Metin Madenciliği ile Benzer Haber Tespiti

Anıl KARADAĞ*, Hidayet TAKÇI*

* Gebze Yüksek Teknoloji Enstitüsü, Bilgisayar Mühendisliği Bölümü, Kocaeli

anil.karadag@gmail.com, htakci@gmail.com

Özet: Bu çalışmada - içeriğini farklı haber sitelerinin farklı kategorilerinde yer alan haberleri tarayarak elde eden - dinamik içerikli bir haber sitesi için benzer haberleri sınıflandıran bir sistem geliştirilmiştir. Çalışmanın amacı, arama işleminin işlevselliğini arttırmak ve birbiriyle ilişkili haberleri tespit ederek haber arayan okuyucuya daha yararlı bilgiler sunmaktır.

Sistemin çalışma prensibi şu şekildedir, önce her habere anahtar kelimelerden (veya etiket) oluşan bir etiket listesi atanır, sonrasında haberlerin etiket listeleri karşılaştırılarak haberler arasındaki benzerlikler değerlendirilir. Etiketler kelime köklerinden meydana getirilmiştir. Kök tespiti için Türkçe dil işleme kütüphanesi Zemberek'ten yararlanılmıştır. Haber benzerliklerinin bulunmasında özellikle "son dakika" kategorisindeki haberler kullanılmıştır.

Anahtar Sözcükler: Metin madenciliği, metin benzerliği, benzer haber tespiti, vektör uzay modeli, kosinüs benzerliği.

Detection Similar News with Text Mining

Abstract: In this study - a system which classificates similar news for a dynamic contents news site, scannes news in different categories and different sources is developed. This study's goal is to increase the functionality of searching and to show more beneficial informations which are interrelated between each other for readers who search news in web.

Principles of the running system are to assign a label list that contains keywords for all news which exist in the system after that to evaluate relation of the news which are interrelated between each other with comparing of every label lists with all the other news's label lists. Label lists are occured by word roots. Zemberek Turkish language processing library is helped for detection word roots. The news category is "last-minute" are used to find similarity of news.

Keywords: Text Mining, text similarity, similar news detection, vector space model, cosine similarity

1. Giriş

Internetin yaygınlaşması ile birlikte yapısal olmayan verilerin (özellikle de metinsel veriler) miktarı çok fazla artmıştır. Metinsel verilerin olmadığı hiçbir veri analizi yeterli sonucu vermeyecektir. Merrill Lynch'e göre kullanılabilen iş bilgilerinin %85'inden fazlası yapısal olmayan verilerden çıkarılmaktadır [1]. Bununla birlikte, metinsel verilerle çalışmak, onları anlamlı hale getirmek kolay bir iş değildir.

Elektronik postalar, kişisel web sayfaları, anket sonuçları, internet sayfalarında yer alan forum bilgileri gibi yapısal olmayan verilerin analizi için veri madenciliğinin metinler üzerinde çalışan, daha farklı özelliklere sahip bir uyarlaması olan metin madenciliği kavramı tanımlanmıştır. Metin madenciliği; yapısal olmayan veriden ilginç, önceden bilinmeyen ve önemsiz olmayan bilgileri keşfeden, çok sayıda dokümanı analiz eden bir teknolojidir.

Bu teknoloji; doğal dil işleme, bilişsel bilimler, makine öğrenmesi ve istatistik gibi disiplinlerden teknikler kullanır.

Her yıl birkaç misline çıkan verilere en hızlı şekilde erişim eskisinden daha fazla önem kazanmıştır. Veriye erişim için genellikle arama motorları kullanılmaktadır. Arama motorlarındaki teknik, sayfaların indekslenmesi ve ardından verilen sorgu ifadesi ile indekslenen sayfa içeriklerinin karşılaştırılması şeklindedir. İndekslenmeyen sayfalara erişim mümkün değildir ve ayrıca, sorgu kelimesinin doğru girilmesi şarttır(öneri yapısı hariç). Internet aracılığıyla bilgi arama işleminde arama motorları dokümanları içindeki kelimelerle, dokümanlar arasındaki ilişkiyi bu kelimelerin tüm doküman topluluğu içindeki istatistikî kullanım örüntülerine göre ifade eder [2]. Bu sistemlerde bazı kısıtlamalar mevcuttur. Bunlar; a) dokümanın içerdiği kelimeler ile arama yapan kişinin sorgu işleminde kullandığı kelimeler farklı olabilir, b) bazı kelimelerde kelimenin tek başına kullanımındaki anlam ile başka bir kelime ile beraber kullanımındaki anlam farklı olabilir, c) kelimelerin doküman içerisindeki değerini hesaplarken sadece kelime sıklığına bakılır. Bu kısıtlardan dolayı arama motorları yapısal olmayan veriye erişimde en iyi yöntem değildir. Metin madenciliği yöntemi ise arama motorlarından çok daha fazlasını sunmaktadır. Metinlerin otomatik olarak özetlenmesi, dokümanlardan dokümanı özetleyici kavramların çıkarılması, benzer dokümanların kümelenmesi ve buna benzer teknikler metin madenciliğinin bize sunduklarıdır. Daha çok miktarda veriye, daha az bilgi ile ulaşmak mümkündür. Metin madenciliği alanındaki başarılı bir çalışma Swanson tarafından yapılmıştır. ARROWSMITH isimli bu çalışma sayesinde, biyoloji ve tıp bilimi ile alakalı iki ayrı makale grubu arasındaki ortak maddeler ve düşünceler bulunabilmiştir (Swanson, Smalheiser, 1999**)** [3].

Yapılan bir diğer çalışmada metin madenciliği teknikleri kullanılarak sistemde bulunan gerçek soru cevapları analiz edilmiş ve yapılan analiz sonunda benzer cevaplar gruplandırılarak her cevap için cevabı veren yanıtlayıcıların yüzdesi hesaplanmıştır [4].

