Arama Motorları

Hayri Sever[#] ve Yaşar Tonta[##]

0 Giriş

Internet kullanıcıları istedikleri bilgilere erişmek için çoğu zaman arama motorlarını kullanmaktadırlar. Herkesin erişimine açık milyarlarca web sayfasını keşfeden, dizinleyen ve kullanıcıların hizmetine sunan Google, Alta Vista, Yahoo!, Superonline, Arabul, e-kolay gibi arama motorları hakkında son yıllarda yoğun araştırmalar yürütülmektedir. Arama motorları bilgi erişim sistemlerinin bir alt konusudur. Bu nedenle çalışmamızda önce bilgi erişim sistemleri kısaca gözden geçirilmekte, daha sonra arama motorları ayrıntılı olarak incelenmektedir.

1 Bilgi Erişim Sistemleri

Bir bilgi erişim sisteminin temel işlevi, kullanıcıların bilgi ihtiyaçlarını karşılamak amacıyla derlemdeki (collection) ilgili (relevant) belgelerin tümüne erişmek, ilgisizleri (non-relevant) de ayıklamaktır (Tonta, Bitirim ve Sever, 2002: 9). Bilgi ihtiyacı doğal dille ifade edilebileceği gibi, dizin terimleri ve bu terimler arasındaki ilişkiler ("ve", "veya", "ve-değil") çerçevesinde de tanımlanabilir. Sorgu dili aracılığı ile tanımlanan bilgi ihtiyacı içerik (özelde arama) belirteçleri ile gösterilir. Bu durum Şekil 1'de verilmektedir. Benzer şekilde, metin nesneleri (belgeler) arka planda işleyen otomatik dizinleme sürecine giriş oluşturur. Metnin adı, yazarı, yayın tarihi, konusu, vb. gibi bilgiler bir belgenin niteliklerini oluşturur. Belgeleri temsil eden terimler ve belge numaraları ters dizin kütüğü (inverted file) halinde düzenlenir.

Kullanıcının bilgi ihtiyacı ve belgeler temsil edilirken (representation): (1) harf olmayan karakterler boşluklarla yer değiştirilir; (2) tek harfli sözcükler silinir; (3) bütün karakterler küçük harfli yapılır; (4) dur listesinde (stop list) adı geçen sözcükler silinir; (5) sözcükler gövdelenir (stemming); ve (6) tek karakterli gövdeler atılır. Dur listesindeki sözcüklerin erişim değeri yoktur. Bu tür sözcükler eldeki derlemden bağımsız olarak oluşturulacağı gibi derlemdeki yüksek sıklıklı terimlerden de seçilebilir. Türkçe gibi sondan eklemeli (agglutinative) dillerde gövdeleme (bir sözcükten çekim eklerinin atılıp, yapım eklerinin korunması) önemlidir. Nitekim Türkçe için geliştirilen bir gövdeleme algoritması (GÖVDEBUL) kullanılarak yapılan çeşitli deneylerde gövdelemenin bilgi erişim performansını %20-%25 civarında artırdığı gözlenmiştir (Duran, 1999; Sezer, 1999; Eroğlu, 2000).

