T.C
SELÇUK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

EDEBİYAT ESERLERİNİN WEB VERİLERİNE DAYANARAK SINIFLANDIRILMASI

Ercan CANHASI

YÜKSEK LİSANS TEZİ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

Konya, 2007

T.C
SELÇUK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

EDEBİYAT ESERLERİNİN WEB VERİLERİNE DAYANARAK SINIFLANDIRILMASI

Ercan CANHASI

YÜKSEK LİSANS TEZİ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

Konya, 2007

Bu tez 12/07/2007 tarihinde aşağıdaki jüri tarafından oybirliği ile kabul edilmiştir.

Prof.Dr.Ahmet ARSLAN	**Prof.Dr. Şirzat KAHRAMANLI**
(A.B.D Bşk.- Danışman)	**(Üye)**

Yrd.Doç.Dr. Nihat YILMAZ

(Üye)

ÖZET

Yüksek Lisans Tezi

EDEBİYAT ESERLERİNİN WEB VERİLERİNE DAYANARAK SINIFLANDIRILMASI

Ercan CANHASI

Selçuk Üniversitesi Fen Bilimleri Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Prof. Dr. Ahmet ARSLAN

2007, 95 Sayfa

Bu tezde bir metin madenciliği uygulaması yapılıştır. Sunulan çalışmada metin sınıflama ve kategorileştirme yöntemleri kullanılarak Internet'te bulunan edebiyat eserlerin sınıflanması gerçekleştirilmiştir. Yapılan işlem aslında edebiyat eserleri hakkında , Internet arama motorları sayesinde web siteler tespit etmek ve birbirleriyle ilgili web sitelerde tekrarlanan kelimeler ile edebiyat eserleri tanımlamak ve sınıflanmaktır. Edebiyat eserlerinin sınıflanması için vektör destek makineleri kullanılmıştır. Deneysel çalışma kapsamında 2 farklı deney sunulmuştur.

Madencilik açısından metnin uygun hale getirilmesi için yapılması gereken en önemli önişlemlerden biri terim seçme ve ağırlıklandırma işlemidir. Bu çalışmada bulanık mantık sistemine dayanan yeni bir terim seçme şeması da sunulmuştur.

Anahtar Kelimeler: Veri madenciliği, Metin madenciliği, Sınıflama, Terim seçme ve ağırlıklandırma, Bulanık mantık.

ABSTRACT

Master Thesis

LITERATURE WORK CLASSIFICATION ON WEB BASED DATA

Selçuk University
Graduate School of Natural and Applied Sciences
Department of Computer Engineering
Supervisor: Prof. Dr. Ahmet ARSLAN
2007, 95 Page

In this thesis one kind of text mining task is solved. In this work, text classification and categorization techniques are used for classification of literature works presented on the Internet. In particular, webpage's ranked by search engines are retrieved and analyzed to classify the literature works in term of word occurrenc frequencies on related pages. In order to classify the literature works the support vector machines are used. As a part of experimental work of this thesis 2 experiments are performed.

From the many available methods for making a text suitable to text mining tasks, the term selecting and weighting schemas are found as most importants. Additionally a new fuzzy logic system based, term selecting scheme is proposed.

Key Words: Data mining, Text mining, Term selecting and weighting, Classifying, Fuzzy Logic.

TEŞEKKÜR

Çalışmalarım boyunca değerli yardım ve katkılarıyla beni yönlendiren tez danışmanım değerli Hocam Prof. Dr Ahmet ARSLAN'a teşekkürü bir borç bilirim.

İÇİNDEKİLER

ÖZET I
ABSTRACT II
TEŞEKKÜR III
İÇİNDEKİLER IV
ŞEKİLLER, RESİMLER VE TABLOLAR V
1. GİRİŞ 1
2. KONU İLE İLGİLİ BİLGİLER 3
3. MATERYAL VE METOT 4
3.1 – Materyal 4
3.1.1 – Veri Madenciliği 4
3.1.1.1 Veri Madenciliğine Genel Bakış (Tarihçe) 7
3.1.1.2 VM Çekirdek Sistemi (MÇS) 8
3.1.1.3 VTBK İle Diğer Disiplinler Arasındaki İlişki 8
3.1.1.3.1 VTBK ile makine öğrenimi arasındaki ilişki 8
3.1.1.3.2 VTBK ile istatistik arasındaki ilişki 9
3.1.1.3.3 VM ile veri tabanı arasındaki ilişki 9
3.1.1.4 Veri Madenciliğinde Karşılaşılan Problemler 10
3.1.1.4.1 Veri tabanı boyutu 10
3.1.1.4.2 Veri madenciliğindeki gürültüler 11
3.1.1.4.3 Null değerler 11
3.1.1.4.4 Eksik veri 12
3.1.1.4.5 Artık veri 13
3.1.1.4.6 Dinamik veri 13
3.1.1.5 Veri Madenciliği Algoritmaları 14
3.1.1.5.1 Hipotez testi 15
3.1.1.5.2 Sınıflama algoritması 15
3.1.1.5.3 Kümeleme algoritması 16
3.1.1.5.4 Eşleştirme algoritması 16
3.1.1.5.5 Zaman serileri arasındaki bağımlılıklar 16
3.1.1.5.6 Sıra örüntüler 17
3.1.1.6 Veri Madenciliğini Etkileyen Eğilimler 18
3.1.2 – Metin Madenciliği 19
3.1.2.1 Giriş 19
3.1.2.2 Metin Madenciliği Tanımı 19
3.1.2.3 İlgili Araştırma alanları 21
3.1.2.4 Metin Kodlama 21
3.1.2.4.1 Metin önişleme 22
3.1.2.4.2 Filtreleme, Lemmatization ve Stemming işlemleri 23
3.1.2.4.3 Endeks Terim Seçimi 23
3.1.2.4.4 Vektör Uzay Modeli 24
3.1.2.4.5 Dilbilimi ile önişleme 25
3.1.2.5 Metin için Veri Madenciliği Metotları 25
3.1.2.5.1 Metin Sınıflama 26
3.1.2.5.2 Endeks Terim Seçimi 26

3.1.2.5.3 Naive Bayes Sınıflayıcı 27
3.1.2.5.4 En yakın komşuluk Sınıflayıcı 27
3.1.2.5.5 Karar Ağaçları 28
3.1.2.5.6 Destek vektör makineleri ve çekirdek metotları 28
3.1.2.5.7 Kümeleme 30
3.1.3 – Web Madenciliği 33
3.1.4 – Bulanık Mantık 37
3.1.4.1 Giriş 37
3.1.4.2 Bulanık Sistemler 41
3.1.4.3 Bulanık Kümeler ve Üyelik Fonksiyonları 44
3.1.4.4 Bulanıklaştırma İşlemi 45
3.1.4.5 Durulaştırma İşlemi 46
3.1.4.6 Temel İşlemler 47
3.1.5 – Terim Seçme ve Ağırlıklandırma Şemaları 49
3.1.5.1 Terim Seçme 49
3.1.5.1.1 Terim Seçme Metotları 49
3.1.5.1.2 Doküman frekansı (DF) 50
3.1.5.1.3 Enformasyon Kazancı 50
3.1.5.1.4 Müşterek enformasyon (Mutual information) 50
3.1.5.1.5 χ^2 istatistiği (CHI) 51
3.1.5.1.6 Terim güçü 51
3.1.5.2 Terim Ağırlıklandırma 52
3.1.5.2.1 İlgili Araştırmalar 53
3.1.5.2.2 İkili ağırlık metodu 53
3.1.5.2.3 Terim frekans (TF) metodu 54
3.1.5.2.4 Klasik TFIDF metodu 54
3.1.5.2.5 CDT metodu 54
3.1.5.2.6 TFRF metodu 55
3.1.5.3 Var Olan Metotlar Üstüne Bir Eleştiri 55
3.2 – Metot 56
3.2.1 Edebiyat Eserlerini Web Verilerine Dayanarak Sınıflandırma 56
3.2.1.1 Sınıflama modeli ve algoritma 59
3.2.1.2 Modelin Genel görünümü ve Akış şemaları 59
3.2.1.3 Sınıflanacak Verilerin Elde Edilmesi 60
3.2.1.4 Sınıflama için edebiyat eserlerin ve sınıfların seçilmesi 61
3.2.1.5 Edebiyat eserleri hakkında bilgi içeren web sitelerin elde edilmesi 62
3.2.1.6 Yardımcı veri tabanı tasarımı 64
3.2.1.7 Gürültü Temizliği ve Önişlemler 65
3.2.1.8 Kelime Vektörün Oluşturulması 68
3.2.1.9 Sınıflama İşlemi 71
3.2.1.10 Deneyler için geliştirilen ve kullanılan yazılımlar ve araçlar 71
3.2.2 – Bulanık Terim Seçme Şeması 73
3.2.2.1 Genel bilgi 73

3.2.2.2 Şemanın yapısı 76
3.2.2.3 Üyelik fonksiyonlar 77
3.2.2.4 Kural tablosu 79
3.2.2.5 Durulaştırma 80
3.2.2.6 Çözüm uzayı 81
3.2.2.7 Bulanık sistem için bir örnek 81
3.2.2.8 Önerilen yeni şemanın avantajları ve dezavantajları 82
4. SONUÇ 84
KAYNAKLAR 86
EK-1 93

ŞEKİL VE ÇİZELGELER

Şekil 3.1 VTBK sürecinde yer alan basamaklar 6

Şekil 3.2 VM MÇS gösterimi 8

Şekil 3.4 Örnek bir doküman ve bu doküman kelime vektör temsili 22

Şekil 3.5 Destek vektör makineler sınıflayıcı 29

Şekil 3.6 : Web Madenciliği Sınıflandırması 34

Sekil 3.7 : Web Madenciliği Sınıflandırması 35

Şekil 3.8 Klasik (Aristo) Mantık Modeli 39

Şekil 3.9 Bulanık Mantık Modeli 39

Şekil 3.10 Genel Bulanık Mantık Sistemi 42

Şekil 3.11 T-S-K Bulanık Mantık Sistemi 42

Şekil 3.12 Genel Bulanık Mantık Sistemi 43

Şekil 3.13 Isı değerlerinin bulanık ve normal kümelerle gösterimi, üyelik fonksiyonları 44

Şekil 3.15 Üçgen üyelik fonksiyonu 45

Şekil 3.16 Yamuk üyelik fonksiyonu 45

Şekil 3.17 Durulaştırma işlemin genel temsili 46

Şekil 3.18 Kelime uzayında doküman vektörlerin gösterimi 53

Şekil 3.19 Çalışmayı temsil eden Akış diyagramı 59

Şekil 3.19 Web site adreslerin elde edileme algoritması 62

Şekil 3.19 Web sitelerin elde edileme algoritması 63

Şekil 3.22 Web sitelerin madenciliğe uygun hale getirilmesi algoritması 67

Şekil 3.22 Terim seçme ve ağırlıklandırma şeması algoritması 69

Şekil 3.23 Terim seçme ve ağırlıklandırma şeması algoritması 70

Şekil 3.24 Doküman – terim uzayında İdeal terim dağılımı 74

Şekil 3.25 Doküman – terim uzayında Gerçek terim dağılımı 74

Şekil 3.26 Terim seçimi 75

Şekil 3.27 B-TSŞ için Bulanık Sistem 76

Şekil 3.28 B-TSŞ için Bulanık Sistem 77

Şekil 3.28 Kategori içi doküman frekansı 77

Şekil 3.29 Kategori dışı toplam doküman frekansı 78
Şekil 3.30 Terim seçme Ağırlık değeri için üyelik fonksiyonları 79
Şekil 3.31 Bulanık sistemin çözüm uzayı 81
Şekil 3.32 Giriş değerlerine göre elde edilecek çıkış değer örneği 82

1. GİRİŞ

İnternetinin doğuşundan bu yana sadece 20 yıl geçmesine rağmen, sahip olduğu çevrim içi bilgi ile metinsel bir devrim gerçekleştirildi. (Michael W. Berry 2003). Artık akıllı arama motorları sayesinde herhangi bir konuda bilgiye ulaşmak sadece bir sorgu yazılması ile mümkündür. Ancak internete yüklenen veri miktarı büyük hızla büyümeye devam ettiği sürece bu verilere ulaşmak da zorlaşacaktır. Geleneksel veri madenciliği yapısal veri kaynakları (veri tabanları) ile çalışmaktadır. Web siteleri ise, çoğu zaman kalıpsız yada yarı-kalıplı kaynaklar olduklarından, veri madenciliği onlara yeterli seviyede uygulanamamıştır. Bu yüzden Internet kaynaklı HTML sayfalarından (içlerindeki düz metinden) bilgi keşfetmek için, veri madenciliği temeline dayanan web ve metin madenciliği kullanılmaktadır.

Çalışmanın temelinde Internet'te mevcut olan fakat önişlemler ve bilgi keşfi yapılarak ulaşılabilecek bilgilerin belirli bir amaç doğrultusunda elde edilip kullanılması ana fikir olarak konulmuş ve bu yol ile yaygın kitapların türlerine göre sınıflandırılması hedeflenmiştir. Burada kastedilen sıradan bir sınıflama görevi yerine web den toplanacak veriler sayesinde bir sınıflama gerçekleştirmek. Çalışmanın önemi de ayni sebeptendir, sınıflamanın web verilerine dayanarak hangi derecede başarılabileceğini test etmek. Bu denemenin gerçekleşebilmesi için gereken yazılımlar ve araçların geliştirilmesi de deneysel çalışmaların bir parçasıdır.

