T.C.
BEYKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ BİLİM DALI

VERİ MADENCİLİĞİNDE BİR METİN MADENCİLİĞİ UYGULAMASI

(Yüksek Lisans Tezi)

Tezi Hazırlayan: **Harun BAYER**

İSTANBUL, 2011

T.C.
BEYKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ BİLİM DALI

VERİ MADENCİLİĞİNDE BİR METİN MADENCİLİĞİ UYGULAMASI

(Yüksek Lisans Tezi)

Tezi Hazırlayan:

Harun BAYER

Öğrenci No:

080820007

Danışman:

Yrd. Doç. Dr. Gökhan SİLAHTAROĞLU

İSTANBUL, 2011

YEMİN METNİ

Yüksek lisans tezi olarak sunduğum "Veri madenciliğinde bir metin madenciliği uygulaması" başlıklı bu çalışmanın, bilimsel ahlak ve geleneklere uygun bir şekilde tarafımdan yazıldığını, yararlandığım eserlerin tamamının kaynaklarda gösterildiğini ve çalışmanın içinde kullanıldıkları her yerde atıf yapıldığını belirtir ve bunu onurumla doğrularım. 01/07/2011.

İMZA

Aday: Harun BAYER

T.C.
BEYKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZ SAVUNMA SINAVI SONUÇ TUTANAĞI

Beykent Üniversitesi
Fen Bilimleri Enstitüsü Müdürlüğü'ne,

Aşağıda tez adı belirtilen yüksek lisans öğrencisi .....08082007..... no'lu .....Harun BAYER.....'ın 18/07/2011 tarihinde yapılan tez savunma sınavı[1] sonucunda ...60....dakika süreyle sunduğu ve savunduğu tezi hakkında[2] oybirliğiyle/~~oyçokluğuyla,~~ Kabul/~~Red/Düzeltme(..... ay içinde)~~ kararı verilmiştir.

Bilgilerinize saygılarımızla arz ederiz.

Anabilim Dalı : .....BİLGİSAYAR MÜHENDİSLİĞİ.....

Programı : ..... BİLGİSAYAR MÜHENDİSLİĞİ.....

Tez Başlığı[3] : .....Veri Madenciliğinde Bir.....
.....Metin Madenciliği Uygulaması
.....
.....

Tez Sınav Jürisi	**Öğretim Üyesi**	**İmza**
Danışman	: Yrd. Doç. Dr. Gökhan Silahtaroğlu	
Üye	: Yrd. Doç. Dr. Rifat Çölkesen	
Üye	: Yrd. Doç. Dr. Ali Seylan	

---

[1] Jüri üyeleri söz konusu tezin kendilerine teslim edildiği tarihten itibaren en geç bir ay içinde toplanarak öğrenciyi tez savunma sınavına alır. Belirlenen günde yapılamayan jüri toplantısı, katılanların hazırladığı bir tutanakla enstitü yönetimine bildirilir. Bu durumda jüri en geç onbeş gün içinde toplanarak adayı tez savunma sınavına alır. Tez savunma sınav süresi en az 45 dakikadır. Yüksek lisans tez savunma sınavı, tez çalışmasının sunulması ve bunu izleyen soru-yanıt bölümlerinden oluşur ve dinleyiciye açıktır. (Beykent Lisansüstü eğitim ve Öğretim Yönetmeliği-Madde30-3)

[2] Tez sınavının tamamlanmasından sonra jüri, tez hakkında "kabul", "düzeltme" veya "red" kararı verir. Jüri başkanı, jüri üyelerince imzalanmış sınav tutanağını, tez sınavını izleyen üç gün içinde ilgili enstitü yönetimine teslim eder. Tezi başarısız bulunan öğrencinin Enstitü ile ilişiği kesilir. Tezi hakkında düzeltme kararı verilen öğrenci en geç üç ay içinde gerekli düzeltmeleri yaparak ve yönetmelikte belirtilen usullere uygun olarak tezini aynı jüri önünde yeniden savunur. Bu savunma sınavında da tezi kabul edilmeyen öğrencinin enstitü ile ilişiği kesilir. (Beykent Lisansüstü eğitim ve Öğretim Yönetmeliği-Madde30-4)

[3] İleride doğabilecek aksaklıkların engellenmesi için tezin başlığının yazılması gerekmektedir.

TEŞEKKÜR

Bana bu konuda çalışma fikri veren, çalışmam süresince fikir ve düşünceleriyle beni yönlendiren değerli danışmanım Sayın Yrd. Doç. Dr. Gökhan SİLAHTAROĞLU' na sonsuz teşekkürlerimi sunarım.

Ayrıca, çalışmam boyunca hoşgörü ve desteğini esirgemeyen değerli aileme teşekkürlerimi sunarım.

VERİ MADENCİLİĞİNDE BİR METİN MADENCİLİĞİ UYGULAMASI

Tezi Hazırlayan: Harun BAYER

ÖZET

Veri madenciliğinin alt dallarından olan metin madenciliği ile yüksek kapasiteli metinler içerisindeki istenilen öz bilgilere ulaşılmaktadır. Verilerin çoğunlukla metinsel halde bulunmalarından ötürü veri madenciliği uygulamalarının büyük bir çoğunluğu metin madenciliği ile gerçekleştirilmektedir. Bu çerçevede metin madenciliği aslında veri madenciliğinin yansımasıdır diyebiliriz.

Bu tez çalışmasında; son yıllarda birçok alanda kullanılan veri madenciliği ve alt dalı olan metin madenciliğinin gelişim süreçleri, kullanılan modeller ve bu uygulamaların çözüm getirebilecekleri alanlar üzerinde durulmuştur. Veri ve metin madenciliği teknikleri ile veritabanlarında bulunan gizli ilişkiler açığa çıkarılabilecektir.

Metin madenciliğinin yaygın olarak kullanıldığı alan; Türkçe metinlerin analiz edilmesidir. Metin madenciliği teknikleri kullanılarak Türkçe kelimelerin anlam haritalarının çıkarılması üzerine iki modülden oluşan bir uygulama geliştirilmiştir. Uygulama bölümünde ilk olarak metin içerisinde yer alan kelimeler, açık kaynak kodlu "zemberek" programı yardımıyla köklerine ayrılmıştır. Köklerine ayrılan kelimeler arasındaki gizli ilişkiler bu tez kapsamı süresince geliştirilen bilgisayar programı yardımıyla keşfedilmiştir. Keşfedilen bu bilgiler ışığında kelimeler arası anlamsal bağ durumu analiz edilmiş ve analiz sonucunda kelimelerin anlam bilgisi tahmini gerçekleştirilmiştir. Ayrıca metin madenciliği alanında, ileriye dönük çalışmalar için bir dizi öneride bulunulmuştur.

Anahtar Kelimeler: Veri madenciliği, Metin madenciliği, Veritabanı, Kelime ilişkileri, Metin analizi.

A TEXT MINING APPLICATION IN DATA MINING

Thesis Written By: Harun BAYER

ABSTRACT

By using text mining which is one of the sub-branches of data mining, the desired core information within the high capacity texts can be reached. Since the data are mostly in text form, most of the data mining applications are performed through text mining. In this scope, we can actually say that text mining is a reflection of data mining.

In this thesis; the development processes of the data mining which has been used in many areas in recent years and the text mining which is the sub-branch of data mining, the models used and the areas where these applications may bring solutions are emphasized. The hidden relations in the databases can be revealed with data and text mining techniques.

The area at which the text mining is commonly used is analysing the Turkish texts. An application consisting of 2 modules has been developed on extracting the significance charts of the Turkish words. At the application phase, the words in the text are separated into their roots by using open source software " zemberek". The hidden relations between the texts which are separated into their roots are discovered with the help of the software developed within the scope of this thesis. In light of these discovered data, the semantic relation status between the words are analysed and the estimation of the meaning of the words is performed as the result of the analysis. Furthermore a range of suggestions have been made for the further studies on text mining.

Key words: Data mining, Text mining, Database, Word Relations, Text Analysis.

İÇİNDEKİLER

TABLOLAR

Sayfa No.

ŞEKİLLER

Sayfa No.

1. GİRİŞ

Son yıllarda, gelişen bilgi teknolojileri sayesinde üretilen veri miktarı da hızla büyümektedir. Veritabanlarının gittikçe büyümesi, gelişen teknoloji ve internet kaynaklı veritabanlarının kendini sürekli olarak yenilemesi sonucunda yüksek boyutlu verilerden istenilen bilgilere ulaşmak gittikçe karmaşık bir durum haline gelmiştir. Bu durum, kurumların istenilen bilgiye ulaşma süreçlerinde zaman ve ekonomik anlamda sıkıntılarla karşılaşmasına neden olmaktadır. Bu gelişmeler sonucunda yüksek boyutlu verilerden anlamlı, doğru ve güvenilir bilgiye hızlı bir şekilde ulaşabilmek için "veri madenciliği" olarak adlandırılan çalışmalara başlanmıştır. Veri madenciliği ile kurumlar gerek kendi veritabanlarından gerekse farklı yüksek kapasiteli veritabanlarından hızlı bir şekilde istedikleri anlamlı bilgiyi keşfedebilecek hale gelmişlerdir.

Veri madenciliği uygulamalarıyla birlikte devasa boyutlu veriler arasında daha önce rastlanmayan, ilişkisiz olabileceği düşünülen ilginç örüntüler keşfedilebilmektedir.

Veri madenciliği uygulamalarında ilişkisel veriye ulaşabilmek için farklı verilerin yapısallaştırılarak uygulamalar için ön işlemden geçirilmesi ve hazır hale getirilmesi gerekir. Bu durumda metin madenciliği devreye girmektedir. Metin madenciliği; yapısal olmayan ya da yarı yapısal olan verilerin yapısallaştırılarak bilginin keşfi için hazır hale getirilmesidir.

Metin madenciliği ile metinsel veriler arasındaki gizli olan potansiyel bilgiler açığa çıkar. Bu bilgi ışığında uzmanlar, kurum ve kuruluşlar için gerekli olan tahminleri üretebilmektedirler.

Çalışmanın ikinci bölümünde; veri madenciliği içerisinde yer alan temel kavramlardan, veri ve bilginin öneminden, Veri madenciliğinin tarihsel gelişiminden, diğer disiplinlerle ilişkisinden ve veri madenciliğinin hangi alanlarda kullanılabileceğinden söz edilmiştir.

Üçüncü bölümde; veri madenciliği uygulama süreçlerinden bahsedilmiştir. Veri madenciliğinde problem tanımlanması, verilerin hazırlanması, veri madenciliği modelleri ve teknikleri, metin madenciliği ön işlemleri, metin madenciliğinde kullanılan yöntemler, bilginin kullanılan yöntemler ile çıkartılması ve metin madenciliği ile Türkçe kelimelerin ilişkisine değinilmiştir.

Tez çalışmasının dördüncü bölümünde; Türkçe kelimelerle gerçekleştirilen uygulamadan bahsedilmiştir. Söz konusu olan uygulama iki modülden oluşmaktadır. Birinci modülde hazır olan kök bulma algoritması, ikinci modülde ise bu tez kapsamı süresince hazırlanan program işleme sokulacaktır. Bu uygulama sonucunda açığa çıkan bilgiler anlamsal bir çerçeve içerisinde sonuçlandırılacaktır.

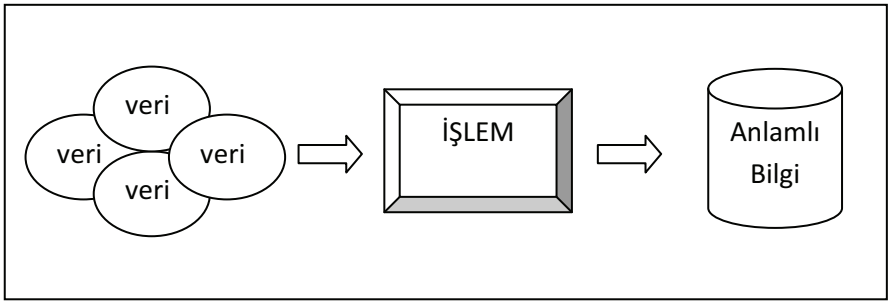
Son bölümde ise tez çalışmasının sonuçları metin madenciliği açısından değerlendirilmiştir. Bunun yanında, çalışmanın faydalı olabileceği alanlar üzerinde durulmuş ve bir dizi öneride bulunulmuştur.

2. VERİ MADENCİLİĞİ

2.1. Veri

Kavramsal olarak, “bir akıl yürütmenin, bir araştırmanın temeli olan ve olduğu gibi kabul edilen öğe” olarak tanımlanan veri; ham, işlenmemiş, kullanılmak üzere olan olay veya durum olabilir [1].

“Veri”, yalnız başına anlamsızdır; ancak bir anlam yüklediğimiz takdirde kullanılabilir bir bilgi olacaktır. İstediğimiz amaç doğrultusunda bilgi oluşturmak için verinin işlenmesi gerekmektedir. Veriyi bilgiye çevirmeye “Veri analizi” denilmektedir. “Bilgi” ise bir soruya yanıt vermek için veriden çıkarılan sonuç olarak tanımlanabilir. Bu bakımdan veri ve bilgi arasında önemli bir iletişim ağı olduğu söylenebilir. Bilgi anlamlı olarak bazen yalın halde bazen de diğer bilgilerle birlikte çeşitli durumlarda istenilen sonuca ulaşmak için kullanılabilir [2].



Şekil.1. Veri ve Bilgi İlişkisi

Günümüzde veri kazanılan bir hazine olarak görülmektedir. Verinin bu derece önemli olması ona olan ilgiyi artırmıştır. Kurum ve kuruluşların problem çözümlerinde verilerden yararlanma ihtiyaçları artmıştır. Veriye olan ilginin artması sonucunda, verinin nasıl kullanılacağı, nasıl yorumlanacağı ve nasıl saklanacağı gibi durumlar ortaya çıkmıştır. Yeni teknoloji ürünü bilgisayarların ekonomik olarak uygun olması, yüksek kapasiteli verilerin bilgisayarlara rahatlıkla kayıt edilmesi ve bu verilerin üzerinde matematiksel, istatistiksel hesapların yapılabilmesi kullanıcıların doğru bilgiye hızlı ve ayrıntılı olarak ulaşabilmelerini sağlamaktadır.

2.1.1. Bilginin Önemi

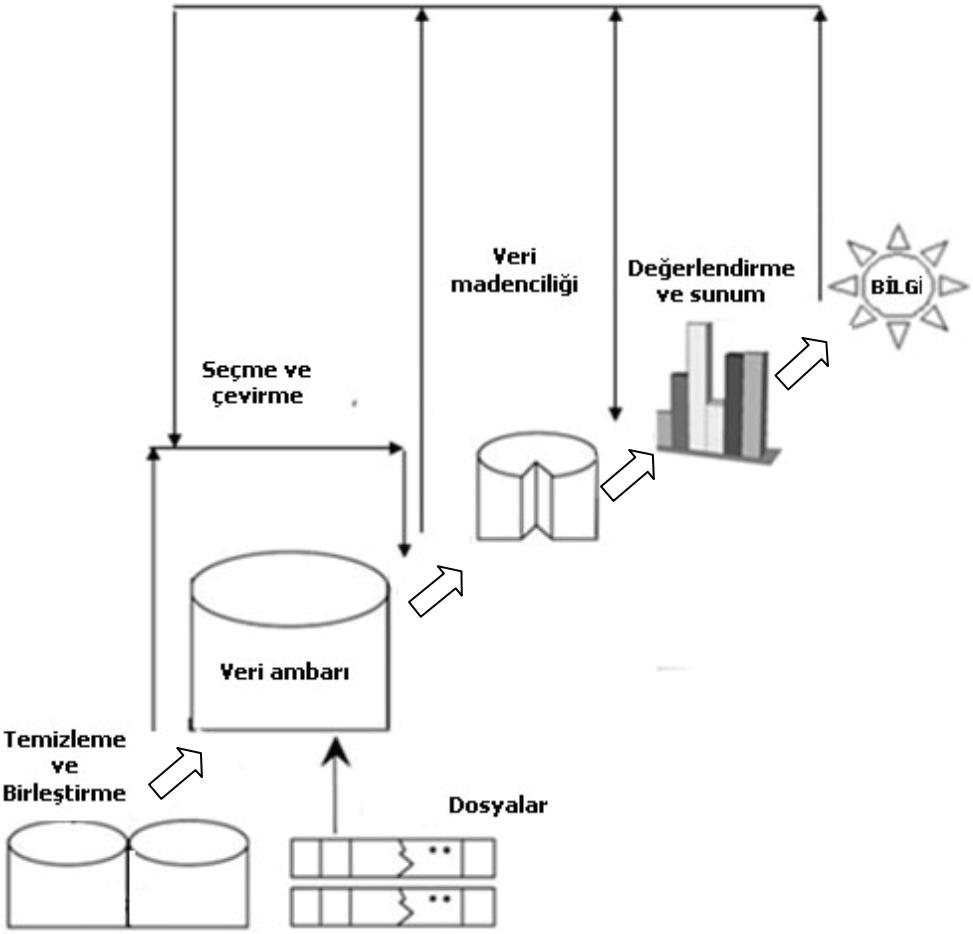
İnsan beyninin "bilgiyi öğrenmede" bir sınırı yoktur. Bilgi, insanlar arasında paylaşıldıkça artar. Bu paylaşım ile insanlar bilgiyi önce alır, işler ve daha sonra kullanır. Bilginin işlenmesinin ardından yeni bilgi oluşur ve bu bilgi tekrar paylaşılarak artar. Öğrenilen bilgilerle insanın yeni bir buluş yapmasının yeni fikirler üretmesinin bir sınırı bulunmamaktadır. İnsan istediği oranda eski bilgilerinden yararlanarak yeni bilgiler üretebilme imkânına sahiptir. Elbette ki bilgi bakımından zengin insanların, düşünce sistemi gelişecektir. Bunun doğal bir sonucu olarak karşılaştıkları problemli durumlara hızlı ve anlamlı çözümler üretebilecektirler. Bilgili insanlar karşılaştıkları güçlük ve sorunlara zihinlerinde sakladıkları bilgileri harmanlayarak, kolaylıkla çözüm üretebilirler. Bilgi olmadığı takdirde, insan gelişimi evresinde karşısına çıkabilecek problemlere çözüm üretemez, problem durumlarında kullanabileceği yeni bilgi şemaları oluşturamaz.

2.1.2. Bilginin Keşfi Süreci

Teknolojinin ilerlemesiyle birlikte her türlü gelişme bilgisayarlar yardımıyla kayıt altında tutulmaktadır. Kaydedilen veriler yüksek boyutlu veritabanları içerisinde gizlenmiş durumdadır. Saklı olan bu bilgilerin çıkartılması sonucunda işe yarar, aktif olarak kullanılabilir, anlamlı ve daha sade bilgilerin gün yüzüne çıkarılması, "bilginin keşfi" olarak bilinmektedir [3]. Gizli bilgiler arasından çıkarılan bu yeni bilgiler, işlenerek işe yarar hale getirilir. Bir süre sonra kullanılan bu bilgiler de veritabanlarında yerlerini alır. Bu döngü sürekli bu şekilde devam etmektedir. Gelişen teknoloji ile birlikte kolayca saklanılabilen bu yeni bilgilere her zaman ihtiyaç olacaktır.

"Bilginin keşfi" sürecinin önemli bir kısmında "veri madenciliği" yer almaktadır. "Veri madenciliği" yerine kavram olarak literatür çalışmalarında "veritabanlarında bilgi keşfi " ifadesi de kullanılmaktadır. "veritabanlarında bilgi keşfi " ifadesi ilk defa "Piatetsky-Shapiro" tarafından 1989 yılında gerçekleştirilen ilk veritabanında bilginin keşfi toplantısında kullanılmış, konuyla ilgili kavram ve

tanımlar ortaya konmuştur. “veri madenciliği“ terimi de “veritabanlarında bilgi keşfi” teriminin bir bileşeni olarak tanımlanmıştır [3].



Şekil.2.Bilginin Keşfi

Kaynak: Han, J., Kamber. M. (2006). Data Mining Concepts and Techniques.

2.2. Veri Madenciliği Kavramı

Gelişen teknolojiyle birlikte bilgisayarlar gündelik hayatımıza daha sık girmektedir. Bu sayede günlük hayatımızda yaptığımız her işlem bilgisayarlarda depolanmaktadır. Örneğin, marketlerden aldığımız her ürün, iade ettiğimiz ürünler, veritabanlarında saklanabilmektedir. Hastanelerde, belediyelerde veya ticarette yaptığımız her işlem artık anında veritabanlarında yerini almaktadır. Gün içerisinde bir mağazaya, alışveriş merkezine girerken veya çıkarken, bir banka içerisinde beklerken yaptığımız her işlem güvenlik sebebiyle kameralar tarafından düzenli

olarak kayıt altına alınmakta ve veritabanlarında saklanmaktadır. Bu şekilde yapılan kayıtların ardından veritabanlarında bir veri yığını oluşacaktır. Bu kayıtlar arasından istenilen bir bilgi çekileceği zaman, hızlı olarak istenilen sonuçlar alınamayabilir. Oluşabilecek bu veri yığınları içerisindeki önemli bilgilerin süzülmesi, çıkartılması gerekmektedir. Düzenli olarak kaydedilen tüm bu veriler, veritabanlarında çıkarılmayı bekleyen değerli bir maden gibi durmaktadır [4]. Bu değerli maden ayıklanıp, işlenildikten sonra anlamlandırılarak işe yarar bilgi haline gelmektedir.

Bilgiye ulaşma sürecinde büyük verilerin hepsinin aynı anda kullanımıyla istenilen bilginin elde edilmesi uzun zaman alacak ve ekonomik açıdan da masrafı fazla olacaktır. Bu durum, büyük veritabanları içerisinden talep edilen bilgilerin, düşük masraflı ve hızlı bir şekilde çekilmesi gerekliliğini karşımıza çıkarmaktadır. Teknolojinin bu doğal gelişim süreci sonucu olarak "veri madenciliği" ortaya çıkmıştır diyebiliriz. Veri madenciliği, bilgiye erişim sürecinde; makine öğrenmesi, veritabanı veya veri ambarı yönetimi, matematiksel ve istatistiksel teknikleri kullanarak önceden tahmin edilemeyecek olan bilgiye ulaşabilmektedir [3].

Bu konuda yapılan bir başka açıklamaya göre [5]; son yıllarda insanların yaptıkları her bankacılık işlemi, her alışveriş, neredeyse insanların her adımı uzaktan algılayıcılar, uydular tarafından kontrol altında tutulmakta ve yapılan birçok işlemin kolayca veritabanlarında saklanılabilir olduğu bilinmektedir. Bu şekilde bilgilerin kayıt altında tutulması veritabanlarının inanılmaz boyutlarda artmasına neden olmaktadır. Yalnızca uydu ve diğer uzay araçlarından elde edilen görüntülerin saatte 50 gigabyte düzeyinde olması, bu artışın boyutlarını daha açık bir şekilde göstermektedir. Hacimlerdeki bu büyük artış sebebiyle depolanan kayıtlar ve olaylar sonucu toplanan bu verilerden nasıl yararlanılacağı konularında araştırmalar yapılmıştır. Bu araştırmalar sonrasında "veritabanlarında bilginin keşfi (knowledge discovery in databases*)"* karşımıza çıkmaktadır. Bilginin keşfi süreci çeşitli basamaklardan oluşmaktadır. Bu süreç içerisinde en önemli basamak veri madenciliği aşamasıdır. Bu önem sebebiyle "veritabanlarında bilginin keşfi" sürecine birçok araştırmacı "veri madenciliği" demektedir [5].

Hızla büyüyen veritabanlarından istenilen bilginin keşfi, bilgilerin birikmesi ve toplanmasından daha önemli bir hal almıştır. Gelişen teknoloji sayesinde veriler

rahatlıkla elektronik depolarda tutulmaktadır. Artık önemli olan bu elektronik depolarda bulunan veriler arasında gizli kalmış bilgilere ulaşmaktır. Veritabanları üzerinde yapılacak birtakım işlemler sonucunda gizli kalmış bu değerli bilgilere ulaşılabilir. Bilim adamları gizli kalmış bilgilerin çıkartılması sırasında önceden "istatistiksel hesaplamalar" ve "makine öğrenmesi" yoluyla elle bazı anlamlı şablonlara ve ilişkilere ulaşılabilmekteydiler. Sonraki yıllarda veri madenciliğinin, büyük miktardaki verinin analiz edilmesinde ve bununla bağlantılı olarak anlamlı şablon ve kuralların keşfedilmesinde faydalı olacağına inanılmıştır [6,7].

Sağlıklı ve istenilen bilgiye hızlı bir şekilde erişim öneminin artması üzerine araştırmacıların bu alandaki çalışmalarının hızla çoğaldığını görmekteyiz. Araştırmacıların yüksek hacimli ve dağınık bilgi veri depoları üzerinde yapmış oldukları çalışmaların neticesinde veri madenciliği ve bilgi keşfi, özellikle elektronik ticaret, bilim, tıp, iş ve eğitim alanlarındaki uygulamalar ile yeni ve dikkat çeken bir araştırma alanı olarak ortaya çıkmıştır. Veri madenciliği, elimizde bulunan sınırsız sayıdaki yüksek kapasiteli bilgilerden, kullanılabilir olan anlamlı bilginin elde edilmesi çalışmalarının tümünü içermektedir. Veri madenciliğinde tümdengelimsel yöntem kullanılmamaktadır [8]. Tümevarım işlemleri formüle edilerek analiz etmeye ve sonuca varmaya çalışılır. Gerekli olan bilgiler seçilerek, tümevarımsal yöntem kullanılarak veriler analiz edilir ve sistematik olarak yeni anlamlı örüntülere ulaşılır.

Veri madenciliği özet olarak çok büyük veritabanlarındaki gizli olan veriler arasındaki ilişkilerin, verilerin birbirlerine olan yakınlık ve uzaklıkların, veriler arasındaki eğilimlerin, veriler arasındaki gizli bilgilerin açığa çıkartılmasıdır. Veriler veritabanlarında işlenmemiş halde bulunurken, veriler üzerinde işlemler yapılmadan işe yarayacak olan bilgiye ulaşılması söz konusu değildir. Veritabanlarında bulunan yapısal olmayan verilerin bir takım yöntem ve teknikler ile işlenmesi ve ardından potansiyel olarak keşfedilemeyen bilgiye otomatik olarak ulaşabilmek için tümevarımsal yöntemler kullanılarak yeni uygulamalar geliştirilmiştir. Veri madenciliği tekniği ile çok boyutlu veritabanları üzerinde bir takım sayısal yöntem ve algoritma uygulamaları yapılarak veriler arasında önceden tahmin edilemeyecek farklı bilgilere ulaşılır. Bu bilgiler, tahmin ve değerlendirme durumlarında ve gelecekle ilgili örüntülerin keşfedilmesi aşamalarında kullanılabilmektedir.

2.3. Veri Madenciliğinin Tarihsel Gelişim Süreci

Bilgisayarlar, ilk çıktığı yıllarda belirli bir takım basit düzeyde işlemler yapabilmekteydi. Bilim adamlarının uğraşları sonucu daha karmaşık yapıda olan problemlerin çözümü için bilgisayarlar geliştirildi. Gelişim sürekli olan bir yapı olduğu için, problemler de bu sürekliliğe paralel bir şekilde yeni bir problem olarak karşımıza çıkmaktadır. Yeni problemlerin giderilmesi için de gelişen bilgisayar teknolojisiyle çalışmalar yapılmakta ve çözümler üretilmektedir.

Teknolojinin gelişim süreciyle birlikte yeni olan sorunlardan birisi de veritabanlarındaki yığınlaşan verilerdir. Son yirmi yıldır veri toplama ve saklama kapasitesinde ciddi bir artış yaşanmaktadır. Öyle ki "gerçekte bilgi miktarı her yirmi ayda bir ikiye katlandığı" belirtilmektedir. Verilerin kayıt altına alınması işlemiyle veritabanlarında bulunan bilgilerin yorumlanması ve değerlendirilmesi süreci eş zamanlı olarak yapılamamaktadır. Saptanan sebeplerden ötürü bu sorunların giderilmesi için akıllı veritabanı analizi ve tahmin edici sistemlerin geliştirilmesi durumu ortaya çıkmıştır. Önemsenen sorunlarla birlikte veritabanları oluşturmak için harcanan zaman ve ekonomik alandaki yatırımlar, yeni çözüm önerilerine yönelimi sağlamıştır [7].

Veritabanlarında saklanan veri miktarı büyüdükçe ve toplanan veriler arasındaki karmaşıklık arttıkça, veritabanlarında saklı olan mevcut veriden anlamlı örüntülerin çıkartılması durumu ortaya çıkmıştır. Yüksek boyutlu verilerin içerisinden istenilen bilgiye ulaşma; zaman ve maliyet açısından belirli bir kayba neden olacaktır. Geçmişteki veritabanı sorgularıyla istediklerini alamayacağını öğrenen kurumlar, istedikleri bilgiye ulaşabilmek için veri madenciliğine yönelmişlerdir [9].

Veri madenciliğinde ilk çalışmaların mantık ve bilgisayar bilimleri alanlarında olduğu bilinmektedir. Mantık ve bilgisayar bilimcileri 1950 ve 1960 yılları arasında yapay zekâ ve makine öğrenmesi modellerini bulmuşlardır [10].

Veri madenciliği çalışmalarının temelinde veriler ve istatistiksel hesaplamalar büyük yer kaplamaktadır. 1960'lı yıllarda verilerin toplanması yönündeki yatırımlar artmış, verilerin saklanmasına önem verilmiştir. Saklanan verilerden istenilen bilgiye

ulaşmak için “veri taraması (data dredging)”, “veri yakalaması (data fishing)” yapılması uygun görülerek, ilk yıllarda veri madenciliğinin bu isimlerle anıldığı bilinmektedir. Bu yıllarda teknolojinin de gelişmesiyle birlikte verilerin güvenliği, saklanması ve analiz edilmesi durumları ortaya çıkmıştır. Bu yıllarda istatistikçiler yeni algoritmalar keşfetmişlerdir. Regresyon analizi, en büyük olabilirlik tahmini, sinir ağları gibi metotlar bu yılların veri madenciliği alanındaki çalışmalarını kapsamaktadır. Bu metotlar ile depolanan büyük veri yığınları analiz edilmeye başlanmıştır. İstatistiksel teknik ve algoritmalar ile devasa boyutlu veritabanlarından istenilen bilgilerin çıkartılması üzerine çalışmalar yapılmıştır [10-12].

1970 ve 1980 yılları arasında araştırmacı E.F.Codd tarafından ilişkisel veritabanı durumunu ortaya koyan “A relational model for large shared data banks” adlı makalesi yayımlandıktan sonra San Jose tarafından California’daki IBM araştırma laboratuarında “SQL” geliştirilmiştir [13]. Bu sırada artan veritabanlarıyla birlikte programlama dilleri de gelişim göstermektedir. Veritabanlarında oluşan yüksek boyutlu verilerin arasından istenilen bilgilerin çekilmesi durumunda programlama dilleri devreye girmektedir. 1980 ‘ li yıllarda ilişkisel (relational) veritabanları oluşturulmuş, SQL’ in geliştirilmesiyle birlikte verilerin anlık ve dinamik analizleri yapılmaya başlamıştır[11].Yine bu yıllarda “Genetic Algorithms”, “K-Means Clustering” ve “Decision Tree Algorithms” gibi algoritmalar geliştirilmiştir [10].

