T.C.

SELÇUK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR TABANLI METİN İŞLEMEDE EN UZUN EŞLEŞME ALGORİTMASININ KARŞILAŞTIRMALI ANALİZİ

Mehmet BALCI

YÜKSEK LİSANS TEZİ

ELEKTRONİK VE BİLGİSAYAR SİSTEMLERİ EĞİTİMİ ANABİLİM DALI

Kasım-2010
KONYA

TEZ KABUL VE ONAYI

Mehmet BALCI tarafından hazırlanan "Bilgisayar Tabanlı Metin İşlemede En Uzun Eşleşme Algoritmasının Karşılaştırmalı Analizi" adlı tez çalışması 13/12/2010 tarihinde aşağıdaki jüri tarafından oy birliği ile Selçuk Üniversitesi Fen Bilimleri Enstitüsü Elektronik ve Bilgisayar Sistemleri Eğitimi Anabilim Dalı'nda YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Jüri Üyeleri	**İmza**
Başkan Prof.Dr. Novruz ALLAHVERDİ	……..
Danışman Yrd. Doç. Dr. Rıdvan SARAÇOĞLU	……..
Üye Yrd. Doç. Dr. Harun UĞUZ	……..

Yukarıdaki sonucu onaylarım.

Prof. Dr. Bayram SADE
FBE Müdürü

TEZ BİLDİRİMİ

Bu tezdeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edildiğini ve tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

DECLARATION PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Mehmet BALCI

22/11/2010

ÖZET

YÜKSEK LİSANS TEZİ

BİLGİSAYAR TABANLI METİN İŞLEMEDE EN UZUN EŞLEŞME ALGORİTMASININ KARŞILAŞTIRMALI ANALİZİ

Mehmet BALCI

Selçuk Üniversitesi Fen Bilimleri Enstitüsü
Elektronik ve Bilgisayar Sistemleri Eğitimi Anabilim Dalı

Danışman: Yrd.Doç.Dr. Rıdvan SARAÇOĞLU

2010, 79 Sayfa

Jüri
Yrd.Doç.Dr. Rıdvan SARAÇOĞLU
Prof.Dr. Novruz ALLAHVERDİ
Yrd. Doç. Dr. Harun UĞUZ

Günümüzde teknolojinin gelişmesi ile birlikte her geçen gün büyük miktarlarda veriler ortaya çıkmaya ve depolanmaya başlanmıştır. Bu verilerden faydalanmanın yolu ise onların verimli bir şekilde organize edilmesi ve yararlı bilgilere dönüştürülmesinden geçmektedir. Bunu amaçlayan veri madenciliğinin bir çeşidi ise metinsel veriler üzerinde çalışan metin madenciliğidir.

Metinsel verilerin incelenmesi ve bu verilerin amaca uygun olarak hızlı bir şekilde kullanılması da bilgi erişim sistemleri için yeni bir problem teşkil etmiştir. Elektronik ortamlarda saklanan metinsel verilerin çoğunun doğal dille yazılmış olmaları da metin madenciliğinde doğal dil işlemenin önemini ortaya koymuş ve metin işleme çalışmalarında metnin yazıldığı dilin yapısının da bilinmesi ihtiyacını doğurmuştur. Türkçe'de köke yapım eki getirilerek oluşturulan yeni kelimeye gövde, bir kelimeye eklenmiş olan çekim eklerinin çıkarılması ile kelimenin gövdesinin bulunması işlemine ise "Gövdeleme" denilmektedir. Anlam olarak hemen hemen aynı ifadeye sahip sözcükler, çekim eki aldıklarında yazılışları tamamen değişmektedir. Bu durumda bu sözcüklerin gövdelerinin bulunması gerekmektedir. Türkçe'nin sondan eklemeli bir dil olduğu göz önünde bulundurulduğunda, verimli bir gövdeleme işleminin metin işlemenin başarısını büyük ölçüde etkileyeceği söylenebilir.

Yapılan bu tez çalışmasında ise, bilgi erişim sistemlerinde yer alan metin işleme süreçlerinden en önemlilerinden biri olan ön işleme süreci ve yine bu sürecin önemli bir parçası olan gövdeleme üzerinde durulmuştur. Tez kapsamında gerçekleştirilen gövdemele yazılımı ile farklı gövdeleme algoritmalarının bilgi erişimindeki performansları karşılaştırmalı olarak analiz edilmiştir.

Anahtar Kelimeler: Metin madenciliği, doğal dil işleme, bilgi erişim sistemleri, gövdeleme, k en yakın komşu.

ABSTRACT

MS THESIS

COMPARATIVE ANALYSIS OF THE LONG MATCH ALGORITHM IN COMPUTER BASED TEXT PROCESSING

Mehmet BALCI

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCE OF SELÇUK UNIVERSITY
THE DEGREE OF MASTER OF SCIENCE IN ELECTRONICS AND COMPUTER SYSTEMS EDUCATION

Advisor: Asst. Prof. Dr. Rıdvan SARAÇOĞLU

2010, 79 Pages

Jury
Asst. Prof. Dr. Rıdvan SARAÇOĞLU
Prof. Dr. Novruz ALLAHVERDİ
Asst. Prof. Dr. Harun UĞUZ

Nowadays, large amount of data has started to arise and stored by development of technology. The way of benefitting these data are to organize them efficiently and convert them to useful information. A kind of data mining that aims this is text minig which works over textual data.

Analysis of the textual data and use of this data quickly, according to the purpose has created a new problem for information retrieval systems. Most of the textual data stored in electronic media also be written in natural language the importance of natural language processing has revealed in text mining, and knowledge of structure of the text is written in language need, has revealed. Turkish affixes to the root of the new words are created by bringing the body, the body of the words finding process by a word that is attached to the removal of the suffixes stemming called. Have almost the same phrase as meaning words, when they have completely changed the spelling of suffixes. In this case, those words must have their body. Turkish is a language to add from the last consideration when, efficient stemming process would affect a large extent the success of text processing can be said.

Contained on information retrieval systems, preprocessing process, Which one of the most important of text processing and an important part of this process was focused on the stemming in thesis. The performance in information retrieval of different stemming algorithms, comparatively analyzed by stemming software in thesis.

Keywords – Text mining, natural language processing, information retrieval systems, stemming, k nearest neighbor.

ÖNSÖZ

Metinsel verilerin günlük hayatımızda ne kadar fazla yer ettiği herkes tarafından bilinen bir gerçektir. Bu gerçekliğe elektronik dünyadaki hızlı gelişmeleri de kattığımızda bilgisayar sistemlerinde saklanan ve işlenen verilerin çoğunun aslında metinsel veriler olduğu görülmektedir. İşte bu verilerin daha verimli bir şekilde işlenilebilmesi için son zamanlarda veri madenciliği içerisinde metin madenciliği çalışmaları hız kazanmıştır. Bu tez çalışmamızda metin madenciliğinin önemli süreçlerinden olan gövdeleme üzerinde durularak literatürde bulunan mevcut gövdeleme algoritmalarından bazıları belge geri getirimi, zaman performansı ve sınıflandırma başarısı bakımından karşılaştırılarak incelenmiştir.

Bu çalışmamda bana yol gösteren ve her türlü bilimsel katkıyı sağlayan değerli hocam ve danışmanım Sayın Yrd. Doç. Dr. Rıdvan SARAÇOĞLU'na teşekkürü bir borç bilirim. Ayrıca çalışmalarımda benden desteğini esirgemeyip kıymetli vakitlerini bana ayıran mesai arkadaşım ve kıymetli dostum Öğr. Gör. Hüseyin ELDEM'e ve kardeşim Abdussamet BALCI'ya çok teşekkür ederim.

Çalışmalarım boyunca her türlü sabrı, hoşgörüyü ve fedakârlığı gösteren eşim Ayşe BALCI'ya ve ailemin tüm değerli fertlerine şükranlarımı sunarım.

Mehmet BALCI
KONYA-2010

İÇİNDEKİLER

SİMGELER VE KISALTMALAR

Kısaltmalar

BES	:	Bilgi Erişim Sistemleri
DKL	:	Durma Kelimeleri Listesi
EUEA	:	En Uzun Eşleşme Algoritması
GKS	:	Gövde Kelimeler Sözlüğü
GTL	:	Gövdelenmiş Terimler Listesi
SUA	:	Sabit Uzunluk Algoritması
VK	:	Veri Kümesi

1. GİRİŞ

Günlük hayatımızın bir parçası olan teknoloji, beraberinde veri miktarında bir patlamayı da getirmiştir. Bu büyük miktardaki verilerden faydalanabilmek için bu veriler kullanışlı bir biçimde saklanmalı, verimli bir şekilde işlenmeli ve bu verilere hızlı bir şekilde erişilmelidir. Yani mevcut ham veri uygun yöntemlerle saklanmalı, işlenmeli ve kullanıcıya ulaştırılmalıdır. Bu gereksinimlerden dolayı ortaya çıkan veri madenciliği büyük miktardaki verilerden faydalı bilgilerin çıkarımı olarak tanımlanabilir (Saraçoğlu, 2007).

Bu verilerin önemli bir kısmını metinler oluşturmaktadır. Özellikle internetin her geçen gün daha da yaygınlaşması ve insanların bilgiye ulaşma isteklerinin de aynı doğrultuda artması neticesinde, doğal dil ile yazılmış veri miktarında da son yıllarda önemli bir atış gözlenmiştir. Bunlara örnek olarak forum siteleri ve elektronik yayınlar gösterilebilir. Tüm dünyadaki bilimsel kuruluşlar ve özellikle üniversiteler bilimsel çalışmalarını elektronik ortamlarda yayınlamaktadır. Bu çalışmalardan faydalanmak isteyen bilim çevreleri de bu metinler üzerinde çalışırken aradıkları bilgiye daha çabuk ulaşmak istemektedirler. İşte bu ihtiyaçtan dolayıdır ki, doğal dil ile yazılmış metinlerin işlenmesi veri madenciliğinin içerisinde ayrı bir çalışma sahası ortaya çıkarmış ve bu çalışma sahasının adına "Metin Madenciliği" ya da "Metin İşleme" (Text Processing) denilmiştir.

1.1. Çalışmanın Amacı ve Önemi

Metin işleme teknikleri, özellikle Bilgi Erişim Sistemleri (BES)'nin temelini teşkil ederken, metinsel verilerin işlenmesinde bilgi çıkarım performansına çok büyük oranda etki eden gövdeleme de hiç şüphesiz büyük önem arz etmektedir. İngilizce üzerine yapılan çalışmaların yanı sıra son yıllarda Türkçe metinlerin gövdelenmesi ile ilgili çalışmalar da git gide yaygın hale gelmiştir.

Bu tez çalışması kapsamında Türkçe metinlerin gövdelenmesi üzerine yapılan çalışmalar incelenmiş ve bu yöntemlerden bazıları kodlanarak deney ortamı meydana getirilmiştir. Oluşturulan yazılım sayesinde bilgi erişiminde farklı yöntemlerin etkisi istatistiksel olarak da ortaya koyulmuştur. Ortaya koyulan bu sayısal verilerle

gövdelemenin etkisi ve gövdeleme için kullanılacak teknikler karşılaştırmalı olarak incelenmiş ve bundan sonraki çalışmalara ışık tutması amaçlanmıştır.

1.2. Literatür Araştırması

Metin madenciliği ve metin madenciliğinin en önemli süreçlerinden olan ön işleme süreçleri hakkında bu güne kadar birçok çalışma yapılmıştır. Bu tez çalışması kapsamında Türkçe metinler için yapılmış ön işleme ve gövdeleme çalışmalarına yer verilmiştir. Bu çalışmaların bazıları kısaca şu şekildedir:

Günümüzde çeşitli alanlardaki bilginin büyük bir kısmı metin belgelerinde yer almaktadır. Bilgi ve belgelerin elektronik ortama aktarılması veya elektronik ortamda saklanılmasından dolayı metinsel veriler hızlı bir şekilde artmaktadır. Bu hızla artan verilere en önemli örnek elektronik postalar ve web sayfalarıdır (Kantardzic, 2003).

Metin madenciliği de bir veri madenciliği yöntemidir ve yapısal olmayan metinlerden bilgi keşfi yapılmasını sağlar. Yaygın olarak aynı konuda yazılmış belgeleri bulmak, birbiriyle ilişkili belgeleri bulmak, kavramlar arası ilişkileri keşfetmek için kullanılır (Yıldırım ve ark., 2008).

Metin madenciliğinin en önemli aşamalarından birisi ön işlemedir. Ön işleme, işlenecek olan metinler üzerinde hazırlık işlemleri yapılarak, daha hızlı işlenmesine ve amaca uygun verilere daha çabuk ulaşılmasına olanak tanır (Türkeş, 2007).

Kısaca gövdenin tanımını yapacak olursak; Türkçe'de köke yapım eki getirilerek oluşturulan yeni kelimeye gövde ismi verilir (Güzel, 2005).

Gövdeleme işlemi dillere göre büyük farklılık göstermektedir. Örneğin İngilizce gibi eklerin kullanımının az olduğu bir dil için yalnızca ekler sözlüğüne bakılarak bir gövdeleyici geliştirmek mümkündür (Porter, 1980).

Türkçe'nin bitişken bir dil olmasından ötürü, eklerin sayısı ve eklenme çeşitleri daha detaylı bir inceleme yapılmasını zorunlu kılar (Jurafsky ve Martin, 2000).

N-Gram yönteminde ise veritabanında tüm kelime gövdeleri tutulmaktadır. Gövdesi aranan kelime bu veritabanındaki gövdelerle karşılaştırılarak gövde araması yapılır. Bu karşılaştırma işleminde ise n-gram eşleştirme teknikleri diye adlandırılan yaklaşımlar kullanılır (Freund ve Willett, 1982). Bir n-gram, bir kelimeden çıkarılmış n

uzunluğunda ardışık karakter dizileridir. Bu yaklaşımdaki temel düşünce ise benzer kelimelerin yüksek oranda aynı n-gram dizilerine sahip olduğudur. Genellikle bu n değeri 2 veya 3 olarak seçilir ve sırasıyla *digrams* ve *trigrams* olarak adlandırılır (Ekmekcioglu ve ark., 1996).

N-gram yöntemi, bir metnin hangi dilde yazıldığını bilgisayar tarafından belirleyebilmek amacıyla da kullanılabilir (Kohonen, 1990).

Biçimbirimsel çözümleyiciler, verilen bir sözcüğün tüm olası kök ve ek birleşimlerini üretirler (Oflazer, 1994).

Biçimbirimsel çözümleyiciler bir sözcüğü kök ve eklerine ayrıştıran sistemlerdir (Kesgin, 2007).

Gövdeleme konusunda Türkçe için ilk çalışma Aydın Köksal tarafından 1981 yılında yapılmıştır (Köksal, 1981). Bu çalışmada kelimenin ilk beş harfi o kelimenin gövdesi olarak kabul edilmiştir. Bu karara birçok kelimenin incelenmesi neticesinde varılmıştır.

Gövdeleme konusunda Solak ve Can'ın yaptığı çalışma şu şekildedir (Solak ve Can, 1994). İlk önce kelime sözlükte aranır, kelime bulunamazsa sonundan bir harf silinir ve ardından gövdeler arasında tekrar aranır. Süreç bu şekilde devam eder.

Bir diğer çalışma ise Adil Alpkoçak, Alp Kut ve Esen Özkarahan tarafından Dokuz Eylül Üniversitesi bünyesinde yapılan "Bilgi Bulma Sistemleri için Otomatik Türkçe Dizinleme Yöntemi" adlı çalışmadadır. (Alpkoçak ve ark., 1995). Bu çalışmada Türkçe'de kullanılan durma kelimelerinin listesi (DKL) oluşturulmuş ve dizinlemede kullanılacak anahtar kelimeleri elde etmek için en fazla eşleşme (maximum match) algoritması kullanılmıştır.

Gökmen Duran tarafından yapılan, "Gövdebul" adlı Türkçe gövdeleme algoritmasında (Duran, 1997) kelime soldan sağa incelenmekte ve elde edilen harf katarı sözlükte aranmaktadır. Ardından biçimbirimsel inceleme yapılmaktadır.

Biçimbirimsel çözümleyici kullanarak gövdeleme işlemi Altıntaş ve Can tarafından gerçekleştirilen çalışmada denenmiştir (Altıntaş ve Can, 2002). Bu çalışmada Oflazer tarafından geliştirilen biçimbirimsel çözümleyici (Oflazer, 1994) kullanılmış ve elde edilen sonuçlar içersinden ortalama gövde uzunluğu ve n-gram yöntemleri ile doğru gövde belirlenmeye çalışılmıştır.

Biçimbirimsel çözümleme sonuçları arasından uygun gövdeyi seçmek için uygulanabilecek yöntemlerden biri de Türkçe için ortalama gövde uzunluğunun

belirlenmesi ve olası gövdeler arasında ortalama uzunluğa en yakın olan gövdenin seçilmesidir (Altıntaş ve Can, 2002).

1.3. Tezin Organizasyonu

Bilgi erişim sistemlerinde metinlerin gövdelenmesinin etkileri üzerine çalışılmış olan bu tez altı bölümden oluşmaktadır. Tez çalışmasının ana hatları aşağıdaki gibidir:

Birinci bölümde, tez çalışmasının konusuna genel bir bakış açısı verilmeye çalışılmıştır. Çalışmanın amacı, geçmişte yapılan çalışmalar ve ortaya koyduğu bilgilere kısaca değinilmiştir.

İkinci bölümde Türkçe'nin yapısı hakkında genel bilgiler verilerek, ek yapısı incelenmiştir. Bu bölümde kelime türemesinde kullanılan tüm yapım ekleri grup grup anlatılarak örneklere de yer verilmiştir. Aynı şekilde, Türkçe'de kullanılan çekim ekleri de örneklerle anlatılarak sondan eklemeli bir dil olan Türkçe'nin eklerle olan ilişkisi ortaya koyulmaya çalışılmıştır.

Üçüncü bölümde BES anlatılmaya çalışılmış ve bilgi erişiminin, bilgi sistemlerindeki yeri, Türkiye'de yapılan bilgi erişim çalışmaları ve doğal dil işleme konularından detaylı bir şekilde bahsedilmiştir.

Dördüncü bölümde "Metin Madenciliği nedir?", "Neden böyle bir çalışmaya ihtiyaç duyulmuştur?" sorularına yanıt aranmış; metin madenciliğinin aşamaları hakkında kısaca bilgi verilerek ön işleme ve gövdeleme konuları detaylı bir şekilde işlenerek metinlerin matematiksel modelinin çıkarılması, terim ağırlıklandırma ve metin sınıflandırma konularına da değinilmiştir.

Beşinci bölümde farklı gövdeleme yöntemlerini kullanan bir BES yazılımı meydana getirilmiş ve 1000 adet belge (tez-makale) üzerinde deneysel uygulamalar yapılmıştır. Deney sonuçları da bu bölümün son kısımlarında rakamlarla sunulmuştur.

Altıncı yani son bölümde ise çalışma sonuçları değerlendirilmiş ve önerilerde bulunulmuştur.

Kaynaklar bölümünde ise bu tez çalışmasında faydalanılan kaynaklar ve referanslara yer verilmiştir.

2. TÜRKÇENİN YAPISI

Türkçe Ural-Altay dil grubuna dâhil eklemeli bir dildir ve sondan eklemeli dillerin tipik bir örneğini teşkil eder. Türkçe, doğal dil işleme ile ilgilenenlerin dikkatini çekecek kadar kurallı bir dildir. Çünkü kuralları oldukça kesindir ve yıllardan beri korunmuştur.

Türkçe'de köke eklenen her ek sözcüğe yeni bir anlam kazandırır. Türkçe'de ekler türlerine göre yapım ekleri ve çekim ekleri olmak üzere ikiye ayrılır. Çekim ekleri kalıplaşmış kullanımlar dışında yeni bir kelime türetmeyen, kelimeler arasında durum, iyelik, çokluk, kip, zaman, şahıs, gibi ilgiler kuran eklerdir. Yapım ekleri ise, eklendikleri kelimeden yeni bir kelime türeten eklerdir.

Türkçe'de yapım ve çekim ekleri sayıca fazladır. Bu zenginlik, diğer dillerde ayrı kelimeler olarak ifade edilen anlamların Türkçe'de bir ek ile ifade edilebilmesini mümkün hâle getirmiştir. Örneğin, "yapmazsın" kelimesi ile "yapamazsın" kelimesi arasında sadece "a" harfi farklıdır ancak taşıdıkları anlam açısından iki kelime arasında belirgin bir ayrım vardır.

Türkçe'de aynı görevde yapım ekleri veya aynı çekim eki arka arkaya gelemez ancak farklı görevde yapım ekleri ve farklı çekim ekleri belirli kurallar içerisinde peş peşe gelebilir.

Türkçe'de çekim eki yapım ekinden sonra gelir ancak çekim ekinden sonra yapım eki gelmez. Yapım ekleri arasında sayılan "-ki" eki bunun istisnasıdır. Ayrıca bazı kalıplaşmış çekim eklerinden sonra yapım eki gelebilir.

Örneğin,

"Ayakkabıcı" kelimesi *ayak+kap+ı+cı* şeklinde kök ve eklerine ayrıştırılabilir. Burada "*-ı"* bir çekim eki olmasına karşın bu kelime içerisinde kalıplaşmıştır ve dolayısı ile sonrasında gelen "*-cı"* yapım ekini alabilmiştir. Ancak "*yemekkabıcı"* gibi bir kullanım günümüz Türkçe'sinde söz konusu değildir.

Eklerin sözcüklere eklenmeleri, aynı zamanda ünlü ve ünsüz uyumu kuralına göre gerçekleşmektedir.

Ünlü uyumunda, ek içerisinde yer alan “a” ve “e” ünlüleri kalınlık-incelik hâline göre, “ı”, “i”, “u”, “ü” ünlüleri ise hem kalınlık-incelik hem de düzlük-yuvarlaklık durumuna göre eklendiği kök ile uyum gösterirler. Bu kurallara göre:

a'dan sonra :a, ı

e’den sonra :e, i

ı’dan sonra :a, ı

i’den sonra :e, i

o’dan sonra :a, u

ö’den sonra :e, ü

u’dan sonra :a, u

ü’den sonra :e, ü

ünlüleri gelebilir. Örneğin *sepet-ler, sepet-çi* kelimelerinde *-ler ve -çi* eklerinde ünlüler bu kurala uygun olarak getirilmiş eklerdir.

Ünsüz uyumu kuralına göre, sözcük sonunda bulunan ünsüz eğer *p, ç, t, k, s, h, ş, f* harflerinden biri ise, *c, d, g* harfleri ile başlayan eklerin ilk ünsüzleri sertleşerek *ç, d, k* haline dönüşür. Örneğin *çarşaf-ta, yavaş-ça, kat-kı* kelimelerindeki ekler ünsüz uyumu gereği sertleşmişlerdir.

Yine ünsüz uyumu kuralına göre, sonunda *p, ç, t, k* harflerinden birini bulunduran bir sözcüğün sesli harf ile başlayan bir ek alması durumunda son harfi yumuşayarak *b,c,d,ğ* harflerine dönüşür. Örneğin *kitab-ı, kâğıd-a, kazığ-ı* kelimelerinde son harfler ünsüz uyumu gereği yumuşamışlardır.

Türkçe’de bir kelimenin anlam ifade eden en küçük parçası kök olarak isimlendirilir. Bu köklere getirilen çeşitli sayıda ve sırada eklerle Türkçe’nin söz varlığı oluşmaktadır. Bu söz varlığı yeni eklemelerle zaman içinde genişlemektedir. Türkçe’de kökler isim ve fiil kökü olarak ikiye ayrılırlar. Yapım eki alarak bir kökten türemiş anlamına gelen *gövde* kelimeler ile kökler arasında kullanım bakımından hiçbir fark yoktur.