Metin madenciliğinde, metinin sayısal gösterimini elde etmek için terim ağırlıklandırma şemaları kullanılır. Metin madencilik yöntemlerinden olan metin sınıflama tekniği metnin çok boyutluluğuyla çalışamamaktadır. Metnin çok boyutlu özelik uzayı sınıflamaya uygun hale getirilmesi için bu özelik sayısının düşürülmesi gereklidir. Boyut indirgeme işlemi aslında metni oluşturan terimlerin içinden bazılarının seçilmesidir. Bu çalışmada yeni bir terim ağırlıklandırma ve seçme şeması önerilmekte. Tez kapsamında yapılan sınıflama görevinde önerilen bu yeni şema eski var olan şemalarla da karşılaştırılmaktadır.

Tezin amacı nispeten detaylı olarak aşağıdaki şekilde açıklanmıştır:

- Genel bir sınıflandırma yapmak
  - Web sitelerinde geçen kelimelere dayanan bir sınıflama modeli geliştirmek.
  - Aşağıda listelenen işlemleri otomatik gerçekleştirecek bir algoritma ve yazılım geliştirmek
    - İlgili web siteleri tespit etmek, yerel veri tabana eklemek
    - Web sitelerini düz metne dönüştürmek
    - Metnin madenciliğe uygun hale getirmek
    - Sınıflama yani öğrenme sürecini gerçekleştirmek
    - Sınıflama modelini test etmek
- Yeni bir terim seçme ve ağırlıklandırma şeması sunmak.
  - Önerilen yeniliğin var olan yöntemlerle karşılaştırmak.

Çalışmanın önemi:

Bu çalışmada denenen sınıflama modeli ve bunun uygulanmasıyla elde edilen sonuçlar çalışmanın önemini yansıtmakta. Çalışmada web sitelerde hali hazırda bulunan veriler kullanılarak sınıflama denemesi yapılmıştır ve sonuç olarak %75 e yakın bir başarıyla gerçekleşen sınıflama, kısmen de olsa web sitelerde bulunan verilere güvenerek sınıflama yapılabileceği gösterilmiştir.
Ayrıca bu çalışmada, metin dokümanlarını oluşturan kelimeler ya da terimlerin filtrelenmesini sağlayacak yeni bir terim seçme metodu ve bu terimlerin ağırlık değerleri ile temsil edilebilmelerini sağlayan ağırlıklandırma şeması önerilmiştir.

Tez metnin ilk ana başlığında teorik bilgiler verilmekte. Çalışmanın dolaylı veya dolaysız ilgili olduğu konular sırayla açıklanmaktalar. İlk bölümde her şeyin kaynağında bulunan veri madenciliği ele alınmıştır. İkinci konuda bu çalışmanın temelinde bulunan fikri bize sunan metin madenciliği konusu anlatılmakta. Devamında web madenciliğine değinilmektedir. Dördünce kısım metin madenciliğinde kullanılan terim seçme ve ağırlıklandırma şemaları anlatılmaktadır. Son bölümü ise bulanık mantığa ayrılmıştır.

Tez metnin ikinci ana başlığı pratik çalışmaları ve uygulamaları anlatılmaktadır. Kitapların sınıflandırılması için hazırlanan sınıflama modeli altyapısı ve gerçek uygulaması ilk bölümde anlatılmakta. İkinci kısımda ise önerilen yeni terim seçme ve ağırlıklandırma şeması tanıtılmaktadır.

3. MATERYAL VE METOT

3.1 Materyal

3.1.1 Veri Madenciliği

Bu bölümde veri madenciliğinden, veri tabanında bilgi keşfinden, kısaca veri madenciliği algoritmalarından, veri madenciliği ön veri işlemeden ve veri madenciliğinde karşılaşılan sorunlardan bahsedilecektir. En son olarak veri madenciliğinde kullanılan araçlara değinilecektir.

Bilgisayar sistemleri her geçen gün hem daha ucuzluyor, hem de güçleri artıyor (Alpaydın 1999). İşlemciler gittikçe hızlanıyor, disklerin kapasiteleri artıyor. Artık bilgisayarlar daha büyük miktardaki veriyi saklayabiliyor ve daha kısa sürede işleyebiliyor Bunun yanında bilgisayar ağlarındaki ilerleme ile bu veriye başka bilgisayarlardan da hızla ulaşabilmek mümkün olabilmektedir. Bilgisayarların ucuzlaması ile sayısal teknoloji daha yaygın olarak kullanılıyor. Veri doğrudan sayısal olarak toplanıyor ve saklanıyor. Bunun sonucu olarak da ayrıntılı ve doğru bilgiye ulaşabiliyoruz.

Örneğin eskiden süper marketteki kasa basit bir toplama makinesinden ibaretti. Müşterinin o anda satın almış olduğu malların toplamını hesaplamak için kullanılırdı. Günümüzde ise kasa yerine kullanılan satış noktası terminalleri sayesinde bu hareketin bütün detayları saklanabiliyor. Saklanan bu binlerce malın ve binlerce müşterinin hareket bilgileri sayesinde her malın zaman içindeki hareketlerine ve eğer müşteriler bir müşteri numarası ile kodlanmışsa bir müşterinin zaman içindeki verilerine ulaşmak ve analiz etmek mümkün olabilmektedir. Bütün bunlar marketlerde kullanılan barkot, bilgisayar destekli veri toplama ve işleme cihazları sayesinde mümkün olmaktadır.

Verilen market örneğinde olduğu gibi ticari, tıp, askeri, iletişim, vb. birçok alanda benzer teknolojilerin kullanılması ile veri hacminin yaklaşık olarak her yirmi ayda iki katına çıktığı tahmin edilmektedir (Frawley 1991).

Verilerin ne kadar hızlı toplandığını ve işlemesinin imkânsız bir noktaya geldiğini en belirgin bir şekilde NASA kurumunda görmekteyiz (Fayyad 2000). NASA'nın kullandığı uyduların sadece birinden, bir günde terabayt'larca veri gelir.

Veri kendi başına değersizdir. İstediğimiz, amacımız doğrultusunda bilgidir. Bilgi bir amaca yönelik işlenmiş veridir. Veriyi bilgiye çevirmeye *veri analizi* veya *bilgi keşfi* (BK) denir. Bu tanımda keşif sözcüğünün kullanılmasının amacı, gizli olan ve daha önceden bilinmeyen örüntülerin bulunmasından kaynaklanmaktadır. Bilgi, bir soruya yanıt vermek için veriden çıkardığımız anlam olarak da tanımlanabilir. Veri sadece sayılar veya harfler değildir; veri, sayı ve harfler ve onların anlamıdır. Veri hakkındaki bu veriye *üstveri* diyoruz.
Veri hacminin hangi boyutlara ulaşabileceği ve bunların işlenmesinin ne kadar güç olduğu kolayca anlaşılabilmektedir. Süper market örneği incelendiğinde, veri analizi yaparak her mal için bir sonraki ayın satış tahminleri çıkarılabilir; müşteriler satın aldıkları mallara bağlı olarak gruplanabilir; yeni bir ürün için potansiyel müşteriler belirlenebilir; müşterilerin zaman içindeki hareketleri incelenerek onların davranışları ile ilgili tahminler yapılabilir. Binlerce malın ve müşterinin olabileceği düşünülürse bu analizin gözle ve elle yapılamayacağı, otomatik olarak yapılmasının gerektiği ortaya çıkar. Veri madenciliği burada devreye girer:

Veri madenciliği büyük miktarda veri içinden gelecekle ilgili tahmin yapmamızı sağlayacak bağıntı ve kuralların bilgisayar programları kullanarak aranmasıdır. Geleceğin, en azından yakın geleceğin, geçmişten çok fazla farklı olmayacağını varsayarsak geçmiş veriden çıkarılmış olan kurallar gelecekte de geçerli olacak ve ilerisi için doğru tahmin yapmamızı sağlayacaktır.

Büyük miktarlarda verinin VT'lerde tutulduğu bilindiğine göre bu verilerin VM teknikleriyle işlenmesine de veri tabanında bilgi keşfi denir (VTBK). Büyük hacimli olan ve genelde veri ambarlarında tutulan verilerin işlenmesi yeni kuşak araç ve tekniklerle mümkün olabilmektedir. Bundan dolayı bu konularda yapılan çalışmalar güncelliğini korumaktadır. Bazı kaynaklara göre; VTBK daha geniş bir disiplin olarak görülmektedir ve VM terimi sadece bilgi keşfi (BK) metotlarıyla uğraşan VTBK sürecinde yer alan bir adımdır (Fayyad et al.,

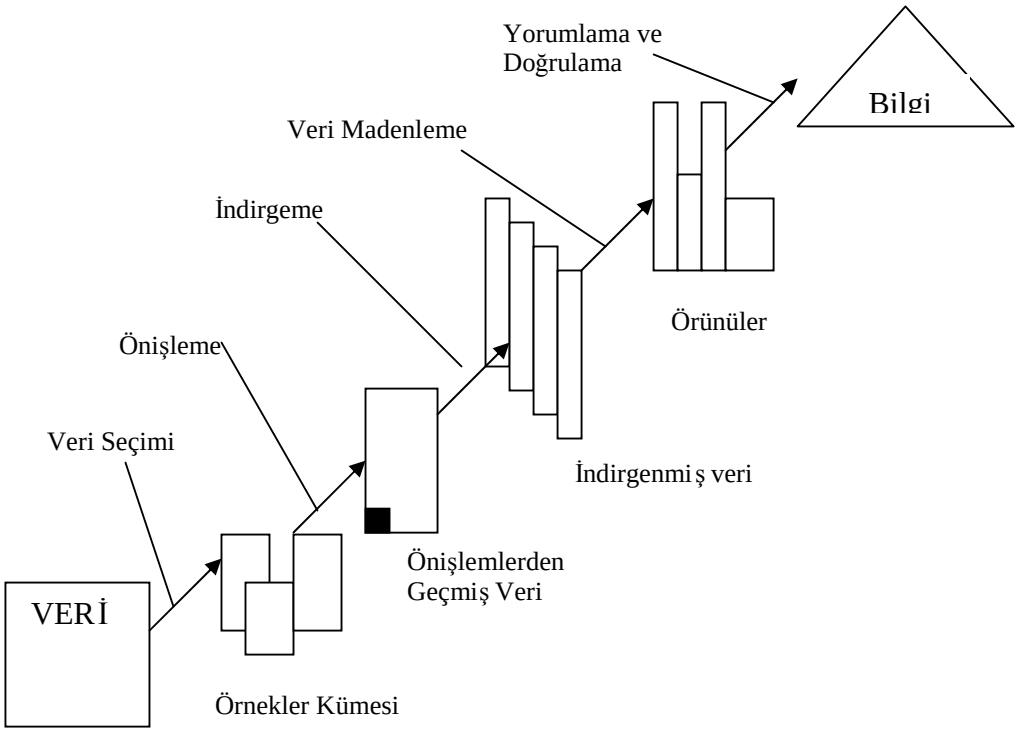
1996a). Prof. Dr. Usama Fayyad'a göre VTBK sürecinde yer alan adımlar şu şekilde sıralanmaktadır (Fayyad et al., 1996b):

1. Veri Seçimi: Bu adım veri kümelerinden sorguya uygun verilerin seçilmesidir. Elde edilen verilere örneklem kümesi denmektedir.

2. Veri Temizleme ve Ön işleme: Örneklem kümesi elde edildikten sonra, örneklem kümesinde yer alan hatalı tutanakların çıkarıldığı ve eksik nitelik değerlerinin değiştirildiği aşamadır. Bu aşama seçilen veri madenciliği sorgusunun çalışma zamanını iyileştirir.

3. Veri Madenciliği: veri temizleme ve ön işlemden geçen örneklem kümesine VM sorgusunun uygulanmasıdır. Örnek VM sorguları: kümeleme, sınıflandırma, ilişkilendirme, vb. sorgulardır.

4. Yorumlama: VM sorgularından ortaya çıkan sonuçların yorumlanma kesimidir. Burada geçerlilik, yenilik, yararlılık ve basitlik açılarından üretilen sonuçlar yorumlanır.Bu basamaklar Şekil 3.1'de ifade edilmiştir.



Şekil 3.1 VTBK sürecinde yer alan basamaklar

VM için yapılan diğer tanımlardan bazıları da şunlardır:

1. Holsheimer tarafından yapılan bir tanıma göre VM, büyük veri kümesi içinde saklı olan genel örüntülerin bulunması olarak açıklanmıştır (Holsheimer and Siebes, 1994).

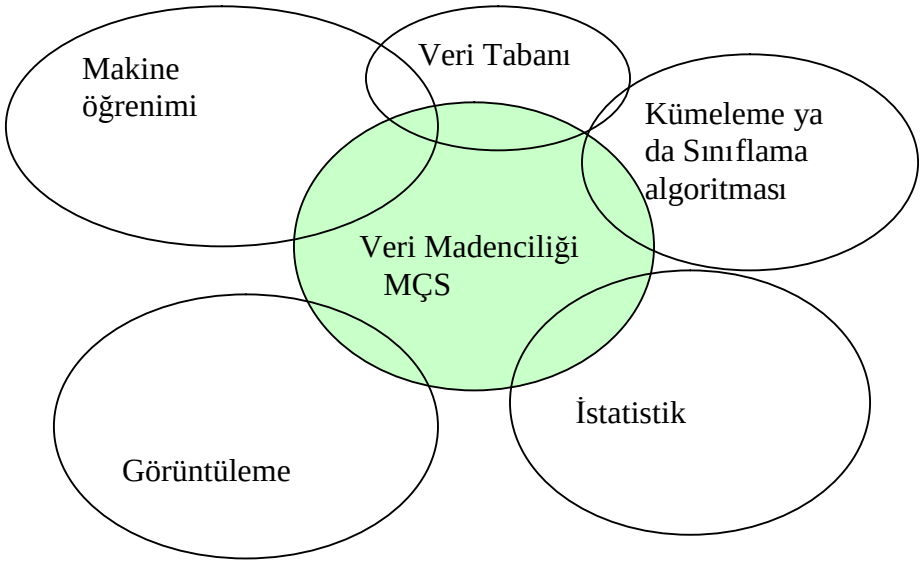
2. VM, önceden bilinmeyen ve potansiyel olarak faydalı olabilecek, veri içindeki gizli bilgilerin çıkarılmasıdır (Frawley et al., 1991).

