Metin Madenciliği ile Benzer Haber Tespiti

Anıl KARADAĞ* Hidayet TAKÇI*

*Gebze Yüksek Teknoloji Enstitüsü, Bilgisayar Mühendisliği Bölümü, Kocaeli

İçerik

- Giriş
- Yapısal Olmayan Veri
- Metin Madenciliği
- Sistemin Yapısı
  - Kullanılan veri seti
  - Haber metinlerinin temizlenmesi
  - Etiket atama
  - Benzer haber tespiti
- Elde Edilen Sonuçlar
- Sorular

Giriş

Dijital depolama ortamlarının kapasitelerinindeki artış ve bilgisayar sistemlerinin kullanım artışları sonucu depolanan veri miktarları büyük boyutlara ulaşmıştır.

Depolanan yapısal olmayan verinin yönetimi bilgi teknolojisinde ciddi problemlerden biridir. Merrill Lynch potansiyel olarak kullanılabilen iş bilgisinin %85'inden fazlasının yapısal olmayan veriden çıkarıldığını tahmin eder [7].

Yapılsal Olmayan Veri

Yapısal olmayan veri (unstructured data) bilgisayarla işlenen bir veri yapısına sahip olmayan ya da makine tarafından kolay okunamayan bilgi Kütlesidir. Yapısal olmayan veri örnekleri;

ses (örnek telefon kayıtları) ve video verileri, e-posta içerikleri, kelime işlemci dokümanları, web sayfalarında yer alam forum verileri anket cevapları, müşteri, kamu kurumları vs. bildirileri, öneri ve şikayet bilgileri, wiki, çevrimiçi chat vb.

Metin Madenciliği

Metin madenciliği; yarı-yapısal ya da yapısal olmayan veriden ilginç, önceden bilinmeyen ve önemsiz olmayan bilgileri keşfeden, çok sayıda dokümanı analiz eden bir teknolojidir.

Temel Yaklaşım

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.

Kullanılan Veri Seti

RSS 2.0 (Really Simple Syndication) dosyalarını destekleyen haber kaynaklarından elde edilen haberler kullanılmıştır. Her haber; başlık, özet, içerik, kaynak, kategori, link, yayınlanma tarihi ve resim bilgileriyle saklanmıştır.

Örnek veri seti

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
Kategori	Dış Haberler
Link	http://www.radikal.com.tr/haber.php?haberno=213701
Yayınlanma Tarihi	2007–02–22 22:53:00
Resim	Boş (NULL)

Haber Metinlerinin Temizlenmesi

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 İşaretlerinin 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.

Etiket Atama Aşaması

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.

İşlem Adımları-1

• 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.

İşlem Adımları-2

Altı karakterli olan kelimeler sözlük tablosunda aranır. Bu tabloda yer alıyorsa bir metinde değeri olmayan ifadeler arasına girer ve ilgil 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.

Gövde tespiti yaklaşımı

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

Örnek

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

Terimlerine ayrılmış örnek haber

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', 'zenginleştirme', 'durdurma', 'verdik', 'süre', 'dol']
Özel isim listesi : ['iran', 'bm', 'güvenlik', 'konsey']

Terimlerine ayrılmış örnek haber-2

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}]

Terim Ağırlıklandırma

Tanımlar:

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

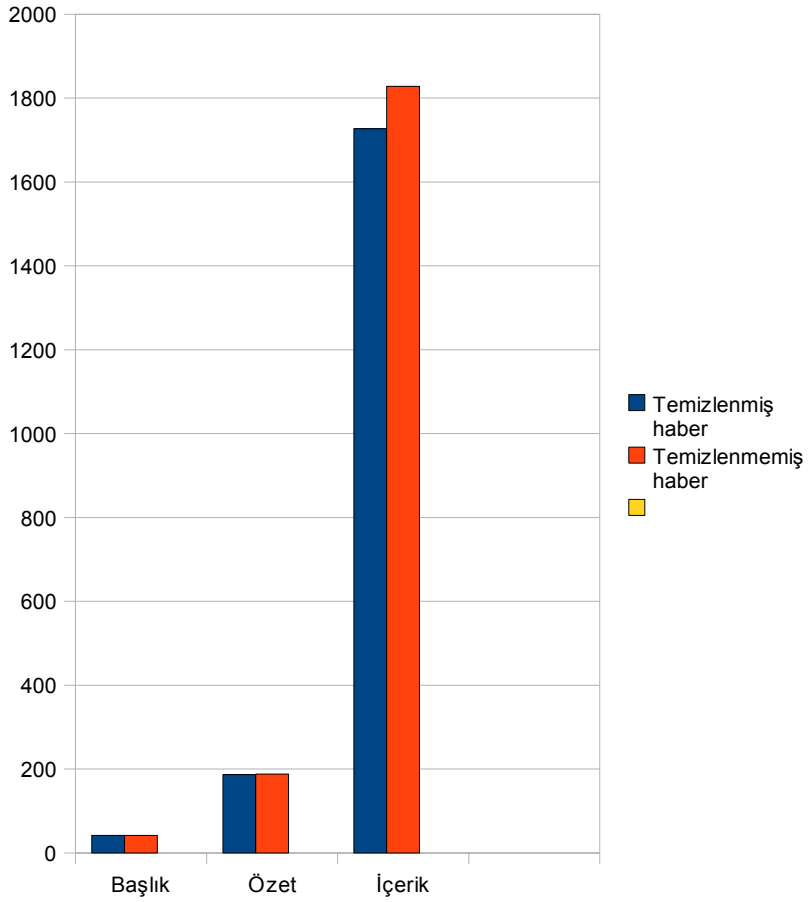
Wd, t : d haberindeki t terimin ağırlığı

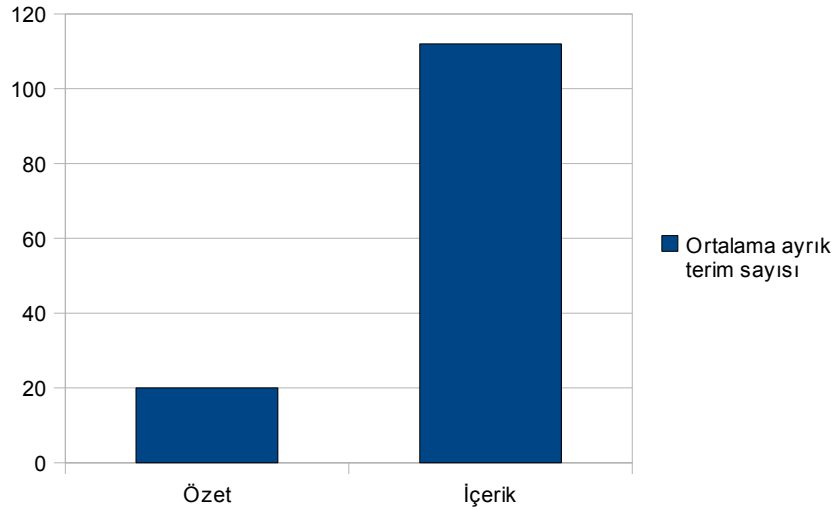
tf x, t : x(etiket, özel, başlık, özet) listesindeki t teriminin geçme sıklığı

