Türkçe Metinlerde Bilgi Erişimi [1]

Fazlı Can, Seyit Koçberber
Bilkent Bilgi Erişim Grubu, Bilgisayar Mühendisliği Bölümü, Bilkent Üniversitesi
canf@cs.bilkent.edu.tr, seyit@bilkent.edu.tr

12 Haziran, 2007

Özet
Bu çalışmada "bilgi erişim" (BE) konusunda Türkçede şimdiye dek yapılmış olan en büyük ölçekli araştırmanın sonuçları verilmekte ve Türkçe için geliştirilecek BE sistemlerinin etkinliğini arttırmada izlenebilecek yaklaşımlar önerilmektedir. Başka araştırmacılarla paylaşmak ve bu konuda yapılacak araştırmaları desteklemek ve özendirmek amacıyla geliştirilen deney derlemi 400 binin üstünde belge ve 72 sorgu içermektedir. Çalışmada, Türkçe belgelere erişimde sözcük kökü kullanımının, sözcükleri olduğu gibi kullanmaya göre önemli ölçüde sistem etkinliğini yükselttiği gösterilmektedir. Derlem indeksi oluşturulurken kök olarak sözcüklerin baştaki harflerini kullanan sözcük kırpma gibi basit bir yaklaşımın, biri sözcüklerin biçim birimlerini öteki sözcüklerin yapısını sayısal olarak inceleyerek kök bulan iki yönteme yakın erişim etkinliği sağladığı gösterilmektedir. Sorgu ve belge uzunluğunun, derlem büyüklüğü artışının ve indeksleme sırasında Türkçede çok kullanılan sözcüklerin kullanımının Türkçe BE sistemlerinin başarımındaki etkileri incelenmektedir. Makalede ayrıca bilgi patlaması kavramı ve Türkçe için BE konusunda şimdiye dek yapılmış olan çalışmalar da kısaca tanıtılmaktadır.

Anahtar Sözcükler
Başarım ölçümü, bilgi erişim, eklemeli diller, deney, deney derlemi yapımı, ölçeklenebilirlik, sözcük kökü bulma, Web.

GİRİŞ

İnsanoğlu yüzyıllar boyu gerçeğe, mutluluğa, güvene ve bolluğa ulaşmak amacıyla bilginin peşinde koşmuştur. Günümüzde, gerek bilgi kaynaklarının sayısı gerekse bu kaynaklarca üretilen bilginin miktarı önemli boyutlara ulaşmış durumdadır. Örneğin, 2003 yılında kişi başına üretilen bilgi miktarının yaklaşık 800MB civarında olduğu tahmin edilmektedir (Varian, 2005). Bu miktarın önemli bir bölümünü gazete haberleri gibi metinler oluşturmaktadır.

Bilgi bolluğu sonucunda çok sayıdaki bilgi kaynağına ve bilgiye dikkat etmek zorunda kalan insan için, karşılaştığı seçeneklerin çokluğu bir gerilim kaynağına dönüşmekte ve bilgi varlığıyla peşinde koşan insanı bunaltmaktadır. Peşinde koşulan bilginin bolluğu, başka bir deyişle "bilgi patlaması", uzun bir zamandan beri insanlar için bir sorun olmuştur (de Solla Price, 1963; Toffler, 1970). Yirminci yüzyılın ikinci yarısında hesaplama biliminin seçkin adlarından Donald E. Knuth insanoğlunun bu konudaki çaresizliğini şu şekilde ifade etmiştir: "Bazen gerçekte kullanabileceğimizden çok daha fazla bilgiyle baş etmemiz gerekebilmektedir; bu durumda yapılabilecek en akıllıca şey, bunun büyük bir kısmını yok saymak ya da yok etmektir." (Knuth, 1973, s. 389).

Bilgi patlaması, önemli miktarı metin halinde olan bilgilerin elektronik ortamda saklanmasını, gerektiğinde bulunup ihtiyaç sahiplerinin dikkatine sunulmasını ve bu işlemin

etkin ve verimli biçimde yapılmasını zorunlu kılmaktadır. İşte bilgi erişim (BE) sistemleri bu işlevi yerine getirmeyi amaçlar. Bunu gerçekleştirmek için bilgi içeren nesnelerin bilgisayar ortamında gösterimini, depolanmasını, organizasyonunu ve ulaşımıyla ilgili olan problemlerin ortak bir çözümünü sağlar. BE sistemleri, kullanıcılar tarafından girilen bir sorguya en çok benzeyen belgeleri sezgisel bir hesaplama yöntemiyle saptar. Sistem tarafından indekslenmiş olan belgeler, girilen sorguya benzerlik sırasına göre kullanıcının dikkatine getirilir (Zobel, Moffat, 2006). Bilgi erişimde "etkinlik", bulunan belgelerin kullanıcının bilgi ihtiyacına uygun olmasını, "verimlilik" ise bulma işleminin hızlı bir biçimde yapılmasını ifade etmektedir. Verimlilik kavramının ikinci öğesi olan bellek kullanımı, bu konuda teknolojide sağlanan gelişmeler nedeniyle her geçen gün önemini biraz daha yitirmektedir.

Makalede, bilgi patlaması problemini ve bilgi erişimi sürecini anlatıldıktan sonra bilgi erişimi konusunda Türkçe için şimdiye dek yapılmış olan çalışmalar kısaca tanıtıldı. Laboratuvar ortamında yapılan BE çalışmalarında kullanılan deney derlemleri bir dizi belge, insanlar tarafından hazırlanmış sorgular ve bu sorgulara karşılık gelen ilgili belgeleri içerir (Sparck Jones, 1981). Makalede Bilkent Üniversitesi Bilgi Erişim Grubu olarak başka araştırmacılarla paylaşmak amacıyla oluşturduğumuz deney derlemi tanıtılıp daha sonra bu derlemle yapmış olduğumuz deneyler ve gözlemlerimiz ayrıntılı olarak anlatıldı. Bu bağlamda, derlem indeksi oluşturulurken kök olarak sözcüklerin baştaki harflerini kullanan sözcük kırpma gibi basit bir yaklaşımın, biri sözcüklerin biçim birimlerini öteki sözcüklerin yapısını sayısal olarak inceleyerek kök bulan iki yönteme yakın erişim etkinliği sağladığı gösterilmektedir. Buna ek olarak, sorgu ve belge uzunluğunun, derlem büyüklüğü artışının ve indeksleme sırasında Türkçede çok kullanılan sözcüklerin kullanımının Türkçe BE sistemlerinin başarımındaki etkileri incelenmektedir.

BİLGİ PATLAMASI VE BİLGİ ERİŞİM SİSTEMLERİ

Bilgi patlaması kavramı bilim dünyasında uzun bir zamandan beri gözlenen bir olgudur. Vannevar Bush 1945 yılında çokça atıfta bulunulan ve klasikleşmiş olan "As we may think" ("Düşünebileceğimiz gibi") adlı makalesinde kullanılabileceğinin çok üstünde bir hızla bilgi üretildiğini vurgulamıştır. O yıllarda ABD'de altı bin bilim adamının yönetimini üstlenmiş olan Bush, bilimin savaş için kullanımı üzerinde çalışmaktaydı. Kişisel olarak ulaşılması gereken bilgilerin çokluğu, Bush'un "Memex" adlı mekanik bir özel dosyalama ve kütüphane sistemi tasarlamasına neden olmuştu. Bu sistemin amacı bir kişiye ait bütün kitap, kayıt, ileti, resim, yazışma gibi belgelerin depolanması ve erişimini sağlamaktı. Bu sistem, sonradan geliştirilecek olan BE ve "hypertext" sistemlerinin ilk habercisi olarak da değerlendirilmiştir.

Günümüzde sıradan insanlar bile kişisel yaşamlarında bir bilgi patlaması yaşamaktadır. Bilgisayar ortamının yarattığı kolaylık, kişisel düzeyde de fazla miktarda bilginin üretilmesine yol açmaktadır. Bu kişisel bilgilerin kolayca ulaşılabilir halde tutulmasını ve kişisel düzeyde yaşanan her şeyin anımsanmasını amaçlayan sistemlerin tasarlandığı, hatta gerçekleştirildiği görülmektedir (Gemmel, Bell, Lueder, 2006).

Günümüzde bilgi patlamasının en kolay izlendiği ortamın Web olduğunu söylemek yerinde olacaktır. Web'deki bilgilerin toplam büyüklüğü ve arama motorlarının Web'deki bilginin ne kadarını indekslediklerine ilişkin olarak hemen her gün bir öncekinden daha büyük sayılar görülmektedir (searchenginewatch.com). Daha önce de belirttiğimiz gibi BE sistemleri bu bilgi bolluğunun yarattığı problemli ortamda bir manada cankurtaran durumuna gelmiştir. Fakat, BE

sistemleri de bilgiye erişimi kolaylaştırması nedeniyle bilgi üretimini teşvik etmekte ve dolaylı olarak yangına körükle gitmektedir.

BE konusunun 1950 yıllarından itibaren bilgi bilimlerinin (information sciences) ve bilgisayar bilimlerinin bir dalı durumuna gelmiş olduğu görülmektedir. İlk kez 1951 yılında bilgi bilimlerinin öncülerinden Calvin Moores "bilgi erişim" ("information retrieval") deyimini kullanmıştır ve Gerard Salton bu alanın babası olarak tanımlanmıştır (Saracevic, 1999).

TÜRKÇEDE BİLGİ ERİŞİM ÜZERİNE YAPILAN ARAŞTIRMALAR

Literatüre bakıldığında, Türkçe de dahil olmak üzere, İngilizce dışındaki dillerde BE konusunda az sayıda araştırma yapıldığı görülmektedir (Can ve diğer., 2007). BE çalışmalarında indeksleme amacıyla sözcük köklerinin kullanılması uzun zamandan beri önemli bir konu olarak yer almaktadır (Frakes, Baeza-Yates, 1992). Türkçe eklemeli bir dil olduğu için, BE uygulamalarında indeksleme sırasında sözcük kökü kullanmanın sistemin başarısını etkilemesi doğal bir beklentidir. Bu beklentinin sonucu olarak, Türkçe için yapılan çalışmalarda kök bulma ve başarıma etkileri BE araştırmalarının en belirgin konusudur.

Türkçe BE konusunda ilk çalışma Aydın Köksal (1981) tarafından bilgisayar bilimlerine ilişkin 570 belge ve 12 sorgudan oluşmuş bir deney derlemiyle gerçekleştirilmiştir. Köksal çalışmasında indeksleme sırasında sözcük kökü olarak her kelimenin ilk birkaç karakterini kullanmış ve deneyler sonucunda en iyi etkinliğin ilk beş karakteri kullanarak ortaya çıktığını gözlemlemiştir.

Solak ve Can ise (1994), 533 gazete haberi ve 71 sorgu kullanarak yaptıkları çalışmada sözcük sonundan her harf atılışını izleyen bir yapı analizi ile sözcük köklerini saptamayı hedeflemişlerdir. Yaklaşımın etkinliği sistem tarafından bulunan ilk 10 ve 20 belgede gözlenen duyarlılıkla (İ10D, İ20D) ölçülmüştür. Sözcükleri oldukları gibi doğrudan kullanarak yapılan indeksleme yaklaşımı yerine sözcük köklerinin indekslenmesi, bu çalışmada da kullanılan yedi farklı sıralama fonksiyonuyla yüzde 0 ila 9 arasında etkinlik artışı sağlamıştır. Bu çalışmada duyarlılık, bakılan belgelerin içindeki kullanıcının bilgi gereksinimini karşılayan belge sayısının bakılan belge sayısına oranını göstermektedir.

