Haber Portallarında Yenilikçi Yaklaşımlar

Fazlı Can[1], Seyit Koçberber[2], Özgür Bağlıoğlu[1], Gönenç Ercan[1,3], Süleyman Kardaş[1], H. Çağdaş Öcalan[1], Erkan Uyar[1], Levent Koç[1]

Bilkent Üniversitesi, Bilgi Erişim Grubu

[1] Bilkent Üniversitesi, Bilgisayar Mühendisliği Bölümü

[2] Bilkent Üniversitesi, Bilgisayar ve Öğretim Teknolojileri Öğretmeliği Bölümü

[3] Bilkent Doğal Dil İşleme Grubu

{canf, ozgurb, ercangu, skardas, hocalan, euyar}@cs.bilkent.edu.tr, seyit@bilkent.edu.tr, lkoc@ug.bilkent.edu.tr

Özet: Çok kaynaktan beslenen bağımsız haber portallarının sağladığı faydalar arasında haber zenginliği, olayları farklı açılardan görme olanağı ve haber sunumundaki tarafsızlık sayılabilir. Web ortamındaki haber kaynaklarının sayısında ve bu kaynaklar tarafından yayımlanan haberlerde aşırı artış gözlenmektedir. Bu durumu kullanıcılara hissettirmeden, gelişmekte olan olaylarla ilgili veya kullanıcı için önemli haberlerin kolayca izlenmesine olanak sağlayacak haber portallarının yapımında yenilikçi yaklaşımların kullanılması gerekmektedir. Bu yaklaşımlar kullanıcılar için geniş bir hizmet yelpazesi sağlayabilecektir. Bu makalede, haber portalı geliştirilmesinde kullanılabilecek olan yenilikçi yaklaşımlar tanımlanıp, bu yaklaşımların araştırma grubumuz tarafından yapımı sürmekte olan büyük ölçekli Bilkent Haber Portalındaki kullanımı anlatılmaktadır.

Anahtar Kelimeler: Bilgi Erişimi, Bilgi Süzme, Eşlenik Bulma, İçerik Çıkartma, Web, Yeni Olay Bulma ve İzleme.

Abstract: Innovative Approaches in News Portals: Users of a news portal that feeds from information served by different news sources are able to view events from different perspectives, and as a result, in a more objective fashion. The number of news publishers and news articles on the Web has been increasing rapidly. This provides a rich environment for users, if handled with care while employing new approaches. News portals can fulfill user information needs by equipping them with tools to track important events and browse events conveniently. Only with such tools, the underlying rich information environment can be transformed into a vast set of concrete information services. In this paper, we introduce some novel approaches that may be used in news portals, which make the readability and traceability of news simpler. These approaches are used in the Bilkent News Portal, whose large-scale implementation is in progress.

Keywords: Information Retrieval, Information Filtering, Near-duplicate Detection, Content Extraction, Web, New Event Detection and Tracking.

1. Giriş

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 erişimi, BE, (information retrieval) konusunda çok sayıda araştırma yapılmıştır. Yeni yayınlanan belgelerin ilgi alanlarına göre kullanıcılara dağıtılması için bilgi süzme, BS, (information filtering) konusu da benzer biçimde araştırmaların yoğun olduğu bir alandır [8].

Internet'in haberlerin yayımında kullanılmasıyla, bilgi ve belge konusunda yaşanan patlamaya benzer bir haber patlaması yaşanmaya başlanmıştır; örneğin NewslsFree Web servisi

43 farklı dilde 12.000'den fazla Web haber kaynağı listelemektedir [10]. Yapımında yenilikçi öğeler içeren Google Haberler yaklaşık 25.000 haber kaynağından beslenmektedir [7]. Günümüzde haber üreten ve yayımlayan 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.

2. Bilkent Haber Portalının Özellikleri

Bilkent Haber Portalı geliştirilmekte olan araştırma konularının uygulamaya çevrilerek genel amaçlı kullanıma sunulan kapsamlı bir haber alma kaynağıdır [1]. Bünyesinde bilgi erişim, yeni olay belirleme ve izleme, bilgi kümeleme ve bilgi süzme gibi birçok sistemi barındırmaktadır. Bu niteliklerinden ötürü varolan sistemlerden ayrılıp, haberleri işleme ve sunma konusuna yeni bir boyut kazandırmaktadır. Kullanıcılar açısından ise kişiselleştirilebilir olması kullanıcıların haber akışı içinde boğulmadan kendi ilgi alanlarındaki haberlere kolaylıkla ulaşabilmelerini sağlamaktadır. Sistemin kişiselleştirilebilir olma özelliği zaman içinde sistemi kullanıcı tercihleri yönünde değiştirmemizi sağlayacak çeşitli istatistiksel bilgilerin toplanmasında kolaylık sağlamaktadır [3].

