Bilgi Dünyası 2008, 9(1):158-178

Yapay Sinir Ağları ile Web İçeriklerini Sınıflandırma*

Web Content Classification Using Artificial Neural Networks

Esra Nergis GÜVEN,**
Hakan ONUR *** **ve Şeref SAĞIROĞLU**** ***

Öz

Internet'in hızlı gelişmesi ve yaygınlaşması elektronik ortamda iş ve işlemleri hızlandırmış ve kolaylaştırmıştır. Elektronik ortamlarda depolanan, taşınan ve işlenen bilgilerin boyutunun her geçen gün artması ise bilgiye erişim ile ilgili birçok problemi de beraberinde getirmiştir. Kullanıcıların elektronik ortamda sunulan bilgilere erişmelerindeki hız ve doğruluk gereksinimi nedeniyle, bu ortamlarda tutulan bilgileri sınıflandırma ve kategorilere ayırma yaklaşımlarına ihtiyaç duyulmaktadır. Sayıları milyonun üzerinde olan arama motorlarının, kullanıcıların doğru bilgilere kısa sürede ulaşmasını sağlaması için her geçen gün yeni yaklaşımlar ile desteklenmesi gerekmektedir. Bu çalışmada, web sayfalarının belirlenen konulara göre sınıflandırılabilmesi için, Çok Katmanlı (MLP) yapay sinir ağı modeli kullanılmıştır. Özellik vektörü içeriğinin seçimi, yapay sinir ağının eğitilmesi ve son olarak web sayfalarının doğru kategorize edilmesi için bir yazılım geliştirilmiştir. Bu zeki yaklaşımın, elektronik ortamlarda bilgilerin

---

* *Bu makale "Değişen Dünyada Bilgi Yönetimi Sempozyumu, 24-26 Ekim 2007, Ankara."da bildiri olarak sunulmuştur.*
** *Gazi Üniversitesi, Bilgisayar Mühendisliği Bölümü, 06570, Maltepe, Ankara. (eng@gazi.edu.tr)*
*** *Gazi Üniversitesi, Bilgisayar Mühendisliği Bölümü, 06570, Maltepe, Ankara. (hakano@adasoft.com.tr)*
**** *Gazi Üniversitesi, Bilgisayar Mühendisliği Bölümü, 06570, Maltepe, Ankara. (ss@gazi.edu.tr)*

kolaylıkla ve yüksek doğrulukla sınıflandırılması, web ortamlarında doğru içeriğe ulaşılması ve birçok güvenlik açığının giderilmesine katkılar sağlayacağı değerlendirilmektedir.

Anahtar sözcükler: *Yapay sinir ağları, Metin gruplama, İçerik sınıflandırma, Web sayfası kategorizasyonu, Bilgi yönetimi.*

Abstract

Recent developments and widespread usage of the Internet have made business and processes to be completed faster and easily in electronic media. The increasing size of the stored, transferred and processed data brings many problems that affect access to information on the Web. Because of users' need get to access to the information in electronic environment quickly, correctly and appropriately, different methods of classification and categorization of data are strictly needed. Millions of search engines should be supported with new approaches every day in order for users to get access to relevant information quickly. In this study, Multilayered Perceptrons (MLP) artificial neural network model is used to classify the web sites according to the specified subjects. A software is developed to select the feature vector, to train the neural network and finally to categorize the web sites correctly. It is considered that this intelligent approach will provide more accurate and secure platform to the Internet users for classifying web contents precisely.

Keywords: *Artificial neural networks, Text categorization, Content classification, Web page categorization, Information management.*

Giriş

Bilgi toplumlarının temel hammaddesi bilgidir. Bilgisayar ve iletişim teknolojileri geliştikçe bilginin üretilmesi, taşınması ve depolanması kolaylaşmıştır. Elektronik ortamların gün geçtikçe yaygınlaşması ve kullanımının artmasıyla birlikte bilgi miktarında da hızlı bir artış gözlenmektedir (Miniwatts, 2006). Bu ortamlarda tutulan bilginin sınıflandırılması ise bilgi denizinde doğru bilgiye hızla erişimi kolaylaştıracak yaklaşımdır.

Bilgi erişim sistemleri temelde kullanıcıların bilgi ihtiyaçlarını karşılaması muhtemel olan ilgili belgelerin tümüne erişir ve ilgili olmayanları da ayıklar. Internet ortamına baktığımızda bu sistemler arama motorları olarak karşımıza çıkar. Arama motorları bilgiye erişim

anahtarı ve yol haritalarıdır. Bilgiye erişmek istediğimizde arama motorlarından oldukça sık faydalanırız. Ancak elektronik ortamlarda taşınan ve depolanan bilgilerin boyutları çok yüksek olup gün geçtikçe hızla artmaya devam etmektedir. Google'ın (2007) Internet'teki en kullanışlı sitelerin en geniş koleksiyonunu sunan ve bir milyondan fazla URL'yi içeren indeksi web ortamlarını daha popüler hale getirmiştir. Bu ortamlarda bilgilerin doğru sınıflandırılması bir zorunluluk haline gelmiştir. Günümüzde bu başarılmış gibi görünse de istediğimiz veya aradığımız bilgiye ulaşmak aslında o kadar da kolay değildir. Hızlı bir şekilde doğru bilgiye erişim için, arama motorlarının kullanımını ve püf noktalarını da iyi öğrenmek gerekir ki bazı durumlarda bu bile yetersiz kalabilir. Arama motorlarını doğru kullanmanın zorlukları karşısında bir adım daha ileri giderek, belirli bir kategori belirtebilmek aramanın daha net sonuçlanmasını sağlayabilecektir. Örneğin, kullanıcının "araba" ve "motor" kelimelerini taradığında erişilen belgeleri bir de kategorilerine göre süzebilmesi, ekonomi grubunda yer alan "araba" ve "motor" kelimeleri geçen belgelere de erişmesini sağlayabilir. Bu nedenle elektronik ortamlarda doğru bilgilerin araştırılması veya doğru bilgilere erişilmesi için her zaman yeni yaklaşımlara ihtiyaç duyulacaktır. Geliştirilecek sınıflandırma yaklaşımlarının hızlı olması ve doğru bilgiye erişim imkânı sağlaması gerekmektedir (Witten, Moffat ve Bell, 1999).