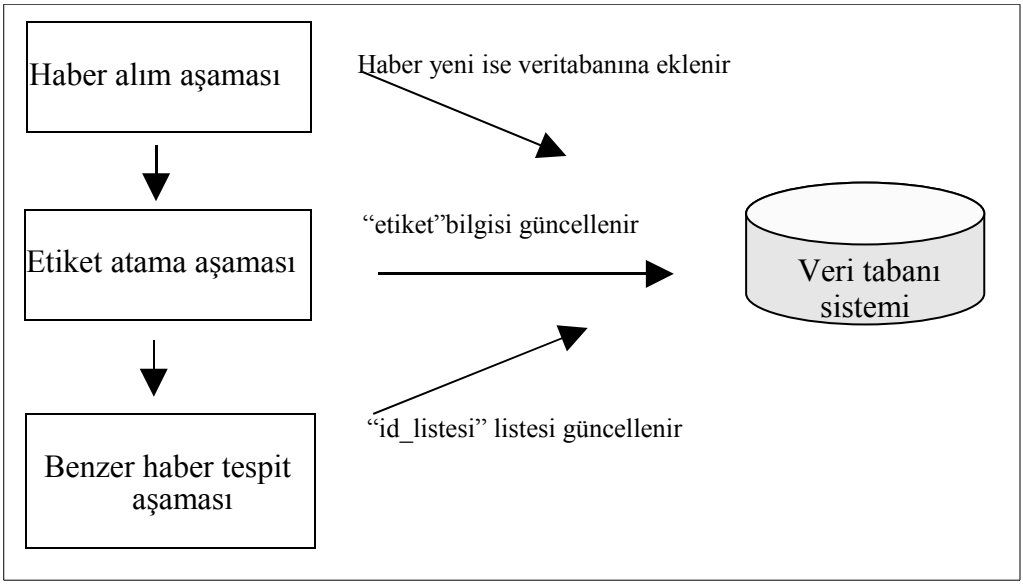
Doküman kümeleme işleminde Gizli Anlambilimsel Dizinleme(GAD) yöntemi kullanılmış, bu yöntem N-Gram kelimeler ile geliştirilmiştir. GAD yöntemi, dokümanları içerdikleri kelimelere göre değerlendirmiş ve doküman topluluğunu bir bütün olarak görmüştür. Böylece bir dokümanda geçen bir kelimenin diğer dokümanlarda geçip geçmediğini ve geçerken ortak kullanıldığı kelimeleri tespit etmesini sağlamıştır.

Çalışmamızda metin madenciliği yöntemlerinden doküman kümeleme yapılacaktır. Deneylerde sekiz farklı kategoriden 50.000 haber içeren bir veri kümesi kullanılmıştır. Makalenin ikinci bölümünde sistemin yapısı açıklanmış, üçüncü bölümünde geliştirilen uygulamadan elde edilen sonuçlara yer verilmiştir. Dördüncü bölümde alınan bir habere etiket atama aşamasında yaşanabilecek sorunlar ve ilişkili haber tespiti aşamasında karşılaşılabilecek durumlar örnek verilerek açıklanmış, beşinci bölümde çalışmadan çıkan sonuçlar verilmiştir.

2. Sistemin Yapısı

Dilde yer alan kavramlar, varlıklar, eylemler, durumlar vb. unsurlar kelimelerle ifade edilir. Bu nedenle bir belgeyi ifade edebilecek en küçük yapı taşı o belgeyi oluşturan kelimelerdir (Dumais vd., 1996) (Rehder vd., 1998). Bu yaklaşımdan yola çıkarak sistemde kayıtlı her haberi temsil edecek terim listesi haber bilgilerinden elde edilir ve bu listeye göre konusal açıdan benzer olan haberler gruplandırılır.

Geliştirilen sistem, haber alma aşamasından oluşan tek aşamalı yapıya iki aşama daha eklemiştir. Bunlar etiket atama aşaması ve benzer haber tespiti aşamalarıdır. Şekil-1'de sistemin mimarisi verilmiştir. Bu mimariye göre; öncelikle haber tablosuna yeni bir haber eklendiğinde bu haberi temsil edecek etiket listesi belirlenir ve daha sonra bu liste ile tabloda kayıtlı haberlerin etiket listesi kıyaslanarak benzer haberler tespit edilir.



Şekil 1. Sistemin mimarisi

Uygulamanın geliştirilmesinde haber depolamak için MySQL veritabanı sunucusu; programlama için Python programlama dili, Türkçe sözcüklerin köklerini bulmada ise Zemberek Türkçe dil işleme çözümleri kütüphanesi kullanılmıştır.

2.1. Veri Seti

Haberler, RSS 2.0 (Really Simple Syndication)[1] dosyalarını destekleyen haber kaynaklarından toplanmıştır. Her bir haber; başlık, özet, içerik, kaynak, kategori, link, yayınlanma tarihi ve resim bilgileriyle tutulmuştur. Toplam haber sayısı yaklaşık olarak 370.000 adettir. Bu haberlerin ilk 50.000 tanesi test amacıyla kullanılmıştır.

Örnek haber bilgisi;

Başlık	Prodi hükümeti bir yıl dayanamadı
Özet	İtalya'da geçen nisanda kurulan solcu Romano Prodi hükümeti, dış politika önergesini Senato'ya kabul ettiremeyince bir yılını dolduramadan istifasını verdi.
İçerik	Boş (Null)
Kaynak	Radikal

1 Netscape firması tarafından geliştirilen XML biçimindeki dosya

Kategori	Dış haberler
Link	http://www.radikal.com.tr/haber.php?haberno=213701
Yayınlanma tarihi	2007–02–22 22:53:00
Resim	*(Yok)

2.2. Haber Metinlerinin Temizlenmesi

Toplanan haber metinleri gereksiz veya hatalı bilgiler (yazım ve noktalama hataları, kodlama hatları) içerebilir. Metin ön işlem aşamasında hatalı sonuçlara sebep olacak faktörler ortadan kaldırılır. Web verileri Html sözdizimindedir. Elde edilen haber bilgileri de web verileri olduğundan yapılacak ilk ön işlem HTML etiketlerinin temizlenmesidir.

Metin temizleme sırasın yapılan işlemler şunlardır;

• Html ifadelerini (a href, br, b, p, font, table, div vb.) temizlemek.

• Html karakter/noktalama işaretleri kodlarını temizlemek. Örneğin “ karakter grubu noktalama işaretlerinden çift tırnağı(“) temsil eder. Bu karakter grubu tespit edilerek ya çift tırnak ile değiştirilir ya da silinir.

• Nokta (.) ve tek tırnak (') dışındaki noktalama işaretleri temizlenir. Bu noktalama işaretlerinin temizlenmemesinin nedeni; nokta, ilgili metni cümlelere ayırmada ayırıcı(separator) olarak kullanılırken, tek tırnak kendisinden sonraki karakterlerin işleme alınmamasını sağlar. Böylece işlenmesi gerekli olmayan daha az veri işlenir.