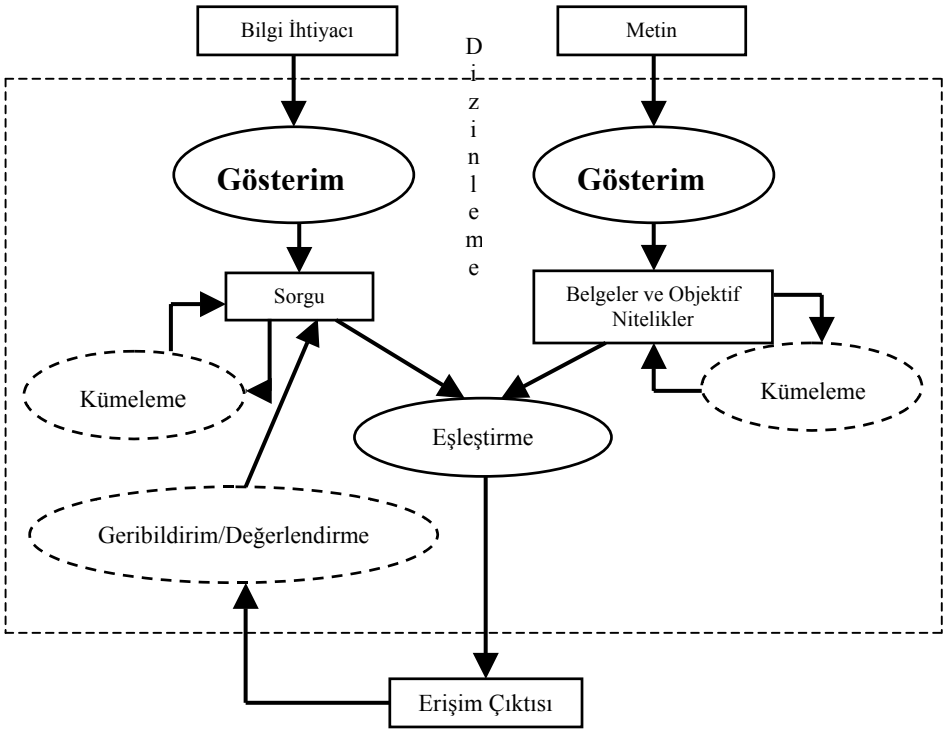
Şekil 1'de verilen kümeleme (clustering) sürecinde belgeler, önceki sorgular kümesi, belge derlemi ve içerik belirteçlerine dayanarak kendi aralarında öbeklendirilir. Farklı amaçlar için farklı kümeleme teknikleri kullanılır. Örneğin, sorgu genişletmek (query expansion) için kullanılan kümeleme tekniğinde içerik belirteçleri eş anlamlılık temelinde öbeklendirilir; ilgili sorgu her eş anlamlı arama belirteci ile ayrı ayrı işletilip erişilen sonuçlar birleştirilir. Öte yandan bir metni daha az sayıda belge belirteciyle göstermek ve dolayısıyla yerden kazanç sağlamak için de kümeleme tekniği kullanılabilir. Belgelerin kümelendirilmesinde amaç bilgi ihtiyacını daha çabuk karşılamaktır. Sorguların kümelendirilmesinde ise amaç, zaman açısından pahalı bir süreç olan geribildirim (feedback) sürecine olan ihtiyacı azaltmak, arama süresini kısaltmak (Raghavan ve Sever, 1995) ve performans etkinliğini artırmaktır (Belkin, Kantor, Fox ve Shaw, 1995).

[#] Bilgisayar Mühendisliği Bölümü, Başkent Üniversitesi, 06530 Bağlıca, Ankara. E-posta: sever@baskent.edu.tr, Web: www.baskent.edu.tr/~sever.

[##] Bilgi ve Belge Yönetimi Bölümü, Hacettepe Üniversitesi, 06532 Beytepe, Ankara. E-posta: tonta@hacettepe.edu.tr, Web: yunus.hacettepe.edu.tr/~tonta/.

Belgeler ve sorgular ortak bir gösterim (representation) dili kullanılarak ifade edilir. Bir erişim fonksiyonu olarak eşleştirme (matching) eldeki belgede geçen terimlerin sorguda geçip geçmediğine bakarak belge ile sorgu arasındaki yakınlığı derecelendirir. Bu derecelendirme sadece 0 (tamamen ilgisiz) ve 1'den (tamamen ilgili) oluşan ikili bir değer alabileceği gibi, kesirli bir sayı da olabilir (az ilgili, çok ilgili).

Bilgi erişim sistemleri için vektör uzayı, olasılıksal, mantıksal, Bayesian ağ, vb. gibi erişim modelleri kullanılır (Salton, 1971; Robertson 1997; McCune, Tong, Dean ve Shapiro,



Şekil 1. Bir bilgi erişim sisteminin işlevsel mimarisi

1985; Rolleke, 1999; Turtle ve Croft, 1992). Erişim fonksiyonu erişim modelinin bir bileşenidir. Örneğin, vektör uzayı modeli erişim fonksiyonu olarak çoğunlukla kosinüs korelasyonunu kullanır. Bayesian ağda olasılıkların yayılımında Bayes Kuramına sıkı bir bağlılık vardır. Benzer şekilde, mantıksal modelde "horn" cümleciklerinin "modus ponens" ilkesine göre yayılımı vardır. Çok sık kullanılan Boole modelinde ise sorgu ifadesi aracılığı ile tanımlanan terimler arasındaki ilişkinin ilgili belgede olup olmadığını küme cebiri kuramı kullanılarak belirlenir.