Bilkent Haber Portalı bağlamında yapımı bitirilmiş veya sürmekte olan önemli uygulama birimleri şu başlıklar altında toplanabilir: 1) Haber kaynağından bağımsız olarak haber sayfalarından içerik (haber resmi ve haber metni) çıkartımı, 2) metin içeriği yaklaşık birbirinin aynı olan haberlerin saptanması, 3) haber portalının ana sayfasında sunulacak haberlerin seçimi, 4) yeni olayların bulunması ve izlenmesi, 5) izlenen olaylardaki yeni gelişmelerin saptanması, 6) mevcut haberlerde arama servisi, 7) bir olayla ilgili veya arama sırasında ulaşılmış haber grupları için özet çıkartımı, 8) kullanıcının ilgilendiği konulardaki yeni haberlerin saptanması ve 9) kullanıcı ihtiyaçlarına göre haber saklama ve kullanıcılar arası haber paylaşımı. Yukarda kısaca tanımlanmış olan uygulamalardan çoğu haber portalı geliştirilmesinde yenilikçi yaklaşımlardır. Bu uygulamaların etkin ve randımanlı çalışmalarını sağlamak amacıyla gerekli durumlarda Türkçe için deney derlemleri oluşturulmuş ve uygulamalar deney gözlemlerine göre geliştirilmiştir [2]. Başka bir deyişle, üzerinde çalışılan yaklaşımlar ilk olarak deneysel olarak çalışılmış, elde edilen en iyi sonuçlar ise pratik olarak portala yansıtılmıştır.

Bir sonraki bölümde geliştirmekte olduğumuz Bilkent Haber Portalında kullanılan yenilikçi yaklaşımlar detaylı olarak anlatılacaktır. Son bölümde ise yapılan çalışmaların sonuçları ve ilerisi için yapılabilecek öneriler anlatılmıştır.

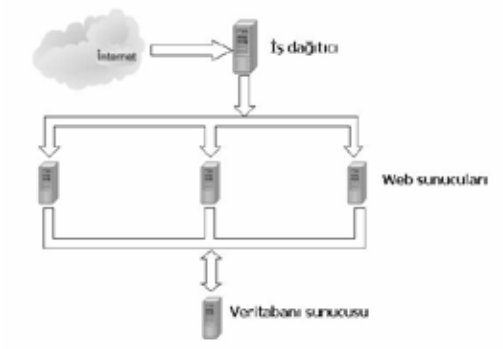
3. Yenilikçi Yaklaşımlar

Bilkent haber portalı tasarlanırken sistemin farklı kullanıcı kitlelerine hitap edeceği düşünülmüştür. Bu bağlamda minimalist bir yapı izlenmiş ve günümüzde de kullanıcı ile etkileşimde olan birçok Web sitesinin en temel tasarım kıstasları olan kullanışlılık ve kolay anlaşılabilirlik ön plana çıkarılmıştır. Sistemde şu anda bütün arayüzler hem İngilizce hem de Türkçe olarak sunulmaktadır. Bu arayüz bir şablon yapısında hazırlanmış olup, dili kolaylıkla değiştirmek mümkündür. Haber portalının arayüzü Şekil 1'de gösterilmiştir. Bilkent haber portalı diğer bir çok haber portalının sağladığı haberlere kolay erişim, olay odaklı haber gruplama vb. gibi genel imkanları bünyesinde barındırdığı gibi, yenilikçi servisler de sunmaktadır. Bu servislerin genel amacı kullanıcının haberleri daha rahat takip etmesini sağlamak, ve incelenen konuların birçok kaynaktan gelen haberlerle geniş bir perspektiften değerlendirilmesine olanak sunmaktır. Kullanıcı bir olayın ilk ortaya çıkışından itibaren ne şekilde geliştiğini kronolojik

bir şekilde gözlemleyebilmekte, belirlediği ilgi alanlarında yazılan haberleri güncel bir şekilde takip edebilmektedir. Kullanıcıya sunulan bütün bu gerçek zamanlı ve deneysel olarak kullanılan yenilikçi servisler bu bölümde anlatılacaktır.