2.3. Etiket Atama

Temizlenmiş haber metinlerinin belli sayıda etiket ile sunulduğu ve bu etiketlerin terim ağırlıklarının hesaplandığı aşamadır. İçerik ya da özet bilgisi Null(boş) olmayan haberlere etiket listesi atanır. 369.000 haberde başlık bilgisinin ortalama uzunluğu 40 karakterdir. Tespit edilecek terim sayısı başlıktaki kelime sayısına eşit olacağından ve bu sayı istenilen etiket sayısından çok az olacağından sadece başlık bilgisi olan haberlere etiket atanmamıştır.

Etiket atama işlem adımları:

• Öncelikle metin token adı verilen bölümlere ayrılır

• Sonra her bir parçanın uzunluğuna bakılır. Uzunluk en az bir karakter olmalıdır.

• Uzunluk kontrolü sonrasında tek karakterli ifadelerin sayı olup olmadığı kontrol edilir. İfade sayı ise doğrudan ilgili listeye(başlık, özet ya da içerik listesine) eklenir. Değilse işleme diğer basamaklarıyla devam edilir.

• Tek başına anlamı olmayan ancak cümle içinde kullanıldığında ilgili cümleye anlam katan edat, bağlaç bv. gibi ifadeler **Sözlük** isimli veri tabanı tablosunda yer alır. Tabloda yer alan en

uzun ifade altı karakterlidir.

• Karakter sayısı birden fazla olan ifadelerin(rakam değilse) işlenme aşamasında en fazla altı karakterli olan kelimeler **Sözlük** tablosunda aranır. Bu tabloda yer alıyorsa bir metinde değeri olmayan ifadeler arasına girer ve ilgili listelere eklenmez.

• Sözlük tablosunda yer almayan kelimeler Zemberek kütüphanesi yardımıyla kök ve eklerine ayrılır. Ek listesinden çekim ekleri kaldırılarak kelimenin kökü ve yeni ek listesi ile yeni kelime üretilir. Bu şekilde ayrıştırılan kelimenin gövdesi(terim) bulunur.

• Zemberek kütüphanesinde yer almayan kelimeler olabilir, kelime listesi çok geniş değildir. Zemberek kütüphanesi tarafından çözülemeyen kelimeler doğrudan listelere eklenir. Özel isim ise özel isim listesine de eklenir.

• Bulunan terim ilgili listeye eklenir. Özel isim ise özel isim listesine de eklenir.

• Oluşturulan listelerdeki terimlerim terim ağırlıkları hesaplanır ve terim ağırlığı büyük olan ilk x terim haberin etiket listesi olarak atanır.

2.3.1. Gövde tespiti yaklaşımı

Haber metinlerindeki kelimelerin Zemberek kütüphanesi ile kök ve ekleri tespit edilir. Ek listesinde bulunan çekim ekleri[2] kaldırılarak yeni ek listesi ve kelimenin kökünden Zemberek aracılığı ile yeni kelime üretilir. Üretilen bu kelime gövde olarak alınır. Yapım ekleri[3] çekim eklerinden daha çeşitli olması ve çekim eklerinin Zemberek kütüphanesindeki karşılığın bulunması bu yaklaşımın uygulanabilirliğini arttırmıştır.

Çekim eklerinin Zemberek kütüphanesindeki karşılığı, kontrol edilecek eki almış kelimenin Zemberek çözümlemesinden gelen ek isminin kontrol edilmesiyle bulunmuştur. İsim çekim eki Hal eklerinden İlgi hali ekini almış "annenin" kelimesi anne (isim kök) + nin (ilgi hali) şeklinde çözümlenir. Zemberek kütüphanesinin verdiği çözümleme sonucu: [ISIM_KOK, ISIM_TAMLAMA_IN]. Böylece İlgi halinin Zemberek karşılığı ISIM_TAMLAMA_ek olduğu tespit edilir.

Zemberek kütüphanesi yardımıyla kök ve eklerine ayırma işleminde bir sözcük için varsa birden fazla öneri sunulur. Verilen önerileri inceleyen çalışma sahibi anlamsal olarak aradığı çözümün hangi öneride yer aldığını bilir. Bilgisayar sistemlerinin anlamsal analiz işlemi yapabilmesi için eğitilmesi, bu konuda yeti kazanması gerekir. Bu işlemde bilgisayarın böyle bir yetisi olmadığından Zemberek kütüphanesi tarafından verilen ilk öneri varsayılan(default) olarak kabul edilir. Örnek bir sonuç;

toplar : top (isim kök) + lar (çoğul eki) verilen bu kelime için Zemberek kütüphanesinden dönen sonuçlar;

topla + r : [FIIL_KOK, FIIL_GENISZAMAN_IR]
top + lar : [ISIM_KOK, ISIM_COGUL_LER]

2 Eklendiği kök veya kelimenin anlamında herhangi bir değişiklik yapmayan, bir ifade katan eklerdir.

3 Eklendiği köke yeni bir anlam kazandıran eklerdir. Kelime türetmede kullanılır.

top+ la + r : [ISIM_KOK, ISIM_DONUSUM_LE, FIIL_GENISZAMAN_IR]
top + lar : [ISIM_KOK, ISIM_KISI_ONLAR_LER]

Aranılan sonuç elde edilen sonuçlardan ikincisidir.

2.3.2. Özel isim tespiti yaklaşımı

Haberde yer alan özel isimler bir listede tutulur. Listeye eklenecek elemanların tespitinde Zemberek kütüphanesinin özel isimler için verdiği OZEL_KOK bilgisinden yararlanılır. Bu şekilde temsil edilemeyen kelimeler için ilk karakterin ASCII(**A**merican **S**tandard **C**ode for **I**nformation **I**nterchange*)* karşılığına bakılır. ASCII değeri 65- 90 arasında yer alıyorsa kelimenin ilk karakteri büyük harftir ve özel isim kabul edilir. Metinler cümle cümle işlendiklerinden cümlelerin ilk kelimeleri için ASCII kontrolü yapılmaz. Cümlelerin ilk kelimeleri özel isim ise bu kelimelerin tespiti yapılamaktadır. Eğer ilk kelime bir kısaltmaysa(örneğin TBMM, TCDD, TSK, TDK) Zemberek kütüphanesi kısaltmalar için ISIM_KOK bilgisi verir. Özel isim listesine eklenmesi gereken bu kelimeler için program içinde bir kontrol yer alır.