1980’li yılların sonlarına doğru “Piatetsky-Shapiro” veritabanlarında bilginin keşfi sürecinin büyük bölümünü ilk defa “veri madenciliği” kavramıyla adlandırmış ve ardından bu alanda yeni teknik ve metotlar geliştirmeye başlamıştır [3]. Veri madenciliği çalışmaları ile eldeki verilerden çok net olmayan ve önceden bilinmeyen potansiyel olarak gizli olan bilgilerin çıkartılmasında ilerleme sağlanmıştır.

1990 yılıyla beraber yapılan tahminlere göre verilerin yirmi ayda bir ikiye katlamakta, 90’ların sonuna doğru ise bu rakamın her bir yılda ikiye katlamakta olduğu bilinmektedir. Bu yıllar da veritabanında bilgi keşfinde çalışmalara devam edilmiş ve büyük veritabanları için “veri ambarı veritabanı” (database warehouses) geliştirilmiştir. Gelişen teknolojiyle birlikte veri madenciliği, veritabanlarında sistematik olarak bir işin parçası olmuştur [10,14,15].

2000'li yıllarda veri madenciliği alanında çalışmalar artmış ve karar destek uygulamalarına yardımcı yöntemlere ağırlık verilmiştir. "Teknolojik gelişmeler, ham verilerin yeni fırsatlar üretmek üzere yönetim ve pazar ihtiyaçlarına yanıt verecek bilgiye dönüştürülmesini kolaylaştırmış ve bir anlamda kurumları veri madenciliği üzerinde çalışmaya mecbur bırakmıştır" [11]. Veri madenciliği alanında yapılan uygulamalardan olumlu sonuçlar alınmasıyla birlikte farklı alanlarda da veri madenciliği teknikleri kullanılmaya başlamıştır. Yapılan araştırmalar veri madenciliğinin önemini göstermektedir. Bu bakımdan 2000'li yılları bilişim sektöründe veri madenciliği devri olarak adlandırmak yanlış olmayacaktır.

2.4. Veri Madenciliğinin Diğer Disiplinlerle İlişkisi

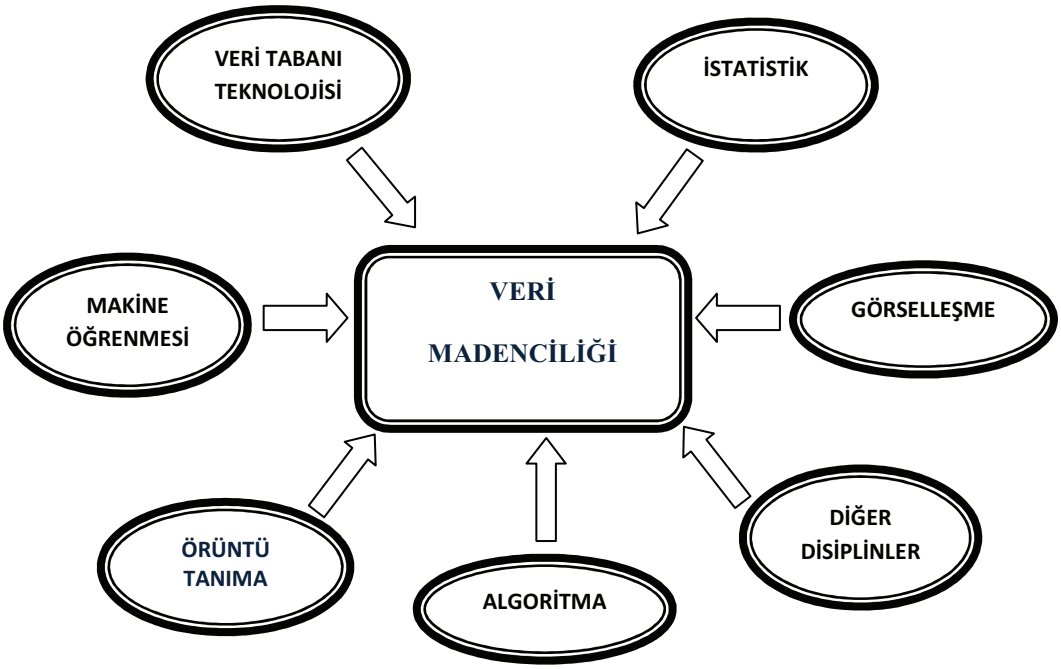
Veri madenciliğinin temeline bakıldığında büyük boyutlu veritabanlarının değerlendirilmesi sürecinde makine öğrenmesi, yapay zekâ ve istatistik gibi uygulamaların kullanılmasıyla ortaya çıktığı bilinmektedir [11].

Disiplinler arası bir yapısı olan veri madenciliğinin temelini oluşturan makine öğrenmesi ve yapay zekâ uygulamalarında kullanılan karar ağaçları, birliktelik kuralı çıkartımı, kümeleme, sınıflandırma ve örüntü tanıma gibi pek çok model ve teknik veri madenciliği alanında da kullanılmaktadır. Makine öğrenmesinde ve istatistiksel uygulamalarda kullanılan veritabanı, veri madenciliğinde kullanılan veritabanlarına göre çok daha küçüktür. Veri madenciliği, makine öğrenmesi yöntemlerine göre kayıp, eksik ve değersiz verileri işlemede daha başarılıdır. Veri madenciliği, çok büyük boyuttaki veritabanları arasındaki veriler arası ilişkileri saptayarak örüntüler oluşturur. Bu ilişki ve örüntülere göre değerlendirme, karar verme ve tahmin süreçlerinde veri madenciliği kullanılmaktadır [16,17].

Veri madenciliğinin diğer disiplinlerle olan ilişkisi, başka bir açıklamaya göre[18]; "veri madenciliği Veritabanı Teknolojisi (Database Technology), İstatistik (Statistics), Yapay Zekâ (Artificial İntelligence), Makine Öğrenimi (Machine Learning), Örüntü Tanımlama (Pattem Recognition) ve Veri Görselleştirmesi (Data Visualization) gibi pek çok teknik alan arasında köprü görevi gören çok disiplinli bir alandır" ifadesiyle belirtilmiştir [18].

Yapay zekâ çalışmalarının dayanağı olan makine öğrenmesi veri madenciliğinin temelini oluşturmuştur. Makine öğrenmesi alanında çalışan uzmanlar ilerleyen zaman içerisinde kestirim algoritmaları bulmuşlardır. Bu algoritmalar sayesinde çıkarsamalar yaparak çözümler üretmeye çalıştıkları literatür çalışmalarında görülmektedir [17-20].

İstatistikçiler ise büyük boyutlu veritabanlarından analiz araçlarının yardımıyla veriler içerisindeki örüntü ve ilişkileri keşfederek, geçerli tahminler yapabilmek için kestirim algoritmalarını kullanmaktadırlar. İstatistikçiler uzun zamandır veriler arasındaki ilişkileri, örüntüleri manuel olarak bulmaya çalışmaktaydılar. Veri madenciliği istatistikçilerin yaptıklarını otomatik olarak yapabilen bir süreç olarak gelişmiştir [19].



Şekil.3. Veri Madenciliğinin Birden Fazla Disiplinden Oluşumu

Kaynak: Han, J. & Kamber M. Second Edition. Data Mining Concepts and Techniques.http://www.cs.sfsu.edu/~huiyang/869-sp2008/lectures/conc-issues-trends.pdf

Veri madenciliği, yapı taşı veriler olan makine öğrenmesi teknikleri, istatistiksel uygulamalar ve veritabanı sistemleri gibi alanlarda pratik çözümler sağladığı için bu alana olan ilgi gittikçe artmaktadır. Veri madenciliği; geleneksel yöntemlere göre çok daha hızlı ve yarı otomatik olarak, gizli ilişki ve örüntülere ulaşabilmesi sebebiyle, yaygın olarak kullanılan bir alan olarak görülmektedir.

Veri madenciliği alanında son yıllarda büyük bir gelişme görülmektedir. Yazılım üreticileri veri madenciliğinin diğer disiplinlerle olan ilişkilerinden yararlanarak bu alanda çalışmalarına ağırlık vermişlerdir. Veri madenciliği uygulamalarının farklı alanlarda da kullanımı ile çözülmesi güç olan durumlar basit bir şekilde çözümlenebilmektedir. Veri alanındaki problemlerin giderilmesi, gizli ilişki ve örüntülerin çıkartılması, istatistiksel tahmin etme, istatistiksel olarak karar verme ve değerlendirme süreçlerinde destekleyici olması, bir bakıma kararı bilimsel olarak vermesi açısından son derece önemli bir yere sahip olan veri madenciliği alanındaki çalışmalar gün geçtikçe çeşitlenerek çoğalmaktadır.

2.5. Kullanım Alanları

Veri madenciliğinin kullanım alanı oldukça geniştir. Durağan verilerden anlamlı bilgilerin otomatik olarak çıkartılması işlemini sağladığı için veri madenciliği birçok alanda kullanılmaktadır.

Veri madenciliğinde zor olan durum, geçmişte neler olduğunun tanımını yapabilecek anlamlı bilgileri içeren model kurmaktır. Bu amaçla geçmişteki durumlardan yola çıkarak yeni durumların nasıl ve ne olacağı tahmin edilmeye çalışılmaktadır [16]. Veritabanlarında tuttuğumuz veriler arasındaki ilişkilere bakılarak yeni örüntüler oluşturulur ve sonrasında oluşan bu anlamlı bilgi şirket ve kurumlarda iş problemlerine çözüm oluşturma aşamalarında da kullanılabilmektedir.

Veri madenciliği bankacılık, astronomi, biyoloji, finans, pazarlama, sigorta, tıp ve birçok başka dalda da uygulanmaktadır. Ayrıca, Amerika Birleşik Devletleri'nde gizli dinlemelerde, vergi kaçakçılığının ortaya çıkarılması gibi durumlarda da kullanıldığı bilinmektedir [5].

Veri madenciliği; metin ve web madenciliği olarak farklı alanlarda da kullanılmaktadır. Metin madenciliği, veri madenciliğinin bir benzeridir. Metin madenciliği ile yapısal olmayan ya da yarı yapısal olan veriler işlenerek analiz ve keşif işlemlerine geçilmektedir. Veri madenciliğinin aslında büyük bölümü metin madenciliği uygulamalarıyla sürmektedir. Metin madenciliği uygulamaları genellikle metin analizi çalışmalarında, metinlerin sınıflandırılması durumunda ve kelime analizleri problemlerinde kullanılmaktadır.

Web madenciliği, internetin temel yapısını oluşturan web sitelerinin içeriklerini inceler. Sitelerdeki başlıklar, sayfalarda bulunan kelimeler, menüler, konu yapısı, resimler vb. gibi içerik bilgileri incelenerek, siteler arasındaki ilişkiler tespit edilir. Bu tespite göre web siteleri sınıflara, kategorilere ayrılabilmektedir [9].

Uygulama alanları	*Uygulama*	*Açıklama*
Perakendecilik	*Benzerlik korunumu* *Çapraz satış*	*Etkin ürünlerin benzerliklerini tespit etmek... Müşteriler için birçok ürün bulunması, ürün satışları arasındaki ilişkiyi tespit etmek.*
Banka ***Kredi Kartı Hareketleri***	*Müşteri ilişki yönetimi* *Müşteri kaybı analizi*	*Müşteri değerlerinin tanımlanması, müşteriler arası benzerlikleri tespit etmek, gelirleri maksimum edebilecek programlar geliştirmek, aldıkları hizmeti iptal etme riski olan müşterileri gösteren raporlar oluşturma.*
Sigorta, Bankacılık, Telekomünikasyon	*Sahtekârlık saptama ve yönetimi*	*Geçmiş veriler kullanılarak sahtekârlık yapanlar için bir model oluşturma ve benzeri davranış gösterenleri belirlemek.*
Hedef Pazar Bulma	*Pazar analizi*	*Benzer özellikler gösteren müşterilerin bulunması, benzer gelir grupları, ilgi alanları, harcama alışkanlıkları. Benzer müşterileri otomatik olarak gruplayarak pazar dilimleri tanımlama ve bu dilimleri pazarlama kampanyaları hazırlarken eğilim analizleri için kullanmak.*
Finans Planlaması ve Bilanço Değerlendirme	*Risk analizi*	*Nakit para analizinin incelenmesi ve kestirimi, talep incelenmesi ve kestirimi, zaman serileri incelenmesi.*

Şekil.4. Veri Madenciliğinin Genel Kullanım Alanları

Kaynak: Kaya, H. ; Köymen, K. (2008). Veri Madenciliği Kavramı ve Uygulama Alanları.

3. VERİ MADENCİLİĞİ UYGULAMA SÜRECİ

3.1. Problemin Tanımlanması

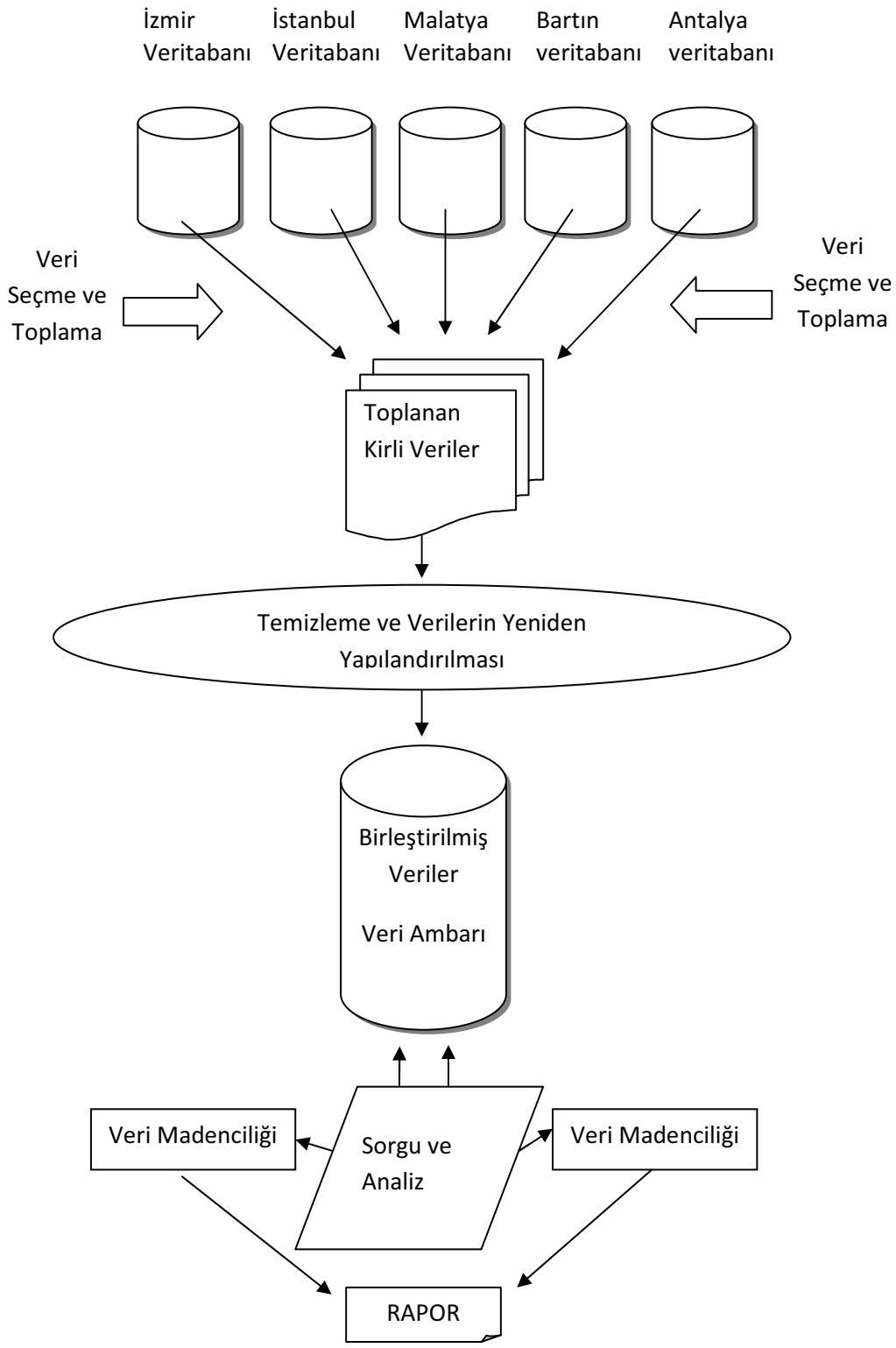
Veri madenciliği uygulama sürecindeki en önemli aşama problemin tanımlanması olarak bilinmektedir. Burada problemin amacı ile tanımı net ve açık bir şekilde ortaya konmalıdır [5]. Problem durumu ortaya konulduktan sonra problemin olası çözümü düşünülürken, verilerin problemin durumuna uygun olup olmadığı, uygulanacak olan yöntem ve tekniklerin istenilen sonuca götürüp götürmeyeceği iyi analiz edilmelidir.

Problemin tanımına göre tasarım planı gerçekleştirilmektedir. Kurulacak olan sistemin çözüm üretebilmesi için problemi ayrıntılı olarak tanımlamak gerekir. Problem tespit edildikten sonraki aşamalarda hazırlanacak olan içerik, uygulanacak modeller ve değerlendirme durumları problemin tanımına göre düzenlenir. Bu aşamada problemin analizinin iyi yapılması gerekmektedir. Problemin analizinin başarılı olmaması durumunda uygulama sürecinin tüm aşamaları yanlış tasarlanabilir. Uygulama sürecinin son aşaması olan değerlendirme sürecindeki testlerle, sürecin en başında tanımlanan problem durumu geçerli olmalıdır.

3.2. Verilerin Hazırlanması

Veri hazırlama süreci; problem durumunun hazırlanmasından sonraki aşamadır. Çalışmalara temel oluşturacak son verilere dönüştürülmesi aşaması buradadır. Modelin kurulumu sırasında çıkabilecek bir sorun sık sık verilerin hazırlanması aşamasına geri dönüş yapılmasına ve verilerin yeniden düzenlenmesine neden olabilmektedir. Veri analizi uzmanlarına burada büyük iş düşer [5].

Problem tüm ayrıntılarıyla tespit edildikten sonra problemin amacına göre ihtiyaç duyulan veriler toplanır. Verilerin hangi kaynaklardan alınacağı kararı, hangi verilerin işe yarar olduğuna karar verilmesi, toplanan verilerin aralarındaki uyuma göre birleştirilmesi, eksik ve kayıp bilgilerin temizlenmesi işlemleri, verilerin ön işlem sürecinde yer almaktadır.



Şekil.5. Veri Madenciliğinde Veri Hazırlama Süreci

Kaynak: Özçakır, F., C.; Çamurcu, A., Y. (2007). Birliktelik Kuralı Yöntemi İçin Bir Veri Madenciliği Yazılımı Tasarımı ve Uygulaması.

3.2.1. Verileri Toplama

Problemin tanımlanmasından sonra çözüm için gereken verilerin toplanması aşamasıdır. Verilerin alınacağı kaynaklar belirlenir. Verilerin toplanmasında şirket, kurumlar kendi veritabanı dışındaki verilerden yani birbiriyle bağlantılı veritabanlarından, oluşturulan veri ambarlarından, hava durumu, harita, nüfus sayımı, internet, sağlık, alışveriş, bankacılık, eğitim, kültürel tabanlı gibi farklı veritabanlarından yararlanabilirler [5, 20, 21].

Kısacası veri madenciliğinde problemle ilgili veritabanı hazırlarken amaca uygun birçok veritabanından veri alınabilir. Düzenli veya düzensiz olarak alınan veriler daha sonra istenilen yapısal veri haline getirilecektir. Daha sonra yapısallaşan veritabanı içerisinde ihtiyaç duyulan bilgiler sorgulanacaktır.

3.2.2. Verilere Değer Biçme

Problemle ilgili olan veriler çeşitli kaynaklardan alınmış olabilir. Farklı veritabanlarından alınan veriler arasında doğal olarak bir uyum sorunu ortaya çıkacaktır. Bu uyumsuzlukların başlıca olası nedenleri veri formatlarının farklı olması, verilerin kayıt tarihlerinin değişik zamanlara ait olması, güncelleme hataları, kodlama farklılıkları, farklı ölçü birimleridir. Farklı kaynaklardan alınan veriler arasındaki uyuma göre verilerin nasıl düzenleneceğine karar verilir. Kurulacak olan modellemeden ne derecede iyi bir sonuç isteniliyorsa, problemin net tanımı yapıldıktan sonra problemle ilgili verilerin büyük bir titizlikle değerlendirilmesinin yapılması gerekmektedir [5, 11, 20].

3.2.3. Veri Temizleme ve Yeniden Yapılandırma

Veri analizinin başarılı ya da başarısız olması açısından bu aşama süreç içerisinde önemli bir yere sahiptir. Bu aşamada farklı kaynaklardan alınan veriler arasındaki uyumsuzluklar giderilerek veri analizi için kullanılacak olan veriler yeniden düzenlenerek tek bir veritabanı oluşturulmaya çalışılır. Değişik veritabanları arasında tutarlılık sağlanmalıdır. Uyumsuzluklar büyük bir titizlikle gözlenmeli ve

giderilmelidir. Uyumsuzluklar rastgele giderildiğinde, sürecin ileriki aşamalarında problemler çıkabilir [5, 22, 23]. Verilerin rastgele değil de belli bir sisteme göre temizlenmesi, indirgenmesi gerekmektedir. Bazı değerleri boş olan her verinin çöpe atılmaması gerekir. Bu boşluklar, gerekli algoritmalar yardımıyla tespit edilerek doldurulmalı ve kullanıma sunulmalıdır.

Gerçeği temsil eden örneklemde bulunan hatalı veya eksik veriler üzerinde işlemlerin yapıldığı bu aşamada, keşfedilen bilginin kalitesini artırma çalışmalarının olduğu bilinmektedir [7,15]. Tipik olarak veri temizleme uyumsuz, tutarsız ve tamamlanmamış verilerin ortadan kaldırılması demektir. Veri temizleme eksik verilerin tamamlanması, 'kirli veri' olarak adlandırılan gürültünün düzeltilmesi ya da ortadan kaldırılması, verilerdeki tutarsızlıkların giderilmesi, hatalı toplanan verilerin temizlenmesi gibi işlemlerden oluşur. Günlük hayattan gelen uyumsuz veriler, tipik değerlerden önemli ölçüde farklıdır. Tamamlanmamış veriler ise; veri girişlerinde yapılan eksikliklerden veya istisnalardan kaynaklanan verilerdir. Bilgi giriş elemanları tarafından ya da müşteriler tarafından bilerek ya da bilmeyerek yanlış veriler kayıt edilebilmektedir. Genellikle yanlış girilen verilerin değerlendirilmeden çıkarılması tercih edilir [11,24,25].

Eksik veriler; veriler arasındaki tutarsızlık nedeniyle atılması, boş değer içeriyorsa silinmesi durumunda, yanlış anlama sonucu kaydedilmemesi veya veri girişi yapan kişinin veriyi önemsiz görmesi sonucunda oluşabilmektedir. Veri temizleme aşamasında eksik veriler tamamlanabilir. Gürültülü veriler düzeltilebilir ve verilerin kendi aralarındaki tutarsızlıklar giderilebilmektedir. Şekil 6' da eksik verilerden oluşan bir tablo görülmektedir.

kayit_ID	musteri_ID	urun_ID	urun_fiyat	urun_kdv
1	1	45	4	8
2	1		100	8
3		42		18
4	2		300	22
5	2	46	360	
6	3	76	45	18

Şekil.6. Eksik Veri Örneği

Gerçekleştirilecek olan sistemin eksik verilerinin yaratacağı sorunları ortadan kaldırmak için bir takım yöntemler kullanılmaktadır. Bu yöntemler aşağıdaki gibidir[4, 24, 25]:

Veritabanında bulunan eksik değerler atılabilir. Veritabanından çıkarılacak olan eksik veriler eğer veri bütünlüğünü etkilemeyecekse çıkarılabilir. Aksi takdirde bu yöntem hem zaman kaybına neden olacak hem de keşfedilecek bilginin kalitesini düşürecektir.

Veritabanındaki eksik değerler elle doldurulabilir. Veritabanındaki eksik değerlerin elle girilmesi zaman alacaktır. Kullanılan veritabanı küçükse ve gerçek hayattaki verilere ulaşma imkânı varsa ve bu verilere kesinlikle ulaşılması gerekiyorsa bu yöntem kullanılabilir.

Tüm verilere aynı bilgi girilir. Örnek olarak doğum tarihi olmayan kişilere, "doğum tarihi yok" şeklinde bir kayıt girilebilir. Yapılan çalışma sonucunda girilen "doğum tarihi yok" bilgisinin anlamlı bir sonuç olduğu çıkarılarak bu kişiler arasında bir ilişki de saptanabilir, benzer özelliklerde oldukları ya da doğum tarihi olmayan kişilerin ürünleri aldıkları saptanabilir. Bu yöntem sayesinde veri madenciliğinin gerçek amacına uygun olarak önceden bilinmeyen durumları ortaya çıkarabilir.

Veritabanındaki eksik olan verilere tüm verilerin ortalama değerinin verilmesiyle sorun giderilebilir. Veritabanındaki veri değeri eğer bilinmiyorsa tüm veri değerlerinin ortalama değeri eksik değerlere girilebilir. Aynı sınıftaki tüm değerlere bakılarak hesaplanan ortalama değer, eksik olan verilerin yerine yazılabilir. Örnek olarak, tüm lüks dairelerin ortalama fiyatının lüks bir dairenin fiyatı olarak işlenmesi verilebilir.

Veri madenciliği algoritması kullanılarak en olası değerin tahmin edilerek eksik verinin giderilmesi gerçekleştirilir. Eksik değerin çıkarımı için K-Ortalama, Regresyon, Karar Ağaçları, Bayesyen Sınıflandırma, Maksimum Beklenti, Zaman Serileri Analizi gibi yöntem ve teknikler veri madenciliği sürecinde kullanılabilmektedir.

“Gürültülü veri” veri değerlerinin yanlış girilmesi, hatalı veri toplama araçları kullanılması, ölçülen değişkende tutarsızlık olması ya da veri toplanması sırasında gelişen sistem dışı hatalar yüzünden oluşmaktadır. Problem için istenilen çözümün kaliteli olması bakımından bu verilerin silinmesi ya da düzeltilmesi gerekmektedir [4]. Verileri temizlemek yalnızca verilerin silinmesi demek değildir. Bazen eksik verileri tamamlamak için bu duruma özel modellemeler yapıldığı bilinmektedir [11]. Kullanılacak olan veri dinamik değerlere sahip ise gürültülü verilerin düzeltilmesi gerekmektedir. Gürültülü verinin giderilmesi için aşağıda bulunan yöntemler kullanılabilir [4,24]:

Kümeleme yöntemi**:** Gürültülü veriler, kümeleme yöntemi ile düzenli hale getirilebilir. Verilerin birbirlerine benzerliklerine göre alt kümeler oluşmaktadır. Kümeleme sırasında bazı veriler hiçbir alt küme içerisinde bulunmayabilir. Bu veriler uç değer olarak kabul edilirse, her biri küme ortalaması değerine ya da alt kümelerin en küçük veya en büyük değerine göre tanımlanır. Bu şekilde veri seti düzenlenmiş olur.

Kümeleme yönteminde veriler kendi içindeki benzerlik ve yakınlıklara göre sınıflandırılırlar. Kümelemenin genel amaçlarından bir tanesi de; yüksek kapasiteli veri yığınları içerisinden tanımlayıcı verileri bularak, veri hacmini daraltarak, aynı küme içerisinde yer alması gereken verilerin belirlenmesi ve bu sayede küme dışında kalan istisnai durumları ortaya çıkarmasıdır.

Bu yöntem ile küçükten büyüğe doğru ya da büyükten küçüğe doğru veriler sıralanır. Atlama yapılmadan sıralama yapılmış veri setinin ortalaması alınır. Bu ortalama veri setinin ortak değeri olarak kullanılır. Örneğin A={3, 4, 5, 7, 8, 2, 9, 10, 16, 30, 17, 21} şeklinde bir veri seti olsun. İlk olarak bu verileri küçükten büyüğe doğru dizilmesi gerekmektedir. Daha sonra dizi, alt kümelerine ayrılır.

A= {2, 3, 4, 5, 7, 8, 9, 10, 16, 17, 21, 30}
A_1={ 2, 3, 4, 5 }
A_2={ 7, 8, 9, 10 }
A_3={ 16, 17, 21, 26}

Oluşan alt kümeler yukarıdaki gibidir. Daha sonra oluşturulan alt kümelerin her birinin sırasıyla ayrı ayrı aritmetik ortalaması alınır. Ortalamalar alındıktan sonra alt küme elemanlarının yerine aşağıdaki gibi o alt kümenin ortalaması yazılır.

A_1={ 3.5, 3.5, 3.5, 3.5 }
A_2={ 8.5, 8.5, 8.5, 8.5 }
A_3={ 21, 21, 21, 21}

Yukarıdaki işlemler yapıldıktan sonra yeni veri seti aşağıdaki gibi olur.

A={ 3.5, 3.5, 3.5, 3.5, 8.5, 8.5, 8.5, 8.5, 21, 21, 21, 21}

Kümeleme yönteminin diğer uygulamasında alt ve üst sınır değerlerine göre düzenleme durumu vardır. Her bir küme için en küçük ve en büyük değerler sınır kabul edilir. Aşağıdaki gibi alt kümelere ayrılan değerler, alt ve üst değer olarak ayrılan hangi değere yakınsa o değeri alır.

A_1={ 2, 3, 4, 5 } ⇨ A_1={ 2, 2, 5, 5 }
A_2={ 7, 8, 9, 10 } ⇨ A_2={ 7, 7, 10, 10 }
A_3={ 16, 17, 21, 30} ⇨ A_3={ 16, 16, 16, 30}

Daha sonra düzenlenen veri seti aşağıdaki gibi sıralanmaktadır:

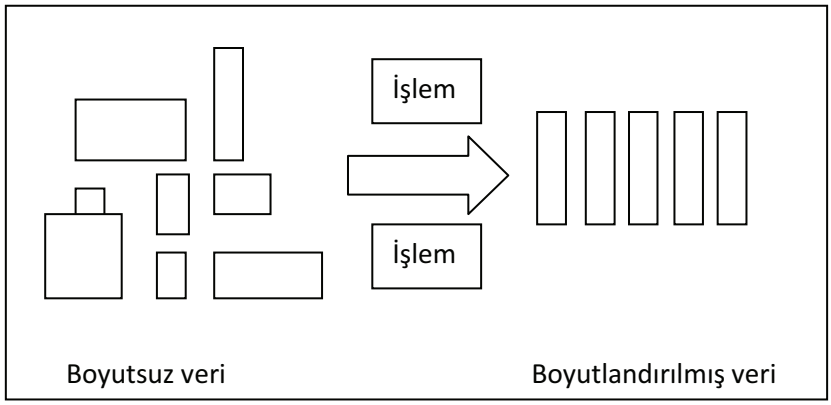
A={2, 2, 5, 5, 7, 7, 10, 10, 16, 16, 16, 30}

Regresyon yöntemi**:** Regresyon yönteminde, veriler bir fonksiyona uydurularak düzeltme işlemi gerçekleştirilir. Uydurulan fonksiyona uymayan değerler aykırı veri olarak kabul edilmektedir [23]. Veri temizleme aşamasında, eksik verilerin tamamlanması durumunda ya da gürültülü verilerin belirlenmesinde regresyon yöntemi kullanılabilmektedir. Eksik değer içeren kayıtlar regresyon yöntemiyle tamamlanabilir veya hiç uygulamaya sokulmadan atılabilmektedir.