Wikipedia'da doküman sınıflandırma/kategorizasyon problemi, "bir elektronik dokümanın içeriğinin bir veya daha çok kategoriye ayrılması işlemi" olarak ifade edilmektedir. Doküman sınıflandırma işlemi danışmanlı ve danışmansız olmak üzere iki şekilde yapılmaktadır. Bu sınıflandırmada karar ağaçları (Moulinier ve Ganascia, 1996), kural öğrenme (Apte, Dameran ve Weiss, 1994), sinir ağları (Ng, Goh ve Low, 1997; Wiener, Pedersen ve Wiegend, 1995), lineer sınıflandırıcılar (Lewis, Schapire Callan ve Papka, 1996), en yakın komşuluk algoritmaları (kNN) (Yang ve Pedersen, 1997), destek vektör makinaları (Joachims, 1997), tf*idf değerleri (terim sıklığı* devrik belge sıklığı), kavram madenciliği (content mining), gizli anlam analizi (LSA-Latent Semantic Analysis) ve Naive Bayes metodları (Lewis ve Ringuette, 1994; McCallum ve Nigam, 1998) gibi farklı yaklaşımlar kullanılmaktadır (bkz. Ruiz ve Srinivasan, 2002).

Metin kategorizasyonu yapan bir sistemin amacı, metni önceden tanımlanmış kategorizasyon şemasına göre ayrı etiketlere ya da kategorilere dâhil etmektir. Bu işaretlemeler, filtreleme veya düzeltme gibi amaçlarla kullanılabilir. Günümüzdeki hızlı bilgi artışında otomatik metin kategorizasyonu önemli bir hedeftir.

Web tarayıcısı kullanan sınıflandırma sistemlerinin çoğunda işlemler insan desteğiyle yapılmaktadır (Shanks ve Williams, 2001). Bu işlemleri elle yapmak iyi bir yaklaşım gibi görünse de doküman sorgularının milyonlara eriştiği bir ortamda bu sistemler pek de işe yaramamaktadır. Bu ortamlarda bu işlemleri yapacak yeni yaklaşımlara her zaman ihtiyaç duyulmaktadır.

Bu çalışmada birçok yöntem sunulmuş olmasına karşın; kategorilerin özellik alanlarının yüksek boyutlu olmaları karşılaşılan temel bir problemdir. Özellik alanını daraltmak veya iyi alt kümeler seçmek, etkin ve başarılı bir uygulama gerçekleştirmek için oldukça önemlidir. Bu anlamda, kategorizasyonu belirleyen özelliklerin seçimleri için birçok yaklaşım ve yöntem mevcuttur (Yu ve Liddy, 1999).

Yapay sinir ağları birçok alanda problem çözümlemeye başarıyla uygulanmış bir yapay zekâ metodudur (Haykin 1994). Problemlere hızlı ve zeki çözüm sağlamaları, az veriyle genelleme yapabilmeleri, öğrenebilmeleri ve giriş ve çıkış verileri mevcut sistemlere genel bir model oluşturabilmeleri, farklı problemlere kolaylıkla uyarlanabilmeleri gibi sebeplerden dolayı bu çalışmada web sayfası kategorizasyonu için yeni bir yaklaşım olarak sunulmuştur. Bu yaklaşımın çalışabilirliğini göstermek için ise WeSaKa isimli bir yazılım geliştirilmiş ve tanıtılmıştır. Bu kategorizasyonda “spor”, “ekonomi” ve “kültür” sınıfları ele alınmış ve belirlenen web sayfalarının hangi sınıfa en yakın olduğu otomatik olarak tespit edilmeye çalışılmıştır.

Literatürdeki mevcut çalışmalarda olduğu gibi bu çalışmada da dokümanlar bir özellik vektörüne dönüştürüldükten sonra yapay sinir ağları ile sınıflandırılmıştır.

Bu bildiride yapay sinir ağları tanıtılmış, sistemin yapısı verilmiş, uygulamada takip edilen adımlar sunulmuş ve geliştirilen arayüz tanıtılmıştır. Son bölümde ise sunulan çalışma farklı açılardan değerlendirilmiştir.

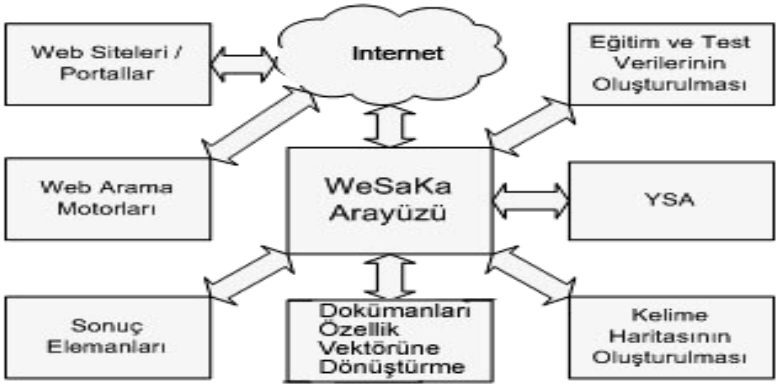
Yapay Sinir Ağları

Yapay sinir ağları (YSA), farklı zeki özellikleri bulundurmasından dolayı pek çok uygulamada kullanılmaktadır (Sağıroğlu, Beşdok ve Erler, 2003). YSA, bir sisteme ilişkin çeşitli parametrelere bağlı olarak tanımlanan girişler ve çıkışlar arasında ilişki kurabilme yeteneğine sahiptir. Bu ilişkinin doğrusal bir formda olması zorunlu değildir. Ayrıca YSA'lar, çıkış değerleri bilinmeyen tanımlanmış sistem girişlerine de uygun çıkışlar üretebilmekte, böylelikle çok karmaşık problemlere bile iyi çözüm olabilmektedirler (Sağıroğlu, Beşdok ve Erler, 2003).

