ASYU 2010
Akıllı Sistemlerde Yenilikler ve Uygulamaları Sempozyumu, 21-24 Haziran 2010, Kayseri & Kapadokya, TÜRKİYE

Web Tabanlı Otomatik Özet Çıkarma Sistemi

M.Vecdi SAMİ[1], Banu DİRİ[2]
[1]Kartek Kart ve Bilişim Teknolojileri Ltd._ti., [2]Yıldız Teknik Üniv., Bilgisayar Müh.Böl.
[1]mehmet.sami@smartsoft-it.com, [2]banu@ce.yildiz.edu.tr

Özet

Hızla gelişen teknoloji dünyasında kullanıcıların talebi, her şeyin hızlı, otomatik ve kullanıcı dostu olması yönündedir. Bu çalışmada Türkçe web sayfalarının belirlenen parametrelere göre özetlenmesi gerçekleştirilmiştir. Özetlenecek metin içerisinde yer alan her cümleye bir puan verip, özetleme oranına göre en yüksek puanlı cümleler seçilerek metnin özeti yaklaşık %59'luk bir başarı ile çıkarılmıştır.

Abstract

In the technology world which grows rapidly users are expecting everything faster, more automatic and more user-friendly. In this study, summarization of Turkish web pages have been performed by using specific parameters. Text to be summarized is being analysed and scores are assigned for each sentence and selection of sentences is being performed by ordering high value scored sentences. This method of text summarization gives us a performance ratio of %59.

1. Giriş

İnternet kullanımının artması ile istenilen bilgiye erişmek, hatta erişilen bilginin öncelikli olarak özetinin ve gerekli görüldüğünde bilginin tamamının okunma isteği günümüzde önemli hale gelmiştir. Metin dosyalarının özetlenmesi konusunda ilk olarak 1959 yılında Luhn [1] çalışmıştır. Bu çalışmada, cümleler içerisindeki kelimelerin kullanım sıklıkları çıkarılmış ve en çok kullanım frekansına sahip kelimelerin yazı hakkında önemli görüşler verdiği öne sürülmüş ve bu kelimelerin geçtiği cümleler seçilerek özetleme yapılmıştır. Edmundson [2] 1969 yılında yaptığı çalışmada, kelime frekanslarına ek olarak ipucu veren ifadelerin (Sonuç olarak, Özetle, Bu makale gösteriyor ki,...), konu başlığını içeren kelimelerin ve cümlenin bulunduğu yerin özetleme işlemi sırasında yeni özellikler olarak kullanılabileceğini göstermiştir. Bu eski yaklaşımların temelindeki düşünceler günümüzde halen metin özetleme araştırmalarında kullanılmaktadır. Ko ve Seo [3] 2008 yılında cümle seçerek özetlemeye istatistiksel yaklaşımların üzerine kavramsal bilgiyi de ekleyerek yeni bir yaklaşım önermiştir. Türkçe üzerine cümle seçerek özetleme yapan ilk çalışmalardan biri Altan'nın [4] Türkçe ekonomi haber makalelerinin özetlerinin çıkarılmasıdır. Metindeki cümleler sahip oldukları özelliklere göre puanlar alıp, en yüksek puana sahip cümleler seçilerek özetleme gerçekleştirilmiştir. Pembe [5] arama motorlarında kullanılmak üzere metnin iç hiyerarşisini de inceleyen bir özetleme sistemi geliştirmiştir. Uzundere ve arkadaşları [6] Türkçe metinler için [4]'deki özetleme kriterlerine yenilerini ekleyerek bir özetleme aracı geliştirmişlerdir.

Bu çalışmada [4]'deki özetleme kriterleri referans alınmış, ancak text dokümanları yerine html (hyper text markup language) sayfalarının özeti çıkarılmıştır. Bu çalışmanın en kritik noktası farklı formatlarda karşımıza çıkan web sayfalarının kullanıcı yönlendirilerek sisteme tanıtılması ve bir sonraki özetleme anında işlemin hatasız olarak gerçekleşmesidir.

Çalışmanın ikinci bölümünde özetleme sistemi hakkında, üçüncü bölümde ise geliştirilen web tabanlı otomatik özet çıkarma sisteminden bahsedilmiştir. Dördüncü bölümde deneysel çalışmalar ve son bölümde de sonuç ve gelecek çalışmalar yer almaktadır.

2. Özetleme Sistemin Yapısı

Geliştirilen özetleme sistemini iki ayrı parça altında inceleyebiliriz. Herhangi bir web sayfasından istenilen bölümün özetlenebilmesi için öncelikle web ortamından özetlenecek olan bilgiye erişmek, ikincisi de metnin akıcı olarak özetinin çıkarılmasıdır.

