Bilgisayar Dünyası
Murat Maga

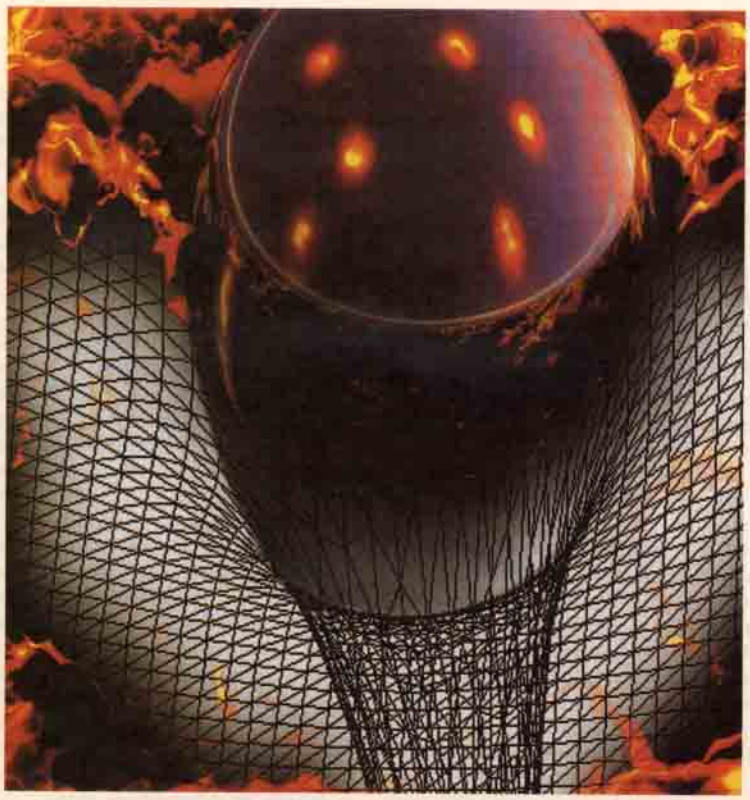
WWW Tarama Araçları

İnternet kullanıcılarının en büyük sorunu yüzbinlerce farklı arşivde tutulan gigabytelarca bilgiye nasıl ulaşabileceğidir. Sayısız arşivdeki bilgilere tek tek bakmak sadece insanın sabır sınırlarını zorlamakla kalmaz, aynı zamanda mümkün de değildir. Bu amaçla İnternet'in ilk günlerinden itibaren arşivlerde tutulan bilgilerin sorgulanabilmesi için onları indeksleyen, Tarama Araçları (Search Engines) adı verilen yazılımlar üretilmiştir. Örneğin, herkesin kullanımına açık FTP arşivlerinde yer alan dosyaları ARCHIE adlı bir program aracılığı ile öğrenebilirsiniz. Dosyanın tam adını ya da adında geçtiği düşünülen sözcükleri verip, ARCHIE aracılığı ile hangi FTP arşivlerinde tutulduğunu öğrenmek mümkün. Benzer şekilde Gopher servislerinde tutulan bilgilerin fihristini tutan VERONICA (Very Easy Rodent Oriented Network Information for Computerized Archives) adlı bir program var. VERONICA, belirli aralıklarla kendisinde kayıtlı bulunan Gopher servislerindeki dizinlerin ve bu dizinlerin içinde yer alan dosyaların listesini çıkarır. Bu liste aracılığı ile yarattığı veritabanıyla da sorgulayıcılara hizmet verir. ARCHIE ve VERONICA dışında WAIS (Wide Area Information Service, Geniş Alan Bilgi Servisi) adında başka bir yazılım var. ARCHIE ve VERONICA'dan farklı olarak, WAIS, sabit bir veritabanı üzerinden sorgulamayı gerçekleştirir. Bu veritabanını WAIS kendisi güncellemez; güncellemenin sistem sorumlusu tarafından gerçekleştirilmesi gerekir.

İnternet'in gelişimine koşut olarak, tarama araçları da gelişti. Yukarıda adı belirtilen bazı tarama araçları birçok yönden kullanıcının istediği esnekliği sağlamıyordu. Özellikle WWW yayıldıktan sonra, kimse VERONICA ya da WAIS gibi karakter tabanlı arayüzlü eski tarama araçlarını kullanmak istemedi. Bu noktadan sonra, WWW üzerindeki bilgilerin sorgulanmasını sağlayan WWW tarama araçları devreye girdi.

WWW'deki tarama araçları, önceki kuşak tarama araçlarına göre çok daha esnek. Tarama için birden fazla sözcük verilebiliyor. Tarama yapılacak alanlar belirlenebiliyor. Ve/veya gibi mantıksal bağlaçlar kullanılabiliyor. Tarama sonunda kaç yanıt istendiği belirtilebiliyor. Sorgulamalar için grafik ortam kullanıldığından işlemler de çok daha rahat.