Literatürde birçok YSA yapısı mevcuttur (Haykin 1994; Sağıroğlu, Beşdok ve Erler, 2003). Sunulan çalışmada Çok Katlı Perseptron (ÇKP) modeli kullanılmıştır. ÇKP, birçok alana uygulanmış olan bir YSA yapısıdır (Sağıroğlu, Beşdok ve Erler, 2003). Birçok öğrenme algoritmasının bu ağı eğitmede kullanılabilir olması, bu modelin yaygın kullanılmasının sebebi olarak açıklanabilir. Şekil 1'de de verildiği gibi bir ÇKP modeli, bir giriş, bir veya daha fazla ara ve bir de çıkış katmanından oluşur. Bir katmandaki bütün işlem elemanları bir üst katmandaki bütün işlem elemanlarına bağlıdır. Giriş katındaki nöronlar tampon gibi davranırlar ve giriş sinyalini ara kattaki nöronlara dağıtırlar. Ara kattaki her bir nöronun çıkışı, kendine gelen bütün giriş sinyallerini takip eden bağlantı ağırlıkları ile çarpımlarının toplanması ile elde edilir. Elde edilen bu toplam, çıkışın toplam bir fonksiyonu olarak hesaplanabilir.

Buradaki fonksiyon, basit bir eşik fonksiyonu, bir sigmoid veya hiperbolik tanjant fonksiyonu olabilir. Diğer katlardaki nöronların çıkışları da aynı şekilde hesaplanır. Kullanılan eğitme algoritmasına göre, ağın çıkışı ile arzu edilen çıkış arasındaki hata tekrar geriye doğru yayılarak hata minimuma düşünceye kadar YSA'nın ağırlıkları değiştirilir. Bu çalışmada ağın çıkışı ile arzu

edilen çıkışlar arasındaki hata tüm giriş seti için bulunduktan sonra ağırlıklar değiştirilmektedir.



Şekil 1: Geliştirilen WeSaKa Yazılımının Blok Şeması

Yapay sinir ağlarında kullanılan çok sayıda öğrenme algoritması bulunmaktadır. Bu çalışmada en fazla 10 epokta öğrenen ve hesaplamalarda çıkartabilen ve hesaplamalarda birçok hususu çözümleyebilen Levenberg-Marquardt (LM) öğrenme algoritması kullanılmıştır (Levenberg, 1944; Marquardt, 1963).

LM metodu, maksimum komşuluk fikri üzerine kurulmuş bir en az kareler hesaplama metodudur (Levenberg, 1944; Marquardt, 1963). Bu algoritma, Gauss-Newton ve En Dik Düşüş (Steepest Descent) algoritmalarının en iyi özelliklerinden oluşur ve bu iki metodun kısıtlamalarını ortadan kaldırır. Genel olarak bu metod yavaş yakınsama probleminden etkilenmez.

E($\underset{\sim}{w}$*)*'nin bir amaç hata fonksiyonu olduğu düşünülürse, *m* tane hata terimi için $e_i^2(\underset{\sim}{w})$ aşağıda verilmiştir.

$$E(\underset{\sim}{w}) = \sum_{i=1}^{m} e_i^2(\underset{\sim}{w}) = \left\| f(\underset{\sim}{w}) \right\|^2 \qquad (1)$$

bu eşitlikte *w* ağırlıkları ifade ederken,

$$e_i^2(\underset{\sim}{w}) \equiv (\underset{\sim}{y}_i - \underset{\sim}{yd}_i)^2 \text{ dir.}$$

Burada, amaç fonksiyonu *f(.)* ve onun Jakobiyeni *J*'nin bir noktada $\underset{\sim}{w}$ bilindiği farzedilir.

LM öğrenme algoritmalarında hedef, parametre vektörü $\underset{\sim}{w}$'nın, $E(\underset{\sim}{w})$ minimum iken bulunmasıdır. LM'nin kullanılmasıyla yeni vektor $\underset{\sim}{w}_{k+1}$, farzedilen vektör $\underset{\sim}{w}_k$'dan aşağıda verilen ifadeden hesaplanır.

$$\underset{\sim}{w}_{k+1} = \underset{\sim}{w}_k + \delta \underset{\sim}{w}_k \qquad (2)$$

burada $\delta \underset{\sim}{w}_k$ aşağıdaki şekilde verilir.

$$(J_k{}^T J_k + \lambda I)\delta \underset{\sim}{w}_k = -J_k{}^T f(\underset{\sim}{w}_k) \qquad (3)$$

Eşitlikte,

J_k: *f* 'in $\underset{\sim}{w}_k$ değerlendirilmiş Jakobyeni,

λ:Marquardt parametresi, ve

I: birim veya tanımlama matrisidir.

Levenberg-Marquardt algoritmasında hesaplama akışı aşağıdaki şekilde özetlenebilir.

(i) $E(\underset{\sim}{w}_k)$ 'yı hesapla,

(ii) küçük bir λ değeri ile başla (mesela λ = 0.01),

(iii) $\delta \underset{\sim}{w}_k$ için Eşitlik (3)'ü çöz ve $E(\underset{\sim}{w}_k + \delta \underset{\sim}{w}_k)$ değerini hesapla,

(iv) şayet $E(\underset{\sim}{w}_k + \delta \underset{\sim}{w}_k) \geq E(\underset{\sim}{w}_k)$ λ'yı 10 kat artır ve (iii)'e git,

(v) şayet $E(\underset{\sim}{w}_k + \delta\underset{\sim}{w}_k) < E(\underset{\sim}{w}_k)$ λ'yı 10 kat azalt, $\underset{\sim}{w}_k$: $\underset{\sim}{w}_k \leftarrow \underset{\sim}{w}_k + \delta\underset{\sim}{w}_k$ 'yi güncelleştir ve (iii)'e git.