3.1.1.1 Veri Madenciliğine Genel Bakış (Tarihçe)

VM yaklaşımı ortaya çıkmadan önce, büyük veri tabanlarından faydalı örüntüler elde etmek için, çevrim-dışı veri üzerinde çalışan istatistiksel paketler kullanılırdı. İstatistiksel yaklaşımların kullanımında bu paketlerin dezavantajları ortaya çıkmaktaydı. Bu dezavantajlardan en önemlisi; istenen verilerin toplanmasından ve amacın belirlenerek istatistiksel yaklaşımların uygulanmasından sonra bir uzman tarafından değerlendirilmesi gerekliliğidir. Başka bir dezavantajı ise her farklı ihtiyaç için bu işlemlerin tekrarlanmasıdır. Bu sorun VTBK'de kısmen aşılmıştır. VTBK (Matheus, 1993) çok büyük hacimli verilerden anlamlı ilişkileri otomatik keşfeder.

3.1.1.2 VM Çekirdek Sistemi (MÇS)

VTBK farklı disiplinleri biraraya getiren bir sistemdir. VM Çekirdek Sistemi (MÇS) (*The Mining Kernel System*) Şekil 3.2'de gösterilmiştir [http://www.pccc.qub.ac.uk/tec/courses/datamining/ohp/dm-OHP-final_2.html, 2002].



Şekil 3.2 VM MÇS gösterimi

VM MÇS'ni oluşturan diğer disiplinlerle VTBK arasındaki ilişki izleyen kesimde anlatılacaktır.

3.1.1.3 VTBK İle Diğer Disiplinler Arasındaki İlişki

3.1.1.3.1 VTBK ile makine öğrenimi arasındaki ilişki

Makine öğrenimi gözlem ve deneye dayalı ampirik kuralların otomatik biçimde bulunması olan VTBK sistemleri ile yakından ilgilidir. Genel olarak makine öğrenimi ve örüntü tanıma alanlarında yapılan çalışmaların sonuçları VTBK'de veri modelleme ve örüntü çıkarmak için kullanılmaktadır. Bu çalışmalardan bazıları:

Örneklerden öğrenme, düzenli örüntülerin keşfi, gürültülü ve eksik veri ve eksik belirsizlik yönetimi olarak sayılabilir.

VTBK'nın makine öğreniminden en büyük farkı aşağıda sıralanmıştır:

- . VTBK büyük veri kümeleriyle çalışabilir,
- . VTBK gerçek dünya verileriyle uğraşır.

Veri görselleştirmede kullanılan yöntemler, VTBK sistemi ile elde edilen örüntülerin, kullanıcıya grafikler aracılıyla sunumunu sağlar.

3.1.1.3.2 VTBK ile istatistik arasındaki ilişki

İstatistik ile VTBK arasındaki ilişkinin ana sebebi veri modelleme ve verideki gürültüyü azaltmadan kaynaklanmaktadır. İstatistiğin VTBK'de kullanılan tekniklerinden bazıları aşağıda sıralanmıştır:

- . Özellik seçimi (Corinna, 1995),
- Veri bağımlılığı (Zhong and Ohsuga, 1994; Shapiro and Matheus, 1992),
- Tanıma dayalı nesnelerin sınıflandırılması (Chan and Wong, 1991),
- . Veri özeti (Shapiro and Matheus, 1992),
- Eksik değerlerin tahmini (Elder-IV and D. Pregibon, 1995),
- Sürekli değerlerin ayrımı (Zhong and Ohsuga, 1994; Fayyad and Iranı, 1993), vb.

3.1.1.3.3 VM ile veri tabanı arasındaki ilişki

VM sorgularına girdi sağlamak amacıyla VT kullanılmaktadır. VT'deki sorgu cümlecikleri VM'nin istediği örneklem kümesini elde etmek amacıyla kullanılmaktadır. Özellikle ilişkilendirme sorgusunda fazla miktarda VT sorgusu yapmak gerekmektedir.

VM, VT'den farklıdır, çünkü VT'de var olan örüntüler için sorgular çalıştırılırken, VM'deki sorgular genelde keşfe dayalı ve ortada olmayan örüntüleri keşfetmeye dayalıdır.

3.1.1.4 Veri Madenciliğinde Karşılaşılan Problemler

Makina öğrenimiyle VM arasındaki farklar sıralanırken şu önemli detay hemen söylenir: MÖ küçük deneysel verilerle uğraşırken VM büyük hacimli gerçek dünya verileriyle uğraşır. Bu fark VM'de büyük sorunlar oluşturur. Bundan dolayı mesela küçük veri setleriyle ve yapay hazırlanmış verilerle doğru çalışan sistemler büyük hacimli, eksik, gürültülü, NULL değerli, artık, dinamik verilerle yanlış çalışabilir. Bundan dolayı bu sorunların aşılması gerekmektedir.

3.1.1.4.1 Veri tabanı boyutu

Veri tabanı boyutu 2. bölümün başında verilen NASA örneğinde olduğu gibi veri hacmi büyük boyutlara ulaşmaktadır (Fayyad et al., 2000). VT'de tutulan verilerin boyutu iki boyutlu olarak genişlemektedir:

- **Yatay Boyut:** nesnelerin özellik sayılarıyla genişlemektedir.
- **Dikey Boyut:** nesnelerdeki kayıt sayısıyla genişlemektedir.

Geliştirilen pek çok algoritma yüzler mertebesindeki verilerle uğraşacak şekilde geliştirildiğinden aynı algoritmanın yüz binlerce kat daha fazla kayıtlarla çalışabilmesi için azami dikkat gerekmektedir. Veri hacminin büyüklüğünden kaynaklanan sorunun çözümü için uygulanacak alternatif çözümlerden bazıları:

- Örneklem kümesinin yatay ve dikey boyutta indirgenmesi,
  - Yatay indirgeme: Nitelik değerlerinin önceden belirlenmiş genelleme sıradüzenine göre, bir üst nitelik değeri ile değiştirilme işlemi yapıldıktan sonra aynı olan çokluların çıkarılma işlemidir.
  - Dikey indirgeme: Artık niteliklerin indirgenmesi işlemidir.
- VM yöntemleri sezgisel/buluşsal bir yaklaşımla arama uzayını taramalıdır, vb.

Örneklem kümesinin geniş olması bulunacak örüntüleri ne kadar iyi tanımlıyorsa, bu büyük kümeyle uğraşma zorluğu da o kadar artmaktadır.

3.1.1.4.2 Veri Madenciliğindeki gürültüler

Veri girişi veya veri toplanması esnasında oluşan sistem dışı hatalara gürültü denir. Veri toplanması esnasında oluşan hatalara ölçümden kaynaklanan hatalar da dâhil olmaktadır. Bu hataların sonucu olarak VT'de birçok niteliğin değeri yanlış olabilir.

Günümüz ticari ilişkisel veri tabanları bu tür hataların ele alınması için az bir destek sunmaktadır. VM'de kullanılan gerçek dünya verileri için bu sorun ciddi bir problemdir. Bu sebepten dolayı VM tekniklerinin gürültülü verilere karşı daha az duyarlı olması gerekir.

Sistemin gürültülü veriye daha az duyarlı olmasından kasıt, gürültülü verilerin sistem tarafından tanınması ve ihmal edilmesidir.

Chan ve Wong (1991), gürültünün etkisini azaltmak için istatistiksel yöntemler kullanmıştır. Sınıflama üzerine yaptığı çalışmalardan tanınan Quinlan'nın gürültünün sınıflama üzerine etkileri konusunda yaptığı çalışmada; etiketli öğrenmede etiket üzerindeki gürültünün öğrenme algoritmasının performansını doğrudan etkileyerek düşürdüğünü tespit etmiştir (Quinlan, 1986).

Tümevarımsal karar ağaçlarında uygulanan metotlar bağlamında gürültülü verinin yol açtığı problemler araştırılmıştır (Quinlan, 1986).

3.1.1.4.3 Null değerler

Eğer VT'de bir nitelik değeri NULL ise o nitelik bilinmeyen ve uygulanamaz bir değere sahiptir. VT'de birincil anahtar haricindeki herhangi bir niteliğin özelliği NOT NULL (NULL olamaz) şeklinde tanımlanmadığı sürece bu niteliğin değeri NULL olabilir.

Kurulacak bir ilişkide kullanılacak verilerin aynı sayıda niteliğe ve NULL olsa bile aynı sayıda nitelik değerine sahip olması gerekir.

Lee NULL değerini ilişkisel veri tabanlarını genişletmek için aşağıdaki üç gruba girecek şekilde ayırmıştır (Lee, 1992):

- Bilinmeyen,
- Uygulanamaz,
- Bilinmeyen veya uygulanamaz.

Bu ayrımda şu ana kadar sadece bilinmeyen değer üzerinde çalışmalar yapılmıştır (Luba and R. Lasocki, 1994; Grzymala-Busse, 1991; Thiesson,1995).

Veri kümelerinde var olan NULL değerleri için çeşitli çözümler söz konusudur [Quinlan, 1986]:

- NULL değerli kayıtlar tamamıyla ihmal edilebilir,
- NULL değerli kayıtlardaki NULL değerleri olası bir değerle günlenebilir. Bu günleme için çeşitli yöntemler söz konusudur:
    - o NULL değeri yerine o nitelikteki en fazla frekansa sahip bir değer veya ortalama bir değer konulabilir,
    - o NULL değeri yerine varsayılan bir değer konulabilir,
    - o NULL değerinin bulunduğu kaydın diğer özelliklerine göre, NULL değerinin kendine en yakın değerle günlenmesi sağlanabilir, vb.

3.1.1.4.4 Eksik veri

VM'de ilişkilerin kurulabilmesi ve istenen problemin çözümüne ulaşabilmek için gereken örneklem kümesindeki 2 boyutun (bölüm 2.4.1'de tanımlanan yatay ve dikey boyutun) eksik olmaması gerekir. Bu boyuttaki eksiklikler şu şekilde olabilir:

- **Yatay boyutta:** Yatay boyuttaki eksiklik, örneklem kümesinde olması gereken nitelik veya niteliklerin olmamasıdır. Örneğin: eğer insanların göz rengiyle

alakalı bir hastalığın neye bağlı olduğu bulunmaya çalışılıyorsa, niteliklerden göz renginin örneklem kümesinde bulunması gerekir.

Dikey boyutta: Dikey boyuttaki eksiklik örneklem kümesindeki kayıtların eksik olmasıdır. Örneğin bir süper markette yaşı 10 ve 25 yaşındaki kişiler her yaptıkları alışverişte bir ürünü sürekli alıyorlarsa, bu örüntünün keşfedilmesi için örneklem kümesinde yeterli sayıda 10-25 yaş aralığına giren kayıtların bulunması gerekir. Eğer örneklem kümesinde bu kayıtlar bulunmazsa gerçek hayatta var olan bir örüntü kaçırılmış olur.

3.1.1.4.5 Artık veri

Artık veri, problemde istenilen sonucu elde etmek için kullanılan örneklem kümesindeki gereksiz niteliklerdir.

Artık nitelikleri elemek için geliştirilmiş algoritmalar, özellik seçimi olarak adlandırılır. Özellik seçimi arama uzayını küçültür ve sınıflama işleminin kalitesini de artırır (Deogun et al., 1995; Kira and Rendeli, 1992; Almuallim and Dietterich, 1991; Pawlak, 1986).

3.1.1.4.6 Dinamik veri

İçeriği sürekli değişen veri tabanlarıdır. Bunlara örnek kurumsal çevrim-içi veri tabanları gösterilebilir. Bir veri tabanındaki içeriğin sürekli değişmesi VM uygulamalarının uygulanabilmesini önemli ölçüde zorlaştırıcı sorunlar doğurmaktadır. Bu sorunlardan bazıları şunlardır:

- Ortaya çıkan VM örüntülerinin sürekli değişim halinde olan verilerden hangisini ifade ettiğinin tespitinin zorluğu ve bu üretilen sonuçların zaman içinde eski üretilen sonuçlardan farkının tespiti ve gereken yerlerin günlenme zorluğu,

- VM algoritmalarının çalışabilmesi için verilerin üzerine okuma kilidi konulması gerektiğinde, bu verilerin başka uygulamalar tarafından değişime

açık olmaması,

- VM algoritmalarının ve çevrim-içi VT uygulamalarının aynı anda uygulanmasından kaynaklanan ciddi performans düşüşlerinin olması, vb.

3.1.1.5 Veri Madenciliği Algoritmaları

Veri madenciliği algoritmaları verilerde var olan bilgiyi anlaşılabilecek kurallar olarak çıkartmaya yarayan metotlardır.

Veri madenciliği algoritmaları genel olarak iki ana gruba ayrılır (Simoudis, 1996): **Doğrulamaya dayalı algoritmalar:** Kullanıcı tarafından ispatlanmak istenen bir hipotez ortaya sürülür ve VM algoritmalarıyla bu hipotez ispatlanmaya çalışılır. Çok boyutlu analizlerde ve istatistiksel analizlerde tercih edilen metottur. Hipotez testi buna örnektir.

