Otomatik Haber Metinleri Sınıflandırma
Automatic Text Categorization of News Articles

M.Fatih Amasyalı, Tülay Yıldırım***

*YTÜ Bilgisayar Mühendisligi Bölümü , mfatih@ce.yildiz.edu.tr
**YTÜ Elektronik Mühendisligi Bölümü ,tulay@yildiz.edu.tr

Özetçe

Bilginin kategorilendirilmiş olması bilgiye erişim zamanını azaltır. Günümüzde Internet en büyük bilgi kaynaklarından biridir. Ancak Internet'teki bilginin büyük bir bölümü doğal dille hazırlanmıştır. Internet'in daha verimli kullanılabilmesi için bu doğal dille yazılmış metinlerin kategorilendirilmesi gerekmektedir. Internet'teki bilgi miktarının büyüklüğü ve artış hızı bu işlemin elle yapılabilmesini neredeyse imkansız hale getirmektedir. Bu noktada otomatik metin sınıflandırma sistremlerine ihtiyaç duyulmaktadır. Diğer dillerin aksine bu konuda Türkçe üzerinde çok az çalışma mevcuttur. Bu çalışmada gazetelerin web sayfalarındaki haber metinleri otomatik olarak sınıflandırılmaya çalışılmıştır. Metinler 5 haber sınıfına ayrılmaya çalışılmış ve %76 oranında başarı elde edilmiştir.

Abstract

To categorize the data reduces the access time. Nowadays, Internet is one of the biggest data resources. However, most of the data in Internet is written by natural language. To use the Internet more efficiently, it needs to be categorized. The amount of data and increment rate is so high that this process can not be done by hand. Hence, the necessity of automatic text categorization systems is increasing. On the contrary of other languages, there is not much study on Turkish texts. In this study, a system is developed for Automatic Text Categorization of News Articles. The articles are classified into 5 different classes and 76% success ratio is achieved.

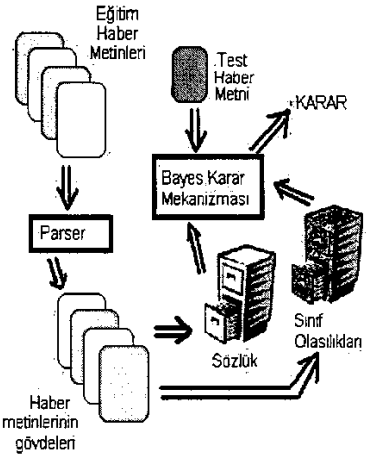
1. Giriş

İçinde arama yapılacak bilgi miktarı ne kadar büyük olursa olsun eğer bilgiler iyi sınıflandırılmışsa istenilen bilgiye erişim çok fazla zaman almamaktadır. Bunun en iyi örneği olarak ilişkisel veri tabanları gösterilebilir. Günümüzde Internet bilginin en önemli kaynaklarından biri haline gelmiştir. Ancak içindeki bilgilere erişmek her zaman o kadar kolay olamamaktadır. Bunun en önemli sebebi Internet'teki bilgilerin iyi kategorilendirilmemiş olmalarıdır. Bunun sebebi ise Internet'teki bilginin insanların günlük doğal dillerini kullanarak oluşturulmuş olmasıdır. Ancak bu büyük bilgi hazinesinden öyle kolayca vazgeçmek mümkün değildir. Internet'i daha verimli kullanabilmek için bu doğal dille yazılmış metinlerin (web sayfalarının) kategorilendirilmesi gerekmektedir. Internet'in ilk yıllarında bu kategorilendirme işlemi arama motorlarındaki uzman toplulkları tarafından elle gerçekleştirilmekteydi. Ancak Internet'teki bliginin bugün ulaştığı boyut ve daha belkide daha önemlisi bilginin artış miktarı bu kategorilendirme(sınıflandırma) işleminin elle yapılmasını imkansız hale getirmiştir. İşte bu problemi çözebilmek için otomatik metin sınıflandırma sistemleri ortaya çıkmaya başlamıştır. Bir metnin yazarını yazının çeşitli özelliklerinden bulabilen sistemler[1Bizim çalışaya referans], Posta kutunuza gelen bir mailin sizin tarafınızdan istenip istenmediğini anlayan sistemler (spam mail belirleyiciler), web sayfalarını belirli kategorilere otomatik olarak atan Google Directory bu tür sistemlere örnek olarak gösterilebilir.

