WWW Tarama Araçları

Yakın zamana değin İnternet'te bilgiye ulaşmanın en yaygın yolu WWW'de dolaşmaktı (surf etmek). Bu yolla yapılan taramalarda bilgi rastlantıyla bulunuyordu. Belirli bir WWW sayfasındaki bağlantıları izleyerek sayfadan sayfaya yapılan taramalarda kullanıcı elbet bir süre sonra istediği bilgiye ulaşacağını umuyordu. Eğer zamanınız varsa WWW'de gezinmek aslında eğlenceli bir iş. Çünkü yeni yerler keşfediyorsunuz. Ancak başınızda bekleyen bir patronunuz ya da sınırlı bir süre içinde yetiştirmeniz gereken bir işiniz için bilgiye ulaşmanız gerekiyorsa o zaman durum değişir; sayfadan sayfaya gezinirken tekrar tekrar aynı sayfalara bağlanıyorsanız keşfetmek zevki yiter, dahası kab°sa dönüşebilir.

Günümüzde, WWW'de yayımlanan bilgileri aramak ve bunlara daha verimli bir şekilde ulaşmak için, yeni yeni araçlar geliştirildi. Bu yazıda WWW'de bilgiye ulaşmada kullanılan 2 ayrı yol üzerinde duracağız: Konulara göre aşamalı bir biçimde hazırlanmış bir klasörde tarama, tarama araçlarıyla yapılan sözcük taramaları. Ancak bu iki yol arasında tercih yapma dışında sözcük tarama araçlarından hangisinin seçilmesi gerektiği de önemli.

Tarama Araçları ve Dizinler

Tarama Araçları: HotBot ya da AltaVista gibi tarama araçları kendi listelerini otomatik olarak yaratıyorlar. Bunlar WWW'de dolaşıp İnternet kaynaklarını buluyor, bunları topluyor ve arama yapan kişi de bu bulunan sonuçlar üzerinde aramasını gerçekleştiriyor.

Eğer bir WWW sayfasında birtakım değişiklikler yapılmışsa doğal olarak tarama aracı bu değişiklikleri buluyor. Bunda sayfa başlığı, ana yazı ve diğer sayfa öğeleri önemli bir rol oynuyor. Klasörler: Yahoo gibi bir klasörün yapılandırılmasında ise insanlar rol oynuyor. WWW arşivlerini yaratanlar bu klasörde kendi WWW arşivlerine giden bağlantılar koymanın yanı sıra bu arşivleri anlatan kısa açıklamalarda bulunuyor; ya da bu sayfaları, daha sonradan gezen klasör editörleri kendi yorumlarını ekliyor. Değişen bir WWW arşivi, klasördeki listelemede hiçbir değişikliğe yol açmıyor.

Melez Tarama Araçları: Bazı tarama araçları kendi dizinlerini tutuyorlar. Bir tarama aracının dizininde WWW arşivinizin olması biraz şans biraz da sayfanızın niteliğine bağlıdır. Kimi zaman kendi WWW arşivinizi bu listeye eklenmesi için form doldurursunuz; ancak bu, listeye gireceği anlamına gelmiyor. Daha sonra dizin editörleri bu formlarda belirtilen sayfalara bir göz atıp sadece ilgi çekici bulduklarını listelerine ekliyorlar.

Tarama araçları üç temel öğeden oluşuyor. Bunlardan ilki örümcek (spider) ya da crawler olarak adlandırılan robotlardır. Bunlar bir WWW arşivini gezer, içindekileri okuyarak, bu sayfalardan diğerlerine giden bağlantıları izlerler. Bu robotlar her ay ya da iki ayda bir gibi belli sürelerle bu sayfalara bir değişiklik olup olmadığını denetlemek için bakarlar.

Bu robotların bulduğu her şey tarama araçlarının ikinci bölümü olan dizinlere gider. Katalog olarak da adlandırılan dizinler, örümceklerin bulduğu her WWW arşivinin bir kopyasının bulunduğu büyük bir kitaba benzetilebilir. Eğer bir WWW arşivi değişirse o zaman bu kitapta yeni bilgilerle güncellenmiş olur.

Bazen bu indekslere yeni arşivlerin ya da bu arşivlerde yapılacak değişikliklerin aktarılması biraz zaman alabilir. Bu arşivlere örümcekler gelmiştir ancak daha henüz dizine geçirilmemiştir. Dizinlenene değin bu yeni WWW arşivleri sözcük tarama araçları ile bulunamaz.