2.3.3. Terim ağırlıklandırma yaklaşımı

Kelime gövdelerinden oluşan haberin başlık, özet ve içerik listeleri[4] elde edildikten sonra içerik ya da özet listesi Null olmayan haberler için bu listeler etiket atamada kullanılır. Etiket tespitinde kullanılacak listedeki terimlerin listede geçme sıklığı tespit edilir ve böylece ayrık terimler belli olur. Ayrık terimleri içeren sözlük yapısı gövde, sayı, başlık, anahtarlarına sahiptir. Gövde; ayrık terimi, sayı; terimin geçme sayısını özet, özel gösterir. Başlık, özet ve özel anahtarları boolean(True, False) şeklindedir. Terim eğer başlıkta, özette ve özel isim listesinde geçiyorsa değeri 1, aksi durumda 0'dır. Örnek bir haber üzerinde gösterirsek,

Haberin başlığı: İran'a tanınan süre doldu
Özeti : BM Güvenlik Konseyi'nin yaptırım kararında İran'a uranyum zenginleştirmeyi durdurması için verdiği süre doldu.
İçeriği : Null

Başlık listesi : ['iran', 'tanınan', 'süre', 'dol']
Özet listesi : ['bm', 'güvenlik', 'konsey', 'yaptırım', 'karar', 'iran', 'uranyum', 'zenginleştirme', 'durdurma', 'verdik', 'süre', 'dol']
Özel isim listesi : ['iran', 'bm', 'güvenlik', 'konsey']

İçeriği Null olduğu için sadece özet bilgisine bakılır. Özet bilgisinden oluşan etiket listesi;
[{'ozel': 1, 'baslik': 0, 'govde': 'bm', 'sayi': 1, 'ozet': 0}, {'ozel': 1, 'baslik': 0, 'govde': 'güvenlik', 'sayi': 1, 'ozet': 0}, {'ozel': 1, 'baslik': 0, 'govde': 'konsey', 'sayi': 1, 'ozet': 0}, {'ozel': 0, 'baslik': 0, 'govde': 'yaptırım', 'sayi': 1, 'ozet': 0}, {'ozel': 0, 'baslik': 0, 'govde': 'karar', 'sayi': 1, 'ozet': 0}, {'ozel': 1, 'baslik': 1, 'govde': 'iran', 'sayi': 1, 'ozet':0}, {'ozel': 0, 'baslik': 0, 'govde': 'uranyum', 'sayi': 1, 'ozet': 0}, {'ozel': 0, 'baslik': 0, 'govde': 'zenginleştirme', 'sayi': 1, 'ozet': 0}, {'ozel': 0, 'baslik': 0, 'govde': 'durdurma', 'sayi': 1, 'ozet': 0}, {'ozel': 0, 'baslik': 0, 'govde': 'verdik', 'sayi': 1, 'ozet': 0}, {'ozel': 0, 'baslik': 1, 'govde': 'süre', 'sayi': 1, 'ozet': 0}, {'ozel': 0, 'baslik': 1, 'govde': 'dol', 'sayi':1, 'ozet': 0}]

4 Python programlama dilinde yer alan, farklı tipte verileri içeren veri tipidir.

Her terim için elde edilen bu bilgiler sonrasında terimlerin metin içerisindeki değerini ifade eden ağırlıkları hesaplanır. Var olan ağırlıklandırma yöntemlerinden(terim frekansı (tf), ters doküman frekansı (idf), terim frekansı-ters doküman frekansı (tf-idf), terim ayrıştırma değeri, olasılıksal terim ağırlıklandırma, tek terim doğruluğu genetik algoritmalar vb.) frekans ağırlıklandırma yöntemi seçilmiştir. Bu yöntemde vektördeki kelimelerin ağırlığı metindeki geçiş sayısına(frekansına) eşittir.

Gün içerisinde en az 24 kez çalışan ve günde ortalama(mayıs ayına kadar) 1500 haber alan bu yapıda, bir haberi temsil edecek etiketleri tespit ederken haber sadece kendinden sorumlu tutulur. Terim ağırlığı hesaplanırken terimin sadece o habere özgü değeri hesaplanabilir. Çok daha az ve uzun sürede güncellenen bir sistemde terimleri ağırlıklandırırken veri tabanı sisteminde bulunan tüm bilgiler kullanılabilir.

Terim ağırlıklarının hesaplanmasında üç yaklaşım denenmiştir. Kullanılan yaklaşımlardaki ortak kavramlar;

Etiket listesi : İçeriği ya da özeti Null olmayan haberin içerik ya özet metinlerindeki terimlerinin listesi.
Özel isim listesi : Haberde geçen özel isimlerin listesi(tekrarsız).
Başlık listesi : Başlık bilgisindeki terimlerin listesi
Özet listesi : Özet bilgisindeki terimlerin listesi
Wd, t : d haberindeki t terimin ağırlığı
tf x, t : x(etiket, özel, başlık, özet) listesindeki t teriminin geçme sıklığı
L x : x(etiket, özel, başlık, özet) listesinin eleman sayısı

İlk yaklaşımda d haberinde geçen t teriminin ağırlığı;

$$Wd, t = (tf\ etiket, t / L\ etiket) + (tf\ özel, t / L\ özel) + (tf\ başlık, t / L\ başlık) + (tf\ özet, t / L\ özet)$$

şeklinde hesaplanır.

İkinci yaklaşımda özel isim ağırlıklandırmasında başlık, özet ve içerikte yer alan özel isimler belirli katsayılarla çarpılır. Uygulanan katsayılar birbirinden farklıdır. Haberin başlık bilgisi özet ve içerik bilgisinden okuyucu gözünde daha önceliklidir, bu nedenle katsayılar farklı belirlenmiştir (k1> k2> k3).

k1 : Başlıkta geçen özel isimlerin katsayısı
k2 : Özette geçen özel isimlerin katsayısı
k3 : Etiket listesinde geçen özel isimlerin katsayısı
tip : Değeri, terim özel isimse 1, değilse 0'dır.

$$Toplam = tf\ başlık, t \bullet k1 + tf\ özet, t \bullet k2 + tf\ etiket, t \bullet k3$$
$$Wd, t = (tf\ etiket, t + tf\ başlık, t + tf\ özet, t + tip \bullet Toplam) / (L\ etiket + L\ başlık + L\ özet)$$

Bir diğer terim ağırlıklandırma yaklaşımı Glasgow modelinden[5] yola çıkılarak oluşturulmuştur.

