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, adres 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 yayınları çerecek 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.