Erişim çıktısı eldeki sorgu ifadesinin belgeler ile eşleştirilmesiyle oluşturulur. Erişim çıktısında belge derlemi içinde sorgu ifadesi ile ilgili olduğu "düşünülen" belgeler kullanıcının bilgi ihtiyacına yakınlık derecesine göre azalan sırada sıralanırlar.

Geribildirim, Şekil 1'de görüldüğü üzere, bir bilgi erişim sisteminin performans etkinliğini artırma sürecidir. Bu süreç genellikle erişim çıktısında yer alan belgelerin kullanıcının bilgi ihtiyacını karşılamadığı durumlarda kullanılır ve performansın etkileşimli bir süreç sonucunda kabul edilebilir bir düzeye çekilmesi hedeflenir (Salton ve Buckley, 1990).

2 Arama Motorları

Arama motorları son elli yıldır geliştirilmekte olan bilgi erişim sistemlerini temel almaktadır. Bununla birlikte, arama motorları gerek mimari açıdan gerekse işlevsel özellikleri açısından bilgi erişim sistemlerinden bazı farklılıklar gösterir.

2.1 Robot

Arama motorlarının esas bileşenlerden birisini örümcek (spider) modülü oluşturur. "Robot" adı verilen bu modül web sayfalarını keşfeder, yerel veri tabanına indirir ve bu sayfaların dizinleme modülü tarafından analiz edilmesini sağlar. Tipik bir robot başlangıç adresi verilen hiper-metin veri tabanını önce enlemesine dolaşır ve her bir düğüme (node) iliştirilen Internet kaynağını dizinlenmek üzere yerel diske yerleştirir (Arasu ve diğerleri, 2001). Internet kaynaklarının sayısı ve değişim hızı indirilecek sayfaların seçimi ve ilgili sayfaların tazelenme sıklığı hakkında bilgi sahibi olunmasını gerektirmektedir.

2.2 Dizinleme

Geleneksel bilgi erişim sistemlerinde dizinlenecek belgeler durağandır; bir belge bir defa dizinlendikten sonra bir daha dizinleme işlemine tabi tutulmaz. Halbuki web kaynakları çok hızlı değişmektedir. Internet ortamındaki bir bağlantının (link) ortalama ömrü 44 gündür (Brake, 1997; Kahle, 1997). Internet'teki bilgi hacminin üssel olarak büyümesi ve Web sayfalarının yarı-ömürlerinin (half-life) günlerle ifade edilmesi mimariyi daha da karmaşıklaştırmaktadır. Arama motorları giderek toplam Internet kaynaklarının daha azını dizinleyebilmektedir. Farklı arama motorları farklı kaynakları dizinlediklerinden, dizinlenen sayfaların çakışma oranlarını tahmin etmek de güçleşmektedir. 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).

Geleneksel yaklaşımda, belgelerin yazım kalitesi (ya da metin değeri) oldukça yüksektir. Halbuki, Web sayfalarında yapılan yazım hataları bir istisna olmanın çok ötesindedir. Başka bir sorun ise tekrarlı sayfaların yüzdesinin giderek artmasıdır. Bir araştırmaya göre Web sayfalarının %30'u tekrarlardan oluşmaktadır. Tekrarlı sayfaların tanınması ve yalnızca bir kez dizinlenmesi birçok araştırmaya konu olmuştur (örneğin, Kobayashi ve Takeda, 2000; Kirsch, 1998).

2.3 Belgelerin Gösterimi

Arama motorları dizinleme hacmini azaltmak için genellikle bir belgeyi tümüyle dizinlemez (Kobayashi ve Takeda, 2000; Laursen, 1998). Tipik olarak, bir Web sayfasının başlık kısmı, üst veri (metadata) belirteçlerinin içerikleri, tam metnin ilk bir-iki paragrafı dizinlenir. Web sayfalarının arama motorlarına hitap eden kısmıyla ilgili ilk adım, HTML 3.2 standardı ile atılmıştır. HTML kodunun başında bulunan ve `<head> . . . </head>` alanı ile sınırlanan üst veri belirteçleri tamamen robotlara hitap etmektedir. Dizinlemede kullanılan ve arama motorlarına hitap eden "tanım", "anahtar sözcük" gibi toplam 16 belirteç Dublin Core Üst Veri Seti'nde tanımlanmıştır (dublincore.org/documents/dces/).