Diğer dillerin aksine Türkçe için bu konuda yapılmış çok az sayıda çalışma bulunmaktadır. Bu nedenle bu konudaki eksikliği bir parça azaltmak için bu çalışmada Türkçe gazetelerin web sayfalarındaki haber metinlerinin otomatik olarak sınıflandırlması gerçekleştirilmiştir.

2. Önerilen Sistem

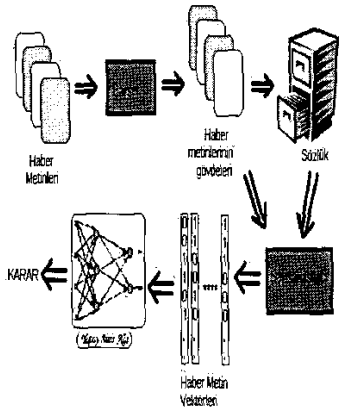
Bu çalışmada Türkçe haber metinlerinin otomatik olarak sınıflandırılması için bir metod geliştirilmiştir. Haber metinlerinin içindeki kelimelerin kendileri değil gövdeleri kullanılmıştır. Haber metinleri temel olarak iki sınıflandırma türüyle sınıflara ayrılmıştır. Bunlardan ilki olan Naive Bayes yöntemi bu alanada en çok başvurulan metodlardan birisidir. Bu metotta eğitimde kullanılan haber metinlerdeki kelimelerin gövdelerinden bir sözlük oluşturulmuş ve sınıf olasılıkları bulunmuştur. Bu metodun genel akış diyagramı Şekil 1'de gösterilmiştir.



Şekil 1: Naive Bayes'le Metin Sınıflandırma.

Kullanılan diğer temel sınıflandırma metodlarında ise metinlerin sayısallaştırılması gerekmektedir. Metinlerin sayısallaştırılma süreci Şekil 2'de gösterilmiştir.

0-7803-8318-4104/$20.00©2004 IEEE



Şekil 2: Metinleri sayısallaştırılmaları(haber vektörleri) ve Yapay Sinir ağıyla sınıflandırılmaları

2.1. Kullanılan Parser ve Gerekliliği

Sistemde Metinlerdeki kelimelerin kendileri yerine gövdelerinin kullanıldığı daha önceden belirtilmişti. Bunun sebebi Türkçe gibi eklemeli dillerde bir gövdenin sonuna birçok farklı ek alarak farklı biçimlerde karşımıza çıkabilmesidir. Örneğin "ababa" kelimesi ile "arabadan", "arabayı", "arabada", ve "arabanın" kelmeleri eğer parser olmasa ayrı ayrı kelimeler olarak görülceeklerdi. Bunun sonucu olarak hem oluşturulan sözlük boyutu çok artacak hemde sınıflandırma başarısı düşecekti. Bu sebeplerden kelimeler bir parser yardımıyla gövde ve eklerine ayrılmış, metinlerde sadece kelimelerin gövdeleri bırakılmıştır. Bu çalışmada yeni bir parser oluşturmak yerine önceden geliştirilmiş bir parser kullanılmıştır[1].

2.2. Karar mekanizmasında kullanılan yöntemler

Önceden bahsedildiği gibi iki temel sınıflandırma metodu kullanılmıştır. İlki Naive Bayes, diğerleri ise MLP ve LVQ olmak üzere Yapay Sinir Ağlarının iki türüdür.

