştir. Aynı şey Arama için de geçerlidir. Çünkü, daha önce de değinildiği gibi, Arama'nın havuza katkısı maksimum 300 belgeyle sınırlıdır. Dolayısıyla, öbek sayısı 300'e çıktığında Arama'nın ilgili belgeleri kapsama oranı bu noktadan itibaren düşmeye başlamaktadır.

Tablo 19'dan da kolayca görülebileceği gibi, öbek sayısına göre arama motorlarının en sık aranan sorular için Türkiye adresli belgeleri kapsama oranları sorulara göre pek değişiklik göstermemektedir. Arama, bu sorular için hemen hemen bütün Türkiye adresli belgeleri

kapsadığından, diğer arama motorlarının kapsama oranları sıfır civarında seyretmektedir. Bunun tek istisnası "oyun" sorusu için Superonline'ın kapsama oranının %8 olmasıdır (bkz. Şekil 15).



Şekil 15. Arama motorlarının "oyun" için öbekteki belge sayısına göre Türkiye adresli belgeleri kapsama oranları

5.3.2 Yenilik Oranları

5.3.2.1 Arama Motorlarının Tüm Belgeler İçin Yenilik Oranları

Sorulara göre dört arama motorunun en sık aranan beş soru için yenilik oranları ile arama motorlarının tüm öbeklerde kaydettikleri (makro ortalama yöntemine göre hesaplanmış) ortalama yenilik oranları Tablo 20'de verilmektedir. Bu oranlar havuzda toplanan (Türkiye adresli ve Türkiye adresli olmayan) tüm ilgili belgelere dayanılarak hazırlanmıştır. Tablo 20'de dikkati çeken önemli noktalar aşağıda özetlenmektedir.

Tablo 20. Arama motorlarının yenilik oranları (Genel)

		Öbek sayısına göre yenilik oranları (%)																					
Sorgu	**Arama Motoru**	50	100	150	200	250	300	350	400	450	500	550	600	650	700	750	800	850	900	950	1000	**Ort.**
mp3	**Arabul**	100	100	99	96	77	64	55	48	43	38	35	32	30	27	26	24	23	21	20	19	49
	Arama	100	100	99	100	95	79	68	60	53	48	43	40	37	34	32	30	28	26	25	24	56
	Netbul	86	80	77	73	69	55	46	39	34	30	27	24	21	20	18	17	16	15	14	13	39
	Superonline	86	80	77	73	74	75	77	79	81	82	83	84	84	85	86	87	88	88	89	89	82
oyun	**Arabul**	90	89	90	79	62	52	45	39	34	31	28	26	24	22	20	19	18	17	16	15	41
	Arama	100	99	99	100	100	100	85	75	66	60	54	50	46	43	40	37	35	33	31	30	64
	Netbul	78	81	81	81	77	64	54	46	41	37	33	31	28	26	24	22	21	20	19	18	44
	Superonline	80	85	85	85	85	87	88	89	90	91	91	92	92	93	93	93	94	94	94	94	90
sex	**Arabul**	100	100	99	99	99	99	99	99	99	99	99	99	98	99	99	99	99	99	99	99	99
	Arama	100	100	100	100	100	95	81	71	63	57	52	47	44	40	38	35	33	31	30	28	62
	Netbul	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Superonline	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
erotik	**Arabul**	100	100	100	100	100	100	99	99	99	99	99	100	100	100	100	100	100	98	93	88	99
	Arama	100	100	100	100	100	100	85	75	66	60	54	50	46	43	40	37	35	33	31	30	64
	Netbul	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Superonline	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
porno	**Arabul**	98	99	98	99	99	99	99	99	99	99	99	99	98	98	98	98	98	96	91	86	97
	Arama	98	98	99	99	82	68	59	51	46	41	37	34	31	29	27	25	24	22	21	20	51
	Netbul	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Superonline	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

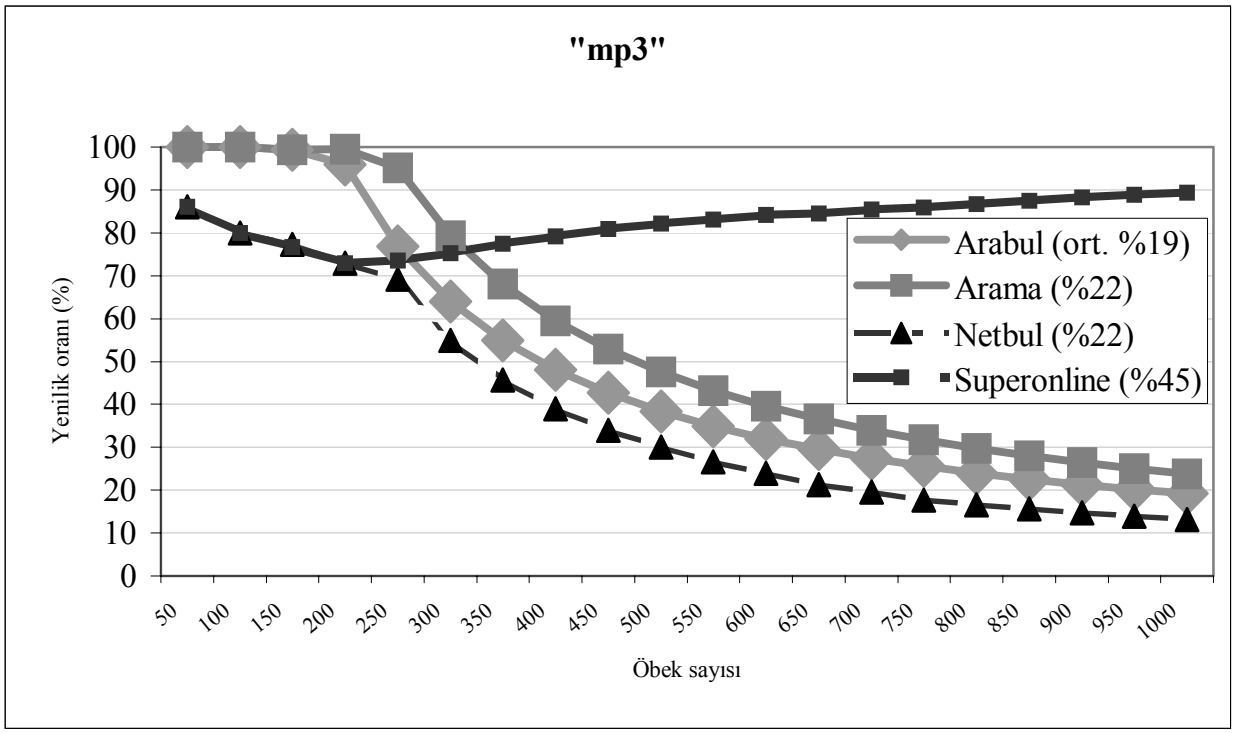
Not: Havuzdaki tüm ilgili belgeler dikkate alınmıştır.

Superonline tüm sorularda en yüksek ortalama yenilik oranına sahiptir. Arabul üç ("sex", "erotik" ve "porno"), Arama iki ("mp3" ve "oyun") soruda ikinci en yüksek ortalama yenilik oranlarına ulaşmıştır. Netbul ise üç soru ("sex", "erotik" ve "porno") için herhangi bir belgeye erişemediğinden ortalama yenilik oranı en düşük olan arama motorudur.

Kapsama oranlarında olduğu gibi, yenilik oranlarının da arama motorlarının havuza katkıda bulundukları toplam belge sayısıyla orantılı olduğu görülmektedir. Tüm sorular için havuza maksimum (1000) belgeyle katkıda bulunan Superonline, ortalama yenilik oranı açısından da birinci sıradadır. Diğer arama motorlarının havuza katkıda bulundukları belge sayıları genelde daha düşüktür (bkz. Tablo 16). Bu nedenle, diğer arama motorlarının yenilik oranları havuza katkıda bulundukları belge sayısına erişene kadar ölçülebilmekte, daha sonraki öbeklerde ise bu arama motorlarının erişebilecekleri "yeni" belge olmadığından, yenilik oranları da doğal olarak giderek düşmektedir. İlk soru ("mp3") için tüm öbeklerde yenilik oranlarını veren Şekil 16'da bu durum açıkça görülmektedir. Çünkü Arabul, Arama ve Netbul bu soru için ilk 250 belgede erişebildikleri tüm yeni belgelere erişmişlerdir. Ancak