2.4 Sıralama ve Bağlantı Analizi

Daha önce de belirtildiği üzere, sorgu makinesi kullanıcının doğal dilde girdiği soru cümlesine dayanarak sorgu ifadesini ya oluşturur ya da kullanıcıdan elde eder. Eşleştirme sonucu sorguyla ilgili olduğu "düşünülen" belgeler azalan önem sırasında kullanıcıya erişim çıktısında sunulur. Sorgu makinesi bu işlevi bir ya da daha fazla erişim fonksiyonu kullanarak gerçekleştirir. Arama motorları durağan belgeler üzerinde çalışan geleneksel bilgi erişim sistemlerinden farklı olarakhiper-metin veri tabanında mevcut sayfalar arası ilişkiler (etiketler, gelen/giden bağlantılar) hakkında da bilgi toplamakta ve her sayfanın içinde bulunduğu çizgenin yapısal/cebrik özelliklerini kaydetmektedirler.

Büyük arama motorları hem ticari sır olması açısından hem de "spam"a yol açmamak için başvurdukları erişim fonksiyonlarını ve dizinleme tekniklerini açıklamamaktadır. Ancak söz konusu arama motorları çoğunlukla akademik ortamda geliştirildikleri için, kullandıkları erişim fonksiyonları tahmin edilebilmektedir. Örneğin, Infoseek arama makinesi

Massachusetts Üniversitesi tarafından geliştirilen INQUERY bilgi erişim sisteminin ticari sürümüdür. Infoseek'de ilgililik hesaplaması kısmen sayfanın başka sayfalar tarafından ne kadar sıklıkla referans verildiğine 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). Alta Vista 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). Kavramlar terimlerin kümelendirilmesine (çevrimiçi eş anlamlı sözlük) dayanır.

Erişim fonksiyonunda bir sorgu ile belge arasındaki benzerlik çeşitli biçimlerde hesaplanabilir. Örneğin, hem sorguda hem de belgede geçen ortak terimler temel alınabilir. Ya da bir belgeyi oluşturan yapısal bileşenlere (başlık, anahtar sözcükler, özet, tam metin, vb. gibi) farklı ağırlıklar verilebilir. Belge başlığında geçen bir terim, belgenin konusunu belirlemede daha ağırlıklı olarak değerlendirilebilir. Erişim fonksiyonu çeşitli belge bileşenlerinin sorgu ile benzerliklerinin toplamı olan bir polinom şeklinde düşünüldüğünde, başlık bileşeninin sorgu ile benzerliği belgenin tam metniyle benzerliği ile aynı kefeye konmayabilir. Ortak bir veri tabanı (ya da belge derlemi) üzerinde farklı erişim modelleri çalıştırılarak elde edilen sonuçların birleştirilmesiyle erişim performansının büyük ölçüde arttığı gözlenmiştir (Lee, 1997, 1995).

