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).