Keşfe dayalı algoritmalar: Doğrulamaya dayalı algoritmaların tersine bu algoritmalarda ortada ispatlanması istenen hipotezler yoktur. Tam tersine bu algoritmalar otomatik keşfe dayanmaktadır. Keşfe dayalı algoritmaların birçok kullanım alanı vardır: istisnai durumların keşfi, karar ağacı, kümeleme gibi algoritmalar bu yaklaşıma göre kurulmuştur.

VM algoritmalarından önemli olanlarına izleyen kesimlerde değinilecektir.

3.1.1.5.1 Hipotez testi

Hipotez testi algoritmaları doğrulamaya dayalı algoritmalardır. Doğrulanacak hipotez VT üzerindeki verilerle belli doğruluk ve destek değerlerine göre sınanır.

- Sınama işlemi uzman tarafından aşağıdaki ihtiyaçlardan dolayı yapılır:
- Bir kural ortaya çıkarılmak istendiğinde,
- Ortaya çıkarılmış bir kuralın budanması veya genişletilmesinde.

3.1.1.5.2 Sınıflama algoritması

Sınıf olmak için her kaydın belli ortak özellikleri olması gerekir. Ortak özelliklere sahip olan kayıtların hangi özellikleriyle bu sınıfa girdiğini belirleyen algoritma, sınıflama algoritmasıdır.

Sınıflama algoritması, denetimli öğrenme kategorisine giren bir öğrenme biçimidir. Denetimli öğrenme, öğrenme ve test verilerinin hem girdi hem de çıktıyı içerecek şekilde olan verileri kullanmasıdır.

Sınıflama sorgusuyla, bir kaydın önceden belirlenmiş bir sınıfa girmesi amaçlanmaktadır (Weiss and Kulikowski, 1991). Bir kaydın önceden belirlenmiş bir gruba girebilmesi için sınıflama algoritması ile öğrenme verileri kullanılarak hangi sınıfların var olduğu ve bu sınıflara girmek için bir kaydın hangi özelliklere sahip olması gerektiği otomatik olarak keşfedilir. Test verileriyle de bu öğrenmenin testi yapılarak ortaya çıkan kurallar optimum sayısına getirilir.

Sınıflama algoritmasının kullanım alanları sigorta risk analizi, banka kredi kartı sınıflaması, sahtecilik tespiti, vb. alanlardır.

3.1.1.5.3 Kümeleme algoritması

Kümeleme algoritması denetimsiz öğrenme kategorisine giren bir algoritmadır. Kümeleme algoritmasındaki amaç verileri alt kümelere ayırmaktır (Michalski and Stepp, 1993). Sınıflama algoritmasında olduğu gibi ortak özellikleri olan veriler bir kümeye girer. Alt kümelere ayrılmak için keşfedilen kurallar yardımıyla bir kaydın hangi alt kümeye girdiği kümeleme algoritması sayesinde bulunur. Kümeleme algoritması genelde astronomi, nüfus bilimi, bankacılık uygulamaları, vb. uygulamalarda kullanılır.

3.1.1.5.4 Eşleştirme algoritması

Eşleştirme algoritması denetimsiz öğrenme kategorisine giren bir algoritmadır. Eşleştirme algoritması sınıflama algoritmasının benzeridir (Seidman, 2000, syf:63). Sınıflama algoritmaları ile eşleştirme algoritmaları arasındaki fark, eşleştirmede sınıflandırmada olduğu gibi bir sınıfa sokulması amaçlanmaz. Eşleştirmedeki amaç örneklem kümesindeki nesnelerin nitelikleri arasındaki ilişkilerin saptanmasıdır.

Nitelikler arasındaki bütün kombinasyonlar çıkarılarak bütün niteliklerin farklı kombinasyonlarındaki farklı değerleri denenerek örüntüler keşfedilmeye çalışılır (Agrawal et al., 1993). Bu, ilişkilendirme algoritmasının sınıflandırma algoritmasından farkıdır. Her bir ilişkilendirme kuralı farklı ifadeleri verecek şekilde ortaya çıkar.

3.1.1.5.5 Zaman serileri arasındaki bağımlılıklar

Zaman serilerindeki örüntü belli bir periyotta, belli bir sıklıkta gerçekleşen olaylardır. Belli frekansla tekrarlanan bu olaylar zaman serileriyle yapılan VM algoritmaları sayesinde keşfedilir. Örneğin, müşteriler e-ticarette yazın yazlık ürünlere, kışın da kışlık ürünlere rağbet gösteriyorsa bu 6 ay periyotla tekrarlanan bir örüntüdür.

3.1.1.5.6 Sıra örüntüler

Belli bir olayın bir başkasını izlemesi sıra örüntülerini oluşturur (Agrawal and Srikant, 1995). Sıra örüntülerine örnek şu şekilde olabilir: Tenis raketi alan birinin az bir süre sonra tenis topu, daha sonra tenis raketinin tellerini ayarlamak için aleti almasıdır. Sıra örüntüleri perakende satış, telekomünikasyon ve tıp alanında kendine geniş bir uygulama alanı bulmaktadır.

3.1.1.6 Veri Madenciliğini Etkileyen Eğilimler

Temel olarak veri madenciliğini 5 ana harici eğilim etkiler (Vahaplar ve İnceoğlu,2001):

a) Veri: Veri madenciliğinin bu kadar gelişmesindeki en önemli etkendir. Son yirmi yılda sayısal verinin hızla artması, veri madenciliğindeki gelişmeleri hızlandırmıştır. Bu kadar fazla veriye bilgisayar ağları üzerinden erişilmektedir. Diğer yanda bu verilerle uğrasan bilim adamları, mühendisler ve istatistikçilerin sayısı hala aynıdır. O yüzden, verileri analiz etme yöntemleri ve teknikleri geliştirilmektedir.

b) Donanım: Veri madenciliği, sayısal ve istatistiksel olarak büyük veri kümeleri üzerinde yoğun işlemler yapmayı gerektirir. Gelişen bellek ve işlem hızı kapasitesi sayesinde, birkaç yıl önce madencilik yapılamayan veriler üzerinde çalışmayı mümkün hale getirmiştir.

c) Bilgisayar Ağları: Yeni nesil Internet, yaklaşık 155 Mbits/sn lik hatta belki de daha da üzerinde hızları kullanmamızı sağlayacak. Bu da günümüzde kullanılan bilgisayar ağlarındaki hızın 100 katından daha fazla bir sürat ve tasıma kapasitesi demektir. Böyle bir bilgisayar ağı ortamı oluştuktan sonra, dagıtık verileri analiz etmek ve farklı algoritmaları kullanmak mümkün olacaktır.

Bundan 10 yıl önceki bilgisayar ağları teknolojisinde hayal edilemeyenler artık kullanılabilmektedir. Buna bağlı olarak, veri madenciliğine uygun ağların tasarımı da yapılmaktadır.

d) Bilimsel Hesaplamalar: Günümüz bilim adamları ve mühendisleri, simülasyonu teori ve deneyden sonra bilimin üçüncü yolu olarak görmektedirler. Veri madenciliği ve bilgi keşfi, bu 3 metodu birbirine bağlamada önemli rol almaktadır.

e) Ticari Eğilimler: Günümüzde ticaret ve isler çok karlı olmalı, daha hızlı ilerlemeli ve daha yüksek kalitede servis ve hizmet verme yönünde olmalı, bütün bunları yaparken de minimum maliyeti ve en az insan gücünü göz önünde bulundurmalıdır. Bu tip hedef ve kısıtların yer aldığı is dünyasında veri madenciliği, temel teknolojilerden biri haline gelmiştir. Çünkü veri madenciliği sayesinde müşterilerin ve müşteri faaliyetlerinin yarattığı fırsatlar daha kolay tespit edilebilmekte ve riskler daha açık görülebilmektedir.

3.1.2 METİN MADENCİLİĞİ

3.1.2.1 Giriş

Metin madenciliği, doğal dil metinleri içinden yararlı bilgi keşfetme amacıyla geliştirilmiş bir özel veri madenciliği kavramıdır. Kısmen yeni bir araştırma alanı olsa da, uzun süredir araştırmalara tabi bir konudur. Belirli bir amaç için kullanışlı bilgi keşfetme için metnin analiz işlemi olarak da tanımlanabilir, metin madenciliği. Veri tabanlarındaki yapısal verilerle kıyasla metin yapısız ve madencilik algoritmaların uygulanabilirliği açısından kullanışsızdır. Yine de metin, modern yaşam ve kültürlerde, resmi bilgi alışveriş araçlarından en önemlisi ve en çok kullanılanıdır.

Bilgisayar ağların bilim ve ekonominin omurgası haline gelmesiyle birlikte makineler tarafından okunabilir dokuman sayısında patlama oldu. Ticari bilgilerin %85 metin formatında olduğunu gösteren değerlendirmeler mevcuttur (Text mining summit conference brochure,2005) Ne yazık ki, dokümanlarda bulunan genelde bulanık ve birçok anlamı olan bağıntıların, değerlendirilmesinde geleneksel mantık-tabanlı programlama paradigmaları yetersiz kalmaktadırlar. Metin madenciliği bu bulanıklık ve çok anlamlılıkla bir taraftan, çok boyutlu veriler ve yapısız doğal dil metinleriyle diğer taraftan, idare edecek metotların birleşimi olarak da görülebilir.

Bu bölümde metin madeniliği kavramın, bilgi keşfetme, bilgi yeniden getirme, makine öğrenmesi, istatistik ve özelikle veri madenciliği, disiplinleri ile ilişkisi acısından tanımı yapılacak. Devamında metin madenciliği metotları ve bunalar hakkında bilimde son durum (state of the art) raporu verilecek.

3.1.2.2 Metin Madenciliği Tanımı

Metin madenciliği metinden bilgi keşfetme işlemidir. Bu işlem makine destekli bir analizdir. Bu tanım ilk olarak Feldman tarafından yapılmıştır (R. Feldman ve I. Dagan, 1995). Metin madeniliği, bilgi keşfetme ve yeniden bulma (information retrivial), bilgi çıkarma ve doğal dil işleme tekniklerini kullanır ve

bunları istatistik, makine öğrenmesi, veri tabanından bilgi keşfetme ve veri madenciliği metotları ve algoritmaları ile birleştirir. Bütün bu yukarıda sayılan disiplinlerle bu kadar iç içe olması onlardan metotlar ve algoritmalar devralması, metin madenciliğinin anlamını ve gerekliliğini sorgulamamıza sebep olabilir. Ancak, metin madenciliğinin sunduğu yenilik de bu nokta da ortaya çıkmaktadır. Metin madenciliği diğer disiplinlerin aksine metin yani yapısal olmayan veriler ile çalışır. Metin madenciliğini tanımlamayı denesek en kolay yol yukarıda sözü geçen ve metin madenciliğinin yakın bağlantısı olan araştırma konularına atıfta bulunmak olurdu. Her bir saha için birer metin madenciliği tanımı ortaya çıkardı.

Metin Madenciliği – Bilgi çıkarma (Information Extraction). Gerçeklerin metinden elde edilmesi.

Metin Madenciliği – Metin Veri Madenciliği. Veri madenciliğine benzer olarak- Kullanışlı, anlamlı örüntülerin bulunması amacıyla makine öğrenmesi ve metin istatistiği alanlarından metotların ve algoritmaların bir uygulamasıdır. Bu amaçla metin ilk önce uygun ön işlemlerden geçer. (U. Nahm ve R. Money, 2002) (R. Gaizauskas, 2003).

Metin Madenciliği – Veri Tabanlarından Bilgi Keşfetme Literatürde metin madenciliği, sadece veri madenciliği ve istatistik algoritmalarını kullanan yeni bir alan olarak tanımlanmaz. Bunun yanında metin madenciliği veri tabanından bilgi keşfi prosesinin bir çok adımını kendi işlemlerinde kullanan bir proses olarak da tanımlanmakta (cri 1999). Hearst ise (M. Hearst 1999) bu tanımı şöyle özetlemekte: geniş metin koleksiyonundan henüz keşfedilememiş bilgilerin bulma işlemi. Kadratoff (1999) ve Gomez (2002) de metin verilerine uygulanan proses yönelimli bir yaklaşım olarak tanımlamakta metin madenciliğini.

3.1.2.3 İlgili Araştırma alanları

Metin betimleme, sınıflama, kümeleme, bilgi keşfi, gizli paternlerin bulunması ve modellenmesi konuları, metin madenciliğinde aktüel araştırmaların yoğunlaştığı konulardır. Bununla birlikte veri madenciliğinde bilinen ve uygulanan metotların ve algoritmaların metin verilerine uygulanabilmesi için gerekli adaptasyonlarının gereksimi de bir ek araştırma alanıdır. (M. Hearst, 1999), (Sparck-Jones ve P. Willett, 1997), (G. Salton ve ark. 1975), (Y. Wilks 1997). İlerleyen bölümlerde sözü geçen araştırma alanlarına bir çok defa atıfta bulunarak daha açıklayıcı bilgiler verilecek.

3.1.2.4 Metin Kodlama

Metin kodlama, geniş metin doküman koleksiyonlarını, yani düz metin verilerini belirli önişlemler ile madencilik prosesine daha uygun hale getirme işlemidir. Daha geleneksel bir yöntem olan kelime çuvalı (bag of words) yönteminde dokümanlar bir kelime kümesi ile temsil edilirler. Geliştirilen yeni yöntemler ise dokümanlardaki sözdizimi, kelime ve cümle yapıları ve anlamlarını göz önünde bulunduran geleneksel yöntemlere kıyasla çok daha gelişmiş ve verimli sonuçlar vermektedirler. Kelime çuvalı yönteminde ise bir dokümanın sadece alt kümesi olan bir küme ile temsil edilmesi yetmemektedir, bunun yanında alt kümenin elemanları olan kelimelerin dokümana göre önemini gösterecek bir sayı (ağırlık) içermeleri lazım. Kelime kümesindeki elemanlar için bir vektör oluşturularak kelimelerin dokumandaki tekrarlanma sayısı uygun kelime endeksine göre vektöre kaydedilmesiyle işlem tamamlanır. Bu mantıkla geliştirilen metin betimleme şemaları vektör uzay modeli (G. Salton 1975), olasılık modeli (S. E. Robertson 1977) ve mantık modelidir (C. J. van Rijsbergen 1986).