2.1 Webden Veriye Erişmek

Bir web sayfasına html belgesi adı verilir ve kendine özgü bir yapısı vardır. Html belgesi html bildirimi, giriş tanımlama bölümü ve sayfa içeriği olmak üzere üç bölümden oluşur. Html belgesi içerisinde yapılmak istenilenleri belirten komutlara tag adı verilir ve bu tag' ler küçüktür (<) ve

büyüktür (>) işaretleri arasında yazılırlar. Açılan her tag'in kapatılması gerekir. Bitiş ile başlama tag'i arasındaki tek fark, bitiş tag'inin önüne (/) işaretinin konmasıdır (<birhtmltagi> tagin etki yapacağı içerik </birhtmltagi>).

Web sayfalarının html mimarisi birbirlerine göre farklılıklar gösterebilmektedir. Bundan dolayı erişmek istenilen web sayfasının esas içerik bölümüne erişebilmemiz için html tag'lerini programda tanımlamamız gerekmektedir. Çünkü herhangi bir html ayrıştırma yöntemiyle doğru içeriğe erişmeniz mümkün olmayacaktır. Web sayfalarının önemli bir kısmında reklamlar, önceki yazılar vb. içerikler yer almaktadır. Eğer biz bir web sayfasında belirli bir bölümden veri çekmek istiyorsak, o bölümün başlangıç ve bitiş tag'lerini belirleyip, sistemimizde tanımlamamız gerekmektedir. Bir diğer önemli konu da web sayfalarının html yapıları her ne kadar farklılık gösterse de başlangıç ve bitiş tag'leri benzerlik gösterebilmektedir. Bu durumda sistemimizde karışıklığa neden olmaktadır. Bunun için tekil (unique) tag'ler tanımlamamız önemlidir ve bu şekilde her web sayfasında çalışabilen bir sistem tasarlanması sağlanmış olacaktır.

Geliştirilen sistemde, html tag'lerinin tutulduğu bir dosya oluşturulmuştur. Özetinin çıkarılması istenen sayfanın html tag'leri bu dosyadaki tag'lerle karşılaştırılıp sayfanın okunabilirliliği denetlenmektedir. Eğer tag'ler dosya içerisinde yer alıyorsa verinin çekilme işlemi başarıyla gerçekleşecektir, aksi durumda ekrana "*Girmek istediğiniz sitenin html tanımlaması yapılmamıştır. Tanımlamak ister misiniz ?*" diye bir mesaj gelecektir. Verdiğimiz "*evet*" cevabı ile kullanıcıya ekrandan html tag'lerinin (başlangıç, bitiş tagleri) tanımlanmasına izin verilmektedir. Böylece tekrar aynı sayfaya erişildiğinde veri çekme işlemi başarıyla gerçekleştirilecektir. Bu tasarım ile kullanıcıya geniş imkanlar sunulmakta ve sadece üzerinde çalışılmış sitelerde değil neredeyse tüm sitelerde özetleme yapabilme imkanı tanınmaktadır.

2.2 Özetleme Süreci

Web'den verinin düzgün bir şekilde çekilmesinden sonra, metnin özetleme süreci başlamaktadır. Bu süreç, konunun belirlenmesi, yorumlama ve üretme olarak üç alt başlıkta ele alınabilir [7].

Konunun Belirlenmesi. Yazının özeti çıkarılırken konusunun belirlemesi önemlidir. Konuyu saptamak için doküman içerisindeki kelime (edat, zamir, bağlaç, sıfat ve hatta fiil hariç) frekanslarının hesaplanması, cümlenin bulunduğu yerin incelenmesi, ipucu veren ifadelerden yararlanılması gibi teknikler kullanılır. Bazen, yazının başlığı, yazının ilk cümlesi, "Özetle", "En önemlisi", "Sonuç olarak" gibi ipucu veren ifadeler de yazıyla ilgili önemli noktaları gösteren işaretler olabilir.

Yorumlama. Cümleler birbirleriyle kaynaştırılarak daha genel cümleler elde edilebilir. Örneğin, "*Ali sınıfa girdi, sırasına oturdu, sınavını bitirdikten sonra eve gitmek için okuldan ayrıldı.*"cümlesinin yorumlanmış hali "*Ali okula gitti*" olacaktır.