L ux : x(etiket, özel, başlık, özet) listesinin ayrık(unique) eleman sayısı

Toplam = tf başlık, t • k1 + tf özet, t • k2 + tf etiket, t • k3
Wd , t = log (tf başlık, t + tf özet, t + tf etiket, t + tip • Toplam + 1) / log (L uetiket + L ubaşlık + L uözet)

Üç yaklaşımın özel isim olan terimlerin ağırlıklarındaki değişimi bir örnekle açıklarsak;
Haber- 1 Başlık terim listesi : ['ırak', 'tecavüz', 'kriz']
Özet terim listesi : ['ırak', 'direnişçi', 'şii', 'polis', 'iki', 'sünni', 'kadın', 'tecavüz', 'sonra', 'intikam', 'saldırı', 'çağrı', 'yap']
Özel isim listesi : ['ırak', 'şii', 'sünni']

Haber- 2 Başlık terim listesi : ['ırak', 'hedef', 'şii', 'az', '41', 'ölü']
Özet terim listesi : ['ırak', 'başkent', 'bağdat', 'iki', 'ayrı', 'intihar', 'saldırı', 'şii', 'hedef', 'alındı', 'saldırı', 'az', '41', 'kişi', 'öl']
Özel isim listesi : ['ırak', 'şii', 'bağdat']
Tablo 1.'de yer alan özel isimlerin ağırlıkları karşılaştırıldığında terimlerin ağırlıklarındaki değişim tespit edilir.

Terimler	1.ağırlıklandırma yaklaşımı sonuçları	2.ağırlıklandırma yaklaşımı sonuçları	3.ağırlıklandırma yaklaşımı sonuçları
Haber–1			
Irak	0.74359	0.37500	0.70184
Şii	0.41026	0.15385	0.42832
Sünni	0.41026	0.15385	0.42832
Haber–2			
Irak	0.56667	0.28571	0.64956
Şii	0.56667	0.28571	0.64956
Bağdat	0.40000	0.13333	0.41629

Tablo 1. Terim ağırlıklandırma yaklaşımı sonuç tahlili

Ağırlıklandırma işleminde özel isimlere, başlığa ve içeriği olan haberlerde özette geçen terimlere artı değer verilmesi planlanmıştır. Web ortamından elde edilen haber bilgilerinde özet bilgisi içerik bilgisinden elde edilmiştir. Haber kaynakları haberlerin özet bilgilerine içerik bilgisinin ilk paragrafı atamakta, yine yazılımda özet bilgisi olmayan haberler için ilk 150 karakter özet bilgisi olarak alınmaktadır. Sonuç olarak özet bilgisi tüm metinden çıkarılan özel bir bilgi değildir. Bu sebepten ağırlık hesaplamalarında özet ile alakalı değişkenler kullanılmamıştır. Kullanılan hesaplama formüllerini yeniden yazarsak;

1.Wd , t = (tf etiket , t / L etiket) + (tf özel , t / L özel) + (tf başlık , t / L başlık)

2. Toplam = tf başlık, t • k1 + tf etiket, t • k3
Wd , t = (tf etiket, t + tf başlık, t + tip • Toplam) / (L etiket + L başlık)

3. Toplam = tf başlık, t • k1 + tf etiket, t • k3
Wd , t = log (tf başlık, t + tf etiket, t + tip • Toplam +1) / log (L uetiket + L ubaşlık)

Kullanılacak hesaplama yaklaşımı seçilirken, açıklanan ağırlıklandırma yaklaşımlarının sonuçları incelenmiştir. Rastgele seçilmiş 2000 haberin (tüm haberlerin etiket listesi Null dan farklı ise işleme alınan benzerlik oranı sayısı= (2000*2000)/2) olacaktır tepiti için sonuçların normal dağılıma uyup uymadıklarına bakılmış ve benzerlikler için standart sapma bilgisi kullanılmıştır. Standart sapması daha büyük olan yöntem daha iyi sonuç verecektir. Tablo 2'de yer alan sonuçlar neticesinde ikinci ağırlıklandırma yaklaşımının kullanımına karar verilmiştir.

1796 haber (1796 • 1795)/2 = 1611910 benzerlik oranı üzerinde	1.ağırlıklandırma yaklaşımı	2.ağırlıklandırma yaklaşımı	3.ağırlıklandırma yaklaşımı
Standart sapma (s)	0.0433	0.045	0.0405

Tablo 2. Terim ağırlıklandırma yaklaşımlarının standart sapma sonuçları

2.3.4. Etiket sayısı tespiti yaklaşımı

Habere atanacak etiketleri içeren etiket listesi hazırlandıktan sonra terim ağırlığına göre büyükten küçüğe sıralanmış terimler arasında ilk x sayıda terim habere atanır. x sayısı belirlenirken; kullanılan tüm haberlerin içeriği olmadığından ve özet bilgisi içerik bilgisinden daha az karakter içerdiğinden her iki durum düşünülmüştür. Tablo 3.'de 50.000 haber üzerinde yapılan sorgu ile elde edilen başlık, özet ve içerik bilgilerinin ortalama karakter sayısı yer alır.

50.000 haber için	Temizlenmiş haberde ortalama karakter sayısı		Temizlenmemiş haberde ortalama karakter sayısı	
Null haberler dâhil	•	✘	✓	✘
Başlık	42	42	42	42
Özet	187	204	188	205
İçerik	1727	1830	1878	1991

Tablo 3. Başlık, özet ve içerik bilgilerinin ortalama karakter sayısı

Etiket sayısı belirlenirken içerik ve özet bilgilerinde yer alan ortalama ayrık terim sayıları incelendi. Tablo- 4 ve Tablo- 5'te ortalama ayrık terim sayıları verildi.

	Ortalama ayrık terim sayısı	Hesaplandığı haber sayısı
Özet	20	28358
İçerik	112	16686

Tablo 4. Ortalama ayrık terim sayısı

İnceleme no	Haber sayısı	Durum	Ortalama ayrık terim sayısı
1	3000	Sıralı	66
2	3000	Sıralı	61
3	3000	Sıralı	61
4	2569	Sıralı	32
5	3000	Rastgele	55
6	1600	Rastgele	53
7	5000	Rastgele	56

Tablo 5. Ortalama ayrık terim sayısı (2)

İncelenen ayrık terim sayıları sonucunda sadece özet bilgisine sahip haberlerin ortalama ayrık terim sayısının 20 olması belirlenecek sayısının bu rakama kısmen yakın olmasını gerektirmiştir. Tablo 4 ve Tablo5'te yer alan en küçük değerlerin(20, sıralı seçilmiş haberlerde 32, rastgele seçilmiş haberlerde 53) ortalaması etiket sayısı olarak belirlenmiştir.