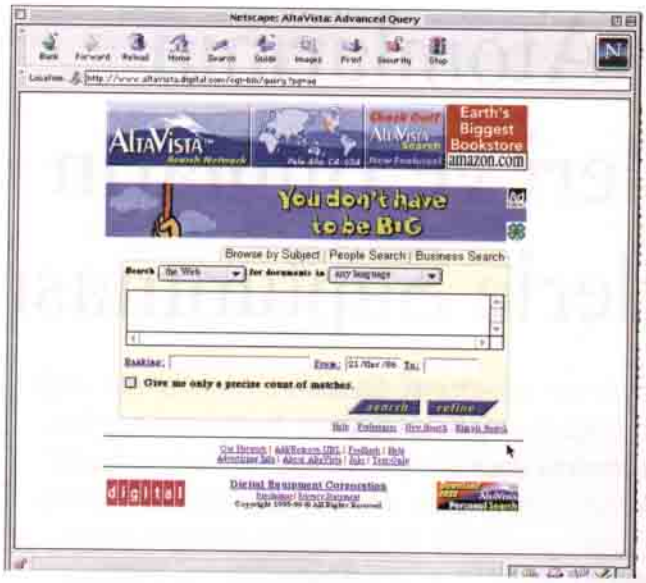
Tarama araçları yazılımları ise tarama aracının üçüncü parçasıdır. Bu dizinlenen milyonlarca sayfa arasından yapılan sorgulamaya uyanları uygunluk sırasına göre belli bir sıralamayla verir.

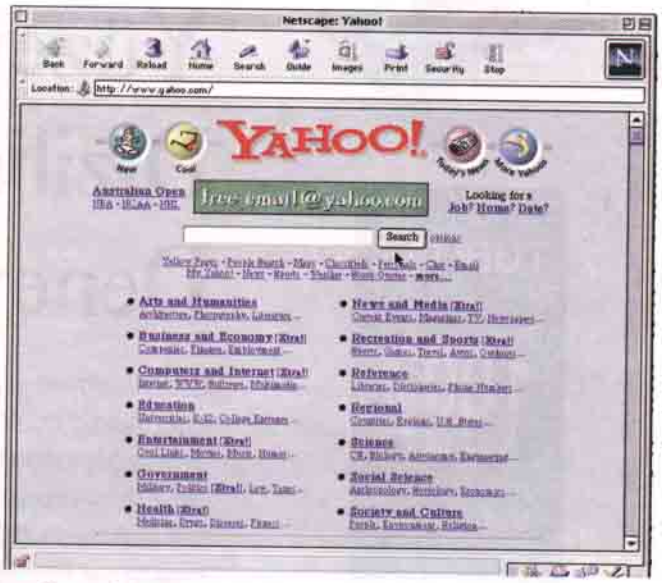
Peki robotlar nereye gideceklerine nasıl karar veriyorlar? Bu kullanılan robota bağlı. Çünkü her robot farklı bir strateji uygular. Genellikle, sunumcu listeleri, "Yeni Ne Var" sayfaları ve WWW'deki en popüler sayfalar gibi birçok yere bağlantısı olan adreslerden başlarlar. Ancak genel olarak birçok WWW dizinleme servisi sizden URL adreslerini girmenizi bekler. Bu girişler belli bir sıra sokulur ve daha sonra dizinin robotu tarafından gezilir.

WWW için zararlılar mı? Eskiden robotlar bilgisayar ağlarını ve sunumcuları için büyük bir sorundu. Ancak günümüzde robotlar daha iyi tasarlanıp, daha profesyonelce kullanıldığından genel olarak sorun yaratmıyorlar.

WWW Arşivlerinin Sıralanması

Bu sıralanan WWW arşivleri elbette ki her zaman doğru olmuyor. Konuyla ilgili olmayan sayfalar da görmeniz mümkün. Bunların arasında hangisinin sizi ilgilendirdiğini bulma bazen





kafa yorucu olabiliyor. Bunun nedeni tarama araçlarının insanlar gibi belli bir yargılarının ve geçmiş deneyimlerinin olmaması. Ancak bu hizmetleri verenler gelişen teknolojiyle bu açıklarını kapatmaya çalışıyorlar.

Peki bu tarama araçları sorgulama sonucunun uygunluğunu nasıl belirliyor? Bunlar, sorgulamadaki girilen sözcüklerin WWW arşivindeki konumu ve kullanılma sıklığı gibi belli kurallar doğrultusunda gerçekleştiriyorlar aramalarını.

Sayfa başlığındaki sözcükler, sorgulamada girilen sözcüklere göre uygunluk derecesi en fazla olan WWW arşividir. Bundan başka tarama aracı, başlıklar ya da ilk birkaç paragraftaki sözcükler gibi sayfanın en üstündeki öğelere bakar.