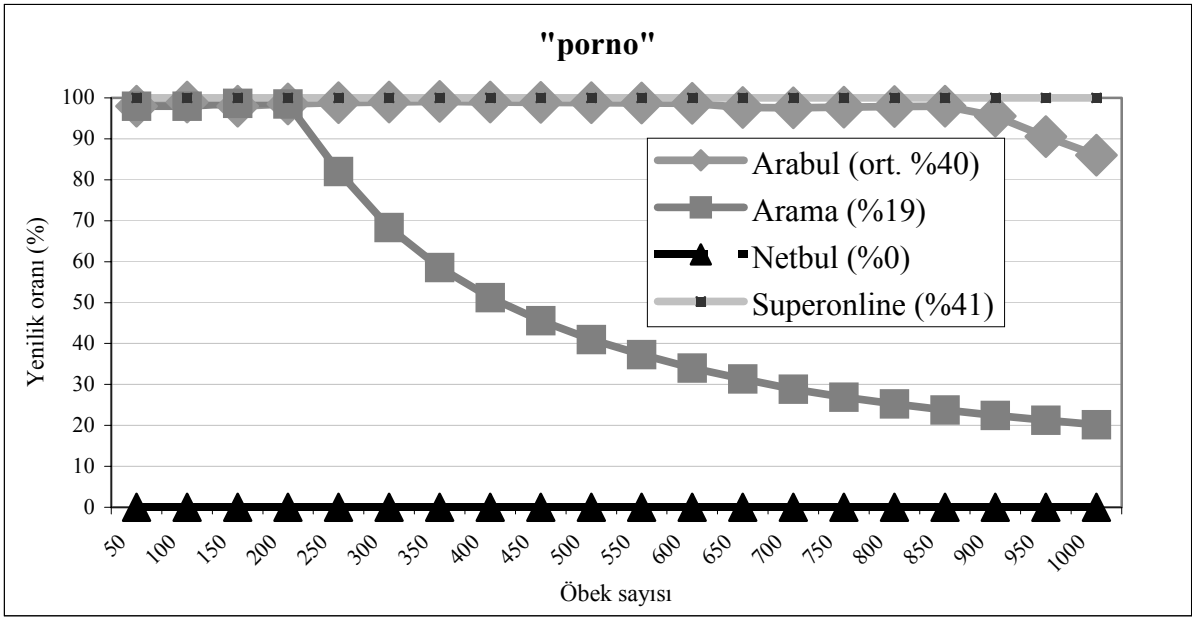
aşağı yukarı bu noktada[20] üç arama motoru havuza katkıda bulundukları toplam belge sayısına eriştiklerinden, yenilik oranları diğer öbeklerde giderek düşmüştür. Superonline ise bu noktadan sonra da "yeni" belgeler bulmaya devam etmiştir. Arama motorlarının yenilik oranlarında ikinci soru ("oyun") için de benzer bir yönelim (trend) izlenmektedir.

Şekil 16'da dikkati çeken bir başka nokta, Arabul ve Arama'nın havuza katkıda bulundukları maksimum sınıra gelene dek eriştikleri belgelerin hemen hemen tamamının "yeni" olmasıdır. Superonline ve Netbul'da ise bu oranlar daha düşük (%80-%85) gerçekleşmiştir. Arama'nın, özellikle "mp3" ve "oyun" soruları için eriştiği tüm belgelerin yeni olduğu gözlenmektedir. Arabul, havuza katkıda bulunduğu belge sayısının nispeten daha yüksek olduğu üç soruda ("sex", "erotik" ve "porno") sürekli yüksek yenilik oranlarına ulaşmıştır ("porno" sorusu için bkz. Şekil 17).



Şekil 16. Arama motorlarının "mp3" sorusu için yenilik oranları

[20] Aslında, havuza katkıda bulunduğu belge sayısı 193 olduğundan, Arabul'da bu düşüş daha erken başlamıştır (bkz. Şekil 16).

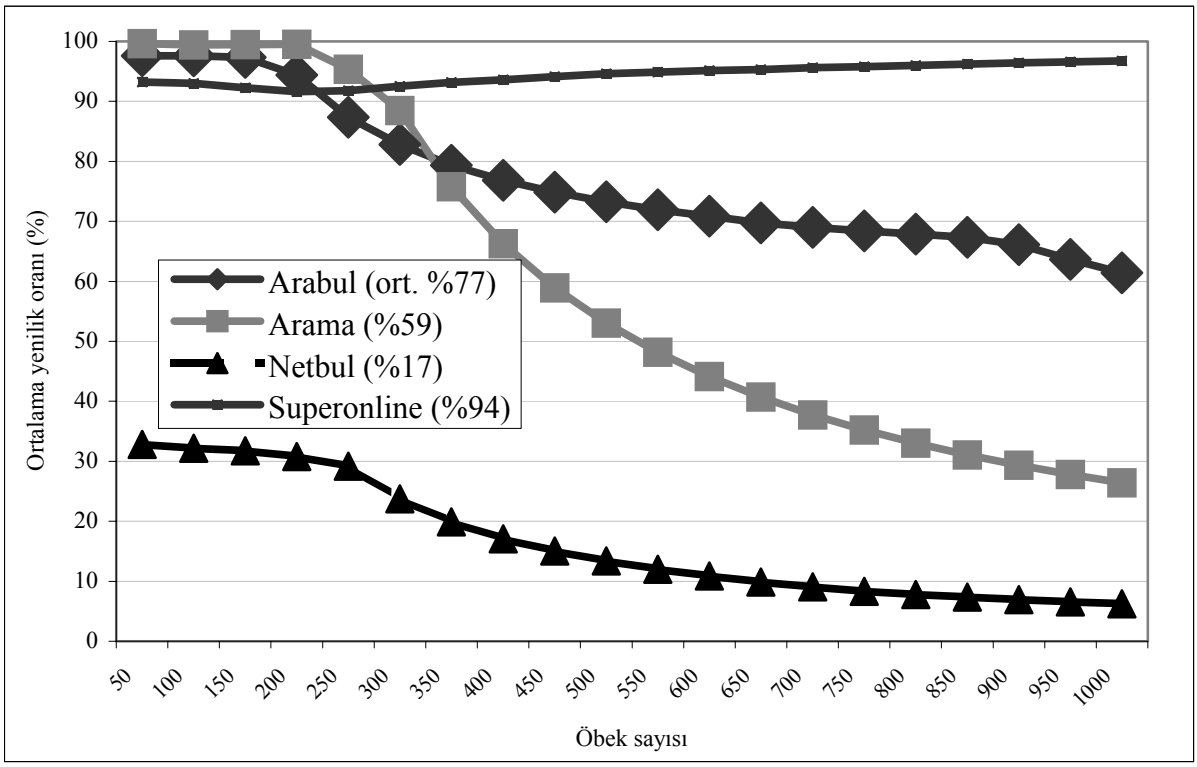


Şekil 17. Arama motorlarının "porno" sorusu için yenilik oranları

Arama motorlarının tüm sorular için kaydettikleri yenilik oranları Şekil 18'de topluca görülmektedir. Şekilden de görülebileceği gibi, Arama ve Arabul, havuza katkıda bulundukları belge sayısı sınırına (yaklaşık 300) gelene dek sürekli yüksek yenilik oranlarına (%100) ulaşmıştır. Superonline'ın ortalama yenilik oranı ise %95 civarındadır. Beş sorudan 3'ü için hiç bir belgeye erişemeyen Netbul'un ortalama yenilik oranı ise havuza katkıda bulunduğu sınıra gelene dek %30 civarında seyretmiştir.

Ortalama yenilik oranlarından, Netbul dışında[21] her arama motorunun eriştiği belgelerin hemen hemen tamamının "yeni" olduğu anlaşılmaktadır. Başka bir deyişle, her soru için farklı arama motorları farklı ilgili belgelere erişmektedir. Bilgi erişim performans değerlendirme araştırmalarında sık rastlanan bu olgu, farklı bilgi erişim sistemlerinin performanslarını karşılaştırmayı çeşitli açılardan güçleştirmektedir. Örneğin, her arama motorunun farklı ilgili belgelere eriştiği bir ortamda arama motorlarının kapsama oranları birbirine yakın çıkmakta, farklı arama motorları tarafından erişilen ilgili belgeler arasındaki çakışma (overlap) oranı sıfıra yaklaşmaktadır. Bunun temel nedenlerinden birisi, hiç kuşkusuz, farklı arama motorlarının farklı belgeleri dizinlemeleridir. Nitekim çalışmamızda, Arama'nın ağırlıklı bir biçimde Türkiye adresli siteleri dizinlerken, Superonline'ın bunun tam tersi bir politika izlediği ortaya çıkmıştır.

[21] Netbul'un ilgili belgelere eriştiği ilk iki soru için yenilik oranının %80-%85 civarında olduğuna daha önce işaret etmiştik. Bu bakımdan aslında bu değerlendirmemize Netbul'u da katmak mümkündür.