Etiket sayısı = [Min(Tablo.4) + Min(Tablo5, sıralı) +Min(Tablo5,rastgele)] / 3
= [20 + 32 + 53] / 3
= 35

İşlemler sonucunda etiket sayısı 35 olarak hesaplanmıştır. Bu sayıdan daha az ayrık terim sayısına sahip haberler için tüm terimler etiket olarak atanır.

2.4. Benzer Haber Tespiti

Bu aşamada birbiriyle benzerlik arzeden haberler gruplanır. Terim ve ağırlıklarından oluşan etiket listesi atanılan bir haber, 'son dakika' veya kendisiyle aynı kategoride olan ve etiket listesi Null olmayan haberlerle eşleştirilir. Son dakika kategorisinde yer alan haberler çeşitli kategorilerde(ekonomi, politika, spor, sağlık, kültür sanat vb.) haber içerdiğinden bu kategorideki haberlerde kullanılır.

Benzerlik hesabında, Dice, Kosinüs ve Jaccard yöntemlerinden Kosinüs yöntemi tercih edilmiştir. Kosinüs benzerlik değeri; doküman vektörleri iç çarpımının doküman boyutları çarpımına bölümü şeklinde elde edilir. Ortak terimlerin koyu renkle belirtildiği etiket listesi verilen haberlerin benzerliğini hesaplarsak;

Haber–1 etiket listesi: **ırak**, 0.64624, sünni, 0.42832, **şii**, 0.42832, tecavüz, 0.39624, yap, 0.27024, çağrı, 0.27024, **saldırı**, 0.27024, intikam, 0.27024, sonra, 0.27024, kadın, 0.27024, **iki**, 0.27024, polis, 0.27024, direnişçi, 0.27024

Haber–2 etiket listesi: **şii**, 0.59810, **ırak**, 0.59810, **saldırı**, 0.41629, bağdat, 0.41629, 41, 0.36673, az, 0.36673, hedef, 0.36673, öl, 0.26265, kişi, 0.26265, alındı, 0.26265, intihar, 0.26265, ayrı, 0.26265, **iki**, 0.26265, başkent, 0.26265

İç çarpım = (0.64624 • 0.59810) + (0.42832 • 0.59810) + (0.27024 • 0.41629)+ (0.27024 • 0.26265)
= 0.896
lHaber–1l = (0.64624 ≤ + 2 • 0.42832 ≤ + 0.39624 ≤ + 9 • 0.27024 ≤)Ω
= 1.264
lHaber–2l = (2 • 0.59810 ≤ + 2 • 0.41629 ≤ + 3 • 0.36673 ≤ + 7 • 0.26265 ≤)Ω
= 1.395

Sim(Haber–1, Haber–2) = iç çarpım / lHaber–1l • lHaber–2l
= 0.896 / [1.264 • 1.395]
= 0.5

Kosinüs hesaplama sonucu belirlenen eşik değerinden büyük ise bu haberler birbirine benzerdir. Eşik değeri tespitinde çıkan sonuçlar değerlendirilmiş ve varsayılan değer 0.5 kabul edilmiştir.

3. Geliştirilen Uygulamadan Elde Edilen Sonuçlar

Bu bölümde sistemin çalışması örnek haberler üzerinden gösterilmiştir. Bulunan sonuçlar ile okuyuculardan alınan sonuçlar karşılaştırılarak sonuçlar değerlendirilmiştir.

Haber bilgisi	İlişkili haberler
Bush'u Irak'ta yalnız bırakıyor 22.02.2007 ABD, istikrarı sağlamak için Irak'a takviye güç göndermeyi planlarken Washington'un bölgedeki müttefikleri birer birer çekiliyor. İngiltere Başbakanı Tony Blair, yaz sonuna kadar Irak'tan 1600 askeri geri çekeceklerini açıkladı. Danimarka da Irak'taki 470 askerini ağustosa kadar çekeceğini duyurdu. Çekilme takvimini dün açıklayan Blair, "Irak savaşı başladığında 40 bin olan ve 2 yıl önce 9 bine düşen, şu anda da 7100 olan asker sayımız, yaklaşık 5 bin 500'e düşecek." dedi. Geri kalan askerlerin, destek ve eğitim görevi yapacağını kaydeden Blair, Irak'taki İngiliz askerî varlığının en azından 2008'e kadar süreceğini de söyledi. Beyaz Saray'ın, "bölgedeki ilerlemenin işareti" olarak değerlendirdiği kararı, Irak hükümeti ise memnuniyetle karşıladı. Irak'taki en büyük müttefikin çekilme takvimini açıklamasının ardından Demokratlar'ın, Bush üzerindeki "Askerleri geri çek." baskısını da artırmaları bekleniyor. Dış Haberler Servisi	ABD, Irak'ta yalnız kalıyor İngilizler de yolcu ABD Başkanı Bush asker takviyesine hazırlanırken, müttefikleri Irak'tan kaçıyor. Britanya Başbakanı Blair 7 bin 200 askerinin büyük kısmını, Danimarka Başbakanı Rasmussen 470 askerinin hepsini çekecek. (Benzerlik oranı 0.54) Blair: Irak'tan 1600 asker çekeceğiz İngiltere Başbakanı Tony Blair, Irak'tan gelecek aylarda 1600 asker çekeceklerini açıkladı. Blair, Avam Kamarasında yaptığı konuşmada, "Irak savaşı ... (Benzerlik oranı 0.57)