Hedef çıkışı hesaplamak için bir YSA'nın ağırlıklarının LM öğrenme algoritması kullanılarak öğretilmesi ağırlık dizisi $\underset{\sim}{w}_0$'a bir başlangıç değerinin atanması ile başlar ve hataların karelerinin toplamı e_i^2'nin hesaplanmasıyla devam eder. Her e_i^2 terimi, hedef çıkış (*y*) ile gerçek çıkış (*yd*) arasındaki farkın karesini ifade eder. Bütün veri seti için e_i^2 hata terimlerinin tamamının elde edilmesiyle, ağırlık dizileri **(i)** den **(v)**'e kadar olan LM öğrenme algoritması adımların uygulanmasıyla daha önce de açıklandığı gibi adapte edilir.

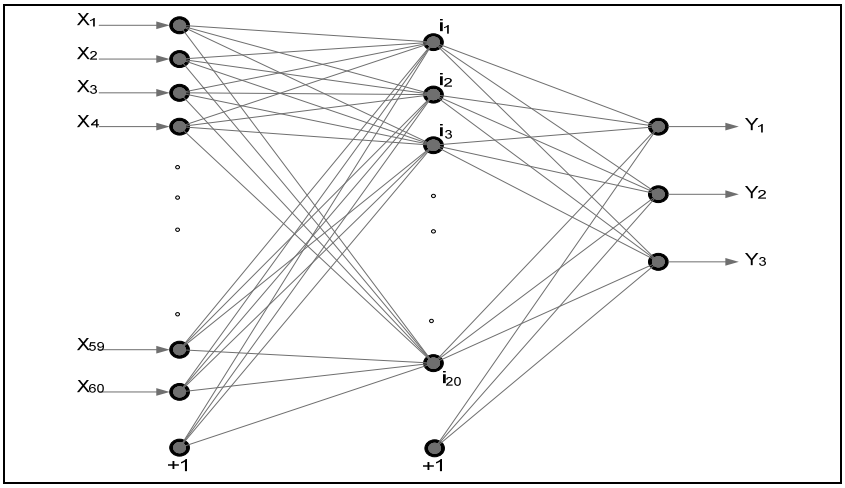
Geliştirilen Sistemin Yapısı ve Uygulanması

Bu çalışmada, doküman sınıflandırılmasının otomatik olarak ve kolaylıkla yapılabilmesi için WeSaKa adını verdiğimiz bir yazılım geliştirilmiştir. WeSaKa'nın geliştirilmesi için Microsoft Visual Basic.NET ortamı tercih edilmiştir.

Şekil 1'de geliştirilen yazılımın blok şeması verilmiştir. Blok şemadan da görülebileceği gibi bir arayüz ile web ortamına veya portallarına erişilebilmekte, kelime haritaları, eğitim ve test verileri oluşturulabilmekte, bu veriler özellik vektörüne dönüştürülebilmekte ve YSA sınıflandırıcı ile dokümanlar otomatik olarak sınıflandırılabilmektedir. Tüm bu işlemlerin sonucunda, WeSaKa aracılığı ile bir web sayfası açıldığında arka planda YSA'ya bir sorgu gönderilmekte ve sayfanın kategorisi hakkında bilgi alınıp kullanıcıya sunulmaktadır.

Blokta verilen YSA yapısının açık şekli Şekil 2'de verilmiştir. Şekil 2'de sunulan YSA yapısında, eğitim ve test kümelerinin kullanımına uygun olarak 60 giriş ve 3 çıkış bulunmaktadır. Giriş ve çıkış arasında 20 nöronlu bir gizli katman kullanılmıştır. Oluşturulan YSA yapısının gizli katmanında ve çıkış katmanında transfer (aktivasyon) fonksiyonu olarak sigmoid fonksiyon kullanılmıştır.

WeSaKa uygulaması ilk çalıştırıldığında, daha önceden hazırlanan kategorileri ve kelime haritasını XML formatında sisteme alır. Bu uygulama için alınan eğitim ve test kategori örnekleri temelde spor, ekonomi ve kültür olmak üzere üç grupta tanımlanmıştır. Tablo 1'de bu örnekler verilmiştir.



Şekil 2: YSA Yapısı

Kelime haritası, YSA'nın girdilerini oluşturan nöronların ifadeleridir. Kelime haritasında yer alan her kelime YSA'ya bir girdi olarak sunulur. Girdilerin değerleri ise ilgili kelimenin metinde kaç kez geçtiğinin toplam kelime sayısına oranıdır. Bu durum aşağıdaki gibi formüle edilebilir:

$$x_i = \frac{T_i}{T} \qquad (4)$$

Burada;

T_i : i kelimesinin tekrarlanma sayısı, ve

T : tüm kelimelerin toplam tekrarlanma sayısıdır.

İlgili alt yapı hazırlıklarını YSA modülü ile kurulan bir bağlantı tamamlar.

Yeni bir web sayfasına girildiğinde veya bir sayfa açılışında önişleme modülüne x_i ($i \in$ Kelime Haritası) matrisi aktarılır ve kelime sayısı uzunluğunda bir matris sonuç elde edilir. Gelen bu sonuçlar ekranda kullanıcı tarafından da görülebilir.

Sistemin doğru çalışabilmesi için ilk olarak kelime haritası oluşturulması gerekmektedir. Kelime haritasının olabildiğince doğru oluşturulması oldukça kritiktir. Söz konusu oluşturma yöntemi için pek çok metod bulunmaktadır (Joachims, 1998).

Türkçe kelimelerin köklerinin bulunması, kelime haritası oluşturulmasında çok önem taşır. Her ne kadar WeSaKa Türkçe kelime kökü bulunması konusunda özel bir işlem yapmasa bile bu konunun önemi açıktır.