Ekmekçioğlu ve Willett'in çalışmasında ise (2000), 6289 haber ve 50 sorgudan oluşan bir deney derlemi kullanılmıştır. Çalışmada, belgelerdeki sözcükler kökleri bulunmadan oldukları gibi indekslenmişlerdir. Oluşturulan indeks üzerinde sorgu sözcüklerini olduğu gibi veya köklerini bularak kullanan iki ayrı yöntem incelenmiş ve kök bulmanın etkileri gözlemlenmiştir. İkinci yöntemde, bir kökten türeyebilen sözcükler bir belgede geçiyorsa, o belge de sorgu sonucu olarak sıralamaya katılmakta ve kullanıcının dikkatine sunulmaktadır. Yazarlar bu yaklaşımın nedeni olarak, Türkçede eklemeler sonucunda elde edilen sözcüklerin sözcük kökünü fazla etkilemediğini göstermişlerdir. Deneylerde sözcük kökü kullanmak "OKAPI" tabanlı sistemin etkinliğini önemli ölçüde arttırmış ve en başta verilen ilk 10 ve ilk 20 belge içinde %32 kadar daha çok ilgili belgeye ulaşmayı sağlamıştır. Burada belirtilmesi gereken bir nokta şudur: Kök kullanmadan doğrudan sözcükleri kullanmak, indeksleme yöntemine bağlı olarak sözcüklerin belge ve sorgulardaki ağırlıklarını yani önemlerini etkileyebilecektir (Salton, Buckley, 1988). Kök bulmak için hazırlanan "lemmatizer" programları (Oflazer, 1994) bu çalışmada da kök bulma yaklaşımlarından birinde kullanılmıştır.

Sever ve Bitirim ise (2003), 2468 yasal belge ve 15 sorgudan oluşan bir deney derlemi ile yaptıkları çalışmada, gerçekleştirmiş oldukları yeni bir kök bulma yönteminin daha önceden

önerilen iki kök bulma yöntemine göre (biri yukarda bahsedilen Solak-Can kök bulma yaklaşımıdır) daha başarılı olduğunu göstermişlerdir. Çalışma; sözcük kökü kullanmanın, sözcükleri olduğu gibi kullanmaya göre sistem etkinliğinde duyarlılığı %25 oranında arttırdığını göstermektedir.

Son olarak, Pembe ve Say'ın çalışmasında ise (2004), araştırmacılar kök bulmayı sorguları zenginleştirme işlemlerinde kullanmışlardır. Bu çalışmadaki deneylerde, çeşitli konulardaki 615 belge ve 5 adet uzun sorgu kullanılarak yedi farklı indeksleme ve erişim kombinasyonunun etkileri incelenmiştir.

Yukarıda anlatılan çalışmaların özeti Tablo 1'de verilmiştir. Bu çalışmaların ortak özelliği, her birinin başka bir deney derlemi ile yapılmış olmasıdır. Şimdiye dek BE araştırmalarında Türkçe için kullanılabilecek standart bir deney derleminin olmaması, farklı araştırmacıların önerdiği yöntemleri karşılaştırmakta bazı zorluklar yaratmaktadır.

TABLO 1. Türkçe BE çalışmaları

Araştırmacılar (Yıl)	**Derlem Konusu**	**Belge Sayısı**	**Sorgu Sayısı**
Köksal (1981)	Bilgisayar bilimleri	570	12
Solak, Can (1994)	Gazete haberleri	533	71
Ekmekçioğlu, Willett (2000)	Gazete haberleri	6289	50
Sever, Bitirim (2003)	Yasal	2468	15
Pembe, Say (2004)	Çeşitli	615	5

TÜRKÇEDE SÖZCÜK KÖKÜ BULMA

Bu çalışmada, Türkçe sözcüklerin kökünü bulmakta dört farklı yaklaşım kullanılmıştır. Bunlar (1) bilgisayar biliminde bir problemi yok sayma anlamına gelen "devekuşu algoritması" yaklaşımıyla "Kök Bulmadan" (KB) sözcüklerin olduğu gibi kullanılması, (2) her sözcükte "Baştan n" (Bn) adet karakterin kök olarak kullanılması, (3) "Sonraki Çeşitlilik" (SÇ) –successor variety- yaklaşımı ve (4) sözcük biçim birimlerine (morpheme) uygun olarak kökün bir "lemmatizer" programı kullanılarak bulunmasıdır. Dördüncü yaklaşım, verilen bir sözcüğün bir sözlükteki karşılığını bulmaya çalıştığı için makalede "Sözcüğün Aslı" (SA) biçiminde tanımlanmıştır.

1. *Kök Bulmadan* (KB): Bu yaklaşımda bütün sözcükler olduğu gibi, yani kök bulmadan kullanılır. Analizlere dahil edilmesindeki amaç, kök bulma algoritmalarının karşılaştırılmasında bir temel ölçü oluşturmaktır.

2. *Baştan n Karakter* (Bn): Bu yöntem yaklaşık olarak kök bulmayı amaçlar ve sözcükleri sondan kırparak baştan n karakterini kök olarak kullanır. Sondan kırpılma ihtiyacı olmayan n veya daha az sayıda karakterden oluşan sözcükler oldukları gibi kullanılırlar. Kök bulma amacıyla sözcüğün ilk 5 karakteri kullanılırsa (n = 5) bu yöntem B5 kısaltması ile gösterilecektir. Yaptığımız deneylerde B3'den B7'ye dek bütün n değerleri incelenmiştir. Bu yöntem, Türkçede eklerin sözcük kökünü çok fazla değiştirmemesi ve uygulama kolaylığından ötürü bu çalışmaya dahil edilmiştir. Ayrıca, Sever ve Tonta (2006) B5, B6 ve B7 yaklaşımlarını bu biçimde elde edilen köklerin ve sözcüklerin dil içindeki dağılımının benzerliği nedeniyle Türkçe için uygun olabilecek yaklaşık kök bulma yöntemi olarak önermektedir. Ancak, çalışmalarında bu öneriyi destekleyecek herhangi bir BE deneyi sonucu vermemişlerdir.

3. *Sonraki Çeşitlilik* (SÇ): Her dile uygulanabilecek olan bu yaklaşımda, elde bir sözcük kümesi olması gerekmektedir (Hafer, Weiss, 1974). Sözcüklerin kökü, eldeki bu sözcüklerin sağladığı bilgi kullanılarak bulunur. Eldeki sözcük sayısının çokluğu doğru sözcük kökünü bulmakta etkili olmaktadır. Bir sözcüğün kökünü bulurken, baştan belli sayıdaki karakteri alınır ve bu öncül ile başlayan sözcüklerde bu öncülü izleyen kaç farklı durum olduğu sayılır. Örneğin, elimizdeki sözcük kümesi {bir, birikim, birlik} ise "bir" öncülünü izleyen üç farklı durum vardır: "bir" sözcüğünde "bir" öncülünü izleyen harf yoktur, "birikim" sözcüğünde "i" ve "birlik" sözcüğündeyse "l" harfi "bir"i izlemektedir. Bizim uygulamamızda, bir sözcük için elde edilmiş en büyük SÇ değerine karşılık olarak bulunan en uzun öncül, sözcük kökü olarak alınmıştır. Bu seçimin nedeni, daha uzun köklerin sözcüğün gerçek anlamını daha iyi yansıtabilmesidir. Uzunluğu dörtten kısa olan sözcükler için ise, bir işlem yapılmadan sözcüğün kendisi kök olarak alınmıştır.

Dilden bağımsız olan SÇ algoritması Türkçeye uygulanırken dilin özellikleri gözetilmiştir. Örneğin, Türkçede kök sonuna gelen ekler nedeniyle kimi harfler değişebilmekte ya da düşebilmektedir. Mesela, "ağaç" köküne getirilen yönelme eki "-a" ile oluşturulan "ağaca" sözcüğünde "ç" harfi "c" ile değiştirilmiştir. Örneğin "burun" sözcüğüne 1. tekil şahıs eki getirilerek oluşturulan "burnum" kelimesinde "u" harfi düşmüştür. Ayrıca Türkçede birleşik sözcükler de vardır. Örneğin "hanımeli" için "hanım" sözcüğünün kök olarak bulunması anlamlı olmayacaktır.

Çalışmada Türkçede en çok rastlanan durum olan sözcük içindeki harf değişiklikleri SÇ algoritmasına yansıtılmıştır. Algoritma, bir harf değişikliği olasılığını görünce bu değişikliğin olma olasılığına bakar. Değişikliğin olma olasılığı, derlemdeki bütün sözcüklerin dağılımına bakılarak hesaplanır. Bu olasılık belli bir eşik (threshold) değerinden fazlaysa söz konusu öncül değişikliğe uğramamış olan öncülün SÇ değerini bir arttırır. Örneğin "ağac**a**" sözcüğündeki son a harfi, "ağaç" öncülünün SÇ değerini bir arttırmak için kullanılır.

4. *Sözcüğün Aslı* (SA): Dördüncü ve sonuncu kök bulma yaklaşımında "lemmatizer" programı (Oflazer, 1994) kullanılarak, sözcüğün asıl kökünün bulunması hedeflenmiştir. Bir "lemmatizer", verilen çekilmiş bir sözcüğün biçim birimlerini (morpheme) inceleyerek sözcüğe karşılık gelen sözlükteki karşılığını bulmayı amaçlar. Bu noktada "lemmatizer"ların kök bulma algoritmaları olmadığını belirtmek yerinde olacaktır. Bir kök bulucu program verilen sözcüğün kökünü bulmaya çalışırken, bir "lemmatizer" verilen sözcüğe karşılık gelen sözlükteki maddeyi ("lemma") bulmaya çalışır. Bunu yaparken aynı zamanda bulunan sözcüğün aslının konuşma parçası olarak görevini (ad, fiil, sıfat, vb.) saptar. İngilizcede bir sözcük için bulunan "lemma", verilen kelimeyle yüzeysel olarak çok benzerlik göstermeyebilir. Örneğin, İngilizce "best" ve "better" sözcükleri için karşılık gelen " "good" (lemma) arasında yüzeysel olarak bir benzerlik yoktur. Türkçede buna benzer durumlarla az karşılaşılmaktadır: Verilen bir sözcüğe karşılık gelen "kök" veya "lemma", gramer kurallarına göre son eklerin çıkartılmasıyla elde edilebilmektedir. Bu nedenle makalenin devamında "lemma" sözcüğü yerine "kök" sözcüğü tercih edilmiştir.

Bir "lemmatizer" yardımıyla "kök" bulma sırasında bir sözcük için birden fazla sonuç söz konusu olabilmektedir. Çalışmamızda doğru kökün seçiminde Altıntaş ve Can (2002) tarafından geliştirilen yönteme göre iki adımda seçim yapılmaktadır: (1) Aday kökler arasından, uzunluğu Türkçe sözlükteki sözcüklerin ortalama uzunluğuna en yakın uzunlukta olan kök seçilir, (2) İlk adımdan sonra yine birden fazla aday kalırsa, bulunan adayların cümle içindeki görevlerine bakılır ve aralarında Türkçede daha sık geçen sözcük türüne karşılık gelen "lemma", eldeki sözcüğün kökü olarak belirlenir.

