Türkçe Haberlerde Yeni Olay Bulma ve İzleme: Bir Deney Derleminin Oluşturulması

New Event Detection and Tracking in Turkish News: Construction of a Test Collection

Fazlı Can, Seyit Koçberber, Özgür Bağlıoğlu, Süleyman Kardaş, H. Çağdaş Öcalan ve **Erkan Uyar**

Bilkent Bilgi Erişim Grubu, Bilgisayar Mühendisliği Bölümü, Bilkent Üniversitesi, Ankara 06800
{canf,ozgurb,skardas,hocalan,euyar}@cs.bilkent.edu.tr, seyit@bilkent.edu.tr

Özet: Yeni olay belirleme ve izleme (YOBİ) sistemleri zaman sırasıyla gelmekte olan haberlerin içindeki yeni olaylara karşılık gelen ilk hikâyeleri ve onların devamı olan haberleri saptamayı amaçlar. Bu bildiride, YOBİ sistemlerinin özellikleri anlatıldıktan sonra, Türkçeye yönelik YOBİ sistemlerinin etkinliğinin ölçümünde kullanılacak bir deney derleminin nasıl geliştirildiği anlatılmaktadır. Derlem, Web'deki beş haber kaynağından indirilen 2005 yılına ait saat ve dakika detayında zaman damgası içeren 200 bini aşkın haber belgesi içermektedir. Bildiride, ilk hikâye ve izleyen haberlerin yarı-özdevimsel olarak insanlar tarafından saptanması için geliştirilmiş olan sistemin tasarımı, yapımı ve haber indiriminin nasıl yapıldığı anlatılmakta, indirilen ve izlenen haberlerin deney derlemi içindeki dağılımıyla ilgili bazı sonuçlar verilmektedir.

Anahtar sözcükler: Yeni olay belirleme ve izleme (YOBİ), bilgi erişim, haber portalı, ilk hikâye

Abstract: A new event detection and tracking system (NEDT) aims to determine the first story and tracking news for the new events among the news articles that arrive in temporal order. In this paper, after introducing NEDT systems we describe the construction of a test collection for measuring the effectiveness of such systems designed for Turkish. The collection contains more than 200 thousand news articles, time stamped at the detail level of hour and minute, downloaded from five different Web news resources. In the paper, we describe the design and construction of a semi-automatic system that helps determine the first story and tracking news, how we download the news, and some results on the distribution of the tracking news in the test collection.

Keywords: first story, information retrieval, new event detection and tracking (NEDT), news portal

Giriş

Bilgi erişim (BE – information retrieval) ve bilgi süzme (BS – information filtering) sistemlerine benzer fakat farklı bir uygulama olan yeni olay belirleme ve izleme (YOBİ) sistemleri son yıllarda Web'deki haber kaynaklarının artmasıyla birlikte önem kazanmıştır. Bir YOBİ ortamında, sisteme zaman sırasıyla sürekli olarak gelmekte olan haberlerin içinde yeni olaylara karşılık gelen haberler ve bu ilk hikâyelerin devamı olan haberler saptanır. Bir YOBİ sistemi bilinmeyenlerin özdevinimli (otomatik) olarak keşfedilmesini amaçladığı için bir veri madenciliği uygulamasıdır (Witten ve Frank, 2000).

YOBİ için geliştirilen algoritmaların haber portallarında kullanımından önce etkinliklerinin deneysel olarak saptanması gerekir. Bu amaçla, insanlar tarafından ilk hikâyesi saptanmış ve izlenmiş haberlerden oluşan deney derlemleri kullanılır (Papka, 1999). Bu derlemler sayesinde, geliştirilen algoritmaların insanlar tarafından saptanmış gerçek duruma ne denli uyum gösterdiği yanlış ikaz (false alarm), Yİ, ve kaçırma oranı (miss rate), KO, gibi çeşitli kıstaslarla ölçülür. Böylece, bu algoritmaların gerçek bir uygulamada kullanıcıların etkinlik beklentilerini ne denli karşılayacağı saptanır. Etkinlik ölçümünde kullanılan Yİ konuyla ilgisiz olduğu halde ilgili sanılarak bulunan haberlerin, KO ise bulunması gerektiği halde kaçırılan haberlerin oranını yansıtır. Sistem geliştirimi sırasında deney derlemi yardımıyla, sistemin çeşitli parametrelerinin elde edilen başarıdaki etkileri saptanır ve bu parametrelerin daha iyi sonuç verecek biçimde seçilmesi sağlanır. YOBİ deney derlemleri sayesinde farklı yaklaşımlarla yapılmış olan sistemlerin etkinlikleri birbirleriyle karşılaştırılabilir. Bu karşılaştırma işlemi ortak kullanılan standart deney derlemleri ile yapıldığı takdirde farklı sistemlerin göreceli etkinliği daha sağlıklı bir biçimde saptanacağı için araştırmacılar ne yapmaları gerektiği konusunda doğru kararlar verebilirler. Standart deney derlemlerinin yapılan araştırmaların düzeyini yükseltici olumlu etkisi literatürde kanıtlanmıştır (Voorhees, 2005).

Bu çalışmada YOBİ sistemlerinin özellikleri açıklandıktan sonra Türkçe haberler için geliştirmekte olduğumuz YOBİ yöntemlerini değerlendirmek amacıyla

kullanacağımız bir YOBİ deney derleminin nasıl geliştirildiği anlatılmaktadır. Derlem, Türkçe YOBİ uygulamalarında bir standart olması amacıyla hazırlanmaktadır. Bildiride YOBİ sistemlerinin özellikleri, bu konuda yapılan çalışmalar, deney derlemi için kaynak seçiminin nasıl yapıldığı, haber kaynaklarından haber indirmek için kullanılan yöntemler ve indirilen HTML dosyalarından haberlerin nasıl ayıklandığı anlatılmaktadır. Daha sonra deney derleminde ilk hikâye ve izleyen haberlerin insan tarafından saptanması için geliştirilmiş olan sistemin tasarımı, yapımı ve izlenen haberlerin derlem içindeki dağılımıyla ilgili örnek bulgular verilmekte ve makale gelecekte yapılabilecek araştırmalara ilişkin notlarla birlikte sonuçlandırılmaktadır.

YOBİ Sistemleri

Internet'in bilgi ve belge paylaşımını elektronik ortamda sağlamasıyla belge sayısında bir patlama yaşanmıştır. Kullanıcıların büyük belge yığınları arasından ihtiyaç duyduğu bilgiye erişimini sağlamak için bilgi erişim (BE) konusunda çok sayıda araştırma yapılmıştır. Yeni yayınlanan belgelerin ilgi alanlarına göre kullanıcılara dağıtılması için bilgi süzme (BS) konusu da benzer biçimde araştırmaların yoğun olduğu bir alandır (Kobayashi ve Takeda, 2000).