3.1 Portalın Mimari Yapısı

Şekil 2'de mimari yapısı anlatılan sistemde sistem yükünün azaltılması için çoklu Web sunucuları kullanımı hedeflenmiştir. Bu yaklaşımın en önemli avantajı gerekli performansı daha ucuza sağlamasının yanısıra yazılım güncellemesi veya olası bir donanım arızasında verilen hizmetin kesilmeden sürdürülebilmesidir. Bir diğer avantajı ise Web kullanıcılarından gelecek talebin artması üzerine ilave Web sunucularının devreye alınması hem daha kolay hem de daha ucuz olacaktır. Portalda bulunan sunucu kümesinde veri depolama ve veritabanı hizmetleri için bağımsız bir sunucu önerilmiştir. Veritabanı sunucusu haberlerin taranması, depolanan haberlere erişim ve yeni haberlerin veritabanına yazılması işlemlerinden başka bir işle meşgul olmayacak, yalnız veritabanı hizmeti verecektir. Bu nedenle başlangıçta 3 adet olan, ancak sayıları ihtiyaca göre artabilecek olan Web sunucuları talep edilen işlemleri paralel olarak yapabilecek şekilde tasarlanmıştır.

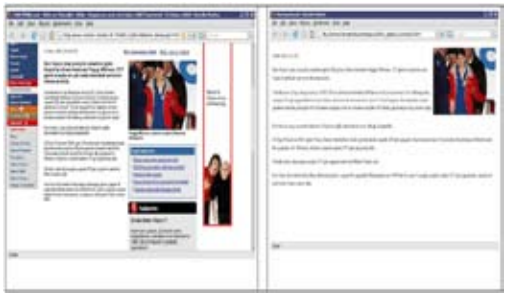


Şekil 2. Portalın Mimarisi.

3.2 Gerçek Zamanda Haber İndirimi ve İçerik Çıkartma

Sisteme haber sağlamak amacıyla geliştirilen haber toplama ve ayıklama alt sistemi çeşitli haber kaynaklarındaki haberleri sistemin kullanabileceği TREC (http://trec.nist.gov/) formatına çevirmekle yükümlüdür. Yapılan işlem şu şekilde özetlenebilir. Belli periyotlarla RSS'leri okunan haber kaynaklarından elde edilen dokümanlar, HTML formatında kaydedilir. Daha sonra kaynaklara özgü yazılan HTML etiketlerini ayıklama yazılımlarıyla sistemin kullanımına hazır TREC formatında dokümanlar elde edilir. Bu işlem gün içinde belli periyotlarla otomatik olarak yapılarak, sistemin güncel kalması sağlanır.

Kullanmakta olduğumuz yaklaşım kısaca üç aşamada çalışmaktadır. İlk önce resim (img) etiketlerini (tag) incelenerek, haberle ilgili olabilecek resmin adresi bulunmaktadır, daha sonra, kaynak kodundaki ilgisiz yerler ve etiketler ayrıştırılmakta ve son olarak, kalan kısımdaki metin yoğunluğuna bakılarak, bir takım sınır değerlere göre haber olma ihtimali yüksek olan bölümler saptanmaktadır. Bu yaklaşımla bir haberin içeriği çıkartılmaktadır. Şekil 3'te yapılan işlemin pratik sonuçları gösterilmektedir. Bu şekilde soldaki resim içerik çıkartma (content extraction) işleminin girdisini, sağdaki resim ise çıktısını göstermektedir.



Şekil 3. Bilkent Haber Portalında içerik çıkartma işlemi örneği (sol: girdi, sağ: çıktı).

3.3 Eşlenik Haberlerin Bulunması ve Ayıklanması

Günümüzde internet ortamındaki haber kaynakları, haberleri ajanslardan almakta ve genellikle bu haberlerin çok az bir kısmını değiştirerek veya hiç değiştirmeden kullanıcılarına

sunmaktadırlar. Bu durum internet ortamında birbirine benzeyen çok sayıda haberin bulunmasına neden olmaktadır. Ayrıca kullanıcının bir sorgu sonucunda çok sayıda eşlenik haberle karşılaşmasına yol açmakta ve benzer şekilde arama motorlarının performansını da düşürmektedir [6]. Bu durum dikkate alınarak, Bilkent Haber Portalı'nda eşlenik haberleri bulup (near-duplicate detection) ayıklayan bir yöntem geliştirilmiştir. Kısaca eşlenik ayıklama süreci şu şekilde çalışmaktadır. Sistemde bulunan her haberin içerdiği tanımlayıcı kelime grupları (adlandırılmış nesneler, vb.) dikkate alınarak bir imzası (signature) çıkarılır ve bu haber-imza ilişkisi dikkate alınarak sorgu sonuçlarında eşlenik haberler ayıklanarak kullanıcıya sunulur.