Şekil 18. Arama motorlarının tüm sorular için ortalama yenilik oranları

5.3.2.2 Arama Motorlarının Türkiye Adresli Belgeler İçin Yenilik Oranları

Sorulara göre dört arama motorunun Türkiye adresli belgeler için yenilik oranları ile her soru için arama motorlarının kaydettikleri (makro ortalama yöntemine göre hesaplanmış) ortalama yenilik oranları Tablo 21'de verilmektedir. Bu oranlar havuzda toplanan sadece Türkiye adresli ilgili belgelere dayanılarak hazırlanmıştır. Tablo 21'de dikkati çeken önemli noktalar aşağıda özetlenmektedir.

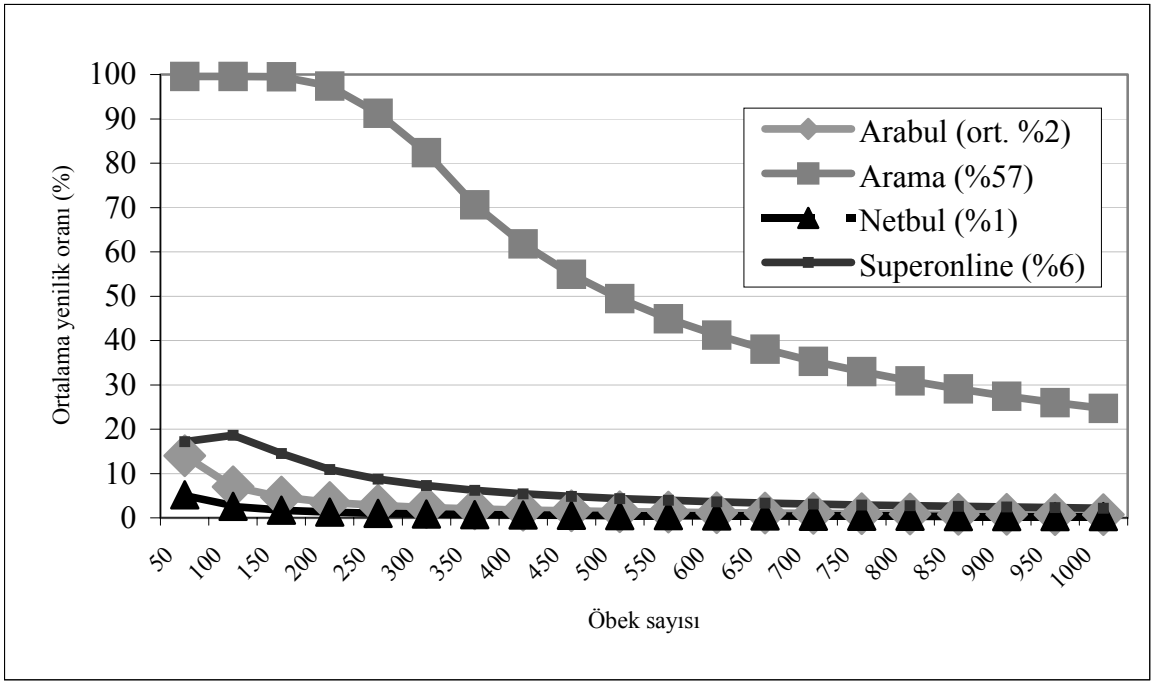
Arama'nın en sık aranan sorular için Türkiye adresli belgeleri bulmadaki tartışmasız üstünlüğü tabloda açıkça görülmektedir. Daha önce de değindiğimiz gibi (bkz. Tablo 17), havuza giren Türkiye adresli belgelerin büyük çoğunluğu Arama'ya aittir. Sadece bir soru ("oyun") için Superonline havuza giren belgelerin yaklaşık yüzde 25'ini sağlamıştır. Bu bakımdan, en sık aranan sorular için yeni belgelerin de Arama tarafından bulunması doğal karşılanmalıdır. Arama motorlarının Türkiye adresli tüm sorular için yenilik oranları Şekil 19'da verilmektedir. Arama, havuza katkıda bulunduğu belge sınırına gelene dek (yaklaşık 300) yüzde yüzlük yenilik oranına ulaşmıştır. Yenilik oranı bakımından Superonline 2.,

Arabul ise 3. sıradadır. Netbul'un yenilik oranı sadece ikinci soru ("oyun") için eriştiği 22 belgeye dayanmaktadır.

Tablo 21. Arama motorlarının Türkiye adresli belgeler için yenilik oranları

		Öbek sayısına göre yenilik oranları (%)																				
Sorgu	**Arama Motoru**	50	100	150	200	250	300	350	400	450	500	550	600	650	700	750	800	850	900	950	1000	**Ort.**
mp3	**Arabul**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Arama	100	100	99	100	86	72	62	54	48	43	39	36	33	31	29	27	25	24	23	22	53
	Netbul	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Superonline	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
oyun	**Arabul**	36	18	12	9	7	6	5	5	4	4	3	3	3	3	2	2	2	2	2	2	6
	Arama	100	99	99	100	100	100	85	75	66	60	54	50	46	43	40	37	35	33	31	30	64
	Netbul	26	13	9	7	5	4	4	3	3	3	2	2	2	2	2	2	2	1	1	1	5
	Superonline	86	93	73	55	44	36	31	27	24	22	20	18	17	16	15	14	13	12	11	11	32
sex	**Arabul**	12	6	4	3	2	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	2
	Arama	100	100	100	100	100	86	74	65	57	52	47	43	40	37	34	32	30	29	27	26	59
	Netbul	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Superonline	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
erotik	**Arabul**	14	7	5	4	3	2	2	2	2	1	1	1	1	1	1	1	1	1	1	1	3
	Arama	100	100	100	100	100	96	82	72	64	57	52	48	44	41	38	36	34	32	30	29	63
	Netbul	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Superonline	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
porno	**Arabul**	8	4	3	2	2	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1
	Arama	98	99	99	88	70	59	50	44	39	35	32	29	27	25	23	22	21	20	19	18	46
	Netbul	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Superonline	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Not: Sadece alan adı ".tr" ile biten ilgili belgeler dikkate alınmıştır.



Şekil 19. Arama motorlarının tüm sorular için Türkiye adresli yeni belge bulma oranları

5.4 Üst Veri Belirteçlerinden Yararlanma

Arama motorlarının Web belgelerinde yer alan HTML üst veri belirteçlerinden erişim için yararlanıp yararlanmadıkları iki küçük uygulamayla test edilmiştir. İlkinde, TKD Web sayfasının (bkz. Şekil 3) anahtar sözcük üst veri alanında yer alan "tkd", "kütüphaneciler", "dokümantasyon" ve "enformasyon" sözcükleri araştırmamızda kullanılan dört Türkçe arama motoruna (Arabul, Arama, Netbul ve Superonline) arama terimleri olarak topluca girilmiştir (8 Ekim 2001). Arama sonuçları Şekil 20'de verilmektedir.

Arama motoru	Sorgu	Erişim Sırası/Erişilen Toplam Belge Sayısı
Arabul	tkd kütüphaneciler dokümantasyon enformasyon	0/0
Arama	tkd & kütüphaneciler & dokümantasyon & enformasyon	0/4
Netbul	+tkd +kütüphaneciler +dokümantasyon +enformasyon	2/2
Superonline	tkd and kütüphaneciler and dokümantasyon and enformasyon	10/32000*

* İlk 20 belge değerlendirilmiştir.

Şekil 20. Türkçe arama motorlarında TKD Web sayfasında yer alan üst veri terimleri ile yapılan arama sonuçları

Üst veri alanlarından erişim amacıyla yararlanan arama motorlarının bu sorgulara karşılık TKD'nin Web sayfasına erişmeleri gerektiği varsayılmıştır. Görüldüğü gibi, Arama ve Arabul, TKD Web sayfasında yer alan sözcükler kullanılarak yapılan aramalarda ilgili sayfaya erişememiştir. Netbul ise ilgili soruya karşılık toplam iki belgeye erişmiş ve bunlardan birisinin TKD Web sayfası olduğu görülmüştür. Superonline ise bu soruya karşılık çok sayıda belgeye erişmiştir. Bu belgelerden ilk 20'si incelenmiş ve istenen TKD Web sayfasına 10. sırada erişildiği görülmüştür.