WWW'deki Lycos, WebCrawler, WWW Worm gibi ilk tarama araçları akademik araştırma projeleri olarak ortaya çıkmıştı. Ancak ortaya çıkan sonuçlar öyle tatmin ediciydi ki, tarama servisleri bir anda İnternet üzerindeki en önemli uğrak yerleri haline geldiler.



Tarama araçlarını yaratmak kuramsal açıdan pek zor gözükmese de, pratikte pek öyle olmadığı görülüyor. Başlangıçta programcısının veritabanında tanımladığı adreslere bakan tarama aracı bu adreslerdeki sayfalarda bulunan metnin indeksini tutuyor. Yine, uğradığı sayfalardaki yeni sayfaların adreslerin -kendi veritabanında yer almıyorsa- kaydediyor. Daha sonra bu yeni adreslere de uğrayıp, aynı işlemleri tekrarlıyor. İlk başta 10 adresle işe başlayan tarama araçları kısa süre içerisinde onbinlere ulaşıyor.

İlk tarama araçları, uğradıkları her sayfadan başlık ya da konu olarak seçtikleri birkaç sözcüğü fihristlerine katmakla yetinirken; yeni kuşak tarama araçları tüm metni indeksine alıyor. Digital firmasının, Alta Vista adlı tarama aracı, aynı anda 1000 tane WWW servisini tarayıp, bulduğu tüm metni veritabanına kaydediyor.

Alta Vista'nın veritabanında yaklaşık 22 milyon WWW sayfası yer alıyor. Bu sayfalarda ise yaklaşık 10 milyar sözcük var. Tüm indeks ise 33 gigabyte tutuyor. Bu devasa veritabanı, Digital firmasının en güçlü bilgisayarlarından birinde tutuluyor. Alta Vista'nın çalıştığı bilgisayarın 6 gigabyte belleği, 10 ana işlemcisi, 210 gigabyte da sabit disk kapasitesi var. Alta Vista bir günde 2 milyonu aşkın tarama isteğine yanıt veriyor. 33 gigabytelık fihrist üzerinde yapılan her bir tarama yarım saniyeden az sürüyor. Tüm bu işlem hacmine karşılık Alta Vista'nın (ve diğer WWW tarama araçlarının) çözemediği önemli bir sorun var. WWW'nin inanılmaz büyüme hızı.

Bazı WWW Tarama Araçlarının Adresleri
http://altavista.digital.com
http://www.inktomi.com
http://guide-p.infoseek.com
http://webcrawler.com
http://wwww.cs.colorado.edu/wwww.html
http://www.lycos.com
http://www.yahoo.com
http://cuiwww.unige.edu/w3catalog
http://www.psu.edu/search.html

Şu anda 100 000 WWW servisinde tutulan 50 milyon sayfa olduğu sanılıyor. Bu sayı, her 9 ayda iki katına çıkıyor. Alta Vista bile, duyurulduğu 15 Aralık 1995 tarihinden bu yana, WWW'nin yarısından azını tarayabilmiş.

WWW tarama araçlarının çözemediği tek sorun bu da değil. WWW üzerinde yer alan her türlü bilginin indeksini tutmak olanak dışı. Şu an için sadece metin dökümanları indekslenebiliyor. World Wide Web'in yeni uzantıları olan VRML ya da JAVA dillerinde hazırlanan belgelerin, ses, görüntü ve canlandırma dosyalarını tarama araçları veritabanlarına alamıyorlar. Aslında, bütün metin dosyalarının da veritabanlarına alınabilindiğinden bahsedemeyiz. İsteğiniz üzerine, bir sorgu sonucunda dinamik olarak üretilen bilgiyi indekslemenin yolu yok. Zaten böyle veritabanları WWW tarama araçlarının "kara listesinde".

Tarama araçları ile ilgili sorunun önemli bir boyutu da, kullanıcıyı doğrudan ilgilendiriyor. Verdiğiniz anahtar sözcüğe göre, bir taramadan on ile onbin arasında değişen miktarda yanıt almanız mümkün. Yanıtların sayısı 25-30'u geçiyorsa, gerçekten aradığınız bilgiye ulaşmak bir sorun haline gelebilir.

Tarama araçları ile ilgili önemli sorunlar bulunsa da, bu servisler şu an için İnternet ve WWW'nin vazgeçilmez parçaları. Daha iyi algoritmalarla çalışan, hızlı araçların yaratılması durumunda daha da iyi hale gelecekler. Ancak özel İnternet şirketlerinin tarama servislerini daha ne kadar "amme hizmeti" olarak götüreceklerini de tartışılır. Kuşkusuz, tarama servisleri bugün İnternet'in en çok ziyaret edilen sayfaları arasında yer alıyor. Bu sayede tarama servisi veren şirketler de WWW sayfalarına aldıkları reklamlardan çok fazla kâr ediyorlar. Ancak taramalardan da sözcük başına para almayı düşünebilirler. Bekleyip, göreceğiz.

Kaynaklar:
New Scientist, 6 Nisan 1996
http://altavista.digital.com