Üretme. En basitinden çok karmaşığına kadar çeşitli üretme metotları içeren, metnin özetlenmiş olan son çıktısıdır. Kullanılabilen metotlar:

1. Birinci aşamada seçilen cümlelerin özet çıktısına eklenmesidir (extraction).
2. En çok kullanılan anahtar kelimelerin veya yorumlanan düşüncelerin özet çıktısına eklenmesidir (topic list).
3. İki veya daha fazla cümlenin birbirine bağlanmasıdır (phrase concatenation).
4. Cümle üretecinin kaynaşan fikirleri veya birbiriyle ilgili düşünceleri giriş olarak alıp, yeni cümleler üretmesidir (sentence generation).

3. Web Tabanlı Otomatik Özet Çıkarma Sistemi

Web tabanlı otomatik metin özetleme sisteminde amaç, farklı html mimarilerine sahip web sayfalarına erişmek, içeriğini indirmek, istediğimiz özetleme oranına göre özetini çıkarmak, kullanıcıya özeti ve istendiği taktirde orijinal dokümanı sunmaktır.

3.1. Puanlamada Kullanılan Özellikler

Cümle seçerek özetlemenin yapılacağı bu sistemde, ilk olarak metin cümlelere ayrılır. Cümlelere ayırma ve *1,2,3,...,n* şeklinde indeks numarası verilmesi, kod içerisinde otomatik olarak gerçekleşmektedir. Alıntı olarak adlandırdığımız aç-kapa tırnak ("...") içerisinde yer alan cümleler bu çalışmada tek bir cümle olarak değerlendirilmiştir. Cümle ayırma işleminden sonra, her cümle önceden belirlenmiş 12 özelliğe göre incelenir ve puanlama yapılarak sayısal bir değer atanır. Puanlamada kullanılan özellikler sırasıyla [6]:

Başlık-Title: Cümlenin, başlıkta ve varsa alt başlıklarda geçen kelimeleri içerip içermediği incelenir.

Yüksek Frekans- High Frequency: Metin içerisinde yer alan her kelimenin (edat, bağlaç, zarf, zamir hariç) frekansı hesaplanır ve yüksek frekanstan düşük frekansa doğru sıralandığında ilk yüzde onda yer alan kelimeler dikkate alınır. Bu kelimelere sahip olan cümlelere bir puan değeri atanır. Kelimenin tipini belirlemede, kökünün çıkarılmasında açık kaynak kodlu *Zemberek* [8] programından yararlanılmıştır.

Yer-Location: Cümlenin metin içerisindeki yerine bakılır. Özellikle giriş veya sonuç paragrafında yer alan cümleler incelenir, çünkü bu bölümlerde yer alan cümleler özet için önem taşımaktadır.

Anahtar Kelimeler- Keywords: Bazen kullanıcıdan anahtar kelimelerin istenmesi daha uygun olabilir. Örneğin, güncel bir haberinin özeti çıkarılmak istenildiğinde kullanıcı dışarıdan *deprem, yangın* gibi kelimeleri girebilir ve sistem bu kelimelerin geçtiği cümlelere puan verir.

Özel İsimler- Uppercase: Haber metinlerinde yer alan özel isimler, haberin içeriği hakkında önemli bilgiler verebilmektedir. Özel isimlerin çıkarılmasında da *Zemberek* [9] programı kullanılmıştır.

Pozitif Kelimeler- Positive Words: Cümlelerin *özetle, sonuç olarak, sonuçta, neticede* gibi toparlayıcı kelimeleri içerip içermediği incelenir.

Negatif Kelimeler- Negative Words: *Çünkü, ancak, öyleyse* gibi kelimeleri içeren cümleler konu hakkında ayrıntılı bilgi veren cümlelerdir ve bu cümlelerin özet metinde yer alması gerekmez.

Sayılar- Numbers: İçerisinde rakam bulunduran cümleler haber metinlerinde önem taşıdığından cümlenin rakam içerip içermediği de kontrol edilmelidir.

Çift Tırnak İşareti- Quotation Mark: Bu işareti içeren cümleler alıntı cümleler olup, haber metin içerisinde önemli olabilirler.

Bitiş İşareti- Ending Mark: Bir cümlenin ünlem işareti veya soru işareti ile bitmesi diğer cümlelere göre daha önemli olduğunun bir işaretidir.

Ortalama Uzunluk- Average Length: Metinde yer alan cümlelerin ortalama uzunluğu hesaplanır. ±1 ortalama uzunluğa sahip cümleler önemlidir.

Gün Ay- Day Month: Tarih bilgisi içeren cümleler önem taşır.