2.1. Yapım Ekleri

Yapım ekleri eklendikleri kelimenin anlamını değiştirerek yeni bir kelime türeten eklerdir. Günümüz Türkçe'sinde, iki yüze yakın yapım eki vardır (Güzel, 2005).

Yapım ekleri, isimden isim yapan, isimden fiil yapan, fiilden isim yapan, fiilden fiil yapan olmak üzere dörde ayrılırlar.

2.1.1. İsimden isim yapan ekler

Bir isme eklenerek yeni bir isim türeten eklerdir. Aynı ek olmamak şartı ile birden fazla isimden isim yapan ek peş peşe kullanılabilir. Bu ekler hem Türkçe kökenli hem de yabancı kökenli kelimelere getirilebilir. Bu ekler ile oluşturulmuş kelimelere örnekler şu şekildedir.

-ay, -ey : düzey, birey

-cı, -ci, cu, cü / çı, çi, çu, çü : arabacı, sucu

-lik, -lık, -luk, -lük : gözlük, kitaplık

2.1.2. İsimden fiil yapan ekler

İsim kök ve gövdelerinden, kökün taşıdığı anlamla ilişkili gövdeler oluşturan eklerdir. Sayıca az olan bir ek türüdür. En işlek olanı *–la, -le* ekidir.

-la, -le : kuzula-, şişmanla-, sabahla-, gözle

2.1.3. Fiilden isim yapan ekler

Fiil köklerine eklenerek kalıcı veya geçici olarak isimler türeten eklerdir. Bu eklerden *–mak, -me, -ış* ekleri hem kalıcı hem geçici isimler türetebilirler.

-acak, -ecek : açacak, gelecek

-ma, -me : bölme, uçurtma

-mak, -mek (geçici) : bilmek, çıkmak, duraklamak

-mak, -mek (kalıcı) : çakmak, yemek

2.1.4. Fiilden fiil yapan ekler

Bu ekler eklendikleri fiilden yeni bir fiil türetebileceği gibi, eklendikleri fiilin anlamını değiştirmeden, özne veya nesnenin fiilin gerçekleşmesi ile ilgili durumunu değiştirerek çatılarını değiştirebilirler.

-ır, -ir, -ur, ür (çatı) : artır-, göçür-

-ala, -ele : kovala-, silkele-

2.2. Çekim Ekleri

Çekim ekleri yapım eklerinde göre daha fazla kullanılan ve eklendikleri kelimelere işlerlik kazandıran eklerdir. Eklendikleri kelimelerden yeni bir kelime türetmezler ancak kelimenin şahsı, iyeliği, zamanı, çokluğu gibi nitelikler üzerinde değişiklik yaparlar. Çekim ekleri, isim çekim ekleri ve fiil çekim ekleri olmak üzere ikiye ayrılırlar.

2.2.1. İsim çekim ekleri

İsimlerin *durum, iyelik, çokluk, soru* ve *şahıs* bilgileri üzerinde değişiklik yapan eklerdir. Bu ek çeşitleri ve ilgili ekler şu şekildedir.

2.2.1.1. Durum ekleri

İsimlerle, diğer isim ve fiiller arasında anlam ilgisi kuran eklerdir.

Belirtme durumu eki : -ı, -i, -u, -ü

Bulunma durumu eki : -da, -de / -ta, -te

Çıkma durumu eki : -dan, -den / -tan, -ten

Eşitlik durumu eki : -ca, -ce / -ça, -çe

İlgi durumu eki : -ın, -in, -un, -ün / -nın, -nin, -nun, -nün

Vasıta durumu eki : -la, -le

Yönelme durumu eki : -a, -e

2.2.1.2. İyelik ekleri

Eklendikleri ismin hangi şahsa veya nesneye ait olduğunu bildirirler.

Birinci tekil kişi : -m

İkinci Tekil Kişi : -n

Üçüncü Tekil Kişi : -ı, -i, -u, -ü / -sı, -si, -su, -sü

Birinci Çoğul Kişi : -mız, -miz, -muz, -müz

İkinci Çoğul Kişi : -nız, -niz, -nuz, -nüz

Üçüncü Çoğul Kişi : -ları, -leri

2.2.1.3. Çoğul eki

Eklendikleri isme çokluk özelliği kazandıran eklerdir. Aynı zamanda, eklendikleri kelimeye bazı özel durumlarda topluluk (Osmanlılar), saygı (Vali Beyler) gibi anlamlar da kazandırabilirler.

Çoğul Eki : -lar, -ler

2.2.1.4. Soru eki

Eklendikleri isimle ilgili bir soru oluştururlar. Dil kuralı gereği eklendikleri kelime ile ayrı kendisinden sonra gelen kelime ile bitişik yazılırlar. Örneğin, "Evde miydiniz?"

Soru Eki : -mı, -mi, -mu, -mü

2.2.1.5. Şahıs eki

Ben : -ım, -im, -um, -üm

Sen : -sın, -sin, -sun, -sün

O : -dır, -dir, -dur, -dür / -tır, -tir, -tur, -tür

Biz : -ız, -iz, -uz, -üz

Siz : -sınız, -siniz, -sunuz, -sünüz

Onlar :-dırlar,-dirler,-durlar,-dürler/-tırlar,-tirler,-turlar,-türler

2.2.2. Fiil çekim ekleri

Fiiller ile isimler ve fiiller arasında geçici anlam ilişkisi kurmaya yarayan eklerdir. Eklendikleri fiilin anlamını kişi ve nesnelere bağlayabilir ya da, fiile *şekil, zaman, kişi* ve *soru* bilgisi eklerler. Şekil ve zaman ekleri *kip ekleri* olarak da isimlendirilirler.

2.2.2.1. Bildirme kipleri

Görülen geçmiş zaman : -dı, -di, -du, -dü / -tı, -ti, -tu, -tü

Duyulan geçmiş zaman : -mış, -miş, -muş, -müş

Şimdiki zaman : - yor

Kesin şimdiki zaman : -mekte, -makta

Geniş zaman : -r

Gelecek zaman : -acak, -ecek

2.2.2.2. Tasarlama kipleri

Emir : Ek kullanılmadığı zaman emir kipidir

Gereklilik : -meli, -malı

İstek : -e, -a

Şart : -se, -sa

2.2.2.3. Birleşik zaman çekimleri

Hikâye : -dı, -di, -du, -dü / -tı, -ti, -tu, -tü

Rivayet : -mış, -miş, -muş, -müş

Şart : -se, -sa

2.2.2.4. Şahıs ekleri

Ben : -m, / -ım, -im, -um, -üm / -ayım, eyim

Sen : -n, / -sın, -sin, -sun, -sün

O : -dır,-dir,-dur,-dür /-tır,-tir,-tur,-tür /-sın,-sin,-sun,-sün

Biz : -k / -ız, -iz, -uz, -üz / -alım, -elim

Siz :-nız,-niz,-nuz,-nüz/-sınız,-siniz,-sunuz,-sünüz/-ın,-in, -un,-ün

Onlar :- lar, ler / -sınlar, -sinler, -sunlar, -sünler

2.2.2.5. Sıfat fiil ekleri

Eklendikleri fiilden geçici sıfatlar oluştururlar. Örneğin, “Okuyacağımız kitap.”

Sıfat fiil ekleri : -acak, -an, -ar, -ası, -dı, -dık, -maz, -mış, -r

2.2.2.6. Zarf fiil ekleri

Eklendikleri fiillerden geçici olarak zarf üreten eklerdir. Bu eklerle oluşturulan kelimeler kalıcı olmaya uygun değillerdir.

Zarf fiil ekleri :-a,-alı,-arak,-dığında,-dıkça,-ı,-ınca,-ıp,-ken,-madan

(Kesgin, 2007)

2.3. Bölüm Sonucu

Gövdeleme işleminde köke eklenen yapım eklerinin yerinde kaldığı ve sadece çekim eklerinin kelimeden çıkarılıp atıldığı düşünüldüğünde, bir kelimenin gövdesinin bulunması için, ilk önce Türkçe'de kullanılan yapım ve çekim eklerinin bilinmesi ihtiyacı ortaya çıkmaktadır. Bu bölümde Türkçe'de kullanılan tüm yapım ve çekim ekleri, türlerine ve kullanıldıkları şahıslara göre gruplandırılarak anlatılmıştır.

3. BİLGİ ERİŞİM SİSTEMLERİ

3.1. Bilgi Sistemleri

Bilgi Sistemleri, bilginin tutarlı ve güvenli bir şekilde saklanıp, kısa sürede ve kolay bir şekilde erişilmesi amacıyla geliştirilmiş sistemlerdir.

Bilgi Sistemleri'nde, uzman olmayan kullanıcılar bile, zorlanmadan bilgi ihtiyaçlarını ifade edebilmeli ve cevap olarak döndürülen bilgi kolay bir şekilde kullanabilmelidir. Kullanıcı üzerinde mümkün olduğunca az sınırlandırma olmalı ve kullanıcının kullanımı kolay ara yüzler üzerinden işlem yapması sağlanmalıdır. Bunun yanında saklanılan bilgide bir kayıp olmamalı, tutarlılık korunmalı, güvenlik sağlanmalı ve kullanıcının bilgi istemiyle eşleşen bilgilerin mümkün olduğunca fazlası çıkış ortamına yansıtılmalıdır.

Bilgi Sistemleri'nde, kullanıcının sorgusuna cevap verme süresi de çok önem taşımaktadır. Kullanıcının her sorgusuna, makul sayılabilecek bir sürede cevap verilmesi daha etkin ve çevrim-içi bir kullanımı olanaklı kılacaktır.

İyi bir Bilgi Sistemi, kullanıcıdan geri-bildirim alarak, ilgili parametreleri güncellemeli ve her seferinde daha mükemmel bir geri getirim çıktısı hedeflemelidir. Kullanıcıdan sorgu sonucuyla ilgili olarak alınan geri-bildirim, kullanıcıyı sıkacak ayrıntı taşımadığı gibi sitemin başarımını artıracak her türlü bilgiyi içermelidir.

Bilgi Sistemleri'ne örnek olarak, Veritabanı Sistemleri (Database Systems), Bilgi Erişim Sistemleri (BES) (Information Retrieval Systems), Uzman Sistemler (Expert Systems), Soru Yanıtlama Sistemleri (Question Answering Systems), Coğrafi ve Kentsel Bilgi Sistemleri (Geographical and Urban Information Systems) sayılabilir.

Veritabanı sistemleri en çok ve en yaygın şekilde kullanılan bilgi sistemleridir. Günümüzde hem metin tabanlı veriler, hem de resim, ses, görüntü gibi veriler üzerinde başarıyla çalışan veritabanı sistemleri bulunmaktadır.

Veritabanı sistemleri, bilgi tutarlılığını ve güvenliğini başarıyla sağlarken, kullanıcının girdiği sorguyla eşleşen bilgilerin hemen hemen hepsini çıkış ortamına yansıtabilmektedir.

Ticari veritabanı sistemlerinde en çok kullanılan modeller, ilişkisel (relational) ve nesneye dayalı (object oriented) veritabanı modelleridir.

Uzman Sistemler ve Soru Yanıtlama Sistemleri, yapay zekâ (artificial intelligence) yöntemleri kullanılarak geliştirilen sistemlerdir. Tıp, iş yönetimi, bilimsel çalışmalar vs. için geliştirilmiş birçok uzman sistem bulunmaktadır. Bu sistemlerde: kural tabanlı (rule-based) ve bulanık mantık (fuzzy logic) yöntemleri çoğunlukla kullanılan yaklaşımlardır (Duran, 1997).

BES, örnek olarak gazete arşivleri, kütüphane katalogları, makaleler, hukuk bilgileri gibi doğal dilde yazılmış metinlerden oluşan verilerin elektronik ortamda saklanması ve erişilmesi için oluşturulan sistemleridir (Sever, 2002).

BES'nin kullanılmasıyla, ofis ve diğer arşiv ortamlarında kullanılan yöntemler terk edilirken, bilgi kayıpları, bilginin istenildiğinde elde edilememesi, bilgi güvenliği gibi sakıncalar da ortadan kaldırılmıştır.

Günümüzde, doğal dil işleme ve anlama üzerinde yapılan çalışmalarla birlikte, kullanıcının bilgi ihtiyacını doğal dil kullanarak ifade edebileceği ara yüzler de geliştirilmeye çalışılmaktadır (Duran, 1997).

3.2. Bilgi Erişim Sistemleri (BES) Nedir?

BES, bilgi vermekten önce, belli bir konuya ilişkin belgelerin varlığını ve nerede bulunabileceğinin bildirilen sistemlerdir. Türkiye'de bu konuda çalışmalar yapmış olan A. Köksal bilgi erişimini : "Bir bilgi erişimi sistemini kullanarak, içerik bakımından araştırılan konu ve kavramlar ile ilgili olarak genellikle varlığı bile bilinmeyen belgelerin izini bulmayı amaçlayan araştırma" biçiminde tanımlamıştır. Buna örnek olarak gazete arşivleri, kütüphane katalogları, makaleler, hukuk bilgileri gibi doğal dilde yazılmış metinlerden oluşan verilerin elektronik ortamda saklanması ve erişilmesi için oluşturulan bilgi erişim sistemleri verilebilir. Günümüzde Türkçe elektronik ortamlarda tutulan belgelerdeki büyük artış, bu belgeleri kullanıcının hizmetine sunacak araçlara/sistemlere gereksinimi artırmıştır.

BES'nde, kullanıcının ihtiyacını ifade eden sorguya karşılık belgeler üzerinde arama yapılarak, kullanıcının ihtiyacı ile ilişkili olduğu düşünülen belgeler döndürülür. Bu işleme yakından bakılırsa belgeleri sisteme sunuldukları halleri ile değil, belgelerin

içeriğini yansıtan belirteç kümesi halinde kullanma zorunluluğu görülür. Bu durum aşağıdaki Şekil 3.1 ‘de de görülmektedir. Bu içerik belirteçlerine anahtar kelime, dizin terimi, tanımlayıcı gibi adlar verilir. İçerik belirteçleri kullanmak, ilk olarak 1950’li yılların sonuna doğru Luhn tarafında önerilmiştir (Lassila, 1998). Her bir dizin terimi belgelerin içeriğini bütünüyle değil, ancak bir yönüyle ifade eder ve bir belge için birçok dizin terimi seçilir. Sorgu ile belge karşılaştırmalarında ise her ikisi için belirlenen içerik belirteçleri arasında birebir eşlemeden çok yeterli oranda benzerlik aranır. Görüleceği üzere erişim işlemlerinde performans dizin terimleri seçimine direk bağlıdır. Belge için dizin terimlerini belirleme işlemine dizinleme denir. Başka bir deyişle dizinleme bir süreçtir ve bu süreç terim-belge, terim-terim ilişkileri ile şekillenir.

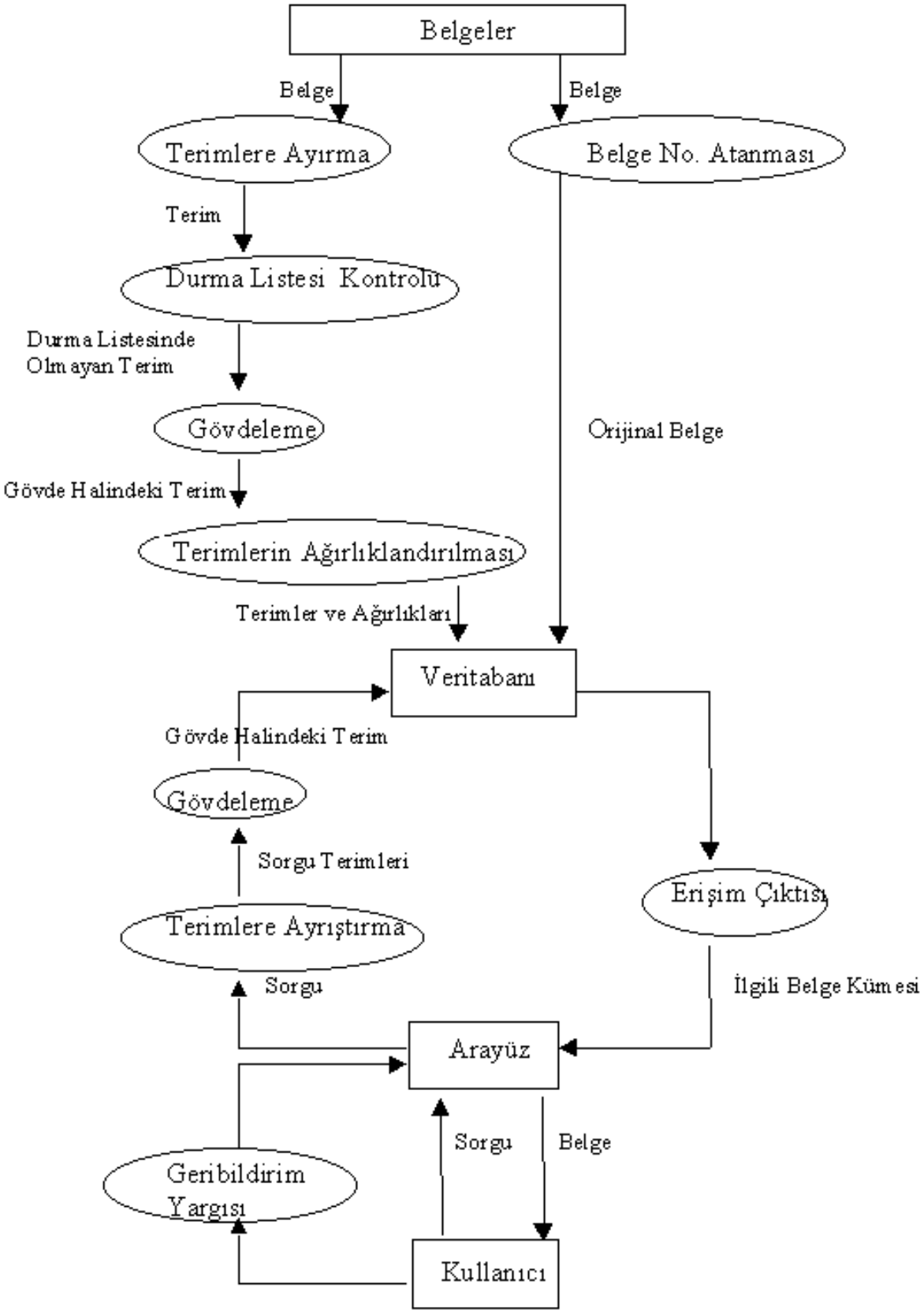
Dizinleme işlemini otomatik ya da elle (manual) gerçekleştirmek mümkündür. Elle dizinlemede konu uzmanları önceden hazırlanmış, terimlerin kullanımını gösteren sözlüklere bakarak ilgili belge için dizin terimlerini seçerler. Bu tarz ile yüksek bir biçimdeşlik (uniformity) ve kalite elde etmek mümkündür. Terimlerin seçimi belli bir konu sözlüğüne göre yapıldığı için kontrollü dizinleme denir. Ancak terim sözlüklerini hazırlama ve BES’nde kullanmanın zorluğu ve maliyetinden kurtulmak için belgeler üzerinde direk işlem yaparak dizin terimlerinin seçildiği otomatik yöntemler geliştirilmiştir. Belge, dizinleme için direk girdi olabildiğinden dizin terimleri daha çeşitlidir ve kullanıcı, sorgusunu daha rahat ifade edebilir. Bu yönteme de kontrolsüz veya otomatik dizinleme denir.

Bir bilgi erişim modelinde, her bir terimin ayrım gücü vardır. Ayrım gücü, terimin belge uzayındaki belgeleri birbirinden ayırabilme/sınıflayabilme gücüdür. Derlem içinde geniş anlamlı, sık kullanımlı terimler için ayrım gücü az; daha dar anlamlı, az kullanımlı terimler için yüksek beklenir. Örneğin bilgisayar terminolojileri ile ilgili belge derlemi için ‘bilgisayar’ terimi, ‘madenleme’ teriminden daha az ayrım gücüne sahiptir. Bilgi erişim terminolojisinde ayrım gücü neredeyse hiç olmayan terimlerin oluşturduğu listeye durma kelimeleri listesi (DKL) denir. Örneğin ‘ve’ terimi içinde geçtiği belgeyi diğer belgelerden ayıramaz. Zira birçok belgede geçebilecek bir terimdir. Aynı biçimde ‘ya da’, ‘ama’…vs gibi kelimeler Türkçe DKL’ne aittir. Liste oluşturulurken öncelikle Türkçe’de yer alan zarf, sıfat, ilgeç, zamir ve bağlaçlar Türkçe’de dizin terimi olarak anlamı olmayan sözcükler olarak algılanmıştır. Terimlerin içinde geçtikleri belge sayısı, o terimin belge sıklığıdır. İdeal olanı dizin terimlerinin orta sıklıklı ve ayrım gücü yüksek terimlerden oluşturulmasıdır. Çünkü sorguyla ilgili

ve ilgisiz olan belgeleri belirlerken, ayrım gücü yüksek terimleri kullanmak daha başarılı sonuçlar üretir.

BES, doğal dilde yazılmış belgeleri yine onların içinden seçilen terimler ile ifade ettiğinden doğal dile bağımlılığı büyüktür. Terimler, doğal dilde yazılı belgelerden çıkartılırken belgelerden elde edildikleri halleri ile dizinde yer alırlarsa aynı gövdeden türemelerine rağmen, farklı çekim eki almış terimler farklı algılanır. Eğer kullanıcı da aynı gövdeye başka bir çekim eki getirerek sorgu hazırlarsa, aynı gövdeden türemiş terimlerin birbiri ile yakınlığı terimler gövdelenmeden yakalanamaz. Dolayısıyla, terimin gerek bir belge için sıklığı gerekse geçtiği belge sayısı gövdeleme yapılması ile yapılmaması tercihine göre çok değişir.

Bir bilgi erişim modelinde, var olan belgelerin oluşturduğu kümeye belge derlemi; sunulan bir sorgu ile ilişkili olarak dönen belgeler listesine erişim çıktısı denir (Sever, 2002).



Şekil 3.1. Bilgi erişim sistemi modeli

3.3. Türkiye'de Bilgi Erişimi Üzerine Yapılan Çalışmalar

Türkiye'de BES'ne ihtiyaç kendini son yıllarda bilgilerin elektronik ortamda tutulmasındaki artışla göstermiştir. Bu eğilim hem devlet kuruluşlarında hem de diğer özel ya da yarı özel kuruluşlarda bilgi bankaları ya da elektronik arşiv adları altında görmek mümkündür. Nitekim, Başbakanlık'a bağlı Mevzuat Bilgi Sistemi, Plastik Sanat Veri Bankası, Resmi Gazete'nin elektronik arşivlenmesi ve sorgulama sistemi hizmete sunulmuştur. Bu sistemler ya Plastik Sanat Veri Bankasında olduğu gibi veriyi biçimlendirmiş ya da düz metin kütüğü üzerinde damga/alt damga (string/substring) taraması gibi daraltılmış metin işleyici olanaklarını desteklemiştir (Duran, 1997).

Türkiye'de BES üzerine yapılan ilk çalışma Aydın Köksal'ın doçentlik tezi olarak gerçekleştirdiği "Özdevimli Deneysel Bir Belge Dizinleme ve Erişim Dizgesi", TURDER'dir (Köksal, 1981). TURDER esas olarak tasarım boyutunda kalan ve kısmen gerçekleştirilmiş bir vektör tabanlı ve geri bildirimsiz BES'dir.