Bu algoritmanın uygulanabilmesi için elbette sözlükte bulunan sözcüklerin ortalama uzunluğu ve Türkçede sözcük türlerinin geçiş sıklığının bilinmesi gerekmektedir. Sözlükte yer alan sözcüklerin ortalama uzunluğu 6.58 harf olarak Altıntaş ve Can (2002) tarafından verilmiştir. Aynı çalışmada sözcük türlerinin sıklığı da verilmiştir. Bu araştırmacılar, belirtilen yöntemle doğru kök bulma oranını deneysel olarak %90 olarak saptamışlardır. Bu sonuç kusursuz olmamakla birlikte pratikte kabul edilebilecek bir başarı durumunu yansıtmaktadır.

Bu çalışmada "lemmatizer" ile kök seçiminde ortalama kök uzunluğu olarak 6.58 ve 5 değerleri kullanılmıştır. Kök uzunluğunun 5 olarak kullanılmasının nedeni, yapılan deneylerde B5 kök bulma yönteminin Bn (baştan n karakter) yaklaşımları arasında en iyi BE etkinlik sonuçlarını sağlamasıdır. Bu iki değer kullanılarak elde edilen kök bulma yöntemleri SA5 ve SA6 olarak adlandırılmıştır. Bunun dışında SA yaklaşımı yabancı sözcükler ve yanlış yazılmış sözcükler için bir "lemma" bulamamakta ve bu durumları "hata" olarak bildirmektedir. Bu hatalı sözcüklerin SÇ yöntemi ile köklerinin bulunmasıyla elde edilen melez yöntem SA-SÇ olarak adlandırılmıştır.

BELGE ERİŞİMİNDE ETKİSİZ SÖZCÜKLERİN SEÇİMİ

Belgelerin içinde fazla sayıda geçen sözcükler erişim sırasında belgeleri ayırt etmede çok etkin değillerdir. Bazı uygulamalar, bu sözcükleri "etkisiz sözcükler" (stopword) listesine koyarak indekslenmesini önlerler. İndeksleme sırasında etkisiz sözcüklerin kullanılmaması indeks dosyalarını küçülttüğü için bellek kullanım ve sorgu işleme verimliliğini arttırabilir. Ancak, bellek üretim teknolojisindeki gelişmeler nedeniyle "bellek verimliliği" kıstası zaman içinde BE sistemlerinin gerçekleştirimindeki önemini yitirmiştir (Witten, Moffat, Bell, 1999). Etkisiz sözcük listesi kullanmanın erişim zamanına etkisi ise yeni önerilen erişim yöntemleri sayesinde ihmal edilebilecek kadar azalmıştır (Zobel, Moffat, 2006). İndekslemede kullanılmayacak etkisiz sözcüklerin seçimi bir takım sezgisel kararlar verilmesini gerektirebilmektedir (Savoy, 1999). Çalışmada etkisiz sözcüklerin seçiminde iki yaklaşım denenmiştir:

Bunlardan birincisi yarı-özdevinimli yaklaşım olup, bu yaklaşımda sözcükler bütün belgelerdeki toplam geçiş frekanslarına göre sıralandıktan sonra belli bir frekansın üzerinde geçen sözcükler etkisiz sözcük listesine alınmış, sonra bunlar arasında bilgi değeri içerenler (örneğin, Türkiye ve Erdoğan – şu andaki başbakanın soyadı –) listeden elle çıkarılmıştır. Listedeki bazı sözcükler içinse bu sözcüklerden türeyebilecek bir takım başka sözcükler bu listeye eklenmiştir. Bu yöntemle oluşturulan etkisiz sözcük listesi 147 sözcük içermektedir ve Ek 1'de verilmiştir. Bu sözcükler kök bulmadan (KB yaklaşımı) yapılan indekslemede gözlenen bütün sözcüklerin, yani tüm belgelerdeki sözcüklerin, %14'ünü kapsamaktadır.

İkinci yaklaşımdaysa, sözcükler birinciye benzer biçimde frekanslarına göre sıralanmış ve hiçbir ayrım yapılmaksızın en başta yer alan, yani en sık geçen, 288 sözcük listeye dahil edilmiştir. Bu seçim yapılırken ilk baştaki sözcüklerin tüm belgelerdeki önemli sayıdaki sözcüğü kapsadığı gözlenmiş, listede aşağıya inildikçe bu kapsama azalmış ve kapsama etkinliğini yitirdiği, yani doyuma ulaştığı zaman sözcük seçimine son verilmiştir. Özdevinimli olarak oluşturulan ikinci liste bütün belgelerdeki sözcüklerin %27'sini kapsamaktadır. İkinci bir etkisiz sözcük listesi üretmemizin nedeni ise, özdevinimli bir biçimde üretilen etkisiz sözcük listelerinin Türkçe BE etkinliğini ne ölçüde etkileyeceğini saptamaktır.

BELGE-SORGU BENZERLİK ÖLÇÜMÜ: SIRALAMA FONKSİYONLARI

İndeksleme sırasında sözcüklere verilecek önemin, ağırlığın, saptanması BE sistemlerinin gerçekleştirimde hem etkinlik hem de verimlilik açısından önem taşımaktadır (Cambazoğlu, Aykanat, 2006; Lee, Chuang, Seamons, 1997; Salton, Buckley, 1988). Bu çalışmada *tf.idf* modeli kullanılmaktadır. Bu gösterimde *tf* ("term frequency") sözcük tekrarını ifade etmekte ve bir belgede daha çok sayıda geçen bir sözcük daha az sayıda geçene göre daha fazla ağırlık kazanmakta, *idf* ise ("inverse document frequency") daha az sayıda belgede geçen sözcüğe daha çok önem ya da ağırlık verildiğini ifade etmektedir.

İndeksleme sırasında sözcüklere önem vermede üç öğe bulunmaktadır: sözcüğün belgedeki frekansı (BF), derlemdeki frekansı (DF), ve ağırlık hesaplanmasında yapılabilecek olan normalizasyondur (N). Normalizasyon, içinde çok sayıda sözcük olan belgelerin aşırı önem kazanmasını önlemeyi amaçlar. İndekslenme amacıyla kullanılan sözcüklerin ($1 \leq k \leq$ toplam sözcük sayısı) belgedeki ağırlığı (a_{bk}) ve sorgudaki ağırlığı (a_{sk}) bu üç öğenin belge ya da sorgudaki ağırlıklarının çarpımıyla elde edilmektedir. Bunun ardından belgeleri sıralamak için kullanılacak olan benzerlik hesabı bir vektör çarpımıyla elde edilmektedir (Salton, Buckley, 1988). Buna göre sorgu (S) ve belge (B) arasındaki benzerlik aşağıda ifade edilen toplama eşittir.

$$benzerlik \quad (S, B) = \sum_{\forall \text{ sözcük}} a_{bk} \cdot a_{sk}$$

Sıralamaya bağlı BE sistemlerinde belgeler sorguya olan benzerliğine göre kullanıcının dikkatine getirilmektedir. Bir sözcüğün B belgesi içindeki ağırlığı üç öğenin çarpımıyla elde edilmektedir: BF x DF x N.

Belge frekansı (BF) için üç olasılık bulunmaktadır ve bunlar (takip eden anlatımda Salton, Buckley, 1988'deki simgeler kullanılacaktır) *b*, *t* ve *n* simgeleri ile gösterilmektedir. Sözcük ağırlığı hesabındaki BF, indekslemede *tf.idf* ile gösterilen yaklaşım *tf* 'ye karşılık gelmektedir. BF'de *b* (binary) ikili ağırlık anlamına gelmektedir ve BF'nin değeri sözcüğün belgedeki geçiş sıklığı göz önüne alınmadan eğer sözcük belgede geçiyorsa 1, sözcük belgede geçmiyorsa 0 olarak alınır. BF'de *t* (term frequency) ise sözcük frekansı anlamına gelmektedir ve BF için *t* seçildiği zaman, BF'nin değeri o sözcüğün belgedeki frekansı olarak alınır. Son seçenek olan *n* (augmented term frequency) genişletilmiş sözcük frekansı anlamına gelip (0.5 + 0.5 x t/maxt) olarak tanımlanmıştır, ki burada *maxt* o belgede en sık geçen sözcüğün frekansıdır.

Derlem frekansı (DF) için de üç seçenek bulunmaktadır ve bunlar *x*, *f* ve *p* olarak gösterilmektedir. DF'de *x* değişiklik yapılmayacağını ifade eder, yani DF için *x* seçildiği zaman DF=1 olarak alınmaktadır. DF için *f* kullanılırsa bu *idf* anlamına gelmektedir, yani daha çok sayıda geçen sözcüklere daha az ağırlık verilmektedir, bu çalışmada *idf* için *ln*(derlemdeki toplam belge sayısı/sözcüğü içeren belge sayısı)+1 formülü kullanılmıştır. DF için kullanılan sonuncu simge, *p* ise, olasılığa bağlı bir hesaplamayı ifade etmektedir ve bu çalışmada kullanılmamıştır.

Normalizasyon için iki olasılık vardır, bunlar *x* ve *c* ile gösterilmiştir. N için *x* simgesi kullanılırsa NC=1 olarak alınır. *c* ise kosinüs normalizasyon yaklaşımını ifade etmekte olup (BF x DF) için hesaplanan ağırlığın Euclid vektör uzunluğu ile normalize edileceğini göstermektedir ve hesaplanmış olan sözcük ağırlıklarının karelerinin toplamının kare köküne eşittir. Sorgu sözcüklerinin normalizasyonu bütün sorgu terimlerinin ağırlığının aynı sayı ile bölünmesini

gerektirdiğinden belgelerin sıralanmasını değiştirmemektedir. Bu nedenle sorgu sözcüklerinin normalizasyonu yapılmamaktadır.

TABLO 2. Deneylerde kullanılan sıralama fonksiyonları

Sıralama Fonksiyonu	**SF1**	**SF2**	**SF3**	**SF4**	**SF5**	**SF6**	**SF7**
Anlamı	*Txc.txx*	*tfc.nfx*	*tfc.tfx*	*tfc.bfx*	*nfc.nfx*	*nfc.tfx*	*nfc.bfx*

Tablo 2'de gösterildiği gibi BF, DF ve N'nin alacağı değerlere göre çeşitli biçimlerde sözcük ağırlıkları hesaplanabilmekte ve böylece farklı sıralama fonksiyonları (SF) elde edilebilmektedir. Örneğin SF1, belge sözcüklerinin ağırlığının hesaplanmasında *txc*'yi (BF= *t*, DF= *x*, N= *c*), sorgu sözcüklerinin ağırlığını hesaplanmasındaysa *txx*'i kullanmaktadır. Bu yöntem, derlemden gelen DF bilgisini kullanmadan, doğrudan kosinüs yöntemiyle benzerlik hesaplamaya karşılık gelmektedir ve BE literatüründe yaygın bir biçimde bilinmektedir. Belge sözcüklerinin ağırlığını saptamada *tfc*, *nfc*'nin, sorgu sözcüklerinin ağırlıklarını hesaplamada ise *nfx*, *tfx* ve *bfx*'in etkin erişim ortamı sağladığı gösterilmiştir. Bu seçenekler altı (SF2-SF7) değişik sıralama fonksiyonunun tanımını sağlamaktadır: "*tfc.nfx*", "*tfc.tfx*", "*tfc.bfx*", "*nfc.nfx*", "*tfc.tfx*" ve "*nfc.bfx*". Bunlar birbirine benzemekle birlikte farklı sıralama fonksiyonlarıdır ve Salton ve Buckley (1988) tarafından BE sistemleri için etkin sonuç verdikleri belirtilerek önerilmiştir. Tablo 2'de yukarıda tanımlanan ve deneylerde kullanılan sıralama fonksiyonlarının tanımı verilmiştir. Bu sıralama fonksiyonları bizim daha önce İngilizce (Can, Özkarahan, 1990) ve Türkçe (Solak, Can, 1994) üzerinde yaptığımız araştırmalarda kullandığımız fonksiyonlardır.