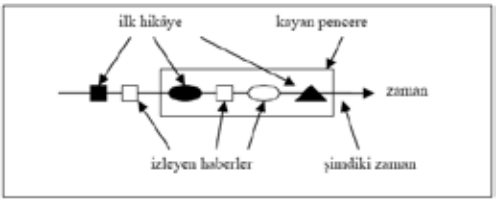
3.4 Portalın Ana Sayfasındaki Haberlerin Seçimi

Genel olarak haber sitelerinin ana sayfalarında gösterilen haberler ve içerikleri editörler tarafından elle saptanmaktadır. Ana sayfadaki haberler bazen sansasyonel bazen de toplumu yönlendirme amaçlı olabilmektedir. Haberlerin bu şekilde belli kişiler tarafından seçilip getirilmesi, tarafsızlığı azaltabilmektedir.

Bilkent Haber Portalında ise ana sayfada getirilen haberler otomatik olarak belirlenmektedir. Genel olarak gündemden uzak olan haberleri eleyip, gündeme en yakın olan ve içerik olarak da en zengin haberlerin seçilebilmesini sağlayacak bir yol izlenmektedir. Bu belirlemede kullanılan yöntem kapsama katsayısı kavramından esinlenerek geliştirilmiştir [5].

3.5 Yeni Olay Belirleme ve İzleme: YOBİ

BE ve BS sistemlerine göre daha yeni bir uygulama olan yeni olay belirleme ve izleme, YOBİ, (topic detection and tracking) sistemleri hem BE hem de BS sistemlerine benzemekle birlikte her ikisinden de amaçları ve geliştirme yöntemleri açılarından farklıdır. Bir YOBİ ortamında sisteme sürekli yeni haber belgesi gelmekte, bu belgelerin içinde yeni olaylara karşılık gelen belgeler saptanmakta ve bu yeni olayların devamı olan haberler bulunarak kullanıcılara iletilmektedir. YOBİ ortamında yeni olayların saptanması ve bu olayların devamının birbirine bağlanarak bir zincir oluşturulması da otomatik olarak yapılmaktadır.



Şekil 4. Yeni haber belirleme ve izleme (YOBİ) (pencere akan zamanla birlikte yeni haberlere doğru sağa kaymaktadır, en yeni haber en sağdadır).

Haberin değeri ve önemi zamanla değişmektedir. Basın yayın kurumları bir haberi en 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 (Şekil 4).

Konu izleme, bir kaç ilk haberle tanımlanmış konularla ilgili olan yeni haberleri saptamayı amaçlayan bir işlemdir. Belirtildiği gibi konu tanımı bir veya daha fazla sayıda haberden oluşur.

Arayüzün bu kısmı, Şekil 5'te verilmiştir, yeni olay belirleme sisteminin saptadığı olayları yeni olaylar ve bütün olaylar başlıkları altında bir "tab" menüde vermektedir. "Yeni Olaylar" ("Recent Events") başlığı altında verilen olaylar en son elde edilen haberler içerisinden belirlenip sunulurken, sistemin aktif olarak çalıştığı süre içerisinde belirlenmiş olayların hepsi "Bütün Olaylar" ("Recent and Past Events") başlığı altında verilmektedir. Kullanıcı bu olayları izleyen haberlere de bu kısımdan ulaşabilmektedir.

Şekil 5. Bilkent Haber Portalı'nda haber izleme sonuçlarının verildiği arayüz.

3.6 İzlenen Haberlerde Yeni Gelişmelerin Saptanması

Bir haber portalında BS veya YOBİ işlemleri sonucunda bulunan haberlerin hepsinin kullanıcı tarafından enteresan bulunması beklenemez. Kullanıcılar belli bir konuyla ilgili olan bütün haberleri değil, söz konusu konudaki gelişimi yansıtan yeni bilgileri içeren haberleri görmeyi tercih edebilirler. Portalda geliştirilmekte olan yeni yöntemlerle bir konuyla ilgili olan haberler içerdikleri yenilik açısından değerlendirilerek yenilik içeren haberler saptanacaktır. Geliştirilecek yöntemler deneysel olarak ölçüldükten sonra, başarılı olanlar portal arayüzüne eklenecek ve kullanıcıların yalnızca yeni bilgi içeren haberleri görmeleri sağlanacaktır.