Türkiye'de bilgi geri-getirimi üzerinde yapılan başka bir çalışma da, Bilkent Üniversitesi'nde A.Solak ve F.Can tarafından gövdelemenin Türkçe metin geri-getirim konusundaki etkisini incelemek için, deneysel amaçlı BES geliştirilmiştir (Solak ve Can, 1994). Asıl amaç Türkçe gövdelemenin bilgi geri-getirim etkinliğini artırmadaki etkisini ölçmek olduğu için, geliştirilen BES kullanıma yönelik olmaktan çok, deneysel amaçlı ve basit bir sistemdir (Duran, 1997).

Türkiye'de yapılan bir başka çalışma da Ebru Sezer tarafından, Hacettepe Üniversitesi'nde yüksek lisans tezi olarak hazırlanmıştır (Sezer, 1999). Bu çalışmada Türkçe üzerinde çalışan, baştan sona yeni bir bilgi geri-getirim sistemi oluşturulmamıştır. Sezer, bu tez çalışması kapsamında, İngilizce tabanında geliştirilmiş SMART Bilgi Erişim Sistemi'nin Türkçe belgeleri sağlıklı işleyecek biçimde güncellenmesi üzerinde çalışmıştır. SMART'ın Türkçe'ye uygun olarak güncellenmesi için gerekli en önemli adımlardan birisi olan terimlerin gövdelenmesi işlemi için ise 1997'de Gökmen Duran'ın yüksek lisans tezinde geliştirdiği "GövdeBul Türkçe Gövdeleme Algoritması" kullanılmıştır. Bu çalışmada amaçlananlar ise şunlardır:

- Farklı bir doğal dil tabanında geliştirilen bir bilgi erişim sisteminin Türkçe yerelleştirme için ihtiyaç duyduğu uyarlamalar ve ekleri ortaya çıkarmak

- Elde edilecek Bilgi Erişim Sistemi ile hem pratik bir ihtiyacı karşılamak hem de deneysel bir ortam sunarak ileride yapılacak çalışmalara açılım sağlamak.
- Türkçe Bilgi Erişim Sistemleri için otomatik gömü üretebilmek.
- Bilgi filtreleme çalışmalarına temel teşkil etmek.

Son yıllarda yapılan önemli çalışmalardan bir tanesi de "Kaşgarlı Mahmut Bilgi Geri Getirim Sistemi" (KMBGS)'dir. 1997 yılında T.C. DPT tarafından desteklenmeye layık görülen KMBGS projesi çerçevesinde, Türkçe tabanlı BES geliştirilmiştir. Bu bazda yapılan çalışmalar üç noktada yoğunlaşmıştır:

1. Tek başına bilgi erişim sistemi,
2. Standart belge modeli,
3. Internet tabanlı arama makinesi.

Birinci kısımda, Türkçe belgeleri saklayan, dizinleyen ve sorgulamaya olanak veren bir boolean bilgi erişim sisteminin gerçekleştirilmesi hedeflenmiştir. Bu kapsamda, Linux platformunda ANSI C dili kullanılarak SMART Bilgi Geri-Getirim Sistemi Türkçe'ye font, gövdeleme ve mesaj düzeyinde yerelleştirilmiş ve üzerinde Türkçe gömü (thesarous) çalışması yapılmıştır. SMART sistemi, 1960'lı yıllardan başlayarak akademik amaçlı sürekli geliştirilen bir laboratuar dizgesi olarak düşünebilir. Bu çalışma bir çok endüstriyel ürünün tasarımına kaynaklık etmiştir; en son örneği HotBoat Arama Makinesi'nde görülebilir.

İkinci kısımda, standart belge modeli oluşturmada makinece anlaşılabilir belge modelleri (belgenin gösterimi ve gerek belge-içi gerekse de belgeler arası ilişkilerin standart bir biçimde tanımlanması) üzerinde çalışılmıştır. Bu kapsamda, World Wide Web (WWW) Konsorsiyumu tarafından desteklenen Resource Description Framework (RDF) modeli temel alınmıştır.

Üçüncü kısımda İnternet kaynaklarına kullanıcıların bilgi ihtiyaçlarını karşılayacak ve aynı zamanda takılar üzerinden sorgulama imkânı veren bir sorgu makinesinin yerelleştirilmesi amaçlandı. Bu kapsamda CNIDR (Center for Networked

Information Discovery and Retrieval) tarafından geliştirilen Isite/Isearch makinası ele alınmıştır (Sever, 2002).

Türkiye'de ilk ticari BG ürünü Boğaziçi Üniversitesi ve Aybim Bilgisayar Tic. Ltd. Şti. tarafından Türk Bilim Vakfı'na 1996 yılında önerilen Gazete Arşiv ve İletişim Dizgesi projesi (GARILDI)'dir (Aybim, 1996).

GARILDI projesi kapsamında, hem sabah gazetesinin hem de diğer basın kuruluşlarının çıkardıkları gazetelerin otomatik olarak arşivlenmesi için resim işleme ve iletişimi sistemi geliştirilmektedir. Basın kuruluşunun kendi ürettiği gazete sayfaları, doğrudan kullandıkları sayfa düzeni programının veri dosyalarından sayfa düzeni, başlık, metin, resim ve alt yazı alanları ayrıştırılarak veri tabanında saklanmaktadır. Diğer gazete sayfaları ise öncelikle tarayıcı aracılığıyla bilgisayar ortamına aktarılarak, sayfa görüntüsü üzerindeki başlık, metin, resim ve alt yazı alanları resim işleme yöntemleri kullanılarak belirlenmekte ve veri tabanına aktarılmaktadır.

Veri tabanındaki yazı ve resimleri adlandırma üç şekilde olmaktadır:

Anahtar Sözcükler: Her bir haber metni için bir ya da çok sayıda (ilgili gazete/dergi adı, tarihi, sayfası, vb.) anahtar sözcük kullanılmaktadır. Bu anahtar sözcüklere göre sorgulama yapılabilmektedir.

Metin Başlıkları: Her bir metnin başlığı veri tabanında bir kayıt alanı olarak saklanmaktadır. Başlık içinde geçen sözcüklere göre sorgulama yapılabilmektedir.

Metin İçeriği: Metin içerisinde geçen bir ya da daha çok sayıda sözcük için ayrıntılı sorgulama yapılabilmektedir.

Projede, fiziksel saklama ortamı olarak CD-ROM kullanılarak, bir nesneye yönelik ve dağıtılmış veritabanı oluşturup, gazete arşivlerinden olası tüm sorguları olanaklı hale getiren bir kullanıcı arabirimi geliştirileceği belirtilmiştir.

GARILDI sisteminin ilk sürümü tamamlanmış ve Sabah gazetesinin internetteki adresine konulmuştur. Tanıtım sayfalarında GARILDI'nın metin içeriğini belirleyen sözcükler, dizin terimleri üzerinden yapılan sorgulamalarda, sorgunun uzun zaman alabileceği belirtilmiştir.

3.4. Doğal Dil İşleme

Doğal dil işleme (Natural Language Processing), ana işlevi doğal bir dili çözümleme, anlama, yorumlama ve üretme olan bilgisayar sistemlerinin tasarımını ve gerçekleştirilmesini konu alan bir bilim ve mühendislik alanıdır. Doğal dil işleme, yapay zekâ (bilgi gösterimi, planlama, akıl yürütme, vb.), biçimsel diller kuramı (dil çözümleme), kuramsal dilbilim ve bilgisayar destekli dilbilim, bilişsel psikoloji gibi çok değişik alanlarda geliştirilmiş kuram, yöntem ve teknolojileri bir araya getirir. 1950 ve 1960'larda yapay zekânın küçük bir alt dalı olarak görülen bu konu, araştırmacıların ve gerçekleştirilen uygulamaların elde ettiği başarılar sonucunda artık bilgisayar bilimlerinin temel bir disiplini olarak kabul edilmektedir. Doğal dil işleme alanındaki araştırmalarda temel amaçlar genellikle şunlar olmuştur:

- Doğal dillerin işlev ve yapısını daha iyi anlamak,
- Bilgisayar ile insanlar arasındaki ara birim olarak doğal dil kullanmak ve bu şekilde bilgisayar ile insanlar arasındaki iletişimi kolaylaştırmak,
- Bilgisayar ile dil çevirisi yapmak.

Doğal dil işleme (Natural Language Processing), önümüzdeki yıllarda insanların bilgisayarlar ile etkileşimlerinde temel bir takım değişiklikler getirmeye aday teknolojilerden biridir. Bilgisayarlar ile doğal dil işleme çok değişik alanlarda uygulama bulmaktadır. Örneğin, çoğumuzun kullandığı sözcük işlemci gibi programlarda bulunan hatalı yazılmış sözcüklerin bulunması ve düzeltilmesi işlevi bu tip uygulamaların en basitlerinden bir tanesidir. Burada, bilgisayar çeşitli nedenlerle (hızlı yazma sırasında hata, doğru yazımı bilmeme, vb.) oluşan yazım hatalarını tespit etmekte ve eğer istenirse kullanıcıya düzeltmede kullanılmak için doğru sözcükler önermektedir. Daha karmaşık bir uygulama olarak bir veri tabanına SQL ile değil de, örneğin Türkçe ile sorgu yöneltmeyi ve sistemin bunu çözümleyerek bir SQL sorgusuna dönüştürüp işledikten sonra sonuçları kullanıcıya vermesini gösterebiliriz. Bilgisayar yardımı ile dilden dile (yarı) otomatik bir şekilde metin çevirisi yapmak, bilgisayar yardımı ile dil öğretmek, bilgisayarların yardımı ile tek veya çok dilli sözcüklere erişmek, doğal dilde cümle ve metin üretmek gibi uygulamaları doğal dil işlemenin en önemli örnekleri

olarak görebiliriz. Çok daha geniş bir bakış açısı ile de konuşma tanıma ve konuşma üretmeyi de - kullandıkları temel teknolojiler oldukça farklı olsa da- bu alan içinde görmek olasıdır. Örneğin, teknolojinin bugün geldiği noktada ABD, Almanya ve Japonya'daki araştırmacılar telefon ile konuşan iki kişinin konuşmalarını anında tanıyıp karşısındaki kişinin diline çeviren, onun anlayabileceği konuşmayı üreten sistemlerin prototiplerini gösterebilmişlerdir. Ancak bu gibi sistemlerin günlük hayatta etkin olarak kullanımları için aradan daha fazla bir sürenin geçmesi gerekecektir. Doğal dil işlemenin bir diğer önemli yönü de, dilbilim kuramlarına deney ortamı yaratarak daha kapsamlı ve çabuk sınanmalarını sağlamaktır. Bu açıdan, doğal dil işleme teknolojisi dilbilimcileri ve bilgisayar bilimcilerini ortak çalışmaya yönlendirmektedir.

Doğal dil işleme ve yakın alanlarda yapılan araştırmalar, bir yanda işlenen dilin yapısal özelliklerinden bağımsız olma iddiasında kuramlar geliştirirken, bir yandan da bunların geniş kapsamlı olarak uygulanması için işlenecek dillere özel kaynakların üzerinde yoğunlaşmaktadır. Ancak şu ana kadar geliştirilen kuramların çoğu genelde İngilizce ve benzeri temel uygulama alanı aldığı için, çeşitli özellikleri ile bu tip dillerden farklı dillere uygulanmalarında sorunlar çıkabilmektedir.

Japonya, ABD, İngiltere, Almanya, Hollanda, Fransa gibi ülkelerde bu teknolojiyi kullanan çeşitli yazılımlar ve bilgisayar sistemleri kullanıcıların hizmetine sunulmuştur. Bilim ve iş alanında her yerde geçerli bir dil olması açısından İngilizce bu gibi ürünlerin en fazla uygulandığı dil olmuştur. Ancak bu teknolojilerin meyvelerini Türkçe'ye uygulamak ve Türkçe'de de araştırma alt yapısı oluşturmak için daha çok çalışma yapılması gerekmektedir (Oflazer ve Bozşahin).

Bu tezin çalışma alanı olan Türkçe metin işleme (text processing) de, doğal dille yazılmış metinleri işleyerek kullanıcıyı istediği bilgiye daha net ve daha çabuk ulaştırma imkânı sunması için oluşturulan teknikleri kapsamaktadır. Yukarıdaki bilgilerden anlaşılacağı gibi, "Metin Madenciliği" ve artık bilgisayar bilimleri içerisinde başlı başına bir disiplin alanı olarak kabul edilen "Doğal Dil İşleme", iç içe girmiş iki alandır.

4. METİN MADENCİLİĞİ

Günümüzde çeşitli alanlardaki bilginin büyük bir kısmı metin belgelerinde yer almaktadır. Bilgi ve belgelerin elektronik ortama aktarılması veya elektronik ortamda saklanılmasından dolayı metinsel veriler hızlı bir şekilde artmaktadır. Bu hızla artan verilere en önemli örnek elektronik postalar ve web sayfalarıdır (Kantardzic, 2003).

Metinsel verilerin yer aldığı veri tabanları kısaca metin veya belge veri tabanları olarak adlandırılır (Mitra ve Acharya, 2003). Bu veri tabanlarında, kitapların elektronik yayınları, dijital kütüphaneler, elektronik postalar, elektronik medya (ortam), teknik veya ticari (mesleki) belgeler, raporlar, araştırma makaleleri, internetteki web sayfaları, vb. şekilde geniş miktarda kullanılabilir bilgi vardır (Saraçoğlu, 2007). Bütün bu geniş miktardaki metinsel bilgiyi saklayan veri tabanlarına belge koleksiyonu adı da verilmektedir. Belge koleksiyonlarından bilgi çıkarımı amacıyla bir veri madenciliği yöntemi olarak metin madenciliği ortaya çıkmıştır.

Metin madenciliği de bir veri madenciliği yöntemidir ve yapısal olmayan metinlerden bilgi keşfi yapılmasını sağlar. Yaygın olarak aynı konuda yazılmış belgeleri bulmak, birbiriyle ilişkili belgeleri bulmak, kavramlar arası ilişkileri keşfetmek için kullanılır. Doğal Dil İşleme (Natural Language Processing), Bilişsel Bilimler (Cognitive Sciences) ve Makine Öğrenmesi (Machine Learning) gibi bilimlerle ortak çalışan bir araştırma alanıdır (Yıldırım ve ark., 2008).

Metin madenciliği, veri madenciliğinin bir alanı olduğuna göre veri madenliği için kullanılan yöntemlerin çoğu metin madenciliği için de geçerlidir. Kullanılan bu yöntemlerden kısaca söz etmek gerekirse, temel olarak beş aşama olarak sınıflandırılabilir:

- Veri seçimi
- Önişleme
- Dönüştürme
- Veri Madenciliği
- Yorumlama/Değerlendirme

Veri Seçimi: Çalışılacak veritabanından veya diğer veri kaynaklarından verilerin seçilerek veri dosyası oluşturulmasıdır.

Ön İşleme: Anormal verilerin kaldırılması, eksik veya hatalı verilerin düzeltilmesi gibi işlemleri içerir.

Dönüştürme: Verilerin kategorize edilmesi, ilgili özelliklerin seçilmesi veya boyut azaltma işlemleridir.

Veri Madenciliği: Bu aşamada uygun bir algoritma seçilerek hazırlanmış verilere uygulanır.

Yorumlama/Değerlendirme: En son aşamada ise keşfedilen bilgiler ve özellikler yorumlanır ve değerlendirilir (Han ve Kamber, 2001).

4.1. Ön İşleme

Metin madenciliğinin en önemli aşamalarından birisi ön işlemedir. Ön işleme, işlenecek olan metinler üzerinde hazırlık işlemleri yapılarak, daha hızlı işlenmesine ve amaca uygun verilere daha çabuk ulaşılmasına olanak tanır. Bu hazırlık için metinler önce cümlelere ve daha sonra da cümleler sözcüklere ayrılır. Bu ayrıştırma işlemi sırasında noktalama işaretleri ve boşluk karakterlerinden faydalanılır. Metinler, cümle dizisi, cümleler ise sözcük dizisi olarak saklanırlar (Türkeş, 2007).

Ön işleme, aşağıdaki aşamaları kapsar:

- Ayrıştırma
    - Noktalama işaretlerinin kaldırılması
    - Tüm kelimelerin tüm harflerinin küçük harf yapılması
- Durma (boş) sözcüklerin atılması
- Gövdeleme

4.1.1. Ayrıştırma

Ön işlemenin ilk aşaması ayrıştırmadır. Ayrıştırma, iki kademeli olarak gerçekleştirilir. İlk önce işlenecek belgedeki noktalama işaretleri temizlenir. Bu işlem yapılırken bulunan noktalama işareti eğer kesme (‘) işareti ise, kesme işaretinden sonra gelen ek, yapım eki olmayacağından yani gövdenin bir parçası olamayacağından dolayı kesme işareti ile birlikte temizlenir (Kesgin, 2007).

İkinci kademe olarak, belge içerisinde geçen tüm kelimelerde bulunan harflerin tamamı küçük harfe dönüştürülür. Böylece sadece ön işleme sırasında değil, tüm metin işleme süreçlerinde karşılaşılabilecek olası karşılaştırma problemleri bertaraf edilmiş olur.

4.1.2. Durma sözcüklerinin atılması

Her dilin yapısına özgü bazı kelimeler vardır. Bilgi erişim terminolojisinde ayrım gücü neredeyse hiç olmayan bu terimler (Sever, 2002) belgeler içerisinde sıkça kullanılmalarına karşın, metinlerin işlenmesinde anlamsal olarak bir ayırıcı özelliğe sahip değildir. İşte bu tip terimlere *boş sözcükler* ya da *durma sözcükleri* adı verilmektedir. Bu terimlerin oluşturduğu listeye de durma kelimeleri listesi (DKL) denir. Durma sözcükleri bir tabloda toplandıktan sonra, ön işleme aşamasındaki belgelerde yer alan kelimeler durma listesindeki sözcüklerle karşılaştırılarak belge durma sözcüklerinden arındırılır.

Türkçe’de kullanılan çok miktarda durma sözcüğü mevcuttur. Bu sözcükler, Adil Alpkoçak, Alp Kut ve Esen Özkarahan tarafından Dokuz Eylül Üniversitesi bünyesinde yapılan “Bilgi Bulma Sistemleri için Otomatik Türkçe Dizinleme Yöntemi” adlı çalışmada belirtilmiştir. Bu çalışmayla ortaya koyulan durma listesi Tablo 4.1’de görülmektedir.

Tablo 4.1. Durma sözcükleri listesi

acaba	biz	genelde	iyi	olsun	zira
acep	bizim	genellikle	içeri	onlar	üstelik
acilen	bizler	gerek	için	oysa	üzere
adeta	bravo	gen	işte	pek	çabucacık
aferin	bu	geç	kadar	pekala	çabucak
ah	bugün	gibi	kah	peki	çabuk
alelacele	bunlar	göre	karşı	sadece	çarçabuk
alelusul	büsbütün	hadi	karşılık	sakın	çok
alenen	bütün	hakkında	karşın	sana	çokça
ama	böyle	hala	katiyen	sen	çoğunlukla
aman	böylece	halbuki	kendi	seyrek	çünkü
ancak	böylelikle	hani	kez	siz	öf
anda	da	hatta	keza	sizler	önce
ansızın	daha	haydi	keşke	sonra	önceden
arada	daima	hayhay	ki	sonradan	önceleri
artık	de	hayır	kim	sonunda	önemli
asla	demek	hem	kuşkusuz	sık	örneğin
ay	demin	hemen	lakin	tam	öte
aynen	denli	henüz	mamafih	tamamen	öyle
aynı	diğer	hep	meğer	tercihan	şayet
ayol	dolayısıyla	hepiniz	mi	tercihen	şekilde
az	dün	hepsi	mu	tesadüf	şey
azıcık	dışarı	her	muhakkak	tüm	şimdi
aşağı	eh	herhalde	mutlaka	usulca	şimdilik
bana	elbet	herkes	mü	uygun	
bari	elbette	herşey	mı	vah	
başka	en	hey	nasıl	vay	
baştan	epey	hiç	ne	ve	
belki	epeyce	hiçbir	neden	veya	
ben	epeydir	iken	nere	ya	
beri	erken	ila	neye	yahu	
besbelli	eskiden	ile	nitekim	yalnız	
bilahare	evet	ileri	niye	yani	
bile	ey	ilgili	niçin	yazık	
bir	eyvah	ilk	o	yaşa	
biraz	eğer	ilkin	of	yine	
birden	fakat	ilkönce	oh	yok	
birdenbire	fazla	inşallah	olan	yukarı	
birbaç	gayet	ise	olarak	yüzden	
birlikte	gene	ister	oldukça	yıl	

4.1.3. Gövdeleme

Türkçe sondan eklemeli bir dil olduğundan dolayı kullanılan kelimelerin yapısı şu şekildedir:

Kök + yapım eki + çekim eki

Örneğin: muhasebe + ci + ler

Kısaca gövdenin tanımını yapacak olursak; Türkçe'de köke yapım eki getirilerek oluşturulan yeni kelimeye gövde ismi verilir (Güzel, 2005). Bir kök veya gövde çekim eki aldığı zaman anlamı değişmese bile yazılışı değiştiği için eski kelimeden farklı yeni bir kelime haline gelir. Metinde hangi kelimenin kaç defa geçtiği sayıldığı zaman, bir kök veya gövdenin çekim eki almış hâli ile yalın hali ve başka çekim eki almış hâlleri farklı kelimeler olarak sayılacaktır. Anlam olarak düşünüldüğü zaman ise farklı yazılışları olan bu kelimeler aynı anlamı ifade etmektedir (Kesgin, 2007).

Yapım eklerinin bir kelimeden yeni bir kelime türetmek için kullanıldıkları düşünüldüğünde, anlamsal olarak kök ile gövdenin birbirinden farklı olduğu gerçeği de ortaya çıkmaktadır. İşte bu anlamsal farklılık, bir kelimenin kökünü bilsek bile kullanıldığı cümleye ve cümle içindeki durumuna göre kelimenin farklı anlamlar alabileceği, bu yüzden de metin madenciliğinde gövdelemenin gerekli olduğunu ortaya koymaktadır.

Yapım ekleri kelimelerin anlamsal olarak farklılaşmasını sağladığı, yani eklendikleri sözcüğe yeni bir anlam kazandırdıkları için yapım eki eklenmiş her sözcük yeni bir gövde olarak kabul edilir. Çekim ekleri ise kelimenin anlamı üzerinde bir değişiklik yapmadan sadece sözcüğün kullanıldığı metne göre uygun hale gelmesini sağlarlar. İşte bu sebeplerden dolayı gövdeleme işlemi bir kelimeden çekim eklerinin çıkarılması olarak tanımlanır.