Yaptığımız deneylerde yukarda tanımlanan sıralama fonksiyonlarına ek olarak SF8 sıralama fonksiyonunu da (Long, Suel, 2003; Witten et al. 1999) kullandık. Bu sıralama fonksiyonuna göre sorgu (S) ile belge (B) arasındaki benzerlik şu biçimde hesaplanmaktadır.

$$SF\ 8 = \sum_{t \in S} \left[\ f_{ss} \cdot \ln\left(1 + N \,/\, f_s\right)\ \right] \cdot \left[\ \left(1 + \ln\ f_{bs}\right) \Big/ \sqrt{B}\ \right]$$

Burada f_{ss}, sözcüğün S sorgusu içinde geçiş frekansını; f_S, sözcüğün bütün derlemdeki geçiş sayısını; f_{bs}, sözcüğün belge içinde geçiş frekansını; N, derlemdeki toplam belge sayısını; B, söz konusu belgedeki toplam sözcük sayısını göstermektedir. Belge sayısı değiştikçe farklılaşan *idf* değerinin sorgu sözcük ağırlığı aracılığıyla (SF8'in ilk çarpanı) dinamik olarak hesaplamaya katılabilmesi, SF8'i özellikle değişen dinamik derlemler için uygun hale getirmektedir.

DENEY DERLEMİ

Bu araştırma için geliştirdiğimiz deney derlemi 408.305 belge ve 72 sorgudan oluşmaktadır. Deney derlemindeki belgeler Milliyet Gazetesinin Web sayfasından (www.milliyet.com.tr) indirilen beş yılın (2001-2005) haberlerini ve köşe yazılarını içermektedir. Bu derlem, makale içinde kısaca *Milliyet* olarak anılacaktır. Yaklaşık 800MB büyüklüğünde olan derlem herhangi bir eleme yapmadan toplam 95.5 milyon (89.33 abecesel, 4.63 sayısal, 1.37 abecesel-sayısal) sözcük içermektedir. Her belgede ortalama 234 sözcük bulunmaktadır. Sözcükler harf ile başlayıp daha sonra harf, rakam, ayraç işareti (') ve çizgi (-) içerebilmektedir. Sözcük sonunda bulunan ' – karakterleri sözcüğe dahil edilmeden iptal edilir. Sesli harflerden a, i ve u'nun uzatmalı durumları (â gibi) farklı harfler olarak kullanılmıştır.

Deney derlemindeki sorgular TREC (http://trec.nist.gov/) yaklaşımında olduğu gibi yazılmış ve değerlendirilmiştir. Bu amaçla, bilgisayar kullanımında deneyimli 33 kişi sorgularını Web ortamında hazırlanan bir arayüzü kullanarak girmiş ve sistem tarafından getirilen belgeleri "evet"

(ilgili), "hayır" (ilgisiz) biçiminde işaretleyerek değerlendirmişlerdir. Sorgu sahibine göstermek için, 8 sıralama fonksiyonu (SF1 - SF8) ve deneylerimize başlarken elimizde olan üç kök bulma yönteminin (KB, SÇ, B6) sağladığı 24 farklı kombinasyonun her biri için elde edilen ilk 100 belge tekrarlar önlenecek şekilde bir havuza konulmuştur. Havuzda bulunan belgeler, önyargı yaratmamak amacıyla sorgu sahibine rastgele bir sırayla gösterilmiştir. Kök bulmadan, yani KB yaklaşımı ile indekslenecek <belge numarası, sözcük ağırlığı> bilgi çifti 67.079.608 taneyken, etkisiz sözcüklerin çıkarılmasıyla bu sayı 61.105.853'e düşmektedir.

KB, SÇ, B6 yaklaşımları ile kök elde edildiğinde, belgelerdeki ortalama farklı sözcük sayısı 150, 134 ve 120 olarak gözlenmiştir. Bu durumlar için ortalama sözcük kökü uzunluğuysa, sırasıyla 9,88; 5,66 ve 7,23 karakter olarak elde edilmiştir. Tersine çevrilmiş (inverted) dosya yapısında her bir sözcük için elde edilen <belge numarası, sözcük ağırlığı> listesinde ise, yine aynı kök bulma yaklaşımları için sırasıyla 61, 55 ve 49 milyon bilgi birimi gözlenmiştir. Bu durum, SÇ ve B6 yaklaşımlarının KB'ye göre %20 (61 milyondan 49 milyona düşmesi nedeniyle) ve %10 (61 milyondan 55 milyona düşmesi nedeniyle) bellek tasarrufu sağladığını göstermektedir.

Sorgular "konu", "tanım" ve "açıklama" alanlarından oluşmaktadır. Sorguların "konu" bölümünde bir kaç sözcükle aranan bilgi belirtilmekte, "tanım" alanında bir iki cümle ile aranan bilgiye açıklık kazandırılmakta ve "açıklama" alanındaysa aranan belgeler için daha fazla açıklayıcı bilgi verilmektedir. Deney derlemindeki sorguların konuları Ek 2'de verilmiştir.

Sorgular için incelenecek belge havuzu hazırlanırken her sorgunun "konu" ve "tanım" alanları alınarak sorgu vektörleri oluşturulmuştur. Ortalama olarak incelenecek belge havuzlarında 466,5 farklı belge ve 104,3 ilgili belge gözlemlenmiştir. Kalite kontrol aşamasında 20 adet sorgu elendikten sonra 72 sorgu kullanım için elimizde kalmıştır. Elenen sorgular ya çok, ya da az sayıda ilgili belge içermektedir (kendi havuzunun %90'i ya da daha fazlası veya %5'i ya da daha azı). Bu türden sorgular, değerlendirmede ayırt edici olamayacağı için deney derlemine alınmamıştır (Carterette, Allan, Sitaraman, 2006). Ortalama sorgu değerlendirme zamanı 130 dakika olarak gözlenmiş olup bütün sorgular için toplam 6923 farklı belge ilgili olarak değerlendirilmiştir.

Makalede sadece "konu" alanı kullanılarak oluşturulan sorgular "kısa sorgu", "konu" ve "tanım" alanları birlikte kullanılarak oluşturulan sorgular "orta boy sorgu" ve bütün sorgu alanları kullanılarak oluşturulan sorgular ise "uzun sorgu" olarak tanımlanmış ve bunlar makalede K_S, O_S ve U_S simgeleriyle gösterilmiştir. Bu sorgularla ilgili bilgiler 3. ve 4. tablolarda verilmiştir. Tablolar K_S'den U_S'ye geçerken sorgulardaki farklı sözcük sayısının ve ortalama sözcük uzunluğunun arttığını göstermektedir.

TABLO 3. Sorgularla ilgili bilgiler

Sorguyla İlgili Bilgi	**En Az**	**En Çok**	**Medyan**	**Ortalama**
Havuz büyüklüğü (farklı belge sayısı)	186	786	458	466,5
Havuzdaki ilgili belge sayısı	18	263	93	104,3
Sorgu değerlendirme süresi * (dk.)	60	290	120	132,4
K_S farklı sözcük sayısı	1	7	3	2,89
O_S farklı sözcük sayısı	5	24	11	12,00
U_S farklı sözcük sayısı	6	59	26	26,11

* Sorgu sahibi tarafından belirtilmiştir.

O_S'ye bakıldığında sorgularda en çok kullanılan sözcüklerin "türkiye'de", "etkileri", "üzerindeki", "türk", "gelen", "son", "türkiye", "avrupa", "meydana" ve "şiddet" olduğu görülmektedir. Bu sözcükler bu sorgu türünde geçen toplam 1004 sorgu sözcüğünün %10,26'sını oluşturmaktadır. En çok kullanılan sorgu sözcüklerinin ve genelde sorgularda kullanılan sözcüklerin kısa olduğu gözlenmiştir. Örneğin, bu sözcüklerden hiçbiri "Avrupalılaştırılamayabilenlerdenmişsiniz" ve benzerleri gibi abartılı sözcükler değildir. Başka bir gözlem de kullanıcıların daha uzun sorgu yazma durumunda daha uzun sözcükler kullanmasıdır. Örneğin K_S'de ortalama sözcük uzunluğu 7,03'den U_S'de bu değer 7,62'ye çıkmaktadır.

TABLO 4. Sorgu sözcük boylarıyla ilgili bilgiler

Sorgu bilgisi	**K_S**	**O_S**	**U_S**
Sözcük sayısı	208	1004	2498
Farklı sözcük sayısı	182	657	1359
Ortalama sözcük uzunluğu (karakter)	7,03	7,57	7,62
Ortalama farklı sözcük uzunluğu (karakter)	7,00	7,75	8,04

DENEY SONUÇLARI

Etkinlik Ölçümleri

Duyarlılık ve anma, belge erişim uygulamalarında en çok kullanılan etkinlik ölçüleridir. Duyarlılık, daha önce de belirtildiği gibi, getirilen belgeler içindeki ilgili belge oranı; anma ise tüm ilgili belgeler içinden getirilebilen ilgili belge oranı olarak tanımlanır. Gerçek uygulama ortamlarında anmanın tam olarak ölçülmesi imkansız değilse bile çok zordur. Getirilen ilgili belgelerin tüm ilgili belgelere oranını saptamak için, her bir sorgu için bütün belgelerin incelenmesi gerekir. Basit ve kolay anlaşılır olması nedeniyle çoğunlukla getirilen ilk 10 (İ10D) veya ilk 20 (İ20D) belge arasındaki duyarlılık ölçüsü kullanılır. Ayrıca, internet kullanıcıları genellikle sorgu sonuçlarının ilk iki sayfasına bakarlar. Bu açıdan da İ10D ve İ20D ölçümleri kullanıcı memnuniyeti ile uyumlu sonuçlar verir. Bununla birlikte, getirilen ilgili belgelerin duyarlılığının ortalama bir ölçüsü olan *Vasati Ortalama Duyarlılık* (VOD: Mean Average Precision MAP) da duyarlılık için güvenilir bir ölçü olarak kabul edilir (Buckley ve Voorhees, 2004; Sanderson ve Zobel, 2005; Zobel, 1998).