İkinci testte ise arama motorları tarafından dizinlendiği kesin olarak bilinen ve anahtar sözcük üst veri alanı dolu olan birer Web sayfası seçilmiştir. Daha sonra ilgili Web sayfasının anahtar sözcük üst veri alanında yer alan sözcükler kullanılarak oluşturulan sorguya karşılık, ilgili arama motorunun bu sayfaya erişip erişemediği kontrol edilmiştir. Test için kullanılan sorguların tamamen üst verileri içermesine ve sorgu içerisinde geçecek olan sözcüklerin belgenin başka bir yerinde (belgenin başlığı, belgenin Web adresi gibi) geçmemesine özellikle dikkat edilmiştir. Her arama motoru için seçilen Web sayfası, ilgili Web sayfasının anahtar sözcük üst veri alanında yer alan sözcükler, bu sözcüklerden seçilerek oluşturulan arama sorgusu, sorgu karşılığında dizinlendiği kesin olarak bilinen sayfaya arama motorunun erişip

erişemediği (erişilen toplam belge sayısı ve ilgili sayfaya kaçıncı sırada erişildiği bilgisiyle birlikte) Şekil 21'de verilmektedir.[22]

Arama Motoru	Seçilen Belge ve Web adresi	"Anahtar Sözcük" Üst Verileri	Sorgu	Erişim Sırası / Erişilen Toplam Belge Sayısı
Arabul	Çiçek Yolla www.cicekyolla.com	çiçek yolla, cicek yolla, özel hediye, ozel hediye, on-line, satış, satıs, satiş, satis, alış-veriş, alis-veris, buket, demet, aranjman, botanik, SSL, SET, Garanti Bankası SET Protokolü, kredi kartı, kradi karti, gül, flower	"çiçek yolla" ve "cicek yolla" ve "özel hediye"	**0/0**
Arama	Çiçek Saksıları ve Bitki Saksıları www.bauhaus.com.tr /ogutler/cicek.htm	Bauhaus, Yapı Market, ogut, öğüt, profesyonel, hobi	Bauhaus & ogut & profesyonel & hobi	**0/0**
Netbul	Sarı Frezya www.sarifrezya.com	Anneler günü, sevgililer günü, sevgiliye özel, flowers to turkey, çiçek, güzel, frezya, cicek, aranjman, tanzim, uygun, cicekler, sevgili, oranj, sarı, somon, zengin, buket, kırmızı, vazo, renkli, sepet, aydınlık, doğal, ısıklı, bitki, sıcak, güneş, karanfil, gelin duvağı, düğün, nikah	+"Anneler günü" +"sevgililer günü" +tanzim +uygun +oranj +buket +sepet	**1/40+**
Superonline	Çikas Çiçek Mağazası www.cikas.com	çiçekler, çiçek, güzel, çikas, pembe gerbera, singapur, tanzim, uygun, süslenmiş, aranjman, gerbera, gül, özel, gün, oranj, sarı, masa, lilyum, geniş, somon, lilyum, zengin, buket, kırmızı, vazo, renkli, sepet, nemli, aydınlık, doğal, ısıklı, bitki, sıcak, güneş, karanfil, gelin, gelin duvağı, düğün, nikah	"pembe gerbera" ve singapur	**1/1**

Şekil 21. Arama motorlarının "anahtar sözcük" üst verilerinden erişim amacıyla yararlanması

Benzeri bir biçimde, üst veri alanlarından erişim amacıyla yararlanan arama motorlarının ilgili sorgulara karşılık, dizinledikleri kesin olarak bilinen Web sayfalarına erişmeleri gerektiği varsayılmıştır. Şekilden de görülebileceği gibi, Arama ve Arabul arama motorları anahtar sözcük üst verilerinden yararlanmamaktadır. Yukarıda değinilen kural (sorguda kullanılan sözcüklerin sadece üst verilerden gelmesi) biraz yumuşatılarak, söz konusu sayfalara erişmek için belgelerin diğer kısımlarında geçen sözcükler de sorgularda kullanılmış ancak ilgili sayfalara erişilememiştir. Öte yandan, Netbul ve Superonline arama motorları üst

[22] Bu bölümde özetlenen iki testle ilgili olarak arama motorlarının eriştikleri belgeler http://cmpe.emu.edu.tr/bitirim/home/ adresinde "Metadata work" başlığı altında kayıtlıdır.

verilerde geçen sözcüklerden oluşturulan sorgulara karşılık ilgili Web sayfalarına ilk sırada erişim sağlamıştır.

Türkçe arama motorlarında erişim amacıyla üst verilerden yararlanma konusunda yapılan bu iki küçük testten elde ettiğimiz bulgular, daha önce özetlenen (bkz. 3.5) yabancı arama motorlarında elde edilen bulgularla paralellik göstermektedir. Her iki testte de Arama ve Arabul'un üst veri alanlarından erişim amacıyla yararlanmadıkları, Netbul ve Superonline'ın ise yararlandıkları ortaya çıkmıştır. Bir başka deyişle, Türkçe arama motorları tasarımcıları anahtar sözcük, tanım vb. gibi üst veri alanlarında verilen erişime yardımcı olabilecek dizin terimlerinden yeterince yararlanmamaktadırlar.

6 SONUÇ VE ÖNERİLER

Bu çalışmada ülkemizde yaygın olarak kullanılan Arabul, Arama, Netbul ve Superonline'a çeşitli türde 17 soru yöneltilmiş ve bu sorulara karşılık erişilen “ilgili” ve “ilgisiz” belgelere dayanarak arama motorlarının çeşitli kesme noktalarındaki duyarlık ve normalize sıralama değerleri açısından performansları değerlendirilmiştir. Arama motorlarının dizinlenen belgeleri ne kadar sıklıkla ziyaret ettikleri ve güncelleştirdikleri erişim çıktılarında yer alan “ölü” (yani erişilemeyen) adreslerin sayısına bakılarak saptanmıştır. . Türkçe arama motorlarında en sık aranan beş sözcük ("mp3", "oyun", "sex", "erotik" ve "porno") dört arama motorunda aranmış ve her arama motorunun kapsama ve yenilik oranları bulunmuştur. Arabul, Arama, Netbul ve Superonline'ın belgeleri dizinlemek amacıyla "anahtar sözcük", "tanım" gibi HTML üst veri (metadata) alanlarından yararlanıp yararlanmadıkları iki küçük deneyle test edilmiştir. Aşağıda araştırma sonuçları kısaca özetlenmekte ve arama motorlarının performanslarını artırmak için bazı öneriler yer almaktadır.

Arama motorlarının eriştikleri ortalama her 6 belgeden birisi (%17) ölü bağlantı içermektedir. Arama motorlarının ölü bağlantı oranları %27 (Arama) ile %4 (Netbul) arasında değişmektedir (Superonline ve Arabul %19). Netbul ile diğer üç arama motorunun ölü bağlantı oranları arasındaki fark istatistiksel açıdan anlamlıdır. Diğer arama motorlarına göre Netbul’da dizinlenen belgeler dizinleme robotları tarafından daha sık aralıklarla ziyaret edilmekte ve erişilemez hale gelen (ölü) adresler daha hızlı güncelleştirilmektedir (bkz. 5.1).

Arama motorları dizinledikleri belgeleri daha sık aralıklarla ziyaret etmelidirler. Erişilemez hale gelen belgeler ya dizinlerden çıkarılmalı ya da bu belgelerin yeni adresleri hızla güncelleştirilmelidir.

Arabul 17 sorudan 6’sı (%35), Netbul ise 17 sorudan 1’i (%6) için hiç bir belgeye erişememiştir. Arama ve Superonline ise tüm sorular için en az bir belgeye erişmişlerdir. Toplam 17 sorudan 3’ünde (%18) hiç bir arama motoru ilgili belgeye erişememiştir. Arabul toplam 17 sorudan 11 (%65), Netbul 8 (%47), Superonline 5 (%29), Arama ise 4 (%24) soruda ilgili belgelere erişememiştir (bkz. 5.2.2.1).

Arama motorlarının bazı sorular için hiç bir belgeye ya da hiç bir ilgili belgeye erişememe nedenleri araştırılmalıdır. Bu sorunu çözmek için daha çok sayıda ve çeşitli belgeler/siteler dizinlenmeli ve erişim algoritmaları sıfır sonuç vermeyecek şekilde iyileştirilmelidir.

Dört arama motorunun soru başına eriştikleri toplam ilgili belge sayısı 10’dur. Arama soru başına ortalama yaklaşık 4, Superonline ise 3 ilgili belgeye erişmiştir. Arabul ve Netbul çok

sayıda soru için hiçbir ilgili belgeye erişemediklerinden soru başına erişilen ortalama ilgili belge sayıları düşüktür (1,5).

Arama motorlarının eriştikleri ortalama her 6 belgeden 5'i ilgisizdir. Arama motorlarının ortalama duyarlık oranları %28 (Arama) ile %11 (Netbul) arasında değişmektedir (Superonline %20, Arabul %15). Kesme noktası yükseldikçe, yani kullanıcının incelediği belge sayısı arttıkça, arama motorlarının ortalama duyarlık değerleri %50 oranında düşmüştür. Bu düşüş Arama'da daha belirgindir (%90). Arama'nın ilk 5 belgedeki ortalama duyarlık değeriyle (%40) Arabul (%16) ve Netbul'un (%13) ortalama duyarlık değerleri arasındaki fark istatistiksel yönden anlamlıdır (bkz. 5.2.2.2). Başka bir deyişle, ilk 5 belgede Arama, Arabul ve Netbul'dan daha fazla sayıda ilgili belgeye erişmektedir. Arama motorlarının daha yüksek kesme noktalarında eriştikleri ilgili belge sayıları ise birbirine benzemektedir.