2.2.1. Learning Vector Quantization(LVQ)

Kohonen tarafından önerilen bir vektör kuantalama metodudur[3]. Eğiticili ve yarışmalı bir öğrenme türüdür. LVQ algoritmasında kuantalanmak istenen bilgiyle aynı boyuttaki betimleyici vektör rasgele olarak seçilir. Eğitim setinin her bir örneği için, bu örneğe en yakın olanı belirlenir ve eğer örnekle betimleyici vektör aynı sınıftan ise betimleyici vektör o sınıfı daha iyi temsil etmesi için eğitim örneğine yaklaştırılır. Eğer farklı sınıflandansalar uzaklaştırılır. Diğer bir ifadeyle her adımda betimleyici vektörlerden biri kazanır ve eğer doğru sınıflandırma yapılmışsa ödüllendirilir, yanlış sınıflandırma yapılmışsa cezalandırılır. Metodun algoritması aşağıda verilmiştir.

[η] öğrenme oranı

[δ] 2.öğrenme oranı

[n] maximum eğitim sayısı

[c] betimleyici vector sayısı

[$\mu_1,\ldots,\mu_c$] betimleyici vektörler (centroids)

[x] eğitim datasından bir örnek

[S(x)] x vektörünün ait oldugu yada betimledigi sınıf olmak üzere

1. $\eta,\delta,n,\mu_1,\ldots,\mu_c$ için ilk değer atamalarını gerçekleştir

2. Eğitim adımları

2.1 X eğitim datasını al

2.2 X e en yakın betimleyici vektörü bul

$(\mu_k): k \leftarrow \arg\min_j \|x-\mu_j\|$ j=1..c

2.3 μ_k nın güncellenmesi:

Eğer x doğru sınıfsa (s(x)=s(μ_k) sınıfları aynı ise)

$\mu_k \leftarrow \mu_k + \eta(x-\mu_k)$ ödüllendir x'e yaklaştır

değilse

$\mu_k \leftarrow \mu_k - \eta(x-\mu_k)$ cezalandır x'den uzaklaştır

2.2.2. Naive Bayes

Naive Bayes Kolay uygulanabilir olduğu kadar üstün performansıylada metin sınıflandırma çalışmalarında en çok kullanılan metodlardan biri haline gelmiştir[2]. Metodda önce tüm eğitim verisindeki metinlerde kullanılan kelimelerden bir sözlük oluşturulur. Daha sonra her bir kelimenin her bir sınıftaki tekrar sayıları(frekansı) bulunur. Sınıflandırılması isetenen yeni bir metin önceden geldiğinde oluşturulan sözlükte var olan kelimelerin herbir sınıftaki frekansları bulunur. Bir metnin C sınıfına dahil olma olasılığı C sınıfının eğitim setindeki oranıyla, metnin içindeki her bir kelimenin C sınıfına ait olma olasılıkları çarpılarak bulunur.

2.3. Vektörel metinlerde boyut azaltma çalışmaları

Sayısallaştırma işlemi sonucunda elde edilen vektörler 2846 boyutludur. Birçok metin sınıflandırma çalışmasında bu problemle karşılaşılmaktadır. Bu boyuttaki verilerle eğitim işlemide test işlemide zaman alıcı işlemlerdir. Bu nedenle verinin daha az boyutta ifade edilmesi gerekmektedir. Bu çalışmada yine metin sınıflandırma işlemlerinde en çok kullanılan boyut azaltma metodlarından olan Information Gain Ölçümleri ve PCA kullanılmıştır.

2.3.1. Principle Component Anaysisl(PCA)

Verilerin birlikte değişimlerini en az miktarda kaybettikleri boyutları seçerek verilerin o boyutlar üzerindeki izdüşümlerini bulan bir metoddur. Bu çalışmada 2846 boyuttan 50'den az sayıdaki boyuta indirgeme yapılmıştır.