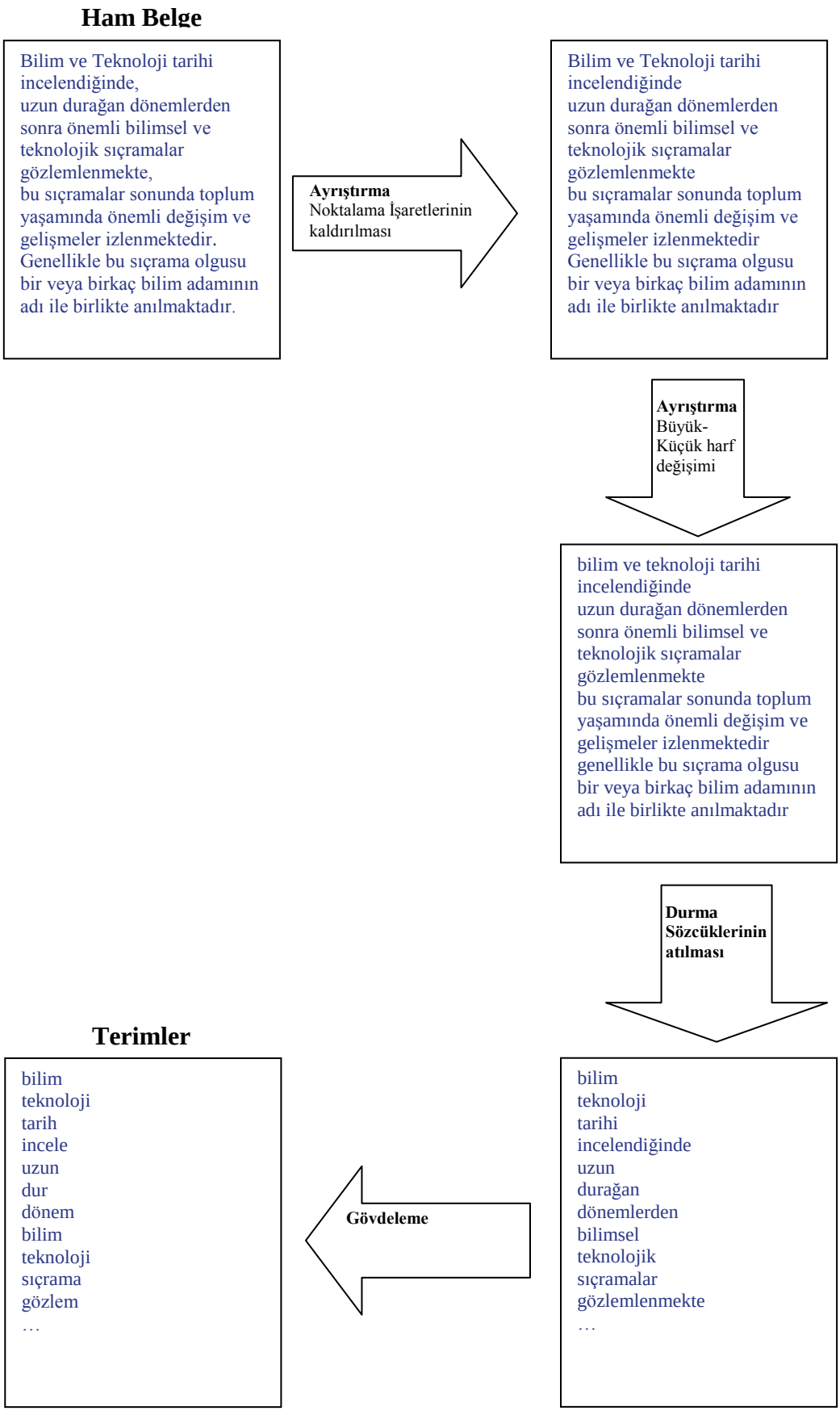
Örnek vermek gerekirse, yapım eki alarak *bilgisayar* kökünden türeyen kelimelerin her birinin farklı bir anlam ifade ettiği görülür.

Bilgisayar : Elektronik bir cihaz
Bilgisayarcı : Bilgisayar satan kişi

Bilgisayarcılık: *Bilgisayarla ilgili uğraş gösteren meslek*

Gövdeleme işlemi dillere göre büyük farklılık göstermektedir. Örneğin İngilizce gibi eklerin kullanımının az olduğu bir dil için yalnızca ekler sözlüğüne bakılarak bir gövdeleyici geliştirmek mümkündür (Porter, 1980). Bu şekilde geliştirilen bir gövdeleyicinin Türkçe gibi çok sayıda ek içeren bir dil için kullanılması mümkün değildir. Türkçe bitişken bir dil olmasından ötürü, eklerin sayısı ve eklenme çeşitleri daha detaylı bir inceleme yapılmasını zorunlu kılar (Jurafsky ve Martin, 2000).

Yukarıda bahsedilen ön işleme süreçleri Şekil 4.1'de örneklenerek şema halinde verilmiştir.



Şekil 4.1. Ön işleme süreçleri

4.1.4. Ön işlemenin önemi

Metin madenciliğinde, işlenecek belge üzerinde yapılan ayrıştırma işlemi (noktalama işaretlerinin kaldırılması ve tüm harflerinin küçük harf yapılması) metnin işlenmesi sırasında karakter bazında oluşabilecek olası kalabalığı önlediği gibi, Türkçe'de aynı amaçla kullanılan bir harfin büyük ve küçük halinin elektronik ortamda farklı ASCII kodlarıyla temsil edilmesinden doğabilecek problemleri de bertaraf etmektedir. Durma sözcüklerinin atılması işlemi ise, tüm kelimeleri terim olarak kabul etmek, yapılacak olan matematiksel modelleme işlemini karmaşık hale getireceğinden ve bu kelimelerin metnin işlenmesinde anlamsal olarak ayıt edici özelliğe sahip olmadıklarından dolayı, bellekte gereksiz yer kaplamalarının işleme sürecinin uzatmaması için kullanılmış bir yöntemdir.

4.1.5. Gövdelemenin önemi

Bir kök veya gövde çekim eki aldığı zaman anlamı değişmese bile yazılışı değiştiği için eski kelimeden farklı yeni bir kelime haline gelir (Kesgin, 2007). Bu yüzden gövdeleme işleminde kelime çekim eklerinden temizlenir. Yapım ekleri ise eklendikleri kelimelere yeni anlam kazandırdıkları için ortaya çıkan kelime artık farklı bir anlama gelen yeni bir kelimedir. İşte bu yüzden gövdelemede, kelime çekim eklerinden arındırılırken, yapım ekleri temizlenmez. Yapım eklerinin anlam olarak farklı yeni bir kelime oluşturduğu düşünüldüğünde, yapım eklerinin atılması gibi bir durumda ortaya çıkacak olan kök ile gövde arasında farklılık olacağı unutulmamalıdır. Gövdeleme işlemi bu anlamsal karışıklığı önlemek ve aynı kökten türetilmiş olsalar bile farklı anlamlara gelen kelimelerin bulunmasını sağlar. Yapılan birçok araştırmaya göre gövde bulma işlemi, bilgi çıkarımının veya metin madenciliğinin verimini artırmaktadır.

4.1.6. Gövdeleme ile ilgili kullanılan metotlar

Gövde bulma işlemleri metnin yazıldığı dille bağlantılı olarak farklılaşmaktadır. Genel olarak gövde bulma ile ilgili metotlar şu şekilde sıralanabilir:

- Eklerin kaldırılması
- Tabloya bakma

- N-Gram yöntemi

Eklerin kaldırılması yaklaşımı iki farklı şekilde kullanılmıştır. Bunlardan birincisi en uzun eşleşmedir (longest match). Bu yöntemde araştırılan kelime kökten başlayıp eşleşen en uzun ekli haline kadar genişletilir. Diğer yöntemde ise, bilinen ekler kaldırılarak gövdeye ulaşılmaya çalışılır. Her iki yöntem için önceden hazırlanmış muhtemel tüm gövdelerin veya eklerin bir listesi bulunmaktadır.

Tabloya bakma yönteminde ise önceden oluşturulmuş bir tabloda muhtemel tüm kelimeler ve bunlara karşılık gelen gövde halleri tutulmaktadır. Bu tablo üzerinden arama yapılarak bir kelimenin gövde hali bulunmuş olur. Tablo 4.2'de örnek bazı kelimeler ve gövde şekilleri verilmiştir.

Tablo 4.2. Örnek kelimeler ve gövdeleri

Term	Stem	Terim	Gövde
engineering	engineer	mühendislik	mühendis
engineered	engineer	mühendisin	mühendis
engineer	engineer	mühendis	mühendis

N-Gram yönteminde ise veritabanında tüm kelime gövdeleri tutulmaktadır. Gövdesi aranan kelime bu veritabanındaki gövdelerle karşılaştırılarak gövde araması yapılır. Bu karşılaştırma işleminde ise n-gram eşleştirme teknikleri diye adlandırılan yaklaşımlar kullanılır (Freund ve Willett, 1982). Bir n-gram, bir kelimeden çıkarılmış n uzunluğunda ardışık karakter dizileridir. Bu yaklaşımdaki temel düşünce ise benzer kelimelerin yüksek oranda aynı n-gram dizilerine sahip olduğudur. Genellikle bu n değeri 2 veya 3 olarak seçilir ve sırasıyla *digrams* ve *trigrams* olarak adlandırılır (Ekmekcioglu ve ark., 1996).

Örneğin BİLGİSAYAR kelimesi için digramlar şunlardır:

B, BI, IL, LG, GI, IS, SA, AY, YA, AR, R

Yine aynı kelime için trigramlar ise şunlardır:

B, *BI, BIL, ILG, LGI, GIS, ISA, SAY, AYA, YAR, AR*, R

Burada '*' boşluğu ifade etmektedir. n karakterli bir kelimede n+1 digram veya n+2 trigram vardır.

Kelimeler arasındaki ilişki ise aşağıdaki formüle göre hesaplanır:

$$S = \frac{2C}{A+B} \tag{4.1}$$

Burada *A*, ilk kelimedeki benzersiz digramların toplam sayısıdır. Yani digramlar oluşturulduktan sonra birbirinin aynı oluşan digramlardan sadece bir tanesi hesaplamaya katılır. *B*, ise ikinci kelimedeki yine benzersiz digramların toplam sayısıdır. *C* ise her iki kelimede ortak bulunan benzersiz digramların toplam sayısıdır (Saraçoğlu, 2007).

N-gram yöntemi, bir metnin hangi dilde yazıldığını bilgisayar tarafından belirleyebilmek amacıyla da kullanılabilir. Bunu kelimeleri oluşturan harflerin yan yana gelme örüntülerine bakarak yapar. Her dilin kelimelerinin 2-gram, 3-gram gibi n-gram kalıpları farklıdır. Bu yaklaşımla bir metnin hangi dille yazıldığı belirlenebilir (Kohonen, 1990).

Türkçe gibi zengin bir biçimbirimsel yapısı olan dillerde gövdeleme işlemi için başvurulan yöntemlerden biri de biçimbirimsel çözümleme kullanmaktır. Biçimbirimsel çözümleyiciler, verilen bir sözcüğün tüm olası kök ve ek birleşimlerini üretirler (Oflazer, 1994).

Biçimbirimsel çözümleyiciler bir sözcüğü kök ve eklerine ayrıştıran sistemlerdir. Türkçe için sık verilen ve özel olarak düşünülmüş bir örnek ve ayrıştırılmış hâli şu şekilde olacaktır.

OSMANLILAŞTIRAMAYABİLECEKLERİMİZDENMİŞSİNİZCESİNE

Bu tek kelimenin kök ve eklerine ayrılmış hâli aşağıdadır. Burada kelimenin kökü olan *osman* Türkçe kurallarına uygun olarak on iki adet ek almıştır (Kesgin, 2007).

OSMAN+LI+LAŞ+TIR+AMA+YABİL+ECEK+LER+İMİZ+DEN+MİŞ+SİNİZ+CESİNE

4.1.7. Gövdeleme konusunda yapılmış çalışmalar

Gövdeleme konusunda Türkçe için ilk çalışma Aydın Köksal tarafından 1981 yılında yapılmıştır (Köksal, 1981). Bu çalışmada kelimenin ilk beş harfi o kelimenin gövdesi olarak kabul edilmiştir. Bu karara birçok kelimenin incelenmesi neticesinde varılmıştır.

Gövdeleme konusunda Solak ve Can'ın yaptığı çalışma şu şekildedir (Solak ve Can, 1994). İlk önce kelime sözlükte aranır, kelime bulunamazsa sonundan bir harf silinir ve ardından gövdeler arasında tekrar aranır. Süreç bu şekilde devam eder.

Bir diğer çalışma ise Dokuz Eylül Üniversitesi bünyesinde yapılan "Bilgi Bulma Sistemleri için Otomatik Türkçe Dizinleme Yöntemi" adlı çalışmadadır (Alpkoçak ve ark., 1995). Bu çalışmada dizinlemede kullanılacak anahtar kelimeleri elde etmek için en fazla eşleşme (maximum match) algoritması kullanılmıştır. Bu yöntemde kelime önce bir sözlükte aranır, bulunamadığı takdirde kelime sonundan bir harf silinerek arama işlemi yeniden yapılır. Süreç bir gövdenin bulunması ya da kelimenin bir harf kalması durumunda sona erer. Bu yöntem uygulama açısından en basit yöntemlerden biri olmakla birlikte, kelime ile ilgisiz gövdeleri sonuç olarak bulma ihtimali vardır. Örneğin, aksamaması kelimesinin gövdesi bu yöntemle bulunmak istendiğinde gövde olarak *aksam* bulunmaktadır. Kelimenin gerçek gövdesi *aksa-mak* fiilinin kökü olan *aksa* kelimesidir.

Gökmen Duran tarafından yapılan, "Gövdebul" adlı Türkçe Gövdeleme algoritmasında (Duran, 1997) kelime soldan sağa incelenmekte ve elde edilen harf katarı sözlükte aranmaktadır. Eğer bir eşleşme bulunur ise kelimenin seçili harf katarı dışında kalan kısmı ek olarak kabul edilerek biçimbirimsel inceleme yapılmaktadır. Yapılan inceleme sonucunda doğru olarak çözümlenebilen sonuçlardan biri gövde olarak seçilmektedir.

Biçimbirimsel çözümleyici kullanarak gövdeleme işlemi Altıntaş ve Can tarafından gerçekleştirilen çalışmada denenmiştir (Altıntaş ve Can, 2002). Bu çalışmada

Oflazer tarafından geliştirilen biçimbirimsel çözümleyici (Oflazer, 1994) kullanılmış ve elde edilen sonuçlar içersinden ortalama gövde uzunluğu ve n-gram yöntemleri ile doğru gövde belirlenmeye çalışılmıştır.

Biçimbirimsel çözümleme sonuçları arasından uygun gövdeyi seçmek için uygulanabilecek yöntemlerden biri de Türkçe için ortalama gövde uzunluğunun belirlenmesi ve olası gövdeler arasında ortalama uzunluğa en yakın olan gövdenin seçilmesidir. Yapılan bir çalışmada, 1 Ocak 1997 ile 12 Eylül 1998 yılları arasında yayınlanan Milliyet Gazetesi'ndeki metinler üzerinde yapılan incelemede, ortalama gövde uzunluğu tüm kelimeler için 4,58, tekil kelimeler için ise 6,58 olarak bulunmuştur (Altıntaş ve Can, 2002).

4.1.8. Biçimbirimsel çözümleme ve sonlu durum makinesi

Türkçe sondan eklemeli bir dil olduğu için ekler açısından oldukça zengindir. Türkçenin biçimbirimsel kuralları karmaşık olmasına karşın, sonlu durum ile tanımlanabilmektedir. Her ek belirli bir anlam taşımakta ve eklendiği kelimeye bu anlamı kazandırmaktadır. Ekler kelimelere eklenirken ünlü ve ünsüz uyum kuralları gereğince bazı harfleri değişir. Ek içerisinde yer alan sesli harfler, kendisinden önce gelen sesli harf ile uyum içinde olmak zorundadır. Bazı belirli durumlarda köklerdeki ve eklerdeki sesli harfler düşerek kelimeden çıkarılır. Benzer şekilde sesiz harfler de kimi zaman değişikliğe uğrayabilir ya da tamamen silinebilir. Bunun yanı sıra yabancı dillerden gelmiş olan kelimeler istisnai durumlar oluşturabilmektedir.

Biçimbirimsel çözümleyiciler tüm bu yukarıda sayılan durumları da göz önünde bulundurarak kelimelerin çözümlemesini yaparlar ve belirli bir düzende sonuçları üretirler. Sonlu durum makinesi ise, biçimbirimsel incelemesi yapılan kelime için tespit edilen eklerin geçerli bir ek olup olmadığını kontrol eder. Eğer ek ya da ekler geçerli ise o zaman olası gövde, mevcut gövdeler arasında aranır ve gövde olup olmadığına karar verilir.

Yukarıda en uzun eşleşme yönteminde bahsedilen ilgisiz gövdeler bulma ihtimali de biçimbirimsel inceleme ve sonlu durum makinesi kullanarak bertaraf edilebilir. Bu durumda;

aksamaması kelimesi için **aksam-*amas*ı** şeklinde yapılan çözümleme sonucunda, sonlu durum makinesi ***amas*ı** ekinin geçerli bir ek olmadığına karar vereceğinden **aksam** olası gövdesi, gövde olarak kabul edilmeyecektir.

Bir sonraki aşamada ise **aksamaması** kelimesi için **aksa-*maması*** şeklinde bir çözümleme ortaya çıkacaktır.

Sonlu durum makinesi ***maması*** ekinin geçerli bir ek olduğuna karar vereceğinden **aksa** olası gövdesi, mevcut gövdeler arasında aranacak ve gövde olarak kabul edilecektir.

4.2. Metinlerin Matematiksel Modeli

Metinsel verilerin bilgisayar sistemleri tarafından işlenebilmesi ve metinler üzerinde veri madenciliği uygulamaları yapılabilmesi için metinlerin matematiksel olarak modellenmesi gereklidir. Bahsi geçen metinsel veriler tez, makale, kitap, e-posta, gazete, dergi, yazışmalar, telefon rehberleri, web sayfaları vs. olabilirler. İşte bu saydığımız tüm metinsel içerikli verileri bünyesinde barındıran yapılara belge, belgeler içerisinde geçen kelime, kelimeler ya da cümleler de terim olarak adlandırılmaktadır.

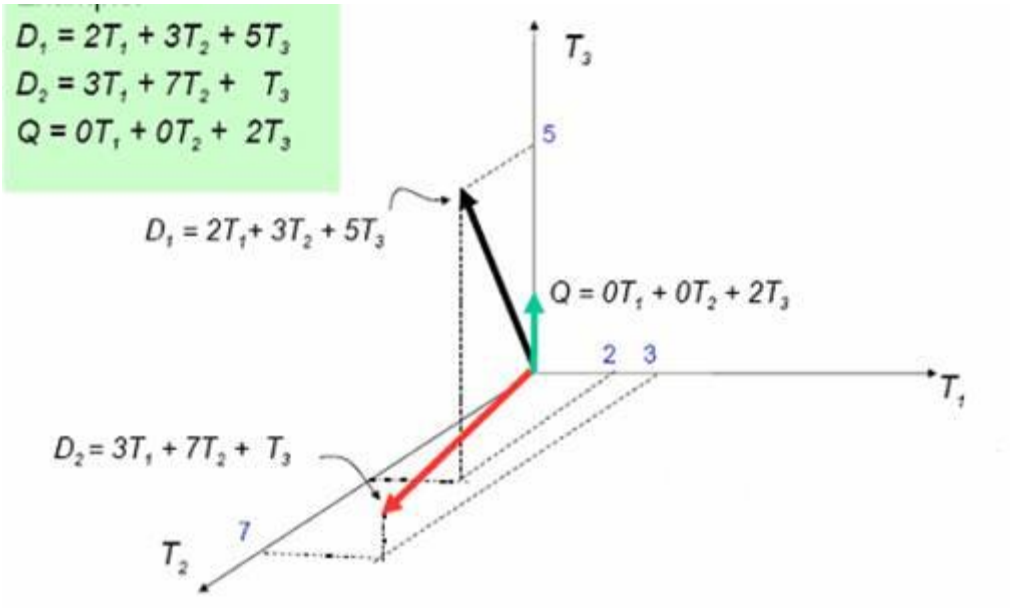
Bir metinsel verinin matematiksel olarak modellenmesinde yukarıda tanımlanan belge ve terim kavramları kullanılmaktadır. Aşağıda kullanılacak olan D belge koleksiyonunu T de terimler kümesini ifade etmektedir.

$$D = d_1, d_2, d_3, ..., d_N \tag{4.2}$$

$$T = t_1, t_2, t_3, ..., t_M \tag{4.3}$$

4.2.1. Vektör uzayı (vector space) modeli

Vektör uzayı modeli, belgeleri oluşturan terimleri ağırlıklandırarak belgeleri birer terim vektörü olarak ele alan modeldir (Eroğlu, 2000). Vektör uzayı modeli, doğal dil belgelerinin çok boyutlu uzayda özel bir anlamını simgelemektedir. Metinlerin matematiksel modelinin çıkarılmasında en yaygın kullanılan model vektör uzayı modelidir.



Şekil 4.2. Vektör uzayı modeli

Dokümanlar Şekil 4.2'de görüldüğü gibi terimlerin vektörleri olarak ifade edilirler. *T*'ler aslında terimleri ifade etmektedirler.

Bu modelde ağırlıklandırma işlemi bir terimin bir belgede kaç defa geçtiği ile ilgilidir. Buna terim sıklığı ya da frekansı *(terf frequency,* ***tf****)* denir. Terim ağırlıklandırma sadece *tf* ile yapılmaz bununla birlikte belge sıklığı *(document frequency,* ***df****)*, ters belge sıklığı *(invers document frequency,* ***idf****)* gibi değerler de kullanılır. Bunlar ileriki bölümlerde detaylı olarak anlatılacaktır.

Vektör uzayı modelinde her belge bir vektör ile temsil edilir. Bir belgenin vektörü oluşturulurken ilk önce tüm terimler gövdelenir ve sırasıyla aşağıdaki adımlar uygulanır:

- Belgenin bulundurduğu terimler ve sıklıkları belirlenir.
- Tüm belgelerdeki terimlerden terim tekrarı olmaksızın bir sözlük oluşturulur.
- Her belge içerdiği terimlere göre bir vektör formuna getirilir.
- Tüm belgelerin vektörleri bir araya getirilerek dizi şeklinde tutulur.

Örneğin,
D1 belgesi: "Bu çalışma çok başarılı bir çalışma"
D2 belgesi: "Az çalışmanın sonucu bu zayıf nottur."

D1 belgesi içinde geçen terimler ve sıklıkları:

bu	çalışma	çok	başarılı	bir
1	2	1	1	1

D2 belgesi içinde geçen terimler ve sıklıkları:

az	çalışma	sonuç	bu	zayıf	not
1	1	1	1	1	1

Belge koleksiyonunda geçen tüm terimlerden oluşturulan terimler sözlüğü:

bu	çalışma	çok	başarılı	bir	az	sonuc	zayıf	not

D1 ve D2 belgelerinin vektör formuna getirilmesi:

	bu	çalışma	çok	başarılı	bir	az	sonuc	zayıf	not
D1:	1	2	1	1	1	0	0	0	0
D2:	1	1	0	0	0	1	1	1	1

Belgelerin vektörleri aşağıdaki gibi oluşur.

D1= (1, 2, 1, 1, 1, 0, 0, 0, 0) (4.4)

D2= (1, 1, 0, 0, 0, 1, 1, 1, 1) (4.5)

4.2.2. Ağırlıklandırma yöntemleri

Bir terimin herhangi bir belgedeki ağırlığı w_i olarak adlandırılmaktadır. Terim ağırlıklandırma için kullanılan yöntemlerden bazılarına burada yer verilecektir.