Metin kodlama, geniş metin doküman koleksiyonlarını, yani düz metin verilerini belirli önişlemler ile madencilik prosesine daha uygun hale getirme işlemidir. Daha geleneksel bir yöntem olan kelime çuvalı (bag of words) yönteminde dokümanlar bir kelime kümesi ile temsil edilirler.

KELİMELER	FREKANS	hale	1
metin	3	getirme	1
kodlama	1	işlemidir	1
geniş	1	geleneksel	1
doküman	1	yöntem	1
koleksiyonların	1	olan kelime	1
düz	1	çuvalı	1
verilerini	1	yönteminde	1
belirli	1	dokümanlar	1
önişlemler	1	kelime	1
madencilik	1	kümesi	1
prosesine	1	temsil	
uygun	1	edilirler	

Şekil 3.4 Örnek bir doküman ve bu dokümanın kelime vektör temisli

3.1.2.4.1 Metin önişleme

Bir dokümanda bulunan bütün kelimelerin elde edilmesi için işaretleme (tokenization) işlemi yapılır. Bu işlem esasında dokumanı oluşturan metinden noktalama işaretleri, bir tek boşluk karakterinden daha fazla beyaz boşlukları ve diğer metin-olmayan karakterleri silme işlemidir. İşlemin sonunda kelimeler ve aralarında birer boşluk karakteri kalır.

Dokuman koleksiyonunu oluşturan bütün dokümanların bileşimi ile oluşan metin X olsun. X metnine uygulanan işaretleme işlemi sonrasında ve kelimelerin bir defa tekrarlanma koşuluyla oluşan kelime kümesine "dokuman koleksiyonunun Sözlüğü" denir.

Yukarda anlatılan algoritmanın daha resmi bir tanımı için ilk önce bazı terimler ve değişkenler tanımlayalım: D dokuman kümesi olsun, ve $T = \{t_1,...,t_m\}$ D küme sözlüğü olsun, o zaman D kümesinde bulunan $d \in D$ dokümanında bulunan herhangi bir $t \in T$ terimin (kelime) frekansı (tekrarlanma

sayısı) tf(d,t) ile gösterilir. O zaman terim vektör tanımı da $\vec{t}_d = (tf(d,t_1),...,tf(d,t_m))$ olur.

3.1.2.4.2 Filtreleme, Lemmatization ve Stemming işlemleri

Sözlük büyüklüğünü yani dokümanı tarif eden çok boyutlu yapının boyut indirgenmesi için, sözlüğü oluşturan kelimeler kümesi filtreleme, lemmatization ve stemming işlemleri ile küçültülmeli.

Filtreleme metotları sözlükte bulunan kelime sayısını, terimlerden bazılarını kümeden tamamen çıkartarak, azaltmakta. Standart filtreleme stop terimlerin filtrelenmesidir. Stop kelime filtrelemesi fikri bazı kelimelerin cümleden çıkartılmaları ile cümlede bir değişikliğe neden olunmayacağıdır. Bu kelimeler genelde bağlaçlar (ve, veya, and, or …), İngilizcide prepositions, articals gibi kelimeler de bu guruba girmektedirler. Bu kelimelerin listeleri Internet ve değişik kaynaklarda mevcutlar. Ayrıca dokümanlar içinde bir kelimenin diğer kelimelere kıyasla çok veya az defa tekrarlanması bu kelimenin ayrıştırıcı bir değeri olmadığından stop kelimesi olarak görülebilir ve filtrelenerek sonuç sözlüğünden çıkartılabilir.

Stemming metodu ise kelimelerin değişik şeklerini tespit edip kökünü bulma işlemi olarak tanımlanıyor. Kelimenin ön ve son eklerinden arınarak sadece kök kelimeye indirgenip öyle sayılması, aynı anlamı olan kelimelerin birkaç defa tekrarlanması önlenir. Sonuçta da amacımız olan sözlük büyüklüğü küçülür.

Lemmatization metotları ise kelimelerin zaman değişiklikleri ve değişik gramer kuraları ile formunu değiştirmesiyle ortaya çıkan farklı ama ayni anlamı olan kelimelerin tespit ederler.

3.1.2.4.3 Endeks Terim Seçimi

Terim sayısını daha da azaltmak için endeks yada kelime seçme algoritmaları kullanılmalıdır. (S. Deerwester ve ark. 1990), (I. H. Witten ve ark. 1999). Bu algoritmaların uygulanması ile sadece seçilmiş yada endekslenmiş terimler ile sözlük oluşmaktadır. En basit kelime seçim metotlarından entropi

temeli kelime seçme algoritmasıdır. Sözlükteki bütün t terimleri için entropi değeri bu formül ile hesaplanır :

$$W(t) = 1 + \frac{1}{\log_2 |D|} \sum_{d \in D} P(d,t) \log_2 P(d,t)$$

burada $P(d,t) = \frac{tf(d,t)}{\sum_{l=1}^{n} tf(d_l,t)}$ dır.

Burada entropi bilindiği üzere bir terimin dokümanları ayırıma başarını temsil eder. Yani bir terim eğer bir doküman kümesinde her dokumanda ayni sayıda tekrarlanıyorsa o terimin ayrıştırma özelliği düşüktür ve dolayısıyla entropi değeri de düşüktür.

3.1.2.4.4 Vektör Uzay Modeli

Vektör uzay modeli çok büyük dokümanları analiz etmenin zorluğu karşısında geliştirilen bir metnin temsil modeldir. Asıl amacı endeksleme ve bilginin yeniden getirilme disiplinlerinde kullanılmak olsa da (G. Salton 1975), bu model metim madenciliği sürecin en önemli adımı oldu. Bu model bir doküman koleksiyonu m-boyutlu uzay vektörü ile temsil eder, yani her dokuman d bir özellik vektörü ile temsil edilir $w(d) = (x(d,t_1),..,x(d,t_m))$.

Vektörün her elemanı bir kelimeyi temsil eder. Dokumanın bu vektore gömülmesi için kullanılabilecek en basit yöntem ikili sistemdir. İkili terim vektöründe her eleman terimi temsil eder ve alabileceği değer bir yada sıfırdır, bir terimin o dokümanda bulunmasını ve terimin sıfır değeri bulunmamasını gösterir. Biraz daha gelişmiş vektör oluşturma metodu ikili sistem yerine bir ağırlıklandırma fonksiyonu kullanmaktır. Kelimenin denk geldiği dokümanda tekrarlanma sayısı (frekansı), yada o dokümana ve diğer dokümanlara göre önemini yansıtan bir fonksiyon (G. Salton ve C. Buckley 1988) vektör oluşturma için kullanılabilir.

3.1.2.4.5 Dilbilimi ile önişleme

Çoğu uygulamalarda bu tür önişlemlere gerek duyulmaz. Ancak diğer yöntemlerin yetmemesi durumunda dilbilgisi kuraları kullanılarak (C. D. Manning ve H. Schutze 2001) terimlerin secimi ve terim vektörü oluşturma işlemleri yapılabilir. Bunun için alta sıralanmış yaklaşımlar kullanılır.

Cümle öğe işaretleme (POS Part-of-Speech tagging) yaklaşımı işe cümle kurmakta kullanılan öğeler (isim, fiil, zarf, bağlaç) tespit edilip önişlemde kullanılırlar.

Metin külçe (text chunking) yaklaşımı ile dokümanda tekrarlanan kelime grubu kalıpları tespit edilirler. Örneğin “cari açık”, her iki kelime tek basına bir anlam taşısa da ikisi bir arda daha çok kullanılmakta.

Parsing yaklaşımında cümle içindeki kelimelerden bir pars ağacı oluşturularak kelimelerin birbirleriyle olan ilişkisi ve kelimelerin cümleye göre önemi incelenir.

Kelime çokanlamlılığı tespiti yaklaşımı ise kelimenin dokunmada farklı anlamlarla bulunmasını tespit etmeyi amaçlar.

3.1.2.5 Metin için Veri Madenciliği Metotları

.

Veri madenciliği metotlarının metin dokümanlarına uygulanmasın temel nedeni metne bir yapı kazandırmak. Dokümanların yapısal olması ulaşım ve kullanım acısından kullanıcıya büyük derecede kolaylık sağlar. Dokümanlara ulaşım örneği çok iyi bilinen ve kütüphanelerde kullanılan endeksleme yöntemidir. Ancak ele yapılması durumunda endeksleme işlemi özelikle de doküman sayısı büyük olan durumlarda zorlaşmaktadır. İçeriği çok sık değişen WWW ve Internet gibi doküman kaynaklarından endeksi manule yenilenmiş yapılar sunmak imkânsızdır. Var olan sistemler genellikle doküman kümelerine anahtar kelimeler atayarak (sınıflama ve kategorileştirme algoritmaları) yada otomatik bir işlem ile dokümanları gruplandırarak endekslemeyi gerçekleştirmektedirler. İlerleyen başlıklar işte bu konular hakkında daha detaylı bilgiler vermektedirler.

3.1.2.5.1 Metin Sınıflama

Metin sınıflama önceden belirlenmiş sınıflara doküman atamayı hedefler (T. Mitchell 1996) Örneğin bir merkeze ulaşılan her bir haberin otomatik bir şekilde "spor", "sanat", "siyaset" gibi etiketlerden birini atama işlemidir. Uygulanacak sınıflama metoduna bağlı kalmadan temelde yapılan işlemi açıklayalım: sınıflama önceden belirlenmiş sınıflara $L \in \mathrm{L}$ atanmış dokumanlar kümesi yani bir eğitim seti tespiti $D = (d_1,...,d_n)$ ile başlar. Bir sonraki adım yeni ve sınıfı beli olmayan dokümanların sınıflanması için kullanılacak modelin temsilidir:

$$f : D \to \mathrm{L}\ f(d) = L$$

Bir sınıflama modelinin başarı ölçümü için sınıfı beli dokümanlardan rasgele ve sayısı küçük olan bir gurup oluşturulur. Bu doküman grubuna test kümesi denir. Test kümesi elemanları sınıflama modeli ile sınıflandıktan sonra gerçek sınıflara atanıp atanmadıkları sayılır ve elde edilen değer sınıflama modelin doğruluk (accuracy) ölçüsüdür.

3.1.2.5.2 Endeks Terim Seçimi

Sınıflanacak dokümanlar genellikle yüz binlerce terim içerdiklerinden sınıflama problemi karmaşıklığı yüksek bir algoritma üstüne kurulması lazım. Sınıflama problemi basitleştirmenin yolu terim sayısını azaltmaktır. En çok kullanılan terim eleme metodu enformasyon kazancı verisidir. Bir t_j terimi için enformasyon kazancı bu formül ile hesaplanır:

$$IG(t_j) = \sum_{c=1}^{2} p(L_c) \log_2 \frac{1}{p(L_c)} - \sum_{m=0}^{1} p(tj = m) \sum_{c=1}^{2} p(L_c \mid t_j = m) \log_2 \frac{1}{p(L_c \mid t_j = m)}$$

3.1.2.5.3 Naive Bayes Sınıflayıcı

Olasılık sınıflama olarak adlandırılan bu sınıflama modeli d_i dokümanı oluşturan kelimelerin olasılık mekanizması ile üretildikleri kabulü ile başlar. $L(d_i)$ sınıfına ait d_i dokümanın içindeki kelimeler sözü geçen sınıfla bir ilişkisi olduğunu farz edelim. Bu ilişki de koşulu dağılım formülü $p(t_1,..,t_{n_i} \mid L(d_i))$ olsun. O zaman Bayesian fomülu şu şekilde tanımlanır (T. Mitchell. 1997):

$$p(L_c \mid t_1,...,t_{n_i}) = \frac{p(t_1,...,t_{n_i} \mid L_c)\,p(L_c)}{\sum_{L \in \mathrm{L}} p(t_1,...,t_{n_i} \mid L)\,p(L)}$$

Burada her bir dokümanın sadece bir L sınıfına ait olmalı ve $t_1,..,t_n$ kelimelerden oluşmalı.

3.1.2.5.4 En yakın komşuluk Sınıflayıcı

Her bir sınıfı tanıyacak acık bir model geliştirmek yerine hedef setindeki sınıflanacak dokümanlara benzeyen ve eğitim setinde bulunan dokümanları tespit edebiliriz. Hedef setinde bulunan dokümanların sınıfı eğitim setindeki dokümanlara benzerliğinden adım adım çıkartılabilir. Eğer k tane benzer doküman tespit edilirse bu yaklaşım k-yakınlıkta komşu sınıflaması (k-nearest neighbor classıfıcation). Burada önemli bir nokta dokümanların benzerliğini değerlendirecek bir ölçü bulmaktır. En basitlerinden karsılaştırılan dokümanların kelime sayıları karşılaştırılmasıdır. Daha gelişmişi kosinüs benzerliğidir. (R. Baeza-Yates ve B. Ribeiro-Neto, 1999)

Bir di dokümanın Lm sınıfına ait olup olmadığını bulmak için eğitim S(di,dj) benzerlik testi ile dj eğitim seti nde bulunan dokümanlar ve di dokumanı karşılaştırılır. İlk k tane en çok benzeyen doküman seçilir.