Arama motorlarının her soruya karşılık az sayıda ilgili belgeye erişmesi dizinlenen toplam belge sayısının azlığından kaynaklanabileceği gibi kullanıcı arabirimlerinin etkin olmamasından ya da ileri tekniklere dayanan bilgi erişim algoritmaları kullanılmamasından da kaynaklanabilir. Arama motorlarının duyarlık değerlerinin düşük olma nedenleri ayrıntılı olarak incelenmelidir.

Arama motorlarının ortalama normalize sıralama değerleri %54 (Arama) ile %20 (Arabul) arasında değişmektedir (Superonline %37, Netbul %30). Arama, erişim çıktılarında ilgili belgeleri Arabul'dan ve Netbul'dan daha üst sıralarda göstermektedir. Arama'nın erişilen ilk 5, 10 ve 15 belgede kaydettiği ortalama normalize sıralama değerleri Arabul'unkilerden, ilk 10 belgede kaydettiği ortalama normalize sıralama değeri Netbul'unkinden daha yüksektir (bkz. 5.2.2.3).

Ortalama duyarlık değerlerinin yüksek olduğu aramalarda ortalama normalize sıralama değerleri de genellikle yüksektir (Pearson's $r = .61$, $p < .05$). Ancak değerlendirilen belge sayısı arttıkça duyarlık ile normalize sıralama değerleri arasındaki ilişki giderek zayıflamaktadır (bkz. 5.2.2.4).

Yapılan araştırmalarda duyarlık ve anma gibi geleneksel performans ölçütleriyle kullanıcı merkezli performans değerlendirmeleri arasında güçlü bir ilişki olmadığı ortaya çıkmıştır. Başka bir deyişle, bazen kullanıcılar az sayıda ilgili belgeye erişen ama bu belgeleri erişim çıktılarının üst sıralarında gösteren bilgi erişim sistemlerini daha başarılı bulabilmektedirler. Bu bakımdan arama motorlarının erişilen ilgili belgeleri ilk sıralarda gösterme konusunda daha fazla çaba sarfetmeleri gerekmektedir.

Arama motorlarının tüm sorular için kaydettikleri ortalama duyarlık değerleri % 0 ile %36 arasında değişmektedir. Hiç bir arama motoru 4. ("Türkçe arama motorlarında performans

değerlendirme"), 12. ("Demirel veya Sezer'in TEMA hakkındaki görüşleri") ve 17. ("TBMM Başkanı Ömer İzgi") sorular için ilgili belgeye erişememiştir. Arama motorları 2. ("barok müzik"), 7. ("DPT") ve 11. ("Demirel veya Sezer") sorularda nispeten daha yüksek sayıda ilgili belgeye erişmişlerdir (bkz. 5.2.2.5).

Arama motorlarının tüm sorular için ortalama normalize sıralama değerleri % 0 ile %66 arasında değişmektedir. Arama motorlarının ortalama normalize sıralama değeri açısından en başarılı oldukları sorular 7., 9. ("uzaylılar") ve 13. ("uzay") sorulardır.

Arama motorlarının hem ortalama duyarlık hem de ortalama normalize sıralama değerleri açısından en başarılı oldukları sorular ise 7., 8. (uzaylı"), 9. ve 11. sorulardır.

Arama motorları, Web'de yaygın olarak kullanılan "internet", "etik", "arama" vb. terimlerin geçtiği spesifik arama sorularında nispeten daha az başarı göstermiştir. Öte yandan tek sözcükten oluşan ya da "VEYA" işleci kullanılan sorularda, erişilen ilgisiz belge sayısı yüksek olmasına rağmen, arama motorları nispeten daha başarılı olmuştur. "VE" işlecinin kullanıldığı sorularda ise başarı oranı daha düşüktür. Arama motorları Boole işleçleri kullanılarak yapılan aramalarda genelde tutarlı sonuçlar vermektedir (bkz. 5.2.3).

Arama motorlarının "VE" işleci kullanılan sorulardaki başarı oranını yükseltmek için daha fazla belge dizinlemeleri gerekmektedir. Tek terimden oluşan sorularda başarı oranları, kullanıcının bu terimleri hangi bağlamda aradıkları belirlenmeye çalışılarak artırılabilir. Web'de yaygın olarak kullanılan terimlerin geçtiği soruları kullanıcının daha spesifik yapmasına olanak verilmelidir. Bir başka yöntem ise, kullanıcıların geçmişte tek sözcükten oluşan bu tür soruları aradıklarında daha çok hangi bağlantıları tıklamış oldukları bilgisine dayanan sezgisel (heuristic) bilgi erişim algoritmaları geliştirmektir.

Arama motorları kullanıcılar tarafından girilen soruları daha iyi analiz etmek ve performansı artırmak için gövdeleme algoritmalarından yararlanmamaktadır. Özellikle Türkçe sözcüklerle yapılan aramalarda gövdeleme algoritmalarının kullanılması arama motorlarının bilgi erişim performansını artırabilir. Bu nedenle Türkçenin dilbilgisi özelliklerini de dikkate alan gövdeleme algoritmaları geliştirilmeli ve kullanılmalıdır.

Türkçe karakter sorunu henüz çözülememiştir. Arama motorları Türkçe karakterler kullanılarak yapılan aramalarda farklı sonuçlar vermektedir (bkz. 5.2.3). Arama motorlarının Türkçe karakter sorununa farklı yaklaşımları kullanıcılar açısından bazı olumsuzluklar yaratmaktadır. Çoğu kullanıcı bu durumun genellikle farkında değildir. Bu bakımdan, kullanıcılar Türkçe karakter kullanımı nedeniyle erişilen belgeleri değerlendirirken zorlanmaktadırlar. Türkçe arama motorlarında yapılan aramaların büyük bir çoğunluğu Türkçeyi kullandıklarından bu sorunun bir an önce çözülmesi gerekmektedir.Web'deki

Türkçe içerik miktarının giderek arttığı düşünülecek olursa, artık arama sorusunda yer alan Türkçe karakterlerin en yakın İngilizce karakterlere çevrilmesi gibi basit yaklaşımlar yerine, gerek Türkçenin dil özelliklerini gerekse kullanıcıların arama davranışlarını da dikkate alan yaklaşımlar yeğlenmelidir.

Türkçe arama motorlarında en sık aranan sözcüklere (“mp3”, “oyun”, “sex”, “erotik” ve “porno”) karşılık erişilen belgelerin büyük bir çoğunluğu (%86) Türkiye adresli değildir. Superonline alan adı “.tr” ile bitmeyen belgelere erişmede tartışmasız bir üstünlüğe sahiptir. Bunda Superonline’ın AltaVista ile işbirliğinin büyük payı olduğu kanısındayız. Türkiye adresli en fazla belgeye erişen arama motoru ise Arama’dır.

Superonline arama motorları arasında en yüksek ortalama kapsama oranına sahiptir. En sık aranan sorulara karşılık erişilen ortalama her 5 ilgili belgeden 2’sine Superonline tarafından erişilmiştir (bkz. 5.3.1.1). Superonline’ı Arabul, Arama ve Netbul izlemektedir. Türkiye adresli belgelerde ise Arama en sık aranan beş soru için ilgili belgelerin büyük bir çoğunluğunu kapsamaktadır. Diğer arama motorlarının Türkiye adresli belgeleri kapsama oranları ihmal edilebilir düzeydedir (bkz. 5.3.1.2).

Arama ve Arabul’un en sık aranan sorular için yenilik oranları %100 civarındadır. Yani, bu arama motorlarının en sık aranan sorulara karşılık eriştikleri belgelerin hemen hemen tümü “yeni”dir. Superonline'ın ortalama yenilik oranı %95 civarındadır. Netbul’da ise bu oranlar %30 civarındadır (bkz. 5.3.2.1).

En sık aranan sorular için Türkiye adresli “yeni” belge bulmada Arama tartışmasız bir üstünlüğe sahiptir. Superonline en sık aranan beş sorudan dördünde Türkiye adresli hiç bir belgeye erişememiştir. En sık aranan sorular için Arabul ve Netbul az sayıda yeni belgeye erişmişlerdir (bkz. 5.3.2.2).

Arama motorları tarafından en sık aranan sorular için bulunan Türkiye adresli ve Türkiye adresli olmayan hemen hemen bütün belgeler yenidir. Bu durum, arama motorlarının eriştikleri ilgili belgeler arasında çok düşük bir çakışma olduğunu göstermektedir. Bir başka deyişle, aynı sorular için her arama motoru birbirinden oldukça farklı belgeler dizinlemekte ve doğal olarak farklı ilgili belgelere erişmektedir.