2.3.2. Informaiton Gain(IG)

Eğitim setindeki verilerin özelliklerinden hangilerinin daha belirleyici olduğunun bulunmasında kullanılan bir ölçüttür[4]. Örneğin Tablo 1'deki gibi bir eğitim seti için A özelliği, B özelliğine göre daha ayırt edici bir özelliktir.

Tablo 1: Eğitim Seti

A	B	Sınıf
X	K	S1
X	M	S1
Y	K	S2
Y	M	S2

Yukarıdaki eğitim setine göre A özelliği X olan örneklerin S1 sınıfından, Y olan örneklerin ise S2 sınıfından oldukları söylenebilir. Ancak B özelliği için böyle bir genelleme mümkün değildir. S eğitim seti içindeki A özelliğinin Information Gain'i Denklem 1'deki şekilde bulunmaktadır.

$$Gain(S, A) \equiv Entrophy(S) - \sum_{v \in Value(A)} \frac{|S_v|}{|S|} Entrophy(S_v) \quad (1)$$

3. Uygulama Sonuçları

Sistemin eğitilmesi için Türkçe gazetelerin web sitelerindeki 5 farklı konudan 10'ar adet haber metni alınmıştır. Test işlemi içinse her bir sınıftan eğitim setinde olmayan 5'er adet makale alınmıştır.

Naive Bayes'le 25 test haberinin 19 tanesi doğru bir şekilde sınıflandırılmıştır. En yüksek sınıflandırma oranları %100'lük başarı ile siyasi ve sağlık içerikli haberlerdir.

Tablo 2'de Uygulanan sınıflandırma metodları ve boyut azaltma metodlarıyla elde edilen sınıflandırma başarıları gösterilmiştir. Sonuçlar LVQ ve MLP için en yüksek performansı elde ettikleri yapılar için verilmiştir.

Tablo 2: Sınıflandırma Sonuçları

*25 test datasında	Ekonomi	Magazin	Sağlık	Siyasi	Spor	Toplam*
Naive Bayes	4	2	5	5	3	19
LVQ 2846 boyut	5	0	5	5	4	19
LVQ 50 boyutlu InfoGain ile	4	5	5	0	5	19
LVQ 50 boyutlu PCA ile	1	2	2	2	0	7
MLP 2846 boyut	3	0	4	2	2	11
MLP 50 boyutlu InfoGain ile	3	2	2	0	1	8
MLP 50 boyut PCA ile	1	1	2	2	2	8

4. Sonuç

Türkçe haber metinlerinin otomatik sınıflandırılması için yapılan bu çalışmada kelimelerin gövdeleri kullanılarak metinlerin özellikleri ortaya çıkarılmıştır. İki temel tür sınıflandırma metodu kullanılnış olup en yüksek performansa LVQ ve Naive Bayes metodlarının eriştiği görülmüştür. Boyut indirgeme çalışmalarının ise performans arttırımına bir katkı sağlamadığı ancak işlem zamanını 50'de 1'e düşürdüğü gözlemlenmiştir. Performansın daha iyileştirilmesi için metinlerin daha başka özelliklerinin kullanılması[4], farklı sınıflandırma metodlarının kullanılması düşünülebilir.

5. Kaynakça

[1] İşler S., Amasyalı M. F., Tatlı E., (2001), "Bir Türkçe Metindeki Kelimelerin Cümle İçindeki Durumlarına Bakılarak Eklerine Ayrılması ", Bilişim Zirvesi'01, İstanbul

[2] http://www.resample.com/xlminer/help/NaiveBC/classiNB_intro.htm

[3] T.Kohonen," Self-Organization and associative Memory",3d ed, 1989, Berlin :Springer-Verlag.

[4] http://acm.armstrong.edu/classes/Y2002/cs5820_fall/notes3.php

[5] Diri B. ve Amasyalı M. F., (2003), "Automatic Authorship Attribution in Turkish Language", ICANN/ICONIP 2003, İstanbul