Kullanılan sözcük sıklığı ise tarama araçlarının göz önünde bulundurduğu bir başka etken. Sorgulama sırasında girilen sözcükler bir sayfada ne kadar çok kullanılırsa o sayfa sorgulama sonundaki listelemede o kadar çok üst sıraya çıkar.

Ancak tarama araçlarının yukarıdaki sorgulamalar dışında başka özellikleri de var. Zaten bu yüzden yapılan bir sorgulamaya her tarama aracı farklı farklı cevap veriyor. Aralarındaki farklardan biri kimi tarama araçlarının daha fazla WWW arşivi dizinlemesi. Bununla birlikte hiçbir tarama aracı da aynı WWW arşivi koleksiyonuna sahip değil.

Bazı tarama araçları kimi arşivleri listeleme sonucunda daha üst sıralara taşıyabilir. Bunun nedeni kimi WWW arşivlerinin daha popüler olması. Tarama araçları bir arşivin popülerliğini, dizininde bulundurduğu arşivlerinden hangi arşive ne kadar bağlantının gittiğini ölçmesiyle belirler.

Kimi melez tarama araçları da, aynı zamanda klasör bulunduranlar, klasörlerinde bulundurdukları ilgi çeken sayfaları daha ön plana çıkarırlar.

Bütün bunların yanında "meta tag"ları kullanan WWW arşivlerini HotBot ve Infoseek gibi tarama araçları, kendi sorgulama sonuçlarında uyumluluk açısından daha üst sıralara taşır. Bunun dışında Excite meta tag'leri sıralandırmasında göz önünde bulundurmaz.

Bu arada tarama araçları kimi sayfaları dizin dışı tutarlar. Bu sayfalarda kimi sözcükler yüzlerce kez kullanılmıştır ki bu şekilde yapılan arama sonucundaki sıralamada bu sayfalar üst sıralara çıkabilsin.

Peki bu tarama araçlarından hangisi en iyisi? Aslında bu tamamen sizin ne istediğinize bağlı. Çünkü yukarıdaki nedenlerden dolayı birbirlerine göre büyük farklılıklar gösteriyorlar. Ancak dizinlerinde bulundurdukları sayfaya baktığımızda AltaVista'nın 100 milyon sayfayla, HotBot (80 milyon) ve Excite'ın (55 milyon) önünde olduğunu görüyoruz. AltaVista, HotBot ve Lycos'un günde taranan sayfa sayısı ise 10 milyon civarında. Sayfaların tazelenme süresi ise AltaVista'da 1 günden 3 aya, Excite 1-3 hafta, HotBot 1 günden 2 haftaya, InfoSeek dakikadan 2 aya kadar, Lycos'da ise 1-2 hafta civarında.

Bu arada Media Metrix ve Relevant Knowledge şirketlerinin araştırmaları sonucu yapılan sıralamaya göre en üst sırada bulunan Yahoo en yakın takipçisi olan Excite'dan iki kat daha fazla ziyaretçi çekiyor. Ziyaretçi sayısı yönünden sırasıyla InfoSeek, Lycos, AltaVista ve WebCrawler bunları izliyor. Elbette ziyaretçi sayısı bir tarama aracının en iyisi olduğunu göstermiyor. Ancak çoğu bilgisayar dergisi bu tarama araçları arasından kendi dergilerinde en iyisini seçiyor. Bu ya derginin editörleri ya da okuyucular tarafından seçiliyor. Sonuçlara baktığımızda dergiler tarafından en çok beğenilen tarama aracı HotBot görünüyor.

Kullanılan Teknoloji

Bu tarama servislerini vermek elbette kolay bir iş değil. Bunun için, en geniş dizin kapasitesine sahip AltaVista 16 tane AlphaServer 8400 sunumcusundan yararlanıyor. Bunların her birinde yaklaşık 8 GB bellek var (kaba bir şekilde 500 tane 16 MB belleğe sahip kişisel bilgisayar diye düşünülebilir). Bu arada yapılan sorgulamalara hızlı cevap verebilmek aynı zamanda hızlı İnternet bağlantısı da gerektiriyor. Örneğin AltaVista saniyede 100 megabit ile DIGITAL Palo Alto gateway üzerinden İnternet'e bağlı.

Alkım Özaygen

Kaynaklar
http://magi.com/~mmelick/it96jan.htm
http://info.webcrawler.com/mak/projects/robots/faq.html
http://www.searchenginewatch.com
http://altavista.digital.com