Bir diğer durum ise veritabanlarındaki bilgiler yenilendikçe değişkenler farklılık gösterebilir. Veriler arasında tutarsızlıklar olabilir. Kimi veriler zamanla değişebilir. Keşif sisteminin verinin zamanla değişebilmesine karşın duyarlı olması gerekmektedir [8]. Bu durumda, keşif sisteminin kendini sürekli olarak

yenileyebilmesi, dirik olması gerekmektedir. Gelebilecek olan farklı veri türleri ya da verilerin içeriklerinin yenilenebilir olması sebebiyle sistemlerin buna göre kurulması gerekmektedir.

Kurumlar kendi veritabanı haricindeki veritabanlarından bilgileri toplayarak, problem durumunun sağlıklı çözümü için geniş veritabanları oluşturmaktadırlar. Toplanan verilerin formatları birbiriyle aynı olmayabilir. Farklı veritabanlarının birleştirilmesiyle "Şema Birleştirme Hataları" (Schema İntegration Errors) oluşabilir. Örnek olarak; bir veritabanında müşteri girişleri 'müşteri ID' şeklinde işlenmişken başka bir veritabanında ise 'müşteri numarası' olarak işlenmiş veri olabilir. Bu tip şema birleştirme hatalarından kaçınmak için metaveriler kullanılır [24]. "Metaveri"; kaydedilen veri hakkında ihtiyaç duyulan tüm bilgilerin tanım bilgisidir. "Metaveriler" veri ambarlarının inşası, bakımı ve değerlendirilmesi durumunda kullanılabilmektedir.



Şekil.7. Verinin Yapısallaştırılması

Bazı algoritmalar, veritabanlarında bulunan sadece sayısal verilerle ilgilenirken bazı algoritmalar ise veritabanlarındaki kategorik metin verileriyle çalışır. Bazı verilerin boyutlarında farklılıklar olabilir. Her farklı veriden eşit ya da benzer seviyede veri alınması gerekmektedir. Bu durumda Şekil 7'de görüldüğü gibi farklı boyutlarda olan çeşitli veritabanlarından topladığımız veriler üzerinde boyutlandırma ve düzenleme işlemleri yapılarak, kullanacağımız algoritmaya uygun hale getirilmesi gerekmektedir. Farklı formattaki verileri de kullanacağımız algoritmanın formatına dönüştürerek, verileri bir anlamda yeniden yapılandırmış oluruz [4].

3.2.4. Verileri Seçme

Verilerin yeniden yapılandırılmasının ardından modele göre veri seçim yapılır. Yapılacak olan analizde kullanılacak olan verilerin belirlendiği aşamadır. Veriler seçilirken kurulacak olan modelle ilgili olmasına dikkat edilir. Seçilen verilerin sayısı önemlidir. Veriler gerektiği kadar olmalıdır. Aksi halde çalışma eksik kalabilir ya da kirli veri oluşumuna neden olabilir [11].

Verilerin seçimi yapılırken anlamı olmayan gereksiz verilerin seçilmemesine dikkat edilmelidir. Bu gereksiz verilerin seçilmesi, problemin giderilmesi için yararlı olacak olan bilgilerin de kullanımını etkileyebilir. Bazı veri madenciliği algoritmaları konu ile ilgisi olmayan verileri, modele girmemesi için otomatik olarak elemektedir. Gereksiz verilerin giderilmesi durumunun algoritmalara bırakılmaması gerekmektedir. Verilerin temizlenmesi ve yeniden yapılandırılması aşamasında bu gereksiz verilerin giderilmesi daha uygun olabilir [5].

3.2.5. Verileri Dönüştürme

Veriler, veri madenciliği uygulamaları için uygun olmayabilir. Modeli daha güçlü yapmak, modelin kalite ve etkililiğini artırmak için bazı verilerin sayısal bir aralığa indirgenmesi gerekmektedir. Kullanılacak olan modele göre içerik değiştirilmeden verilerin uyumlu olması için format değişikliği yapılmaktadır. Bu aşamada veriler veri madenciliği için uygun hale getirilir [17]. Veri dönüştürme işlemlerinden bazıları şunlardır; düzeltme, özniteliklerin oluşturulması, kümeleme, birleştirme, genelleştirme ve normalleştirme olarak açıklanmaktadır [5].

Veri dönüştürme uygulamasında en çok normalleştirme yönteminin kullanıldığı bilinmektedir. Verilerin 0,0 - 1,0 gibi aralıklara ölçeklenmesi işlemine “normalizasyon” denir. Normalizasyon işleminin çeşitli teknikleri bulunmaktadır. Bunlardan bazıları aşağıda anlatılmaktadır [4,23,24]:

Min-maks yöntemi: Orijinal olan verileri doğrusal olarak normalize eder. ‘Min’ bir verinin alabileceği en küçük değeri, ‘maks’ ise verinin alabileceği en büyük değeri ifade etmektedir. Veriler 0-1 aralığına ya da 0-1 den farklı aralıklara

dönüştürülebilirler. Bir verinin ‘min-maks’ yöntemiyle indirgenmesi formülü aşağıdaki gibidir;

$$S' = \frac{S - \min}{\text{maks} - \min}(\text{yeni}_{\text{maks}} - \text{yeni}_{\text{min}}) + \text{yeni\_min}$$

S′: Verinin normalize edilmiş değeri

S : Verinin orijinal değeri

Z skor yöntemi: Bu yöntemde ortalama ve standart sapma değerleri kullanılmaktadır. Bu yöntemin formülü aşağıdaki gibidir;

$$S' = \frac{S - \text{ort}}{\sigma}$$

σ : Standart sapma

Ort: İlgili alanın ortalaması

Ondalık ölçme yöntemi: Kullanılacak olan verinin ondalık kısmında değişiklik yapılarak normalizasyon gerçekleştirilir. Normalizasyonu gerçekleşecek olan verinin ondalık nokta sayısı, değişkenin maksimum mutlak değerine bağlıdır. Formül aşağıdaki gibidir [23];

$$S' = \frac{S}{10^a}$$

a: maks(|S′ |)< 1 olacak şekilde en küçük tamsayı

3.3. Veri Madenciliği Modelinin Seçilmesi

Veri madenciliği problemleri için birden fazla çözüm yöntemi ve teknikleri bulunmaktadır. Bir problemin çözümü için farklı teknik ve algoritmalar model içerisinde kullanılabilmektedir. Tespit edilen ve tanımlanan problemin çözümü için en uygun modelin bulunabilmesi farklı modellerin denenmesi ile gerçekleştirilebilir. Model, problem için tanımlanan veriler ile sıkı bir ilişki içerisindedir. Veri hazırlama sürecinin dinamik olması ve kurulması düşünülen modelin yenilenme ya da değişme durumları göz önüne alınırsa, model seçim sürecinin sürekli yinelenen

bir süreç olduğu anlaşılacaktır [20]. Veri analizi çalışmaları devam ederken sürekli model kurulum aşamasına dönüş yapılabilmektedir.

Kurulması düşünülen modelin anlaşılabilirliliği önemlidir. Araştırmacılar için uygulamada ilgili kararın niçin verildiğinin anlaşılması önemli olabilir. Bunun için problem tanımına göre birden çok model arasından en uygun olan model ya da modeller seçilmelidir. Modelin kurulumu ve veriler üzerinde uygulamanın gerçekleştirilmesinden sonra daha önce edinilen bilgiler ve test sonuçlarına göre modelin değerlendirilmesi yapılır. Modelin değerlendirilmesi yapılırken problem durumu için belirlediğimiz hedeflerle modelin uyumlu olup olmadığına bakılır. Kurulmuş olan modelin hedefleri ne düzeyde karşıladığı saptanır. Yeterli zaman ve bütçe olması durumunda gerçek verilerle önceden model testi yapılabilir. Model değerlendirme durumunda ayrıca gelecekte oluşabilecek verileri kullanabilme durumları üzerinde de çalışmalar yapılmaktadır [11].

Veri madenciliği probleminin çözümü için kullanılan modeller temelde ikiye ayrılmaktadır. Bunlardan birincisi, elde edilen örüntülerden sonuçları önceden kestirilemeyen veri kümeleri için bilginin tahmininde kullanılan "tahmin edici (predictive)" model, ikinci model ise elde edilen verinin tanımlanmasını sağlayan "tanımlayıcı (descriptive)" modeldir. Kullanılan bu iki modelde fonksiyonlarına göre kendi içinde alt kısımlara ayrılmaktadır. Sınıflama (Classification) ve Regresyon (Regression) tahmin edici modeller; Kümeleme (Clustering), Birliktelik Kuralı (Association Rule) ve Ardışık Zamanlı Örüntü (Sequential Pattern) tanımlayıcı modeller olarak bilinmektedir [3,5].

3.3.1. Tahmin Edici Model (Predictive Model)

Bu modeller geçmiş verilerden yararlanarak gelecek ile ilgili durumların saptanmasında, tahmin edilmesinde kullanılan modellerdir. Veritabanından çıkarılan desenler, örüntüler ve ilişkiler geleceğin tahmini için kullanılır. Bu modelde kullanıcılar, girmesi gereken bazı alanları bilmeseler bile sistem girilmesi gereken değerleri önceki verilere bakarak tahmin ederek doldurabilir [8]. Tahmin edici modelde örneğin, "bu işlemde bir dolandırıcılık var mıdır ?" gibi bir soruya yanıt

aranabilir veya müşterinin kredi kartı talebinde bulunmasının ardından, geçmiş bilgilerine bakılarak alacağı krediyi bankaya geri ödeyebilmesinin tahmin edilmesi gibi durumlar incelenebilmektedir.

Tahmin edici modeldeki durumu bir insanın öğrenmesine benzetebiliriz. Bireyin küçük yaşta çevresinde gördüklerini duyduklarını yani öğrendiklerini aslında şematik olarak zihnindeki veritabanına yerleştirilir. Birey yeni karşılaştığı durumlara zihnindeki veritabanından bilgiler çekerek anlamlandırmalar yapar. Örnek olarak cinsiyet ayrımını öğrenen bir çocuk hafızasına yani veritabanına bu bilgiyi kaydeder. Çocuğa daha önce hiç karşılaşmadığı bir kız çocuğu gösterildiği takdirde hafızasındaki şema içerisinde yer alan bilgilere bakarak kendisine gösterilen çocuğun kız olduğuna karar verir. Çocuğun bu yaptığı işlem aslında bir sınıflandırma, tahmin etme durumudur [4].

Metinsel veriler üzerinde tahmin edici model uygulanırken en temel husus veriler arasındaki gizli ilişkilerin açığa çıkarılmasında izlenilen yoldur. Bu tez çalışmasında metinsel verilerin birlikte geçme durumları temel alınmıştır. Veritabanı içerisinde yer alan kelimeler arasındaki ilişkiler incelenerek çıktılar alınacaktır. Ortaya çıkan sonuçlara göre gerekli tahmin ve değerlendirmeler yapılabilecektir.

3.3.1.1. Sınıflandırma (Classification)

En temel veri madenciliği uygulamalarından biri olan sınıflandırma kategorik sonuçları tahmin etmek için kullanılmaktadır. Sınıflandırma ile yeni bir nesnenin özellikleri tanımlandıktan sonra bu nesne, belirli bir sınıfa yerleştirilmektedir. Burada önemli olan nesneyi yerleştireceğimiz sınıfların özelliklerinin önceden açık ve net bir şekilde bilinmesi gerekmektedir. Örnek olarak kredi kartı başvurusu yapacak olan kişilerin düşük, orta ve yüksek riskli olarak sınıflandırılmasının yapılması verilebilir[11]. Başka bir örnek verilecek olursa; genç erkeklerin küçük spor araba satın alması, yaşlı zengin erkeklerin büyük ve lüks araba satın almaları sınıflandırma kestirimine uygun olacaktır.

Sınıflandırma modeli üç aşamadan oluşmaktadır [26,27];

İlk aşamada her nesnenin sınıf etiketi olarak tanımlandığı ve bu tanımlanan etikete göre sınıfının olduğu varsayılmaktadır. Modelin oluşumunda kullanılacak olan verilerin oluşturduğu kümeye öğrenme kümesi denilmektedir. Eğer sınıf etiketleri önceden bilinmiyorsa "denetimsiz öğrenme (unsupervised learning)", sınıf etiketleri önceden biliniyorsa "denetimli öğrenme (supervised learning)" olarak bu adımda yer almaktadır.

Denetimli öğrenmede; öğrenciye nesneler ve nesnelerin özellikleri ve yine bu nesnelerin tanımlanmış, gelecek aşamalarda tahmini istenecek olan değişkenler verilmektedir. Denetimsiz öğrenmede nesnelerin özellikleri bilinirken, tahmin için kullanılacak olan nesnelerin isimleri verilmemektedir [4]. Ayrıca denetimsiz öğrenmede herhangi bir organize olmadan, yöntem kendi yolunu bulabilmektedir.

İkinci aşamada model, eldeki verilerle uygulamaya konulur. Test örneği rastgele seçilmektedir. Öğrenme kümesinden bağımsızdır. Sınıf etiketi bilinen küme ile model kullanılarak oluşturulan sınıf etiketi karşılaştırılır. Modelin doğruluğu sınıflandırılmış test kümesi örneklerinin toplam test kümesi örneklerine oranıyla belirlenir.

Son aşamada ise modelin kullanımından sonra daha önce bilinmeyen ve görülmemiş "veri sınıf etiketi" tahmini yapılmaktadır.

Sınıflandırma modelinin uygulandığı alanlardan bazıları şunlardır; potansiyel müşteriler için düzenlenen kampanyalar, kredi başvurusunda risk değerlendirme, hastalık tanısı, metin madenciliği uygulamalarında gazete haberlerini konularına göre ayırma uygulamaları, web sitelerini kategorilerine göre sınıflandırma.

3.3.1.2. Regresyon (Regression)

Yaygın olarak kullanılan bir modelleme uygulamasıdır. Regresyon, "Herhangi bir değişkenin bir veya birden fazla değişkenle aralarındaki ilişkinin matematiksel olarak denklem şeklinde yazılmasıdır." Yazılan bu denklem regresyon denklemi olarak adlandırılmaktadır [4].

Regresyon analizi süreklilik gösteren değerlerin tahmininde kullanılan bir modeldir. Regresyon analizinde amaç modele girecek değerler ile çıktı olarak alınacak değerler arasında yüksek ilişkili bir sistem kurarak, sonuç ilişkilerine göre verimli tahminler yapılmasıdır. Tahmin edici değişkenler olarak bilinen girdiler "bağımsız değişken", tahmin edilecek olan değişkenler yani sonuç "bağımlı değişken" olarak bilinmektedir [11,19].

Modelde yalnızca bir tane bağımsız değişken kullanılıyorsa tek değişkenli regresyon analizi, birden fazla bağımsız değişken kullanılıyorsa çok değişkenli regresyon analizi yapıldığı bilinmektedir. Tek değişkenli regresyon analizinde bir bağımsız değişken ile bir bağımlı değişken arasındaki ilişki incelenmektedir. Çok değişkenli regresyon analizinde ise bir adet bağımlı değişken ve birden fazla bağımsız değişkenin modelde kullanıldığı bilinmektedir [26].

3.3.2. Tanımlayıcı Model (Descriptive Model)

Tanımlayıcı model; veritabanlarında bulunan veriler arasındaki önceden bilinmeyen ilişkilerin bulunması, strateji geliştirme ve karar verme gibi süreçlerde kullanılmaktadır. "Çocuk bezi alan müşterinin, mama alma olasılığı diğerlerinden 3 kat fazladır" yargısına varılabilmesi tanımlayıcı modele örnek olarak verilebilir [4].

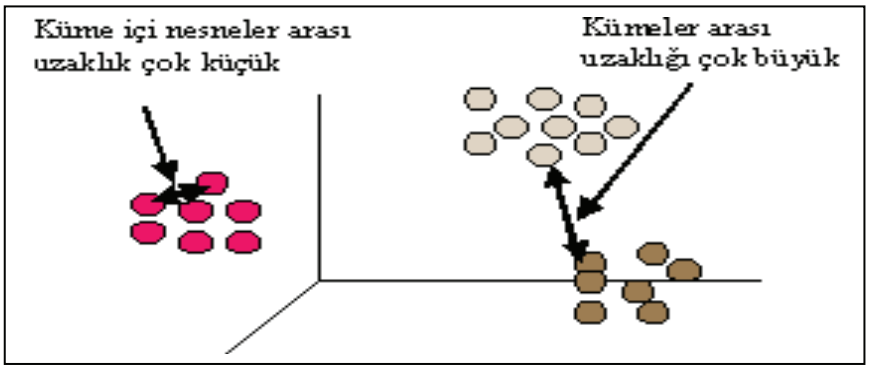
Tanımlayıcı model uygulamaların amacı kesinleşmiş belirli bir hedefi tahmin etmek değildir. Bu modelin amacı veritabanında yer alan veriler arasındaki ilişkileri ve bağlantıları ortaya çıkarmaktır. Bulunan ilişki ve bağlantılara göre yorum yaparak verilerin özelliklerini tanımlamayı gerçekleştirmektedir. Bu sayede verinin ne tür özellikte olduğu bilinir ve bu veri yeni bir veritabanına katıldığı takdirde nasıl bir etki yapacağı konusunda da karar almaya destek olmaktadır [11]. Bu model, ilişkisel veritabanlarındaki gizli birliktelikleri, bağıntıları ve desenleri ortaya çıkararak keşfin yapılmasına olanak sağlamaktadır.

3.3.2.1. Kümeleme (Clustering)

Kümeleme modelinde nesnelerin birbirlerine olan uzaklık ve yakınlıklarına göre gruplara ayrılması sağlanır. Eldeki veriler incelenerek birbirine benzeyen veriler bir kümeye farklı olanlar ise başka bir kümeye yerleştirilirler. Bu sayede heterojen bir veri gurubundan homojen alt kümeler elde edilmektedir. Kümeleme modeli genelde sınıflama sorunlarının çözümünde kullanılmaktadır. Bu model; geniş veri yığınları için tanımlayıcı veriler oluşturarak, veriler içerisinde bulunan doğal kümeleri ortaya çıkarmaktadır. Benzer kümede olması gereken verilerin belirlenmesinde ve küme içerisinde yer almayan, farklı olan verilerin tanımlanmasında kullanılmaktadır [4].

Kümeleme analizinde önceden sınıflandırılmış veri kümeleri yoktur. Verilerin önceden hangi kümede olacağı veya kümelemenin hangi değişkenlerin niteliklerine göre yapılacağı bilinmemektedir. Alan uzmanı tarafından veya bilgisayar yazılımı yardımıyla kümeleme yapılacak veriler ve küme dışında kalacak veriler belirlenmektedir [11]. Her bir kümeye dâhil olan elemanlar o kümenin özelliklerini göstermektedir. Her bir sınıf içerisindeki verilerin nitelik değerleri, o sınıfı belirleyen özellikleri tanımlamaktadır [7].

Şekil 8'de gösterildiği gibi küme içindeki nesneler arasındaki yakınlık veya benzerlik fazla olduğu zaman nesneler aynı küme içerisinde bulunmaktadır. Bir grup içerisinde yer alan bir nesne, diğer grup içerisinde yer alan nesneden farklı nitelikler taşıyorsa farklı kümelerde bulunmaktadır [10].



Şekil.8. Küme Yapısı
Kaynak: Kaya, H., Köymen, K. (2008). Veri Madenciliği Kavramı ve Uygulama.

Kümeleme modeli, sınıflandırma modelinden farklıdır. Sınıflandırma modelinde verilerin önceden sınıfları belirlidir. Yeni bir veri geldiği zaman bu verinin hangi kümeye yerleştirileceği bilinir. Kümeleme modelinde ise önceden belirlenen bir etiket veya sınıf durumu yoktur. Sınıfları bilinmeyen veriler yakınlık ve uzaklıklarına göre kümelere ayrılmaktadır. Bazı uygulamalarda kümeleme modelinin, sınıflamanın ön işlemi olarak görüldüğü bilinmektedir [18].

Kümeleme modeli birçok disiplin tarafından yaygınca kullanılmaktadır. Modelin kullanıldığı alanlardan bazıları; matematik, coğrafya, bilgisayar, tıp, internet, biyoloji ve makine öğrenmesidir. Bilgisayar alanında; resim, ses, karakter tanıma, biyoloji alanında; bitki ve hayvan sınıflandırılması, istatistik alanında; çok değişkenli tahmin ve örüntü oluşturmada, tıp alanında; hastalık tanı teşhislerinde, coğrafya alanında; yeryüzü şekillerinin niteliklerine göre sınıflandırılmasında kullanılabilmektedir. Ayrıca veri madenciliğinin fonksiyonları olan web madenciliği alanında; internet üzerindeki sayfaların haber gruplarına göre kümelenmesinde ve metin madenciliği alanında arşivdeki benzer verilerin gruplandırılmasında uygulanmaktadır [4, 11, 18].

3.3.2.2. Birliktelik Kuralları (Association Rules)

Birliktelik kuralları veri madenciliğini en iyi örnekleyen modellerden biridir. Büyük veritabanları içerisinde bulunan farklı veriler arasında ilişkilerin bulunması birliktelik analizi olarak bilinmektedir. Bu analiz veriler arasındaki potansiyel ilişkilerin saptanmasıyla gerçekleştirilmektedir. Birliktelik analizi, bir veri kümesi içerisinde yer alan verinin başka bir veriyle yüksek sıklıkta birlikteliğinin keşfedilmesidir [11,18].

Müşteri bir ürün aldığı zaman, bu ürünle birlikte başka hangi ürünleri de satın aldığının tespit edilmesi için bu analizden yararlanılmaktadır. Ürün satışlarından bahsedilmesinden anlaşılacağı gibi birliktelik kurallarının ticari veritabanlarında daha çok kullanıldığı bilinmektedir [18,19]. Birliktelik kurallarının kullanım alanının geniş olmasının sebebi ticari alanda başarılı sonuçlar vermesinden kaynaklanmaktadır.

Yarı otomatik olarak hem zaman hem de ekonomik anlamda kazanç yaratan bu model metin madenciliği alanında da kullanılabilir bir uygulamadır. Bu model ile yüksek boyutlu metinlerin analizi gerçekleştirilebilir. Metin içerisinde geçen bir kelimeyle birlikte sık geçen diğer kelimelerde bulunarak, metinde yer alan kelimelerin anlamsal yapıları çıkartılabilmektedir. Ayrıca eksik verilerin tespit edilmesinde metin madenciliği modellerinden birliktelik analizi kullanılabilmektedir.

Birliktelik kurallarının uygulandığı en klasik örnek, "market sepeti" uygulamasıdır. Bu uygulamada müşterilerin hangi ürünleri beraber aldığının analizi yapılmaktadır. Buradaki amaç; müşterilerin aldıkları ürünler arasındaki pozitif ve negatif ilişkileri belirleyerek, müşterilerin satın alma alışkanlıklarını tespit etmektedir. Bu tespit sonrasında elde edilen bilgiyle market sahipleri, müşterilerine ürün satışında farklı promosyon stratejileri geliştirerek uygulayabilirler. Market sahipleri, birlikte satılma yüzdesi yüksek çıkan ürünler bilgisine sahip olduğu takdirde, bu ürünleri aynı sıradaki reyonlara yerleştirerek ürünlerin satışını artırabilirler. Örneğin; "bir müşteri süt satın aldığında sütün yanında ekmek alma olasılığı nedir?" Eğer süt ile birlikte ekmek satın alan müşterilerin oranı yüksek ise ekmek ve süt birbirine yakın reyonlara yerleştirilmelidir. Aynı şekilde deterjan alan müşterinin yumuşatıcı alma oranın da yüksek olduğu belki de mantıksal olarak ta çıkartılabilmektedir. Fakat daha karmaşık durumlar olabilir; süt alan müşterinin yumuşatıcı alması gibi daha önceden kestirimi yapılamayacak ilginç ilişkiler birliktelik analizi yoluyla saptanabilmektedir. Market yöneticileri bu gizli bilgileri elde ederek ürün satışını artırmak için müşterilere farklı kampanyalar düzenleyebilir, ürün kataloglarında birlikte satılan ürünleri aynı sayfaya koyabilir, çapraz satış fırsatları yaratabilirler [4,11,12,28]. Genelde ürün satış gücünün artırılmasında kullanılan birliktelik analizi uygulaması verimlilik sağlanacak her alanda kullanılabilmektedir. Sepet analizi dışında; mühendislik, fen ve sağlık sektörleri alanlarında da kullanılmaktadır [29].

3.3.2.3. Örüntü Tanıma ve Ardışık Zamanlı Örüntüler (Pattern Recognition and Sequential Patterns)

Örüntü tanıma; daha önce belirlenmiş bir nesnenin, elde bulunan veri grupları içerisinde bulunup bulunmamasının tespiti olarak bilinmektedir. Herhangi bir yazılı metni tanımak, parmak izi bulmak, metnin çok benzerini bulmak, yüz tanıma, kan hücrelerinin karşılaştırılması örüntü konularına girmektedir. Örneğin, basketbol oyununda bir oyuncunun basket potasına yönelirken yaptığı hareketleri izleyerek bu oyuncunun daha sonraki sürelerde potaya yönelirken yapacağı adımların önceden bilinmesi örüntü tanıma olarak bilinmektedir [4]. Başka bir deyişle; eski davranış ya da verilere bakılarak yeni olası durumların tahmin edilmesi durumudur denilebilir.

Ardışık zamanlı örüntülerde ise durum birbirini izleyen örüntülerin kullanılması ile veya gözlem sonuçlarının zaman olarak sıralanmasıyla oluşur. Bu sıralama dikkate alınarak veri ilişkileri tanımlanmaktadır. Birbirleriyle ilişkisi olan değişkenlerin ardışık zaman aralıklarında aktifleşmesiyle gerçekleşmektedir. 'Banka taksitlerinden iki veya daha fazlasını geç ödeyen müşterilerin büyük ihtimalle kanuni takibe gidiyor olduklarının belirlenmesi' , 'düşük yağlı veya yağsız yoğurt alan müşteriler %80 ihtimalle diyet süt de satın alacaktır' şeklinde ilişkilerin açığa çıkarılması ardışık zamanlı örüntü uygulamalarına örnek olarak verilebilir [2,11]. Ayrıca ardışık zamanlı örüntü analizleri, gelecek değerlerin tahmininde, hisse değerleri, hava durumu tahmini vb. durumlarda da kullanılabilmektedir.

3.4. Veri Madenciliğinde Kullanılan Teknikler ve Algoritmalar

Veri madenciliği, kayıtlı olan veriler arasındaki gizli ilişkilerin keşfedilmesi ve karar alma sürecinde kullanılan bir uygulamadır. Veritabanlarında bulunan gizli örüntülerin çıkarılması için bir takım veri madenciliği teknik ve algoritmaları kullanılmaktadır. Model seçimi başlığında anlatılan modellerin kullandığı başlıca teknik ve algoritmalar aşağıda açıklanmaktadır.

3.4.1. Karar Ağaçları (Decision Trees)

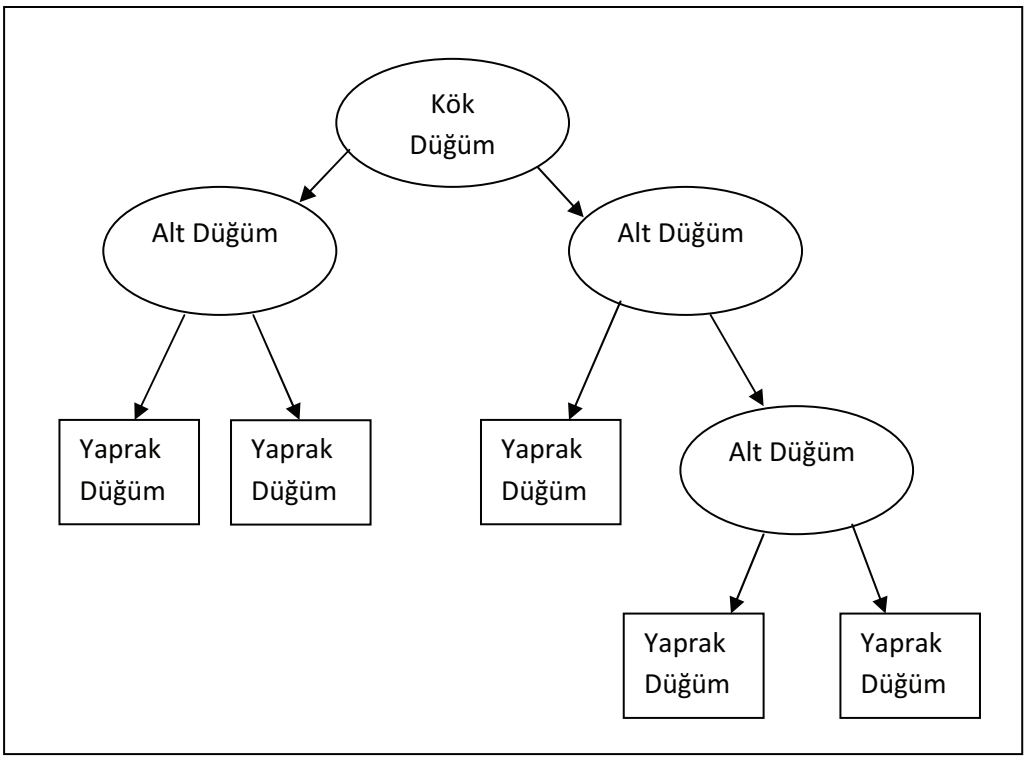
Karar ağaçları tekniğinin Bierman ve Friedman tarafından 1973 yılında önerildiği bilinmektedir [30]. Karar ağaçları, yeni kuşak veri madenciliği tekniklerinden biridir. İstatistiksel ve yapay sinir ağlarındaki yöntemlere göre yorumlanması daha kolaydır. Uygulanması ve yorumlanmasının basit olması bakımından sıklıkla tercih edilen güvenilir sınıflandırma tekniklerinden birisi olarak bilinmektedir [2].

Bu tekniğin uygulaması bir ağacın dallarına ayrılması gibidir. Yapı olarak adından da anlaşılacağı gibi bir ağaca benzemektedir. Ağaçtaki her bir dal ve yaprak birer sınıflandırma sorgusu olacak şekilde düşünülür. Ağaç oluşturulduktan sonra, kökten yapraklara doğru kurallar oluşturulabilir. Kurallar genelde kullanıcıların kolaylıkla anlayabileceği (if-then) türünde yapılardan oluşmaktadır. Veritabanlarına kolay entegre edilmesi, görsel olarak sınıflandırma yapması bakımından da yaygın olarak kullanılan bir teknik olarak görülmektedir [19].