Haber bilgisi	İlişkili haberler
Doğuş Didim'de marina açıyor Doğuş Grubu, 2003'te D-Marin Turgutreis ile başladığı yat limanı işletmeciliğine, Didim'de kuracağı ve tamamlandığında Türkiye'nin üçüncü büyük yat limanı kapasitesine sahip olacak yatırımıyla devam ediyor.	Doğuş'tan Didim'e 52 milyon dolarlık marina Doğuş Grubu'nun 52 milyon dolar yatırımla kuracağı Türkiye'nin üçüncü büyük yat limanı D-Marine Didim'in temeli dün atıldı. Grup, geçen yıl hizmete giren Turgutreis Yat Limanı'nın ardından, Didim Marina ve önümüzdeki dönem açacağı Dalaman Yat Limanı'yla birlikte toplam 200 milyon dolar yatırım yapmayı planlıyor.(Benzerlik oranı 0.59)
	Doğuş'tan yaz turizmine 200 milyon dolarlık yatırım Doğuş Grubu'nun 52 milyon dolarlık yatırımla kuracağı Türkiye'nin üçüncü büyük yat limanının temeli Didim'de atıldı. Grup, 200 milyon dolarlık yatırım yapacak. (Benzerlik oranı 0.58)

3.1. Başarı Testi

Bu bölümde iki ve üç haberden oluşan haber grupları on iki kişiden oluşan iki farklı okuyucu grubu tarafından incelenmiş ve içerik açısından haberlerin benzerliğine karar verilmiştir. Verilen iki haberin konu açısından benzerliği okuyucular tarafından değerlendirilmiş ve bu sonuçların ortalaması Tablo- 6'da verilmiştir. Aynı işlem üç haberden oluşan bir diğer haber grubu üzerinde yapılmış farklı on iki kişiden oluşan okuyucu topluluğunun değerlendirme sonuçları ve bu sonuçların ortalaması da Tablo- 7'de verilmiştir.

Kişi No	Yüzde	Kişi No	Yüzde
1	90	7	60
2	90	8	30
3	70	9	70
4	90	10	50
5	70	11	40
6	80	12	40
Ortalama= 780 / 12 = 65			

Tablo 6. Okuyucudan alınan benzerlik yüzdesi

On iki okuyucunun %65 benzer olduğuna karar verdiği bu haberler, yazılım tarafından %63 benzer bulundu.

Kişi No	Yüzde			Kişi No	Yüzde		
	1-2	1-3	2-3		1-2	1-3	2-3
1	80	5	10	7	85	0	0
2	100	0	0	8	90	5	5
3	80	10	10	9	80	5	5
4	60	0	0	10	90	0	0
5	90	0	0	11	80	0	0
6	90	0	0	12	80	0	0
1.ve 2. haber için ortalama= 1005 / 12 = 84							
1.ve 3. haber için ortalama= 25 / 12 = 2.1							
2.ve 3. haber için ortalama= 30 / 12 = 2.5							

Tablo 7. Okuyucudan alınan benzerlik yüzdesi

Okuyucular tarafından 1. ve 2. haber %84, 1. ve 3. haber %2.1 ve 2. ve 3. haber %2.5 benzer bulunmuştur. Geliştirilen yazılımda benzerlikler 1. ve 2. haber arasında %95, 1. ve 3. haber arasında%32, 2. ve 3. haberlerin arasında ise %33 çıkmıştır.

Elde edilen bu sonuçları yorumlarsak birinci tablodan elde edilen başarı daha yüksektir. İkinci tabloda 1. ve 2. haber için sonuçlar yakın olsa da diğer sonuçlar daha uzak çıkmıştır. Uzak çıkan sonuçlar yazılımın bir eksikliği olarak yorumlanmamalıdır. Makine hesabı insanların fark edemediği ilişkileri ortaya çıkarabilir. Okuyucu gözünde sıfır ilişkiye sahip iki haber aslında birbiriyle alakalı durumlar içerebilir.

4. Karşılaşılabilecek Olumsuz Durumlar

4.1. Etiket Atanamama

Haber bilgilerini işleme sırasında bazı sorunlar yaşanabilir. Karakter kodlamasından kaynaklanan sorunlar bunların başında gelir. Çalışma sırasında bu tarz hatalarla sıkça karşılaşılmıştır. İstenmeyen karakterlerin tamamını işlemin en başında bilmek zordur, yazılımın geliştirilmesi aşamasında ortaya çıkan aksamalarda sorunlu karakterler tespit edilmiştir. Örneğin "" karakterini içeren haber metinlerinde bu karakter tespit edilememişse ve temizlenememişse etiket atama aşamasında işlem tıkanır.

İstenmeyen karakter kodlamasına sahip örnek haber verisi;

"Genelkurmay BaÅŸkanlÄ±ÄŸÄ±, askerÃ® personelin sandÄ±ÄŸa gitmesi iÃ§in seÃ§im

haftasÄ±nda bÃ¼tÃ¼n kamp ve dinlenme tesislerini kapatma kararÄ± aldÄ±. AskerÃ® personel, 22 Temmuz'un Ã¶ncesindeki ve sonrasÄ±ndaki dÃ¶rt gÃ¼n iÃ§inde yÄ±llÄ±k kamp ve tatil hakkÄ±nÄ± kullanamayacak. Sabah gazetesinde yayÄ±nlanan habere gÃ¶re, konuyla ilgili dÃ¼zenleme tÃ¼m kuvvet komutanlÄ±klarÄ±na duyuruldu. AskerÃ® personel yoÄŸunlukla temmuz ve aÄŸustos aylarÄ±nda askerÃ® kamplardan 15 gÃ¼nlÃ¼k sÃ¼relerle yararlanÄ±yordu. KararÄ±n, 100 binden fazla rÃ¼tbeli personeli ve ailelerini ilgilendirdiÄŸi belirtiliyor. ...”

4.2. Benzerlik Oranı Hesabında Farklı Sonuçların Elde Edilmesi,

Dilimizde yer almayan ancak dilimize yerleşen bazı kavramlar Türkçeleştirilmektedir. Benzer haber tespiti aşamasında yabancı kökenli sözcüklerin Türkçeleştirilmesinden kaynaklı hesaplama sorunları yaşanabilir. Bu durumu karşılaşılmış bir örnekle açıklarsak;

Haber- 1) Euro Bölgesi'nde enflasyon düştü
Etiket listesi : euro, 0.54994, 1, 0.35542, yüzde, 0.35542, ayın, 0.35542, ab, 0.35542, birlik,0.35542, avrupa, 0.35542, enflasyon, 0.33719, gerile, 0.22424, 8, 0.22424, nisan, 0.22424, oran, 0.22424, ortalama, 0.22424, açıklanan, 0.22424, olarak, 0.22424, 9, 0.22424, mart, 0.22424, ülke, 0.22424, 13, 0.22424, kullanan, 0.22424