4.2.2.1. Terim sıklığı

Ağırlık belirleme yöntemlerinin en basiti terim sıklığı (*term frequency*) **tf** olarak bilinen terimin belge içindeki bulunma sayısıdır. Bu yönteme göre,

tf_i, d belgesinde t_i teriminin sıklığı;

w_i, d belgesinde t_i teriminin ağırlığı;

olarak düşünüldüğünde ağırlıklandırma aşağıdaki formül ile hesaplanır.

$$w_i = tf_i \tag{4.6}$$

4.2.2.2. Terim sıklığı ters belge sıklığı

Metin madenciliğinde sıkça kullanılan ağırlıklandırma metotlarından birisi de terim sıklığı ters belge sıklığı (*term frequency invers document frequency*) **tfidf** olarak bilinen yöntemdir. Bu yöntemde ağırlık değerinin hesaplanması sadece terim sıklığına bağlı değildir. Terim sıklığının yanında, bir terimin en az bir kere geçtiği belge sayısını veren belge sıklığı (*document frequency*) **df**, ve bu değerin de kullanılmasıyla hesaplanan ters belge sıklığı (*invers document frequency*) **idf** değerleri de kullanılmaktadır. Ters belge sıklığı aşağıdaki formül ile hesaplanır.

$$idf_i = \log\left(\frac{|D|}{df_i}\right) \tag{4.7}$$

Yukarıdaki formülde kullanılan **log**, 10'luk tabanda logaritmayı, **|D|**, belge koleksiyonundaki toplam belge sayısını, $\mathbf{df_i}$ de ilgili terimin en az bir kere geçtiği belge sayısını yani belge sıklığını göstermektedir.

Tüm bu bilgilere dayanarak terimin belge içindeki ağırlığı aşağıdaki formül ile hesaplanır.

$$w_i = tf_i \cdot idf_i \tag{4.8}$$

4.3. Belgeler Arasında Benzerlik

İki belgenin birbirine benzerliğinin bulunabilmesi için öncelikle belgelerin matematiksel modelinin (vektörlerinin) oluşturulmuş olması gereklidir. Daha sonra oluşturulan bu vektörlerin birbirlerine olan uzaklıkları hesaplanır. En basit yaklaşım ise, iki belgeyle ilgili olan iki vektör arasındaki Öklid uzaklığını bulmaktır (Saraçoğlu, 2007). Bu yöntem benzerlik tabanlı arama ve belge sınıflandırmasında kullanılmaktadır. İki belge vektörü arasındaki Öklid uzaklık aşağıdaki formülle bulunur:

$$Öklid(d_q, d_i) = \sqrt{\left[\sum_{j=1}^{M} (w_{q,j} - w_{i,j})^2\right]} \tag{4.9}$$

Benzerlik tabanlı aramalar için aranan sorgu ifadesinin vektörüyle belge koleksiyonunda yer alan belgelere ait vektörlerin Öklid uzaklıkları hesaplanarak sorgu ile arasında en az uzaklık olan belge en çok benzeyen belge olarak geriye döndürülmektedir.

Belge sınıflandırmalarında ise, Belge koleksiyonunda bulunan tüm belge vektörlerinin birbirleriyle olan Öklid uzaklıkları bulunarak belgelerin aynı sınıftan olup olmadıklarına karar verilen bir dizi işlem yapılmaktadır.

4.4. Metin Sınıflandırma

Metin sınıflandırma, yazılı belgelerin içeriklerine bağlı olarak belirli sınıflara atanması işlemine verilen isimdir. Metin sınıflandırma işlemine örnek olarak bir kaynaktan gelen haberlerin konularına göre ayrıştırılması işlemi verilebilir (Kesgin, 2007). Metin sınıflandırmasında kullanılan farklı yöntemler mevcuttur.

Metin sınıflandırma için uygulanabilecek yöntemlerden ilki bilgi mühendisliği yaklaşımıdır (Joachims, 2002). Bu yöntemde sınıflandırma kuralları uzmanlar tarafından oluşturulur ve yeni gelen belgeler bu kurallara göre sınıflandırılabilir. Sınıflandırma kurallarının uzmanlar tarafından el ile oluşturulması yöntemi zor ve zaman alıcı bir işlem olacaktır. Editörler tarafından sınıflandırma amaçlı sorguların oluşturulması için iki günlük bir zaman dilimi gerekebilmektedir (Jackson ve Moulinier, 2002). Bu sebepten bu yöntem birçok uygulama sahası için çok verimsiz ve elverişsiz

olacaktır. Örneğin, çok fazla sayıda sınıfın olduğu bir durumda bu sınıflar için kuralları belirlemek çok güç olabilir. Bunun yanı sıra, kendi belgelerini sınıflandırmak isteyen bir kullanıcı için uzman bilgisi var olmayacaktır. Ayrıca, sınıfların değişmesi durumunda kuralların gözden geçirilmesi ve tekrar oluşturulması gerekecektir (Kesgin, 2007).

Metin sınıflandırma işlemi için makine öğrenmesi yöntemlerini kullanmak tüm bu bahsedilen sorunların çözümü olabilir. Makine öğrenmesi yöntemlerinden en çok kullanılanlarından biri tümevarımsal öğrenme yöntemidir (Jackson ve Moulinier, 2002). Bu yöntemde bir uygulama sınıflandırma yapmaktan çok, bir öznitelik uzayına bağlı olarak hazırlanmış ve etiketlenmiş verilerden sınıflandırma kurallarını öğrenebilir. Bu tip yöntemlere *denetimli öğrenme* ismi verilmektedir (Kesgin, 2007).

Bu bölümde makine öğrenmesi yöntemlerinden iki tanesine Naive Bayes Sınıflandırıcı ve K-NN (En Yakın Komşu) yöntemlerine yer verilecektir. Yöntemler birbirinden farklı olsa da ortak olan özellikleri ikisinde de sınıflandırmaya tabi tutulacak metinlerin matematiksel modelinin olmasıdır.

4.4.1. Naive Bayes algortiması

Naive Bayes, kolay uygulanabilir olması önemli bir avantajı olan olasılık tabanlı bir sınıflandırma metodudur (Joachims, 1997). Metotta önce tüm eğitim verisindeki belgelerde kullanılan kelimelerden bir sözlük oluşturulur. Daha sonra her bir kelimenin her bir sınıftaki tekrar sayıları (frekansı) bulunur. Buradan yola çıkarak her bir kelimenin her bir sınıfa ait olma olasılıkları hesaplanır. Sınıflandırılması istenen yeni bir belge, önceden oluşturulan sözlükte var olan kelimelerine göre şu şekilde sınıflandırılır:

Bir belgesinin bir sınıfına dâhil olma olasılığı; o sınıfının eğitim setindeki oranıyla, belgenin içindeki her bir kelimenin o sınıfına ait olma olasılıklarının çarpılması suretiyle elde edilir.

Yukarıda özetlendiği üzere bu yöntem metnin olasılıklı bir modelini kullanır. $C=\{c_1,c_2,..,c_m\}$ kategorileri göstermek üzere her bir metinsel belge belirli bir kategoriye atanmıştır. Bir d' belgesinin c_j sınıfında olma ihtimalin $\Pr\ c_j \mid d'$ 'nin hesaplanması aşağıdaki şekildedir. Bayes kuralına göre belgenin en yüksek olasılık $\Pr\ c_j \mid d'$ değerine sahip olduğu sınıf atanacağı sınıfı gösterir.

$$H_{BAYES}\ d' = \arg\max_{c_j \in C} \Pr\ c_j \mid d' \tag{4.10}$$

$\Pr\ c_j \mid d'$ değeri ise aşağıdaki gibi hesaplanır.

$$\Pr\ c_j \mid d' = \frac{\Pr\ c_j \cdot \prod_{i=1}^{|d'|} \Pr\ w_i, c_j}{\sum_{c' \in C} \Pr\ c' \cdot \prod_{i=1}^{|d'|} \Pr\ w_i, c'} \tag{4.11}$$

$\Pr\ c_j$ sınıf olasılığını gösterir ve şu şekilde hesaplanır:

$$\Pr\ c_j = \frac{|c_j|}{\sum_{c' \in C} |c'|} = \frac{|c_j|}{|D|} \tag{4.12}$$

Burada $|c_j|$, c_j sınıfındaki toplam belge sayısını , $|D|$ise eğitim setinin tümünü ifade eden D'deki toplam belge sayısını göstermektedir (tüm belgelerin sayısı).

$\Pr\ w_i, c_j$ ise *i* numaralı kelimenin (terimin) c_j sınıfındaki olasılığını gösterir ve aşağıdaki gibi hesaplanır:

$$\Pr\ w_i, c_j = \frac{1 + TF\ w_i, c_j}{\sum_{w' \in d' \in C_j} TF\ w', c_j} \tag{4.13}$$

$TF\ w, c_j$ ifadesi *w* kelimesinin (teriminin) c_j sınıfında görülme sayısıdır.

Son olarak formül aşağıdaki şekle dönüşmüş olur (Saraçoğlu, 2007).

$$H_{BAYES}\ d' = \arg\max_{c_j \in C} \frac{\Pr\ c_j \cdot \prod_{i=1}^{|d'|} \Pr\ w_i, c_j}{\sum_{c' \in C} \Pr\ c' \cdot \prod_{i=1}^{|d'|} \Pr\ w_i, c'} \quad (4.14)$$

4.4.2. K-NN (En Yakın Komşu) algoritması

Naive Bayes sınıflandırıcılar, sınıflandırma kurallarını eğitim verisini inceleyerek öğrenirler. En yakın komşu algoritmaları ise, eğitim verisini incelemek yerine tümevarımla öğrenirler (Kesgin, 2007).

K-NN (En Yakın Komşu) algoritması sorgu vektörünün en yakın K komşuluktaki vektör ile sınıflandırılmasının bir sonucu olan denetlemeli öğrenme algoritmasıdır. Bu algoritma ile yeni bir vektörü sınıflandırabilmek için doküman vektörü ve eğitim dokümanları vektörleri kullanılır. Bir sorgu örneği verilir, bu sorgu noktasına en yakın K tane eğitim noktası bulunur. Sınıflandırma ise bu K tane nesnenin en fazla olanı ile yapılır. K-NN uygulaması yeni sorgu örneğini sınıflandırmak için kullanılan bir komşuluk sınıflandırma algoritmasıdır (Pilavcılar, 2007).

En yakın komşu tabanlı sınıflandırıcılar, eğitim süresince, eğitim kümesinde yer alan tüm belgeleri belleklerinde tutarlar. Sınıflandırılmak üzere bir B belgesi geldiği zaman, sınıflandırıcı bu belgeye en yakın K adet komşu belgeyi seçer. Daha sonra komşu belgeleri dâhil oldukları sınıflara bakarak bu belgeyi bir veya daha çok sınıfa atarlar (Kesgin, 2007).

K-NN algoritmasının çalışabilmesi için belgeler arasında bir benzerlik ilişkisinin bulunması şarttır. Bu benzerlik ilişkisi, belge vektörleri arasındaki Öklid uzaklıklar bulunarak ya da Kosinüs bezerlikleri hesaplanarak kurulur. Buradan yola çıkarak sınıflandırılacak olan sorgu ya da belgenin eğitim belgelerinden hangilerine ne kadar yakınlıkta olduğu tespit edilir ve istenilirse en yakın belgenin sınıfı ne ise yeni sorgu ya da belge de bu sınıfa atanabilir. Sınıflandırmaya tabi tutulan sorgu ya da belge tek bir sınıfa atanabileceği gibi, kendisine yakınlık bakımından en yakın olan **n** tane belgenin sınıfına da atanabilir.

Bu tez çalışması kapsamında belge sınıflandırma için K-NN algoritması kullanılmıştır. K-NN algoritmasında kullanılacak belge vektörleri, Terim Sıklığı Ters

Belge Sıklığı ağırlıklandırma yöntemiyle oluşturulmuş, vektörler arasındaki benzerlik ilişkileri de Öklid uzaklıklar hesaplanarak kurulmuştur.

4.5. Bölüm Sonucu

Metinsel verilerin işlenmesinde, bilgisayar sistemleri onlara anlamlar yükleyerek bir ayrıştırma işlemi yapamaz. Bilgisayarlar, bunun yerine metinlere sadece şekilsel açıdan bakarak bir değerlendirme yapabilirler. Bu da onların bir makine olduğunun en büyük göstergesidir. İşte çağımızın bu "akıllı" makinelerinin insanların kullandığı doğal diller ile yazılmış metinleri anlayıp yorumlayabilmeleri, yine insanoğlunun onlara aktaracağı bazı dilbilgisi kurallarıyla mümkün olabilmektedir. Bu kuraları makinelere aktarmanın tek yolu ise kodlamadır. Doğal dillerde kullanılan kurallar kodlama yardımıyla bilgisayarlara aktarılırken en çok üzerinde çalışılan ve çok fazla önem arz ettiği düşünülen algoritmalar gövdeleme algoritmaları olmuştur. Bu bölümde metin madenciliğinin en önemli süreçlerinden birisi olan ön işleme ve ön işleme içerisinde önemli bir yere sahip olan gövdeleme üzerinde durulmuş, bu güne kadara yapılan gövdeleme çalışmaları ve geliştirilen bazı gövdeleme algoritmaları hakkında bilgiler verilerek ön işlemenin ve gövdelemenin önemine vurgu yapılmıştır.

Ayrıca metinsel belgelerin işlenmesinde onların matematiksel modellere dönüştürülmesinin gerekliliği ve vektör uzayı modeli anlatılmıştır. Matematiksel olarak modellenecek belgelerin vektörlere dönüştürülmesinde ağırlıklandırmanın nasıl yapılacağı ve ağırlıklandırması yapılan belgelerin de aralarındaki benzerliklerin nasıl bulunacağı konuları da işlenmiştir.

Yukarıda sayılan tüm süreçleri de içerisine alan metinlerin sınıflandırılması konusuna iki farklı yöntemle (Naive Bayes ve K-NN (En Yakın Komşu) Algoritması) değinilmiştir.

Bu bölümün sonunda, ön işleme ve ön işleme içerisinde de gövdelemenin metin madenciliğinin başarısını çok fazla etkileyen süreçler olduğu, doğal dille yazılmış metinlerin bilgisayarlar tarafından anlamsal olarak da kategorize edilebilmesi için kelimelerin mutlaka çekim eklerinden arındırılarak ve yapım ekleriyle kazandıkları yeni anlamaları da muhafaza edilerek kullanılabilmesi için gövdelemenin gerekli olduğu sonuçları çıkarılmıştır. Bununla birlikte veri madenciliğinde metinlerin işlenebilmesi için onların matematiksel modelleme yoluyla ifade edilerek kullanılmaları gerektiği de görülmüştür.

5. DENEYSEL ÇALIŞMA

Tez kapsamında yukarıda bahsedilen Türkçe gövdeleme yöntemlerinin somut bir şekilde incelenmesi düşünülmüş ve bu amaçla bir yazılım geliştirilmiştir. Geliştirilen yazılımda gövdeleme algoritmalarından En Uzun Eşleşme ve Sabit uzunluk algoritmasının 4, 5, 6 ve 7 harfli gövdeleme yöntemleri kullanılmıştır.

Aynı yazılım bünyesinde belge koleksiyonunda bulunan tüm belgelerin matematiksel modeli her iki gövdeleme algoritması için vektör uzayı modeli uygulanarak ayrı ayrı çıkarılarak K-NN (En Yakın Komşu) sınıflandırma algoritması kullanılarak iki gövdeleme algoritmasının sınıflandırma üzerindeki başarı oranları hesaplanmıştır.

5.1. Gövdeleyici ve Sınıfandırıcı Yazılım

Yazılım gerçekleştirilirken kullanılan programlama dili C# .NET, yazılan toplam kod satırı sayısı 3890'dır. Bir bilgi erişim sistemi ve belge sınıflandırıcı özelliği de taşıyan bu deneysel yazılımın çalışması ileriki paragraflarda anlatılmıştır.

Gövdeleme, metin işleme süreçlerinde performansı etkileyen en önemli süreçlerden biridir. Bu çalışmamızda gövdeleme işlemini gerçekleştirmek üzere ham belgeler ve sorgu ifadeleri üzerinde ön işleme aşamalarından, önce *ayrıştırma* süreçleri – noktalama işaretlerinin atılması ve tüm harflerin küçük harfe dönüştürülmesi - , sonra da *durma kelimelerinin atılması* süreci işletilmiş, belgelerdeki ve sorgu ifadesindeki gövdelenecek kelimeler bulunmuştur.

Bu çalışmamızda, Tablo 4.1'de gösterilen 224 kelimelik listeye 11 adet kelime daha (Bkz.Tablo 5.1.) ilave edilmiş ve 235 kelimelik bir durma listesi elde edilerek durma kelimelerinin atılması sürecinde içerisinde kullanılmıştır. Mevcut listeye eklenen kelimeler aşağıdakilerdir:

Tablo 5.1. Eklenen durma sözcükleri listesi

var	benim	onların
güya	senin	yeniden
değil	onun	nerede
kaç	sizin	

Ön işlemenin en son ve en önemli aşaması olan gövdeleme için ise *en uzun eşleşme (Maximum Match) ve sabit uzunluklu gövdeleme* yöntemleri kullanılmıştır. Sabit uzunluklu yöntemde her bir terim için gözde uzunluğu 4, 5, 6 ve 7 harf olan dört farklı gövdeleme uygulaması ayrı ayrı test edilmiştir.

Gövdeleme işlemlerinin dışında aynı yazılım bünyesinde metin sınıflandırma uygulaması da yapılmıştır. Metin sınıflandırması için öncelikle belge koleksiyonumuzda yer alan 1000 adet belgenin her iki gövdeleme algortiması (En Uzun Eşleşme ve Sabit Uzunluk Algoritması) için, ağırlıklandırma yöntemlerinden Terim Sıklığı Ters Belge Sıklığı kullanılarak vektör uzayı modeli ile matematiksel modelleri çıkarılmıştır. Daha sonra belge benzerlikleri Öklid uzaklığı yöntemi kullanılarak tespit edilmiş ve metin sınıflandırması için K-NN (En Yakın Komşu) algoritması kullanılmıştır.

5.1.1. Veri Kümesi

Tez çalışması kapsamında gerçekleştirilen yazılımın En Uzun Eşleşme ve Sabit Uzunluklu Gövdeleme yöntemlerini test etmek için üzerinde çalışacağı belgelerden oluşan bir veri kümesi (VK) oluşturulmuştur. Veri kümesinde,

- İşletme,
- İktisat (Ekonomi),
- Sanat tarihi,
- Osmanlı tarihi,
- Selçuklular tarihi,
- Eğitim bilimleri,
- Sebze - meyve gıdası ve üretimi,
- Fizik,
- Kimya,

- Biyoloji,
- Tıp,
- Arkeoloji,
- Psikoloji,
- Sosyoloji,
- Bilişim.

gibi farklı 15 (on beş) disiplin alanından çok sayıda tez ve makaleden oluşan toplam 1000 adet belge bulunmaktadır. VK'nde bu belgelere ait tüm metin içeriklerinin yer alması yazılımdaki süreçlerin işletilmesinde hantal bir yapı ortaya çıkaracağından, dolayısıyla daha fazla zaman kaybına yol açacağından, veri kümesine belgelerin içeriğini en iyi temsi eden:

- Belge (Tez-Makale) adı,
- Özet,
- Anahtar kelimeler,

kısımları yerleştirilmiştir.

Oluşturulan veri kümesindeki belgeler, ayrıştırma (noktalama işaretlerinin atılması ve tüm harflerin küçük harfe dönüştürülmesi), durma kelimelerinin atılması gibi ön işleme süreçlerinden geçtikten sonra kelime kelime ayrılarak gövdeleme işlemine hazır hale getirilir.

Veri kümesine ait bazı sayısal veriler aşağıdaki Tablo 5.2'de görülmektedir:

Tablo 5.2. Veri kümesi sayısal bilgileri

Belge Sayısı	1000
Toplam Kelime Sayısı	327.636
Gövdelenecek Toplam Ham Terim Sayısı *(Durma kelimeleri atıldıktan sonra)*	266.182
Durma Kelimeleri Sayısı	61.454
Belge Başına Ortalama Kelime Sayısı	327,6
EUEA ile Gövdelenmiş Terim Sayısı	238.496
Belge başına ortalama EUEA ile Gövdelenmiş Terim Sayısı	238,5
EUEA ile Birbirinden Farklı Gövdelenmiş Terim Sayısı	7.065
SUA (4 Harfli) ile Birbirinden Farklı Gövdelenmiş Terim Sayısı	11277
SUA (5 Harfli) ile Birbirinden Farklı Gövdelenmiş Terim Sayısı	18595
SUA (6 Harfli) ile Birbirinden Farklı Gövdelenmiş Terim Sayısı	27158
SUA (7 Harfli) ile Birbirinden Farklı Gövdelenmiş Terim Sayısı	35057

5.1.2. Gövde Kelimeler Sözlüğü

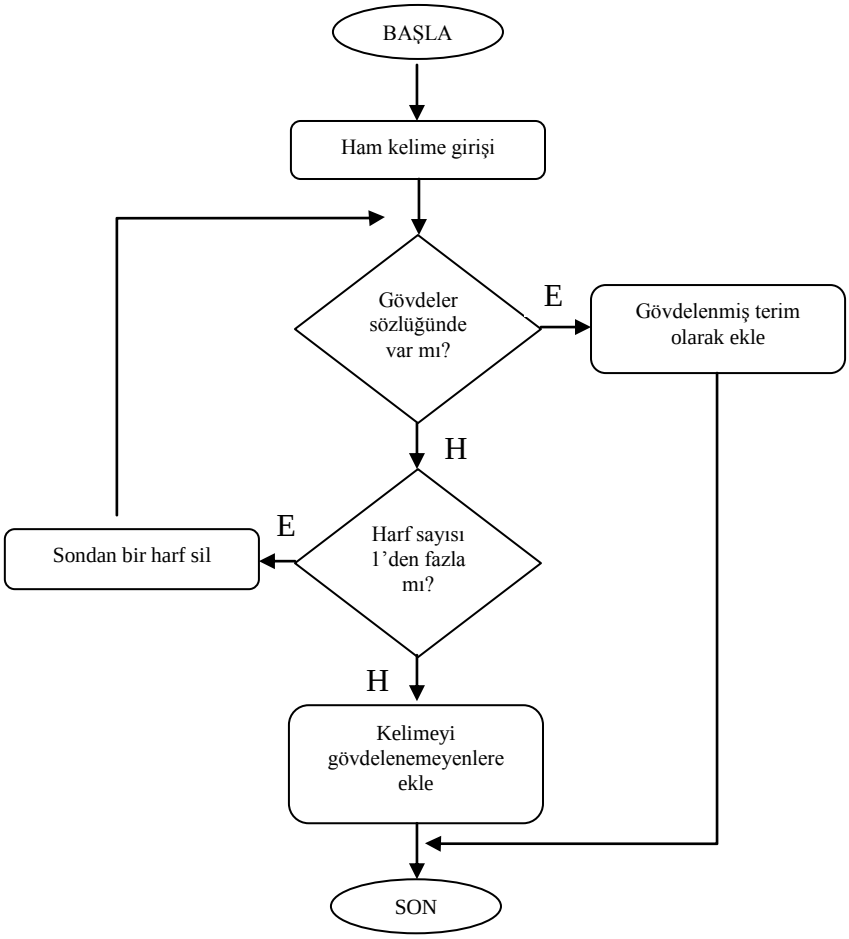
Ayrıştırma (noktalama işaretlerinin atılması ve tüm harflerin küçük harfe dönüştürülmesi), sonra da durma kelimelerinin atılması işlemleri yapıldıktan sonra elde edilen ham terimlerin gövdelemesi işlemine geçilir. En Uzun Eşleşme Algoritması (EUEA) kullanılarak yapılan gövdeleme işlemlerinin tamamında, süreç içerisinde elde edilen verilerin gövde olup olmadığını test etmek için bir Gövde Kelimeler Sözlüğü (GKS) kullanılır. Uygulama içerisinde kullanılan GKS, Türkçe'de yer alan 16.006 adet kelimenin gövde hallerini içermektedir.