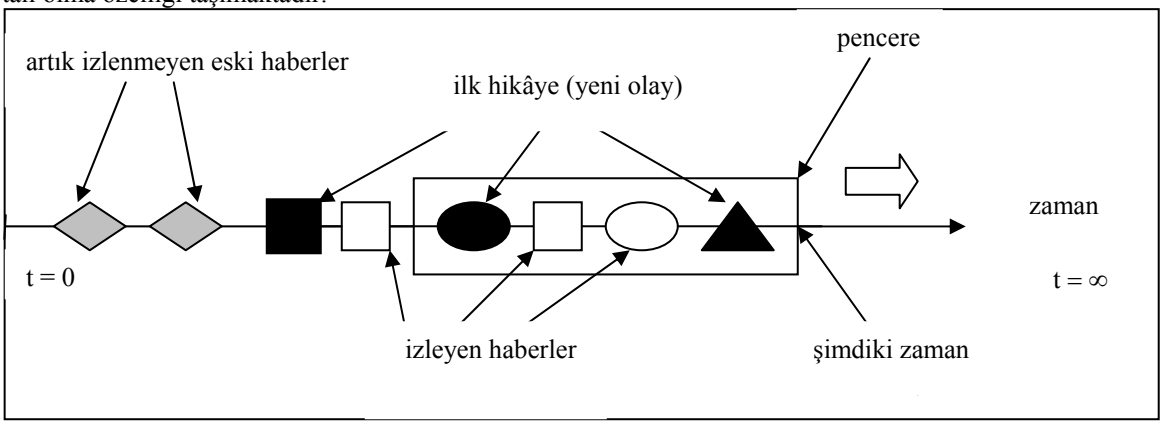
Internet'in haberlerin yayınlanmasında kullanılmasıyla, bilgi ve belge konusunda yaşanan patlamaya benzer bir haber patlaması yaşanmaya başlanmıştır. Günümüzde haber üreten ve yayınlayan muhabir, haber ajansı, basın yayın organı kanallarına yakın gelecekte yeni haber üretim kanalları eklenecektir. Haber kaynağı sayısındaki patlamayla birlikte tüm haberlerin izlenmesi yerine, yeni haberlerin kullanıcılara bildirilmesi ve birbirinin devamı olan haber zincirlerinin takip edilmesini sağlayacak servislerin geliştirilmesi zorunlu olmaktadır. Örneğin NewsIsFree Web servisi 43 farklı dilde 24.000'den fazla Web haber kaynağı listelemektedir (NewsIsFree, 2007; Radev, Otterbacher, Winkel ve Balir-Goldensohn, 2005). Google News (2007) 4500 değişik haber kaynağını tarayan Türkçe dışında çok sayıda dilde YOBİ hizmeti sunan bir haber portalı olma özelliği taşımaktadır.

Haberin değeri ve önemi zamanla değişmektedir. Basın yayın kurumları bir haberi diğerlerinden önce duyurabilmek için kıyasıya bir yarış içindedir. Haberde bulunan bu zaman boyutu ve buna bağlı olarak ortaya çıkan aciliyet özelliği, yeni olayların saptanması uygulamasında da göz önüne alınmalıdır. Haber geçmiş haberlerle benzerlik gösterse bile, yeniliğini ve tazeliğini ön plana çıkarmak için haberlere bir zaman penceresinden bakmak gerekmektedir.

YOBİ'leri BE sistemlerinden ayıran en temel özellik, haberin yapısında önemli bir öge olan yer, kişi ve zaman bilgileridir. YOBİ'lerde BE sistemlerinde kullanılan benzerlik hesaplama yaklaşımından farklı bir benzerlik ölçüsü geliştirilmesi veya bu türden (kişi, zaman, vb. gibi) bilgilere benzerlik ölçümünde daha fazla ağırlık verilmesi etkinliği artırmaktadır (Kumaran ve Allan, 2004).

Yukarıda anlatılan nedenlerle YOBİ yeni araştırma olanakları içeren bir alandır. Araştırmalar sonunda geniş toplum kesimlerinin kullanacağı, haber tüketicilerinin (news consumers) yeni haberlere daha etkin ve randımanlı bir şekilde erişmesini sağlayacak sistemler ortaya çıkmaktadır.

Yeni olay belirleme ve izleme araştırmalarında "yeni olay" (new event) belli bir zamanda ve belli bir yerde gerçekleşen aksiyon olarak tanımlanmıştır. Yeni olaylar beklenmedik bir biçimde oluşabileceği gibi zamanı önceden bilinen haberler de olabilir (örneğin, Türkiye'de 2007 yılında yapılacak olan genel seçimler). Bu noktada birbirine benzeyen ama farklı olan "olay" ve "konu" (topic) kavramlarını tanımlamakta yarar vardır. Örneğin "deprem" bir "konudur," fakat "17 Ağustos 1999" depremi oluştuğu anda "yeni bir olay"dır. Zaman geçtikçe bir aksiyon sonuçlarıyla birlikte bir konu haline dönüşebilir. Örneğin, ABD'nin Irak'ı işgali ilk günlerinde "yeni olay" tanımına uymaktadır. Ancak, zaman içinde bu durum bir "konu" haline dönüşmüştür. Olayların yeniliği göreceli ve bir zaman çerçevesi içinde geçerli olduğu için, yeni olay kavramına bir zaman penceresi içinde bakmak gerekir (Şekil 1).



Şekil 1. Yeni olay belirleme ve izleme (YOBİ).

Konuyla İlgili Çalışmalar

Yeni olayların belirlenmesi problemine yönelik ilk çalışmalar 1996 yılında Carnegie Mellon Üniversitesi, DARPA (US Department of Defense Advanced Research Projects Agency), Dragon Systems ve Amherst'teki Massachusetts Üniversitesi tarafından yürütülen konu belirleme ve izleme, Topic Detection and Tracking ([TDT], 2007) projesiyle başlatılmıştır. Bu projeden önce bu konuda yapılan çalışmalara literatürde az rastlanmaktadır (Papka, 1999, s. 9).

YOBİ konusundaki ilk kapsamlı araştırmalardan biri TDT projesinin kapsamı içindeki Papka'nın (1999) çalışmasıdır. Papka kümeleme (clustering) algoritmalarını YOBİ'de kullanmış ve özel isimlerin haberlerde kullanımına önem vererek sistem etkinliğini yükseltmiştir.