3.1.2.5.5 Karar Ağaçları

Karar ağaçları bir kuralar kümesidir. Bu kurallar ağaç topolojisi şeklindendirler. Kökten başlayarak belirli bir sıra içinde uygulanan kulalar sayesinde bir karara verilir, burada amaç sınıflama olduğuna göre bir dizi testten sonra sınıflama gerçekleşir. Burada önemli olan bu kural setinin nasıl oluştuğu ve ağacın nasıl kurlduluğudur. Bu işlem aslında öğreneme işlemi olarak adlandırılır. Bir eğitim seti ile başlar ve böl ve yönet yöntemiyle çalışan bir öğrenim sürecidir.Bir M eğitim seti içinden bir ti terimi seçilir, bu terim eğitim setindeki dokümanların sınıfını tarifeme özelliği bulunmalıdır yani enformasyon kazancı yüksek bir terim olmalıdır. Ti terimini içeren dokümanlar M+ olarak, terimi içermeyen dokümanlar ise M- olarak adlandırılarak eğitim seti ikiye bölünür. Bu süreç recursif bir şekilde eğitim setindeki bütün dokümanların birer sınıfa atanmasıyla biter. Öğrenme ile oluşturulan ağaç artık sınıflamayı gerçekleştirecek yapıdır. Sınıfı beli olmayan dokümanları sınıflamak için bu yapıdaki testlerden geçirmek yetmektedir.

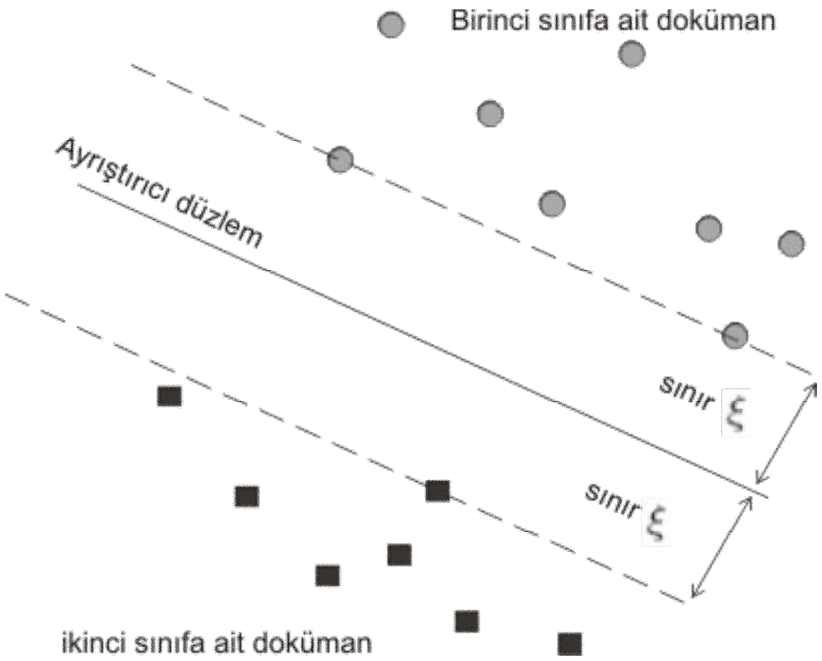
Karar ağaçları standart veri madenciliği tekniğidir (J. R. Quinlan 1986). Metnin çok boyutlu doğası yüzünden metin madenciliğine uygulanmaları daha zordur.

3.1.2.5.6 Destek vektör makineleri ve çekirdek metotları

Destek vektör makinesi (SVM), son zamanlarda metin sınıflama görevlerinde geniş kullanım bulan bir gözetmeli sınıflama algoritmasıdır (T. Joachims, 1998), (S. Dumais ve ark 1998), (E. Leopold ve J. Kindermann, 2002). Her zamanki gibi bir d dokümanını oluşturan terimlerin ağırlık değerlerinden bir vektör oluşturularak başlanır. Tek vektör ile sadece iki sınıf tespit edilir. Pozitif sınıf L1 (y = +1) ve negatif sınıf L2 (y = -1). Giriş vektörler uzayında alttaki formülde y = 0 tanımlayarak bir düzlem tanımlanabilir.

$$y = f(\vec{t}_d) = b_0 + \sum_{j=1}^{N} b_j t_{dj}$$

VDM algoritması eğitim setinden çekilen pozitif ve negatif eğitim elemanların arasında bulunan bir düzlem bulmaktadır. Bj parametresi ise en yakın pozitif ve negatif eğitim elemanı ve bu düzlem arasındaki uzaklık olan ξ değerini maksimuma çekecek şekilde uyarlanmaktadır.



Şekil 3.5 Destek vektör makineler sınıflayıcı

Burada ayrıştırıcı düzlemden ξ uzaklığında bulunan dokümanlar destek vektörleri olarak adlandırılırlar ve esas ayrıştırma düzlemin uzaydaki yerini bunlar belirlerler. Genellikle dokümanların küçük bir kısmı destek vektörüdür. Sınıfı beli olmayan yeni bir dokümanın vektörü td olsun. Bu dokümanın vektörü f(td)>0 ise o zaman doküman L1 sınıfına tersi durumunda L2 sınıfında atanır. Ancak bazen dokümanlar vektörleri lineer bir düzlem ile ayrıştırılamazlar, o zaman bazı doküman vektörleri bilerek yanlış tarafa atanırlar.

Giriş vektörlerinin bulunduğu uzayı ayrıştıracak düzlem lineer olmadığı durumlarda, destek vektör makinelerinin aynen uygulanabilmesi için giriş parametreleri bir dönüşüm ile lineerleştirilmekteler. Örneğin özellik haritası tanımlanır.

$$\phi(t_1,..,t_N) = (t_1,...,t_N, t_1^2, t_1t_2,...,t_Nt_{N-1}, t_N^2)$$

3.1.2.5.7 Kümeleme

Kümeleme metotları geniş doküman kümelerde benzer içeriklere sahip doküman gurupları tespit etmeyi amaçlar. Kümeleme işlemleri sonucunda kümeler oluşur. Her küme d sayısında doküman içerir. Kümeye ait nesneler, bu örnekte dokümanlar, diğer kümedeki nesnelere göre benzerlik göstermemeleri gerek. Bir kümeleme işleminin başarısı kümeler arası farkın büyüklüğüne bağlıdır. Yani bir kümeye ayıt nesne, doküman diğer kümelerdeki nesnelerden ne kadar farklı ise kümeleme işlemi de o derecede başarılıdır. Kümeleme metotları doküman uzayındaki dağılımını kullanarak guruplar oluştururlar. Kümeleme algoritmaları parametre olarak bir farklılık ölçüsü ve dokümanların özelliklerini kullanarak işlem yapmaktadırlar. Değişik kümeleme algoritmaları ve bunların performans testleri (SEK03) çalışmada mevcuttur.

Hangi yöntem olursa olsun kümeler birbirine benzer özellik gösteren nesnelerden oluşturulur. Böylece kümeler kendi içinde aynı özelliği taşıyan nesneleri içermiş olur. Manhattan ve Euclid uzaklık fonksiyonları çoğunlukla benzerliklerin bulunmasında kullanılır. Uzaklık fonksiyonunun sonucu yüksek bir değer ise az benzerlik, düşük bir değer ise çok benzerlik olduğunu ifade eder. P-boyutlu veri nesneleri $i:(x_{i1}, x_{i2},..., x_{ip})$, $j:(x_{j1}, x_{j2},..., x_{jp})$ için aşağıda verilen uzaklık fonksiyonları tanımlanabilir.

Euclid Uzaklık fonksiyonu:

$$d_{ij} = \sqrt{\left|x_{i1} - x_{j1}\right|^2 + \left|x_{i2} - x_{j2}\right|^2 + ... + \left|x_{ip} - x_{jp}\right|^2}$$

Manhattan Uzaklık Fonksiyonu:

$$d_{ij} = \left|x_{i1} - x_{j1}\right| + \left|x_{i2} - x_{j2}\right| + ... + \left|x_{ip} - x_{jp}\right|$$

Veri kümeleri için uygulanacak uzaklık fonksiyonlarının verimleri farklı olabilir, bundan dolayı Euclide ve Manhattan'ın haricindeki uzaklık fonksiyonları bazı veri kümeleri için daha uygun olabilir.

Kümeleme Analizinin özellikleri aşağıda kısaca özetlenmiştir:

- Denetimsiz (unsupervised) öğrenmedir.
- Kümelerin yapılarını doğrudan verilen veriden bulmadır.
- Önceden tanımlanan sınıf ve sınıf-etiketli öğrenme örnekleriyle çalışmamaktadır.
- Bir veri madenciliği fonksiyonudur.
- Kümeleme;
    - Veri dağılımını anlamada fayda sağlar,
    - Her bir kümenin özelliklerini izler.

Kümeleme probleminde;

- Bir d-boyutlu metrik uzayda n veri noktası verilmiştir,
- Veri noktaları k küme içine paylaştırılır.

Kümeleme algoritma türleri aşağıdaki gibi sınıflandırılmıştır:

- Bölümleme Kümeleme Algoritması (Partioning Clustering Algorithms)
- k-Ortalamalar
- k-medoid (CLARANS)
- EM (Beklenen Maksimizasyon) ALGORİTMASI
- Hiyerarşik kümeleme Algoritması
- Toplayıcı (Agglomerative) Hiyerarşik Kümeleme Algoritması
- Bölücü (divisive) Hiyerarşik Kümeleme Algoritması

En çok kullanılan kümeleme algoritması K-ortalamalar algoritmasıdır. K-ortalamalar algoritması bölümleme (partitioning) yöntemleri oalrak adlandırılan algoritmalardan biridir. Bölünme kümeleme problemi şöyle ifade edilmiştir: d

boyutlu metrik uzayda verilen n nesne – bu durumda doküman, aynı kümedeki nesneler diğer kümelerdekine kıyasla daha benzer olacak şekilde k kümeye yerleştirerek bölümlenmesinin yapılmasıdır. K değeri probleme göre bilinebilir yada bilinmez. Hata kareler ölçütü gibi bir kümeleme ölçütünün olması gerekir.

3.1.3.1 Web Madenciliği

Veri madenciliği ve Web son zamanların geçerli iki araştırma sahasıdır. Bu iki sahanın doğal kombinasyonu Web madenciliği olarak adlandırılır. Veri madenciliği uygulamalarından biri olan Web madenciliği, Web verileri üzerinde veri madenciliği fonksiyonlarını yerine getirir (Özakar ve Püskülcü 2002).

Birçok yazara göre web madenciliği terimi ilk kez Etzioni tarafından 1996'da ortaya atılmıştır. Bu bildiride Etzioni Web madenciliğinin veri madenciliği tekniklerini kullanarak Word Wide Web'de bulunan dosya ve servislerden otomatik olarak paternler bulmak ve öngörülmeyen bilgiye ulaşmak olduğunu iddia etmektedir (Etzioni 1996). Araştırmacıların çoğu çalışmalarında bu tanımlamayı esas almışlardır Burada bu işlemlerden bazılarının rahatlıkla arama motorları tarafından yapılabileceği akla gelebilir. Bu durumda Web Madenciliğine ihtiyaç duyulmasının iki sebebi vardır. Bunlar:

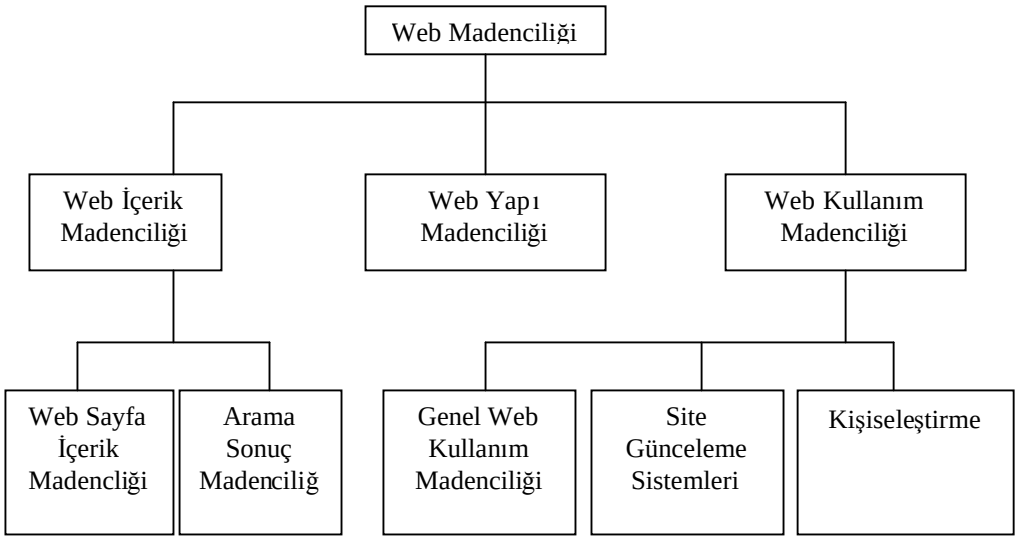
1. Google, Yahoo gibi arama motorlarını kullanıldığında genelde iki çeşit sorunla karşılaşılır: "Veri madenciliği" ile ilgili dokümanlar araştırılırken sonuç olarak çok fazla doküman listelenebilir ama bunların birçoğu araştırılan konuyla yeteri kadar ilgili değildir. Ayrıca dokümanlar sıralanırken araştırılan konuyla en çok ilgili olandan en az ilgili olana doğru sıralanmış değildir. Ancak dokümanlar incelendikten sonra istenilen sırada konuyla ilgili siteler bulunabilmektedir.