Daha önce de belirttiğimiz gibi bu çalışmada, ilgili belgelerin saptanmasında kullanılan incelenecek belge havuzlarının oluşturulmasında KB, B6 ve SÇ sözcük kökü bulma yöntemleri kullanıldı. Fakat, yalnız bu yöntemlerle elde edilen sorgu ve ilgili belge dağarcığı, daha sonra bütün yöntemlerin başarım değerlerinin hesaplanmasında kullanıldı. Böylece, incelenecek belge havuzlarının hazırlanmasında kullanılmayan sıralama fonksiyonu ve kök bulma ikilileri için de aynı ilgili belge bilgileri kullanılmış oldu. Bu uygulama, incelenecek belge havuzlarının hazırlanmasında kullanılmayan yöntemlerin aleyhine bir durum yaratabilirdi. Bu gibi durumlarda kullanılmak üzere, Buckley ve Voorhees (2004) tarafından deney derlemi hazırlanma aşamasında kullanıcılara gösterilmeyen belgeleri dikkate almayan ve ikili tercih (bpref) olarak adlandırılan yeni bir başarım ölçüsü önerilmiştir. Biz de deneylerimizde bu ikili tercih başarım ölçüsünü kullandık. Etkinlik ölçüm değerlerini hesaplamak için, 'trec' değerlendirme paketinin çalışmanın yapıldığı zamanki en son sürümü olan 8.1 sürümü kullanıldı.

Genel Değerlendirme İçin Sözcük Kökü Bulma Yöntemi Seçimi

Genel değerlendirme sürecini kolaylaştırmak amacıyla, önce baştan belirli sayıda harf alma ve sözcüğün aslının bulunması yöntemleri için kendi gruplarında en iyilerini grup temsilcisi olarak seçtik. Tablo 5, bütün baştan belirli sayıda harf alarak sözcük kökü bulma yöntemleri ile SF8 için farklı etkinlik ölçülerine ait değerleri içermektedir. Bütün sıralama fonksiyonu ve sözcük kökü bulma ikilileri içinde en iyi sonucu SF8 verdiği için daha sonraki deneylerde bu sıralama fonksiyonunu kullandık. İkili tercih başarım ölçüsüne göre B4 ve B5, diğer baştan belirli sayıda harf alma yöntemlerine göre en iyileridir. Örneğin, bu iki yöntem B3'e göre %5 ile %6 arasında daha iyidirler. Bu iki yöntemden birini seçmek için VOD, İ10D ve İ20D başarım değerlerini de dikkate aldık. VOD ölçüsüne göre B5'in başarımı, B5 incelenecek belge havuzlarının oluşturulmasında kullanılmamış olmasına rağmen, incelenecek belge havuzlarının oluşturulmasında kullanılan B6 yönteminden %5 daha fazladır. B4 ile B6 nın başarım karşılaştırmaları için de yaklaşık olarak aynı durum söz konusudur. VOD başarım ölçüsüne göre B3 ve B7 bariz olarak en kötüleridir. B4 ve B5'in ikili tercih ve VOD değerleri birbirine çok yakın olmasına rağmen, B5'in İ10D ve İ20D başarım değerleri B4'e göre %5 daha iyidir. Bu gözlemler sonucunda B5 yöntemini baştan belirli sayıda harf alma yöntemlerinin temsilcisi olarak seçtik. Ancak, SF8 ile İ10D ve İ20D için her sorgunun ayrı olarak değerlendirildiği iki uçlu t testi sonuçları B4 ve B5 arasında önemli bir başarım farkı olmadığını göstermektedir.

Benzer bir biçimde sözcüğün kökünü bulma yöntemleri arasında SA5'in başarım değeri, SA6'nın başarım değerinden biraz daha iyidir. Sözcüğün kökünü bulma yöntemi olarak SA-SÇ, hem SA5 hem de SÇ nin avantajlarına sahiptir. Büyük bir oranda olmasa da SA-SÇ SA5'e göre biraz daha iyi sonuçlar elde etmiştir. Sonuç olarak, son değerlendirmeler için KB, SÇ, B5 (baştan belirli sayıda harf alma yöntemleri temsilcisi) ve SA-SÇ (sözcük aslı bulma yöntemleri temsilcisi) yöntemlerinin kullanılmasına karar verilmiştir.

TABLO 5. SF8 ile B3-B7 yöntemleri için elde edilen değişik etkinlik ölçüsü değerleri

Yöntem	İkili Tercih	VOD	İ10D	İ20D
B3	0,4120	0,3134	0,5139	0,4757
B4	0,4382	0,4013	0,5625	0,5361
B5	0,4322	0,4092	0,5917	0,5653
B6	0,4014	0,3885	0,5667	0,5382
B7	0,3901	0,3658	0,5556	0,5181

Genel Değerlendirme: Sözcük Kökü Bulma ve Sıralama Fonksiyonlarının Etkileri

İkili tercih ölçüsüne göre KB, SÇ, B5, SA-SÇ ve SA5 (SA5'i SA-SÇ ile karşılaştırmak için dahil ettik) için başarım değerleri Tablo 6'da verilmiştir. Aynı tablo; SA-SÇ, SÇ ve B5'in KB ye göre başarım artış oranları ile SA-SÇ'nin SÇ ve B5'e göre başarım artış oranlarını da içermektedir. Tablo 6'da, bütün sözcük kökü bulma yöntemlerinin KB yöntemine göre önemli ölçüde daha iyi başarım elde ettiği görülmektedir. Daha kolay karşılaştırılabilmeleri için KB, B5, SÇ ve SA-SÇ yöntemlerinin ikili tercih başarım değerleri çubuk grafik olarak Şekil 1'de gösterilmiştir. Bütün sıralama fonksiyonları arasında en iyi başarım değerine sahip SF8 ile KB

yöntemine göre B5 %32,78, SA-SÇ %38,37 ve SÇ ise %32,23 daha iyi başarım değeri elde etmişlerdir.

Deneylerimizde gözlemlediğimiz en etkin kök bulma yöntemi SA-SÇ'dir. Tablo 6'da gösterilen SF8 ile B5, SA-SÇ ve SÇ'nin başarım karşılaştırmalarına göre, çok önemli seviyede olmasa da, SA-SÇ SÇ den %4,65 ve B5'ten de %4,21 daha iyi başarım elde etmiştir (SA-SÇ/SÇ ve SA-SÇ/B5 sütunlarına bakınız). Ancak bu başarım artışları, istatistiksel olarak anlamlı olabilecek düzeyde değildir. SA-SÇ kadar iyi olmasa da, SÇ ve daha basit olan B5 yöntemleri de etkilidir. B5 de istatistiksel olarak anlamlı olmayacak bir seviyede SÇ'den daha iyidir.

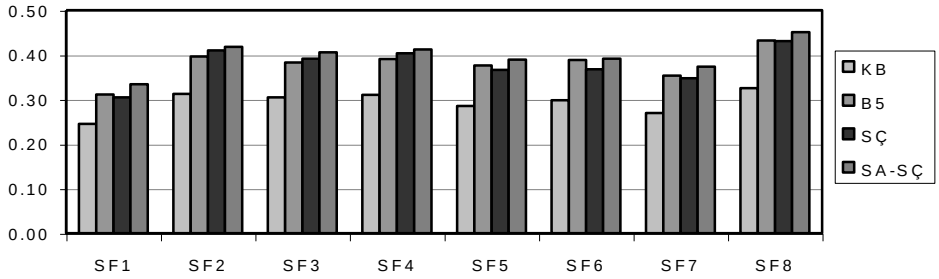
B5, SA-SÇ ve SÇ nin SF1 den SF8'e kadar bütün sıralama fonksiyonlarının başarım değerleri ile karşılaştırılması sonucunda (Tablo 6'ya bakınız) B5 ile SÇ arasında istatistiksel olarak anlamlı bir başarım farkı görmedik. Ancak, B5 ile SA-SÇ ve SÇ ile SA-SÇ arasında iki uçlu t testi ($p < 0.001$) sonuçlarına göre önemli bir başarım farkı varmış gibi görünmektedir. Fakat, bu yöntemlerin her bir sorgu için elde edilen ikili tercih başarım değerleri kullanılarak hesaplanan iki uçlu t testi sonuçları, aralarında istatistiksel olarak anlamlı bir başarım farkı olmadığını göstermektedir. Bu sonuçlar, B5 gibi basit bir sözcük kesme yöntemi ile, SÇ gibi dile özgü istatistikleri kullanan daha dikkatli bir sözcük kesme yöntemi ve SA-SÇ gibi daha gelişmiş bir kök bulma yönteminin benzer başarım değerleri elde ettiğini göstermektedir. Elde edilen sonuçlar, başka bir eklemeli dil olan Fince üzerinde Kettunen, Kunttu ve Jarvelin (2005) tarafından yapılan araştırma sonuçları ile aynı doğrultudadır. Kettunen ve arkadaşları çalışmalarında kök bulma yöntemi ile basit bir sözcük kesme yönteminin benzer erişim etkinlikleri sağladığını belirtmişlerdir.

TABLO 6. Sözcük kökü bulma yöntemlerinin ikili tercih başarım değerleri ile birbirlerine göre başarım artış oranları

SF*	**KB**	**B5**	**SÇ**	**SA-SÇ**	**SA5**	**SA-SÇ /KB**	**SÇ/KB**	**B5/KB**	**SA-SÇ /SÇ**	**SA-SÇ /B5**
SF1	0,2452	0,3108	0,3046	0,3339	0,3275	36,18	24,23	26,75	9,62	7,43
SF2	0,3124	0,3961	0,4096	0,4175	0,4095	33,64	31,11	26,79	1,93	5,40
SF3	0,3045	0,3823	0,3908	0,4054	0,3992	33,14	28,34	25,55	3,74	6,04
SF4	0,3099	0,3905	0,4030	0,4122	0,4045	33,01	30,04	26,01	2,28	5,56
SF5	0,2849	0,3764	0,3663	0,3890	0,3805	36,54	28,57	32,12	6,20	3,35
SF6	0,2982	0,3883	0,3678	0,3908	0,3847	31,05	23,34	30,22	6,25	0,64
SF7	0,2692	0,3532	0,3477	0,3734	0,3642	38,71	29,16	31,20	7,39	5,72
SF8	0,3255	0,4322	0,4304	0,4504	0,4447	38,37	32,23	32,78	4,65	4,21
Ortalama	0,2854	0,3715	0,3675	0,3922	0,3861	35,08	28,38	28,93	5,26	4,79

*__SF__: sıralama fonksiyonu., __KB__: KB'nin ikili tercih değeri, __SA-SÇ/SÇ__: SA-SÇ'nin SÇ'ye göre yüzde olarak ikili tercih değeri artışı.

Sonuçlar SF1'in bütün sözcük kökü bulma yöntemleri için en kötü başarım değerlerini elde ettiğini göstermektedir. SF1 sıralama fonksiyonu içinde, belge sözcük vektörlerini standart duruma getirme özelliği yoktur. Ayrıca SF1, '*idf*' değerini hesaplamalarda kullanmaz. SF2, SF8 dışındaki diğer sıralama fonksiyonlarından daha iyi sonuçlar elde etmiştir. SF1-SF7 sıralama fonksiyonları için benzer sonuçlar daha önce de rapor edilmiştir (Can ve Özkarahan, 1990). Bu sıralama fonksiyonları için elde edilen göreceli başarım değerleri daha önce rapor edilen sonuçlar ile uyumludur. Çalışmalarımız, "derlem"e belge ekleme ve çıkarmalarda yeniden sözcük ağırlık hesabı yapılmasını gerektirmeyen SF8 in en iyi sonucu elde ettiğini göstermektedir. Bu özelliğin, dinamik olarak büyüyen belge erişim uygulamaları için pratik bir değeri vardır.