Geliştirilen yöntemin INQUERY bilgi erişim sistemi üzerinde çalışması ve bu sisteme bağlı parametreleri en iyileştirmeyle kullanması nedeniyle araştırmanın sonuçlarını tekrarlamak ve başka uygulamalara yansıtmak zordur. Aynı proje çerçevesinde Carnegie Mellon Üniversitesinde Yang, Pierce ve Carbonell (1998) tarafından yapılmış olan çalışmada ağaç yapılı (hierarchical) ya da tek düzeyli kümeleme algoritmaları anında ya da geriye dönük olarak ilk haber saptamada kullanılmıştır. Yang, Carbonell, Brown, Pierce, Archibald ve Liu (1999) yaptıkları çalışmada pratik uygulamalar için birtakım önemli noktalara işaret etmektedirler. Bunların bazıları küme temsilinin kullanıcıların aramalarını kolay biçimde yapabilmeleri için düzenlenmesi, bulunan haber kümelerinin kullanıcılar için özetlenmesi ve kullanıcılar tarafından verilecek olan bilgilerin küme yapısına yansıtılmasıdır.

YOBİ sürecinin birinci adımı olan "ilk hikâye"yi bulmanın zorlukları Allan, Lavrenko ve Jin (2000) tarafından incelenmiştir. Çalışmalarında amaç ilk hikâyeyi bulmak olup, "ilk hikâye" yakalama, problemi izleme (tracking) problemine indirgenmektedir. Deneylerde önceki haberlerin hiçbirine yeterince benzemeyen (bir eşik değeri –threshold- kullanarak) bir haber yeni bir olay olarak alınmaktadır. Çalışma ayrıca ilk haberlerin bulunmasının bir izleme problemi olarak çözüldüğü varsayımı ile yanlış ikazın üst sınırının nasıl hesaplanacağını da göstermektedir.

Yeni olayların saptanmasında rol oynayacak çeşitli etmenler (sözcüklere verilecek önem ağırlığı, önemsiz kelimelerin ayıklanması vb. gibi), yeni olay belirleme üzerindeki etkileri ve haber vektörlerinin farklı biçimlerde üretilip olumlu yanlarının birleştirilmesi gibi yaklaşımlar Kumaran ve Allan (2004) tarafından incelenmiştir. Çalışmada ad verilmiş nesnelere daha fazla önem verilmesi önerilmekte ve bu yaklaşımın olumlu etkileri gösterilmektedir. Hatzivassiloglou, Gravano ve Maganti (2000) tarafından yapılan çalışmada dört kümeleme yöntemi ve bazı dil özelliklerinin haber vektörlerine taşınması halinde başarıma (performans) etkileri ölçülmüştür.

Konuyla ilgili Türkiye'de yapılmış saptayabildiğimiz az sayıdaki çalışmadan ilki Kurt'un (2001) Türkçe'de YOBİ konusunda yaptığı yüksek lisans tezidir. Bu çalışma yaklaşık 47.000 belge içermektedir. Geliştirilen yöntem ile 15 haberin gelişimi izlenerek değerlendirilmiştir. Haberler için dört kaynak kullanılmış olup bu kaynakların derlem içindeki katkıları %69, %18, %11 ve %2 biçiminde dağılmıştır (Kurt, 2001, s. 55-56). Kullanılan 15 haberin toplam haberlerdeki belge payı %2,8'dir (Kurt, 2001, s. 38). Tezde geliştirilen yöntem van Rijsbergen'de de anlatılan (1979, s. 52) tek geçişli bir kümeleme algoritması ile k-NN kümeleme algoritmasının bir birleşimidir (Kurt, 2001, s. 37).

Türkiye'de YOBİ konusunda saptayabildiğimiz diğer çalışma Vural'ın (2002) yüksek lisans tezidir. Vural kapsama katsayısı kavramından (Can ve Özkarahan, 1990; Can, 1993) ve TREC (Text Retrieval Conference, http://trec.nist.gov/) katılımcıları tarafından da kullanılan TDT (2007) deney derleminden yararlanmıştır.

Kullanılan Haber Kaynaklarının Seçimi

Deneyler için 2005 yılına ait haberlerin alınmasına karar verilmiştir. Derlemin kapsamış olduğu zaman süresi ve sonuçta elde edilen haber sayısı daha önce bu konuda yapılmış olan başka çalışmalarla uyumlu ve hatta daha uzun ve daha kapsamlı olduğu söylenebilecek niteliktedir (TDT, 2007; Kumaran, Allan ve McCallum, 2004; Papka, 1999). Kapsamış olduğu zaman aralığı sonuçta gerçekleştirilecek olan gerçek zamanlı uygulama ortamının özelliklerini yansıtmaktadır.

Çalışmada kullanılacak deney derleminin hazırlanmasında aşağıda belirtilen Web haber kaynaklarından yararlanılmış ve bu kaynaklardaki 2005 yılına ait saat ve dakika damgalı bütün haberler indirilmiştir.

- CNN Türk (http://www.cnnturk.com);
- Haber 7 (http://www.haber7.com);
- Milliyet Gazetesi (http://www.milliyet.com.tr);
- TRT (http://www.trt.net.tr);
- Zaman Gazetesi (http://www.zaman.com.tr).

Bu haber kaynakları

- Gün bilgisine ek olarak haberlerin yayınlanma zamanını istenilen ayrıntıyla (saat, dakika) vermektedir. Bu ayrıntı YOBİ işleminde önemlidir. Yararlanılan kimi haber kaynaklarında bu durum "son dakika" vb. gibi haberler için geçerli olabilmektedir. Bu gibi durumlarda sadece bu türden ayrıntılı zaman bilgisi içeren haberler deney derlemine dahil edilmiştir;
- Görüşleri doğrultusunda haberlere verdikleri önem ve farklı yorumlarıyla çeşitlilik sağlarlar. Farklı görüşlere sahip oldukları söylenebilecek haber kaynaklarını kullanmak, haberlerin çeşitli yönlerini yakalayabilecek bir sistemin geliştirilmesini destekleyecektir.