2. Arama motorlarında yine "veri madenciliği" konusunun araştırıldığı varsayılırsa, bu konu ile yakından ilgili olan makine öğrenmesi , bilgi keşfi ile ilgili dokümanlar içerisinde "veri madenciliği " kelimeleri geçmediği için sonuç olarak listelenmeyecektir. Bu sebeple son zamanlarda araştırmacılar veri madenciliği kavramını Web'e uyarlamışlardır (Sakiroglu ve ark. 2003).

Web madenciliği kabaca Web'ten faydalı bilginin keşfi olarak da tanımlanabilir. Bu tanım içinde otomatik tarama, bilgi alma ve kullanılabilir kaynakların milyonlarca web sitesi veya online veritabanlarından seçilmesi web içerik madenciliği konusuna girerken bir veya birçok web sunucu veya online

servisten kullanıcı erişim desenlerinin analiz ve keşfi Web kullanım madenciliği konusuna girmektedir (Takcı ve Sogukpınar, 2002).

Web üç tip veri bulundurur; içerik, Web log dosyaları ve Web yapı verisi. Sekil 3.8 ‘de madencilik yapılabilecek verinin sınıflandırması incelenmektedir. Bunlar Web içerik madenciliği, Web yapı madenciliği ve Web kullanım madenciliğidir.



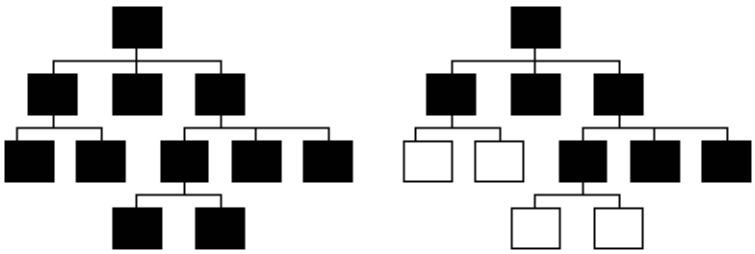
Şekil 3.6: Web Madenciliği Sınıflandırması

1. Web içerik Madenciliği: Web içerik madenciliği temel olarak Internet de saklı bilgiyi bulma üzerine yoğunlaşmıştır (arama motorları, vs.). Kısaca konusu, site içeriğidir. Adından da anlaşılacağı gibi web dokümanlarının içeriklerini yorumlamak ile ilgilenir. Web içerik madenciliği akıllı yazılım ajanları (web robotları, web örümcekler vs.) daha doğrusu makine öğrenimi veya yapay zeka ile ilgilidir. Son zamanlarda dokümanlardan bilgi çıkarma için XML de kullanılmaya başlanmıştır. Burada; saniyede binlerce web sayfasını inceleyen geniş ölçekli programlara “derleyici” (Crawler) denilmektedir (Belen ve ark. 2003). Web içerik verilerinin çoğu belli bir düzene sahip olmayan düz metinlerdir. Lycos, Alta Vista, Web Crawler gibi bilinen çeşitli arama motorları bu tekniklerden faydalanırlar.

Web içerik madenciliği, arama motorlarındaki yapının genişletilmiş hali olarak düşünülebilir. Internet de arama yapılırken birçok teknik kullanılmaktadır.

Bu tekniklerden, klasik arama motorlarında en çok kullanılan kelime tabanlı arama yaklaşımıdır. Bunun dışında, içerik hiyerarşisi, kullanıcı davranışları ve sayfalar arası link ilişkileri de kullanılan en temel yaklaşımlardandır.

Derleyicide, çekirdek URL adres setine bakarak değerlendirme başlamakta ve çekirdek URL adreslerindeki linkler kaydedilip arama bu linklerden devam etmektedir. Web'deki muazzam büyük yapı, özelleşmiş derleyici yapılarının geliştirilmesine neden olmuştur. Sekil 3.9'da genel derleyici ve özelleşmiş derleyicilerdeki arama mantığı görülmektedir. Şekildeki siyah gölgeli kısımlar derleyicinin değerlendirmeye aldığı sayfaları temsil etmektedir. Buna göre özelleşmiş derleyici bir sayfayı ilgili bulduysa sayfanın linklerini değerlendirmeye almakta, aksi halde diğer sayfaları değerlendirmeye geçmekte bir alt seviyeye inmemektedir (Dunham 2003).



Sekil 3.7: Web Madenciliği Sınıflandırması

2. Web Yapı Madenciliği: Web yapı madenciliği sitenin yapısal dizaynını iyileştirmek için kullanılır. Web sayfaları arasındaki bağlantılarını (hyperlink) ilişkilerini keşfetmekle ilgilenir. Yani HTML kodlarındaki <a href> etiketleri arasında yer alan veriyi yorumlar. Web içerik madenciliği web sayfasının içeriği ile ilgilenirken, web yapı madenciliği doğrudan web sayfaları arasındaki bağlantılar ile ilgilenir (Sakiroglu ve ark. 2003).

3. Web Kullanım Madenciliği: Web kullanım madenciliği; bir veya birçok web sunucudan kullanıcı erişim desenlerinin otomatik keşfinin ve analizin yapıldığı bir tip veri madenciliği etkinliğidir. Birçok organizasyon pazar analizleri için geliştirdikleri stratejileri ziyaretçi bilgilerine dayanarak yerine getirir.

Organizasyonlar günlük operasyonlarla her gün yüzlerce MB veri toplamaktadır. Bu bilgilerin çoğu web sunucuların otomatik olarak tuttuğu günlük dosyalarından elde edilir. Günlük dosyalarında, istemcinden sunucuya gönderilen her bir istek bir kayıt olarak tutulur (Takcı ve Sogukpınar 2002).

Web verilerinin analizi sonucunda bir ziyaretçinin sitede kalma süresi, hizmet stratejileri, etkin kampanyalar ve diğerleri bulunabilir. Ayrıca siteye bağlanan bir kullanıcının hangi amaçla siteye bağlandığı, kötü niyetli bir kullanıcı olup olmadığı da bulunabilmektedir. Bir elektronik ticaret sitesi için en iyi müşteri veri madenciliği sayesinde bulunabildiği gibi bir "hacker" da aynı yöntemlerle bulunabilir.

Web kullanım madenciliği baslıca üç fazdan oluşmaktadır: (Belen ve ark. 2003)

1. Ön İsleme : Ön isleme veri kaynağından alınan verinin desen bulmaya hazır hale getirilmesi adımıdır. Belki de web kullanım madenciliğinin en önemli aşamasıdır. Çünkü etkili bir şekilde yapıldığından zaman ve kaynak tasarrufu sağlayacaktır. Bu adımda esas olarak veri gürültüden temizlenir.

2. Desen Bulma: Veri madenciliğinde desen bulmak için kullanılan bir çok yöntem ve algoritma vardır ve bunların çoğu web kullanım madenciliğinde de kullanılmaktadır.

3. Desen Analizi: Desen analizi web kullanım madenciliğinin son adımıdır. Desen analizinin amacı bulunan desenlerden ilginç olmayan desenleri elemektir. Desen analizinin en çok karşılaşılan sekli SQL gibi bilgi sorgulama dilleri ile yapılan uygulamalardır. Bir başka yöntem ise verilerin veri küplerine yüklenerek OLAP işlemlerinin yapılmasıdır.

Web içerik madenciliği dokümanların içinden bilgi çıkarırken web kullanım madenciliği kullanıcıların erişimlerinden bilgi çıkarmaktadır. Erişimlere dayalı bilgilerle kullanıcı davranışları bulunabilmekte ve kişiye özel hizmet olanağı sağlanabilmektedir.

3.1.4.1 BULANIK MANTIK

3.1.4.1 Giriş

Bilimde, özelikle de bilim dallarının gövdesi sayılan matematikte, görecelilik yada belirsizlik genelde istenmeyen ve kaçınılan bir durum olarak tanımlanır. Klasik bilimcilere göre, bilim bütün ortaya koyduğu açıklamalarda kesinlik için uğraşmalıdır ve bundan dolayı da belirsizlik bilimsel olmayan bir şey olarak kabul görmüştür. Ancak bununun yanında bir de alternatif bakış açısı vardır ve buna göre, belirsizlik sadece kaçınılması mümkün olmayan bir durum değil aynı zamanda büyük bir fayda alanı açan ve zerinde çalışılması gereken bir durumdur.

Newton fiziğinde, belirsizliğe yer vermeyen matematiksel analizin rolü istatistiksel mekanikte, olasılık teorisi tarafından karşılanmış ve bu teori aslında belirli bir tipteki belirsizliklerin giderilmesini amaçlamıştır (Klir ve Yuan, 1995).

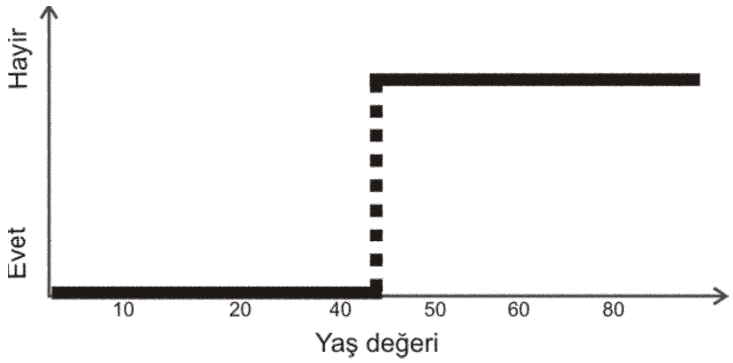
Bulanık Mantığın tarihi çok eskilere dayanmaktadır. Aristoteles'in "Var ya da Yok" yasalarına karşın Heraclitus, bir şeyin hem doğru hem yanlış olabileceği fikrini ortaya sürmüştür. Plato ise bu durumu ileriye götürerek "doğru" ve "yanlış" olmanın dışında, doğru ve yanlışın iç içe olduğu üçüncü bir durumdan bahseder. Ancak ilk kez Lukasiewicz 1900'lerin başında "olası" kavramını ortaya atmıştır. Bu kavram Bulanık Mantığın temelini oluşturur. Lukasiewicz, Doğru ile Yanlış arasında sonsuz farklı değer olduğundan bahsetmiş ve ancak bu mantık uygulamalarda çok başarı elde edememiştir. (Koyuncu E. 2004)

1930'larda ünlü Amerikan filozofu Max Black tarafından belirsizliği açıklayıcı öncü kavramlar geliştirilmiş olsa bile, bugün 1965'te Lütfü Askerzade tarafından yayınlanan makale modern anlamda belirsizlik kavramının değerlendirilmesinde önemli bir nokta olarak kabul edilmiştir. Askerzade, bu makalede kesin olmayan sınırlara sahip nesnelerin oluşturduğu bulanık küme teorisini ortaya koymuştur. Askerzade'nin bu makalesinin önemi sadece ihtimaller teorisine karşı duruşu ile ilgili değil, ayrıca ihtimaller teorisini oluşturan Aristo mantığına karşı da bir meydan okuma şeklinde gelişmiştir. Bulanık küme

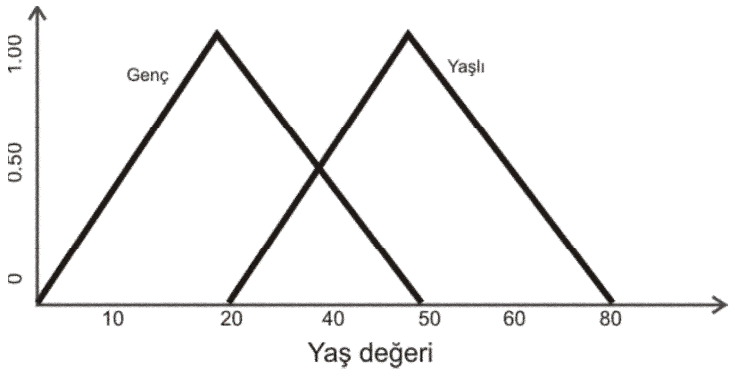
teorisinin üyelikten üye olmamaya dereceli geçişi ifade etmesindeki yeteneği geniş faydaları olan bir yetenektir. Bize, belirsizliğin ölçülmesinde güçlü ve anlamlı araçlar sunmasının yanı sıra, doğal dilde ifade edilen belirsiz kavramların anlamlı bir şekilde temsilini de vermektedir. Fakat Aristo mantığı üzerinde temellenen klasik küme teorisi, verilen bir alana ait bütün bireyleri incelenen özelliğe göre ikiye ayırır: kümeye ait olan elemanlar ve ait olmayan elemanlar. Kümeye üye ve üye olmayan elemanlar arasında kesin ve belirsiz olmayan bir ayrım vardır. Doğal dilde ifade edilen ve üzerinde çalışılan çoğu sınıflandırma kavramı, bu türde bir karakterde değildir. Örneğin; uzun insanlar kümesi, birden çok büyük sayıların oluşturduğu küme, gibi kavramlar klasik kümenin öngördüğü şekilde incelenemezler. Bu kümeler, kesin olmayan sınırlara sahip olarak kabul edilir ve üyelikten üye olmamaya geçişin dereceli olduğunu göz önüne alınarak işlem yapılır (Klir ve Yuan, 1995).

Bulanık Mantık ile çözümleme yapılırken problemin çözümü için gerekli tecrübe, gözlem bilgileri gibi klasik mantıkta ve matematik formüleri ille problemin çözümünde yararı olmayan bilgiler kullanılmaktadır. Bulanık Mantık işlemlerinde, yargılar öznel (bilirkişi) gözlemlere dayanır ve bu yargılar matematiksel olarak analize hazır bir şekilde modellenir. Matematik kural olmaktan çıkar, hayat için birkaç rakam oluverir. (Koyuncu E. 2004)

Bulanık mantık ve geleneksel Aristo mantığının karşılaştırmışını alttaki şeklerde görebilirsiniz. İlk grafikle belirsizliği,göreceliği, öznelliği içermeyen kesin bir çerçeve içinde bir kişinin yaşına göre genç olup olmadığını gösteren bir grafik görülmekte. Buradaki sorun bir kişinin yaşı 45 iken genç sayılması bir sonraki sene 46 yaşında yaşlı kabul edilmesi. Gündelik hayata bu ve benzeri kavramlar belirli bulanıklık içerirler. İkinci şekil ise geleneksel mantığa karşı genç ve yaşlı kavramları, gerçekteki bulanıklığı ve kesinsizliği içererek yansıtmakta.