ŞEKİL 1. SF1-SF8 için KB, B5, SÇ ve SA-SÇ nin ikili tercih değerleri.

Makalenin sonraki bölümlerinde, başarım artışı karşılaştırmaları için en iyi sonucu elde eden SF8 ile KB, B5 ve SA-SÇ sözcük kökü bulma yöntemlerini kullandık. B5'in başarım değerleri SÇ ile çok yakın olduğu için SÇ yi değerlendirmelere almadık.

Etkisiz Sözcükler ve Erişim Etkinliği Arasındaki İlişki

Bu bölümde, etkisiz sözcük listesinin erişim başarısına etkilerini inceledik. Deneylerde Ek 1'de verilen yarı otomatik, yarı el ile oluşturulmuş etkisiz sözcük listesini kullanarak ve etkisiz sözcük listesi kullanmadan ikili tercih değerlerini ölçtük. Tablo 7'de verilen sonuçlara göre, etkisiz sözcük listesinin başarım üzerinde istatistiksel olarak anlamlı bir etkisi yoktur. Sorguları hazırlayanlara etkisiz sözcükler hakkında bir şey söylenmemiştir. Yönlendirme olmadan oluşturulan sorgularda, etkisiz sözcüklerin çok sıkça kullanılmadığı görülmüştür. Örneğin, K_S içindeki ortalama etkisiz sözcük sayısı 1,74'tür.

TABLO 7. KB, B5 ve SA-SÇ yöntemleri için etkisiz sözcük listesi varken ve yokken (KB*, B5*, SA-SÇ*) elde edilen ikili tercih değerleri

KB	KB*	B5	B5*	SA-SÇ	SA-SÇ*
0,3255	0,3287	0,4322	0,4330	0,4504	0,4524

Yukarıdaki sonuçların elde edildiği deneylerde, önce etkisiz sözcükleri atıp daha sonra kök bulma işlemine geçtik. Ek olarak, önce sözcük kökü bulup sonra etkisiz sözcükleri atarak da deneyler yaptık. Bu deneylerde B5 kök bulma yöntemiyle elde edilen sözcük kökünü yine aynı yöntemle kökleri bulunmuş etkisiz sözcükler arasında aradık. Bu deneylerde de istatistiksel olarak anlamlı bir başarım farkı gözlemlemedik

Otomatik etkisiz sözcük listesi oluşturmanın başarım üzerindeki etkilerini gözlemlemek için de, belgeler içinde en çok geçen ilk 288 sözcüğü etkisiz sözcük listesi olarak kullanarak deneyler yaptık. Bu deneylerde de belge erişim etkinlik ölçümünde bir fark gözlemlemedik.

Deneylerdeki gözlemlerimizden, etkisiz sözcük listesi kullanmanın Türkçe belge erişim uygulamalarında istatistiksel olarak anlamlı bir başarım artışı sağlamadığı sonucuna vardık. Fakat bu sonuç, bizim kullandığımız '*tf.idf*' modelinden kaynaklanmış da olabilir. Örneğin, Savoy (1999) Fransızca belgeler üzerinde yaptığı araştırmada etkisiz sözcüklerin '*tf.idf*' modelinde bir etkisi olmadığını, ancak, "OKAPI" modelinde bir etkisi olduğunu belirtmiştir. Bizim bulgularımız da Savoy'un bulgularıyla uyumludur.

Ölçeklenebilirlik

Belge derlemlerinin dinamik yapısı nedeniyle, ölçeklenebilirlik belge erişim sistemleri için önemli bir konudur. Ölçeklenebilirlik araştırmaları için, tüm deney derleminden her defasında 50.000 belge artırarak farklı büyüklükte sekiz deney derlemi oluşturduk. İlk derlem; haberler geliş zamanına göre sıralandığında baştan 50.000 belgeyi, ikinci derlem ise ilki üzerine eklenmiş sıradaki 50.000 belge olmak üzere 100.000 belgeyi içermektedir. Diğer derlemler de, bir önceki derleme sıradan yeni 50.000 belge eklenerek oluşturulmuştur. Son derlem bütün belgeleri içermektedir.

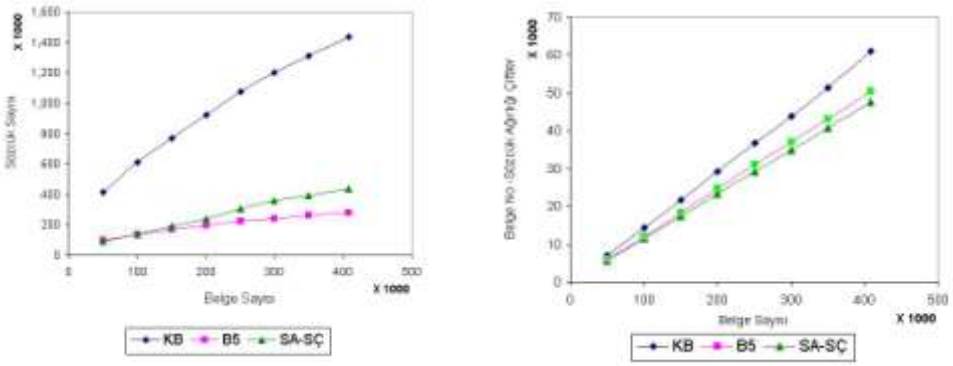
Başarım hesaplamaları aşamasında her derlem için, bu derlemde en az bir ilgili belgesi bulunan derlem için geçerli sorgular dikkate alındı. Örneğin ilk 50.000 belgeyi içeren ilk derlem içinde, 15 adet sorguya uygun belge olmadığı için yalnız 57 adet sorgu kullanıldı. Tablo 8, her alt derlem için kullanılan sorgu özelliklerinin derlem büyüklüğü ile orantılı olduğunu göstermektedir. Örneğin sorgu başına derlemdeki ortalama ilgili belge sayısı, derlem büyüklüğünün her artışında yaklaşık olarak 10 artmaktadır (11,0; 21,5; 34,0 …). Bu deney derlemlerinin benzer özellikler taşıdığını gösterir.

TABLO 8. Artan derlem büyüklüğü için sorgu ve ilgili belge özellikleri

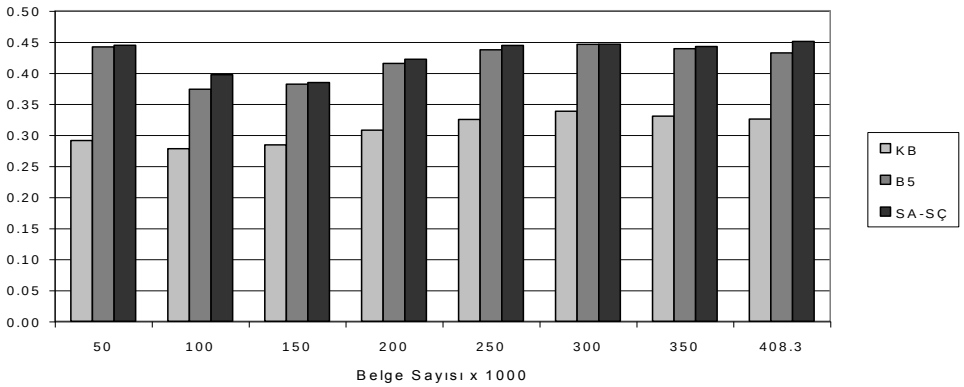
Belge Sayısı	Geçerli Sorgu Sayısı	Toplam Farklı İlgili Belge Sayısı	Sorgu Başına İlgili Belge Ortalaması	Sorgu Başına İlgili Belge Medyanı
50.000	57	719	10,72	11,0
100.000	62	1380	21,08	21,5
150.000	63	2014	30,55	34,0
200.000	64	2944	44,33	45,5
250.000	68	3764	56,51	56,5
300.000	70	4794	71,45	66,0
350.000	71	5725	86,29	79,0
408.305	72	6923	104,30	93,0

Artan belge sayısı için erişim ortamının diğer özelliklerindeki değişim, Şekil 2'de grafik olarak gösterilmiştir. Şekil 2'ye göre derlemdeki belge sayısı arttıkça, farklı sözcük sayısı da artmaktadır. Fakat, B5 ve SA-SÇ sözcük kökü bulma yöntemleri için farklı sözcük sayısındaki artış bir doyuma erişmektedir. Doyuma erişim B5 için daha barizdir. İncelenen bütün kök bulma yöntemleri için <belge numarası, sözcük ağırlığı> çiftlerinin sayısı belge sayısı artıkça doğrusal

olarak artmaktadır. Şekil 2'de grafiksel olarak gösterilen belge numarası-sözcük ağırlığı listesi uzunlukları KB yöntemi için çok sayıda kısa liste olduğunu göstermektedir.



ŞEKİL 2. Artan derlem büyüklüğüne göre farklı sözcük ve ilgili belge-sözcük ağırlığı ikili sayıları.



ŞEKİL 3. Derlem büyüklüğü artışına göre SF8 ile KB, B5 ve SA-SÇ yöntemleri için ikili tercih değerleri.

Derlem büyüklüğüne göre KB, B5 ve SA-SÇ yöntemlerinin ikili tercih başarım değerleri Şekil 3'te verilmiştir. İlk derlem büyüklüğü artışında, daha sonraki üç artışa göre daha iyi bir başarım artışı elde edilmiştir. İkinci derlem büyüklüğü artışıyla elde edilen 100.000 belge için bir başarım azalışı gözlendikten sonra, başarım artışı devam etmekte ve 250.000 belgeden sonra değişmeyen bir başarım gözlenmektedir. Bunun nedeni, derlem büyüklüğünün belirli bir değeri aşmasıyla derlem özelliklerinin kararlı bir duruma erişmesidir.

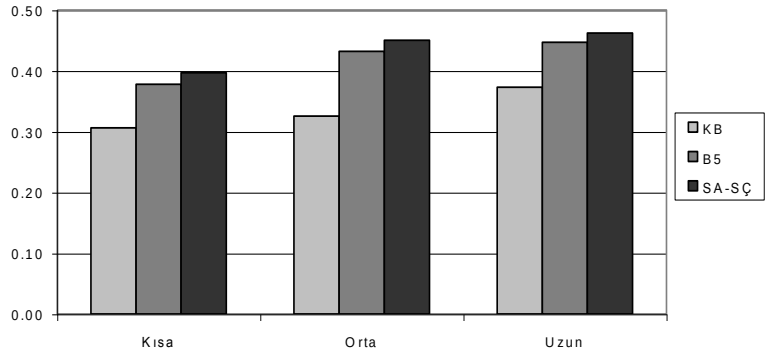
Bu çalışmada KB, B5 ve SA-SÇ yöntemlerinin derlem büyüklüğü açısından göreli başarımlarını inceledik. Sıralama fonksiyonlarının yaklaşık aynı düzeyde etkinlik başarımları elde ettiklerini gözledik. Beklentimizin aksine, Türkçenin özelliklerine göre düzenlenmiş olan SA-SÇ yöntemi yüksek derlem büyüklüklerinde beklenen başarım artışını sağlayamamıştır. Çok

basit olarak baştan 5 harfi alan B5 ve SA-SÇ metotları, bütün derlem büyüklükleri için eşit kabul edilebilecek erişim etkinlik başarımları elde ettiler.