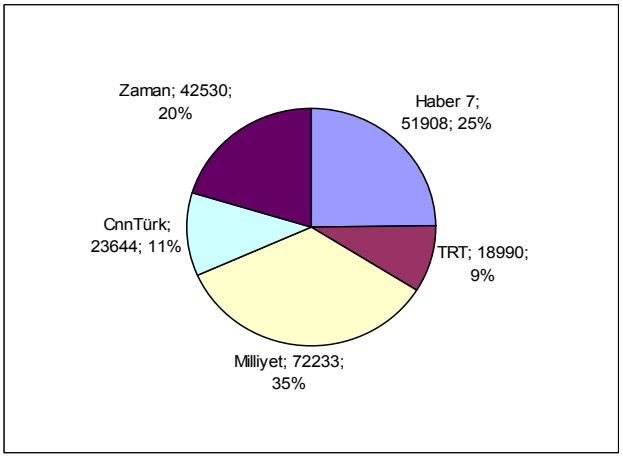
Bu haber kaynaklarından deney derlemi oluşturmak amacıyla indirilen haberlerle ilgili özet bilgi Tablo 1'de verilmiştir. Tabloda da görüldüğü gibi indirilen sayfalar (içerdikleri resim, reklam, vb. gibi bilgilerden ötürü) indekslemek için üretilen XML dosyalarına göre yaklaşık yirmi (17.4) kat daha büyüktür. Tablo 1'in üçüncü sütunu ve Şekil 2 haberlerin kaynaklar arasındaki dağılımını göstermektedir. Haberlerin en çoğu (%34,5) Milliyet gazetesinden, en azıysa (%9,1) TRT'den indirilmiştir. Derlemde toplam 209.305 haber bulunmaktadır. Haberlerin bütün zamanlara eşit dağıldığı varsayılırsa saat başına yaklaşık 24 ve her 150 saniyeye 1 haber düşmektedir. Ancak haberlerin kaynaklarına tek tek bakıldığında ya da tüm haber kaynaklarına bir arada bakıldığında, farklı günlerde farklı sayılarda haber olduğu görülmektedir. Örneğin hafta sonları daha az sayıda habere rastlanmaktadır (bkz. Şekil 3). Bu gözlemler, deney derleminin gerçek bir uygulamada karşılaşılacak durumları temsil ettiğini gösterecek niteliktedir. Haber kaynaklarının farklı günlerde farklı haber üretmesi, yeni olay tanımada kullanılabilecek olan pencerenin zaman uzunluğu yerine haber sayısıyla tanımlanmasının daha uygun olacağına işaret etmektedir. Benzer bir biçimde Yang, Pierce ve Carbonell (1998) sabit sayıdaki haberi (400) kayan zaman penceresi tanımında kullanmışlardır.

Haber İndirme Yöntemleri

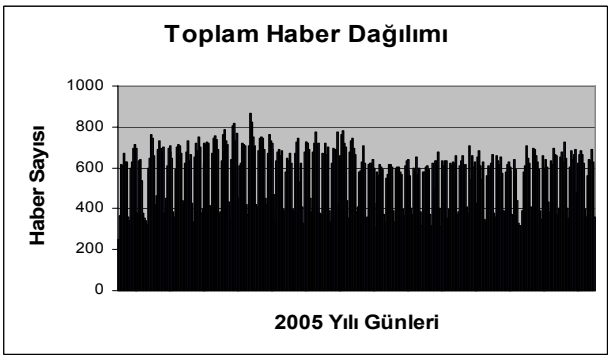
Bu bölümde haber kaynaklarından haber indirme yöntemleri ve Web sayfalarının indirilmesinden sonra oluşturulan dosyaların HTML etiketlerinden ayıklanması işlemi anlatılmaktadır. YOBİ çalışmalarında unutulmaması gereken bir nokta haber kaynaklarının haber Web sayfalarında kullandıkları HTML etiket yapılarını değiştirebilecekleridir. Bu nedenle haber portalı uygulamalarında Web sayfalarındaki HTML etiketlerini ayıklayacak olan alt sistemin esnek olması ve bir kaynakla ilgili olağandışı bir durum gördüğünde sistemin personeli uyarması gerekmektedir. Yazılım bu gibi durumları düşünerek geliştirilmelidir.

CNN Türk

CNN Türk Web sayfasında geçmiş tarihlere ait haberlere erişim her haber için farklı olarak verilmiş olan bir numara ile sağlanmaktadır. Ancak, erişim haberin kategorisine bağlı olarak belirlenen dokuz adet alt kategoriye ait Web adresinden biri üzerinden sağlanmaktadır. Son güne ait haber numarası ve kategori eşlemesi Web sitesinde mevcuttur. Ancak, Web sitesinde geçmiş tarihler için kategorilere ait haber numaralarını saptamak veya haber numarasının kategorisini saptamak için bilgi mevcut değildir. Bu nedenle 2005 yılına ait haberlere erişimde CNN Türk Web sayfasında bulunan arama motorunu kullanarak 2005 yılına ait haberlerin linklerini elde edecek bir yazılım sistemi geliştirilmiştir.



Şekil 2. Haberlerin kaynaklar arasındaki dağılımı[1]



Şekil 3. Haberlerin yıl içindeki günlere sayısal dağılımı

Tablo 1. İndirilen haberlerle ilgili özet bilgi

Haber Kaynağı	Haber Sayısı	Yüzde Katkısı	İndirilen Bilgi (MB)	Ayıklanmış Bilgi (MB)	Yaklaşık Kelime Sayısı
CNN Türk	23.644	11,3	1008,3	66,8	271
Haber 7	51.908	24,8	3629,5	107,9	238
Milliyet Gazetesi	72.233	34,5	508,3	122,5	218
TRT	18.990	9,1	937,9	18,3	121
Zaman Gazetesi	42.530	20,3	45,3	33,7	97
Hepsi birlikte	209.305	100,0	6051,7	348,3	200

Yazılım sistemi 2005 yılına ait haberleri dört adımda indirmektedir. Bu adımlar aşağıda anlatılmıştır.

1. Harf ve rakamlardan oluşan ve ilk 2 karakteri belli olan kısmî uyum (partial match) sorguları (aa*, ab*, …) oluşturulur ve CNN Türk arama servisine gönderilir. Sorgu sonucuna göre her biri 200 adet haber içerecek şekilde sorgu sayfası sonuçlarını indirecek linkler hazırlanır;
2. Adım 1'de oluşturulmuş her bir linke ait haber özetlerini gösteren Web sayfaları indirilir;
3. Adım 2'de indirilen sorgu sonuç sayfaları taranarak haber linkleri saptanır. Haber linklerinde haberin tarihi de bulunmaktadır. Haber tarihi ve linki daha önceki sorgularda elde edilmişse dikkate alınmaz, yeni bir haberse indirilecek haber linkleri arasına eklenir;
4. Adım 3'de saptanmış haber linkleri içinden 2005 yılına ait olanlar her bir haber ayrı bir dosyada olacak şekilde indirilir.

Örneğin, içinde "az" ile başlayan kelime geçen haberlere erişmek için aşağıdaki link kullanılır:

http://www.cnnturk.com/arama/arama.asp?PID=3&QU=az*

"az" ile başlayan 201. ile 400. sıradaki haberlere erişmek için ise aşağıdaki link kullanılır:

http://www.cnnturk.com/arama/arama.asp?PID=3&QU=az*&MR=200&SH=201

CNN Türk için uygulanan bu yöntemle elde edilen haber numaraları incelendiğinde 2005 yılına ait en küçük ve en büyük haber numarası arasında indirilmemiş çok sayıda haber olduğu saptanmıştır. Tarama sisteminde yukarıda anlatılan yöntemle erişim bilgisi (path) bulunamayan haber numaralarının gerçekte bir habere karşılık gelip de bu haberin arama sistemine dahil edilmemiş olma olasılığı incelenmektedir. CNN Türk'de indirilen haberler arasında İngilizce haberler de bulunmaktadır. Bu haberler İngilizcede sık kullanılan "the" ve benzeri kelimeler kullanılarak ayıklanmıştır. Eğer haberlerde bu kelimeler belli sayıdan fazla geçerse bunlar veri tabanına eklenmemektedir; çünkü Türkçe haberlerde de, mesela bir filmi anlatıyorsa ("The Independence Day" gibi) bu türden sözcükler geçebilmektedir.