WeSaKa'da kelime haritası oluşturma işleminde uzman görüşüne ihtiyaç duyulmuştur. Bunun için önceden tanımlanan kategorilere uyan pek çok sayfa ziyaret edilmiş, her sayfada bulunan kelimeler ayrıştırılarak bir küme oluşturulmuştur. Bu küme oluşturulurken ilgili her kelimeye kaç kez rastlandığı ve bu kelimelerin hangi kategoriler altında bulunduğu bilgisi de tutulmuştur. Daha sonraki adımda küme 500 kelimeye yaklaşınca öncelikle kelimeler sıralanmış ve üç karakterden kısa olan kelimeler ("ile", "de", "da" vs.) kümeden çıkartılmıştır. Kalan kelimeler alfabetik sıraya göre dizilmiş ve birbirinin kökü olabilecek kelimeler korunup diğerleri kümeden çıkartılmış, korunan kelimelerin görülme sayısına çıkartılan kelimeler eklenmiştir. Kelime kökleri belirlenirken o kelimenin metin içerisinde olabileceği her duruma kök olabilecek bir kelime seçilmiştir. Örneğin "kültür sanat" kategorisinde sıkça geçen "müzik" kelimesi her zaman müzik kökü ile değil "müziğe", "müziği" gibi ek almış hallerde de bulunur. Bu nedenle müzik kelimesinin kökü kelime haritası için "müzi" olarak alınmıştır. Bu şekilde bir kabul yaparken belirlenen kökün başka bir kelimenin kökü veya tamamı olmamasına da dikkat ederek karışıklığa yola açması engellenmiştir. Son olarak küme görülme sıklıklarına göre sıralanmış her bir kategoride sık görülen ve kategoriyle birebir ilişkili kelimelerden 20'şer adet

alınmış, kelime haritası oluşturulmuştur. Bu kelime haritası Tablo 1'de görülebilir.

Kelime haritası oluşturulurken Naive Bayesian ve SVM (destek vektör makineleri) kullanılabilecek olmasına rağmen kelime kökü çıkartma algoritması eksikliğinden dolayı insana bağımlı yöntem tercih edilmiştir.

İşlemlerin kolay anlaşılması için yapılan çalışmalar farklı başlıklar altında aşağıda sunulmuştur.

Test ve Eğitim Kümesinin Oluşturulması

Eğitim kümeleri oluşturulurken girilen web sayfasında kelime haritasındaki her kelimenin veya kelime kökü ile başlayan kelimelerin sayısı her kelime haritası maddesi için ayrı ayrı belirlenmiştir.

Bu belirleme işleminden sonra toplam kelime sayısı hesaplanıp, bulunan değerler kelime sayısına bölünmüş ve sonuçlardan 1x60'lık bir matris oluşturulmuştur. Bu matrise ilk üç sütun olarak da her bir kategoriye uyma oranları verilmiştir. Spor kategorisinde olduğu düşünülen bir belge için [1 0 0] matrisi eklenmiştir.

Test ve eğitim kümelerinin oluşturulması temelde aynı sistemle yapılmıştır. Tek fark, danışmanlı öğrenme tekniğine uygun olarak eğitim kümesinde olması gereken kategori sonucu sisteme verilmiş olup, test kümesinde bu veri sağlanmamış, aksine elde edilen sonuçlar YSA'dan sorgulanmıştır. Çalışma sonucunda oluşturulan eğitim kümesi girişleri (kelime haritası) Tablo 1'de gösterilmiştir.

Tablo 1: Kelime Haritası

Kategori 1 (spor)		Kategori 2 (ekonomi)		Kategori 3 (kültür)	
KELİME	TEKRARLAMA	KELİME	TEKRARLAMA	KELİME	TEKRARLAMA
fark	24	devlet	43	aşk	15
fikstür	33	düşüş	10	bahar	5
forma	10	düzey	13	bale	9
futbol	34	eğilim	8	belgesel	5
galibiyet	8	ekonomi	143	film	17
gol	38	enflasyon	12	güzel	9
hakem	32	firma	8	kitabı	11
kart	23	fiyat	15	klasik	5
lider	7	fuar	2	konser	11
lig	119	ihale	13	koro	11
maç	57	ihraç	3	müzi	11
menajer	4	kalkınma	19	ödül	21
pozisyon	6	kamu	14	öykü	24
puan	63	petrol	13	resim	20
saha	16	taahhüt	5	sergi	12
spor	171	tutma	3	şarkı	5
stad	9	ürün	28	şiir	14
şampiyon	25	yatırım	129	tiyatro	21
teşvik	5	yükseliş	4	tür	74
transfer	11	yüzde	124	yaz	23

YSA'nın Eğitilmesi

Daha önceden belirlenen üç kategoriden 10'ar adet belge için toplamda 60 girişli 30 örnekten oluşan bir eğitim kümesi oluşturulmuştur. YSA sınıflandırıcının eğitiminde kullanılan verilere örnekler Tablo 2'de verilmiştir. Oluşturulan bu eğitim kümesi YSA'ya sunulmuş ve Levenberg-Marquardt öğrenme algoritması kullanılarak eğitim yapılmıştır. Eğitim sırasında nöronlar arasındaki

ağırlıklara ilk değer olarak [-1,+1] arasında rastgele değerler atanmıştır.

Yaklaşık 350 tekrardan sonra eğitim tamamlanmış ve toplam mutlak hata oranı $4.5e^{-10}$ civarına indirilmiştir. Eğitim süresi P4 2GHz, 512MB RAM'li bir sistemde 1 saat sürmüş olsa da eğitim sonuçları istenilen düzeye ulaşmıştır.