Haber- 2) Avro bölgesinde enflasyon düşüşe geçti
Etiket listesi : avro, 0.54364, 1, 0.35542, yüzde, 0.35542, ayın, 0.35542, ab, 0.35542, birlik 0.35542, avrupa, 0.35542, enflasyon, 0.33333, gerile, 0.22424, 8, 0.22424, nisan, 0.22424, oran, 0.22424, ortalama, 0.22424, açıklanan, 0.22424, olarak, 0.22424, 9, 0.22424, mart, 0.22424, ülke, 0.22424, 13, 0.22424, kullanan, 0.22424

Euro kelimesi İngilizce kökenli bir sözcüktür. Ekonomi çevrelerinin bu sözcüğü kullanması sonucu Türk Dil Kurumu (TDK) bu sözcüğe Türkçe karşılık olarak “Avro” kelimesini belirlemiştir. Ancak Avro ile ilgili haberlerin bir kısmında hala “Euro” kullanılır. Etiket listelerinde yer alan “Avro” ve “Euro” sözcükleri eşleme (matching) aşamasında farklı terim olarak algılanırlar. Bu algılama sonucunda bu iki haberin benzerliği 0.83 olmuştur. Etiket listeleri dikkatle incelenirse Avro ve Euro süzcükleri dışındaki tüm terim ve terimlerin ağırlıkları birbirine eşittir. Eğer yazılım Avro ve Euro sözcüklerini terim karşılaştırma aşamasında aynı kabul etse benzerlik 0.99 olacaktır. Türkçe karşılığı olan bazı yabancı kökenli sözcükler; distribütör(dağıtıcı), Euro(Avro), trend(eğilim).

5. Sonuç ve Öneriler

Farklı kaynaklardan toplanılan metin verileri üzerinde işlem yapmak zor bir iştir. Web ortamından elde edilen verilerin işlenmesi ise daha zordur. Bilgilerin temin edildiği sayfaların karakter kodlaması, metinlerde yer alan HTML sözdizimi ifadeleri ve karakter kodları, Türkçe metinlerde yer alan yazım ve noktalama yanlışları, bu bilgilerin işlenmesinde sorun yaratabilecek özelliklerdir.

Haber depolayarak arşiv oluşturma ve depolanan haberleri değerlendirme maksadıyla 2007 yılı şubat ayında geliştirilmeye başlanan haberara.net çalışması birçok kaynaktan aynı ya da farklı konularda yayınlanan haberleri toplar. Haber arama servisinin işlevselliği arttırmak ve birbiriyle ilişkili haberler tespit etmek amacıyla bu çalışma hazırlanmıştır. Çalışma, aynı ya da farklı kaynaktan alınan aynı kategorideki benzer konulu haberleri gruplandırır. Yayınlanan haber bilgilerine benzer haber bilgisinin eklenmesi amaçlanmıştır. Elde edilen sonuçları itibariyle başarılı bir çalışmadır.

Hazırlanan uygulama, veri tabanı sisteminde yer alan ve yer alacak habere onları temsil edecek belli sayıda etiket atar. İlgili haber tespiti aşamasında etiket bilgilerini kullanarak haberler arasındaki ilişkiyi yorumlar. İstatistiktiksel benzerlik tespiti yapar.

Bu tarz çalışmalarda terimlerin anlamsal özelliklerinden de yararlanılmalıdır. Türkçe zengin bir dil olduğundan ve yabancı kökenli sözcüklerin yer edindiği bir dil olduğundan terim eşleme aşamasında çalışmamızda olmayan kontroller kullanılabilir. Yazılışları aynı anlamları farklı Türkçe sözcükler ve Türkçe'ye kazandırılmış ancak özgün halde kullanılan yabancı sözcüklerin kontrol edilmesi yararlı olabilir. Terim ağırlıklandırmada frekans ağırlıklandırma yaklaşımına göre üç farklı hesaplama yaklaşımı denenmiştir. Terim ağırlıklandırmada kullanılacak verinin özelliklerine göre daha farklı bir ağırlıklandırma tercih edilebilir.

6. Kaynaklar

[1] Grobernik M., Mladenic D., "Text-mining Tutorial", **J. Stefan Institute**, Slovenia

[2] Berry M. W., Drmac Z. and Jessup E. R., "Matrices, Vector Spaces, and InformationRetrieval", **SIAM Review**, 1999

[3] ARROWSMITH http://kiwi.uchicago.edu/webwork/PURPOSE.html

[4] Mizrahi A.R., Weisenstern A.M, "Survey System", 2003

[5] http://www.miislita.com/term-vector/term-vector-4.html

[6] Zhao Z., Liu H., "Searching for Interacting Features", **Department of Computer Science and Engineering Arizona State University**

[7] Yair Even- Zohar, "Introduction to Text Mining", **Automated Learning Group National Center for Supercomputing Applications University of |||inois**

[8] Pilavcılar İ.F., "Metin Madenciliği ile Metin Sınıflandırma", **FBE, Yıldız Teknik Üniv.**, Yüksek Lisans Tezi, 2007.

[9] Güven A., Bozkurt O.Ö. ve Kalıpsız O., "Gizli Anlambilimsel Dizinleme Yönteminin N-gram Kelimelerle Geliştirilerek, İleri Düzey Doküman Kümelemesinde Kullanımı ", **Bilgisayar Müh. Bölümü, Yıldız Teknik Üniversitesi**

[10] Güven A., "Türkçe Belgelerin Anlam Tabanlı Yöntemlerle Madenciliği ", **FBE, Yıldız Teknik Üniversitesi**, 2007

[11] Garcia Dr. Edel, "Term Vector Calculations A Fast Track Tutorial", 2005

[12] Han J. ve Kamber M. “Data Mining: Concepts and Techniques”, **Morgan Kaufmann**, San Francisco 2000

[13] Mooney Raymond J. ve Nahm Un Yong, “Text Mining with Information Extraction” **Department of Computer Sciences, University of Texas**, Austin

[14] Hearst M., http://people.ischool.berkeley.edu/~hearst/text-mining.html

[15] Blumberg R., Atre S., “The Problem with Unstructured Data”, **DM Review Magazine**, 2003

[16] https://zemberek.dev.java.net/surumler/v04/zemberek_0.4.0.html

[17] Kovalerchuk B., Vityaev E., “Data mining in finance: advances in relational and hybrid methods”. **Kluwer Academic Publishers**, 2002.

[18] Karadağ A., Uyarer T., http://haberara.net, 2007