Haber 7

Haber 7 Web sitesinde gün boyu verilen haberlere bir numara verilmekte ve daha sonra bu haberler arşivde saklanmaktadır. Haber numarası bilinen bir habere doğrudan haber numarası verilerek erişilebilmektedir. Örneğin, haber numarası 100000 olan bir habere aşağıdaki linkten erişilmektedir.

http://www.haber7.com/haber.php?haber_id=100000

Haber 7 Web sitesinden haber indirmek için hazırlanan robot program verilen iki haber numarası arasındaki tüm haberleri indirecek şekilde düzenlenmiştir. 2005 yılına ait haberlerin en küçük ve en büyük haber numaraları saptanarak haber indirme programına parametre olarak verilmiştir. İndirilen dosyalar haber numarası, dosya adı olacak şekilde saklanmıştır.

Haber 7 Web sunucusu aynı IP numarasından kısa bir süre içinde gelen çok sayıda haber indirme isteğiyle karşılaştığında haber indirmeye çalışan bilgisayarın IP numarasını kara listeye eklemekte ve bir süre bu adresten gelen isteklere cevap vermemektedir. Hazırlanan robot programa cezalı duruma düşmeyi önleyici parametreler eklenmiştir.

Milliyet Gazetesi

Milliyet Gazetesi Web sitesinde günlük gazetede çıkan haberlerin yanı sıra son dakika haberleri adı altında haberin oluş zamanını da içerecek şekilde günlük liste verilmektedir. Önceki tarihlere ait listelere

http://www.milliyet.com.tr/YYYY/AA/GG/son

linkiyle erişilebilmektedir. Bu linkde "YYYY" yıl, "AA" ay ve "GG" gün olacak şekilde son dakika haberlerine erişilecek tarihi temsil etmektedir. Örneğin, 23 Nisan 2005 tarihindeki son dakika haberlerinin listesine erişmek için aşağıdaki link kullanılır.

http://www.milliyet.com.tr/2005/04/23/son/

Bir güne ait son dakika haberleri "dünya," "ekonomi," "siyaset," "spor," "Türkiye" ve "yaşam" olacak şekilde altı kategoriye ayrılmıştır. Haberlere Türkçeye özel harf kullanmadan kategori adının ilk üç harfi kullanılarak ve her gün için baştan başlayacak şekilde iki basamaklı bir haber numarası verilerek oluşturulan statik linklerle erişilmektedir. Örneğin, "Türkiye" kategorisinde 25/03/2005 tarihindeki 45 numaralı habere

http://www.milliyet.com.tr/2005/03/25/son/sontur45.html

linkiyle erişilmektedir.

Milliyet Gazetesi son dakika haberlerini indiren robot programa parametre olarak başlangıç ve bitiş tarihleri verilmektedir. Robot program indirdiği haberleri kategori bilgisini de dosya adına ekleyerek saklamaktadır.

TRT

TRT Web sitesinde eski tarihli haberlere erişim Haber 7'dekine çok benzemektedir. Örneğin, haber numarası 120000 olan bir habere aşağıdaki linkten erişilmektedir.

http://www.trt.net.tr/wwwtrt/hdevam.aspx?hid=120000

TRT haber sitesinden haber indirmek için hazırlanan robot program verilen iki haber numarası arasındaki tüm haberleri indirecek şekilde düzenlenmiştir. 2005 yılına ait haberlerin en küçük ve en büyük haber numaraları saptanarak haber indirme programına parametre olarak verilmiştir. İndirilen dosyalar haber numarası dosya adı olacak şekilde saklanmıştır.

Zaman Gazetesi

Zaman Gazetesi sitesinde bir güne ait haberlerin tümüne bir defada erişilebilmektedir. Örneğin, 10 Nisan 2005 tarihine ait bütün haberlere aşağıdaki linkten erişilmektedir.

http://www.zaman.com.tr/sdk_hepsi.php?trh=20050410

Zaman Gazetesi sitesinden haber indirmek amacıyla hazırlanan robot program parametre olarak bir tarih aralığı kabul etmektedir. İndirilen haberler her gün için ayrı bir dosya olacak şekilde saklanmaktadır.

Zaman Gazetesi, biz 2005 yılına ait haberleri indirdikten sonra Web sitesinde servis veren yazılımını değiştirmiştir. Yeni yazılım için haber indirme robot programı TRT için geliştirilen haber indirme programına benzer olarak haber numarası aralığıyla çalışmaktadır.

İndirilen Sayfalardan Haberlerin Ayıklanması

Haber kaynaklarından haberleri içeren Web sayfalarının indirilmesinden sonra gelen aşama bu HTML dosyalarından HTML etiketlerinin (HTML tag) ve haber dışındaki reklam ve diğer sayfalara olan linklerin ayıklanması işlemidir. Haber sayfalarının Web programları tarafından oluşturulması, sayfa kaynak kodunda belirli karakter dizilerinin ayıklanmak istenilen bilgiler çevresinde tutarlı olarak gözlenmesini sağlar. Bu nedenle aynı kaynaktan gelen HTML dosyaları sınırlı sayıda çeşitlilikle kendi içinde tutarlı bir yapıya sahiptir. Bu dosyalardan haberlere ait haberin zamanı (gün/ay/yıl saat:dakika detayında), başlığı ve içeriği olmak üzere üç temel bilgi ayrıştırılır.

Haber özelliklerini ayıklama işlemini yapan program Java dilinde yazılmıştır. Java'yı seçmemizin başlıca nedeni karakter dizgisi (string) işlemlerinin bu dilde oldukça kolay bir biçimde yazılabilmesidir. Ayrıca Java dilinde yazılmış HTML ayıklamaya yarayan birçok kütüphane de mevcuttur. Bu amaçla literatürde mevcut hazır kütüphanelerden biri sayfadaki HTML etiketlerini temizlemek ve haberle ilgili gerekli bilgileri sayfadan almak için kullanılmıştır. Bu kütüphane sourceforge.net'in "HTML Parser" kütüphanesidir (HTML, 2007).

"HTML Parser" kütüphanesi HTML sayfalarını doğrusal veya iç içe biçimde ayıklamak için kullanılmaktadır. Ana kullanım amacı sayfalardaki belli bilgileri ayıklamak, HTML etiketlerini sayfanın yapısından çıkarmak, HTML sayfalarının yapısını değiştirmek ve yenilemektir. Bu kütüphane hızlı ve doğru biçimde ayıklama işlemini geçekleştirmektedir.