Karar ağaçları uygulamasında iki aşamalı bir işlem yürütülmektedir. İlk aşamada, önceden oluşturulan bir eğitim verisi model oluşturmak için sınıflama algoritması tarafından analiz edilir. Bu aşamaya öğrenme aşaması da denilmektedir. İlk aşamanın sonunda öğrenilen model, karar ağacına veya sınıflama kurallarına dönüştürülmektedir. İkinci aşamada ise bir test verisi sınıflama kurallarına veya karar ağacına uygulanarak doğruluk belirlenir, test örneklerinde belirlenen sınıf ile modelin tahmin ettiği sınıf karşılaştırılmaktadır. Eğer modelin doğruluğunun kabul oranı yüksek ise, oluşturulan kurallar yeni verilerin sınıflandırılmasında kullanılabilmektedir [18].

Karar ağacı oluşturulurken cevabı veritabanında bulunan sorular ve alınan cevaplar önemlidir. Ağaç yapısı, alınan bu soru ve cevaplara göre ilerler. İlk olarak kök düğümü oluşturulmaktadır. İlk düğüm yani kök düğümü oluşturmak en önemli noktadır. Burada seçilecek olan değişkenin, seçilecek olan veritabanını kabaca ikiye bölebilmesi gerekmektedir. Kök düğümden sonra ağacın alt dallarında da yine aynı yol izlenir. Şekil 9'da gösterildiği gibi en sonunda sınıfı temsil eden bir yaprağa ulaşılmaktadır. Yaprağa ulaşılması artık tahmini sınıfın bulunduğunun göstergesidir.

Çünkü artık sorular sorulamadığı için dallara ayrım yapılamamaktadır. Burada farklı algoritmalar farklı türde ağaçlar oluşturabilirler. Ancak her ağaç kendisi gibi farklı sınıflandırmalara sebep olacaktır. Bu nedenle ağaç oluşturulurken düğümlerin yapılaşması ve düğümlerin dallara ayrılmasının tespiti iyi yapılmalıdır[4].



Şekil.9. Karar Ağacı Şeması

Karar ağacı oluşturulurken kullanılan algoritmanın ne olduğu önemlidir. Kullanılan algoritmaya göre ağacın şekli değişebilmektedir. Değişik ağaç yapıları da farklı sınıflandırma sonuçları verebilmektedir. Kök düğüm ağaç yapısında önemli bir yer tutar. Kök düğüme göre izlenecek yol ve dolayısıyla sınıflandırma değişecektir. Kök düğüm veya diğer alt düğümlerin o düğümden dallara ayrıldığında veritabanını kabaca eşit parçalara bölecek şekilde yapılandırılması gerektiği bilinmelidir. Örneğin veritabanında bulunan cevap katılıyorum/katılmıyorum gibiyse iki eşit parçaya, katılıyorum/katılmıyorum/Belli değil gibi üç değişkenli ise mümkün olduğunca üç eşit parçaya bölünmesi istenmektedir. Amaç istenen en kısa cevaba ya da sınıfa ulaşmaktır [4].

Karar ağacına dayalı algoritmaları kod yapısı kaba olarak aşağıdaki gibi çalışmaktadır [31]:

D: Veritabanı, T: Ağaç
T=0; Başlangıç durumu ağaç boş küme
Dallanma kriterlerini belirle
Dallanma kriterlerine göre T kök düğüm belirle
Dallanma kriterlerine göre kök düğümü dallara ayır
Tüm dallar için
Do
Düğüm oluşturmak için değişken belirle
IF(durdurma kriterine ulaşıldı)
Yaprak ekle, Dur
Else
Return

Belirli bir sınıfın muhtemel üyesi olacak elemanların belirlenmesi, çeşitli durumların yüksek, orta ve düşük risk grupları gibi kategorilere ayrılması, gelecekteki olayların tahmin edilebilmesi için kurallar oluşturulması, sadece belirli alt gruplara özgü ilişkilerin tanımlanması, kategorilerin birleştirilmesi gibi alanlarda karar ağaçları yaygın olarak kullanılmaktadır [5].

Bireylerin kredi geçmişlerine bakılarak kredi kararı verilmesi (Credit Scoring), geçmişte işletmeye en faydalı bireylerin özellikleri değerlendirilerek işe alma süreçlerinin belirlenmesi, tıbbi gözlem verilerinden yararlanarak en etkili kararın verilmesi gibi durumlarda kullanılabilmektedir. Ayrıca parametrik modellerin kurulmasında kullanılmak üzere çok sayıdaki değişkenden en önemlilerinin seçilmesi, üretim verileri incelenerek ürün hatalarına yol açan değişkenlerin belirlenmesi gibi karar verilecek durumlarda da iyi sonuçlar verebilmektedir [3].

3.4.2. Naive Bayes

İstatistiksel sınıflandırma modeli tekniklerinden biri olarak kullanılan "Naive Bayes" tekniği, olasılık hesaplama temeline dayanan bir tekniktir. Daha önceden kayıtlı olan belirli bir sınıfa yerleştirilmiş olan veriler kullanılarak, yeni verilerin mevcut olan sınıflardan hangisine dâhil olacağı olasılığının hesaplanması yöntemidir[4].

Bu teknik, herhangi bir anlaşılmaz veya belirsizlik taşıyan bir durumun modelini oluşturarak, bu durumla ilgili evrensel doğrularla ve gerçekçi gözlemler ışığında belirli sonuçlara ulaşabilmektedir. Çıkan sonuçlar üzerinde olasılık hesaplamaları yapılarak belirsizlikler giderilmeye çalışılır. Bu algoritma ile örneğin, sağlık sektöründe bir kişinin tahlil sonuçları değerlendirilerek bir hastalığa yakalanmış olup olmama olasılığının tahmini yapılabilir. Ayrıca müşterilerin harcama artışlarının belirlenmesi de bu tekniğe örnek olarak gösterilebilir. Belirli adet müşterinin yaptıkları alışveriş verilerinin yanında müşterilerin eğitimi, yaş, cinsiyet, gelir gibi verileri de müşterilerin harcama artışlarının olasılıklarının çıkarılmasında kullanılabilmektedir [11].

"Naive Bayes" tekniği, metin belgelerinin sınıflandırılmasında; metin madenciliği uygulamalarında da sıklıkla kullanılmaktadır. Büyük veritabanları arasından hızlı olarak istenilen kuralların çıkarılmasında da başarılı sonuçlar aldığı literatür taramalarında gözlemlenmiştir.

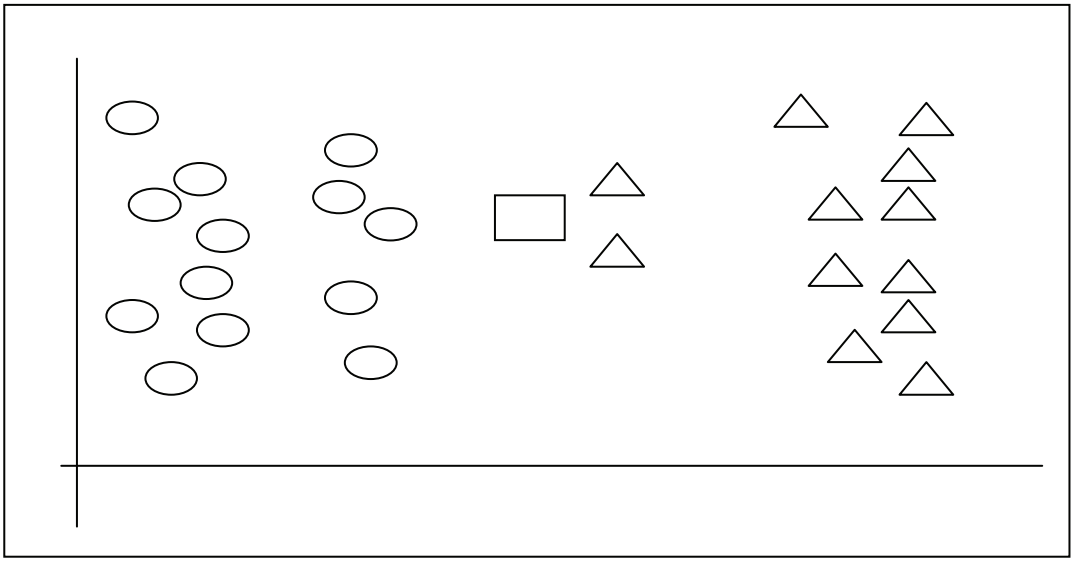
3.4.3. K-En Yakın Komşu (K-Nearest Neighbor)

Mesafeye dayalı sınıflandırma tekniklerinden birisi olan "K- En Yakın Komşu" tekniği günümüzde bilgisayarların ucuzlaması, kapasitelerin artması sebebiyle ve özellikle de çok işlemcili sistemlerin yaygınlaşmasıyla birlikte tercih edilen tekniklerden biridir. Veriler arasındaki ilişkilerin, hızlı ve kolay bir şekilde çıkartılması nedeniyle tercih edilmektedir [2].

Bu teknikte sınıflandırma yapılırken veritabanındaki her bir verinin diğer verilere olan uzaklıkları hesaplanır. Ancak, bir kayıt için diğer kayıtlardan yalnızca

‘k’ adedi göz önüne alınmaktadır. Kullanılan algoritmanın isminden de anlaşılacağı üzere bu, ‘k’ adet kayıtlı nokta, mesafesi hesaplanan noktaya diğer kayıtlara nazaran daha yakındır. Burada ‘k’ değeri önceden seçilmektedir. ‘K’ değerinin küçük olması, noktaların birbirine yakın olanlarının veya benzeyenlerinin aynı sınıfa toplandığı, ‘k’ değerinin büyük olması ise farklılıkları çok olan noktaların, birbirinden uzak olan değerlerin toplanması anlamına gelmektedir. Tipik ‘k’ değerleri 3,5 ve 7’ dır [4].

Örneğin, şekil 10’da bulunan dikdörtgen simgenin etrafındaki hangi nesnelere yakın olduğu, şekilde bulunan nesnelerin hangisinin sınıfına ait olduğu araştırılsın.



Şekil.10. Veri Kümesi

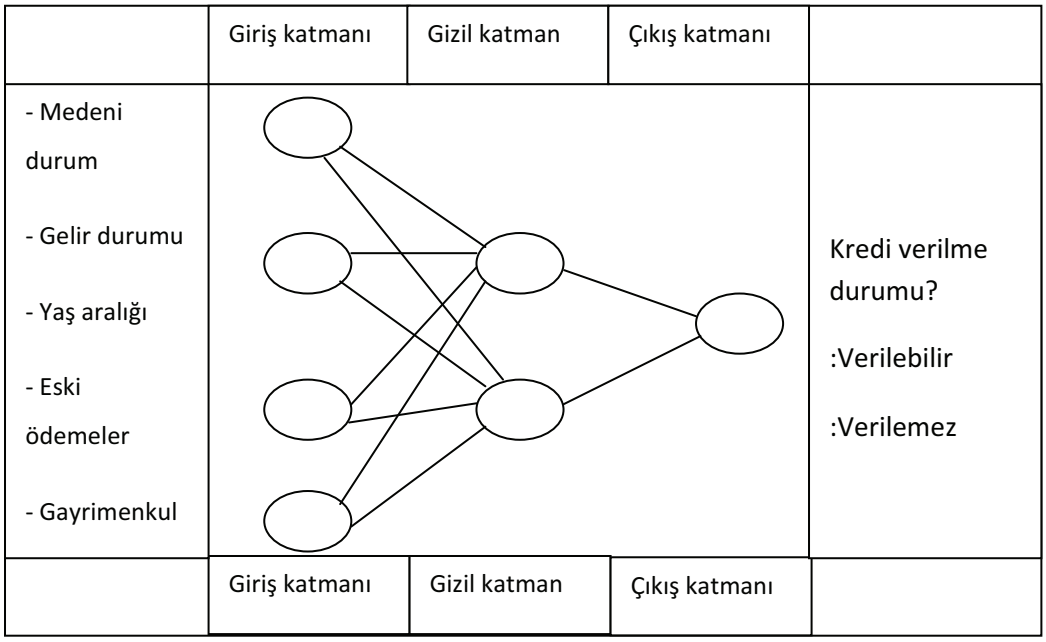
Kaynak: Silahtaroğlu, G., *Kavram ve Algoritmalarıyla Temel Veri Madenciliği.*

Karenin sınıflandırılması yapılırken en önemli nokta k’ ya verilecek olan değerdir. Burada k=2 olarak tanımlandığı takdirde; kareye en yakın iki nesne üçgen olduğundan dolayı karenin üçgen sınıfına ait olduğu söylenebilir. Eğer k = 7 olsaydı, bu durumda kareye en yakın 5 nesne yuvarlak, kareye en yakın iki nesne ise üçgen olduğu için karenin yuvarlak sınıfına dâhil olacağı söylenebilmektedir [4]. Bu teknik, coğrafi bilgi sistemlerinde sıklıkla kullanılmaktadır. Belirlenen bir noktaya en yakın kent, istasyon vs. belirlenmesi aslında “k-en yakın komşu” tekniğinin temelini oluşturmaktadır.

3.4.4. Yapay Sinir Ağları (Artificial Neural Networks)

Yapay sinir ağları, beyin hücreleri olan nöronların çalışma prensibinden esinlenerek geliştirilmiş bir modeldir [32]. Yapay sinir ağlarının 1980 ‘sonrası yaygınlaştığı bilinmektedir [2].

Yapay sinir ağının mimarisi, nöronlar arasındaki bağlantılara bakılarak anlaşılabilir. Yapay sinir hücresi bağlantılarının her birinin kendine ait bir ağırlığı vardır. Bu ağırlıklar öğrenme mekanizması ile belirlenebilmektedir. Katmanlar (nöronları bir araya getiren parçalar) durumlara göre girdi, gizli ve çıktı olabilir. Şekil 11 ‘de gösterildiği gibi girişlerin olduğu nöron katmanına giriş katmanı, çıkışların olduğu katmana da çıkış katmanı denilmektedir. Bu iki katman arasında ise gizil katman yer almaktadır. İki katmandan oluşan sinir ağları; yalnızca giriş ve çıkış katmanlarından oluşmaktadır [30].



Şekil.11. Yapay Sinir Ağ Yapısı

Bu teknik sinir ağının katmanlarından, girdi ve çıktı arasında gelişen birtakım hesaplamaları birleştirerek sonuçlandıran bir uygulamadır. Yapay sinir ağları tekniği uygulama ve yorumlanma bakımından zor bir uygulamadır. Bu teknikte doğru bir model kurulabilmesi için ağın eğitiminin iyi yapılması gerekmektedir. Fazla eğitilmiş

bir ağ tahmin kabiliyetini kaybederken, az eğitilmiş bir ağ ise yanlış tahmin yapabilmektedir [19].

Model kurulduktan sonra sürekli olarak eğitim verilerinin girişi yapılır. Uygulama sonrasında elde edilen sonuçlar ile gerçek sonuçların karşılaştırılması yapılmaktadır. Bu karşılaştırma sonucunda modelde gerekiyorsa iyileştirmeler yapılır. Hata seviyesi minimuma ulaştığı zaman model tamamlanmış olur [11].

Yapay sinir ağlarında kullanılan algoritmalar sınıflandırma modelinde kullanıldığı gibi örüntü ve kümeleme uygulamalarında da kullanılmaktadır. Yapay sinir ağlarının uygulandığı başlıca problem durumları olarak; hisse senedi tahmini, denetim, kredi değerlendirmesi, el yazısı tanıma, parmak izi tanıma, ses tanıma, meteorolojik tahmin yorumlama, hastalık tanısı gösterilebilir [11,26].

3.4.5. Apriori

"Apriori", birliktelik analizinde tekrarlanan öğeleri bulmak için yaygın olarak kullanılan algoritmalardan birisidir [18]. "Agrawal" ve "Srikant" tarafından geliştirilen apriori algoritmasının, bağlantı kurallarının çıkarılmasında en etkili algoritmalardan olduğu bilinmektedir [4]. Algoritmanın ismi, yaygın nesnelerin bir önceki bilgilerinin kullanılmasından ötürü, "prior" anlamında bir apriori olarak tanımlanmaktadır [33].

Bu uygulamada, ilk olarak geniş nesne kümeleri oluşturulur. Geniş nesne kümeleri kullanılarak, bu algoritma yoluyla istenilen kuralların çıkarımı sağlanabilmektedir [34]. Geniş nesne kümelerini ortaya çıkartan algoritmalar veritabanını birçok kez tararlar. Algoritmanın yaptığı ilk taramada bir elemanlı minimum destek seviyesi, kullanıcı tarafından önceden oluşturulan eşik değeri ile karşılaştırılmaktadır. Eşik değerini geçen geniş nesnelerin olup olmadığına bakılır. Bundan sonraki taramalar, bulunan geniş nesnelerle başlar ve yeniden geniş nesne kümeleri oluştururlar. Burada bu geniş nesne kümelerine aday nesne kümeleri denilmektedir. Taramanın sonunda hangi aday nesnelerin gerçekten geniş olup olmadığı kontrol edilmektedir. Nesnenin geniş olabilmesi için destek değerinin,

önceden hesaplanan eşik değerinden büyük olması gerekmektedir [11]. Her yeni tarama bir önceki tarama sonuçlarında bulunan geniş nesne kümelerinden başlanarak yapılmaktadır. Tarama sonuçlarında elde edilen bu geniş nesne kümeleri her defasında önceden hesaplanan eşik değeriyle karşılaştırılır ve nesne kümelerinin destek değerleri hesaplanır. Bu tarama işlemi, oluşturulan nesne kümelerinin tümünün desteklerinin kullanıcının girdiği eşik değerinden küçük olduğu duruma kadar devam etmektedir. Bir başka deyişle tarama işlemi; başka yeni geniş nesne kümeleri oluşturulamayana kadar devam etmektedir [4].

Apriori algoritması sıklıkla perakendecilik sektöründe, tıp, eğitim, internet gibi alanlarda kullanılmaktadır. Bu alanlarda apriori algoritmasının kullanımıyla çeşitli birliktelik kuralları çıkartılır. Bu kurallara göre de keşfe götürecek olan tahminler uzman kişiler tarafından yapılmaktadır. Ayrıca bu uygulama ile, metin içerisindeki sık geçen kelimelerin tespit edilebilir, çıkan sonuçlar çalışmalarda kullanılabilir [33].

3.5. Metin Madenciliği

Veri madenciliği, veritabanlarında bulunan gizli kalmış ilişkilerin ve potansiyel bilginin bazı uygulamalar sonucunda açığa çıkarılması olarak bilinmektedir. Veri madenciliği, erişilebilir ve kullanılabilir veri üzerinden işlemlerini gerçekleştirmektedir. Veri madenciliği uygulamalarının birçoğunda işlemler veritabanlarında bulunan metinsel veriler üzerinden olmaktadır. Veritabanlarında bulunan metinler üzerinden işlemlerin yapılması da metin madenciliğinin veri madenciliği içerisinde büyük bir paya sahip olduğunun göstergesidir. Bu sebeple, veri madenciliği uygulamalarının adının geçtiği yerde metin madenciliğinden de bahsetmek uygun olacaktır.

Bilgisayarlı sistemlerden önce bilgiye erişim için elle indeksleme yapılıyordu. 1996-2001 yılları arasında yayınlanan 164.000 periyodik yayın olduğu düşünülürse, elle indeksleme yapmanın zor, zaman alıcı ve yeterli olmadığı görülmektedir. Şu an internette 2 milyardan fazla web sayfası bulunduğu bilinmektedir. İnternetteki bu kadar çok bilgi yığınlarının indekslenmesinin ve indekslerin güncellenmesinin elle

yapılmasının zorluğunun yanında indeksleme uzmanlarının da öznel yorumlarının indekslemeye karışması sebebiyle aranılan bilgilere ulaşılmasının ötesinde, bulunan bilgiler yanıltıcı sonuçlara da neden olabilmektedir. Bu durumda istenilen bilginin çıkarımının sağlanmasında indekslemenin elle yapılması durumu; zaman kaybı, ekonomik kayıp ve sektörde kalite kaybını gündeme getirmektedir. Bu kayıpların oluşmaması ve binlerce bilgi arasından otomatik olarak istenilen bilgiye ulaşabilmek için doküman madenciliği ya da metin madenciliği adı altında çalışmaların yapıldığı bilinmektedir [35].

Günümüzde internetin yaygınlaşması ve kişisel bilgisayarların hızla artmasıyla, tutulan kayıtlar, sanal ortamdaki kütüphanelerin artması, her gün binlerce haberin haber sitelerine girilmesi, kelimelerin otomatik olarak indekslemesini yapan 'google' gibi arama motorlarının veritabanlarında tuttukları verilerin gün geçtikçe artması, iletişim alanındaki artış gibi durumlar sonucunda metin madenciliği çalışmaları artmıştır [36]. İhtiyaçları karşılamak için doğan metin madenciliği ile yüksek kapasiteli metinlerin analizi ve bu metinler arasında bulunan gizli ilişkilerin keşfi yapılabilmektedir [37].

Bir başka ifadeye göre [38]; yeni nesil teknolojilerin gelişmesiyle birlikte işlemlerimizin birçoğunu bu teknolojik aletler yardımıyla kolaylıkla yapabilmekteyiz. Bu teknolojik aletlerin kullanımının artması ve internete erişimin kolaylığı ile kullanıcıların yaptıkları davranış kayıtlarının veritabanlarındaki yeri de gittikçe artmaktadır. Bu sebeple çevremizde oluşan karmaşık ve büyük bilgi kaynakları keşfedilmeyi beklemektedir. Veritabanlarında kayıtlı olan verilerin çoğunluğu metinsel formatta olduklarından dolayı veri madenciliğinin alt dalı olan metin madenciliği çalışmalarının yoğunluğu artmaktadır. Ancak günümüzde metin madenciliğine destek olabilecek uygulamalar yetersiz olarak görülmektedir [38].

Literatür çalışmasında yer alan bazı bilgi çıkarımı uygulamalarından söz edilecek olursa; Bilgi çıkarımı konusunda ilk çalışmanın 1959 yılında "Luhn" adlı bir bilim adamı tarafından gerçekleştirildiği bilinmektedir. Bu çalışmada "Luhn", sözcüklerin cümleler içindeki kullanılma frekansına bakarak hareket etmiştir. En çok kullanılan sözcüklerin o yazı hakkında en önemli görüşleri verdiğini söylemiştir.

Daha sonraları “Edmundson”, Luhn’un yöntemini geliştirerek çalışmalarında kullanmıştır [39].

“Harris”, kelime anlamları üzerine yaptığı çalışmada “iki kelimenin birlikte geçtiği doküman/cümle sayısının iki kelimenin benzerliğiyle doğru orantılı olduğunu öne sürmüştür”. Kelimelerin anlamsal benzerliklerinin bulunması için çalışmalar yapan “Yuhua Li” ve “Jay J. Jiang”, büyük metin kütüphanelerinde kelimelerin birlikte geçme sıklıklarını ve kelimelerin “Wordnet” hiyerarşisinde birbirlerine olan uzaklıklarını bir arada kullanarak kelimelerin benzerliklerini ölçtükleri, çıkan sonuçları insan deneklerinin yanıtlarıyla karşılaştırdıkları bilinmektedir [40].

3.5.1. Metin Madenciliğinde Kullanılan Yöntemler

Veritabanlarında kayıtlı olan biçimsiz ve karmaşık yani yapısal olmayan verilerin ya da yarı yapısal olan verilerin yapısallaştırılmasında metin madenciliği teknikleri sıklıkla kullanılmaktadır. Verilerin yapısallaştırılması sonucunda veriler uygulamalarda kullanılabilir hale gelir [41,42].

Metin madenciliği yapısal olmayan verilerle uğraşırken, veri madenciliği yapısal olan verilerle teknik çözümlerle ilgilenmektedir [43]. Metin madenciliği yapısal olmayan verilerden bilgi çıkarılmasında birtakım yöntemler kullanır. Yapısal olmayan metin verisinden içerik çıkarmak için kullanılan geleneksel yöntemlerden bazıları şunlardır; anahtar kelimeler, mantıksal aramalar, istatistiksel veya olasılıksal algoritmalar, sinir ağları ve kalıp keşfedici sistemler gibi dilbilimsel olmayan yöntemler. Bu yöntemler karakter eşleştirme yöntemiyle çalışmaktadır. Bundan dolayı içeriği açıklayıcı şekilde, istenilen sonuçları yeterince tatmin edici olmasa da kullanılmaktadır [9].

Bir başka ifadeye göre [35]; geleneksel bilgiye ulaşma yöntemleri sayesinde ana konu başlıklarına yönelik aramalarda yeterli ve başarılı sonuçlar alındığı bilinmektedir. Örneğin google gibi arama motorlarında “matematik problemleri” aranmasında hiçbir sıkıntı yoktur. Matematik problemleri konusunda google karşımıza sorgu sonucu olarak istenilen dokümanları çıkaracaktır. Yalnız bu

dokümanlar belki de yüzlerce sayfa olarak karşımıza çıkacaktır. Bu durumda aramanın daha özel hale getirilip sorgulanması ihtiyacı doğacaktır. Yani sorgulamanın daha özele inilerek "matematik yaş problemleri" gibi anahtar kelimelerle aranması istenilen bilgiye ulaşmada kolaylık sağlayacaktır [35].

Geleneksel yöntemlerden birçoğu halen kullanılmaktadır. Bunlardan bazıları olan; anahtar kelimeler, sinir ağları, olasılıksal yöntemler literatür taramasında birçok çalışmada görülmektedir.

Geleneksel bilgiye ulaşma sistemlerinde karşılaşılan birtakım problemlerden bazıları aşağıda açıklanmaktadır [35];

Birinci problem yazarların dokümanları oluştururken kullandığı kelimelerle, sorgu yapan kullanıcıların girdiği kelimeler yapı olarak aynı olsa da anlam olarak farklılıklar gösterebilmesi durumudur. Bunun sebebi Türkçede kelimeler çok miktarda değişik anlamlar içerebilmektedir. Aynı yapıdaki kelimelerin farklı kültürlerde ve farklı disiplinlerde farklı anlamlar çıkarabileceği de unutulmamalıdır [43]. Bu problem durumunun giderilmesinde doküman içerisinde aranan kelimenin başka hangi dokümanlarda yer aldığı ve hangi kelimelerle birlikte geçtiği göz önüne alınmaktadır [42]. Bu tez çalışmasının uygulamasında da bir kelimenin hangi kelimelerle birlikte metin içerisinde ne sıklıkla geçtiği bulunabilecek açığa çıkan ilişki sonuçlarına göre sorgulanacak kelimenin hangi anlamlarda olabileceği tahmin edilebilecektir.

Ortak kelimeler içeren dokümanların mantıksal olarak benzeştiği, ortak kelime içermeyen dokümanların ise farklı konular olduğu gözlenmektedir. Metin analizinde kullanılan yöntemlerden olan; gizli anlambilimsel dizinleme yöntemiyle benzerlikleri olan kelimeler önceden kümelenmektedir. Bir kelime sorgusuyla doküman aranır. Belki de doküman içerisinde sorgu yaptığımız kelime bulunmayabilir. Ancak önceden dizine attığımız benzer olan kelimeleri sorguladığımız dokümanda görebiliriz. Sorguladığımız kelimeyle benzerlikleri olan kelimelerinde konuyla ilişkili olduğu söylenebilmektedir. Böylece dokümanın içerisinde yer alan kelimelerle, kullanıcının sorgu yaparken kullandığı kelimeler

farklı olsa da konunun benzer kelimeler yardımıyla karşımıza çıkacağı bilinmektedir[42].

İkinci problem, kelimelerin yan yana gelerek farklı anlamlar çıkarabilmesi durumudur. Bir kelimenin cümle içerisinde kullanılırken fiile yakınlığına bağlı olarak anlama kattığı değer farkı olabilmektedir. Bu problem durumunun çözümünde gizli anlambilimsel yollar kullanılabilmektedir. Problemin çözüme kavuşturulmasıyla birlikte giderek büyüyen veri yığınları arasından güçlü ilişkileri çıkartabilmekte kolaylaşacaktır [35,42].

Yüksek kapasiteli metinler içerisinden kaliteli, net ve açık bilgilere ulaşabilmenin bir diğer yolu dilbilimsel yöntemlerdir. Dilbilimsel yöntemlerde akıllı olarak kelimelerden anlamlar çıkarılabilmektedir. Aynı zamanda kelimeler sınıflandırılabilmektedir [9]. Metin madenciliğinde doğal dil işleme yöntemiyle bilgi çıkarımında verimli sonuçlar alındığı bilinmektedir [44].

Yapay zekâ uygulamasından yardım alan doğal dil işleme yöntemleri ile birçok problem durumu hızlı bir şekilde giderilmiştir. Bilgisayar teknolojisinin günlük hayatımızdaki yerinin artmasıyla birlikte bu yeni gelişen yöntemlerde hayatımızda önemli bir yer edinmektedir. Gün geçtikçe ülkemizde de metin madenciliği üzerine çalışmalar artmaktadır. Bu çalışmaların hem geleneksel yollarla hem de dilbilimsel yollarla yapıldığı literatür çalışmalarında görülmektedir.

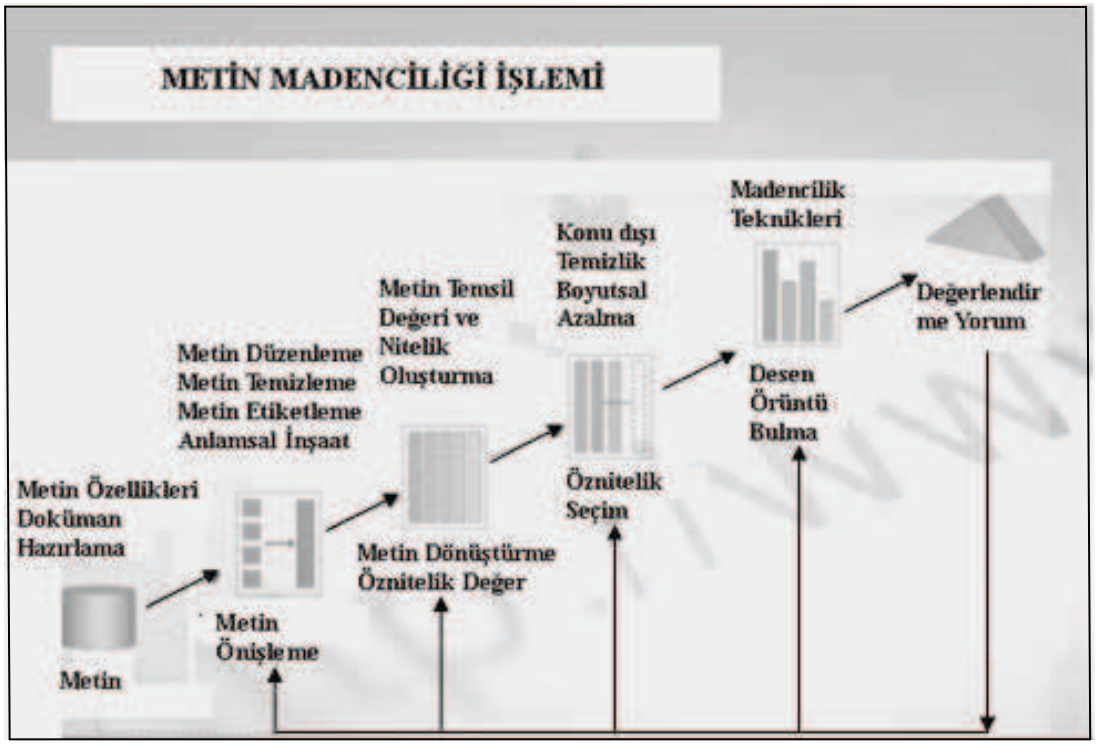
3.5.2. Metin Madenciliği Uygulama Süreci

Doküman madenciliği olarak da bilinen metin madenciliğindeki ana amaç; dokümanlar arasında ayrıca elle bir tasnif yapılmadan, otomatik olarak metnin analizinin yapılabilmesidir.