Lx : x(etiket, özel, başlık, özet) listesinin eleman sayısı

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

Etiket Sayısının Belirlenişi

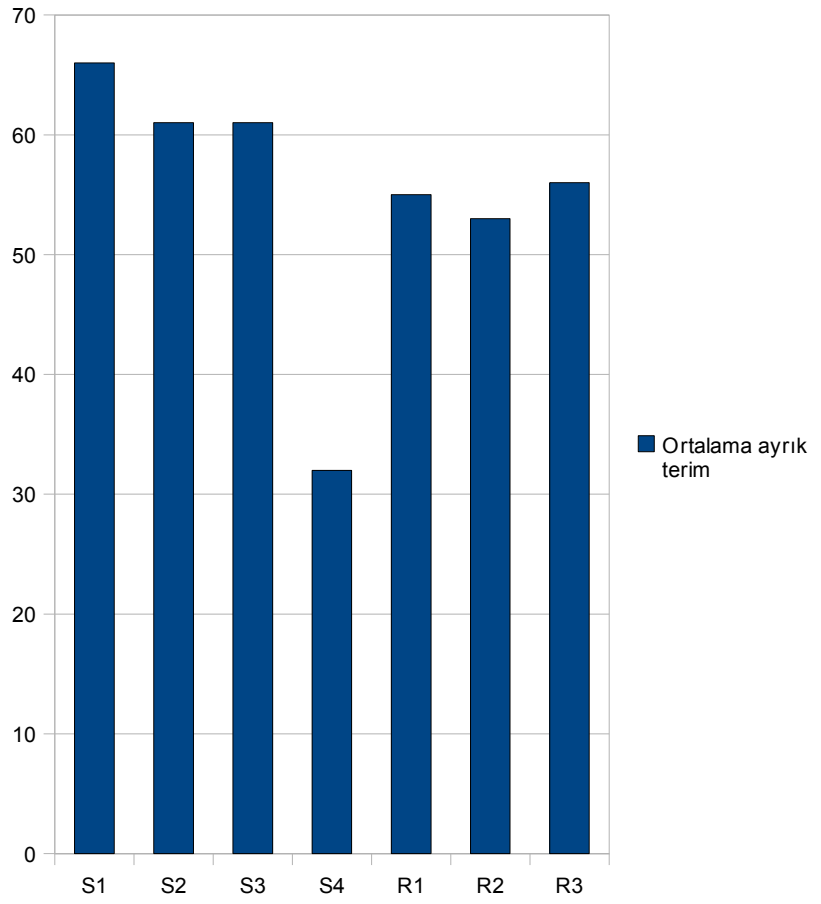




Etiket Sayısının Belirlenişi-2

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



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.

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.

Örnek

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

Örnek-Devam

İç çarpım = (0.64624 • 0.59810) + (0.42832 • 0.59810) + (0.27024 • 0.41629) + (0.27024 • 0.26265) = 0.896

lHaber–1l = (0.64624² + 2 • 0.42832² + 0.39624² + 9 • 0.27024²)½ = 1.264

lHaber–2l = (2 • 0.59810² + 2 • 0.41629² + 3 • 0.36673² + 7 • 0.26265²)½ = 1.395

Sim(Haber–1, Haber–2) = iç çarpım / lHaber–1l • lHaber–2l

= 0.896 / [1.264 • 1.395]

= 0.5

Sonuçlar

Haber: Doğuş Didim'de marina açıyor

Doğuş Grubu, 2003'te D-Marin Turgutreis İle başladığı yat limanı işletmeciliğine, Didim'de kuracağı ve tamamlandığında Türkiye'nin üçüncü büyük yat limanı kapasitesine sahip olacak yatırımıyla devam ediyor.

Sonuçlar - Benzerleri

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)

Kaynaklar

Grobernik M., Mladenic D., “Text-mining Tutorial”, J. Stefan Institute, Slovenia

Berry M. W., Drmac Z. and Jessup E. R., ”Matrices, Vector Spaces, and InformationRetrieval”, SIAM Review, 1999

ARROWSMITH http://kiwi.uchicago.edu/webwork/PURPOSE.html

Mizrahi A.R., Weisenstern A.M, “Survey System”, 2003

Zhao Z., Liu H., “Searching for Interacting Features”, Department of Computer Science and Engineering Arizona State University

Yar Even- Zohar, “Introduction to Text Mining”, Automated Learning Group National Center for Supercomputing Applications University of |||inois

Kaynaklar-devam

Güven A., “Türkçe Belgelerin Anlam Tabanlı Yöntemlerle Madenciliği ”, FBE, Yıldız Teknik Üniversitesi, 2007

Garcia Dr. Edel, “Term Vector Calculations A Fast Track Tutorial”, 2005

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

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

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

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