"HTML Parser" kütüphanesi kullanılarak her bir haber kaynağı için özelleştirilmiş yöntemlerle haberler ayıklanmaktadır.

İlk ve İzleyen Haberlerin Saptanması

Deney derleminin oluşturulmasında Amerika Birleşik Devletleri'nde DARPA tarafından desteklenen TDT projesine benzer bir yol izlenmiştir. TDT projelerinde, yıllar içinde çeşitli biçimlerde deney derlemi geliştirilmiştir. Önceleri bütün haberleri gözden geçirerek yaratılan deney derlemleri (Papka, 1999) daha sonra yine insanlar tarafından fakat yazılım yardımıyla da yapılmıştır (TDT, 2004). Bizim yaklaşımımız TDT projesinde 2004'de kullanılan yaklaşımdan esinlenerek gerçekleştirilmiştir. İzleyen haberleri bulmak için bir BE sistemi kullanıldığından ötürü yarı-özdevimsel bir niteliktedir (TDT, 2004).

Yeni olayların ve onları izleyen haberlerin saptanması için hazırladığımız sistem ETracker (Event Tracker), Microsoft .net ortamında C# ile geliştirilmiş bir Web uygulamasıdır. Çalışmalarımızı proje sonrası aşamalarda İngilizce haberleri de kapsayacak şekilde genişletmeyi planlamamız nedeniyle ETracker'in kullanıcı arayüzü İngilizcedir.

Bu çalışmada eldeki bütün haberler Türkçe bilgi erişimde etkinliği kanıtlanmış bir yöntemle indekslenmiştir. Bu yöntem Can, Koçberber, Balçık, Kaynak, Öcalan ve Vursavaş'ın (2006; yayın aşamasında) yaptığı çalışmada MF8 olarak tanımlanmaktadır. MF8 bilgi erişimde *tf.idf* (term frequency * inverse document frequency) yaklaşımını kullanmaktadır (Frakes ve Baeza-Yates, 1992). *tf.idf* yaklaşımına göre bir terimin değeri bir belgede ya da sorguda geçiş sıklığıyla doğru orantılı, derlemdeki farklı belgelerde geçiş sıklığıyla ise ters orantılıdır. İndekslenen kelimelerin kökleri bulunmuştur. Kök bulma amacıyla bilgi erişimde iyi sonuç verdiği gösterilen ilk beş karakter yöntemi kullanılmıştır (Can ve diğerleri, yayın aşamasında). İndeksleme sırasında yine aynı çalışmada verilen indekslenmeyecek kelimeler listesi (stopword list) kullanılmıştır.

MF8'e göre bir Q sorgusu ile bir d_j belgesi arasındaki benzerlik şu biçimde hesaplanmaktadır:

$$MF8 = \sum_{t \in Q} \left(\left(1 + \ln f_{dt}\right) / \sqrt{D} \right) \cdot \left(f_{qt} \cdot \ln\left(1 + N / f_t\right) \right) \quad (1)$$

Bu formülde

- f_{dt}: t teriminin d_j içindeki toplam geçiş sayısını;
- D: d_j içindeki toplam terim sayısını (yani d_j'nin içerdiği terimlerin f_{dt} değerlerinin toplamını);
- f_{qt}: t terimin Q içindeki geçiş sayısını;
- N: derlemdeki toplam belge sayısını;
- f_t: t terimin derlemdeki bütün belgelerde geçme sayısını

ifade etmektedir.

Aşağıdaki iki bölümde önce haber profillerinin ne olduğu, sonra bu haber profilleri yardımıyla ilk haberin ve onu izleyen haberlerin nasıl bulunduğu ve bu amaçla geliştirilen yazılım, ETracker, anlatılmaktadır.

Haber Profilleri

Deney derlemini oluşturmak için önce seçilen her bir haber için bir profil hazırlanmıştır. Bir haber profili şu maddelerden oluşmaktadır (örnek bir profil Şekil 4'de verilmiştir).

- **Başlık (Topic Title):** Olayı çağrıştıracak ve kolayca akılda kalan on kelimeden az bir cümle ya da kelimeler grubu;
- **Olay Tanımı (Event Summary):** Haber başlığını ayrıntılı hale getiren 1-2 cümle ile olayın tanımı;
- **Ne (What):** Olay sırasında ne olduğu;
- **Kim (Who):** Olayı gerçekleştiren veya olaydan etkilenen kişiler;
- **Ne zaman (When):** Olayın gerçekleştiği zaman;
- **Nerede (Where):** Olayın gerçekleştiği yer;
- **Sayı (Topic Size):** Tahmini haber sayısı;
- **Tohum (Seed):** Konu ile ilgili ilk haber (sistemdeki belge numarası);
- **Haber Türü (News Type):** Haberin türü ya da türleridir.

Haber türü, belirlenmiş 13 haber sınıfı (doğal afetler, kazalar, vb. gibi) (TDT, 2004) arasından seçilerek işaretlenir. Bu amaçla gerekiyorsa birden fazla haber türü seçilebilir.[1]

ETracker Sistemi

ETracker sistemi, birbirini izleyen ve aşağıda ayrıntıları verilen dört adımda ilk haber ve izleyen haberleri insan denetimi altında bulmaktadır. Değerlendirici ("annotator"), önce haber profilini ve tohum haberi (ilk hikâye-ilk haber, "first story") sistemin yardımıyla etkileşimli bir süreçten sonra belirler. Tohum sisteme girilen sorgular yardımıyla belirlenir. Her adımda değerlendirici, sistem tarafından bulunan haberlerin, tohumla ilgili olup olmadığını belirler. Sistemden gelen sonuçlar, ilk adımda 200 haber (belge), sonraki adımlardaysa sırasıyla 300 ve 400 haberle sınırlıdır ve o adımda kullanılan sorguya MF8 kullanılarak hesaplanan benzerliğine göre sıralanarak verilir.