Bu sonuçlar belirli bir konuda ilgili belgelerin tümüne erişebilmek için birden fazla Türkçe arama motoru üzerinde arama yapılması gereğini ortaya çıkarmaktadır. Bazı arama motorlarının çok az sayıda Türkiye adresli belge dizinledikleri görülmektedir. Türkiye’yle ilgili bazı sorularda (“Atatürk ve Fikriye Hanım”, “Ömer İzgi” vb. gibi) arama motorlarının nispeten daha başarısız olmalarının nedenlerinden birisi de kanımızca budur. Türkçe arama motorlarının dizinledikleri Türkiye adresli belge sayıları artırılmalıdır.

Arama ve Arabul arama motorları HTML belgelerinde yer alan “anahtar sözcük” ve “tanım” üst veri (metadata) alanlarında geçen terimleri dizinlememekte ve erişim amacıyla bu terimlerden yararlanmamaktadır. Netbul ve Superonline’ın ise bu alanları dizinledikleri ve erişim amacıyla kullandıkları ortaya çıkmıştır (bkz. 5.4). İlgili alanlara girilen bilgilerin arama motorlarını yanıltmak amacıyla zaman zaman kötüye kullanıldığı bilinen bir gerçektir. Ancak arama motorlarının bu tür kötüye kullanımları eleyecek daha akıllı arama algoritmaları geliştirmeleri ve bu alanlarda yer alan erişim açısından değerli bilgilerden yararlanmaları gerektiği kanısındayız.

KAYNAKÇA

Adalı, S., Bufi, C. ve Temtanapat, Y. (1997). Integrated search engine. Xindong Wu et al. (Eds.), *1997 IEEE Knowledge and Data Engineering Exchange Workshop: Proceedings: November 4, 1997, Newport Beach, California* içinde (s. 140-147). Los Alamitos, CA: IEEE Computer Society Press.

Akal, F. (2000). *Kavram tabanlı Türkçe arama makinası*. Yayımlanmamış yüksek lisans tezi, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü.

Alsaffar, A.H., Deogun, J.S., Raghavan, V.V. ve Sever, H. (2000, March/June). Enhancing concept-based retrieval based on minimal term sets. *Intelligent Information Systems Journal*, 14(2/3), 155-73.

Alsaffar, A.H., Deogun, J.S., Raghavan, V.V. ve Sever, H. (1999). Concept based retrieval by minimal term sets. Z.W. Ras ve A. Skowron (Eds.), *Foundations of Intelligent Systems: 11th International Symposium on Methodologies for Intelligent Systems (ISMIS'99). Warsaw, Poland. June 8-11, 1999* içinde (s. 114-123). Berlin: Springer-Verlag.

Aslantürk, O. (2000, Ekim-Aralık). Türkçe tabanlı arama araçlarının karşılaştırılmasında yöntem tanımı ve popüler arama araçları üzerine bir deneme. *Düşünceler,* No. 54-55, s. 3-19.

Badino, G.N. (2001). *Approximate text searching.* Unpublished doctoral thesis, University of Chile, Santiago, Chile. [Çevrimiçi]. Elektronik adres: http://citeseer.nj.nec.com/navarro98approximate.html [30 Ekim 2001].

Belkin, N.J., Kantor, P., Fox, E.A. ve Shaw, J.A. (1995). Combining the evidence of multiple query representations for information retrieval. *Information Processing & Management*, 31, 431-448.

Bergman, M.K. (2001, August). The deep Web: Surfacing hidden value. (White Paper), *The Journal of Electronic Publishing*, 7(1). [Çevrimiçi]. Elektronik adres: http://www.press.umich.edu/jep/07-01/bergman.html [10 Aralık 2001].

Berners-Lee, T., Cailliau, R., Groff, J. ve Pollermann B. (1992). World Wide Web: The information universe. *Electronic Networking: Research, Applications and Policy*, 1(2), 74-82.

Berners-Lee, T., Cailliau, R., Luotonen, A., Nielsen, H.F. ve Secret, A. (1994, August). The World-Wide Web. *Communications of the ACM*, 37(8): 76 - 82.

Bharat, K. ve Broder, A. (1998). A technique for measuring the relative size and overlap of public Web search engines. *Computer Networks and ISDN Systems*, 30, 379-388.

Blair, D.C. (1990). *Language representation in information retrieval.* Amsterdam: Elsevier, 1990.

Blair, D.C. ve Maron, M.E. (1985, March). An evaluation of retrieval effectiveness for a full-text document retrieval system. *Communications of the ACM,* 28(3), 289-299.

Bollmann-Sdorra, P. ve Raghavan, V.V. (1993). On the delusiveness of adopting a common space for modeling IR objects: Are queries documents? *Journal of the American Society for Information Science*, 44, 579-587.

Bollmann-Sdorra, P., Raghavan, V.V. ve Sever, H. (1999). Term preference weight. Yasemin Topaloğlu, Mustafa Türksever ve Aylin Kantarcı (Eds.), *Proceedings of 14th International Symposium on Computer and Information Sciences (ISCIS'99), October 18-20, 1999, Izmir, Turkey* içinde (s. 360-369). İzmir: Ege Üniversitesi.

Brake, D. (2001). Lost in Cyberspace. *New Scientist Magazine* [Çevrimiçi]. Elektronik adres: http://www.newscientist.com/ ; http://www.well.com/~derb/lost.html [30 Eylül 2001].

Chen, H. ve Lynch, K.J. (1992). Automatic construction of networks of concepts characterizing document databases. *IEEE Transactions on Systems, Man, and Cybernetics*, 22(5), 885-902. [Çevrimiçi]. Elektronik adres: http://ai.bpa.arizona.edu/go/intranet/papers/Automatic-92.pdf [9 Ekim 2001].

Chu, H. ve Rosenthal, M. (1996). Search engines for the World Wide Web: a comparative study and evaluation methodology. Steve Hardin (Ed.), *Global Complexity: Information, Chaos and Control. ASIS '96: Proceedings of the 59th ASIS Annual Meeting, Baltimore, Maryland, October 21-24, 1996* içinde (s. 127-135). Medford, NJ: American Society for Information Science. [Çevrimiçi]. Elektronik adres: http://www.asis.org/annual-96/ElectronicProceedings/chu.html [12 Şubat 2002].

Clarke, S.C. ve Willet, P. (1997, July/August). Estimating the recall performance of Web search engines. *Aslib Proceedings*, 49(7), 184-189.

Cooper, W.S. (1995 January). Some inconsistencies and misidentified modeling assumptions in probabilistic information retrieval. *ACM Transactions on Information Systems,* 13(1): 100-111.

Crestani, F., Lalmas, M., Van Rijsbergen, C.J. ve Campbell, I. (1998). Is this document relevant?... Probably: A survey of probabilistic models in information Retrieval. *ACM Computing Surveys*, 30(4), 528-552.

Crouch, C.J. ve Yang, B. (1992). Experiments in automatic statistical thesaurus construction. Nicholas Belkin, Peter Ingwersen ve Annelise Mark Pejtersen (Eds.), *SIGIR '92: Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval: Copenhagen, Denmark, June 21-24, 1992* içinde (s. 77-88). New York, NY: Association for Computing Machinery.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. ve Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41, 391-407.

Deogun, J.S., Sever, H. ve Raghavan, V.V. (1998). Structural abstractions of hypertext documents for Web-based retrieval. A Min Tjoa ve Roland R. Wagner (Eds.), *Proceedings: Ninth International Workshop on Database and Expert Systems*

Applications, August 26-28, 1998, Vienna, Austria içinde (s. 385-390) Los Alamitos, CA: IEEE Computer Society.

Deutsch, P. (1992, Spring). Resource discovery in an Internet environment – The Archie approach. *Electronic Networking: Research, Applications and Policy,* 2(1), 45-51.

Doorenbos, R.B. Etzioni, O. ve Weld, D.S. (1996). A scalable comparison-shopping agent for the World-Wide Web. (Tech. Rep. No. UW-CSE-96-01-03). Department of Computer Science and Engineering, University of Washington.

Dublin Core Metadata Initiative. (1998). *Dublin Core Element Set, Version 1.0.* [Çevrimiçi]. Elektronik adres: http://www.purl.org/dc. [25 Aralık1999].

Duda, R.O. ve Hart, P.E., (1973). *Pattern classification and scene analysis.* New York: Wiley.

Duran, G. (1999). *GövdeBul: Türkçe gövdeleme algoritması.* (Yayımlanmamış yüksek lisans tezi), Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, Ankara. [Çevrimiçi]. Elektronik adres: http://ata.cs.hun.edu.tr/~km/gokmen/index.html [Eylül 15, 2001].

Etzioni, O. ve Weld, D. (1994, July). A softbot-based interface to the Internet. *Communications of the ACM,* 37(7), 72-76.