Tablo 2. Eğitim ve Test Verileri Vektörleri

Örnek	**Giriş verileri** (kelime haritasındaki kelimeler için elde edilen değerler)	**Çıkış verileri** (Kategori)
1	0,0,0,0,0,0,0,0,0.00711743772241993,0.00355871886120996,0,0,0,0,0.00355871886120996,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.00355871886120996,0,0,0,0,0,0,0,0,0,0,0.0249110320284698,0,0.00355871886120996,0.00711743772241993,0.0391459074733096	0,1,0
2	0,0,0,0,0,0,0,0,0.00338983050847458,0,0,0,0,0,0.00338983050847458,0.0101694915254237,0,0,0.0169491525423729,0,0,0.00338983050847458,0,0,0,0,0,0,0,0,0,0.00677966101694915,0,0.0203389830508475,0,0,0,0,0,0,0,0,0,0.00338983050847458,0.0101694915254237,0,0.00677966101694915,0,0,0,0,0,0,0,0,0.00338983050847458,0,0.00338983050847458,0,0.00338983050847458	1,0,0
...	…	…
…	…	…
30	0,0.00803212851405622,0,0,0.00401606425702811,0,0,0,0.00401606425702811,0,0.0120481927710843,0,0.0200803212851406,0,0.00803212851405622,0,0,0,0,0.00803212851405622,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.00401606425702811,0,0,0,0,0,0,0,0.00401606425702811,0,0,0,0,0,0,0,0,0,0.00401606425702811,0.00401606425702811,0,0.00803212851405622,0,0	0,0,1

YSA'nın Test Edilmesi

YSA'nın test edilmesi için Tablo 3'te verilen gazetelere ait web siteleri kullanılmıştır. Farklı kategorilere ait ölçülen değerler de bu tabloda verilmiştir.

Tablo 3. YSA Test Sonuçları

		Ölçülen Değerler		
Web Sitesi	Kategori	Spor	Ekonomi	Kültür
http://www.hurriyet.com.tr	Spor	1.0000	0.0000	0.0000
http://www.hurriyet.com.tr	Spor	0.9998	0.0000	0.1191
http://www.hurriyet.com.tr	Spor	1.0000	0.0000	0.0000
http://www.hurriyet.com.tr	Spor	1.0000	0.0000	0.0000
http://www.hurriyet.com.tr	Spor	1.0000	0.0000	0.0000
http://www.milliyet.com.tr	Spor	1.0000	0.0000	0.0000
http://www.milliyet.com.tr	Spor	1.0000	0.0000	0.0000
http://www.milliyet.com.tr	Spor	1.0000	0.0000	0.0000
http://www.milliyet.com.tr	Spor	0.9995	0.0032	0.0000
http://www.milliyet.com.tr	Spor	0.9754	0.0000	0.0026
http://www.zaman.com.tr	Spor	0.9982	0.0000	0.5260
http://www.zaman.com.tr	Spor	0.2371	0.0000	0.9742
http://www.zaman.com.tr	Spor	0.3084	0.0000	0.9936
http://www.zaman.com.tr	Spor	1.0000	0.0000	0.0000
http://www.zaman.com.tr	Spor	0.9984	0.0000	0.3168
http://www.ntvmsnbc.com	Spor	1.0000	0.0000	0.0120
http://www.ntvmsnbc.com	Spor	1.0000	0.0000	0.0000
http://www.ntvmsnbc.com	Spor	0.7847	0.0000	0.9980
http://www.ntvmsnbc.com	Spor	1.0000	0.0000	0.0000
http://www.ntvmsnbc.com	Spor	1.0000	0.0000	0.0000
http://www.hurriyet.com.tr	Ekonomi	0.0000	0.9781	0.0027
http://www.hurriyet.com.tr	Ekonomi	0.0000	1.0000	0.0000
http://www.hurriyet.com.tr	Ekonomi	0.0000	0.9986	0.0001
http://www.hurriyet.com.tr	Ekonomi	0.0000	0.9835	0.0018
http://www.hurriyet.com.tr	Ekonomi	0.0000	0.9995	0.0259
http://www.milliyet.com.tr	Ekonomi	0.0000	1.0000	0.0001
http://www.milliyet.com.tr	Ekonomi	0.0000	1.0000	0.0000
http://www.milliyet.com.tr	Ekonomi	0.0000	0.8289	0.8330
http://www.milliyet.com.tr	Ekonomi	0.0000	0.9845	0.0022
http://www.zaman.com.tr	Ekonomi	0.0000	0.8757	0.0447
http://www.zaman.com.tr	Ekonomi	0.0000	0.0383	0.9242
http://www.zaman.com.tr	Ekonomi	0.0000	1.0000	0.0000
http://www.zaman.com.tr	Ekonomi	0.0000	0.9752	0.0012
http://www.zaman.com.tr	Ekonomi	0.0000	1.0000	0.0023
http://www.ntvmsnbc.com	Ekonomi	0.0022	0.1447	0.0028
http://www.ntvmsnbc.com	Ekonomi	0.0001	0.9997	0.0000
http://www.ntvmsnbc.com	Ekonomi	0.0000	1.0000	0.0000
http://www.ntvmsnbc.com	Ekonomi	0.0000	1.0000	0.0000
http://www.ntvmsnbc.com	Ekonomi	0.0012	0.8720	0.0134
http://www.ntvmsnbc.com	Ekonomi	0.0095	0.9642	0.0000
http://www.hurriyet.com.tr	Kültür Sanat	0.0072	0.0000	0.9994

http://www.hurriyet.com.tr	Kültür Sanat	0.0000	0.0000	1.0000
http://www.hurriyet.com.tr	Kültür Sanat	0.0000	0.0000	1.0000
http://www.hurriyet.com.tr	Kültür Sanat	0.0000	0.0000	1.0000
http://www.hurriyet.com.tr	Kültür Sanat	0.0000	0.0000	1.0000
http://www.milliyet.com.tr	Kültür Sanat	0.0000	0.0000	1.0000
http://www.milliyet.com.tr	Kültür Sanat	0.0001	0.0000	1.0000
http://www.milliyet.com.tr	Kültür Sanat	0.0000	0.0000	1.0000
http://www.milliyet.com.tr	Kültür Sanat	0.0206	0.0000	0.9941
http://www.zaman.com.tr	Kültür Sanat	0.0000	0.0000	1.0000
http://www.zaman.com.tr	Kültür Sanat	0.0000	0.0000	1.0000
http://www.zaman.com.tr	Kültür Sanat	0.0000	0.0000	1.0000
http://www.zaman.com.tr	Kültür Sanat	0.0001	0.0000	1.0000
http://www.zaman.com.tr	Kültür Sanat	0.0000	0.0000	1.0000
http://www.ntvmsnbc.com	Kültür Sanat	0.0041	0.0000	1.0000
http://www.ntvmsnbc.com	Kültür Sanat	0.0000	0.0000	1.0000
http://www.ntvmsnbc.com	Kültür Sanat	0.0441	0.0000	0.9843
http://www.ntvmsnbc.com	Kültür Sanat	0.0000	0.0000	1.0000
http://www.ntvmsnbc.com	Kültür Sanat	0.0001	0.0000	0.9993
http://www.ntvmsnbc.com	Kültür Sanat	0.0000	0.0000	1.0000