İlgili haberin ilk hikayesini (tohumunu) bulmak için kullanıcılar ETracker sisteminin BE sistemi özelliklerini kullanarak izlemek istedikleri habere ilişkin sorgularını girerler ve zaman sırasına göre listelenen haberler arasından tohum niteliği taşıyan haberi saptarlar. Bu işlem birinci adımdan önce yapılır. Değerlendirme sırasında izlenecek olan adımların tanımı ve "kalite kontrol" işlemi şu biçimde tanımlanmıştır:

- *Adım-1 (tohum ile ara)*: Değerlendirici, tohumu sorgu olarak kullanarak ETracker sisteminde arama yapar. ETracker en çok benzeyen 200 haberi sıralayarak değerlendiriciye sunar. Değerlendirici bu dokümanların tohumla ilgili olup olmadığına karar verip "Evet" ya da "Hayır" etiketini verir.
- *Adım-2 (profil ile ara)*: Değerlendirici, haber profilinden oluşturulan sorgu için sistem tarafından bulunan sonuçları değerlendirir. (Bu ve daha sonraki adımlarda kullanıcının daha önceki adım(lar)da "Evet/Hayır" biçiminde işaretlediği haberlerin linkleri yeşil ve kırmızı renklerle gösterilir.)
- *Adım-3 (ilgili haberleri kullanarak ara)*: Değerlendirici, ilk iki adımda "Evet" diye belirlenen ve bu adımlarda en başta sıralanan üçer belgeyi ETracker'a birer sorgu olarak gönderir (ilk sıralarda gelen belgelerin çakışması durumunda daha alt sıralardaki belgeler kullanılır ve toplam altı belge seçilir). Bu adımda elde edilen sonuçlar bir bilgi kaynaştırma (data fusion) algoritması (reciprocal rank) yardımıyla tek bir sıra haline sokulur (Nuray ve Can, 2006) ve değerlendiricinin dikkatine getirilir. Bilgi kaynaştırma işlemi, daha çok sorgu tarafından daha önlerde getirilen haberleri sıralamada daha üst sıralara yerleştirir.
- *Adım-4 (yaratıcı sorgular ile ara)*: Değerlendiricilerin kendi sorgularını (yaratıcı sorgular) oluşturmaları istenmektedir. Bu adıma kadar, değerlendiriciler olay hakkında uzmanlaştıkları için geriye kalan ilgili belgelerin saptanmasında değişik kelimeler, özel isimler, belirli haberler v.s. kullanabilirler. Bu adımda, değerlendiriciler istedikleri sayıda sorgu ile çalışabilirler.

Şekil 4. Örnek profil: "Sahte rakı"

[1] "Haber türü" maddesi ileride yapılması söz konusu olabilecek sınıflandırma araştırmasına gerekli bilgiyi sağlamayı amaçlamaktadır.

Kalite kontrol amacıyla değerlendiricinin profil üzerinde yaptığı incelemeyi denetlemek için sistemden "ilgili" olarak işaretlediği tüm belgeler alınır, sonra bu belgelerin ilgili olup olmadığına karar verilir. Denetleyen kişi yanlış ikaz (false alarm) durumlarını tespit edip, "Evet" diye yanlışlıkla etiketlenen belgeleri "Hayır" diye düzeltir. Bunun yanı sıra ilgisiz haberler arasından rastgele on haber seçilir, bu haberlerin arasından en fazla bir tane ilgili haber çıkması beklenir. Birden fazla ilgili haber çıkarsa, incelenen profil için değerlendirmenin tekrarlanması istenir. Böylece, profillerle ilgili haberlerin gözden kaçırılması önlenmeye çalışılır.

Deney Derlemindeki Profiller

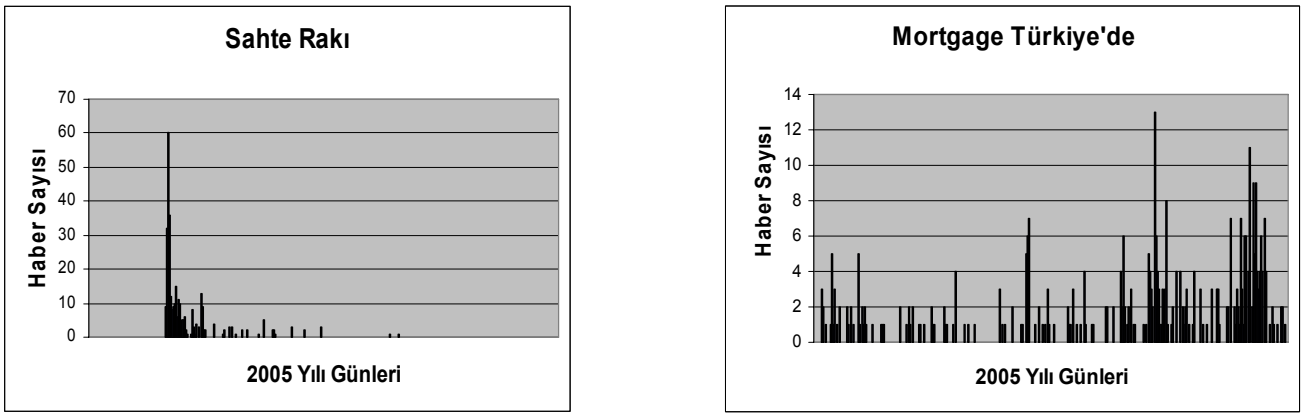
ETracker sistemi kullanılarak üretilmiş haber profillerinden "sahte rakı" Şekil 4'de, "sahte rakı" ve "mortgage Türkiye'de" profilleriyle ilgili haberlerin yıl içindeki dağılımı Şekil 5'de, bazı örnek haber profillerine ait özellikler de Tablo 2'de verilmiştir. Halen derlemdeki haber profilleri yeni yapılan eklemelerle genişletilmektedir.

Şu ana kadar sistem 39 kullanıcı tarafından kullanılmıştır. Kullanıcı başına ortalama 2,4 profil düşmektedir. Kullanıcıların çok olması derlemin sağlıklı olması açısından da önemlidir. Her bir profil için ortalama 114 dakika harcanmıştır. En fazla profil 17 profille "şiddet ya da savaş haberleri" ile ilgilidir. Bu türü "Ünlüler/İnsanlarla ilgili haberler," "Kazalarla ilgili haberler" izlemektedir. Şu anda sistemde değerlendirilmiş ve kontrol edilmiş 90 profil yer almaktadır.

Sonuç ve Araştırma Olasılıkları

Geliştirmekte olduğumuz deney derlemi büyüklüğü ve kapsamıyla Türkçe için yapılacak olan YOBİ çalışmaları için ilk standart olması amacıyla hazırlanmaktadır. Derlem, başka araştırmacıların YOBİ çalışmalarını kolaylaştıracak ve yeni araştırmaları teşvik edecek niteliktedir. TDT'den (2004) esinlenerek geliştirilen derlem oluşturma yaklaşımı verimli ve etkin bir yöntem olması nedeniyle yeni YOBİ deney derlemlerinin hazırlanmasında kullanılabilir.

YOBİ araştırmalarının bir özelliği de kazanılacak olan deneyimin yeniliklerin saptanmasının (novelty detection) önemli olduğu başka uygulama alanlarına da aktarılabilir olmasıdır. Bunlar arasında istihbarat bilgilerinde yeni gelişmelerin saptanması, ticari veri madenciliğinde yeni alış veriş alışkanlıklarının belirlenmesi ve e-posta ortamlarında yeni başlatılan yazışma konularının saptanarak izlenmesi sayılabilir. Bu esneklik YOBİ alanını araştırmacılar için daha çekici bir hale getirmektedir.