5.2. En Uzun Eşleşme Algoritması

Bu yöntemde kelime önce belgeye ait olan kelimelerin gövdelerinin bunduğu bir sözlükte aranır, bulunamadığı takdirde kelime sonundan bir harf silinerek arama işlemi yeniden yapılır. Süreç bir gövdenin bulunması ya da kelimenin bir harf kalması durumunda sona erer. Bu yöntem uygulama açısından en anlaşılır yöntemlerden biri olmakla birlikte, gövdeleme sırasında her bir terim için defalarca sözlükte arama işlemi

yapıldığı için zaman açısından da kötü bir performansa sahiptir. Ayrıca kelime ile ilgisiz gövdeleri sonuç olarak bulma ihtimali de vardır. Örneğin, *aksamaması* kelimesinin gövdesi bu yöntemle bulunmak istendiğinde gövde olarak *aksam* bulunmaktadır. Kelimenin gerçek gövdesi *aksa-mak* fiilinin kökü olan *aksa* kelimesidir.

En Uzun Eşleşme Algoritmasının akış diyagramı Şekil 5.1'de gösterilmiştir.



Şekil 5.1. En uzun eşleşme algoritması akış diyagramı

Yukarıda kullanılan gövdeleme algoritmasında işlenen metinde geçen bir kelimenin gövdeleme aşamaları aşağıdaki gibi gerçekleşmiştir.

En Uzun Eşleşme Algoritması Nasıl Çalışır?

Örnek Ham Kelime: incelendiğinde

- incelendiğinde → gövdeler sözlüğünde var mı? →**YOK**

  → harf sayısı 1'den fazla mı? → **EVET** → sondan bir harf sil

- incelendiğind → gövdeler sözlüğünde var mı? →**YOK**

  → harf sayısı 1'den fazla mı? → **EVET** → sondan bir harf sil

- incelendiğin → gövdeler sözlüğünde var mı? →**YOK**

  → harf sayısı 1'den fazla mı? → **EVET** → sondan bir harf sil

- incelendiği → gövdeler sözlüğünde var mı? →**YOK**

  → harf sayısı 1'den fazla mı? → **EVET** → sondan bir harf sil

- incelendiğ → gövdeler sözlüğünde var mı? →**YOK**

  → harf sayısı 1'den fazla mı? → **EVET** → sondan bir harf sil

- incelendi → gövdeler sözlüğünde var mı? →**YOK**

  → harf sayısı 1'den fazla mı? → **EVET** → sondan bir harf sil

- incelend → gövdeler sözlüğünde var mı? →**YOK**

  → harf sayısı 1'den fazla mı? → **EVET** → sondan bir harf sil

- incelen → gövdeler sözlüğünde var mı? →**YOK**

  → harf sayısı 1'den fazla mı? → **EVET** → sondan bir harf sil

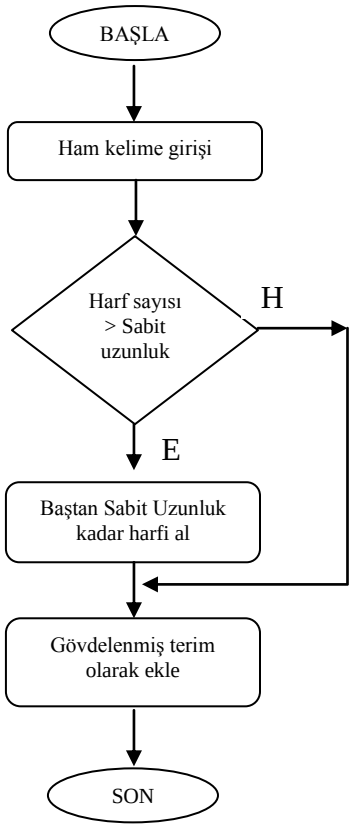
- incele → gövdeler sözlüğünde var mı? →**VAR** → ***gövde* = *incele***

incelendiğinde kelimesinin gövdesi ***incele*** olarak bulunur.

5.3. Sabit Uzunluk Algoritması

Sabit Uzunluk Algoritması (SUA)'nda gövdelenecek her bir terimin gövdesi bulunurken o terime ait sabit sayıdaki harf gövde olarak kabul edilir. Bu yöntemi kullanan araştırmacılar yaptıkları çalışmalarda farklı sayıda harfi gövde kabul ederek çalışmalar yapmışlardır. Aşağıda bahsedilecek olan deneysel çalışmada, bu gövdeleme yöntemi 4, 5, 6 ve 7 harfli gövdeleme yapılarak dört farklı şekilde denenerek bilgi erişimi üzerindeki performansları incelenmiştir.

Sabit Uzunluk Algoritmasının akış diyagramı Şekil 5.2'de gösterilmiştir.



Şekil 5.2. Sabit uzunluk algoritması akış diyagramı

5.4. Gövdeleme Algortimalarının Gerçekleştirilmesi

Eğer veri kümesi üzerinde bir belge araması yapılacak ise önce sorgu ifadesi girilir ve hangi gövdeleme algoritması (EUEA: En Uzun Eşleşme Algoritması, SUA: Sabit Uzunluk Algoritması) kullanılarak bilgi erişimi yapılmak isteniyorsa o seçilir, ardından butonuna basılır.

Program ilk olarak kullanıcı tarafından girilen sorgu ifadesini sırasıyla metin ön işleme süreçlerine alır. Önce sorgu ifadesi içerinse geçen tüm noktalama işaretlerini temizler. Daha sonra sorgu ifadesinin tamamında geçen tüm harfleri küçük harfe dönüştürür. Böylece sadece ön işleme sırasında değil, tüm metin işleme süreçlerinde karşılaşılabilecek olası karşılaştırma problemleri bertaraf edilmiş olur. Bu işlemlerden sonra sorgu metni kelimelere ayrılır. Ön işlemenin diğer önemli bir adımı da durma kelimelerinin atılması işlemidir. Durma kelimeleri (Bkz. Tablo 4.1) de sorgu

ifadesinden çıkarıldıktan sonra gövdelenecek ham terimler elde edilmiş olur. Aşağıda Tablo 5.3'te örnek bir sorgu ifadesi ve işleyen süreçler gösterilmiştir:

Tablo 5.3. Sorgu ifadesinin işlenmesi

Örnek Sorgu:	Küreselleşme, kurumsallaşma ilişkisinin incelenmesi ve teknoloji kullanımı.
Noktalama işaretlerinin atılması:	Küreselleşme kurumsallaşma ilişkisinin incelenmesi ve teknoloji kullanımı
Tüm harflerin küçük harfe dönüştürülmesi:	küreselleşme kurumsallaşma ilişkisinin incelenmesi ve teknoloji kullanımı
Kelimelere ayrılması:	küreselleşme kurumsallaşma ilişkisinin incelenmesi ve teknoloji kullanımı
Durma kelimelerinin atılması sonucu ham terimler:	küreselleşme kurumsallaşma ilişkisinin incelenmesi teknoloji kullanımı

Yukarıda bahsedilen süreçler hem En Uzun Eşleşme hem de Sabit Uzunluklu Gövdeleme yöntemleri için çalışmaktadır.

EUEA seçilerek arama işlemi başlatıldığında önceki süreçlerde elde edilen ham terimler, yukarıda bahsedilen *en uzun eşleşme algoritması* kullanılarak gövdelenir ve her ham terimin bir gövdelenmiş terim hali oluşturulur. Aşağıdaki Tablo 5.4'te örnek sorguya ait ham terimlerin En Uzun Eşleşme algoritmasına göre gövdelenmiş halleri bulunmaktadır.

Tablo 5.4. Ham terimlerin EUEA ile gövdelenmesi

Ham Terimler	EUEA ile Gövdelenmiş Terimler
küreselleşme	küresel
kurumsallaşma	kurumsal
ilişkisinin	ilişki
incelenmesi	incele
teknoloji	teknoloji
kullanımı	kullan

Yukarıda sorgu ifadesi için işletilen süreçlerin tamamı (ayrıştırma, durma kelimelerinin atılması ve gövdeleme) veri kümesinde yer alan tüm belgeler için de aynen işletilir. Tabiî ki burada sorgu ifadesinden farklı olarak bir belgeye ait olan bir değil üç grup veri (belge adı, özet, anahtar kelimeler) görüyoruz. Süreçler başlatılmadan önce bu üç grup veri birleştirilerek bir belgeye ait olan tek bir veri kümesi oluşturulur. Bu veri kümeleri oluşturulurken bir belgeye ait olacak kelime sayısı en fazla 1000 kelime ile sınırlandırılmıştır. Her belge için kelime sayısının maksimum 1000, belge sayısının da 1000 olduğu düşünüldüğünde veri kümesinde yer alan 1000 x 1000 toplam 1.000.000 (en fazla) kelime üzerinde işlem yapabilecek bir bilgi erişim sistemi tasarlanmıştır. Yani yukarıdaki 7 (yedi) kelimeden oluşan sorgu örneği için işletilen tüm süreçler veri kümesi üzerinde yer alabilecek (en fazla) 1.000.000 kelime için işletilebilecektir.

Programın kullandığı bir diğer gövdeleme algoritması Sabit Uzunluk Algoritması (SUA)‘dır. Bu yöntemin işletilmesinde de gövdeleme işlemine geçilmeden, ilk olarak kullanıcı tarafından girilen sorgu ifadesi sırasıyla metin ön işleme süreçlerine alınır. Önce sorgu ifadesi içerisinde geçen tüm noktalama işaretleri temizlenir. Daha sonra sorgu ifadesinin tamamında geçen tüm harfler küçük harfe dönüştürülür. Ardından kelimelere ayrılan sorgu ifadesi içerisinde geçen durma kelimeleri temizlenir. Böylece gövdelenmek üzere ham terimler elde edilir.

SUA ile, gövdelemeye hazır olan ham terimlerin ilk 4, 5, 6 ve 7 harfleri alınır ve aynı yöntemle dört farklı gövdeleme işlemi yapılarak gövdelenmiş terimler listesi (GTL) elde edilir (Bkz. Tablo 5.5).

Tablo 5.5. Ham terimlerin SUA ile gövdelenmesi

Ham Terimler	4 Harfli Gövdeler	5 Harfli Gövdeler	6 Harfli Gövdeler	7 Harfli Gövdeler
küreselleşme	küre	küres	kürese	küresel
kurumsallaşma	kuru	kurum	kurums	kurumsa
ilişkisinin	iliş	ilişk	ilişki	ilişkis
incelenmesi	ince	incel	incele	incelen
teknoloji	tekn	tekno	teknol	teknolo
kullanımı	kull	kulla	kullan	kullanı

5.4.1. Örnek uygulamalar

Oluşturulan gövdeleyici yazılımın VK üzerinde bilgi erişimi yapabilmesi için örnek sorgulamalar yapılmış ve bu örnek sorgular hem EUEA hem de SUA ile işletilerek deney sonuçları aşağıda verilmiştir.

- **İlk Sorgulama:**

Sorgu İfadesi: "Turizm sektöründe otellerin sorunları üzerine çözümler var mı?"

Tablo 5.6. İlk sorgu ifadesinin işlenme süreçleri

Sorgu:	Turizm sektöründe otellerin sorunları üzerine çözümler var mı?
Noktalama işaretlerinin atılması:	Turizm sektöründe otellerin sorunları üzerine çözümler var mı
Tüm harflerin küçük harfe dönüştürülmesi:	turizm sektöründe otellerin sorunları üzerine çözümler var mı
Kelimelere ayrılması:	turizm sektöründe otellerin sorunları üzerine çözümler var mı
Durma kelimelerinin atılması sonucu ham terimler:	turizm sektöründe otellerin sorunları üzerine çözümler

En Uzun Eşleşme Algoritmasının Uygulanması

EUEA hem sorgudaki ham terimlerin hem de veri kümesindeki ham terimlerin gövdelenmesi için kullanılır. Bu gövdeleme işlemleri esnasında bir çok kez GKS'ne bakılarak o anki verinin gövde olup olmadığı sorgulanır. Eğer veri GKS'de bir kelime ile eşleşirse GTL'ne eklenir.

Tablo 5.7. İlk sorgu ifadesi için GKS ile karşılaştırma sayıları

Durma kelimelerinin atılması sonucu ham terimler	**EUEA ile gövdelenmiş terimler**	**GKS ile karşılaştırma sayıları**
turizm	turizm	0
sektöründe	sektör	5
otellerin	otel	6
sorunları	sorun	5
üzerine	üz	6
çözümler	çözümle	2
	Toplam:	**24**
Terim Başına Ortalama GKS Karşılaştırma Sayısı:		**24 / 6= 4**

Yukarıdaki Tablo 5.7'de görüldüğü gibi sadece sorgu ifadesindeki ham terimlerin gövdelenmesinde 24 (yirmi dört) kez GKS ile karşılaştırma işlemi yapılmıştır. GKS'nde 16.006 adet gövde kelime olduğu ve her verinin bu gövde kelimelerin her biri ile karşılaştırılacağı düşünülürse, sadece sorgu ifadesi için;

24 x 16.006 = 384.144 (5.1)

adet karşılaştırma işlemi yapılmıştır.

Benzer şekilde veri kümesinde yer alan **266.182** adet gövdelenmemiş ham terimin her birisi için de GKS ile karşılaştırmalar yapılmıştır. Sorgu ifadesinin gövdelenmesi ile hesaplanan terim başına ortalama GKS karşılaştırma sayısı temel alındığında;

Ham terim sayısı: **266.182**
Terim başına GKS ile ortalama karşılaştırma sayısı: **4**
GKS'ndeki kelime sayısı: **16.006**

Yukarıdaki sayısal veriler ışığında veri kümesinin gövdelenmesi sırasında;

266.182 x 4 x 16.006 = 17.042.036.368 (5.2)

adet karşılaştırma işlemi yapılmıştır.

Sonuç olarak ilk sorguda GKS ile yapılan toplam karşılaştırma sayısı;

17.042.036.368 + 384.144 = 17.042.420.512 (5.3)

olarak bulunur.

Sorgudan elde edilen gövdelenmiş terimler ile veri kümesinden elde edilen gövdelenmiş terimler karşılaştırılmıştır.

Veri kümesindeki gövdelenmiş terim sayısı: **238.496**
Sorgudaki gövdelenmiş terim sayısı: **6**

Sorgudaki terimler ile veri kümesindeki terimlerin karşılaştırma sayısı

238.496 x 6 = 1.430.976 (5.4)

EUEA'nda toplam karşılaştırma sayısı: **17.043.851.488**

Yukarıdaki bazı sayısal bilgileri verdikten sonra sorgu sonucunda meydana gelen bilgi erişim çıktılarına bakalım:

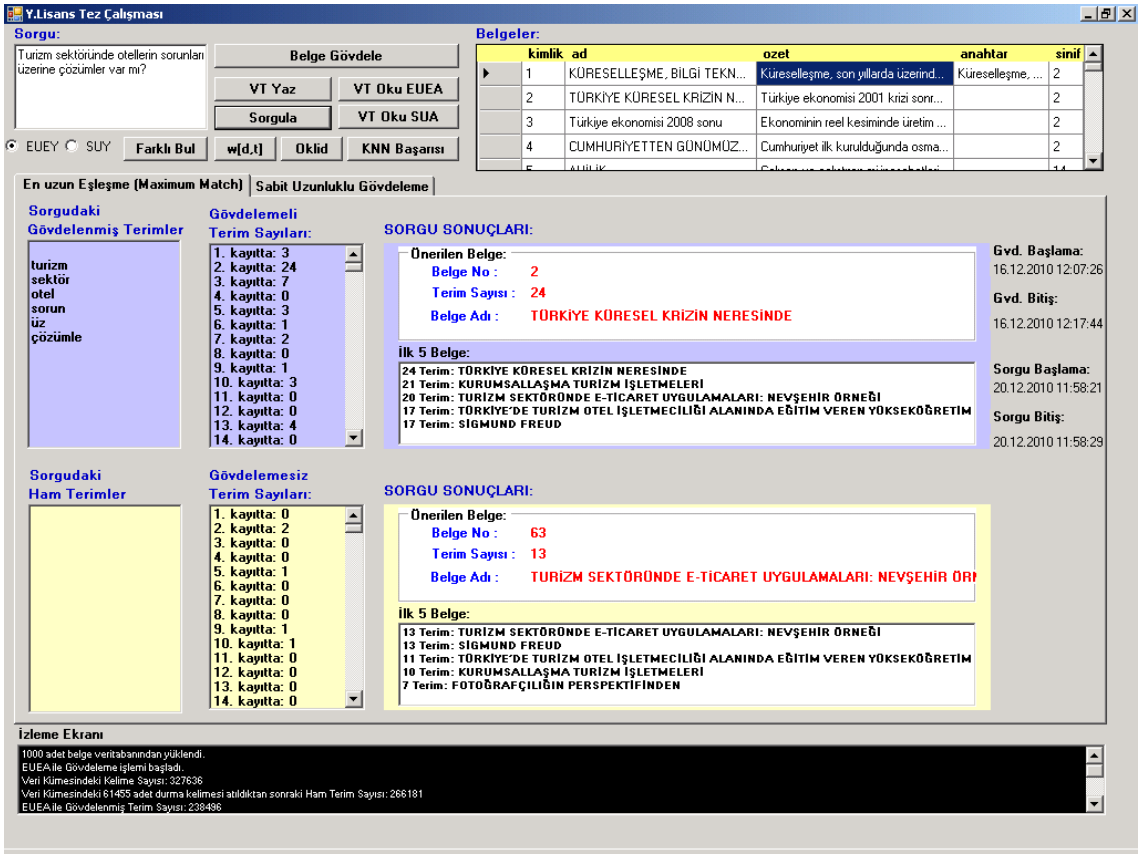
Sorgudaki gövdelenmiş terimler ile veri kümesindeki gövdelenmiş terimlerin karşılaştırılması sonucunda en fazla terimin eşleştiği ilk beş belge ile ilgili bilgiler aşağıda Tablo 5.8'de verilmiştir.

Tablo 5.8. İlk sorgudan EUEA ile elde edilen bilgi erişim çıktıları

Eşleşen Gövdelenmiş Terim Sayısı	Belge No	Belge adı	Sorguyla İlgisi Var mı?
24	2	Türkiye Küresel Krizin Neresinde?	Evet
21	205	Kurumsallaşma, Turizm İşletmeleri	Evet
20	63	Turizm Sektöründe E-Ticaret Uygulamaları: Nevşehir örneği	Evet
17	43	Türkiye'de Turizm Otel İşletmeciliği Alanında Eğitim Veren Yükseköğretim Kuruluşlarındaki Eğitimcilerin Turizm Mesleki Eğitiminin Etiksel Açıdan İncelenmesine Yönelik Bir Alan Araştırması	Evet
17	973	Sigmund Freud	Hayır
Sorguyla İlgili Kayıtlarda Eşleşen Toplam Terim Sayısı : 82			

EUEA ile ilk sorgulamanın yapılması ve bilgi erişim çıktılarının elde edilmesi **3 saat 49 dakika 13 saniye** sürede tamamlanmıştır.

Uygulamanın ekran çıktısı Şekil 5.3'te görülmektedir.



Şekil 5.3. İlk sorgu için EUEA ekran çıktısı

Sabit Uzunluk Algoritmasının Uygulanması

SUA hem sorgudaki ham terimlerin hem de veri kümesindeki ham terimlerin gövdelenmesi için kullanılır. Bu gövdeleme işlemleri esnasında GKS kullanılmaz. Bu gövdeleme yönteminde ham terimlerin başından itibaren belli sayıda karakteri alınarak ham terimin gövdelenmiş hali elde edilir. Bu yöntem kullanılırken sorgudaki gövde uzunluğu 4, 5, 6 ve 7 harfli olacak şekilde dört farklı gövdeleme yapılmış ve bunun akabinde dört farklı erişim çıktısı elde edilmiştir.

Tablo 5.9. İlk sorgudaki ham terimlerin SUA ile gövdelenmesi

Ham Terimler	4 Harfli Gövdeler	5 Harfli Gövdeler	6 Harfli Gövdeler	7 Harfli Gövdeler
turizm	turi	turiz	turizm	turizm
sektöründe	sekt	sektö	sektör	sektörü
otellerin	otel	otell	otelle	oteller
sorunları	soru	sorun	sorunl	sorunla
üzerine	üzer	üzeri	üzerin	üzerine
çözümler	çözü	çözüm	çözüml	çözümle

Yukarıdaki Tablo 5.9'da sorgu ifadesi içinde geçen terimlerin SUA'na göre (4, 5, 6 ve 7 harfli) gövdelenmiş halleri görülmektedir. Aynı gövdeleme yöntemi kullanılarak veri kümesinde bulunan **266.182** adet ham terim de gövdelenerek, aynı sayıda gövdelenmiş terim elde edilmiştir.

Kabul edilen her bir gövde uzunluğu (4, 5, 6 ve 7) için sorgudan elde edilen gövdelenmiş terimler ile veri kümesinden elde edilen gövdelenmiş terimler karşılaştırılmıştır.

Veri kümesindeki gövdelenmiş terim sayısı: **266.182**
Sorgudaki gövdelenmiş terim sayısı: **6**

Bir gövde uzunluğu için yapılan karşılaştırma sayısı:

266.182 x 6 = 1.597.092 (5.5)

Dört gövde uzunluğu için yapılan toplam karşılaştırma:

1.597.092 x 4 = 6.388.368 (5.6)

Yukarıdaki bazı sayısal bilgileri verdikten sonra sorgu sonucunda meydana gelen bilgi erişim çıktılarına bakalım:

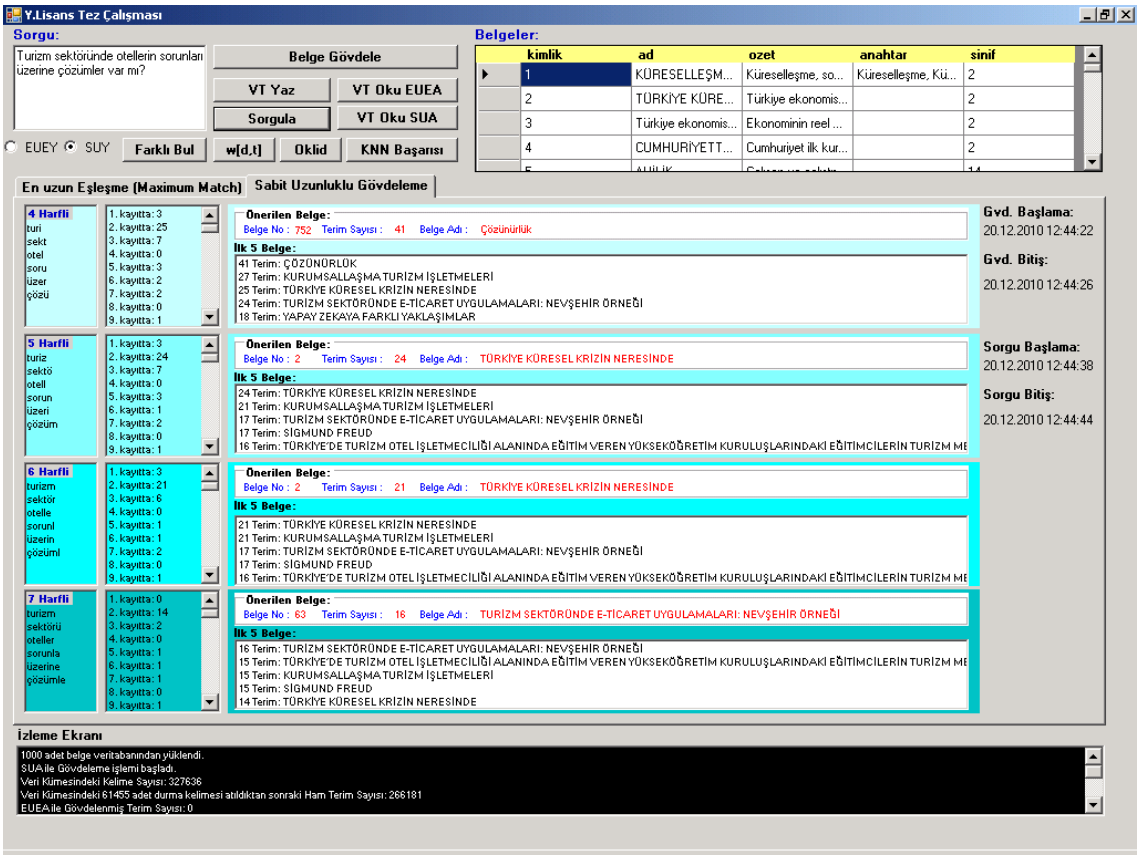
Sorgudaki gövdelenmiş terimler ile veri kümesindeki gövdelenmiş terimlerin karşılaştırılması sonucunda en fazla terimin eşleştiği ilk beş belge ile ilgili bilgiler aşağıda Tablo 5.10'da verilmiştir.