Metin analizi uygulamalarında genelde anahtar kelimeler, sık geçen kelimeler, metin içerisinde birlikte geçen kelimeler dikkate alınarak işlemler gerçekleştirilmektedir. Otomatik olarak çıkarılan kelimeler tekrar bir modele sokularak ilişkileri meydana çıkartılabilmektedir. Modelden çıkan sonuçlara göre de uzman kişi gerekli tahminlerini gerçekleştirmektedir [21].

Metin madenciliği uygulamalarının amaçlarından bir tanesi metinden anlamlı ve nitelikli özet bilginin çıkartılmasıdır. Böylece metnin içerdiği asıl içerik anlaşılacaktır. Binlerce doküman arasından toplanan veriler gruplandırılarak alanlarına göre kategorilere yerleştirilmektedirler. Veriler üzerinde madencilik uygulama ve teknikleri kullanıldıktan sonra gizli olan potansiyel bilgiler ortaya çıkabilecektir [9,45].

Metin madenciliği uygulamasının başka bir amacı da müşterilerin ürünleri almaları veya terk etmeleri gibi müşteri davranışlarının tahmin edilmesi durumudur. Bu amaçla hareket eden uygulamaların daha yaygın olduğu bilinmektedir[9].



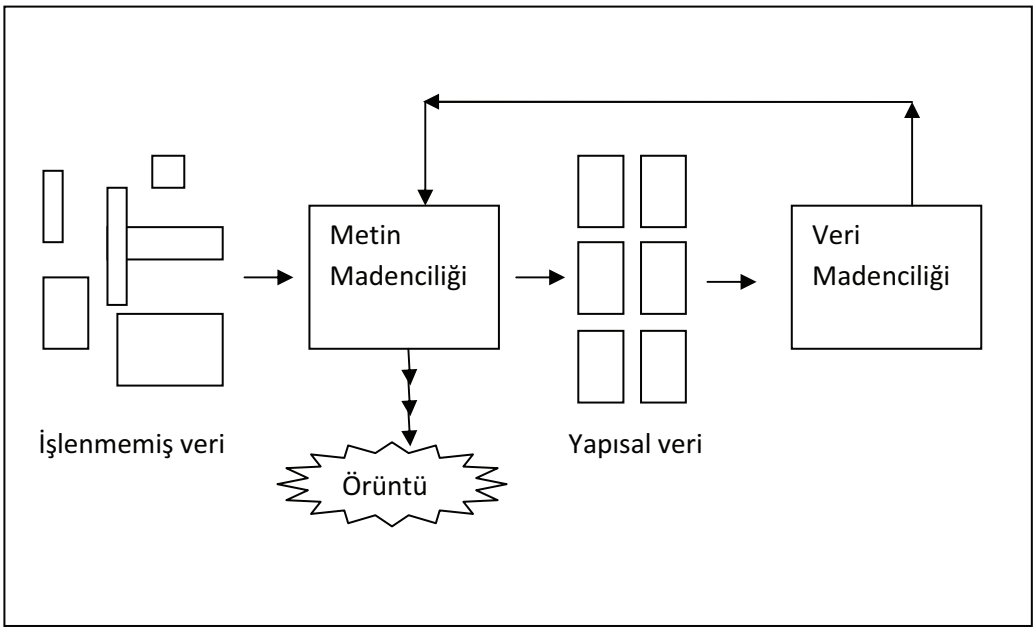
Şekil.12. Metin Madenciliği Uygulama Süreci

Kaynak: Wasilewska A. , Bahar 2006. www.cs.sunysb.edu/~cse634/presentations/ TextMining.pdf

Metin madenciliği uygulama süreci bir anlamda da veri madenciliğinde verilerin ön işleme alındığı aşamadır denilebilir. Yapısal olmayan ya da yarı yapısal olan verinin işlenmesinde metin madenciliği uygulamaları yer almaktadır. Metinlerden elde edilen içerik, bu problem alanının çözümünde kullanılacak olan

tahminsel veri madenciliği modellemesinde girdi olarak işlemlere tabi tutulacak ve anlamlı bilgi çıkarımı gerçekleştirilebilecektir.

Aşağıda yer alan şekil 13 üzerinde de gösterildiği gibi, Metin madenciliği yapısal olmayan veriyi düzenleyerek ve organize ederek, veri madenciliği için kullanılacak olan veriyi hazırlamakta ve daha sonra elde edilen sonuçlar metnin analizinde kullanılmaktadır.



Şekil.13. Metin ve Veri Madenciliği Arasındaki İlişkisel Durum

3.5.2.1. Verilerin Hazırlanması

Metin madenciliğinde bilgi çıkarımı yapacağımız konunun iyi bilinmesi gerekmektedir. Problem alanı olarak görülen konu ile toplanacak olan konular birbirleriyle örtüşmelidir. Ayrıca metin madenciliği hangi alanda uygulanacak ise o alanla ilgili yeterince doküman toplanması gerekmektedir. Burada ne kadar çok veri olursa o kadar çok kaliteli bilgi çıkarımı söz konusu olacaktır. Ayrıca toplanılan veriler farklı veritabanlarından olursa çok daha kaliteli bilgilere ulaşılabilecektir.

Gerçek hayattaki uygulamalar makine öğreniminde olduğu gibi yalnızca sembolik veya kategorik veri türleri ile değil, aynı zamanda tamsayı, kesirli sayılar, çoklu ortam verisi, yardımlı metin, tıp alanında bilgi içeren veri, sanatsal veri

tipindeki farklı yapısal olmayan ya da yarı yapısal verilerle işlemler yapılmasını gerektirebilir. Uygulamada kullanılacak olan veri saklandığı ortama göre ilişkisel veritabanlarında, nesneye yönelik veritabanlarında, internet kaynaklı veritabanlarında da olabilmektedir. Saklandığı ortama göre çeşitlilik gösteren veri tipleri bulunmaktadır [7]. Örneğin, coğrafi alanla uğraşan bir kurumun veritabanında sıklıkla harita kayıtları bulunabilir.

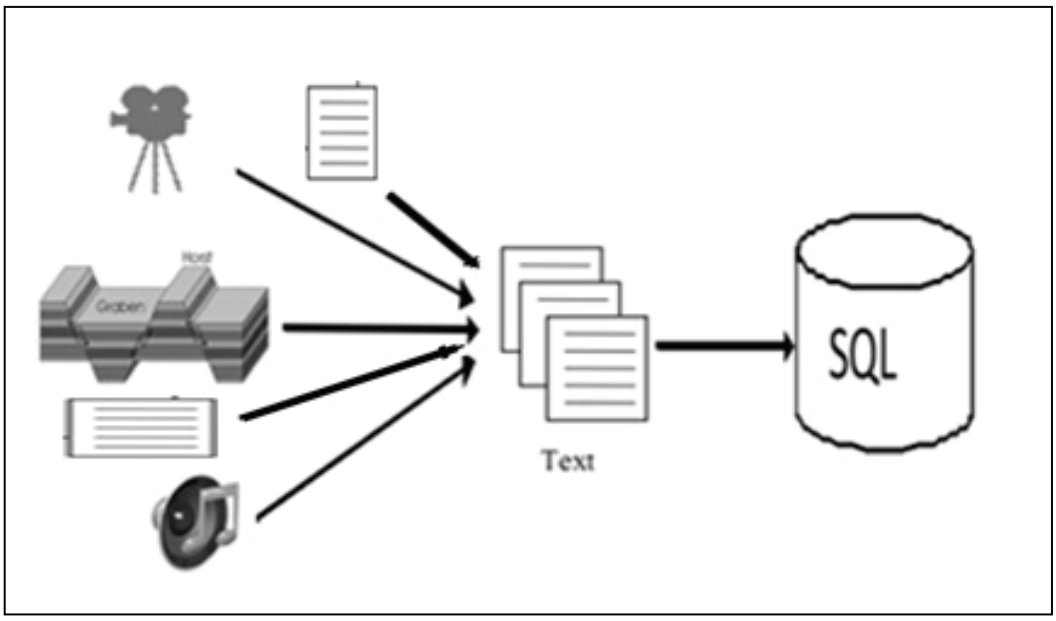
İnternet kullanımının artmasıyla, otomatik olarak kendini indeksleyen web sayfalarının çoğalmasıyla birlikte veritabanı kaynağı sıkıntısı çekilmemektedir. Burada önemli olan nokta şudur; İnternet adresli kaynakların çıkaracağı sıkıntılar iyi tespit edilmelidir. İnternetten alınacak veri karmaşık, düzensiz, olabilmektedir. Sonuçta her kişisel bilgisayar kullanıcısı istediği gibi internete veri yükleyebilmektedir. Burada önemli olan diğer nokta ise daha önce de bahsedildiği gibi seçilecek olan internet kaynağının geçerliliği ve güvenilirliliğinin olmasına dikkat edilmelidir. Her verinin toplanması, kaynak olarak veritabanlarına alınmasıyla ciddi zaman alıcı sıkıntılar yaşanabilir. Bu olumsuz durumun oluşmaması için kullanılacak kaynak uzmanlar tarafından seçilmeli ve kayıt altına alınmalıdır.

Literatür çalışmalarında, metin madenciliği uygulamasında kullanılan kaynaklardan bazıları da kurumların kullanmış oldukları veritabanlarıdır. Buradan alınan veri kaynaklarında karşılaşılabilecek sorunlardan bazıları da eksik verilerin olması, girilmeyen kayıtların (boş değerler) bulunması, gürültülü verilerin olması, verilerde belirsizlik olması ve güncel olmayan verilerin olmasıdır. Bu gibi durumlarda da veritabanı derleyicilerine büyük iş düşmektedir.

3.5.2.2. Dönüşüm ve Temizleme

Problem olan alanda uygulanacak metin madenciliği için, çeşitli kaynaklardan verilerin toplanması gerekmektedir. Bu çeşitli kaynaklardan alınan verilerin hepsi metinsel formatta olmayacaktır. Veri dönüştürmenin ilk aşaması olarak, Şekil 14'de olduğu gibi bu farklı kaynaklardan alınan ses, resim, video vs. gibi farklı formatlara sahip verilerin metinsel formata dönüştürülmesi gerekmektedir.

Çeşitli kaynaklardan alınan veriler bir ön işleme tabii tutulacaklardır. İlişkili ya da ilişkisiz veritabanlarından toparlanan anlamlı ya da anlamsız veriler burada ön işlemden geçirilir.



Şekil.14. Veri Temizleme Yapısı

Problem alanıyla ilgili olan veriler toparlanarak metinsel veriye dönüştürülür. Basılı ortamda bulunan metinsel verilerin de dijital ortama çevrilerek uygulamada kullanılmak üzere kayıt altına alınması gerekmektedir.

İnternet kaynaklı veritabanlarından alınan verilerin metinsel formata çevrilmesiyle birlikte önceden var olan ya da çeviri işlemi sonrasında oluşabilecek yanlış yazım hatalarına, noktalama işaretleri gibi hatalara dikkat edilmesi gerekmektedir.

Düzenlenen verilerden sonra istenilen veritabanının oluşturulması durumunda dikkat edilmesi gereken hususlardan bazıları şunlardır [45,46];

Metinsel verilerde tek başına anlam ifade etmeyen kelimelerin kayıtlardan silinmesi gerekebilir. Bu kelimelerin bilgi erişim sistemlerinde bir işe yaramadığı önceden bilinmektedir. Bu kelimelerin ayrım yaptırma güçleri zayıftır. Örneğin; ve, ki, ile, sonra kelimelerinin anlamsal bir belirleyiciliği bulunmamaktadır.

Türkçe problem durumlarına çözüm getirilmesi düşünülüyorsa, yabancı dildeki kelimelerin temizlenmesi gerekmektedir.

Yarım kalan sözcüklerin bir algoritma yardımıyla ya tamamlanması ya da metinden atılması gerekmektedir. Yalnız sözcüğün metinden atılması durumunda metinden çıkartılacak olan bilgi yeteri kadar kaliteli olmayabilir.

Veritabanı oluştururken cümleler kelimelere parçalanırken cümle sonlarına dikkat edilmelidir. Her nokta '.' İşareti cümle sonu değildir. Kısaltmalarda, tarih formatı gibi yerlerde nokta işareti kullanıldığına dikkat edilmelidir. İşe yaramayacağına inanılan tarihler veritabanından silinebilir.

3.5.2.3. Metinsel Verinin Uygulamaya Konulması

Metin madenciliği uygulamasında kullanılacak veriler üzerinde gerekli temizlik, düzenleme ve dönüştürme işlemleri gerçekleştirildikten sonra analizi yapılacak olan metin, bir ön işlem sürecinden geçirilir. Bu ön işlemde, metin içerisinde yer alan kelimelerin türetilmiş kısımlarından arıtılarak, türemiş haldeki anlam yapısını değiştirmeyecek biçimde köklerine ayırma işlemi gerçekleştirilir. Kelime köklerinin bulunması için de bir takım kök bulma algoritmaları kullanılmaktadır.

Kelime köklerinin bulunmasında bilinen pek çok algoritma vardır. Yalnız bunlardan yabancı kaynaklı olanların Türkçe diline uygulanması mümkün olmamaktadır. Yabancı dillerde genelde ön ek, iç ek ve son ek bulunabilirken, Türkçede ise kelimeler yapısı itibariyle sondan ek alabilmektedir. Bu durumda yapılacak olan çalışmadan verim almak isteniyorsa Türkçe kök bulma algoritmaları kullanılabilir ya da kelime köklerinin bulunduğu bir kök sözlüğü oluşturularak verimli bir sonuç elde edilebilir [45].

Metin analizi çalışmalarında seçilecek metinlerin temsil değeri yani metnin hangi konuda ne ile ilgili olduğunun tespit edilmesi gerekmektedir. Bu işleme "metin temsil değerlemesi (Text Representation)" denilmektedir. Metnin temsil değerinin belirlenmesi için "vektör uzayı temsil tekniği (Vector Space Representation)" iyi sonuçlar verdiği bilinmektedir [45].

Bilgiye erişimin babası olarak bilinen ‘Gerard Salton’ tarafından, bilim dünyasına kazandırılan “vektör uzayı” tekniği bilgi keşfinde yaygın olarak kullanılmaktadır [46]. Vektör uzay tekniğinde her nesne vektör yapısında tanımlanmaktadır. “Nesnelerin sahip oldukları farklı özellikler, vektör uzayının eksenlerini oluşturmakta ve her nesne sahip olduğu özelliklere göre vektör uzayında belirli bir konuma sahip olmaktadır” [44]. En çok bilinen yöntemlerden olan vektör uzayı tekniği, kelimelerin yazılı olduğu dokümanı hangi oranda temsil ettiğini bulmaya yarar. Belirli bir kelime girilerek yapılan sorgulama sonucunda dönen sonuç kümelerindeki her bir dosyanın aranan kelime ile farklı oranlarda temsil yetisine sahip olduğu söylenebilir. Sorgulanan kelimenin kaç kez tekrar edildiğine bakılarak kelimelerin ilişki düzeyleri ve benzerlikleri hesaplanabilir. Ayrıca gruplara ayrılan dokümanların benzerliklerinin bulunmasında, “cosinüs benzerliği” tekniği de kullanılabilmektedir [44, 45].

Vektör uzayı tekniğinde yazılışları aynı fakat anlamları farklı olan kelimelerin sorun yaratacağı düşünülebilir. Bir kelimenin birden fazla anlam alabileceği durumlar göz önünde bulundurulmalıdır. Bu sorunun çözümü için sistem kelime ağırlıklarını belirlerken gerekirse istenilen kelimenin ağırlığını düşürebilir, bu kelimeyle yazılışları aynı olup farklı anlamlarda kullanılan diğer kelimelerin de ağırlıklarını arttırabilmektedir [46]. Ayrıca, bir kelimenin birden çok anlam taşıyabileceği durumlarda eğer kelimelerin birbirleriyle ne sıklıkla geçtiği bilgisi çıkartılabilirse, bu alandaki anlam karmaşası giderilebilir.

“Verilerin sahip oldukları boyutlar arttıkça genellikle çok az sayıda boyut doğrudan kümelerle ilgili olur ancak ilgisiz boyutlardaki veri, çok fazla gürültüye sebep olabilir ve bu da keşfedilecek kümelerin gizlenmesine sebep olur.” Bir diğer husus ise boyut büyüdükçe verinin seyrekleşmesidir. Veri boyutlarının hızla arttığı şu günlerde yüksek boyutlu veritabanlarından bilgi çıkarımı zaman almakta ve zorlaşmaktadır. Verinin boyutu artınca istenilen doğru sonucun bulunma ihtimali de azalacaktır. Bu sebeple boyut küçültme algoritmaları kullanılarak veritabanının boyutunda azaltma yapılabilir [47].

Metin madenciliğinde kullanılacak olan model ve bu modele uygun yöntem ve teknikler hangi alanda bilgi çıkarımını gerçekleştirecekse o alandaki verileri kolaylıkla kullanabilmelidir. Metin madenciliği uygulamalarında kullanılan başlıca yöntem ve tekniklerden bazıları şunlardır;

Makine öğrenmesi yöntemlerinden olan Naive Bayes yöntemi, Sınıflandırma modelinde kullanılan en başarılı tekniklerden birisidir. Bu yöntem bir doküman içerisindeki verileri birbirinden bağımsız olarak düşünmektedir. Bir kelimenin diğer kelimelerle ilişkisi bu yöntemde önemli değildir. Bu ilişkisiz durum olumsuz gibi görünse de Naive Bayes yöntemi metin madenciliği ve klasik veri madenciliği uygulamalarında başarılı sonuçlar verebilmektedir [43,46].

K- En Yakın komşuluk algoritmasının kullanılması kolaydır. Bu uygulama yeni sorgu örneğini sınıflandırmak için kullanılan bir sınıflandırma algoritmasıdır. Bu uygulama daha önce klasik veri madenciliği teknik ve algoritmaları içerisinde detaylı olarak incelenmiştir.

Rocchio algoritmasında, her kategoriye ait eğitim örneklerden alınarak prototip bir doküman vektörü oluşturulur. Hangi kategoriye ait olduğu sorgulanan dokümanın, oluşturulan prototipe olan mesafesine bakılır ve buna göre süzme işlemi gerçekleştirilir. Burada verilerin eğitilmesi gayet hızlı olmaktadır [46].

Metin madenciliğinde kullanılan bir diğer metod ise “Destek vektör makineleridir”. Doğrusal olarak ayırt edilebilen ve edilemeyen veri kümelerinin sınıflandırılmasında etkili bir yöntemdir. Yüksek boyutta doğrusal sınıflandırma işlemi yapılabilmektedir. Metin sınıflandırma da olduğu kadar farklı pek çok alanda da kullanılmaktadır [27, 46].

İnternetin yaygınlaşması, çeşitli veri kaynaklarının artması gibi veritabanlarını büyütecek gelişmelerin olması sonucunda yapısal olmayan verilerin artmasıyla birlikte metin madenciliği alanında da sürekli bir gelişme, yenilenme olacağı kaçınılmazdır. İnternet kullanıcılarının artması, her ortamda kayıt yapılması gibi etmenlerin, veritabanlarındaki veri sayılarını artırması ile daha da büyüyecek olan veritabanları için belki de şimdi kullanılan yöntemler yeterli olmayacaktır. Bu durumda metin madenciliği alanında uygulanan tekniklerin gelecek durumları daha

iyi tahmin edebilmesi ve tanımlayabilmesi için sürekli gelişen, dinamik bir yapıda olması gerekmektedir.

3.5.2.4. Bilgi Çıkarımı

Yüksek boyutlu veritabanlarından, insan müdahalesini asgari seviyeye indirerek, otomatik olarak çalışarak, işe yarayacak olan bilginin çıkartılması işlemini metin madenciliği algoritmaları başarıyla sonuçlandırmaktadırlar.

Metin madenciliğinin son aşaması olan bilgi keşfi, beklenmedik ilişkilerin çıkarılması gibi durumlar burada gerçekleşmektedir. Analiz ve sınıflandırma işlemleri gerçekleştirilmiş olup bilginin yapılandırılmış kümeler içinden çıkarımı söz konusudur. Bazen bu kümeler arasından çekilecek olan bilgiler, çözüm isteyen şahsın ihtiyacına göre farklı şekillerde olabilmektedir. Çıkarılan örüntü ve ilişkilere dayalı olarak metin analizi uzmanları isteğe göre bilgileri bu sonuçlardan çıkartırlar ve "bilginin keşfi" başarıyla gerçekleştirilmiş olur.

3.5.3. Uygulama Alanları

Metin madenciliği uygulamaları birçok alanda kullanılabilmektedir. Metin madenciliği uygulamalarının kullanıldığı başlıca alanlar aşağıdaki gibidir [9,45]:

Müşteri ilişkileri yönetiminde (Customer Relationship Management); müşterilerden elde edilen çeşitli metinsel verilerden nitelikli bilgiler çıkartılır ve bu bilgilere göre terk etme ve çapraz satış durumları tahmin edilebilir. İnsanların yeni tüketim alışkanlıklarının keşfinde, kişi profili, içerik analizi yapılarak kişiye özel kampanya üretilmesinde, müşterilerin internetteki firmalar ve ürünleri hakkındaki görüşlerinin tespitinde kullanılabilmektedir. Ayrıca; Sağlık ve Biyoloji alanlarında madencilik çalışmaları sonucunda tanı, tedavi bilgilerinin tahminsel kararı verilmesi, bu alandaki yeniliklerin takip edilmesi ve birimlere özel gruplandırılması, hastalık raporlarıyla hastalığı tetikleyen bilinmeyen etmenlerin tahmini gibi alanlarda da rastlanmaktadır.

Güvenlik alanında; sağlık, sigorta ve hükümet tarafından toplanan yüksek boyutlu veriler arasından anormal ilişkiler aranarak dolandırıcılık tespiti yapılır. Polis vaka kayıtlarıyla yeni vaka kayıtlarının ilişkilendirilmesi, dolandırıcılık yapacak olan kişilerin yazışmalarının analiz edilmesiyle dolandırıcılık şebekesinin ortaya çıkartılması gibi uygulamalarda kullanılmaktadır.

Pazar araştırmalarının etkililiği, verilen bir metinden özet çıkarma, farklı kaynaklı aynı konulu haberlerin tespiti, bir metnin farklı bir dile otomatik çevrimi, akademik bir çalışmanın çalıntı olup olmadığının tespiti, isimsiz bir metnin yazarının tespiti gibi durumlarda da sıkça metin madenciliği uygulamalarına başvurulmaktadır.

3.6. Metin Madenciliği ve Türkçe Kelime Yapılarının İlişkisi

Türkçenin en önemli özellikleri arasında yer alan çok anlamlılık durumu dili zenginleştirmektedir. Metin analizi uygulamasında kelimelerin çok anlamlı olması bir problem durumu olarak görülmektedir. Bu yüzden çok anlamlılık araştırmacılar için önemli bir kavramdır.

Bir kelimenin yapısal olarak birden fazla anlamı araştırılması durumunda bu çalışma da olduğu gibi farklı veritabanlarından çeşitli veriler toplanabilir ve bu farklı kelimeler içerisinde sorgulanacak olan kelimenin bağlantıları araştırılabilir. Açığa çıkabilecek bilgiler ışığında problemli kelimenin farklı anlamları tahmin edilebilir.

Metin madenciliği alanında bazı araştırmacılar, Türkçe kelimelerin köklerine inerek, çalışmalarını geliştirmek zorundadırlar. Bu çerçevede Türkçe kelimelerin anlamlı köklerine kadar çözümlenmesi gerekmektedir. Türkçe kelimelerin köklerinin bulunması için literatürde birtakım çalışmalar yer almaktadır. Bu tez çalışmasının uygulamasının birinci modülünde de kelimeler, kök bulma algoritması sayesinde köklerine ayrıştırılacaktır.

4. TÜRKÇE'DE BİRLİKTE KULLANILAN SÖZCÜKLERİN METİN MADENCİLİĞİ YÖNTEMİYLE ANALİZİ

4.1. Çalışmanın Amacı

Günümüzde yüksek kapasiteli, düzenli veya düzensiz metin yığınları içerisinde saklı olan bilgiyi çıkarabilmek için metin madenciliği tekniklerinden yararlanılmaktadır. Bu çalışmanın amacı da metinsel veriler içerisindeki kelimelerin birlikte geçme sıklıklarının bulunması ve bunun sonucunda yapısal halde bulunan kelimelerin gerçek anlam haritasının çıkarılmasını sağlamaktır. Bu çerçevede Türkçe kelimeler arasındaki gizli ilişkilerin anlamsal bir harita yardımıyla analiz edilmesi ve analiz çıktılarına göre de kelimelerin anlam tahminleri gerçekleştirilmiş olacaktır.

4.2. Çalışmanın Önemi

Bu çalışmanın veri madenciliği ve alt dalı olan metin madenciliği tekniklerinin hangi alanlarda nasıl katkı sağlayabileceği bilgisinin verilmesinin yanı sıra bilgisayar, sağlık ve alışveriş başta olmak üzere birçok alanda bu tekniklerden yararlanmak isteyen araştırmacılara veri madenciliği yöntem ve tekniklerinin anlatılmasıyla birlikte yol gösterme konusunda önemli bir değeri vardır. Ayrıca çalışılan konunun yapı taşının kelimeler olması itibariyle Türkçe alanında çalışan araştırmacılara ve Türk dili ve edebiyatı dünyasına da destek sağlayabilecek önemli bir kaynak olacağı düşünülmektedir.

4.3. Çalışmanın Kapsam ve Kısıtları

Bu çalışmada teknolojinin hızla ilerlemesiyle ihtiyaç haline gelen veri madenciliği ve metin madenciliği yöntem ve teknikleri ile büyük bir problem alanı olan Türkçe kelimelerin anlam belirsizlikleri incelenmiştir.

Türkçe kelimeler arasındaki gizli ilişkiler metin madenciliği teknikleriyle analiz edilmiş olup, kelimelerin birlikteliklerine göre yakınlıklar bulunmuş, anlam tahminleri gerçekleştirilmiştir. Uygulama içerisinde temel örneklerden başlanmış ve farklı kapasiteli örneklerle sonuçlar tutarlı hale getirilmiştir.

Fen bilimleri alanındaki çoğu çalışmada olduğu gibi bu çalışmanın da bazı kısıtları vardır. Bu kısıtlardan aşağıda bahsedilecektir:

Veri toplama aşamasında İnternet tabanlı kaynaklardan yararlanılmıştır. Bu durumda şöyle bir sorun ortaya çıkmaktadır; İnternet kullanıcılarının özgür bir şekilde internet ağına veri yüklemeleri sonucunda ortaya çıkan düzensiz veriler sorun yaratmıştır. İnternet ortamına kimin nasıl veri yüklediğinin bilinmemesi sebebiyle internetten alınan verilerin kullanılabilirliliği şüphe yaratmaktadır. Kirli verinin çok olması zaman ve ekonomik anlamda süreçte problemlerle karşılaşılmasına yol açmıştır. Verilerin toplama ve hazırlama aşaması tez hazırlama sürecinde büyük bir paya sahiptir.

İnternet ortamından alınan veriler büyük bir veritabanı oluşturduğu için, tek tek bütün kelimelerin Türkçe kelime olup olmadığı sorgulanmamış, bu kelimelerin Türkçe kurallarına uygun olarak yazılıp yazılmadığı kontrol edilmemiştir. Bu durum göz önünde bulundurularak çoğunlukla Türkçe kaynaklı edebi ve sanat eserleri tercih edilmiştir.

Çalışmacının, Türkçe dilbilimsel konulara hâkimiyetinin zayıf olması sebebiyle kelime alt yapıları detaylı olarak çalışmada incelenmemiştir. Kaldı ki dilbilim konuları başlı başına incelenmesi gereken uzun soluklu araştırmaları kapsamaktadır.

Kök bulma algoritmasına koyulacak olan metin dosyasının "UTF-8" formatında olması gerekir. Aksi takdirde Türkçe karakterler tanımlanmayacaktır.

Bu uygulamada 13.926 metin sayfası; 4.156.123 kelime toplanmıştır. Fakat teknolojik imkânların kısıtlı olması sebebiyle ve algoritma yapısının tam performans gösterememesinden ötürü bu yüksek veritabanından performans alınamamıştır. Bu sebeple toplanılan veritabanı içerisinden rastgele 78.112 kelimelik bir veritabanı oluşturulmuş, uygulama bu veritabanı üzerinden devam etmiştir.

İlk modülde Java programlama dili ile kodlanmış Türkçe doğal dil işleme kütüphanesi olan Zemberek uygulaması kullanılmıştır. İkinci modülde ise "Microsoft

Visual Basic 2008" ile "Windows" platformunda çalışılmıştır. Veritabanı olarak "Microsoft SQL Server Express 2005" ve "Microsoft Office Access 2007" programlarından yararlanılmıştır. Bu uygulama süresi de veritabanının büyüklüğüne bağlı olarak artmaktadır. Veritabanının büyük olmasından ötürü kullanılan memory alt yapısı, oluşturulan büyük ölçekli sorgu karşısında yetersiz kalmaktadır.

Bu çalışmada kullanılan bilgisayar sisteminin özellikleri; İntel Core 2 Duo T7300 2.0 800Mhz, 4mb Cache, 2gb ddr2 Ram, işletim sistemi Windows XP' dir. Bu uygulamanın tam anlamıyla verimli olabilmesi için donanımsal özelliklerin günümüzde kullanılan en iyi performansı sağlayabilecek sistemden oluşması gerekmektedir.

4.4. Verilerin Toplama Süreci

İnternet üzerinden toplanılan veriler ayrı klasörlerde kategori haline getirilmiştir. Toplanılan bu veriler seçilirken genelde Türkçe yazım kurallarına uyumlu olabilecek metinler seçilmiştir. Yani bu metinler çoğunlukla roman, masal, hikâye, fıkra, gezi yazısı, deneme, destan, makale, mektup, biyografi, mizah gibi çeşitli edebi eserlerden alınmıştır. Ayrıca sinema, spor, eğitim, teknoloji alanlarından da çeşitli metinler işlemde kullanılmak üzere kayıt altına alınmıştır.