3.2. Cümlelerin Seçimi

Metin içerisindeki her cümle mevcut 12 özelliğe göre incelenip, bir puan değeri verilir. Bir cümle tek bir özellikten puan alabileceği gibi birden fazla özellikten de puan alabilir. Bunun için tüm özelliklere bir ağırlık değeri atanmalıdır (Tablo 1). Çalışmamızda kullanılan ağırlık değerleri [6]'da olduğu gibi sezgisel olarak belirlenmiştir. Örneğin, *başlık* özelliği 20, *negatif kelimeler* -20 ve içerisinde sayısal bir değer bulundurma özelliğide 3 puan olsun. İlgili cümle, başlıkta geçen iki kelimeyi, negatif kelimeler listesinde yer alan kelimelerden bir tanesini ve tarih bilgisini içermiş olsun. Bu durumda cümlenin puanı da (20 * 2) + (-20 * 1) + (3 * 1) = 23 olarak hesaplanacaktır. Her cümlenin puanı hesaplandıktan sonra, kullanıcı tarafından seçilmiş olan özetleme yüzdesine göre, en yüksek puana sahip cümleler, orijinal metin içerisinde bulunma sırasına göre sırasıyla özet metinde yer alırlar.

Tablo 1 Özelliklerin Ağırlık Değeri

Özellikler	**Ağırlık Değeri**
Başlık	20
Yüksek Frekans	10
Giriş	20
Sonuç	2
Anahtar Kelimeler	8
Özel İsimler	3
Pozitif Kelimeler	15
Negatif Kelimeler	-20
Sayı	3
Çift Tırnak	2
Bitiş İşareti	2
Ortalama Uzunluk	10
Tarih	5

4. Deneysel Sonuçlar

Bu bölümde, seçilen 10 farklı site üzerinden özetleme yapılarak sistemin başarı sonuçları değerlendirilmiştir. On farklı siteden indirilen her yazı, 15 farklı kişiye verilerek cümle seçme yöntemi ile özetlerinin çıkarılması istenmiştir. Bu işlem sonrasında elimizde (10*15=150) 150 adet farklı özet bilgisi ile sistemin test edilmesi aşamasına geçilmiştir.

Sistemin gerçek kullanıcılardan gelen her özet için test edilmesinde Tablo 1'deki giriş parametre değerleri kullanılmıştır. Bunun haricinde sistem parametrelerinden olan *özetleme oranı* gerçek kişinin seçmiş olduğu özet cümle sayısının, tüm metindeki cümle sayısına oranı alınarak hesaplanır. Örneğin, özeti çıkarılacak metinde 35, çıkarılan özet içerisinde de 12 cümlenin yer aldığını düşünelim. Böylece özetleme oranı %34 (12/35) olarak sisteme girilir ve sistemin seçtiği özet cümleler ile gerçek kullanıcının seçmiş olduğu cümlerin karşılaştırması yapılarak, her özet metni için başarı hesaplanır. Sisteme, seçimlik olarak özetleme oranı yerine, özet metinde kaç cümle olması gerektiğide verilebilir.

Sistemin başarısı her bir metin için ayrı ayrı hesaplanıp ortalaması alınmıştır. Denemelerde tüm metinler için aynı puanlama parametreleri kullanılmıştır. Kullanıcıların ve sistemin seçtiği cümleler baz alınarak sistemin başarısı hesaplanmıştır. Her bir metnin ve sistemin ortalama başarısı Tablo 2'de verilmektedir..

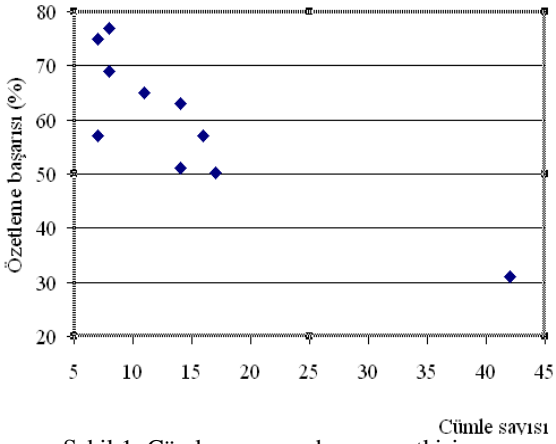
Puanlamalar özetin belirlenmesinde önemli olup, özetin istenen parametrelere göre çıkarılmasını sağlar. Bu tercihin kullanıcıya bırakılarak parametrik bir yapı kurulması, sistemin başarımını

arttırmaktadır. Kullanıcı farklı bir puanlama belirlemediği takdirde sistem üzerindeki başlangıç parametre değerleri kullanılmaktadır.