Tablo 5.10. İlk sorgudan SUA ile elde edilen bilgi erişim çıktıları

ÇIKTI SIRASI	4 HARF			5 HARF			6 HARF			7 HARF		
	Terim Sayısı	Belge No	Sorgulya ilgisi var mı?	Terim Sayısı	Belge No	Sorgulya ilgisi var mı?	Terim Sayısı	Belge No	Sorgulya ilgisi var mı?	Terim Sayısı	Belge No	Sorgulya ilgisi var mı?
1	41	752	Hayır	**24**	2	Evet	**21**	2	Evet	**16**	63	Evet
2	**27**	205	Evet	**21**	205	Evet	**21**	205	Evet	**15**	43	Evet
3	**25**	2	Evet	**17**	63	Evet	**17**	63	Evet	**15**	205	Evet
4	**24**	63	Evet	17	973	Hayır	17	973	Hayır	15	973	Hayır
5	18	996	Hayır	**16**	43	Evet	**16**	43	Evet	**14**	2	Evet
İlgili kayıtlarda eşleşen toplam terim sayısı	**76**			**78**			**75**			**60**		

SUA ile ilk sorgulamanın yapılması ve bilgi erişim çıktılarının elde edilmesi **6 saniye** sürede tamamlanmıştır.

Uygulamanın ekran çıktısı Şekil 5.4'te görülmektedir.



Şekil 5.4. İlk sorgu için SUA ekran çıktısı

EUEA ve SUA Erişim Çıktılarının Değerlendirmesi

EUEA'nın toplam eşleşen terim sayısı 82 olduğuna göre;

Sabit Uzunluk Algoritmasının, En Uzun Eşleşme Algoritmasına yakınlaşma değerleri aşağıdaki gibi hesaplanır

4 Harfi gövdeleme için: 76 / 82 = 0,93 (%93)

5 Harfi gövdeleme için: 78 / 82 = 0,95 (%95)

6 Harfi gövdeleme için: 75 / 82 = 0,91 (%91)

7 Harfi gövdeleme için: 60 / 82 = 0,73 (%73)

En Uzun Eşleşme Algoritmasında elde edilen çıktılara göre, **2** numaralı **"Türkiye Küresel Krizin Neresinde"** adlı belge, en fazla eşleşen terime sahip, yani aranan konuya en uygun belgedir. Sabit Uzunluk Algoritması sonuçlarına göre de en yakın sonucu (%95) üreten 5 harfli gövdeleme yönteminde de aynı belgenin en uygun belge olduğunu görüyoruz.

- **İkinci Sorgulama:**

Sorgu İfadesi: "Nükleer enerji gerçekten zararlı olsaydı, ülkemizde uygulanmazdı. "

Tablo 5.11. İkinci sorgu ifadesinin işlenme süreçleri

Sorgu:	Nükleer enerji gerçekten zararlı olsaydı, ülkemizde uygulanmazdı.
Noktalama işaretlerinin atılması:	Nükleer enerji gerçekten zararlı olsaydı ülkemizde uygulanmazdı
Tüm harflerin küçük harfe dönüştürülmesi:	nükleer enerji gerçekten zararlı olsaydı ülkemizde uygulanmazdı
Kelimelere ayrılması:	nükleer enerji gerçekten zararlı olsaydı ülkemizde uygulanmazdı
Durma kelimelerinin atılması sonucu ham terimler:	nükleer enerji gerçekten zararlı olsaydı ülkemizde uygulanmazdı

En Uzun Eşleşme Algoritmasının Uygulanması

İlk sorgu için uygulananlar bu sorgu için de aynen uygulandığı için detaya tekrar girilmeyecektir.

Tablo 5.12. İkinci sorgu ifadesi için GKS ile karşılaştırma sayıları

Durma kelimelerinin atılması sonucu ham terimler	**EUEA ile Gövdelenmiş terimler**	**GKS ile karşılaştırma sayıları**
nükleer	nükleer	0
enerji	enerji	0
gerçekten	gerçekten	0
zararlı	zarar	3
olsaydı	ol	6
ülkemizde	ülke	6
uygulanmazdı	uygula	7
	Toplam:	**22**
Terim Başına Ortalama GKS Karşılaştırma Sayısı:		**22 / 7= 3,14**

Yukarıdaki Tablo 5.12‘de görüldüğü gibi sadece sorgu ifadesindeki ham terimlerin gövdelenmesinde 22 (yirmi iki) kez GKS ile karşılaştırma işlemi yapılmıştır. GKS’de 16.006 adet gövde kelime olduğu ve her verinin bu gövde kelimelerin her biri ile karşılaştırılacağı düşünülürse, sadece sorgu ifadesi için;

22 x 16.006 = 352.132 (5.7)

adet karşılaştırma işlemi yapılmıştır.

Veri kümesi için yapılan GKS ile karşılaştırma sayıları ilk sorguda hesaplanarak **17.042.036.368** olarak bulunmuştu.

Sonuç olarak ikinci sorguda toplam yapılan GKS ile karşılaştırma sayısı;

17.042.036.368 + 352.132 = 17.042.388.500 (5.8)

olarak bulunur.

Sorgudan elde edilen gövdelenmiş terimler ile veri kümesinden elde edilen gövdelenmiş terimler karşılaştırılmıştır.

Veri kümesindeki gövdelenmiş terim sayısı: **238.496**
Sorgudaki gövdelenmiş terim sayısı: **7**

Sorgudaki terimler ile veri kümesindeki terimlerin karşılaştırma sayısı

238.496 x 7 = 1.669.472 (5.9)

EUEA’nda toplam karşılaştırma sayısı: **17.044.057.972**

Yukarıdaki bazı sayısal bilgileri verdikten sonra sorgu sonucunda meydana gelen bilgi erişim çıktılarına bakalım:

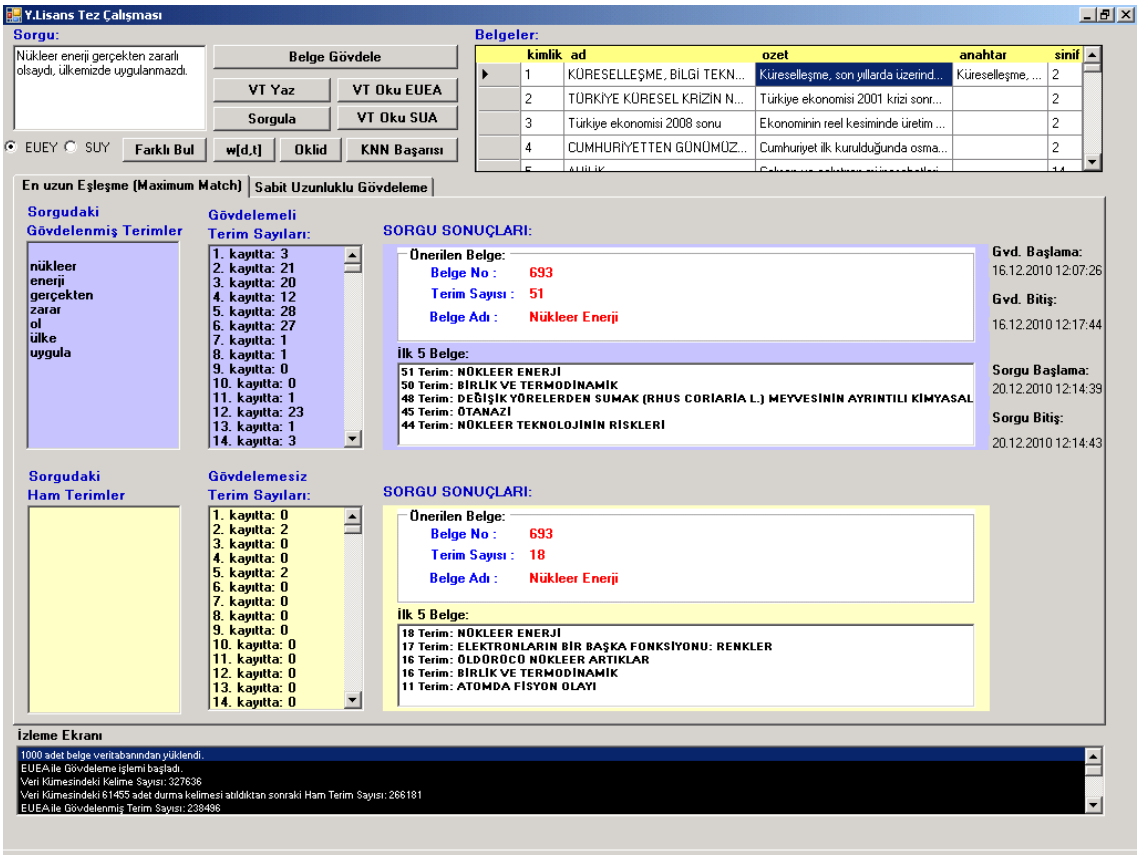
Sorgudaki gövdelenmiş terimler ile veri kümesindeki gövdelenmiş terimlerin karşılaştırılması sonucunda en fazla terimin eşleştiği ilk beş belge ile ilgili bilgiler aşağıda Tablo 5.13'te verilmiştir.

Tablo 5.13. İkinci sorgudan EUEA ile elde edilen bilgi erişim çıktıları

Eşleşen Gövdelenmiş Terim Sayısı	Belge No	Belge adı	Sorguyla İlgisi Var mı?
51	**693**	**Nükleer Enerji**	**Evet**
50	**725**	**Birlik ve Termodinamik**	**Hayır**
48	**303**	**Değişik Yörelerden Sumak (Rhus Coriaria L.) Meyvesinin Ayrıntılı Kimyasal Bileşimi ve Oleorezin Üretiminde Kullanılması Üzerine Araştırma**	**Hayır**
45	**971**	**Ötanazi**	**Hayır**
44	**768**	**Nükleer Teknolojinin Riskleri**	**Evet**
Sorguyla İlgili Kayıtlarda Eşleşen Toplam Terim Sayısı : 95			

EUEA ile ilk sorgulamanın yapılması ve bilgi erişim çıktılarının elde edilmesi **3 saat 59 dakika 8 saniye** sürede tamamlanmıştır.

Uygulamanın ekran çıktısı Şekil 5.5'te görülmektedir.



Şekil 5.5. İkinci Sorgu için EUEA ekran çıktısı.

Sabit Uzunluk Algoritmasının Uygulanması

Bu yöntem kullanılırken sorgudaki gövde uzunluğu 4, 5, 6 ve 7 harf olacak şekilde dört farklı gövdeleme yapılmış ve bunun akabinde dört farklı erişim çıktısı elde edilmiştir.

Tablo 5.14. ikinci sorgudaki ham terimlerin SUA ile gövdelenmesi

Ham Terimler	4 Harfli Gövdeler	5 Harfli Gövdeler	6 Harfli Gövdeler	7 Harfli Gövdeler
nükleer	nükl	nükle	nüklee	nükleer
enerji	ener	enerj	enerji	enerji
gerçekten	gerç	gerçe	gerçek	gerçekt
zararlı	zara	zarar	zararl	zararlı
olsaydı	olsa	olsay	olsayd	olsaydı
ülkemizde	ülke	ülkem	ülkemi	ülkemiz
uygulanmazdı	uygu	uygul	uygula	uygulan

Yukarıdaki tabloda sorgu ifadesi içinde geçen terimlerin SUA‘na göre (4, 5, 6 ve 7 harfli) gövdelenmiş halleri görülmektedir. Aynı gövdeleme yöntemi kullanılarak veri kümesinde bulunan **266.182 adet** ham terim de gövdelenerek, aynı sayıda gövdelenmiş terim elde edilmiştir.

Kabul edilen her bir gövde uzunluğu (4, 5, 6 ve 7) için sorgudan elde edilen gövdelenmiş terimler ile veri kümesinden elde edilen gövdelenmiş terimler karşılaştırılmıştır.

Veri kümesindeki gövdelenmiş terim sayısı: **266.182**
Sorgudaki gövdelenmiş terim sayısı: **7**

Bir gövde uzunluğu için yapılan karşılaştırma sayısı:

266.182 x 7 = 1.863.272 (5.10)

Dört gövde uzunluğu için yapılan toplam karşılaştırma:

1.863.272 x 4 = 7.453.088 (5.11)

Yukarıdaki bazı sayısal bilgileri verdikten sonra sorgu sonucunda meydana gelen bilgi erişim çıktılarına bakalım:

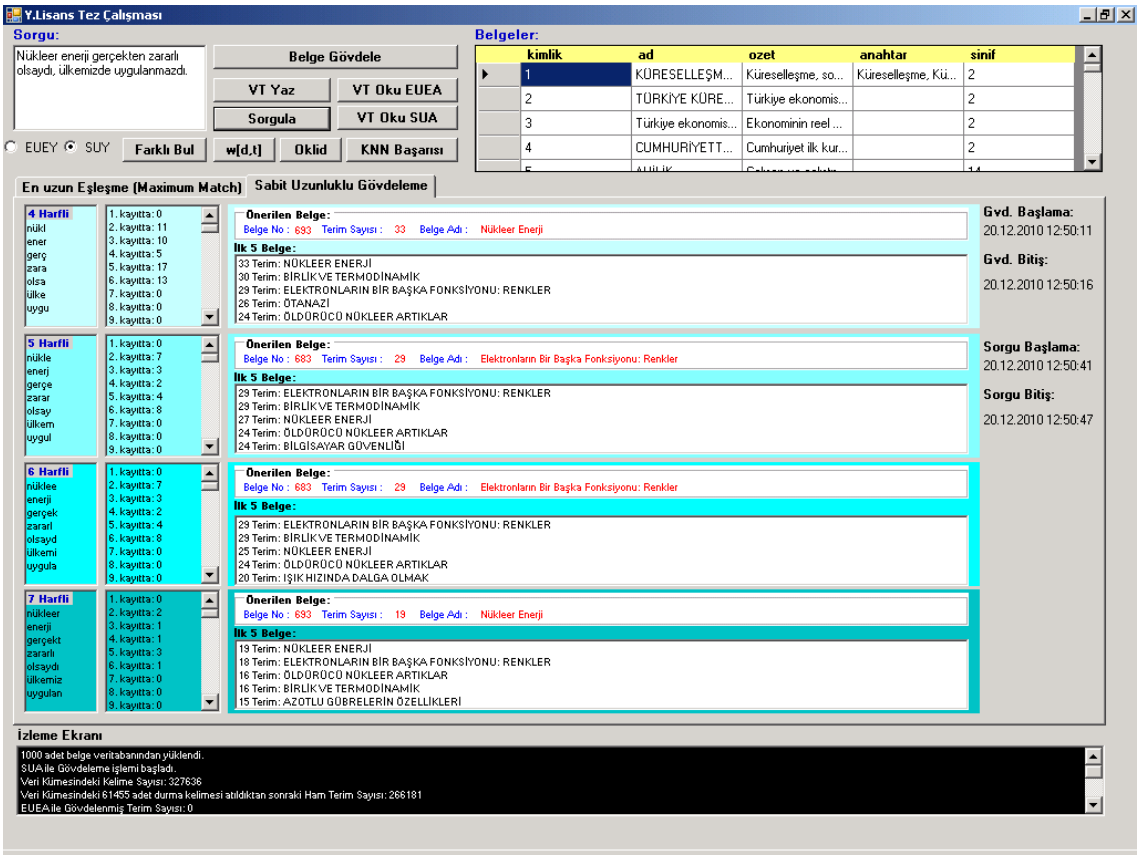
Sorgudaki gövdelenmiş terimler ile veri kümesindeki gövdelenmiş terimlerin karşılaştırılması sonucunda en fazla terimin eşleştiği ilk beş belge ile ilgili bilgiler aşağıda Tablo 5.15’te verilmiştir.

Tablo 5.15. İkinci sorgudan SUA ile elde edilen bilgi erişim çıktıları

ÇIKTI SIRASI	4 HARF			5 HARF			6 HARF			7 HARF		
	Terim Sayısı	Belge No	Sorgulya ilgisi var mı?	Terim Sayısı	Belge No	Sorgulya ilgisi var mı?	Terim Sayısı	Belge No	Sorgulya ilgisi var mı?	Terim Sayısı	Belge No	Sorgulya ilgisi var mı?
1	**33**	693	Evet	29	683	Hayır	29	683	Hayır	**19**	693	Evet
2	30	725	Hayır	29	725	Hayır	29	725	Hayır	18	683	Hayır
3	29	683	Hayır	**27**	693	Evet	**25**	693	Evet	**16**	715	Evet
4	26	971	Hayır	**24**	715	Evet	**24**	715	Evet	16	725	Hayır
5	**24**	715	Evet	24	988	Hayır	20	671	Hayır	15	943	Hayır
İlgili kayıtlarda eşleşen terim sayısı	**57**			**51**			**49**			**35**		

EUEA ile ilk sorgulamanın yapılması ve bilgi erişim çıktılarının elde edilmesi **7 saniye** sürede tamamlanmıştır.

Uygulamanın ekran çıktısı Şekil 5.6'da görülmektedir.



Şekil 5.6. İkinci sorgu için SUA ekran çıktısı

EUEA ve SUA Erişim Çıktılarının Değerlendirmesi:

EUEA'nın toplam eşleşen terim sayısı 95 olduğuna göre;

Sabit Uzunluk Algoritmasının, En Uzun Eşleşme Algoritmasına yakınlaşma değerleri aşağıdaki gibi hesaplanır

4 Harfi gövdeleme için: 57 / 95 = 0,60 (%60)

5 Harfi gövdeleme için: 51 / 95 = 0,54 (%54)

6 Harfi gövdeleme için: 49 / 95 = 0,52 (%52)

7 Harfi gövdeleme için: 35 / 95 = 0,37 (%37)

En Uzun Eşleşme Algoritmasında elde edilen çıktılara göre, **693** numaralı **"Nükleer Enerji"** adlı belge, en fazla eşleşen terime sahip, yani aranan konuya en uygun belgedir. Sabit Uzunluk Algoritması sonuçlarına göre de en yakın sonucu (%60) 4 harfli gövdeleme yöntemi üretmiştir. Bu yöntem aynı belgenin en uygun belge olduğu bilgisinde de EUEA ile örtüşüyor.

5.4.2. Genelleme yapılması

Çalışmanın bu kısmında çok sayıda farklı sorgu kullanılarak her sorguda SUA'nın, EUEA'na kaç harfli gövde uzunluğu ile ne kadar yakınlaşabildiği üzerinde durulmuştur. Yapılan deney sonuçları aşağıdadır (Bkz.Tablo 5.16).

Toplam 15 adet sorgu kullanılarak SUA'nın EUEA'na yaklaşma değerleri, ilgili belgelerdeki eşleşen toplam terim sayıları oranlanarak bulunmuştur.

Tablo 5.16. SUA'nın EUEA'na yaklaşma değerleri

Sorgu No	4 HARFLİ GÖVDELEME	5 HARFLİ GÖVDELEME	6 HARFLİ GÖVDELEME	7 HARFLİ GÖVDELEME
1	1,291666667	1,041666667	0,916666667	0,916666667
2	0,914529915	0,35042735	0,307692308	0,102564103
3	1,055555556	1,055555556	0,805555556	0,583333333
4	0,775510204	0,469387755	0,414965986	0,408163265
5	1,208791209	0,747252747	0,736263736	0,736263736
6	0,794117647	0,735294118	1,029411765	1
7	1,081081081	1,027027027	0,972972973	0,810810811
8	1,028571429	0,994285714	0,96	0,525714286
9	1,117647059	1	0,960784314	0,68627451
10	1,5	1,392857143	1,285714286	1,053571429
11	0,763888889	0,972222222	0,972222222	0,458333333
12	1,172413793	0,931034483	0,931034483	0,793103448
13	0,542857143	0,971428571	0,971428571	0,914285714
14	0,903225806	0,451612903	0,322580645	0,322580645
15	1,292682927	1,414634146	1,243902439	1,219512195
Toplam:	**15,44253932**	**13,5546864**	**12,83119595**	**10,53117748**
Ortalama:	**1,0295026**	**0,9036458**	**0,8554131**	**0,7020785**

15 adet farklı sorgu oluşturulup her biri VK üzerinde hem EUEA, hem de SUA kullanılarak ayrı ayrı işletilmiştir. Her sorgu için SUA'nın 4, 5, 6 ve 7 harfli gövdeleme yöntemlerinin bilgi erişim çıktılarındaki ilgili belgelerdeki eşleşen toplam terim sayıları

alınarak, EUEA'nın bilgi erişim çıktılarındaki ilgili belgelerdeki eşleşen toplam terim sayısına oranlanmıştır. Bu hesaplamaya bir örnek vermek gerekirse;

Örneğin, 14 numaralı sorgu ifadesi için; SUA ile eşleşen toplam terim sayıları;

4 HARFLİ GÖVDELEME	5 HARFLİ GÖVDELEME	6 HARFLİ GÖVDELEME	7 HARFLİ GÖVDELEME
56	**28**	**20**	**20**

EUEA kullanılarak eşleşen toplam terim sayısı ise **62** olarak bulunmuştur.

Bu sayısal veriler kullanılarak oranlama yapıldığında EUEA'na yaklaşma değerleri aşağıdaki gibi bulunur.;

4 HARFLİ GÖVDELEME	5 HARFLİ GÖVDELEME	6 HARFLİ GÖVDELEME	7 HARFLİ GÖVDELEME
56/62=0,903225806 (% 90)	**28/62=0,451612903 (% 45)**	**20/62=0,322580645 (% 32)**	**20/62=0,322580645 (% 32)**

5.5. Metin Sınıflandırma

5.5.1. Vektör uzayı ile metinlerin matematiksel modelinin oluşturulması

Metinlerin matematiksel modelinin oluşturulması için belgelerde geçen tüm terimler her iki gövdeleme algoritmasıyla da ayrı ayrı gövdelenir. Gövdelenen bu terimlerden terimlerin tekrarı olmaksızın bir terimler sözlüğü meydana getirilir. Meydana getirilen bu terimler sözlüğü yazılım bünyesinde tek boyutlu bir dizide tutulmaktadır. Sonra her belgede geçen terimler ve terimlerin belge içindeki tekrarlanma sayıları tespit edilerek bir matris oluşturulur. Buna Terim Sıklığı (***tf***) Matrisi denir. Bu matriste yer alan her satır, bir belgeye ait olan vektör uzayı modeli kullanılarak oluşturulmuş vektörü temsil eder.