Sorgu Uzunluğunun Etkileri

Belge erişim uygulamalarında kullanıcının ihtiyaçlarına bağlı olarak farklı uzunlukta sorgular olabilir. Örneğin, detayları belirli bir konuyu araştıran kullanıcılar daha çok kelime içeren sorgu girişi yaparken, daha genel konularda araştırma yapan kullanıcılar kısa sorgular kullanır. Etkin bir belge erişim sisteminin farklı uzunlukta sorgular için de etkin sonuçlar üretebilmesi beklenir. Bu nedenle, bu çalışma kapsamında sorgu uzunluğunun başarım üzerindeki etkileri de incelenmiştir. Sorguların uzunluklarına göre gruplanması daha önceki test derlemi başlığı altında anlatılmıştır (Tablo 3 ve 4'e bakınız).

Deney sonuçları Şekil 4'te özetlenmiştir. Şekil 4'te görüldüğü gibi K_S ile O_S arasında B5 için %14,4, SA-SÇ için ise %12,5 gibi önemli bir başarım artışı vardır. Sorgu uzadıkça başarım artışı eğilimi önemli bir seviyede olmasa da devam etmektedir. KB için ise K_S ile O_S arasında %6,23, O_S ile U_S arasında ise %14,59 başarım artışı vardır. Bu KB'nin, sorgu uzunluğu artışı ile daha fazla başarım elde ettiğini gösterir. Sözcüğün kökünün bulunmadan olduğu gibi kullanılmasının başarım üzerindeki kötüleştirici etkisi, uzun sorgularla kısmen ortadan kaldırılmış olur. Deneylerde sorgu uzunluğu ile verimlilik arasında doğrusal bir bağıntı olmadığını gözledik. Sorgu uzunluğu arttığında başarımda başlangıçta bir iyileşme olmakta, sonra başarım artmadan aynı seviyede kalmaktadır. Fakat, KB yönteminde başarım artışı sorgu uzunluğu artışıyla birlikte artmaya devam etmektedir.



ŞEKİL 4. SF8 için farklı sorgu uzunluklarına ait ikili tercih değerleri.

Sorgu uzadıkça başarımın artma nedeni olarak, uzun sorguların kullanıcının ihtiyacını daha doğru ve duyarlı olarak tanımlaması gösterilebilir. Başka çalışmalarda da artan sorgu uzunlukları için benzer sonuçlar rapor edilmiştir. Örneğin Can, Altıngövde ve Demir (2004) Financial Times TREC derlemi üzerinde sorgu uzunluğu artışı için benzer sonuçlar rapor etmişlerdir.

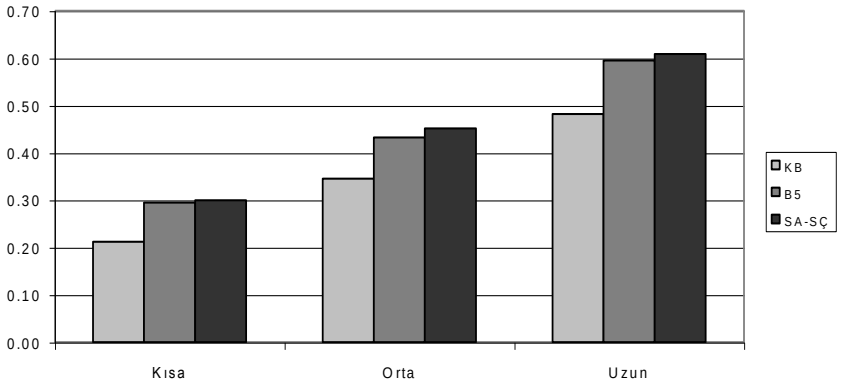
Belge Uzunluğunun Etkileri

Gerçek uygulama ortamlarında farklı uzunlukta belgeler olabilir. Belge uzunluğunun başarım üzerindeki etkilerini incelemek için Milliyet derlemini belge uzunluğuna göre üç alt derleme böldük. İlk derlem 100 sözcüğe kadarlık kısa belgeleri, ikinci derlem 101 ile 300 arasında sözcük içeren orta uzunlukta belgeleri, üçüncü derlem ise 300'den fazla sözcük içeren

uzun belgeleri içermektedir. Ölçeklenebilirlik deneylerinde olduğu gibi sorguları ve ilgili belgeleri alt derlemlere dağıttık. Ölçenebilirlik deneylerinin aksine, bütün sorgular için her üç alt derlemde de ilgili belge vardır. Tablo 9'da da görüldüğü gibi ilgili belgelerin çoğu, orta uzunlukta belgeleri içeren alt derlemdedir. Bunun nedeni kısmen orta uzunlukta belgeleri içeren derlemin neredeyse tüm derlemin yarısını oluşturması, kısmen de çok kısa belgelerin sınırlı bilgi içermesi, çok uzun belgelerin de çok genel bilgi içermesidir.

TABLO 9. Farklı belge uzunluklarını içeren derlemlerin özellikleri

Derlem Belge Türü	Belge Sayısı	Geçerli Sorgu Sayısı	Toplam Farklı İlgili Belge Sayısı	Sorgu Başına İlgili Belge Ortalaması	Sorgu Başına İlgili Belge Medyanı
Kısa	139.130	72	1864	27,50	18,5
Orta	193.144	72	3447	52,14	45,0
Uzun	76.031	72	1612	24,67	21,0



ŞEKİL 5. Farklı belge uzunluğu alt derlemleri için SF8'e ait ikili tercih başarım değerleri.

Deneylerde her sorgunun "konu tanımı" alanı kullanıldı. Şekil 5'te elde edilen ikili tercih başarım değerleri grafik olarak gösterilmiştir. Şekil 5'e göre belge uzunluğu arttıkça, bütün sözcük kökü bulma yöntemleri için ikili tercih ölçüsüyle hesaplanan etkinlik değerleri de istatistiksel olarak anlamlı ölçüde ($p < 0,001$) artmaktadır. Etkinlik değerlerindeki bu sürekli artışın nedeni uzun belgelerin içeriklerini daha iyi yansıtıyor ve bu nedenle erişim aşamasında daha iyi ayrıştırılabiliyor olmalarıdır.

SONUÇLAR VE ARAŞTIRMA OLASILIKLARI

Bu çalışmada çok büyük ölçekli bir Türkçe metin derlemi üzerinde kapsamlı bir belge erişim yöntemleri incelemesi yapılmıştır. Çalışma, hem derlem büyüklüğü hem de kapsadığı konular açısından bir ilktir. Sonuçları, kullandığımız erişim ve bilgi organizasyonu yöntemleri ile sınırlı olmak üzere şöyle özetlenebilir:

- etkisiz sözcük listesinin erişim sistemlerinin etkinliği üzerinde bir etkisi yoktur;

- sistem etkinliği açısından; sözcüğün baştan belirli sayıda harfini alma, sözcük derlemi istatistiklerini kullanarak sözcüğün baştan belirli bir kısmını alma ve ayrıntılı bir analiz ile sözcüğün aslının bulunması yöntemleri çok yakın başarım değerleri vermektedir;
- uzun sorgular sistemin etkinliğini artırmaktadır. Ancak, sorgu uzunluğunun artması ile etkinliğin artması arasında bir doğru orantı yoktur;
- uzun belgelerde etkinlik artmaktadır.

Çalışmamızın bir sonucu olarak; sözcükleri köklerini bularak kullanma yönteminin, Türkçe belge erişim sistemlerinin etkinliği için gerekli olduğunu söyleyebiliriz. Sözcüğü olduğu gibi kullanan KB yöntemi ile karşılaştırıldığında, B5 sözcük kökü bulma yöntemi %33, SA-SÇ sözcük kökü bulma yöntemi ise %38 etkinlik artışı sağlayabilmektedir.

Çalışma sonuçlarının belge erişim sistemi geliştiricileri için birçok olumlu ve pratik çıkarımı vardır. Belge erişim sistemlerinde etkisiz sözcük listesi kullanmamanın başarımı düşürücü bir etkisi yoktur. Aksine, etkisiz sözcük listesinin olmaması, belge erişim sistemi kullanıcılarının bilerek etkisiz sözcük içeren sorgular oluşturmaları durumunda da istedikleri sonucu elde etmelerini sağlar (Witten ve diğer., 1999). Ayrıntılı bir sözcük kökü bulma yöntemi yerine, çok basit bir şekilde, sözcüğün başındaki harfleri kullanma yaklaşımı, hem kolayca programlanabilir hem de sözcük kökü bulmadan sözcüğü olduğu gibi kullanma yöntemine göre önemli bir başarım artışı sağlar. Uzun sorgularda elde edilen daha yüksek bir etkinlik, kullanıcıların beklentileriyle uyuştuğu için istenen bir özelliktir.

Deneylerde, sıralama fonksiyonu SF8'in önemli ölçüde daha iyi erişim başarımı sağladığı gözlenmiştir. '*idf*' kullanımının sözcük ağırlıklarının hesaplanmasını kolaylaştırması nedeniyle, SF8 gerçek hayattaki belge erişim uygulamalarına daha uygundur. Bu çalışmanın en önemli katkılarından biri de, diğer araştırmacılarla paylaşıma açacağımız çok büyük ölçekteki deney derlemidir.

Bu çalışma birçok yönde geliştirilebilir. Örneğin sözcüğün kökünü bulma işlemi, birleşik sözcükleri de göz önüne alacak şekilde geliştirilebilir. Türkçe belge erişim araştırmaları kapsamında "OKAPI" (BM25), dil modelleme (Zobel, Moffat, 2006), müşterek bilgi modeli (Turney, 2002) ve gruplama tabanlı belge erişim (Can ve Özkarahan, 1990; Can ve diğer., 2004; Altıngövde, Özcan, Öcalan, Can, Ulusoy, 2007) gibi farklı modellerin de uygulanması bunlardan bazılarıdır.

TEŞEKKÜR

Deneyleri gerçekleştiren araştırma grubumuzun üyeleri öğrencilerimiz Erman Balçık, Cihan Kaynak, H. Çağdaş Öcalan ve Onur M. Vursavaş ile, kullanılan sorguları hazırlayan ve değerlendiren meslektaş ve öğrencilerimize, değerli yorum ve önerileri için Sengör Altıngövde'ye ve makalenin düzeltimi için zamanını esirgemeyen ve büyük özen gösteren İrfan Karakoç'a teşekkür ederiz.

Bu çalışma, 106E014 numaralı proje ile TÜBİTAK tarafından kısmen desteklenmiştir; çalışmada verilen buluşlar, öneriler ve sonuçlar yazarlara ait olup destekleyen kuruluşla bir ilgisi yoktur.