Geliştirilen yaklaşımla, Hürriyet, Zaman, Radikal gazeteleri web siteleri ile NTVMSNBC web sitesindeki denemelerde %99'a varan oranlarda mutlak sonuç elde edilmiştir. Burada sonuçları verilmemiş olsa da bazı gazetelerin web sitelerinde yapılan testlerde başarı oranının %80'lere düştüğü görülmüştür.

Başarı oranlarındaki değişimin birçok nedeni vardır: Kelime haritası oluşturulurken yapılan tarama miktarı, test sonucu alınan sayfadaki metnin aslında içerik olarak farklı kategoriye ait olması, sayfanın bir bölümünde bulunan bilgilendirme niteliğindeki metinlerin yanıltıcı olması, eğitim kümesi oluşturulurken yapılan taramanın geniş kapsamlı olmaması vb. gibi. Ancak bu sebeplerin kolaylıkla ortadan kaldırılması ve başarı oranı düşük olan sayfalarda da bu oranın yükseltilmesi mümkündür.

Uygulama Arayüzü

Web sayfalarını kategorize etme işlemi için YSA'da kullanılacak olan eğitim ve test kümelerinin oluşturulması ve ilerleyen aşamalarda ise sayfaların hangi kategoride olduğunun gösterilmesi için gerçekleştirilen çalışmanın arayüzü Şekil 3'te gösterilmiştir.

Şekil 3. Uygulama Arayüzü

Eğitim ve test kümelerini oluşturmak amacıyla gezilecek olan web sayfalarının adresleri URL satırına yazılarak o sayfadaki kelimeler ayrıştırılmış ve üç karakterden uzun olan kelimeler "Parse to Cat" tuşu ile tutulmuştur. Aynı kategoriden birçok sayfa gezilerek bu kategoride en çok rastlanan kelimeler ekran çıktısında görülen "S" tuşuyla sıralatılarak ait olduğu kategori için tekrarlama sayısıyla birlikte kaydedilmiştir. Belirlenen üç kategori için farklı sayfalar gezilerek elde edilen özellik matrislerinden eğitim kümesi oluşturulması için "Save ANNT" ile kaydedilmiştir. Kelime haritasında belirlenen 60 kelimenin x_i değerlerini, 30 farklı sayfa örneği için içeren bu küme YSA'nın eğitilmesinde giriş olarak kullanılmıştır. Çıkış olarak ise ait olduğu kategoriye göre [1 0 0], [0 1 0] veya [0 0 1] kullanılmıştır. Daha önceden tanımlanmış olan YSA yapısına elde edilen giriş ve çıkışlar verilerek "Train" tuşu ile

YSA'nın eğitimi gerçekleştirilmiştir. Eğitim sonucunda elde edilen YSA modeli eğitim sonrası gezilecek olan web sayfalarının kategorize edilmesi sırasında test için kullanılmıştır.

Sonuç olarak Şekil 3'te görülen program arayüzü hem eğitim kümesinin oluşturulması ve YSA'nın eğitilmesi hem de kategorizasyon sonuçlarının görülmesinde büyük kolaylıklar sağlamıştır. Bu nedenle yeni eğitim kümesi oluşturulması ya da var olan eğitim kümesine yeni kelimeler eklenmesi mümkündür. Eğitim kümesinde değişiklik yapılması durumunda, YSA'nın tekrar eğitilerek yeni ağırlıkların kaydedilmesi gerekmektedir. Böylece bundan sonra ziyaret edilecek sayfalar yeni eğitim kümesine göre değerlendirilecektir.

Değerlendirme ve Sonuç

Bu çalışmada, web ortamlarında bulunan verilerin doğru bir şekilde otomatik olarak sınıflandırılabilmesi için YSA temelli zeki bir sınıflandırma yaklaşımı ve bu yaklaşımın kolaylıkla uygulanabilmesi için WeSaKa isimli bir yazılım başarı ile gerçekleştirilmiş ve uygulanmıştır.

Uygulama sırasında belirlenen bu kategorilerin sayısında ve tanımlamasında herhangi bir sınır yoktur. Ancak yeni kategoriler tanımlandıkça YSA yapısının çıktı sayısı değiştirilmeli ve eğitim yenilenmelidir.

Bu çalışmada uygulamanın başarılı olabilmesi için iyi bir şablon seçilmesi gerektiği, şablon içeriğinin fazla olmasının sistemin öğrenmesini zorlaştırdığı gibi gereğinden az olmasının da YSA sonuçlarındaki hata oranının artmasına yol açtığı tespit edilmiştir.

Literatürde vurgulandığı gibi konuların birbirine yakınlığı da sonuçlarda önemli rol oynamıştır. Spor konuları ekonomi ve kültür konularından uzak olduğu için oldukça yüksek başarı sağlanabilirken, sınıflandırmada ekonomi ve kültür haberleri genelde iç içe olduğundan hata oranının arttığı gözlemlenmiştir.

Farklı sayfa örneklerinde doğruluk oranını Tablo 3'te gördüğümüz WeSaKa'nın kategorilere göre ortalama sonuçları

Tablo 4'te verilmiştir. Elde edilen sonuçlardan görülmüştür ki günümüzde oldukça önemli bir yer alan hatta ihtiyaç haline gelen metin kategorizasyonu üzerine yapılan bu uygulamanın kısa zamanda gerçekleştirilebilmesi ve sonuçlarının başarıyla elde edilmesi, hem YSA'nın gücünü bir kez daha farklı bir uygulamada göstermekte hem de doküman kategorizasyonunda farklı ufuklar açılmasına büyük katkılar sağlayabilmektedir.