5.5.2. Ağırlıklandırma işlemleri

Bir önceki aşamada her belgenin matematiksel modeli çıkarıldıktan sonra, belgelere ait olan vektörler oluşturulmuştu. Şimdi oluşturulan belge vektörleri daha sağlıklı sonuçlar vermesi için ağırlıklandırma işlemine tabi tutulacaktır. Yazılımın uygulanmasında yukarıdaki bölümlerde anlatılan ağırlıklandırma yöntemlerinden Terim Sıklığı Ters Belge Sıklığı (*Term Frequency Invers Document Frequency*, ***tfidf***) kullanılmıştır.

$$idf_i = \log\left(\frac{|D|}{df_i}\right) \tag{5.12}$$

Bu yöntemin uygulanmasında yukarıda 5.12'de gösterilen ters belge sıklığı (*invers document frequency*, ***idf***) formülü kullanıldığından ilk önce bu formülde yer alan $\boldsymbol{df_i}$ (*Belge sıklığı*)'nin hesaplanması gerekmektedir. Bu değer ilgili terimin en az bir kere geçtiği belge sayısını göstermektedir. Yine bu formülde yer alan **|D|** değeri de toplam belge sayısını göstermektedir.

Sonuç olarak, yukarıdaki tüm değerler kullanılarak 5.13'de gösterilen aşağıdaki formül ile her belge vektörü için ağırlıklandırılmış vektörler elde edilir.

$$w_i = tf_i \cdot idf_i \tag{5.13}$$

5.5.3. Öklid uzaklıkların hesaplanması

Belgeler arasındaki benzerlik hesaplama yöntemlerinden en kolay ve yaygın olan yöntem belge vektörlerinin üç boyutlu uzaydaki Öklid uzaklıklarının hesaplanması yöntemidir. Metin sınıflandırması için belgeler arasındaki benzerliklerin de bulunması gerektiği için her bir belgenin diğer belgelerle olan Öklid uzaklıkları hesaplanmıştır. Bu hesaplamada aşağıdaki formül kullanılmıştır.

$$Öklid(d_q, d_i) = \sqrt{\left[\sum_{j=1}^{M} (w_{q,j} - w_{i,j})^2\right]} \tag{5.14}$$

5.5.4. K-NN algoritmasının uygulanması

Bu tez çalışmasında gövdeleme yöntemlerinin kıyaslanması amacıyla, gövdelemenin sınıflandırma üzerindeki etkisi de araştırılmıştır. Bu bağlamda metin sınıflandırma algoritmalarından K-NN algoritması kullanılmıştır. Bu algoritma belgelerin birbirine yakınlığı prensibine dayandığı için yukarıda anlatılan Öklid mesafeleri hesaplanan belgeler test ve eğitim belgeleri olarak gruplandırılarak n kere çapraz doğrulama **(n Fold Cross Validation)** yöntemine tabi tutulmuştur.

n Kere Çapraz Doğrulama **(n Fold Cross Validation):**

Veri seti rastgele n gruba ayrılır. 1. grup test için ayrılırken geriye kalan gruplarla model kurulur. Kurulan model teste ayrılan veriler için kullanılır. Süreç n defa tekrar eder ve sırasıyla tüm guruplar (veri setinin tamamı) test için kullanılmış olur.

5.5.5. Gövdeleme yöntemlerinin sınıflandırma başarıları

Yukarıda anlatılan Vektör Uzayı ile Metinlerin Matematiksel Modelinin Oluşturulması, Ağırlıklandırma İşlemleri, Öklid Uzaklıkların Hesaplanması, K-NN Algortimasının Uygulanması aşamaları hem EUEA hem de SUA'nın 4, 5, 6 ve 7 harfli kullanımları için ayrı ayrı uygulanmıştır. Bu çalışmamızda sınıflandırıcı ile 15 sınıfa ait 1000 belge içerisinden en fazla geçen 5 sınıfa (Eğitim Bilimleri, Fizik, Sebze ve Meyve Gıdası, İşletme, İktisat(Ekonomi)) ait 607 belge seçilmiştir. Her belge vektörünün Öklid mesafesi bakımından en yakın olduğu belge tespit edildikten sonra test belgesi kendisine en yakın olan eğitim belgesinin sınıfına atanır. Eğitim süreci tüm belgelerin k-NN ile yeniden sınıflandırması tamamlanınca sona erer.

Seçilen belgeler üzeride, n=8 seçilerek çapraz doğrulama uygulanmış ve çapraz doğrulama sonucunda sınıflandırma başarıları aşağıda Tablo 5.17'de görüldüğü gibi gerçekleşmiştir:

Tablo 5.17. Gövdeleme Algoritmalarının k-NN Sınıflandırma Başarısı Oranları

Gövdeleme Yöntemi	k-NN'de kullanılan k parametresi değeri	Başarı (%)
EUEA	9	72
SUA (4 harf)	8	64
SUA (5 harf)	5	60
SUA (6 harf)	9	68
SUA (7 harf)	9	72

5.6. Bölüm Sonucu

Yukarıda sonuçları verilen deneysel uygulamalar ile, EUEA kullanılarak gövdeleme ve SUA kullanılarak gövdeleme yapılması halinde, BES'nin çıktıları oluşturulurken, eşleşen terim sayılarında nasıl bir durum meydana geldiği ve gövdeleme algoritmalarının metin sınıflandırmasında nasıl etkili oldukları gözlenmiştir. Yapılan uygulamalarda, farklı sorgulara farklı sayılarda eşleşen terim sayılarına sahip belgeler çıktı olarak sunulmuştur. Bu farklılaşmayı ortadan kaldırmak ve genel bir hüküm verebilmek için, çok sayıda sorgunun, her iki algoritmaya (SUA için 4, 5, 6 ve 7 harfli gövdeler) ait erişim çıktıları tek tek hesaplanarak ortalaması alınmış ve Tablo 5.16'da gösterilmiştir. Tablo 5.16'ya bakarak şu sonuçları çıkarabiliriz:

Bilgi Erişim Sistemleri'nde, En Uzun Eşleşme Algoritması kullanarak gövdeleme yöntemi ile Sabit Uzunluk Algoritması kullanarak gövdeleme yöntemi (ham terimin ilk 4, 5, 6 ve 7 harfi gövde kabul edilerek) belgelerdeki eşleşen terim sayıları bakımında kıyaslanmış ve sonuç olarak, EUEA'na en yakın sonucu veren SUA yöntemi ham terimin ilk 4 ve 5 harfini gövde kabul ederek yapılan gövdeleme yöntemleri bulunmuştur.

Ayrıca, k-NN sınıflandırma uygulanması sonucunda en başarılı sınıflandırma sonucunu veren En Uzun Eşleşme Algoritması'dır. Sabit Uzunluk Algoritması gövdelemesinde kullanılan gövde uzunlukları içerisinden en başarılı sınıflandırma En Uzun Eşleşme ile aynı başarıya sahip yedi harfli gövdelemeye aittir. Aynı sınıflandırma işleminde sınıflandırma başarısı en düşük olan beş harfli gövdeleme yöntemi olarak bulunmuştur. Yukarıdaki Tablo 5.17'den anlaşılacağı üzere beş, altı ve yedi harfli

gövde uzunlukları için gövde uzunluğunun artması sınıflandırma başarısını artırırken, sadece 4 harfli gövdeleme için aynı durum söz konusu olmamıştır. Sınıflandırma başarısı ile sabit uzunluklu gövdelemede kullanılan gövde uzunluğu tamamen doğru orantılı değildir. Sınıflandırma başarısının gövde uzunluklarıyla tamamen doğru orantılı olarak değişmemesinin sebebi, veri kümesindeki ham terimlerin aldıkları ekler (yapım eki ve çekim eki) ve meydana gelen ses olayları ile ilgili olduğu kadar, belge koleksiyonundaki belgelerin ait oldukları sınıf bilgilerinin manüel olarak yani bir uzman görüşü olmaksızın belirlenmesiyle de alakalıdır.

Sonuç olarak, k-NN sınıflandırması başarı oranlarına bakılarak "En Uzun Eşleşme Algoritması'na en yakın sonucu Sabit Uzunluk Algoritması'nın yedi harfli gövdelemesi vermiştir." denilebilir.

Yukarıda her iki gövdeleme algoritmasıyla da yapılan iki adet örnek sorgulamaya ait gövdeleme ve sorgulama işlemlerinin sayısal verileri aşağıda görülmektedir:

- **İlk sorgu için:**

  EUEA Gövdeleme Süresi : 10 dk 18 s.
  EUEA Sorgulama Süresi : 7 s.
  SUA(Dört gövde uzunluğu için) Gövdeleme Süresi : 4 s
  SUA(Dört gövde uzunluğu için) Sorgulama Süresi : 6 s

  SUA'nın EUEA'na yakınlaşma oranları:
  - 4 Harfli Gövdeleme: %93
  - 5 Harfli Gövdeleme: %95

- **İkinci sorgu için:**

  EUEA Gövdeleme Süresi : 10 dk 18 s.
  EUEA Sorgulama Süresi : 4 s.
  SUA(Dört gövde uzunluğu için) Gövdeleme Süresi : 5 s
  SUA(Dört gövde uzunluğu için) Sorgulama Süresi : 6 s

  SUA'nın EUEA'na yakınlaşma oranları:
  - 4 Harfli Gövdeleme: %60
  - 5 Harfli Gövdeleme: %54

EUEA, kelimelerin gövdelenmesinde bir sözlükten faydalanılmasından dolayı anlamsal olarak uygun bir gövdeleme yöntemi olmasına karşın, gövdelemenin tamamlanması için harcanan süre bakımından uygun bir performansa sahip değildir. Bu performans kaybının sebebi de yine her defasında sözlükteki gövde kelimeler ile karşılaştırma yapılmasıdır. Yukarıda anlatılan örnek sorgulamalarda sözlükle yapılan karşılaştırma sayıları da hesaplanmıştır. Örnek olarak, ilk sorguda geçen gövdeleme sürelerine baktığımızda SUA için gövdeleme süresi olan **4s** 4, 5, 6 ve 7 harfli gövdeleme işlemlerinin tamamı için geçen süredir. Aritmetik olarak sadece bir SUA ile gövdeleme için geçen süre **1s** diyebiliriz. EUEA ile SUA'yı kıyasladığımızda, **10dk 18s** ile **1s** arasındaki çok büyük zaman farkı vardır.

Sadece zaman, bir algoritmanın diğerinin yerini tutması için tabiî ki yeterli bir veri değildir. Bunun için, algoritmaların bilgi erişimi açısından birbirlerine makul derecede yakın sonuçlar üretebilmesi gerekmektedir. SUA'nın EUEA'na yaklaşma değerleri hem örnek iki sorgunun hem de genellemede kullanılan on beş sorgunun sonucunda (Bkz. Tablo 5.16) verilmiştir.

Gövdeleme algoritmaları kıyaslanırken her iki algoritmanın da sınıflandırma başarıları incelenmiş ve en fazla belgeye sahip 5 sınıfa ait 607 adet belge için EUEA ve SUA'nın sınıflandırma başarıları ***n fold validation*** ile n=8 seçilerek hesaplanmıştır. Sınıflandırma başarılarına ait sayısal veriler yukarıda Tablo 5.17'de gösterilmektedir.

6. SONUÇ ve ÖNERİLER

Günümüzde teknolojinin hızla gelişmesi çok yoğun miktarda veri artışını da beraberinde getirmiştir. Bu yoğun veri artışından en fazla etkilenen de metinsel veriler olmuştur. Metinsel verilerin incelenmesi ve bu verilerin amaca uygun olarak hızlı bir şekilde kullanılması da bilgi erişim sistemleri için yeni bir problem teşkil etmiştir. Elektronik ortamlarda saklanan metinsel verilerin çoğunun doğal dille yazılmış olmaları da metin madenciliğinde doğal dil işlemenin önemini ortaya koymuş ve metin işleme çalışmalarında metnin yazıldığı dilin yapısının da bilinmesi ihtiyacını doğurmuştur.

Yapılan bu çalışmada ise metin işleme süreçlerinden en önemlileri olan ön işleme süreçleri ve yine bu süreçlerden en önemlisi olan gövdeleme üzerinde durulmuştur. Metinsel ifadelerde geçen kelimelerin gövdelenmesi başlı başına incelenmesi gereken bir süreç olup, metnin yazıldığı dile göre de farklı yaklaşımlarla değerlendirilmesi gerekmektedir. Bu tez çalışmasında, gövdeleme çalışmaları incelenmiştir. Gövdeleme çalışmalarında Türkçe'nin sondan eklemeli bir dil olması göz önünde bulundurularak mevcut çalışmalardan bir tanesi olan *En Uzun Eşleşme (longest match) Algoritması* ve kelimelerin sabit uzunlukta alınarak gövdelenmesi prensibini temel alan *Sabit Uzunluk Algoritması* kullanılarak bir gövdeleyici ve metin sınıflandırıcı yazılım uygulaması yapılmıştır. BES'nde gövdeleme yöntemlerinin iki tanesini sayısal veriler elde ederek karşılaştırmamızı sağlayan bu yazılım uygulamasında gövdeleme sürecinden önce ön işlemenin diğer süreçleri de uygulamaya dâhil edilmiştir.

Bu tez çalışmasında, metinsel verilerin sadece elektronik ortamlarda muhafaza edilmesinin yanı sıra, etkin bir şekilde kullanılması gerektiği, bunun için de BES geliştirilirken, bu sistemlerin programlanmasında gövdeleme algoritmalarının kullanılması ve kullanılan farklı gövdeleme yöntemlerinin erişim performansına etkilerinin de göz ardı edilmemesi gereken bir gerçeklik olduğu bir kez daha ortaya koyulmuştur.

Sonuç olarak, EUEA ile doğru bilgi çıkarımları elde edilebildiği, fakat EUEA'nın zaman performansının bilgi erişim sistemlerinde kullanıcının isteklerine uygun sürede cevap veremeyecek kadar kötü olduğu görülmüştür. SUA'nın doğru bilgi çıkarımındaki EUEA'na yaklaşma oranları 4, 5, 6 ve 7 harfli sabit uzunluklu gövdelemeler yapılarak ayrı ayrı test edilmiş ve EUEA'na terim eşleşmesi bakımından en yakın değerleri 4 harfli ve 5 harfli SUA'nın verdiği görülmüştür. Birbirine yakın

bilgi çıkarım sonuçları veren bu iki algoritma arasında kullanıcıya cevap verme süreleri, EUEA için **10dk 18 s**, SUA için **1 s** olduğuna göre aralarında çok büyük bir fark olduğu açıkça görülmektedir.

İşte bu sebeplerden dolayı sonuç olarak, "Bilgi erişim sistemlerinde, en uzun eşleşme algoritması yerine, onunla hemen hemen aynı sonucu çok daha kısa ve makul sürede üreten sabit uzunluk algoritmasının uygun gövde uzunluğu seçilerek kullanılması bilgi erişim performansı açısından daha uygundur." diyebiliriz.

Bu çalışmada ayrıca, gövdeleme yöntemlerinin sınıflandırma başarıları da k-NN sınıflandırması kullanılarak incelenmiştir. Bu inceleme sonuçları Tablo 5.17'de sayısal verilerle sunulmuştur. Bu verilerden görüldüğü üzere beş, altı ve yedi harfli gövde uzunlukları için gövde uzunluğunun artması sınıflandırma başarısını artırırken, sadece 4 harfli gövdeleme için aynı durum söz konusu olmamıştır. Sınıflandırma başarısı ile sabit uzunluklu gövdelemede kullanılan gövde uzunluğu tamamen doğru orantılı değildir. Sınıflandırma başarısının gövde uzunluklarıyla tamamen doğru orantılı olarak değişmemesinin sebebi, veri kümesindeki ham terimlerin aldıkları ekler (yapım eki ve çekim eki) ve meydana gelen ses olayları ile ilgili olduğu kadar, belge koleksiyonundaki belgelerin ait oldukları sınıf bilgilerinin manüel olarak yani bir uzman görüşü olmaksızın belirlenmesiyle de alakalıdır.

Sınıflandırma işleminin sonucu olarak, "k-NN sınıflandırması başarı oranlarına bakılarak En Uzun Eşleşme Algoritması'na en yakın sonucu Sabit Uzunluk Algoritması'nın yedi harfli gövdelemesi vermiştir." denilebilir.

Bu tez çalışması literatürde yer alan mevcut gövdeleme yöntemlerinden iki tanesinin zamanlama performanslarını karşılaştırmalı olarak incelemiştir. Bu çalışmada kullanılan belge koleksiyonu oluşturulurken bir çok bilim dalından belgelere yer verilmiştir. Aynı çalışma, daha fazla gövdeleme algoritmasını içerecek şekilde ve daha fazla belge bulunduran bir veri kümesi üzerinde genişletilerek uygulanabilir. Bu genişletilmiş uygulamada, her bir belgenin dâhil olduğu sınıfların, o bilim dallarında uzman kişiler tarafından incelenmesi sonucunda tespit edilmesi ve bir belgenin birden fazla sınıfa dâhil olabildiği bir ortam oluşturulması çalışmanın daha sağlıklı sonuçlar üretebilmesi açısından önerilebilir.

KAYNAKLAR

Alpkoçak A., Kut A., Özkarahan E., 1995, An Interactive Document Indexing Software for Turkish Language, Bilişim Bildirileri, Dokuz Eylül Üniversitesi, İzmir.

Altintas, K., Can, F., 2002, Stemming for Turkish : a comparative evaluation. *Proceedings of the 11th Turkish Symposium on Artificial Intelligence and Neural Networks (TAINN)*,pp 181-188, Istanbul / Turkey.

Aybim Bilgisayar Tic. Ltd. Şti., 1996, Gazete Arşiv ve İletişim Dizgesi (GARILDI), Türk Bilim Vakfı Proje Önerim Raporu, İstanbul.

Duran G., 1997, Gövdebul: Turkish Stemming Algorithm, Yüksek Lisans Tezi,Hacettepe Üniversitesi,Bilgisayar Mühendisliği Bölümü, Ankara, Türkiye.

Ekmekcioglu F., Lynch M., Willett, P., 1996, Stemming and n-gram Matching for Term Conflation in Turkish Texts, Information Research, Vol. 2, No:2.

Eroğlu M., 2000, A Study On The Effects Of Stemming And Thesaurus For Retrieving Information In Turkish Documents, Master Thesis, Hacettepe University, Computer Enginering, Ankara, Türkiye.

Freund, G.E. and Willett, P., 1982, Online Identification of Word Variants and Arbitrary Truncation Searching Using a String Similarity Measure, Information Technology: Research and Development, Vol. 1, pp. 177-187.

Güzel, A., 2005. Üniversiteler İçin Türk Dili Ders Kitabı, Başkent Üniversitesi, Ankara.

Han, J., ve Kamber, M., 2001, "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers.

Jackson, P., Moulinier, I., 2002, Natural language processing for online applications: text retrieval, extraction, and categorization, Amsterdam**.**

Joachims T., 1997, Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. In Proceedings of the International Conference on Machine Learning (ICML'97), 143-151.

Joachims, T., 2002, Learning to classify text using support vector machines, Kluwer Academic Publishers, Boston.

Jurafsky, D. and Martin, J., 2000, Speech and Language Processing, Prentice Hall, New Jersey.

Kantardzic M., 2003, Data Mining:Concepts, Models, Methods, and Algorithms, IEEE Pres, Wiley Interscience Publications.

Kesgin F., 2007, Topıc Detectıon System For Turkısh Texts, Master Thesis, Graduate School of Natural and Applied Sciences, Istanbul Technical University,Istanbul.

Kohonen, T., 1990, "The Self-Organizing Map," Proceedings of the IEEE, vol. 9, pp.1464-1479.

Köksal A., 1981, Tümüyle Özdevimli Deneysel Bir Belge Dizinleme ve Erişim Dizgesi, TBD 3. Ulusal Bilişim Kurultayı, Ankara.

Lassila O., 1998, Web Metadata : A Matter of Semantics. IEEE Iternet Computing, pp. 30-37.

McCune B. P., Tong R. M., 1985, Dean J. S. and Shapiro D. G., RUBRIC: A System for Rule-Based Information Retrieval, IEEE Trans. On Software Engineering, 11(9), pp. 939-944.

Mitra S. and Acharya T., 2003, Data Mining: Multimedia, Soft Computing and Bioinformatics, Wiley Interscience Publications, New Jersey.

Oflazer, K., 1994, Two-level Description of Turkish Morphology, *Literary and Linguistic Computing*, Vol. 9, No:2.

Oflazer K.(*), Bozşahin H.C.(**), Natural Language Processing in Turkish, (*)Bilkent University, Ankara. (**) Midle East Technical University,Ankara.

Pilavcılar İ., 2007, Metin Madenciliği ile Metin Sınıflandırma(KNN Algoritması) – 3, Yazılım Mühendisliği İleri Seviye Makaleleri http://www.csharpnedir.com.

Porter, M.F., 1980, An Algorithm For Suffix Stripping, *Program*, 14(3):130-137.

Salton G. and Mc Gill M.J., 1983, Introduction to Modern Information Retrieval.McGraw-Hill, New York.

Saracoğlu R., 2007, Searching For Similar Documents Using Fuzzy Clustering, PhD Thesis, Graduate School of Natural and Applied Sciences, Selçuk University, Konya.

Sever H., 2002, Kaşgarlı Mahmut Bilgi Geri Getirim Sistemi (KMBGS) Proje no: 97K121330 Sonuç Raporu, Bilgisayar Mühendisliği Bölümü Bilgi Erişim Araştırma Grubu, Hacettepe Üniversitesi, Ankara.

Sezer E., 1999, SMART Bilgi Erişim Sistemi'nin Türkçe Yerelleştirilmesi ve Otomatik Gömü Üretimi, Yüksek Mühendislik Tezi, Hacettepe Üniversitesi, Bilgisayar Mühendisliği Bölümü, Ankara, Türkiye.

Solak, A., 1994, Can, F., Effects of stemming on Turkish text retrieval, *Proceedings of the Ninth Int. Symp. on Computer and Information Sciences*., pp. 49-56 Antalya, Turkey.

Türkeş M.K., 2007, Phrase Based Indexıng In Informatıon Retrıeval, Yüksek Lisans Tezi, Graduate School of Natural and Applied Sciences, Istanbul Technical University,Istanbul.

Yıldırım P.(*), Uludağ M.(**), Görür A.(*), 2008, Hastane Bilgi Sistemlerinde Veri Madenciliği, Akademik Bilişim Konferansları'08, (*) Çankaya Üniversitesi, Bilgisayar Mühendisliği Bölümü, Ankara. (**) European Bioinformatics Institute, Cambridge, UK.

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı : Mehmet BALCI
Uyruğu : T.C.
Doğum Yeri ve Tarihi : Konya, 02/04/1982
Telefon : 0.338.226 20 88
Faks : 0.338.226 20 80
e-mail : mehmetbalci@kmu.edu.tr

EĞİTİM

Derece		**Adı, İlçe, İl**	**Bitirme Yılı**
Lise	:	İmam Hatip Lisesi, Alanya, ANTALYA	1999
Üniversite	:	Selçuk Üniversitesi, KONYA	2006
Yüksek Lisans	:		
Doktora	:		

İŞ DENEYİMLERİ

Yıl	**Kurum**	**Görevi**
2009	Karamanoğlu Mehmetbey Üniversitesi, Meslek Yüksek Okulu, Bilgisayar Teknolojileri Bölümü, KARAMAN	Öğretim Görevlisi
2008-2009	Seydişehir Anadolu Ticaret Meslek Lisesi, Bilişim Teknolojileri Alanı, KONYA	Bölüm Şefi
2006-2009	Seydişehir Anadolu Ticaret Meslek Lisesi, Bilişim Teknolojileri Alanı, KONYA	Teknik Öğretmen
2003-2006	Babil Bilgi Teknolojileri, KONYA	Firma Sahibi

UZMANLIK ALANI

- Görsel Programlama
- Veritabanı Yönetim Sistemleri
- Delphi Programlama
- İşletim Sistemleri

YABANCI DİLLER

- İngilizce