Foltz, P.W. (1996). Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments and Computer, 28*(2), 197-20. [Çevrimiçi]. Elektronik adres: http://www-psych.nmsu.edu/~pfoltz/reprints/BRMIC96.html [12 Ekim 2001]

Frank, A. (1996, June). Internet services. *LAN Magazine/Network Magazine*, [Çevrimiçi]. No. 94. Elektronik adres: http://www.networkmagazine.com/article/NMG20000727S0006 [27 Şubat 2002]

Frei, H.P. ve Stieger, D. (1995). The use of semantic links in hypertext information retrieval. *Information Processing & Management*, 31, 1-13.

Furner, J., Ellis, D. ve Willet, P. (1996). The representation and comparison of hypertext structures using graphs. Maristella Agosti ve Alan Smeaton (Eds.) *Information Retrieval and Hypertext* içinde (s. 75-96). Boston, MA: Kluwer.

Gordon, M. ve Pathak, P. (1999). Finding information on the World Wide Web: The retrieval effectiveness of search engines. *Information Processing & Management*, 35, 141-180.

Graham, I.S. (1997). *HTML sourcebook*. New York, NY: Wiley. [Çevrimiçi]. Elektronik adres: http://www.w3.org/MarkUp/ [30 Eylül 2001].

Gudivada, V.N., Raghavan, V.V., Grosky, W.I. ve Kasanagottu, R. (1997, September/October). Information retrieval on the World Wide Web. *IEEE Internet Computing*, 1(5), 58-68.

Hawking, D., Craswell, N., Thislewaite, P. ve Harman, D. (1999). Results and challenges in Web search engine evaluation. *Proceedings of the eighth Text REtrieval Conference*

(TREC-8), Gaithersburg, Maryland, November 17-19, 1999 içinde [Çevrimiçi]. Elektronik adres: http://trec.nist.gov/pubs/trec8/t8_proceedings.html [11 Eylül 2001].

Henshaw, R. (2001, September). What next for Internet journals? Implications of the trend towards paid placement in search engines. *First Monday*, [Çevrimiçi]. 6(9). Elektronik kopya: http://www.firstmonday.dk/issues/issue6_9/henshaw [2 Şubat 2002].

Henshaw, R. ve Valauskas, E.J. (2001). Metadata as a catalyst: Experiments with metadata and search engines in the Internet journal *First Monday*, *Libri*, 51(2), 86-101.

Howe, W. (2001, August 31). A brief history of the Internet. [Çevrimiçi]. Elektronik adres: http://www.walthowe.com/navnet/history.html [2 Şubat 2002].

Inktomi Corp., (2000). Web surpasses one billion documents. [Çevrimiçi]. Elektronik adres: http://www.inktomi.com/new/press/2000/billion.html [2 Şubat 2002].

Internet Society. (2000, January 18), What is Internet? [Çevrimiçi]. Elektronik adres: http://www.isoc.org/internet, [2 Şubat 2002].

Jansen, B., Spink, A., Bateman, J. ve Saracevic, T. (1998). Real life information retrieval: A study of user queries on the Web. *SIGIR Forum*, 32(1), 5-17.

Jansen, J. (1996). Using an intelligent agent to enhance search engine performance. (1996). *First Monday*, [Çevrimiçi] 2(3). Elektronik adres: http://www.firstmonday.dk/issues/issue2_3/jansen/index.html. [2 Şubat 2002].

Kahle, B. (1997 March). Preserving the Internet. *Scientific American* [Çevrimiçi] 276(3), 82-83. Elektronik kopya: http://www.sciam.com/0397issue/0397kahle.html [10 Aralık 2001]

Kahle, B. (1996, April 11). Archiving the Internet. [Çevrimiçi]. Elektronik adres: http://www.archive.org/sciam_article.html [2 Şubat 2002].

Kartal, M. (1998). *Bilimsel araştırmalarda hipotez testleri: Parametrik ve nonparametrik teknikler*. 2. bs. Erzurum: Şafak.

Kirsch, S. (1998). Infoseek's experiences searching the Internet, *SIGIR Forum*, 32(2), 3-7.

Klarin, S., Pavelić, D. ve Pigac, S. (2001). Croatian remote access electronic serials: Results of a survey. *Austrian Metadata Workshop, 18 May 2001, Vienna*. [Çevrimiçi]. Elektronik adres: http://www.cscaustria.at/vortrag/Austrian_Metadata_Workshop2001_Willer1.ppt [14 Şubat 2002].

Kleinberg, J.M. (1998). Authoritative source in a hyperlinked environment. *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms* içinde (s. 668-677). New York, NY: Association for Computing Machinery.

Koehler, W. Digital libraries and World Wide Web sites and page persistence. *Information Research* [Çevrimiçi]. 4 (4), 1999. Elektronik kopya: http://information.net/ir/4-4/paper60.html [25 Ocak 2002].

Korfhage, R.R. (1997). *Information storage and retrieval.* New York: Wiley.

Köksal, A. (1987 Ocak-Mart). Bilgi erişim sorunu ve bilgi erişim dizgelerine ilişkin temel kavramlar. *Bilişim*, s. 18-25.

Köksal, A. (1979). *Bilgi erişim sorunu ve bir belge dizinleme ve erişim sistemi tasarımı.* (Yayımlanmamış doçentlik tezi), Ankara: Hacettepe Üniversitesi.

Kredel, H., Meuer, H.-W., Schumacher, R. ve Strohmaier, E. (2000) Internet and WWW - An introduction. [Çevrimiçi]. Elektronik adres: http://www.uni-mannheim.de/rum/dokus/intro.htm [2 Şubat 2002].

Kobayashi, M. ve Takeda, K. (2000, June). Information retrieval on the Web. *ACM Computing Surveys*, 32(2), 144-172.

Küçük, M.E., Olgun, B. ve Sever, H. (2000). Application of metadata concepts to discovery of Internet resources. Tatyana Yakhno (Ed.), *Advances in Information Systems: First International Conference, ADVIS 2000, Izmir, Turkey, October 25-27, 2000 Proceedings (LNCS)* (Vol. 1909) içinde (s. 304-313). Berlin: Springer Verlag.

Laursen, J.V. (1998, February/March).Somebody wants to get in touch with you: Search engine persuasion. *Database*, s. 43-46 [Çevrimiçi]. Elektronik adres: http://www.onlineinc.com/database/DBtocs/DBtocfeb98.html [5 Ekim 2001].

Lawrence, S. ve Giles, C. L. (1999). Accessibility of information on the Web. *Nature,* 400, 107-109.

Lawrence, S. ve Giles, C. L. (1998, April 3). Searching the World Wide Web. *Science* [Çevrimiçi]. 280(5360), 98-100. Elektronik adres: http://www.neci.nec.com/~lawrence/science98.html [14 Şubat 2002].

Lee, J.H. (1997). Analysis of multiple evidence combination. N.J. Belkin, A.D. Narasimhalu ve P. Willet (Eds.), *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, Pennslvania, USA, July 1997* içinde (s. 267-275). New York: ACM Press.

Lee, J.H. (1995). Combining multiple evidence from different properties of weighting schemes. Edward A. Fox, Peter Ingwersen ve Raya Fidel (Eds.), *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'95), Seattle, Washington, USA, July 9-13, 1995* içinde (s.180-188). New York, NY: ACM Press.

Leighton, H.V. ve Srivastava, J.V. (1999). First 20 precision among World Wide Web search services (search engines). *Journal of the American Society for Information Science*, 50, 870-881.

Leuski, A. (2001). Interactive information organization: Techniques and evaluation, (Unpublished doctoral dissertation). University of Massachusetts, Amherst, MA. [Çevrimiçi]. Elektronik adres: http://ciir.cs.umass.edu/ [23 Eylül 2001].

Lynch, C. (1997 March). Searching the Internet. *Scientific American*, 52-56. [Çevrimiçi]. Elektronik adres: http://www.sciam.com/0397issue/0397lynch.html [12 Şubat 2002].

McCune, B.P., Tong, R.M., Dean, J.S. ve Shapiro, D.G. (1985). {RUBRIC}: A system for rule-based information retrieval. *IEEE Transactions on Software Engineering*, 11(9), 939-944.

Maron, M.E. (1984). Probabilistic retrieval models. Brenda Dervin ve Melvin J. Voigt (Eds.), *Progress in Communication Sciences* Vol. 5 içinde (s. 145-176). Norwood, NJ: Ablex.

Mettrop, W. ve Nieuwenhuysen, P. (2001). Internet search engines -- Fluctuations in document accessibility. *Journal of Documentation,* 57, 623-51.

Mitchell, T.M. (1997). *Machine learning*. New York: McGraw Hill.