Ek 1. İndekslemede kullanılmayacak etkisiz sözcükler

ama	böylece	eden	Hiç	mi	olsun	tarafından
ancak	bu	ederek	Hiçbir	mu	olup	üzere
arada	buna	edilecek	İçin	mü	olur	var
ayrıca	bundan	ediliyor	İle	nasıl	olursa	vardı
bana	bunlar	edilmesi	İlgili	ne	oluyor	ve
bazı	bunları	ediyor	İse	neden	ona	veya
belki	bunların	eğer	İşte	nedenle	onlar	ya
ben	bunu	etmesi	İtibaren	o	onları	yani
beni	bunun	etti	İtibariyle	olan	onların	yapacak
benim	burada	ettiği	Kadar	olarak	onu	yapılan
beri	çok	ettiğini	Karşın	oldu	onun	yapılması
bile	çünkü	gibi	Kendi	olduğu	öyle	yapıyor
Bir	da	göre	kendilerine	olduğunu	oysa	yapmak
birçok	daha	halen	Kendini	olduklarını	pek	yaptı
biri	de	hangi	Kendisi	olmadı	rağmen	yaptığı
birkaç	değil	hatta	Kendisine	olmadığı	sadece	yaptığını
Biz	diğer	hem	Kendisini	olmak	şey	yaptıkları
bize	diye	henüz	Ki	olması	siz	yerine
bizi	dolayı	her	Kim	olmayan	şöyle	yine
bizim	dolayısıyla	herhangi	Kimse	olmaz	şu	yoksa
böyle	edecek	herkesin	Mı	olsa	şunları	zaten

Ek 2. Deneylerde kullanılan sorgu konuları

Sorgu No	Sorgu Konusu	Sorgu No	Sorgu Konusu
1	Kuş Gribi	37	Türkiye'de Mortgage
2	Kıbrıs Sorunu	38	ABD Afganistan Savaşı
3	Üniversiteye Giriş Sınavı	39	Yüzüklerin Efendisi-Kralın Dönüşü
4	Tsunami	40	Beyin Göçü
5	Mavi Akım Doğalgaz Projesi	41	Aile Kadın Şiddet
6	Deprem Tedbir Önlem	42	Sporcuların Doping Yapması
7	Türkiye PKK Çatışmaları	43	Ozon Tabakasındaki Delik
8	Film Festivalleri	44	Rusya'da Okul Baskını
9	Bedelli Askerlik Uygulaması	45	İstanbul'da Bombalı Saldırı
10	Stresle Başa Çıkma Yolları	46	Sakıp Sabancı'nın Vefatı
11	Şampiyonlar Ligi	47	Ecevit Sezer Çatışması
12	17 Ağustos Depremi	48	Kıbrıs Türk Üniversiteleri
13	Türkiye'de İnternet Kullanımı	49	Türkiye'de 2003 Yılında Turizm
14	Amerika Irak İşgal Demokrasi Petrol	50	Türkiye'nin Nükleer Santral Çalışmaları
15	Türkiye'de Futbol Şikesi	51	Hızlı Tren Kazası
16	Fadıl Akgündüz	52	YÖK'ün Üniversitelerimiz Üzerindeki Etkisi
17	İşsizlik Sorunu	53	İbrahim Tatlıses'in Kadınları
18	2005 F1 Türkiye Grand Prix	54	Parçalanmış Aileler
19	Ekonomik Kriz	55	Aile İçi Şiddet
20	Nuri Bilge Ceylan	56	Türkiye'de Kanser
21	Türkiye'de Meydana Gelen Depremler	57	Futbol Terörü ve Holiganizm
22	ABD-Irak Savaşı	58	Türkiye'de İkinci El Otomobil Piyasası
23	Hakan Şükür'ün Milli Takım Kadrosuna Alınmaması	59	Tarihi Eser Kaçakçılığı
24	Avrupa Birliği, Türkiye ve İnsan Hakları	60	Festival
25	Turizm	61	Türkiye'de Bayram Tatillerinde Meydana Gelen Trafik Kazaları
26	Türkiye'deki Sokak Çocukları	62	Öğrenmeyi Etkileyen Faktörler
27	Türk Filmleri ve Sineması	63	Kekik Otu
28	Pakistan Depremi	64	Telif Hakları
29	Sanat Ödülleri	65	İnternet ve Toplum
30	Avrupa Birliği Fonları	66	Tarım Hayvancılık Sorunları
31	Futbolda Şike	67	İran'da Nükleer Enerji
32	Milletvekili Dokunulmazlığı	68	Satranç
33	2001 Erkekler Avrupa Basketbol Şampiyonası	69	Kalıtsal Hastalıklar
34	2002 Dünya Kupası	70	Hiperaktivite ve Dikkat Eksikliği
35	Bilişim Eğitimi ve Projeleri	71	Lenf Kanseri
36	Global Isınma	72	28 Şubat Süreci

KAYNAKÇA

Altıngövde, I. S., Özcan, R., Öcalan, H. C., Can, F., Ulusoy, O. (2007). Large-scale cluster-based retrieval experiments on Turkish texts. (Poster makale.) *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (ACM SIGIR '07), basılacak.

Altıntaş, K., Can, F. (2002) Stemming for Turkish: A comparative evaluation. *Proceedings of the 11th Turkish Symposium on Artificial Intelligence and Neural Networks (TAINN 2002)* (s. 181-188) İstanbul: İstanbul University Press.

Bitirim, Y., Tonta, Y. ve Sever, H. (2002). Information retrieval effectiveness of Turkish search engines. *Lecture Notes in Computer Science*, 2457, 93-103.

Buckley, C. ve Voorhees, E. M. (2004). Retrieval evaluation with incomplete information. *Proceedings of the 27th International Conference on Research and Development Information Retrieval* (ACM SIGIR '04) (s. 25-32). Sheffield: ACM.

Bush, V. (1945). As we may think. *The Atlantic Monthly*, 176(1), 101-108.

Cambazoglu, B. B. ve Aykanat, C. (2006). Performance of query processing implementations ranking-based text retrieval systems using inverted indices. *Information Processing and Management*, 42(4), 875-898.

Can, F. ve Özkarahan, E. A. (1990). Concepts and effectiveness of the cover coefficient-based clustering methodology for text databases. *ACM Transactions on Database Systems*, 15(4), 483-517.

Can, F., Altıngövde, I. S., & Demir, E. (2004). Efficiency and effectiveness of query processing in cluster-based retrieval. *Information Systems*, 29(8), 697-717.

Can F. (2006). Turkish information retrieval: Past changes future. *Lecture Notes in Computer Science*, 4243, 13-22.

Can, F., Koçberber, S., Balcık, E., Kaynak, C., Öcalan, H. C. ve Vursavaş, O. M. (2006). First large-scale information retrieval experiments on Turkish texts. (Poster makale.) *Proceedings of the 29th International ACM SIGIR Conference on Research and Development Information Retrieval* (ACM SIGIR '06) (s. 627-628). Seattle: ACM.

Can, F., Koçberber, S., Balcık, E., Kaynak, C., Öcalan, H. C. ve Vursavaş, O. M. (2007). Information retrieval on Turkish texts. *Journal of the American Society for Information Science and Technology*, (yeniden gönderim için düzeltiliyor).

Carterette, B., Allan, J., & Sitaraman, R. K. (2006). Minimal test collections for retrieval evaluation. *Proceedings of the 29th International Conference on Research and Development in Information Retrieval* (ACM SIGIR '06) (s. 268-275). Seattle: ACM.

de Solla Price, D. (1963). *Little science, big science… and beyond*. Columbia University Press, New York.

Ekmekcioğlu, F.C. ve Willett, P. (2000). Effectiveness of stemming for Turkish text retrieval. *Program*, 34(2), 195-200.

Frakes, W. B. ve Baeza-Yates, R. (1992). *Information Retrieval: Algorithms and Data Structures*. Prentice Hall.

Gemmell, J., Bell, G. ve Lueder, R. (2006). MyLifeBits: a personal database for everything. *Communications of the ACM*, 49(1) 89-95.

Hafer, M. A. ve Weiss, S. F. (1974). Word segmentation by letter successor varieties. *Information Storage and Retrieval*, 10, 371-385.

Knuth, D. E. (1973). *The Art of Computer Programming, volume 3: Sorting and Searching*. Addison-Wesley, Reading, MA.

Köksal, A. (1981). Tümüyle özdevimli deneysel bir belge dizinleme ve erişim dizgesi: TÜRDER. *TBD 3. Ulusal Bilişim Kurultayı*, 6-8 Nisan, Ankara, 37-44.

Lee, D. L., Chuang, H. ve Seamons, K. (1997). Document ranking and the vector-space model. *IEEE Software*, 14(2), 67–75.

Long, X. ve Suel, T. (2003). Optimized query execution large search engines with global page ordering. *Proceedings of the 29th Very Large Data Bases Conference* (VLDB 2004) (s.129-140). Berlin: Morgan Kaufmann.

Oflazer, K. (1994). Two-level description of Turkish morphology. *Literary and Linguistic Computing*, 9(2), 137-148.

Pembe F. C. ve Say ACC (2004). A linguistically motivated information retrieval system for Turkish. *Lecture Notes in Computer Science*, 3280, 741-750.

Salton, G. ve Buckley, C. (1988). Term weighting approaches automatic text retrieval. *Information Processing and Management*, 24(5), 513-523.

Sanderson, M. ve Zobel. J. (2005). Information retrieval system evaluation: effort, sensitivity, and reliability. *Proceedings of the 28th International Conference on Research and Development Information Retrieval* (ACM SIGIR' 05) (s. 162-169). Salvodor: ACM.

Saracevic, T. (1999). Information Science. *Journal of the American Society for Information Science*, 50(12) 1051-1063.

Sever, H. ve Bitirim Y. (2003). FindStem: analysis and evaluation of a Turkish stemming algorithm. *Lecture Notes in Computer Science*, 2857, 238-251.

Sever, H. ve Tonta, Y. (2006). Truncation of content terms for Turkish. *CICLing, Mexico* (basılacak).

Solak, A. ve Can, F. (1994). Effects of stemming on Turkish text retrieval. *Proceedings of the Ninth Int. Symp. on Computer and Information Sciences* (ISCIS '94) (s. 49-56). Antalya.

Sparck Jones, K. (1981). Retrieval system tests. K. Sparck Jones (Ed.), *Information retrieval experiment* (s. 213-255). London: Butterworths.

Toffler, A.: Future Shock. Bantam Books, New York , 1990 (ilk basılış tarihi: 1970).

Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting of the Assoc. for Computational Linguistics* (ACL 2002), (s. 417-424). Philadelphia.

Varian, H. R. (2005). Universal access to information. *Communications of the ACM*, 48(10) 65-66.

Witten, I. H., Moffat, A. ve Bell, T. C. (1999). *Managing gigabytes: Compressing and indexing documents and images, 2nd ed*. San Francisco, CA, Morgan Kaufmann.

Zobel, J. (1998). How reliable are the results of large-scale information retrieval experiments? *Proceedings of the 21st International Conference on Research and Development Information Retrieval* (ACM SIGIR' 98) (s. 307-314). Melbourne: ACM.

Zobel, J. ve Moffat, A. (2006). Inverted files for text search engines. *ACM Computing Survey*, 38(2), 1-56.

1

Türk Dil Kurumu (TDK) tarafından basılması planlanan bir kitap için, TDK'nın daveti üzerine, hazırlanan bu makale daha sonra yayımlanan *JASIST*'deki makalemizin öncülü olma niteliğindedir: Can, F. , Kocberber, S., Balcik, E., Kaynak, C., Ocalan, H. C., Vursavas, O. M. "Information retrieval on Turkish texts.*" Journal of the American Society for Information Science and Technology*. Vol. 59, No. 3 (February 2008), pp. 407-421). Söz konusu kitap yayımlanmamıştır. Web'e konuluş tarihi: 8 Kasım, 2015.