4.5. Çalışmanın Uygulama Süreci

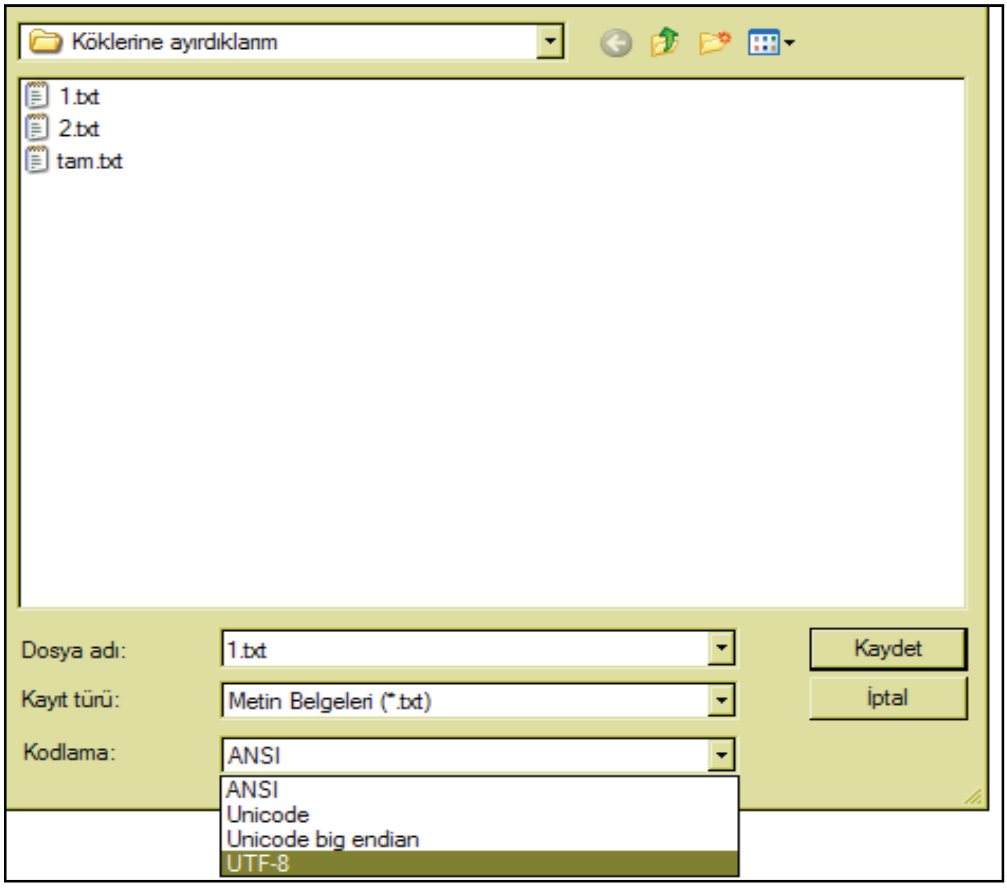
Bu çalışmada, metinlerde yer alan kelimelerin köklerinin bulunması için zemberek kök bulma algoritması ve köklerine ayrılan kelimelerin birlikteliklerinin bulunması için ise tez süresince hazırlamış olduğum bilgisayar programı kullanılacaktır. Kelimelerin birlikteliklerine bakılarak anlam haritası oluşturulacak ve bu anlam haritasına göre de kelimelerin gerçek anlam bilgisi tahmini yapılabilecektir.

Tez kapsamı süresince hazırlanılan programın arka planında; veritabanından gelen bilgilerin dizilerde tutulması işlemi, dizilerde tutulan verilerin belirli bir

algoritma sürecinden geçirilmesi ve bu süreçte eşleştirme yönteminden faydalanılarak sonuçların program arayüzü'ne yansıtılması aşamaları yer almaktadır.

4.5.1. Kelime Köklerinin Bulunması

İnternet üzerinden alınan dokümanların dosya uzantıları farklı olabilmektedir. "pdf", "doc", "html", "txt", gibi farklı formatlarda metin dokümanlarıyla karşılaşılmıştır. Tüm metinler Şekil 15 üzerinde gösterildiği gibi "UTF-8" kodlama türüne çevrilerek, dosya uzantısı "txt" olacak biçimde kayıt edilmiştir. Kök bulma algoritması bu kodlama türü ile tüm karakterleri tanımakta sorunsuz bir şekilde çalışmaktadır.



Şekil.15.UTF-8 Kaydı

Format değişikliği yapıldıktan sonra metinler bir ön işlem sürecinden geçirilmiştir. Metinde yer alan şekil, tablo gibi yazı haricindeki bilgiler silinmiştir. Yazılar üzerinde de birtakım düzeltmeler yapılmıştır. Kelimeler, kök bulma

algoritmasının kütüphane süzgecinden geçeceği için üzerinde çokta fazla değişiklik yapılmamıştır.

Metinlerin hepsinin bir anda kökü bulunamamıştır. Çünkü kök bulma programı her durdurulduğunda, metinleri baştan incelemeye başlamaktadır. Bunun için toplanılan veriler bölümlere ayrılmıştır. Bölmelere ayrılan metinler sırasıyla "c" sürücüsü içerisinde 'tez.txt' dosyasına kaydedilmiştir.

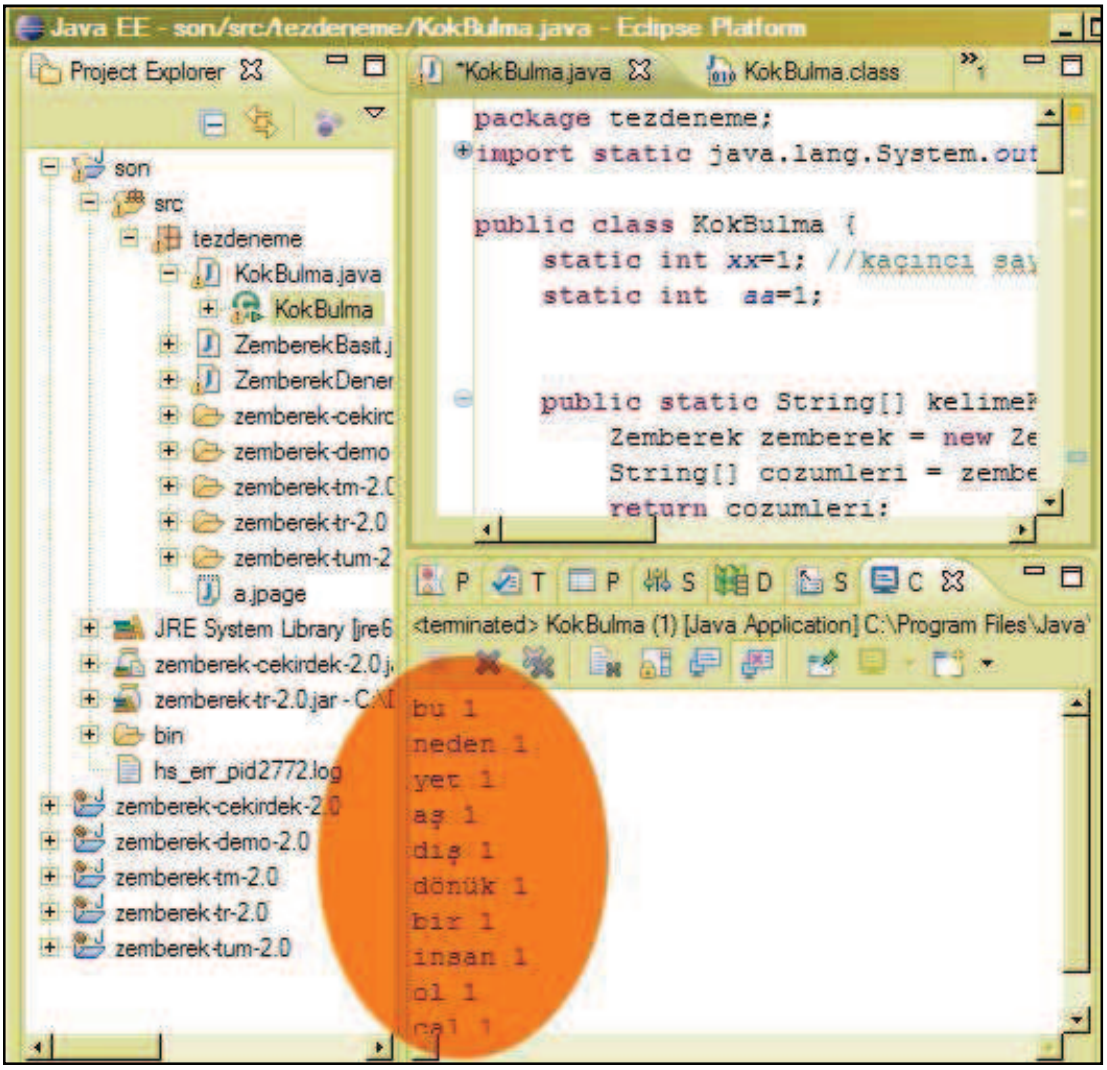
Zemberek yazılımı, Şekil 16'daki kod bloğunda görüldüğü gibi bu adresten işlenecek olan dosyayı bulur, okuma ve köklerine ayırma işlemlerini gerçekleştirir. Daha sonra kökleri bulunan kelimeleri, yine Şekil 16'daki kod bloğunda görüldüğü üzere c sürücüsü içerisinde 'sonuc.txt' dosyası olarak kayıt eder.

```
String filePath = "c:\\tez.txt";
String fileWrite = "C:\\sonuc.txt";
try{
        KokBulma.OkuveAyir(filePath,fileWrite);
```

Şekil.16. Kod Eklentisi

Yeni bir işlemde c sürücüsü içerisinde bulunan 'tez.txt' içerisindeki metin silinir, köklerine ayrılacak olan yeni metin bu dosyaya eklenir ve kelimeler sırasıyla şekil 17'de görüldüğü gibi köklerine ayrılırlar. Kök bulma işlemi bu şekilde döngüsel olarak veritabanındaki son kelimeye kadar devam etmektedir.

Bu aşamada işlemlerin hızlı ilerleyebilmesi için bilgisayarın donanımsal özelliklerinin hızlı performans yapısına sahip olması gerekmektedir. Aksi takdirde işlemler veritabanının büyüklüğüne bağlı olaraktan ciddi boyutlarda zaman alacaktır.

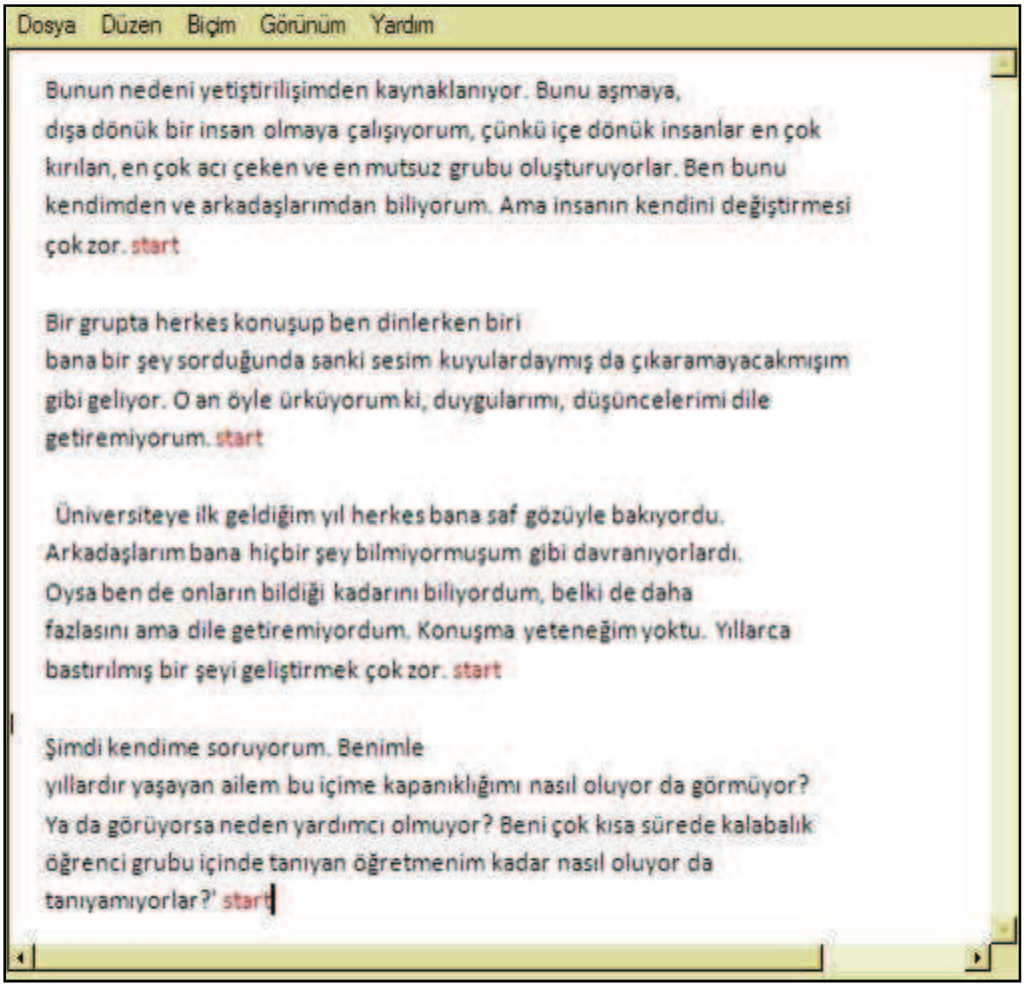


Şekil.17.Kök Bulma İşlemi

Uygulamanın ikinci modülündeki programa uygun verileri hazırlaması açısından zemberek yazılımına bazı eklentiler yapılmıştır. Kelime köklerinin sayfa sayfa bulunabilmesi ve bu bağlamda ikinci uygulamada sorgulanacak olan kelimenin hangi sayfalarda hangi kelimelerle tekrarlandığını gösterebilmek açısından sayfa sonlarına durma noktası eklenmiştir. Durma noktası olarak yabancı kelime olan ‘start’ kelimesi, denemeler sonucu bulunmuş ve kullanılmak üzere seçilmiştir.

Dokümanlar içerisinde önceden ‘start’ kelimesinin olabileceği akla gelebilir. Tüm dokümanlarda ‘start’ kelimeleri, durma noktaları eklenmeden önce araştırılmıştır, varsa ‘start’ kelimesi Türkçe karşılığı olan ‘başla’, ‘başlama’ gibi yerine göre türetilerek düzeltmeler yapılmıştır ya da metin içerisinden çıkarılmıştır.

Bu düzeltmeler yapıldıktan sonra ‘start’ kelimesi tüm sayfa sonlarına manuel olarak eksiksiz bir şekilde yerleştirilmiştir.



Şekil.18. Durma Noktası Eklenmiş Metin Örneği

Toplanılan metin parçalarının tamamının sayfaya yerleştirilmesinin imkânsız olmasından dolayı, şekil 18’de görüldüğü üzere, her paragraf bir sayfa olarak varsayılacaktır. Bu durumda yukarıdaki şekilde 4 adet ‘start’ bulunduğundan ötürü 4 adet sözde sayfa olduğu anlaşılacaktır. Görüldüğü gibi sözde sayfa sonlarında durma noktaları yerleştirilmiştir. Zemberek açık kaynak kodlu yazılımına sayfa sonlarını belirtmek ve kelime köklerinin yanına sayfa numaralarını yazdırmak için şekil 19’daki gibi bir kod eklentisi yapılmıştır.

```
if(x!=0){
out.println(cozumleri[x-1].toString()+ " "+ xx );
wline.write(cozumleri[x-1].toString());
if (cozumleri[x-1].toString().equals("start"))
{
aa+=1;}
if (cozumleri[x-1].toString().equals("start")){
xx+=1;}
wline.newLine();}
sayac++;
```

Şekil.19. Durma Noktası Kod Eklentisi

Tablo 2'deki kod bloğu sayesinde yazılım sayfa sonlarındaki "start" kelimelerine göre köklerin yanına Tablo 3 'teki gibi sayfa numarası ekleyecek ve her "start" görüşünde bir sayfa numarası artırarak son "start" kelimesine kadar numaralandırmaya devam edecektir.

Kök bulma işlemi gerçekleşmesi durumunda yazılım "start" kelimelerini de sıralanmış kökler arasında gösterecektir. "Start" kelimesi kök bulma işlemleri gerçekleştikten sonra program "Microsoft Office Acces 2007" programı yardımıyla otomatik olarak veritabanından silinmiştir.

Şekil 18 ile verilen metnin, kelime köklerine ayrılmış görüntüsü aşağıda bulunan Tablo 1 üzerinde gösterildiği gibidir. Burada görüldüğü üzere kelimeler çok fazla hata payı olmadan köklerine ayrılmıştır. Ancak, tek başına anlamı olmayan kelimeler ve bağdaştırma kelimeleri de sonuç olarak karşımıza gelmektedir.

Tablo.1. Metin Kökleri

bu 1	ama 1	start 2	yılla 3
neden 1	insan 1	üniversite 3	bastır 3
yet 1	kendi 1	ilk 3	bir 3
kaynaklan 1	değ 1	gel 3	şey 3
bu 1	çok 1	yıl 3	geliştir 3
aş 1	zor 1	herkes 3	çok 3
dış 1	start 1	ban 3	zor 3
dönük 1	bir 2	saf 3	start 3
bir 1	grup 2	göz 3	şimdi 4
insan 1	herkes 2	bak 3	kendi 4
ol 1	kon 2	arkadaş 3	sor 4
çal 1	ben 2	ban 3	ben 4
çünkü 1	din 2	hiçbir 3	yılla 4
iç 1	biri 2	şey 3	yaşa 4
dönük 1	ban 2	bil 3	aile 4
insan 1	bir 2	gibi 3	bu 4
en 1	şey 2	davran 3	içim 4
çok 1	sor 2	oy 3	kapanık 4
kır 1	sanki 2	ben 3	nasıl 4
en 1	ses 2	de 3	o 4
çok 1	kuyu 2	onlar 3	da 4
acı 1	da 2	bildik 3	gör 4
çek 1	çıkar 2	kadar 3	ya 4
ve 1	gibi 2	bile 3	da 4
en 1	gel 2	belki 3	gör 4
mutsuz 1	o 2	de 3	neden 4
grup 1	an 2	daha 3	yardım 4
oluş 1	öyle 2	fazla 3	ol 4
ben 1	ürk 2	ama 3	ben 4
bu 1	ki 2	dile 3	çok 4
kendi 1	duygu 2	getir 3	kıs 4
ve 1	düşünce 2	kon 3	süre 4
arkadaş 1	dile 2	yetenek 3	kalaba 4
bile 1	getir 2	yok 3	start 4

4.5.2. Kelimeler Arasındaki İlişkiler ve Anlam Birliktelikleri

Kelimelerin gizli ilişkilerinin açığa çıkarılması işlemi bu modül içerisinde gerçekleştirilecektir. Sayfa içerisinde yer alan kelimelerin anlamsal ilişkilerinin bulunması aşamasında, sorgusu gerçekleştirilecek olan kelimenin diğer kelimelerle birlikte geçme sıklıklarına bakılacaktır. Bunun sonucunda kelimeler arasındaki anlamsal ilişkiler tespit edilmiş olacak ve sorgusu gerçekleştirilen kelimenin anlam tahmini yapılabilecektir.

Örneğin, koyun kelimesinin iki tane kökü vardır. Birincisi "koy" kökü diğeri "koyun" köküdür. Burada "koyunculuk" kelimesinin köküne ayrılması durumu göz önünde bulundurulursa, kelime kökünün "koyun" olarak bulunabilmesi için sayfa içerisinde hangi kelimelerle sık geçtiğine bakılabilir. Koyunculuk kelimesinin besicilik, süt, kıvırcık, merinos, yün vb. kelimelerle metin içerisinde sık geçtiği bilinirse, bu kelimenin "koyun" kökünde kalmasına dikkat edilebilir ve bu kelimenin anlamı "küçükbaş hayvan" olarak tahmin edilebilir.

Koyun kelimesi birde "koy" olarak; "göl, deniz veya okyanusların karaların içine doğru yaptığı görece sığ girinti" anlamında çözümlensin. Turistik tanıtım yapan makale içerisinden aşağıdaki gibi bir metin parçası ele alınacak olursa;

"Körfezin manzarası, yeşil ve mavinin eşine az rastlanır güzellikteki dostluğunu gözler önüne seriyor. Körfez, denizi ve kumu; çevresindeki çam ağaçları ile eşsiz güzelliğe sahip sayısız koydan oluşuyor. Göl ve denizlerin içeri doğru yaptığı girintiler sonucu oluşan koylara turizmin yeni gözdesi Bedirhan köyünde daha çok rastlanmaktadır. Bu köyde oturanlar 'koyun' etrafına yerleşmişlerdir. Buradaki halk deniz turizmi, balıkçılık ve küçükbaş hayvancılık yaparak geçimlerini sağlarlar. Küçükbaş hayvanı olan köylü sayısı bir hayli fazladır. Buna rağmen balıkçılık daha fazla gelişmiştir. Gökova körfezinin güney kıyısında bulunan 'koylar' o bölgede oturanların denizle olan yakınlıklarını artırmıştır. Körfezde onlarca ufak 'koyun' bulunması deniz turizmini de geliştirmiştir. Gökova körfezi 'koylarıyla' meşhur bir körfezimizdir…"

Yukarıdaki metin içerisinde “koyun”, “koylar” gibi “koy” kökünden türemiş kelimeler bulunmaktadır. Öncelikle metin parçası kök bulma algoritması içerisine yerleştirilir. Tablo 2’ de gösterildiği gibi metin içerisinde yer alan kelimelerin kökleri, yanlarında bulunan sayfa numaralarıyla birlikte zemberek yazılımı yardımıyla çıkartılır.

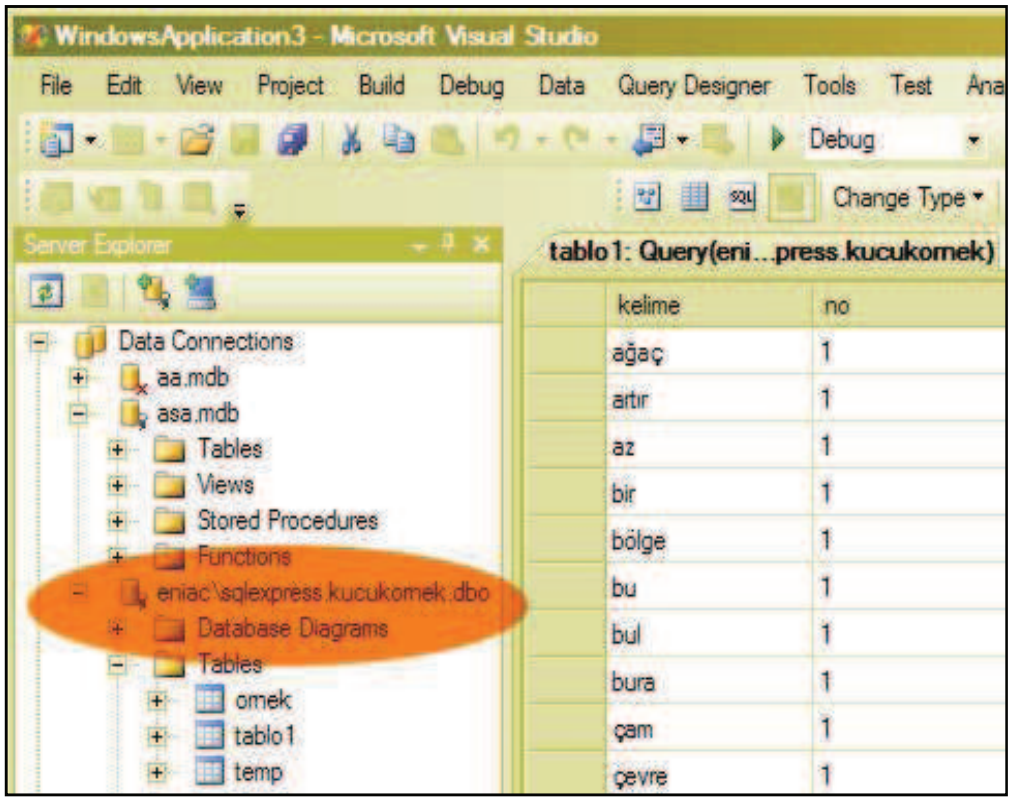
Tablo.2. Tek Sayfa İçerisindeki Metin Kökleri

körfez 1	koy 1	bura 1	körfez 1
manzara 1	oluş 1	halk 1	güney 1
yeşil 1	göl 1	deniz 1	körfez 1
ve 1	ve 1	turizm 1	onlar 1
mavi 1	deniz 1	balık 1	ufak 1
eşin 1	içeri 1	ve 1	koyun 1
az 1	doğru 1	küçükbaş 1	bul 1
rast 1	yap 1	hayvan 1	deniz 1
güzel 1	girinti 1	yap 1	turizm 1
dost 1	sonuç 1	geçim 1	de 1
gözle 1	oluş 1	sağ 1	geliştir 1
ön 1	koy 1	küçükbaş 1	körfez 1
ser 1	turizm 1	hayvan 1	koy 1
körfez 1	yen 1	o 1	meşhur 1
deniz 1	gözde 1	köy 1	bir 1
ve 1	köy 1	sayı 1	körfez 1
kum 1	daha 1	bir 1	kıyı 1
çevre 1	çok 1	hayli 1	bul 1
çam 1	rast 1	faz 1	koy 1
ağaç 1	bu 1	buna 1	o 1
il 1	köy 1	rağmen 1	bölge 1
eş 1	otur 1	balık 1	otur 1
güzel 1	koyun 1	daha 1	deniz 1
sahip 1	etraf 1	faz 1	o 1
sayısız 1	yerleş 1	geliş 1	yakın 1
			artır 1

Kelime köklerini çıkaran yazılım, kök kelimeleri c sürücüsü içerisindeki “sonuc.txt” dosyasına kayıt etmektedir. Kelimeler köklerine ayrılırken, metin içerisindeki kelime sırasına göre kök bulma işlemini gerçekleştirmektedir. Kayıtlı dosya içerisindeki kelimeler analiz işlemi için veritabanına aktarılır.

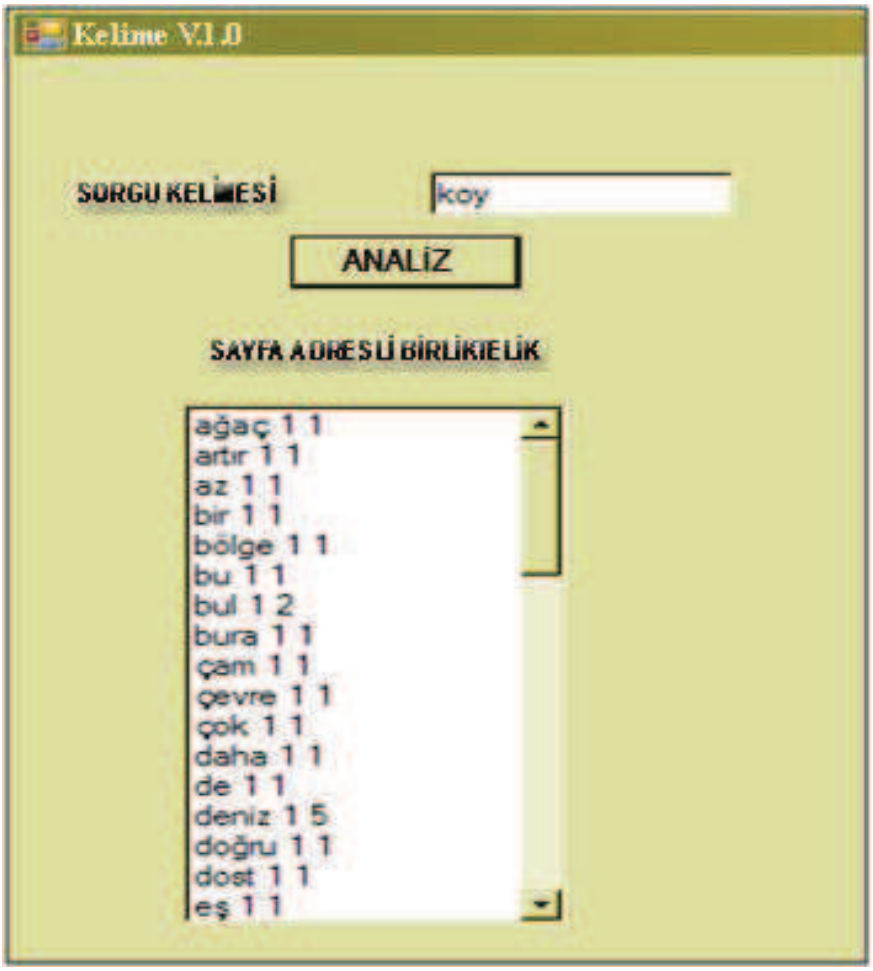
Kelimeler, Şekil 20 üzerinde gösterildiği gibi, “Windows Application” paneli üzerinde yapılan çalışmada, “eniac\sqlexpress.kucukornek.dbo” adlı veritabanında tutulur.

Şekil 20’deki veritabanı içerisinde bulunan “tablo1” adlı tablonun ilk sütunu içerisindeki veriler kelimeleri temsil etmektedir. İkinci sütun içerisindeki veriler ise kelimelerin hangi sayfada olduğunu gösteren numara kayıtlarını göstermektedir.



Şekil.20. Veritabanı Bilgileri

Veritabanına entegre edilen veriler ilk olarak sıralı olarak sorgulanır. Kelime analizi programı çalıştırıldığında şekil 21'de gösterildiği gibi karşımıza çıkan arayüzde sorgulanacak olan "koy" kelimesi programın ilgili kısmına girilir.



Şekil.21."Koy" Kelimesi Sonuçları

"koy" kelimesi sorgulandıktan sonra sonuç;

"Kelime + birlikte geçtiği sayfa numarası + sorgulanan kelimeyle birlikte kaç defa geçtiği" şeklinde bir satır sıralamasıyla karşımıza çıkacaktır. "Koy" kelimesinin tüm kelimelerle birliktelikleri aşağıda bulunan Tablo 3 üzerinde gösterilmektedir.

Tablo.3. Sorgulanan Kelimenin Birliktelik Sonuçları

ağaç 1 - 1	geçim 1 - 1	o 1 - 1
artır 1 - 1	geliş 1 - 1	oluş 1 - 2
az 1 - 1	girinti 1 - 1	onlar 1 - 1
balık 1 - 2	göl 1 - 1	otur 1 - 2
bir 1 - 2	gözde 1 - 1	ön 1 - 1
bölge 1 - 1	gözle 1 - 1	rağmen 1 - 1
bu 1 - 1	güney 1 - 1	rast 1 - 1
bul 1 - 2	güzel 1 - 2	sağ 1 - 1
buna 1 - 1	halk 1 - 1	sahip 1 - 1
bura 1 - 1	hayli 1 - 1	sayı 1 - 1
çam 1 - 1	hayvan 1 - 2	sayısız 1 - 1
çevre 1 - 1	içeri 1 - 1	ser 1 - 1
çok 1 - 1	il 1 - 1	sonuç 1 - 1
daha 1 - 2	kıyı 1 - 1	turizm 1 - 3
de 1 - 1	körfez 1 - 6	ufak 1 - 1
deniz 1 - 5	köy 1 - 3	ve 1 - 4
doğru 1 - 1	kum 1 - 1	yakın 1 - 1
dost 1 - 1	küçükbaş 1 - 2	yap 1 - 3
eş 1 - 1	manzara 1 - 1	yeleş 1 - 1
eşin 1 - 1	mavi 1 - 1	yen 1 - 1
etraf 1 - 1	meşhur 1 - 1	yeşil 1 - 1
faz 1 - 2		

Normalde metin sayfası içerisinde “koy” kelimesinden türeyen “koyun” kelimesi de yer almaktadır. “Koyun” kelimesinin anlamsal ilişkilerini görebilmek için “koy” kökünün birlikteliklerine bakılır. Tablo 3’te görüldüğü gibi sorgulanan “koy” kök kelimesinin “deniz” kelimesiyle 5 defa, “körfez” kelimesiyle 6 defa geçtiği sonucu çıkmıştır. Metin sayfaları içerisinde “koy” kelimesi “hayvan” ve “küçükbaş” kelimeleriyle de 2 defa geçtiği görülmektedir. Diğer kelimelerle de sıklıkla 1 veya 2 defa geçmiştir. Buradan çıkarılan sonuca göre sorgulanan kelime ile yakınlığı bulunan kelimeler arasında anlamsal bir ilişki olduğu tespit edilmiştir.