Şekil 5. Örnek iki profil için haberlerin günlük dağılımı.

Tablo 2 . ETracker sistemiyle üretilmiş haber profillerinden örnekler

Haber Profillerinin Bazıları	İzleyen Haber Sayısı	Haber Ömrü (Gün)	İlk *n* Günde İzleyen Haber Sayısı			
			***n*=100**	***n*=50**	***n*=25**	***n*=10**
Londra metrosunda patlama	454	175	440	419	376	236
Sahte rakı	323	182	316	291	255	197
400 koyun intihar etti	10	8	10	10	10	10
Mortgage Türkiye'de	375	356	60	41	25	13
Onur Air'in Avrupa'da yasaklanması	159	203	154	154	148	105
İlk yüz nakli	14	17	14	14	14	10
Attila İlhan vefat etti	40	69	40	37	36	32
Şu anda Toplam Profil Sayısı : 90 Ortalamalar:	75	97	64	54	47	36

Teşekkür

Çalışmadaki profilleri yazan ve değerlendiren meslektaş ve öğrencilerimize teşekkür ederiz. Bu çalışma, 106E014 numaralı projeyle TÜBİTAK tarafından kısmen desteklenmiştir. Çalışmada verilen öneriler ve sonuçlar yazarlara ait olup destekleyen kuruluşla bir ilgisi yoktur.

Kaynakça

Allan, J., Lavrenko, V. ve Jin, H. (2000). First story detection in TDT is hard. *Proceedings of the 9th Conference of Information and Knowledge Management (ACM CIKM'00)* içinde (s. 374-381). McLean, VA: ACM. 6 Temmuz 2007 tarihinde http://delivery.acm.org/10.1145/360000/354843/p374-allan.pdf?key1=354843&key2=7708273811&coll=GUIDE&dl=GUIDE&CFID=23203773&CFTOKEN=18908848 adresinden erişildi.

Allan, J., Papka, R. ve Lavrenko, V. (1998). On-line new event detection and tracking. W.B. Croft ve diğerleri (Eds.), *Proceedings of the 21st International Conference on Research and Development in Information Retrieval (ACM SIGIR '98)* içinde (s. 37-45). Melbourne: ACM.

Can, F. (1993). Incremental clustering for dynamic information processing. *ACM Transactions on Information Systems, 10*, 143-164.

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*, 483-517.

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

Can, F., Koçberber, S., Balçık, E., Kaynak, C., Öcalan, H.C., ve Vursavaş, O.M. (yayın aşamasında). Information retrieval on Turkish texts. *Journal of the American Society for Information Science and Technology.*

Frakes, W.B. ve Baeza-Yates, R. (1992). *Information retrieval: Data structures and algorithms*. Englewood Cliffs, NJ: Prentice-Hall.

Google News. (2007). 28 Ocak, 2007 tarihinde http://news.google.com/ adresinden erişildi.

HTML parser. (2007). 13 Ocak 2007 tarihinde http://htmlparser.sourceforge.net/ adresinden erişildi.

Hatzivassiloglou, V., Gravano, L. ve Maganti, A. (2000). An investigation of linguistic features and clustering algorithms for topical document clustering. *Proceedings of the 23rd International Conference on Research and Development in Information Retrieval (ACM SIGIR '00)* içinde (s. 224-231). Athens: ACM.

Kobayashi M. ve Takeda, K. (2000). Information retrieval on the Web. *ACM Computing Surveys, 32*, 144-173.

Kumaran, G. ve Allan, J. (2004). Text classification and named entities for new event detection. *Proceedings of the 27th International Conference on Research and Development in Information Retrieval (ACM SIGIR '04)* içinde (s. 297-304). Sheffield: ACM.

Kumaran, G., Allan, J. ve McCallum, A. (2007). *Classification models for new event detection.* (CIIR Technical Report). 7 Ocak 2007 tarihinde http://ciir.cs.umass.edu/pubfiles/ir-362.pdf adresinden erişildi.

Kurt, H. (2001). *On-line new event detection and tracking in a multi-resource environment.* Yayımlanmamış Yüksek Lisans Tezi, Bilkent Üniversitesi, Ankara. 12 Şubat 2007 tarihinde http://citeseer.ist.psu.edu/kurt01line.html> adresinden erişildi.

NewsIsFree. (2007). 12 Şubat 2007 tarihinde http://newsisfree.com adresinden erişildi.

Nuray, R. ve Can, F. (2006). Automatic ranking of information retrieval systems using data fusion. *Information Processing & Management, 42*, 595-614.

Papka, R. (1999). *On-line new event detection, clustering, and tracking.* Yayımlanmamış Doktora Tezi. University of Massachusetts, Amherst.

Radev, D., Otterbacher, J., Winkel, A. ve,Balir-Goldensohn, S. (2005). News InEssence: Summarizing online news topics. *Communications of the ACM*, *48*(10), 95-98.

TDT 2004: Annotation manual: Version 1.2 - August 4, 2004. (2004). 9 Ocak 2007 tarihinde http://projects.ldc.upenn.edu/TDT5/Annotation/TDT2004V1.2.pdf adresinden erişildi.

Topical Detection and Tracking (TDT). (2007). 12 Şubat 2007 tarihinde http://www.nist.gov/speech/tests/tdt/ adresinden erişildi.

Van Rijsbergen, C.J. (1979). *Information retrieval* (2nd ed.). London: Butterworths. 12 Şubat 2007 tarihinde http://www.dcs.gla.ac.uk/Keith/Preface.html> adresinden erişildi.

Vural, A., (2002). On-line new event detection and clustering using the concepts of the cover coefficient-based clustering methodology. Yayımlanmamış Yüksek Lisans Tezi, Bilkent Üniversitesi, Ankara. 12 Şubat 2007 tarihinde www.cs.bilkent.edu.tr/tech-reports/2002/BU-CE-0218.pdf adresinden erişildi.

Voorhees, E. (2005). TREC: Improving information access through evaluation. *Bulletin of the American Society for Information Science and Technology*, *32*(1). 6 Temmuz 2007 tarihinde http://www.asis.org/Bulletin/Oct-05/voorhees.html adresinden erişildi.

Witten, I.H. ve Frank, E. (2000). *Data mining: Practical machine learning tools and techniques with Java implementations*. San Francisco: Morgan Kaufmann.

Yang, Y., Carbonell, J., Brown, R., Pierce, T., Archibald, B. ve Liu, X. (1999). Learning approaches for detecting and tracking news events. *IEEE Intelligent Systems, 14*, 32-43.

Yang, Y., Pierce, T. ve Carbonell, J. (1998). A study on retrospective and on-line event detection. W.B. Croft ve diğerleri (Eds.), *Proceedings of the 21st International Conference on Research and Development in Information Retrieval (ACM SIGIR '98)* içinde (s. 28-36). Melbourne: ACM.