2.5 Etkinlik

Genelde bilgi erişim sistemlerinin, özelde arama motorlarının performans etkinliği tipik olarak *anma*, *duyarlık* ve *posa* (ya da yanlış alarm) ölçütleri ile ölçülür. Bu ölçütleri Tablo 1'de verilen ikili sınıflamaya dayanarak tanımlamak mümkündür. Belirli bir soruya karşılık sistem tarafından derlemden erişilen belgeler ilgili ve ilgisiz olmak üzere ikiye ayrılır. Örneğin, '*a*' erişilen ve kullanıcının ilgili bulduğu belge sayısını, '*b*' erişilen ancak kullanıcının ilgisiz bulduğu ("false drops") belge sayısını, '*a*+*b*' ilgili ya da ilgisiz erişilen toplam belge sayısını, '*a*+*c*' ise derlemdeki erişilen ya da erişilemeyen toplam ilgili belge sayısını verir. Buna göre *anma*, sistem tarafından erişilen ilgili belgelerin (*a*) derlemdeki toplam ilgili belgelere (*a*+*c*) oranını verir (Van Rijsbergen, 1979: 10) (bkz. Tablo 1). *Duyarlık*, sistem tarafından erişilen ilgili belgelerin (*a*) erişim çıktısında yer alan (ilgili ve ilgisiz) toplam belgelere (*a*+*b*) oranını verir (Van Rijsbergen, 1979: 10). Anma ve duyarlık değerleri 0 ile 1 arasında değişmektedir. Bu değerler ne kadar yüksek olursa bir bilgi erişim sisteminin etkinliği de o kadar yüksektir (Salton, 1989). Posa ise, sistem tarafından ilgili olduğu varsayılıp erişilen (*b*) ve fakat gerçekte ilgisiz olan belgelerin toplam ilgisiz belgelere (*b*+*d*) oranını verir. Bu oran "bir sistemin ilgisiz belgeleri ne derece sağlıklı olarak reddettiğini ölçer" (Blair, 1990: 116).

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

3 Sonuç

Hızlı bilgi artışıyla başa çıkmaya çalışan Internet kullanıcılarının durumu "yangın hortumundan su içmeye çalışan" kimselere benzetilmektedir. Arama motorları kullanıcılara bu konuda yardımcı olan en önemli bilgi keşfetme ve erişim araçlarıdır. Kullanıcıların işlerini daha da kolaylaştırmak amacıyla arama motorlarıyla ilgili olarak çeşitli konularda

(belge keşfetme, dizinleme, arayüz tasarımı, gövdeleme ve erişim algoritmaları, performans ölçümü vb. gibi) yapılan araştırmalar sürdürülmelidir.

4 Kaynakça

Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A., Raghavan, S. (2000). Searching the Web. [Çevrimiçi]. Technical Report. Stanford University. http://dbpubs.stanford.edu:8090/pub/2000-37 [30 Haziran, 2003].

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 [30 Haziran 2003].

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

Brake, D. (1997). Lost in Cyberspace. *New Scientist Magazine* [Çevrimiçi]. Elektronik adres: http://www.newscientist.com/ ; http://www.well.com/~derb/lost.html [30 Haziran 2003].

Duran, G. (1999). *GövdeBul: Türkçe gövdeleme algoritması.* (Yayımlanmamış yüksek lisans tezi), Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

Eroğlu, M. (2000). *Gövdelemenin ve gömünün Türkçe bir bilgi erişim sistemi üzerindeki etkisinin araştırılması.* (Yayımlanmamış yüksek lisans tezi), Hacettepe Üniversitesi Fen Bilimleri Enstitüsü.

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.

Kahle, B. (1997, March). Preserving the Internet. *Scientific American* [Çevrimiçi] 276(3), 82-83. Elektronik adres: http://www.hackvan.com/pub/stig/articles/trusted-systems/0397kahle.html [30 Haziran 2003].

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

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: ACM.

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

Laursen, J.V. (1998, February/March).Somebody wants to get in touch with you: Search engine persuasion. *Database*, 21(1): 42-46.

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 [30 Haziran 2003].

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.

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.

Raghavan, V.V. ve Sever, H. (1995). On the reuse of past optimal queries, In: *Proceedings of 18th ACM International Conference on Research and Development in Information Retrieval (SIGIR'95), Seattle, WA, USA, July 1995*, pp. 344-351.

Robertson, S.E. (ed.), (1997). Special issue on OKAPI, *Journal of Documentation*, 53 (1).

Rolleke, T. (1999). POOL: *Probabilistic object-oriented logical representation and retrieval of complex objects; A model for hypermedia retrieval*. (Unpublished Ph.D. thesis), University of Dortmund.

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.

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.

Tonta, Y., Bitirim, Y. ve Sever, H. (2002). *Türkçe arama motorlarında performans değerlendirme.* Ankara: Total Bilişim.

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

Van Rijsbergen, C.J. (1979). *Information retrieval*. London,: Butterworths. [Çevrimiçi]. Elektronik adres: http://www.dcs.gla.ac.uk/Keith/Preface.html [30 Haziran 2003].