Bu kapsamda portalda, bir konuyu takip eden haberler içinde yenilik içeren haberlerin saptanması ve bu haberlerin bir zaman çizgisi üzerinde kullanıcıya sunumu sağlanacaktır. Bu amaçla algoritmalar geliştirilerek etkinlik ve randımanları Türkçe haber ortamında hazırlanan deney derlemi kullanılarak ölçülecektir. Buradaki amaç bir konu için izlenerek bulunmuş haberlerin arasında yeni gelişmelerin saptanmasıdır. Daha önceden proje çalışanlarının katkısıyla geliştirilen kapsama katsayısı (cover coefficient) kavramı bu amaçla kullanılacaktır [5]. Bu aşamada kapsama katsayısı kavramını kullanan iki yaklaşım ve cümle düzeyinde analizini de esas alan üçüncü bir yaklaşım YB işlemlerinde kullanılacaktır.

Yeni gelişmelerin saptanması konusunda bu kısımda anlatılanlar, hala üzerinde çalışılmakta olup deneylerde belirlenen en iyi sistemin Bilkent Haber Portalına taşınması düşünülmektedir.

3.7 Arama Servisi

Bilkent Haber Portalı'nda haber arşivi üzerinde arama yapabilmeyi sağlayan bilgi erişimi sistemi açık kodlu Lemur [9] paketi kullanılarak geliştirilmiştir. Bu amaçla kullanılmakta olan arayüz Şekil 6'da verilmiştir. Gün içinde belli periyotlarla edinilen yeni haberler arttırmalı olarak indekslenerek bilgi erişim sisteminin güncelliği sağlanmaktadır. Arama sonuçları tarihe göre ya da dokümanların sorguya olan ilişkisine göre sıralanabilmekte ve ayrıca getirilen sonuçlar isteğe bağlı olarak çok yakın benzerlik gösteren eş dokümanlardan ayıklanabilmektedir. Kullanılan indeksleme algoritması bahsi geçen özelliklerin yüksek hızda gerçekleşmesini sağlamaktadır. Ağır yük altında ve geniş belge koleksiyonlarında tutarlı bir performans sergilemesinin de temel nedenidir.

Çalışmakta olan bilgi erişim sistemi konuyla ilgili yapmış olduğumuz deneylerin sonuçlarından hareketle "kök" bulma algoritması olarak kelimenin ilk 5 karakterini kullanılmaktadır [4].

Şekil 6. Bilkent Haber Portalı bilgi erişim sonuçlarının verildiği arayüz.

3.8 Haber Grupları için Özet Çıkarımı

İnternetin gelişmesiyle artık bilgiye erişim çok kolay hale gelmiştir. Ve aynı konu hakkında birbirine benzeyen bir çok bilgiye ulaşmak kolaylaşmıştır. Bu bağlamda Bilkent Haber Portalında kullanıcılara belli bir konuya ait haber grupları için özet çıkarabilme olanağı da sağlanacaktır. Bu şekilde kullanıcılar, bir konu ile ilgili olarak kısa zamanda genel bilgi sahibi olabileceklerdir. Özet çıkartımı henüz deneysel aşamadadır. Portala deneylerdeki başarımı gözlendikten sonra eklenmesi düşünülmektedir.

3.9 İlgiye Yönelik Haber İzleme

Bilgi süzme (information filtering) işlemi kullanıcının kendi belirlediği konulardaki haberleri takip edebilmesi için geliştirilmiştir. Bu işlem, literatürde çok çeşitli yöntemlerle yapılmış ve uygulanmıştır. Bu yöntemlerdeki ortak özellikler kullanıcı ilgisinin belirlenmesi, güncel dokümanların kullanıcının ilgisine sunulması ve bunu takiben kullanıcı profillerinin güncellemesidir.