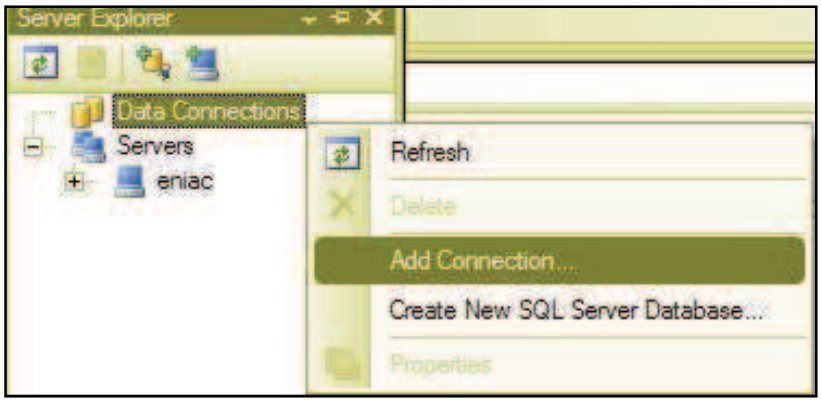
Bunun yanında “koyun” kelimesinin metin içerisinde gerçekte ilk anlamının “deniz kenarındaki sığ girinti” anlamında olduğu tahmin edilmiştir.

Birkaç ufak örnek verildikten sonra uygulamanın yüksek kapasiteli veritabanıyla gerçekleştirilen işlemlerine geçelim.

Türkçe kelimeler anlam bakımından çok zengin bir yapıya sahiptir. Kullandığımız kelimelerin çoğunun birbirleriyle gizli veya açık ilişkileri bulunmaktadır. Kelimeler birbirlerinden beslenerek anlamlı örüntüler oluşturabilirler. Kelime birlikteliklerine göre, kelimeler arasında olan, tahmin edilemeyecek uzak veya yakın ilişkiler bulunabilmektedir.

Kelimelerin farklı anlamlarının tespitinin güvenirliliğini yüksek seviyede sonuçlandırabilmek için 13.926 sayfalık metin sayfasından 4.156.123 kelimelik bir veritabanı oluşturulmuş fakat elimizdeki donanımsal imkânların kurulan algoritma yapısına tam uygun olmaması nedeniyle bu veritabanı içerisinden rastgele 78.112 kelimelik bir örneklem alınmış ve uygulamaya koyulmuştur. İşleme koyulan bu veritabanı ile kelime ilişkilerinin tespitinde verimli sonuçlar alınabilmiştir.

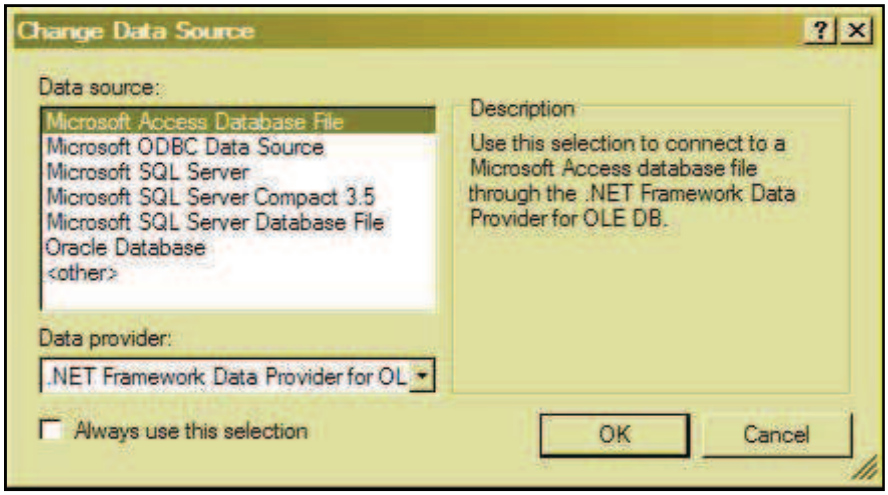
Yüksek kapasiteli veriler, daha önceki örneklerde anlatıldığı şekilde bir ön işlem sürecinden geçirilmiş ve “temel.mdb” olarak “C:\Documents and Settings\eniacx\Desktop” adres yoluna kayıt edilmiştir.



Şekil.22. Yeni Veritabanı Bağlantısı

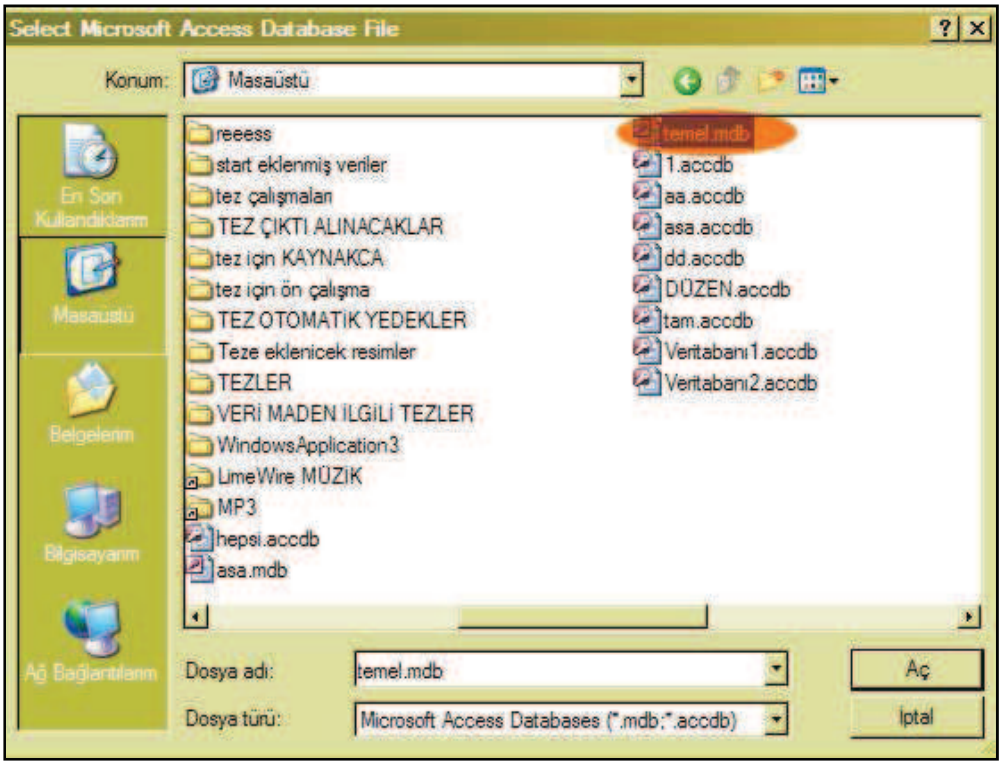
Oluşturulan veritabanına ulaşabilmek için yeni bir veritabanı bağlantısı kurmamız gerekmektedir. Program üzerinden yeni bir veritabanına bağlanabilmek için şekil 22’deki gibi Server Explorer penceresinde bulunan “Data Connections”

sekmesi üzerinde farenin sağ tuşu tıklanır ve açılan menüden "Add Connection" sekmesi tıklanır. Açılan pencerede bağlanılacak veritabanın dosya türü seçilir. Şekil 23 üzerinde de görüldüğü gibi "Microsoft Access Database File" seçilmiştir. Seçim onaylandıktan sonra veritabanı seçimi yapılması istenir.



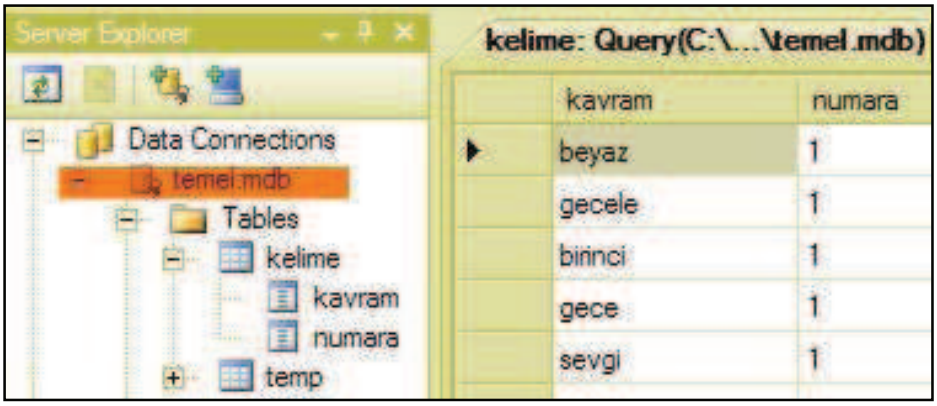
Şekil.23. Veritabanı Dosya Türü

Dosya türü seçildikten sonra şekil 24 ile gösterildiği gibi "temel.mdb" olarak kayıt ettiğimiz veritabanı seçilir.



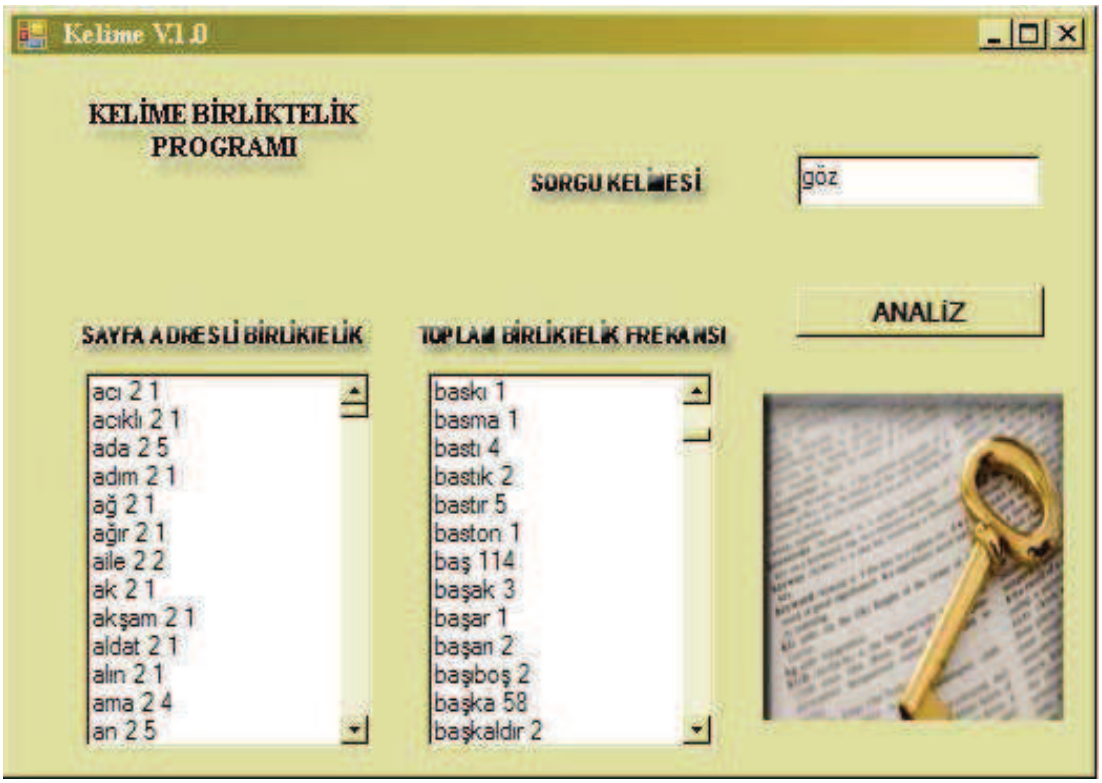
Şekil.24. Veritabanı Seçimi

Seçme işlemi onaylandıktan sonra "temel" adlı veritabanı şekil 25 ile gösterildiği gibi program içerisine entegre edilmiş olur.



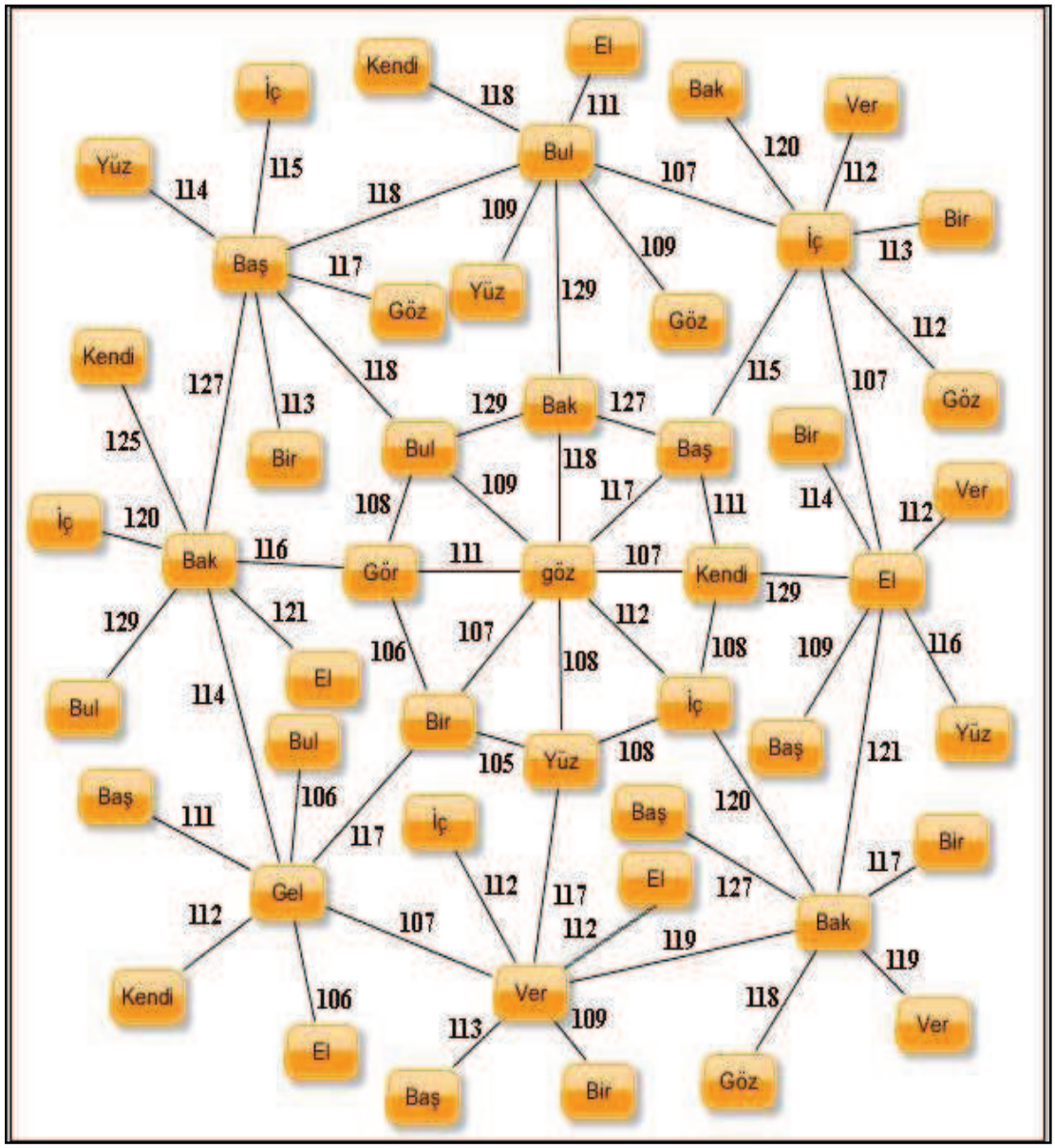
Şekil.25."Temel" Veritabanı

Veritabanı içerisinde yer alan bir kelimenin, şekil 26'da gösterildiği gibi sorgu kelimesi olarak ilgili alana girilmesi gerekmektedir. Sorgu kelimesi alanına "göz" kelimesi girilmiştir. "Göz" kelimesinin anlamsal olarak kompleks yapıda ve anlamsal ilişkilerinin üst seviyede olması, bu kelimenin örnek sorgu kelimesi olarak seçilmesinde etkili olmuştur.



Şekil.26. Kelime Birliktelik Programı Arayüzü

Kelimeler arası bağlantıların karmaşıklığına bakılmaksızın, geliştirilen program sonuçlarına göre kelimelerin birlikte geçme frekansları sonuçlandırılmış, bu çıkan değerlere göre de ilişkisel kelime haritası, düğümler ve düğümler arası bağlantılardan oluşturulmuştur.

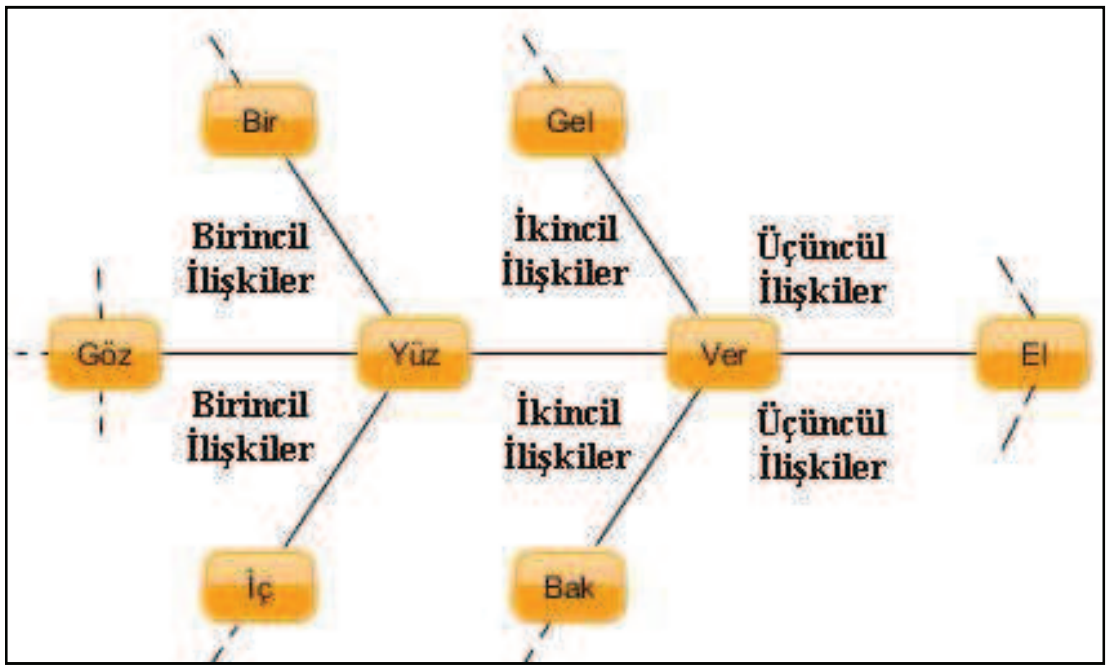


Şekil.27. Kelime Anlam Haritası

"Göz" kelimesinin diğer kelimelerle bağlantısı incelendiğinde karşımıza Şekil 27'deki gibi ilişkisel bir kelime haritası çıkacaktır. Burada kelimelerin birbirleriyle olan gizli veya açık ilişkileri detaylı olarak görülmektedir.

Kelime anlam haritası oluşturulurken kelime anlamlarını değiştirmeyecek "ve", "veya", "ama" gibi kelimeler sık geçmesine rağmen harita üzerinde gösterilmemiştir ve işlem dışı bırakılmıştır.

Kelime anlam haritasının merkez düğümü içerisinde sorgu kelimesi olan "göz" yer almaktadır. Merkezdeki kelime ile üst düzey yakınlığı bulunan 8 kelime birincil ilişkiler içerisinde düğümler halinde gösterilmektedir. Birincil ilişkiler içerisinde yer alan kelimeler; "gör", "yüz", "bak", "bul", "baş", "kendi", "iç", "bir" olarak sonuçlandırılmıştır.



Şekil.28.İlişkisel Boyut

Merkez düğümdeki kelimeyle en sık geçen 8 kelimenin kendi aralarındaki anlamlı ilişkileri de bulunmuş ve harita üzerinde gösterilmiştir. Birincil ilişkiler içerisinde yer alan kelimelerin sayısının artırılması anlamsal yapıyı daha da güçlendirebilecektir fakat tez çalışmasına yansıtılması konusunda boyutsal açıdan çizge sorun çıkarmaktadır.

Şekil 28'de kelimelerin ilişkisel seviyelerinin gösterilmiştir. Bunun yanında her ilişki alanındaki kelimelerin birbirleriyle olan bağlantıları da haritaya yansıtılmıştır.

Birincil ilişkiler içerisinde yer alan düğümlerin her biriyle en sık geçen kelimelerden yeni birer düğüm oluşturulmuştur. Bu oluşturulan alandaki bağlantılar ise ikincil ilişkiler olarak değerlendirilmektedir. İkincil ilişkiler içerisindeki düğümlerin her biri arasında bağlantı kurulmuş ve birlikte geçme frekansları tespit edilmiştir.

İkincil ilişkiler içerisinde yer alan kelimeler ile en sık geçen kelimeler arasında da bir bağlantı kurulmuştur. Bu bağlantı alanı da üçüncül ilişkiler olarak değerlendirilmiştir.

"Göz" kelimesinin en çok geçtiği ilk 8 kelimenin birliktelik sonuçları aşağıda yer almaktadır:

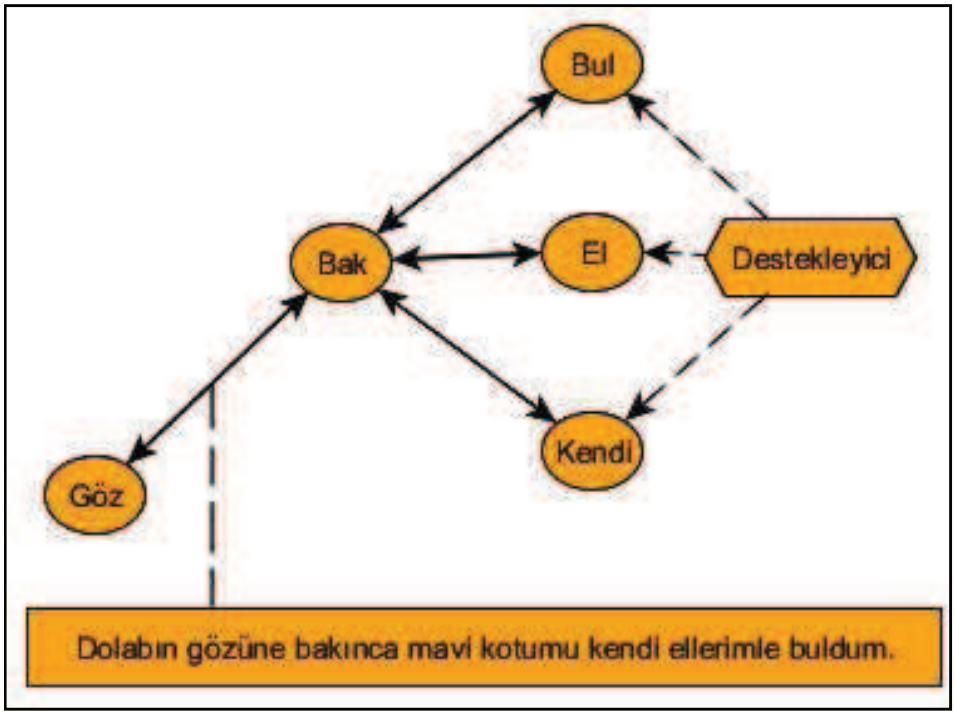
- "göz" kelimesi ile " bak" kelimesinin birlikteliği "118"
- "göz" kelimesi ile " baş" kelimesinin birlikteliği " 117"
- "göz" kelimesi ile "iç" kelimesinin birlikteliği "112"
- "göz" kelimesi ile "gör" kelimesinin birlikteliği "111"
- "göz" kelimesi ile "bul" kelimesinin birlikteliği "109"
- "göz" kelimesi ile " yüz" kelimesinin birlikteliği "108"
- "göz" kelimesi ile "kendi" kelimesinin birlikteliği "107"
- "göz" kelimesi ile "bir" kelimesinin birlikteliği "107"

Bu sonuçlara göre sorgulanan "göz" kelimesinin en sık geçtiği kelime birincil ilişkiler içerisindeki "bak" kelimesidir. Bu durumda "göz" kelimesi ile "bak" kelimesi arasında en üst seviyede bir ilişki olduğu söylenebilir. Bu kelimelerin metin sayfaları içerisinde birlikte kullanılma oranlarının yüksek olması, bu kelimeler arasında anlamsal bir bağ ilişkisi olduğunun güçlü bir kanıtıdır.

Sorgulanan kelimenin birincil ilişkiler içerisindeki diğer kelimelerle olan ilişkilerine bakılırsa; bu kelimelerin de sorgulanan "göz" kelimesi ile ilişkilerinin güçlü olduğu söylenebilmektedir.

Sorgulama sonucunda açığa çıkan bu bilgiler ışığında, sorgulanan "göz" kelimesinin ilk anlamının tahmini yapılabilmektedir.

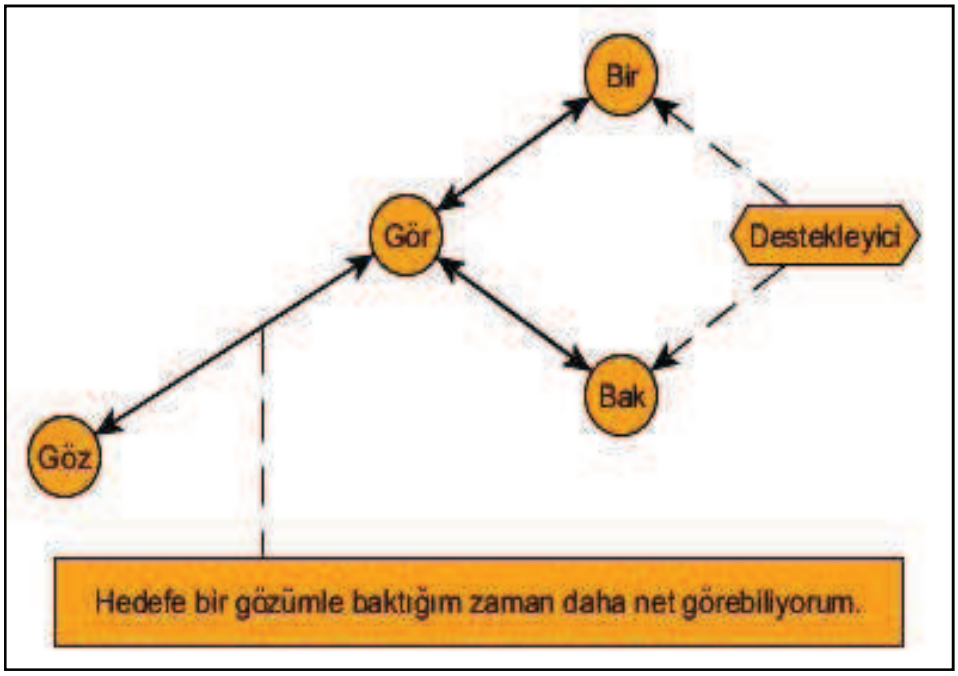
İlk anlam tespitinde üç destekleyici kelime görev almıştır. Bu kelimeler "bul", "el" ve "kendi" olmak üzere şekil 29'da gösterilmektedir. Destekleyici kelimeler, kelimenin ilk anlamının tahmin edilmesinde yardımcı olmuştur.



Şekil.29."Göz" Kelimesi İlk Anlam Tahmini

"Göz" kelimesi ile ilişkisi en üst seviyede olan ve "bak" kelimesi ve destekleyici kelimeler, cümle içerisinde birlikte kullanıldığı takdirde, "göz" kelimesinin ilk anlamının "masanın gözü, çekmece" anlamında olabileceği sonucuna varılabilmiştir.

Destekleyici kelimelerin, metin sayfaları içerisindeki kendi aralarındaki birliktelikleri sorgulandığı takdirde, bu kelimelerin de merkezdeki "göz" kelimesi ile ilişkilerinin kuvvetli olduğu görülmektedir.



Şekil.30."Göz" Kelimesi İkinci Anlam Tahmini

"Göz" kelimesi ile ilişkisi "bak" kelimesi kadar yüksek olmayan "gör" kelimesi, ikinci anlam tahmininde anahtar kelime olacaktır.

İkinci anlam tahmininde iki adet destekleyici kelime kullanılmıştır. Şekil 30'da gösterildiği gibi "Bak" ve "Bir" kelimeleri anlam tahmininde destekleyici olarak görev yapmaktadır.

"Göz" kelimesi ile üst seviyede ilişkisi bulunan "gör" kelimesi ve "bir" ile "bak" destekleyici kelimeleri, cümle içerisinde birlikte kullanıldığı takdirde, "göz" kelimesinin ikinci anlamının "görme fonksiyonunu gerçekleştiren beş duyu organından biri" olabileceği tahmin edilebilmektedir.

Kelime anlam tahminlerinin artması için kullanılan metin sayısının artırılması gerekmektedir. Veritabanı ne kadar çok büyütülürse o kadar fazla kelime anlamı tahmini yapılabilir. Yapılacak elektronik bir dil sözlüğünde, olabilecek kelime anlamlarının tespitinde, yüksek kapasiteli veritabanı içerisinde kelimelerin birlikteliklerine bakılarak bu yöntem ile birden fazla anlam tahmin edilebilir.

Bu çalışmanın uygulaması, elektronik bir sözlük içerisinde kullanılabilir. Sistem, sözlük altyapısına entegre edildiği takdirde yardımcı algoritmalarla birlikte çalışarak kullanışlı bir sözlük oluşturulabilir. Önerilen sözlük arayüzü aşağıdaki gibi oluşturulabilir.

Göz:

1) Masanın gözü, çantanın gözü, çekmece.
   Dolabın gözünde dünden kalmış bir poğaça vardı.
2) Görme fonksiyonunu gerçekleştiren beş duyu organından biri.
   Bilgisayar başında çok vakit geçirdiğim için gözlerim ağrıyor.