Tablo 2: Sistemin başarısı
(Cs: cümle sayısı, Öo: özetleme oranı, Öss: özet oranının standart sapması, S%: sistemin başarısı, Bss: başarının standart sapması)

No	Cs	Öo	Öss	S%	Bss
1	14	0,44	1,82	50,93	18,1
2	17	0,40	2,52	50,13	14,1
3	11	0,53	2,18	64,87	17,0
4	14	0,46	2,03	63,00	10,1
5	7	0,57	1,13	75,00	17,5
6	8	0,48	1,85	77,20	19,1
7	8	0,53	0,59	68,73	10,2
8	16	0,38	2,26	57,07	12,2
9	7	0,46	0,68	57,47	13,4
10	42	0,15	3,09	30,93	14,4
Ort	14,40	0,44	1,81	**59,53**	14,6

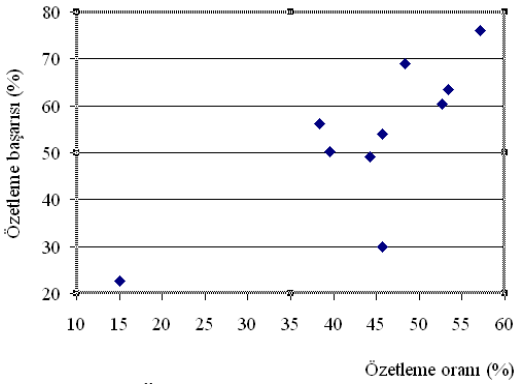
Şekil 1, cümle sayısı ile sistemin özetleme başarısının değişimini göstermektedir. Sistemin başarısını yükselten parametrelerden biri de özetleme oranıdır. Bu değer ne kadar büyükse başarı da o kadar yüksek olmaktadır (Şekil 2).



Şekil 1: Cümle sayısının başarıya etkisi

6. Sonuç

Geliştirilen sistem, Türkçe web sayfalarındaki metinlerin otomatik olarak özetinin çıkarılmasını amaçlamaktadır. Özetleme işlemi cümle seçilerek gerçekleştirilmektedir. Cümlelerin seçilmesinde, her cümlenin sahip olduğu puan önem taşımaktadır. Cümlelere puan verilirken 12 farklı özellik tespit edilmiş ve bu özelliklere sezgisel bir yaklaşımla ağırlık değerleri verilmiştir. Bu değerlerin değiştirilebilme imkanı ve özetleme oranı kullanıcının seçimine bırakılmıştır. Çalışmamızda 10 farklı siteden indirilen yazıların, 15 farklı kişi tarafından özetlenmesi istenmiş ve özet sonuçları sistemin verdiği özetler ile karşılaştırılmıştır. Sistemin özetleme başarısı yaklaşık %59 olarak elde edilmiştir. Cümle sayısının ve özetleme oranının sistemin performansına etkisi incelendiğinde uzun metinlerde performansın düştüğü, özetleme oranı arttırıldığında performansın yükseldiği görülmüştür.



Şekil 2: Özetleme oranının başarıya etkisi

Gelecek çalışmalarımızda, web sitelerinin uyumunu daha iyi inceleyip kullanıcıya çok iş düşmeyecek şekilde özetleme sistemini geliştirmek olacaktır. Ayrıca, sistemde kullanılan 12 adet özelliğin ayrı ayrı değerlendirilip hangi özelliklerin sisteme olumlu, hangilerinin olumsuz etki yaptığını tespit ederek sistemin performansına olumsuz etki eden özellikleri belirleyip sistemden çıkarmayı hedeflemekteyiz.

7. Kaynaklar

[1] H. P., Lunh, "The Automatic Creation of Literature Abstracts", *IBM Journal*, p:159-165, 1958

[2] H.P., Edmundson, "New Methods in Automatic Abstracting", *Journal of the ACM*, Vol.16(2), p:264-285, 1969

[3] Y. Ko and Y. Seo, "An effective sentence-extraction technique using contextual information and statistical approaches for text summarization", *Pattern Recognition Letters*, 2008, Vol. 29, Issue 9, p. 1366-1371.

[4] Z. Altan, "A Turkish Automatic Text Summarization System", *Proceedings of the Artificial Intelligence and Applications*, 2000

[5] C. Pembe, T. Güngör, "Automated Query-biased and Structure-preserving Text Summarization on Web Documents", INISTA, 2007

[6] E. Uzundere, E. Dedja, B. Diri ve M.F. Amasyalı, Türkçe Haber Metinleri için Otomatik haber Özetleme, *Akıllı Sistemler ve Uygulamaları*, ASYU, 2008, Isparta

[7] E. Hovy, C.Y., Lin, "Automated Text Summarization in SUMMARIST", *Annual Meeting of the ACL Proceeding of a Workshop*, 1998

[8] http://code.google.com/p/zemberek/