Montague, M. ve Aslam, J.A. (2001). Relevance score normalization for metasearch. *Tenth International Conference on Information and Knowledge Management, Atlanta, Georgia, USA, November 5-10, 2001* içinde (s. 427-433). New York: ACM.

Notess, G.R. (2001). Joining the in-crowd. *EContent*, 24(3), 60.

Olgun, B. ve Sever, H. (2000). Kaynak keşif yeteneğinin artırılması için Internet kaynaklarının içeriklerinin standart biçimde tanımlanması. *Bilgi Dünyası*, 1, 56-88.

O'Neill, E.T., Lavoie, B.F. ve McClain, P.D. (1998). An analysis of metadata usage on the Web. 1998. [Çevrimiçi]. Elektronik kopya: http://www.oclc.org/research/publications/arr/1998/ oneill_etal/metadata.htm [12 Kasım 2001]

Oppenheim, C., Morris, A. ve McKnight, C. (2000). The evaluation of WWW search engines. *Journal of Documentation*, 56, 190-211.

PC Computing Online. (1996 August). Search engines. [Özel sayı]. [Çevrimiçi]. Elektronik adres: http://www.zdnet.com/pccomp/webmap/spmaps/map0896/search.html [12 Eylül 2001]

Qin, J. ve Wesley, K. (1998 September). Web indexing with meta fields: a survey of Web objects in polymer chemistry. *Information Technology and Libraries,* 17(3), 149-156.

Robertson, S.E. ve Jones, K.S. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27, 29-146.

Rocchio, Jr., J.J. (1971). Evaluation viewpoints in document retrieval. G. Salton (Ed.), *The SMART retrieval system: Experiments in automatic document processing* içinde (s. 324-336). Englewood Cliffs, NJ: Prentice-Hall.

Salton, G. (1989). *Automatic text processing.* Massachusetts: Addison-Wesley.

Salton, G. (Ed.) (1971). *The SMART retrieval system: Experiments in automatic document processing*. Englewood Cliffs, NJ: Prentice-Hall.

Salton, G. ve Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41, 288-97.

Salton, G. ve Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24, 513-523.

Salton, G., Fox, E.A. ve Wu, H. (1983). Extended boolean information retrieval. *Communications of the ACM*, 26(11):1022-36.

Salton, G., Wong, A. ve Yu, C.T. (1976). Automatic indexing using term discrimination and term precision measurement. *Information Processing & Management*, 12, 43-51.

Savoy, J. ve Picard, J. (2001). Retrieval effectiveness on the Web. *Information Processing & Management*, 37, 543-569.

Sezer, E. (1999). *SMART bilgi erişim sisteminin Türkçe yerelleştirilmesi ve otomatik gömü üretimi.* (Yayımlanmamış yüksek lisans tezi). Hacettepe Üniversitesi Fen Bilimleri Enstitüsü. [Çevrimiçi]. Elektronik adres: http://ata.cs.hun.edu.tr/~km/belgeler.html [15 Eylül 2001].

Silverstein, C., Henziger, M., Marais, H. ve Moricz, M. (1999). Analysis of a very large Web search engine query log. *SIGIR Forum*, 33(1), 6-12.

Soydal, İ. (2000). *Web arama motorlarında performans değerlendirmesi.* (Yayımlanmamış bilim uzmanlığı tezi). Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü. Ankara.

Sparck Jones, K. (1971). *Automatic keyword classification for information retrieval*. London: Butterworths.

Srinivassan, P. (1992). Thesaurus construction. Bill Frakes ve Ricardo Baeza Yates (Eds.), *Information retrieval data structures & algorithms* içinde (s.161-218). Englewood Cliffs, New Jersey: Prentice-Hall.

Sullivan, D. (2001, June 26) How search engines work. [Çevrimiçi]. Elektronik adres: http://www.searchenginewatch.com/webmasters/work.html [2 Şubat 2002].

Sullivan, D. (2000) Search engine features for searchers. [Çevrimiçi]. Elektronik adres: http://www.searchenginewatch.com/facts/ataglance.html, [20 Aralık 2000].

Svenonius, E. (2000). *The intellectual foundations of information organization*. Cambridge, MA: MIT Press.

Tennant, R., Ober, J. ve Lipow, A.G. (1996). *Internet el kitabı*. (Çev.Y. Tonta ve diğerleri) Ankara: Türk Kütüphaneciler Derneği.

Tonta, Y. (1995). Bilgi erişim sistemleri. *Türk Kütüphaneciliği,* 9, 302-314.

Tonta, Y. (1992). Analysis of search failures in document retrieval systems: a review. *The Public-Access Computer Systems Review,* 3(1): 4-53, 1992. [Çevrimiçi]. Elektronik kopya: http://info.lib.uh.edu/pr/v3/n1/tonta.3n1 [2 Şubat 2002]

Tonta, Y. (1991, April). A study of indexing consistency between Library of Congress and British Library catalogers. *Library Resources & Technical Services,* 35, 177-185.

Tonta, Y. (1990). Konu erişimi: Kütüphane kataloglarında yapılan konu aramaları üzerine bir deneme. *Türk Kütüphaneciliği,* 4(2): 60-69.

Turner, T.P. ve Brackbill, L. (1998). Rising to the top: Evaluating the use of the HTML META tag to improve retrieval of World Wide Web documents through Internet search engines. *Library Resources and Technical Services,* 42(4), 258-271.

Turtle, H. ve Croft, B. (1991). Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3), 187-222.

Ünver, Ö. ve Gamgam, H. (1999). *Uygulamalı istatistik yöntemler*. 3. bs. Ankara: y.y.

Van Rijsbergen, C.J. (1979). *Information retrieval*. London,: Butterworths. [Çevrimiçi]. Elektronik adres: http://www.dcs.gla.ac.uk/Keith/Preface.html [23 Eylül 2001].

Voorhees, D.E.M. ve Harman, D. (2000). Overview of the Fifth Text Retrieval Conference (TREC-5). E.M. Voorhees ve D.K. Harman (Eds.), *Information Technology: Proceedings of the Eighth Text REtrieval Conference (TREC-8), Gaithersburg, Maryland, November 20-22, 1996* içinde (s. 1-28). [Çevrimiçi]. Gaithersburg, MD: U.S. Department of Commerce. Elektronik adres: http://trec.nist.gov/pubs/trec5/papers/overview.ps.gz [12 Şubat 2002]. (NIST Special Publication 500-238 The Fifth Text REtrieval Conference (TREC 5) Elektronik adres: http://trec.nist.gov/pubs/trec5/t5_proceedings.html [12 Şubat 2002]).

Voorhees, D.E.M. ve Harman, D. (1999). The Text Retrieval Conference (TREC): History and plans for TREC-9. *SIGIR Forum*, 32(2), 12-15. [Çevrimiçi]. Elektronik Kopya: http://trec.nist.gov/ , [27 Ekim 2001]

W3C. (1997). *Extensible Markup Language (XML).* [Çevrimiçi]. Haz. T. Bray, J. Paoli ve C.M. Sperberg-McQueen. Elektronik adres: http://www.w3.org/TR/PR-xml-971208. [25 Aralık 1999].

W3C. (1999). *Resource Desciption Framework (RDF) Model and syntax specification.* [Çevrimiçi]. Hazl. O. Lassila ve R.R. Swick. Elektronik adres: http://www.w3.org/TR/REC-rdf-syntax. [24 Şubat.1999].

Wong, S.K.M. ve Yao, Y.Y. (1990). Query formulation in linear retrieval models. *Journal of the American Society for Information Science*, 41, 334-341.

Wong, S.K.M., Ziarko, W. Raghavan, V.V. ve Wong, P.C.N. (1989). Extended Boolean query processing in the generalized vector space model. *Information Systems*, 14(1), 47-63.

Yao, Y.Y. (1995). Measuring retrieval effectiveness based on user preference of documents. *Journal of the American Society for Information Science*, 46, 133-145.

Yu, C.T. ve Lee, T.C. (1986). Non-binary independence model. Fausto Rabitti (Ed.), *1986 ACM Conference on Research and Development in Information Retrieval, Palazzo dei Congressi, Via Matteotti, 1, Pisa, Italy, September 8-10, 1986* içinde (s. 265-268). Baltimore, MD: Association for Computing Machinery.

Yuwono, B. ve Lee, D.L. (1996). Search and ranking algorithms for locating resources on the World Wide Web. Stanley Y.W. Su (Ed.), *Proceedings of the Twelfth International Conference On Data Engineering February 26 - March 1, 1996, New Orleans, Louisiana* içinde (s. 164-171). Los Alamitos, CA: IEEE Computer Society Press.

DİZİN

Koyu harfle yazılan sayfalar ilgili terimin yoğun olarak geçtiği sayfaları göstermektedir.

A

B

Ç

D

E

F

G

H

I

İ

R

S

T

X

Y