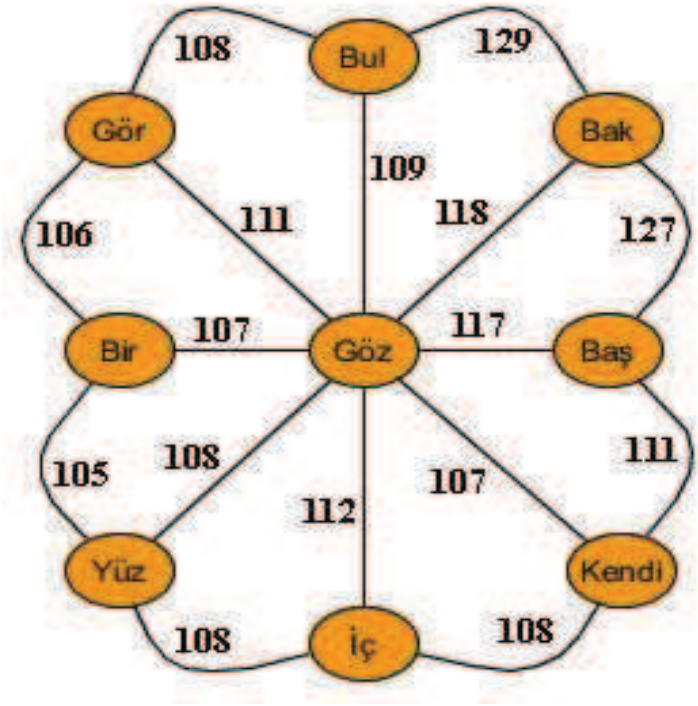
Yüz :

1) Doksan dokuzdan sonra gelen sayının adı.
   Hikmet Bey'in kurum ve edası, her zamankinden belki yüz kat üstündü.
2) Başta, alın, göz, burun, ağız, yanak ve çenenin bulunduğu ön bölüm, surat, çehre.
   Kırışıklıklar arasında kaybolmuş bir yüzüm var.
3) Yorgana ve yastığa geçirilen kılıf.
   Yorganın beyaz renkli yüzü kirden kararmıştı.

Koy:

1) Denizin, gölün küçük girintiler biçiminde karaya doğru sokulduğu bölümü.
   Sandalını Kaşık Adası'nın bir küçük koyuna çekti.
2) Koyun : Küçükbaş hayvan.
   Koyunlar kuzulamaya başladı.
3) Bir şeyi bir yere bırakmak.
   Öteki elini doktorun omzuna koydu.

Şekil.31.Örnek Sözlük Arayüzü



Şekil.32. Birincil İlişkiler

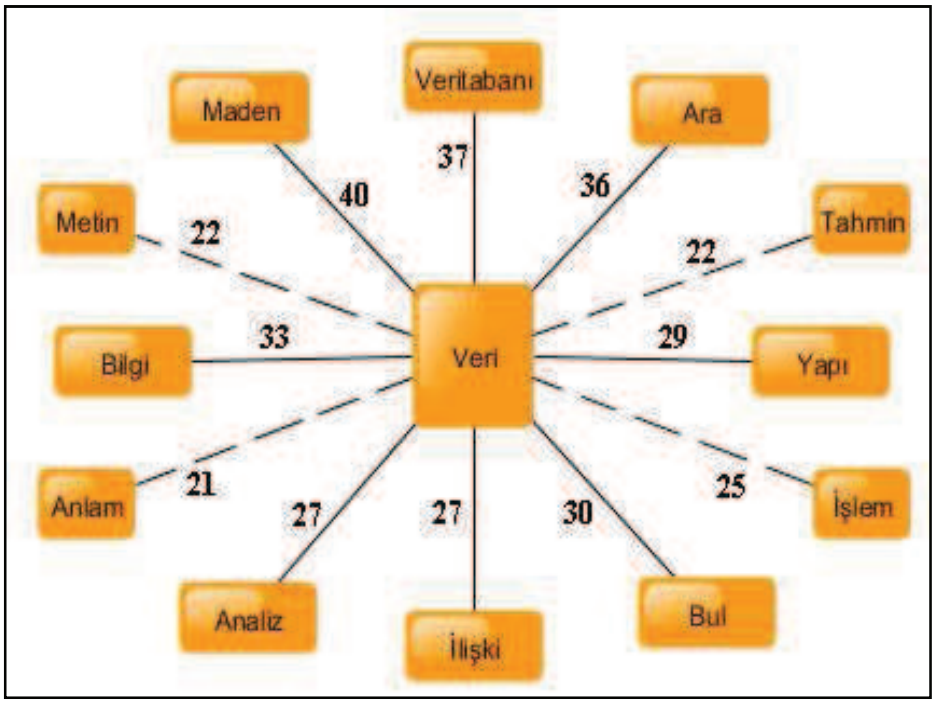
Anlam haritasından çıkartabileceğimiz bir diğer husus da sorgulanan bir kelime ile sık geçen kelimelerin ayrıca kendi aralarındaki birlikteliklerdir. "Bak" kelimesinin, "göz" kelimesi ile "118" defa geçtiği bilinmektedir. Bunun yanı sıra "bul" kelimesinin de "109" defa "göz" kelimesi ile birlikte geçtiği bilinmektedir. Aynı şekilde birincil ilişkiler içerisindeki diğer kelimelerin de sorgu kelimesi ile ilişkileri şekil 31'de görüldüğü üzere güçlüdür.

Sorguladığımız kelime ile birlikte geçen "bak" ve "bul" kelimesinin metin sayfaları içerisinde birlikte geçme durumunu incelediğimizde; bu kelimelerin birlikte geçme frekanslarının "129" olduğu tespit edilmiştir. Bu durumda, "göz" ve "bak" kelimesinin beraber geçtiği yerlerde "bul" kelimesinin de sıklıkla geçebileceği söylenebilmektedir. Birincil ilişkiler içerisindeki diğer kelimelerin de aralarındaki birlikteliklere göre aynı şekilde yorum yapılabilir.

Hazırlanılan programı, belirli bir konusu olan bu tez metninin literatür bölümünde işleme alalım.

Tez çalışmasının literatür bölümündeki metinsel veriler, yukarıdaki örneklerde anlatıldığı gibi ön işlem süreçlerinden geçirilir, veritabanına dönüştürülür ve daha sonra zemberek kök bulma algoritması yardımıyla köklerine ayrılır.

Köklerine ayrılan kelime veritabanı, kelime birliktelik programına entegre edilir. Program çalıştırıldıktan sonra, bu tez çalışmasının yapı taşı olan "veri" kelimesi sorgu olarak girildiği takdirde karşımıza şekil 32'deki gibi ilişkisel bir kelime haritası çıkacaktır.



Şekil.33."Veri" Kelimesi Anlamsal İlişkileri

Şekil 32'deki bilgiler ışığında, sorgulanılan "veri" kelimesi ile üst düzeyde ilişkisi olan 8 kelime kesiksiz ok işaretleriyle gösterilmiştir. "Veri" kelimesi ile ilişkileri bir alt kademede olan diğer 4 kelimenin ilişkileri de kesikli olarak gösterilmiştir. "Veri" kelimesi ile "maden", "veritabanı", "bilgi", "analiz", "ilişki", "ara", "bul", "yapı" kelimelerinin birlikte geçme frekanslarına göre ilişkileri kuvvetlidir diyebiliriz. Bunun yanında "metin", "anlam", "tahmin", "işlem" kelimeleri de "veri" kelimesi ile iyi bir ilişki göstermiştir.

Çıkan bu sonuçlar ışığında, "veri" kelimesinin anlamının "bir araştırmanın, bir tartışmanın, bir muhakemenin temeli olan ana öğe, bilgi " olabileceği tahmin

edilebilir. “Veri” ve “bilgi” kelimelerinin birlikte geçme frekanslarına bakıldığı takdirde de aralarında kuvvetli bir anlamsal bağ olduğu görülebilmektedir.

Çıkan bu sonuçların yanı sıra bir başka önemli sonuç bulunmuştur. Bu tez çalışmasının anahtar kelimeleri; Veri madenciliği, Metin madenciliği, Veritabanı, Kelime ilişkileri, Metin analizi olarak “veri” kelimesi araştırması yapılmadan önce belirlenmiştir.

“Veri” kelimesinin anlamsal haritasına bakıldığı takdirde, “veri” kelimesi ile aralarında anlamsal bağ bulunan kelimelerin, bu tez çalışmasının anahtar kelimelerine benzediği sonucu çıkartılmıştır. Bu durumda kullanılan program, tez çalışmasının anahtar kelimelerini büyük ölçüde tahmin edebilmiştir.

4.6. Elde Edilen Sonuçların Değerlendirilmesi

Metin madenciliği modellemelerinden birliktelik analizi kullanılarak kelimelerin birbirleriyle sık geçme durumlarının araştırması yapılmış, kelimeler arsında olan ilişkiler tespit edilmiştir. 78.112 kelimelik veritabanı üzerinde sorgulanan 30 adet kelimenin ilişki durumlarının analizi sonucunda; geliştirilen sistemin, kelimelerin anlamlarını Tablo 4’ de gösterildiği gibi % 86’lık bir oranla tahmin etme başarısı gösterebilmiştir.

Tablo.4. Sistem Verimliliği

Sistem Duyarlılık Analizi			
		Pozitif Değer	Negatif Değer
Toplam veri	: 78.112		
		86%	14%
Toplam sorgu kelimesi	: 30		

Çalışma içerisinde gösterilen “göz” kelimesinin metin sayfaları içerisindeki diğer kelimeler ile aralarındaki ilişkiler ortaya çıkartılmıştır. Metin sayfaları içerisinde “göz” kelimesinin en sık geçtiği kelimeler arasında anlamsal bir bağ olduğu keşfedilmiştir. Kelimeler arasındaki bu ilişkilere göre destekleyici kelimeler

yardımıyla “göz” kelimesinin birinci ve ikinci anlamları tahmin edilmiş ve sözlükteki gerçek anlamları bulunabilmiştir.

Bu tez çalışmasının literatür bölümündeki metin sayfaları üzerinde “veri” kelimesi ile yapılan araştırma sonuçlarına göre;

“Veri” kelimesinin literatür sayfaları içerisinde en sık geçtiği kelimeler tespit edilmiştir. “Veri” kelimesi ile literatürde en sık geçen kelimeler arasında anlamsal bir bağ olduğu tespit edilmiştir. Kelimeler arasındaki birlikteliklere göre “veri” kelimesinin gerçek anlam tahmini yapılmıştır. Sorgulanan kelimenin sözlükteki gerçek anlamıyla tutarlı olduğu görülmüştür. Bunun yanında “veri” kelimesi ile en sık geçen kelimelerin, bu tezin anahtar kelimelerini oluşturabildiği ve bu sayede anahtar kelimelerinin tahmin edilmesinde de büyük ölçüde doğru sonuçlar verebileceği kanısına varılmıştır.

Anlam haritası üzerinde bulunan düğümler içerisindeki bazı kelimelerin birlikte geçme oranları yüksek olduğu halde aralarında bir anlam ilişkisi kurulamamaktadır. Bunun nedeni kelimelerin belirli bir standarda sahip olmamasından kaynaklanmaktadır. Ayrıca bu durum, yazarların bazı kelimeleri sıklıkla kullanmasından kaynaklanabilir.

Veritabanlarında bulunan anlamsız kelimelerin bir program yardımıyla otomatik olarak çıkarılması sonucunda anlamsız kelimelerin işleme alınmaması sağlanabilir. Bu duruma alternatif çözüm önerisi olarak kullanılan dil kütüphanesinin yalnızca anlamlı kelimeleri barındırması ile de ulaşılabilir. Anlamsız kelimelerin sonradan manuel olarak işlem dışında bırakılması ayrı bir zaman alacaktır.

Kullanılan kelime sayısının artırılması ve daha karmaşık ilişkilerin harita üzerinde gösterilebilmesidir. Bu sayede kelimeler arası ilişkiler daha çok boyutlandırılabilecek ve ilişkilerin uzaklık ve yakınlıkları daha keskin çizgilerle ayrılabilecektir.

5. SONUÇLAR ve ÖNERİLER

Günümüz dünyasında bilgisayar kullanımının artması, internet kullanımının yaygınlaşması sonucunda yaptığımız hemen her işlem manyetik ortamlarda kayıt altına alınmaktadır. Teknolojinin gelişmesiyle birlikte çok sayıda karmaşık veri, manyetik depolarda saklanmakta ve bu verilerin kapasiteleri hızlıca büyümektedir.

Hızla büyüyen veritabanları içerisinden ihtiyaç duyulan bilgilerin çıkarılması bir hayli karmaşık ve maliyetli hale gelmiştir. Bu problemli durumların ve ihtiyaçların giderilmesi, anlamlı bilgilerin yüksek kapasiteli veriler arasından hızlı bir şekilde keşfi sürecinde veri madenciliği yöntem ve teknikleri ortaya çıkmıştır.

Veri madenciliği ile devasa boyuttaki verilerin analiz edilmesi sonucunda gizli ilişkiler ortaya çıkarılmış, istenilen anlamlı bilgiler keşfedilebilmiştir. Veri madenciliğinin hemen her alanda kullanımının artması ile veri madenciliği çalışmaları zenginleşmiş ve çeşitli alt dallara ayrılmıştır.

Veri madenciliğinin alt dalı olan metin madenciliği çalışmaları ile metinsel veriler içerisindeki potansiyel bilgiler keşfedilmektedir. Literatür taramalarında metin madenciliği çalışmalarının daha çok metin analizi ve metin sınıflandırmaları üzerinde durulduğu görülmüştür. Kelimeler, metinsel verilerin önemli bir parametresi olması sebebiyle metin analizi çalışmalarında büyük rol oynamaktadır.

Bu uygulamada, metin madenciliği modellemelerinden birliktelik kuralları modeli kullanılarak Türkçe kelimeler arasındaki gizli ilişkiler keşfedilip, analiz edilmeye çalışılmıştır.

Birliktelik kuralları yöntem ve teknikleriyle literatürde birçok çalışma yer almaktadır. Bu modellemenin kullanıldığı en yaygın çalışmalar; metin analizleri ve metinlerin kelimeler yardımıyla sınıflandırılması işlemleridir. Bu çalışmada diğer çalışmalardan farklı olarak bir kelimenin diğer kelimelerle birliktelikleri bulunmuş, sözlükteki olabilecek anlamları % 86'lık bir başarı oranıyla tahmin edilebilmiştir.

İnternet ortamından alınan çeşitli veriler metin madenciliği süreçlerinden geçirilmiş, tez kapsamı süresince hazırlanan "kelime birliktelik programı" yardımıyla kelimeler arasındaki gizli ilişkiler ortaya çıkarılabilmiştir. Program çıktılarına göre kelimeler arası ilişkileri gösteren bir anlam haritası tasarlanmıştır. Tasarlanan harita, üç katmanlı olarak analiz edilmiştir. Analiz sonuçlarına göre metin sayfaları içerisinde en sık geçen kelimelerin birlikteliklerinin kuvvetli yapıda olduğu ve bu sık geçen kelimeler arasında anlamsal bir bağ olduğu sonucuna varılmıştır. Bu bilgiler ışığında, kelimelerin anlam tahminlerinin, kelimeler arasında olan yakın veya uzak ilişkilere göre yapılabileceği söylenebilmektedir.

Ayrıca, belirli bir konuya ait olan bu tez çalışmasının literatür kısmının verileri veritabanına dönüştürülmüş ve oluşturulan bu veritabanı uygulamada kullanılan modüller aracılığıyla analiz edilmiştir. Analiz sonuçlarına göre metin sayfaları içerisinde sık geçen kelimeler arasında anlamsal bir bağ olduğu bilgisine bir kez daha ulaşılmıştır. Bu bilgilere göre kelimelerin gerçek anlamları tahmin edilebilmiştir. Bunun yanında belirli bir konuya ait olan metin sayfalarının anahtar kelimeleri, analiz sonucunda açığa çıkan kelime ilişkilerine göre tahmin edilmiş olup, bu tahminlerin yüksek oranda doğru olduğu sonucuna ulaşılmıştır. Anahtar kelimelerinin tahmin edilmesi ile de metin dokümanının anlatmak istediği kompleks düşünce ve fikirler çözümlenebilmektedir.

Yapılan bu tez çalışması birden fazla problem durumuna açıklık getirebilmiştir. Tamamlanmış olan bu çalışmanın, bilgisayarcılar ve Türk dili ve edebiyatı alanında çalışan araştırmacıların yapmış olduğu araştırmalara katkı sağlayabilecek durumda olması nedeniyle önemli bir kaynak olabileceği düşünülmektedir. Araştırmacılar, bu çalışma üzerinde farklı algoritma ve teknikler kullanarak yeni çalışmalar meydana getirebilir, bu çalışma amacından yola çıkarak yeni keşifler yapabilirler.

Bundan sonraki aşamada bu analiz sisteminden faydalanılarak dil sözlüğü programlarının anlamlı çeviriler yapabilmesi sağlanabilir. Dil sözlüğü programı içerisine bu tez çalışmasının sistemi entegre edildiği takdirde, veritabanından çekilecek bilgilere göre kelime tanımlamaları yapılabilir. Kelimelerin birlikte geçme

sıklıklarına göre sorgulanan kelimenin birden fazla anlamı var ise bu anlamlar tahmin edilebilir, çeviri sonucu olarak çıkarılabilir.

Sistemin kullanılabileceği başka bir alan ise; günümüzde internetin yaygınlığı ve insanların birçok veriyi arama motorları sayesinde aradığı bilinmektedir. Arama motorlarının alt yapısına eğer sistem organize bir şekilde entegre edilebilirse; arama motorlarının optimizasyonu sağlanabilir, aranılan bilgilere kısa sürede erişilebilinir. Örnek verilecek olursa; google arama motorunun yaptığı gibi arama çubuğuna aranılacak kelime girildiği takdirde, aranılan kelimenin yanında çıkan öneri panelinde yardımcı kelimeler belirir. Bu yardımcı kelimelerin belirlenmesinde bu tez çalışmasının uygulaması kullanılabilir. Bu sayede istenilen öz bilgiye erişim daha kısa sürede sağlanacaktır.

Yapılabilecek yeni bir uygulama önerisi ise; artık internet sitelerinin birçoğunda sıklıkla reklam alanlarının olduğunu görebilmekteyiz. İnternet sayfalarında verilen bu reklamların neye göre verildiği önemlidir. Çok kullanıcılı bir internet sitesi içerisinde spor, sağlık, magazin, politika, kültür gibi alanların olduğu kabul edilirse; bu alanlardaki reklam alanlarının her sayfasında, aynı satış ürünleriyle karşılaşılması satış verimini düşürebilir. Örneğin; Bir web sayfası içerisinde "arabalarla" ilgili bir makale yer almış olsun. Bu makale içerisinde en çok "araba" kelimesinin geçtiği varsayımına dayanarak, "araba" kelimesi ile en sık geçen kelimeler tespit edilirse, kelimeler arası ilişkilere göre web sayfası içerisinde reklam alanları oluşturulabilir. "Araba" kelimesinin web sayfası içerisinde sık geçmesine göre; "oto aksesuarları", "yedek parça ürünleri", "araba bakımları", "otomobil satışları" gibi alanlarda reklamlar, site içerisinde gerekli görülen yerlerde sunulabilir.

Getirilebilecek son bir öneri ise; bir yazara ait metin sayfaları içerisindeki anahtar kelimeler tespit edilerek, yazarın gerçekte ne demek istediği fikri, bilimsel olarak açığa çıkartılabilir. Metinsel veriler arasındaki ilişkilere bakılarak çıkarılan anahtar kelimeler sayesinde kompleks fikir ve düşünceler keşfedilebilir.

KAYNAKLAR

[1] Bilgi Türleri.(1986) *Büyük Larousse Sözlük Ve Ansiklopedisi*, (Cilt 23, s. 12164-12165). İstanbul, İnterpress Basın ve Yayıncılık.

[2] Alpaydın, E. (2000). "*Zeki Veri Madenciliği, Ham Veriden Altın Bilgiye Ulaşma Yöntemleri".* Bilişim 2000 Eğitim Semineri, Boğaziçi Üniversitesi Bilgisayar Mühendisliği Bölümü.

[3] Albayrak, A.S. , Yılmaz, Ş.K., 2009. *Veri Madenciliği, Karar Ağacı Algoritmaları ve İMKB Verileri Üzerine Bir Uygulama*. Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Dergisi, 14(1), 31-52, 2010. SDÜ Elektronik Dergi Sistemi.

[4] Silahtaroğlu, G., *Kavram ve Algoritmalarıyla Temel Veri Madenciliği*, Papatya Yayıncılık, İstanbul, 2008.

[5] Akpınar, H., 2000. *Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği.* İstanbul Üniversitesi İşletme Fakültesi Dergisi, 29(1), 1-22, 2010. ULAKBİM.

[6] Han, J. and Kamber M., 2006. *Data Mining Concepts and Techniques*, 2010. http://www.worldcat.org/title/data-mining-concepts-andtechniques/oclc/475299457/ viewport

[7] Sever, H. ve Oğuz, B., 2002. *Veri Tabanlarında Bilgi Keşfine Formel Bir Yaklaşım: Kısım 1- Eşleştirme Sorguları ve Algoritmalar*. Bilgi Dünyası, 3(2), s.173-204., 2010, ULAKBİM.

[8] İnceoğlu, M.M. ve Vahaplar A., 2001."*Veri Madenciliği ve Elektronik Ticaret".* VII. Türkiye' de İnternet Konferansı. Ege Üniversitesi, İzmir, Türkiye.

[9] Dolgun, Ö.M., Özdemir, T.G. ve Oğuz D., 2009. *Veri Madenciliği'nde Yapısal Olmayan Verinin Analizi: Metin ve Web Madenciliği.* İstatistikçiler Dergisi 2(2), s.48-58, 2010.

[10] Kaya, H. Ve Köymen, K., 2008. *Veri madenciliği Kavram ve Uygulama alanları.* Doğu Anadolu Bölgesi Araştırma ve Uygulama Merkezi Dergisi 6(2), s.159-164. , 2010.

[11] Argüden, Y. Ve Erşahin, B., *Veri madenciliği:Veriden Bilgiye, Masraftan Değere,* Arge danışmanlık ve Alkim kağıt san. Ve Tic. A.Ş.,İstanbul, 2008. http://www.scribd.com/doc/39058683/Veri-Madencili%C4%9Fi.

[12] Akyokuş S., "*Veri Madenciliği Yöntemlerine Genel Bakış"*, Türkiye Bilişim Derneği Veri Madenciliği Günü, Atatürk Üniversitesi, Erzurum, 2006.

[13] Worsley, J. and Drake J., 2002. Practical PostgreSQL, 2010. http://chestofbooks.com/computers/databases/postgresql/practicalpostgresql/index.html.

[14] Lori Bowen Ayre,. 2006. *Data Mining for Information Professionals.* http://www.slideshare.net/Tommy96/data-mining-for-information-professionals-lori-bowen-ayre.

[15] Tolun, M.R. ve Sever, H. 2006. *Veri Madenciliği.* Gündem Dergisi, 24, s.9-13., Çankaya Üniversitesi Yayınları, 2010.

[16] Doğan, B., *Zeki Öğretim Sistemlerinde Veri Madenciliği Kullanılması,* Doktora Tezi, Marmara Üniversitesi, 2006.

[17] Akman, M. *Veri Madenciliğine Genel Bakış ve Random Forests Yönteminin İncelenmesi: Sağlık Alanında Bir Uygulama*, Yüksek Lisans Tezi, Ankara Üniversitesi, 2010.

[18] Özekeş, S., 2003. *Veri madenciliği Modelleri ve Uygulama Alanları.* İstanbul Ticaret Üniversitesi Dergisi, 3, s.65-82, 2010.

[19] Koyuncugil, A.S. ve Özgülbaş, N., 2009. *Veri Madenciliği: Tıp ve Sağlık Hizmetlerinde Kullanımı ve Uygulamaları.* Bilişim Teknolojileri Dergisi, 2(2), s.21-32, 2010, Gazi Üniversitesi Bilişim Teknolojileri Dergisi Veritabanı.

[20] Ayık, Y.Z., Özdemir, A. Ve Yavuz, U., *Lise Türü ve Lise Mezuniyet Başarısının, Kazanılan Fakülte İle İlişkisinin Veri Madenciliği Tekniği İle Analizi.* Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 10(2), s.441-454, 2010.

[21] Akgöbek, Ö. ve Çakır, F. "*Veri Madenciliğinde Bir Uzman Sistem Tasarımı",* Akademik Bilişim 09-XI. Akademik Bilişim Konferansı Bildirileri, Harran Üniversitesi, Şanlıurfa, 2009.

[22] Döşlü, A., *Veri Madenciliğinde Market Sepet Analizi ve Birliktelik Kurallarının Belirlenmesi*, Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi, 2010.

[23] Öğüdücü, Ş.G. *Veri Madenciliği: Veri Ön İşleme*, http://ninova.itu.edu.tr/tr/dersler/bilisim-enstitusu/195/bbl-606/ekkaynaklar?g8396, 2010.

[24] Oğuzlar, A.(2003). *Veri Ön İşleme.* Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 21, s.67-76, 2010.

[25] Artinyan E.N., *Veri analizi-veri kalitesi ve Bütünlüğü*, http://www.denetimnet.net/UserFiles/Documents/Makaleler/BT%20Denetim/Veri_Analizi_Veri_Kalitesi_ve_B%C3%BCt%C3%BCnl%C3%BC%C4%9F%C3%BC.pdf, 2010.

[26] Bilekdemir, G. *Veri Madenciliği Tekniklerini Kullanarak Üretim Süresi Tahmini ve Bir Uygulama*, Yüksek Lisans Tezi, Dokuz Eylül Üniversitesi, 2010.

[27] Öğüdücü, Ş.G., *Veri madenciliği: Temel Sınıflandırma Yöntemleri*, http:// ninova.itu.edu.tr/tr/dersler/bilisim-enstitusu/195/bbl-606/ekkaynaklar?g8396, 2010.

[28] Timor, M., Şimşek, U.T., (2008). *Veri Madenciliğinde Sepet Analizi ile Tüketici Davranışı Modellemesi*. İstanbul Üniversitesi İşletme İktisadi Enstitüsü Dergisi, 19(3), s.3-10, 2010.

[29] Birant, D., Kut, A., Ventura, M., Altınok H., Altınok B., Altınok E., Ihlamur M., *"İş Zekası Çözümleri İçin Çok Boyutlu Birliktelik Kuralları Analizi"*. Akademik Bilişim Konferansı, Muğla Üniversitesi, 2010.

[30] Çinko, M., (2006). *Kredi Kartı Değerlendirme Tekniklerinin Karşılaştırılması*. İstanbul Ticaret Üniversitesi Sosyal Bilimler Dergisi, 5(9), s.143-153, 2010.

[31] Tuncer, T., Tatar, Y., "*Karar Ağacı Kullanarak Saldırı Tespit Sistemlerinin Performans Değerlendirmesi".* IV. İletişim Teknolojileri Konferansı, s.77-82, Çukurova Üniversitesi, Adana, 2009.

[32] Özdamar, E.Ö*., Veri madenciliğinde Kullanılan Teknikler ve Bir Uygulama*, Yayınlanmış Yüksek Lisans Tezi, Mimarsinan Üniversitesi, 2002.

[33] Özçakır, F.C., Çamurcu A.Y., *Birliktelik Kuralı Yöntemi İçin Bir Veri Madenciliği Yazılımı Tasarımı ve Uygulaması*. İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 6(12), s.21-37, 2009.

[34] Çıngı, H., *Veri Madenciliğine Giriş*, Anket Düzenleme Ders Notları, http://yunus.hacettepe.edu.tr/~hcingi/ist376.htm.

[35] Güven, A., Bozkurt, Ö.Ö., Kalıpsız, O., "*Veri Madenciliğinin Geleceği".* Akademik Bilişim Konferansı, Dumlupınar Üniversitesi, Kütahya, 2007.

[36] Güven, A., Bozkurt, Ö.Ö., Kalıpsız, O., "*Gizli Anlambilimsel Dizinleme Yönteminin N-Gram Kelimelerle Geliştirilerek İleri Düzey Doküman Kümelemesinde Kullanımı".* Çukurova Üniversitesi Türkoloji Araştırmaları Merkezi, Dilbilim Araştırmaları. 2010.

[37] Hearst,M., A., *Untangling Text Data Mining. Annual Meeting of the Association for Computational Linguistics*, University of Maryland,(1999). http://people.ischool.berkeley.edu/~hearst/papers/acl99/acl99-tdm.html.

[38] Oliveira, D.D., Baião, F., Mattoso, M., (2009). "*MF-Ontology: an Ontology for the Text Mining Domain".* Universidade Federal do Rio de Janeiro, www.cos.ufrj.br/uploadfiles/1234273167.pdf.

[39] Uzundere, E., Dedja, E., Diri, B., Amasyalı, M.F., "*Türkçe Haber Metinleri için Otomatik Özetleme"*. Akıllı Sistemlerde Yenilikler ve Uygulamaları Sempozyumu, Süleyman Demirel Üniversitesi, 2008.

[40] Amasyalı, M.F, Beken, A., "*Türkçe Kelimelerin Anlamsal Benzerliklerinin Ölçülmesi ve Metin Sınıflandırmada Kullanılması".* SIU, Antalya,2009.

[41] Mooney R. J. and Nahm, U.Y., "*Text Mining With İnformation Extraction".* Department of Computer Sciences, University of Texas, Austin.

[42] Karadağ, A., Takçı, H*., " Metin Madenciliği ile Benzer Haber Tespiti"*. Akademik Bilişim Konferansı, Muğla Üniversitesi, 2010.

[43] Pilavcılar, İ.F*., Veri madenciliği İle Metin Sınıflandırma*. Yayınlanmış Yüksek Lisans Tezi. Yıldız Teknik Üniversitesi Fen bilimleri Enstitüsü, 2007.

[44] İlhan, S., Duru, N., Karagöz, Ş., Sağır, M., "*Metin Madenciliği ile Soru Cevaplama Sistemi",* Elektrik Elektronik ve Bilgisayar mühendisliği sempozyumu, Bursa, 2008.

[45] Erol, U., *Metin Madenciliği Süreçleri*, http://metinmadenciligi.com, 2010.

[46] Adsız, A., *Metin Madenciliği.* Yüksek Lisans Projesi, Ahmet Yesevi Üniversitesi, 2006.

[47] Yıldız, K., Çamurcu, Y., Doğan, B., "*Veri Madenciliğinde Temel Bileşenler Analizi ve Negatifsiz Matris Çarpanlarına Ayırma Tekniklerinin Karşılaştırmalı Analizi",* Akademik Bilişim Konferansı, Muğla Üniversitesi, 2010.

ÖZGEÇMİŞ

Harun BAYER, 1984 yılında İstanbul' da doğdu. Güngören Teknik ve Meslek Lisesi Bilgisayar bölümünde lise eğitimini tamamladıktan sonra 2004 yılında Sakarya Üniversitesi Eğitim Fakültesi Bilgisayar Öğretim Teknolojileri ve Öğretmenliği bölümünde eğitime başladı. Lisans süresince bilgisayar uygulamaları konularında ve eğitim bilimleri konularında 4 yıl süreyle eğitim aldı. Bilgisayara olan merakının yanında sosyal ve kültürel organizasyon işlerine de ilgi duymaktaydı. 2006 yılında sosyal ve kültürel Genç Öğretmenler Kulübünü kurdu ve 5000 ' e yakın üyesi bulunan bu kulübü 2 yıl süreyle yönetti. 2008 yılında Lisans eğitimini tamamladı. Eğitim ve bilgisayara karşı özel bir ilgisi olması sebebiyle, iş hayatının da bu alanlarda olmasına özen gösterdi. Bu istekler doğrultusunda 2008 yılında Beykent üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği anabilim dalında yüksek lisans eğitimine başladı. Yüksek lisans eğitimine devam ederken aynı zamanda liselerde bilgisayar öğretmenliğine devam etti. Bilgisayar sektöründeki çalışmalarına devam etmektedir.