Kullanıcı kendi ilgi alanlarını belirledikten sonra, okuduğu haberleri, belirlediği ilgi alanlarına ekleyerek, ilgi alanlarını güncelleyebilir. Ayrıca, bu yönteme alternatif olarak kullanıcının ilgi alanına ait kelimeleri de girebilmesine olanak sağlanmıştır. Bu şekilde takip edilmek istenilen konular sisteme kaydedilip ilgili haberler kolaylıkla izlenebilir. Böylece bilgi süzme veya YOBİ ile ilgilenilen haberlere erişilir. Sistem ilgi alanlarına eklenen dokümanlardan kilit kelimeler seçerek veya kullanıcının girdiği kelimelerle yeni gelen dokümanların ilgi alanlarıyla ilişkili olup olmadığına karar verir.

3.10 Kullanıcıya Yönelik Diğer Özellikler

Daha önce anlatılan özelliklerin yanında Bilkent Haber portalı kişisel olarak da bazı avantajlar sunmaktadır. Bu bağlamda kullanıcılar portala girdiklerinde ilgilerine göre haberleri takip edebildikleri gibi, başka kullanıcılara haber önerebilmekte, önerdiği haberlere yorum ekleyebilmektedir. Ayrıca beğendiği haberleri saklayabilmektedir. Buradaki amaç portalın kullanıcıya yönelik olarak kişiselleştirilmesidir.

4. Sonuç

Gerçekleştirilen haber portalı taşıdığı özelliklerle araştırma ve ticari uygulamalarda örnek olabilecek niteliktedir. Geliştirilen prototip haber portalı çeşitli özgün yaklaşımları bir arada sunmaktadır. Bu haber portalı ayrıca bizim için bir laboratuvar ortamı sağlamıştır. Portal sayesinde çeşitli yenilikçi yaklaşımlar denenebilmektedir. Bu şekilde pratikte olumlu sonuçlarını gördüğümüz yaklaşımları araştırma ile deneyerek, mühendislikten bilime geri dönen bir geri aktarım halkasını tamamlayacağımızı düşünmekteyiz. Haber portalı aynı zamanda bu türden uygulamalarda çıkabilecek problemleri pratik ortamda tanımlamamıza ve böylece yeni araştırma konuları saptamamıza yardımcı olmaktadır.

Ayrıca bu kapsamda yapılması düşünülen nihai çalışmanın amacı kişiye özgü Web-gazetesi hazırlayabilmektir. Bu sayede kullanıcılar haber okurken ilgi alanlarına ağırlık verebilecek, haberlerin bütününü belli bir düzen içerisinde görerek olayları geniş bir bakış açısından değerlendirebilecek, benzer olayları görerek karşılaştırabilecektir. Bu şekilde sağlanacak yapının ise insanların en etkin şekilde, tarafsız ve etraflıca haber almasını sağlayarak gündemi takip etmesini kolaylaştıracağı öngörülmektedir.

Teşekkür

Bu çalışma TÜBİTAK tarafından 106E14 numaralı proje ile desteklenmiştir ve TÜBİTAK tarafından 108E074 numaralı proje ile desteklenmektedir.

Kaynaklar

[1] Bilkent News Portal. http://news-portal.bilkent.edu.tr/PortalTest son ulaşıldığı tarih: 17 Mart, 2009. [2] Can, F., Koçberber S., Bağlıoğlu, O., Kardaş, S., Öcalan, H. C., Uyar,

E. Türkçe haberlerde yeni olay bulma ve izleme: Bir deney derleminin oluşturulması. Değişen Dünyada Bilgi Yönetimi Sempozyumu, s. 5059, 2007.

[3] Can F., Koçberber S., Bağlıoğlu O., Kardaş S., Öcalan H. C., Uyar E. Bilkent news portal: a personalizable system with new event detection and tracking capabilities. ACM SIGIR Konferansı, s. 885, 2008.

[4] Can F., Koçberber S., Balçık E., Kaynak C., Öcalan H. C., Vursavaş O. M., Information retrieval on Turkish texts. Journal of the American Society for Information Science and Technology, 59(3): 407-421, 2008.

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

[6] Chowdury A., Frieder O., Grossman D., McCabe M. C. Collection statistics for fast duplicate document detection. ACM Trans. on Information Systems, 20(2):171-191, 2002.

[7] Google Haberler. http://news.google.com.tr/ son ulaşıldığı tarih: 11 Mart, 2009.

[8] Kobayashi M., Takeda K. Information retrieval on the Web. ACM Computing Surveys, 33(2): 263-311, 2000.

[9] Lemur. http://www.lemurproject.org son ulaşıldığı tarih: 17 Temmuz, 2008.

[10]NewsIsFree. http://newsisfree.com son ulaşıldığı tarih: 11 Mart, 2009.