Tablo 4. Kategorilere Göre Skorlar

Kategoriler	Spor	Ekonomi	Kültür Sanat
Spor	0,915075	0,000160	0,197115
Ekonomi	0,000650	0,882145	0,092720
Kültür Sanat	0,003815	0,000000	0,998855

Burada eğitimin uzun süre alması bir dezavantaj gibi görünse de test işlemlerinde bu sürenin saniyeler mertebesinde olduğunu belirtmekte fayda vardır.

Kelimelerin köklerinin ve kelime haritalarının çıkartılmasının önemi sonuçlar incelendiğinde daha iyi anlaşılmaktadır. Gerçekleştirilen çalışmanın planlanan sonraki adımı kelime haritası alanında uygulamayı daha da geliştirmektir. Bu alanda yapılacak olan çalışma, sınıflandırılacak kategori sayısının artması durumunda eğitim ve test veri kümelerinin hazırlanmasında büyük hız sağlayacaktır.

Bu çalışmanın geliştirilmesinde karşılaşılan güçlükler farklı ortamları biraraya getiren bir arayüz geliştirilmesi, kelime haritası oluşturulurken dikkat edilmesi gereken hususların incelenmesi ve kategori için belirleyici olacak olan optimum kelimelerin bulunması, oluşturulan kelime haritası için eğitim setinin oluşturulması, bu eğitim setinin uygulanacağı YSA yapısının belirlenmesi ve en uygun yapıyı bulmak için denemeler yapılması olarak söylenebilir. Bu sırada karşılaşılan en büyük güçlük eğitim aşamasının uzun

zaman alması olmuştur. Fakat yapılan çalışmalardan sonra en iyi sonucu veren katman ve nöron sayıları, aktivasyon fonksiyonu ve öğrenme algoritması tercih edilmiştir.

Bu çalışma gelecekte daha büyük sayıda verilerle ve farklı web sitelerinde test edilecektir.

Kaynakça

Apte, C., Damerau, F. ve Weiss, S.M. (1994). Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems, 12*, 233–251.

Google. (2007). 20 Nisan 2007 tarihinde http://www.google.com.tr/intl/tr/why_use.html adresinden erişildi.

Haykin, S. (1994). *Neural networks: A comprehensive foundation.* New York: Macmillan College.

Joachims, T. (1997). *Text categorization with support vector machines: Learning with many relevant features* (Technical Report LS-8 Report: 23). Dortmund: University of Dortmund.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. C. N'edellec ve C. Rouveirol (Ed.), *Proceedings of the European Conference on Machine Learning* içinde (s. 137-142). Berlin: Springer.

Levenberg, K. (1944). A method for the solution of certain nonlinear problems in least squares. *Quarterly of Applied Mathematics*, *2*, 164-168.

Lewis, D. ve Ringuette, M. (1994). A comparison of two learning algorithms for text categorization. *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR'94)* içinde (s. 81-93). Las Vegas.

Lewis, D.D., Schapire, R.E., Callan, J.P. ve Papka, R. (1996). Training algorithms for linear text classifiers. *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* içinde (s. 298-306). New York: ACM.

Marquardt, D.W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, *11*, 431-441.

McCallum, A. ve Nigam, K. (1998). A comparison of event models for naive Bayes text classification. *Learning for Text Categorization: Papers from the 1998 Workshop* içinde (s. 41-48). San Francisco, CA: AAAI Press.

Miniwatts International Inc. Internet Usage Statistics: The Big Picture. (2006). 01 Aralık 2006 tarihinde http://www.internetworldstats.com/stats.htm adresinden erişildi.

Moulinier, I. ve Ganascia, J.G. (1996). Applying an existing machine learning algorithm to text categorization. S. Wermter, E. Riloff ve G. Scheler (Ed.), *Connectionist, statistical, and symbolic approaches to learning for natural language processing* içinde (s. 343-354). Heidelberg: Springer Verlag.

Ng, H.T., Goh, W.B. ve Low, K.L. (1997). Feature selection, perceptron learning, and a usability case study for text categorization. N.J. Belkin, A.D. Narasimhalu, P. Willett ve W. Hersh (Ed.), *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* içinde (s. 67-73). Philadelphia, PA: ACM.

Ruiz, M.E. ve Srinivasan, P. (2002). Hierarchical text categorization using neural networks. *Information Retrieval*, *5,* 87-118.

Sağıroğlu, Ş., Beşdok, E. ve Erler, M. (2003). *Mühendislikte yapay zekâ uygulamaları I: Yapay sinir ağları*. Kayseri: Ufuk Kitabevi.

Shanks, V. ve Williams, H.E. (2001). Fast categorisation of large document collections. *Proceedings: Eight Symposium on String Processing and Information Retrieval November 13-15, Laguna de San Rafael, Chile* içinde (s. 194-204). San Rafael, Chile: IEEE Computer Society.

Wiener, E.D., Pedersen, J.O. ve Weigend, A.S. (1995). A neural network approach to topic spotting. *Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval (SDAIR'95)* içinde (s. 317-332). Las Vegas.

Witten, I.H., Moffat, A. ve Bell, T.C. (1999). *Managing gigabytes: Compressing and indexing documents and images*. San Francisco, CA: Morgan Kaufmann.

Yang, Y. ve Pedersen, J.O. (1997). A comparative study on feature selection in text categorization. *Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97)* içinde (s. 412-420). San Francisco, CA: Morgan Kaufmann.

Yu, E.S. ve Liddy, E.D. (1999). Feature selection in text categorization using the Baldwin effect. *Proceedings of IJCNN '99 (International Joint Conference on Neural Networks)* içinde (s. 2924-2927). Washington, DC: IEEE Press.