Şekil 3.8 Klasik (Aristo) Mantık Modeli



Şekil 3.9 Bulanık Mantık Modeli

Bulanık Mantığın uygulama alanları günden güne genişliyor. Bulanık mantığı konu eden çok sayıda makale de yayınlanmakta. Endüstriyel Sistem modellemelerinden, yazılım geliştirmeye; otomatik kontrol sistemlerinden, veri analizine; yöneylem araştırma tekniklerinden, sosyolojik değişim kurallarını izleme gibi birçok alanda Bulanık Mantık uygulamalarını başarılı bir şekilde görmek mümkün. Özellikle Modern Kontrol Sistemleri ve Otomatik Kontrol Sistemleri, Örnek olarak, günümüzde Robotik Hareket Sistemlerinin karmaşık kontrol problemleriyle çoğunlukla Bulanık Mantık ilgileniyor (Koyuncu E. 2004).

Bulanık mantığın bu kadar uygulamaya yatkınlığı gerçek hayat problemlerinin matematiksel modeller ile temsil edilmesi güç olduğu durumlarda, basit if-then kuralarla tarif edilebilmesinden gelir.

Bulanık mantığın çok değişik alanlarda ve bu kadar çok kullanılmasının sebeplerini sıralayalım: bulanık mantığı anlamak kolaydır, bulanık mantık esnektir, bulanık mantık açık olmayan data bulunması durumunda tolerans gösterir bulanık mantıkla karmaşık lineer olmayan fonksiyonlar modellenebilir, bulanık mantık geleneksel kontrol teknikleriyle birlikte kullanılabilir, bulanık mantığın ana fikri konuşma dilidir. (Matlab 6.5)

Standart olasılık teorisi, kesinliğin bulunmadığı durumlarda iyi bir çözüm sağlar. Fakat bir çok basit problemde olasılık teorisinin hiçbir çözüm önerisi yoktur. Birkaç tipik örnek aşağıda verilmiştir.

1) Yağmurun yağma olasılığı nedir?
2) Otelden havaalanına taksiyle ne kadar sürer?
3) Robert genellikle işten saat 18:00'da döner. Saat 18:30'da evde olma olasılığı nedir?
5) Zayıflık çekicidir. Cindy zayıftır. Cindy'nin çekiciliği hakkında ne söylenebilir?

Bu tip sorularla insanlar rutin olarak karşı karşıya gelebilir ve cevap verirler. Cevaplar sayı değildir. Cevaplar bulanık algılamanın dil ile ifadesidir. "Çok yüksek değil, yaklaşık 0.8, hemen hemen olanaksız" vb. gibi insanlar gündelik hayatında karşılaştığı bu gibi problemlere cevap bulurlar (Askerzade, 2001).

Bulanık mantığın genel özellikleri Askerzade tarafından şu şekilde ifade edilmiştir:

- Bulanık mantıkta, kesin değerlere dayanan düşünme yerine, yaklaşık düşünme kullanılır.
- Bulanık mantıkta her şey [0,1] aralığında belirli bir derece ile gösterilir.
- Bulanık mantıkta bilgi büyük, küçük, çok az gibi dilsel ifadeler şeklindedir.
- Bulanık çıkarım işlemi dilsel ifadeler arasında tanımlanan kurallar ile yapılır.
- Her mantıksal sistem bulanık olarak ifade edilebilir.

- Bulanık mantık matematiksel modeli çok zor elde edilen sistemler için çok uygundur.
- Bulanık mantık tam olarak bilinmeyen veya eksik girilen bilgilere göre işlem yapma yeteneğine sahiptir (Elmas, 2003).

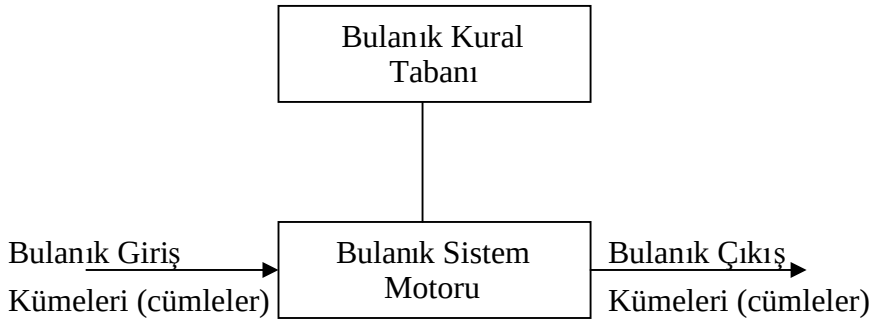
3.1.4.2 Bulanık Sistemler

Bulanık sistemler kural-temelli sistemlerdir. Bu kuralar probleme özgü tanımlanırlar ve uzman kişilerin problem ile ilgili tecrübelerini yansıtırlar. Bulanık sistemlerin merkezinde Eğer-İse (If-Then) ifadeler tablosu bulunur.

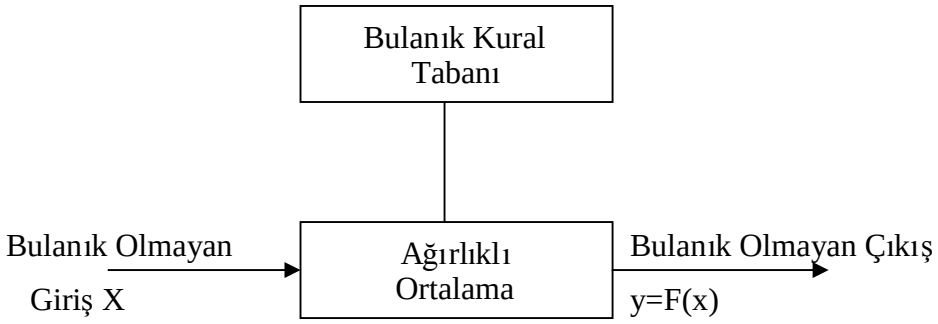
Örneğin: Eğer hizmet kötü ise fiyat ucuzdur, cümlesi bir bulanık sistem kuralı olsun. Bu kuraldaki "hizmet" giriş "fiyat" ise çıkış parametresidir. "Kötü" ve "ucuz" kelimeleri de belirsizlik ve görecelilik içeren iki kavram oldukları için üyelik fonksiyonları ile temsil edilen giriş ve çıkış değerlerdirler.

Bulanık sistemlerin literatürde çoğunlukla kullanılan üç şekli vardır. Genel bulanık sistem, Takagi-Sugeno-Kank bulanık sistemi (T-S-K) ve bulanıklaştırma durulaştırma birimli bulanık sistemi (Wang, 1997).

Geleneksel bulanık sistemlerdeki giriş ve çıkışlar bulanık kümeler yani cümlelerden oluşurlar (Şekil 3.9). Bu sistemler bu şekilde mühendislik uygulamalarında kullanılamazlar. Mühendislikte bulanık giriş ve çıkışların kullanılabilmesi için bu değerlerin bulanıklığı ortadan kaldırılarak gerçek matematiksel değerle dönüştürülmeleri gerek.

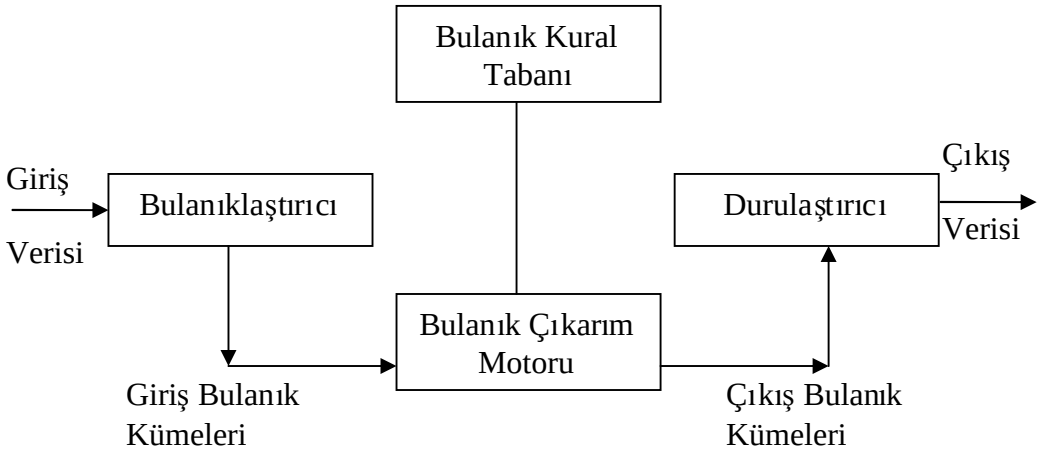


Şekil 3.10 Genel Bulanık Mantık Sistemi



Şekil 3.11 T-S-K Bulanık Mantık Sistemi

Bu problemi çözmek üzere Takagi, Sugeno ve Kang, giriş ve çıkış değerlerinin geleneksel değerler almasını sağlayan bir bulanık sistem kurdular. Takagi-Sugeno-Kang bulanık sistemi Şekil 3.10'de görülmektedir. Bu sisteme bir örnek verelim : EĞER hizmetin x değeri yüksek İSE fiyat değeri y = cx' tir. T-S-K bulanık sisteminin ana problemi, İSE kısmından sonra matematik formül bulunmasıdır ve sözel bilgileri modelleyememesi ve giriş-çıkış değişkenleri arasında yazılması mümkün olan tüm kuralların son kısımlarının bulanık olmaması dolayısı ile yazılamamasıdır. Bu sistem bulanık mantığın çok yönlü uygulamalarında araştırmacılara gerekli kolaylığı sağlamamaktadır (Wang, 1997.



Şekil 3.12 Genel Bulanık Mantık Sistemi

Bütün bu elverişsiz durumları ortadan kaldıran sistem 'Bulanıklaştırma-Durulaştırma' birimli bulanık sistemdir (Şekil 3.12). Mühendislik istemlerinde kullanılabilmesi için genel bulanık sistemlere, giriş kısmına gerçek değerleri bulanık kümelere çevirecek bulanıklaştırıcı ve çıkışa da bulanık kümeleri gerçek değerlere çevirecek durulaştırıcı eklemektir (Wang, 1997).

Bulanık sistemlerin başlıca özellikleri arasında en önemli, konu olarak, çoklu girdileri, kural tabanı ve çıkarım motoru ile işleyerek tek çıktı haline dönüştürmesi gelir. Bazı özel durumlarda, çıktılar birden fazla olabilir. Ancak, hemen her mühendislik çalışmasında en az bir tane çıktı bulunur. Bulanık sistem doğrusal olmayan bir şekilde girdileri oluşturan değişkenleri, Çıktı değişkenine dönüştürerek, sistemin davranışını tespit eder. Böylece bilgi tabanının doğrusal olmayan dönüşümlere maruz bırakılması ile istenen sonuçlara ulaşmak için incelenen sistemin kontrol altına alınması mümkün olmaktadır. Bulanık sistemler sayesinde mühendislikte görüntü işleme, zaman serileri esaslı tahmin yapmak, kontrol sorunlarını çözmek ve haberleşme yani iletişim konularında uygulamalar yapmak mümkün olmaktadır. Bunun dışında bulanık sistemler mühendislik, tıp, sosyoloji, psikoloji, işletme, uzman sistemler, yapay zekâ, sinyal işlenmesi, ulaştırma, kavşak sinyalizasyon onu gibi birçok alanda rahatlıkla kullanılabilir. (Zekai Ş. 2001)

3.1.4.3 Bulanık Kümeler ve Üyelik Fonksiyonları

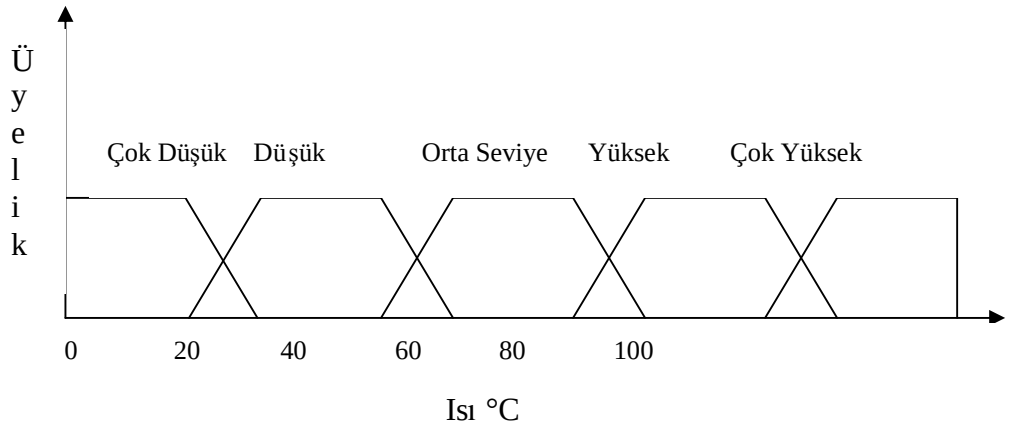
Klasik küme teorisinde elemanlar ya kümeye dâhildir ya da değildir. Bu karakteristik fonksiyonla aşağıdaki gibi ifade edilebilir:

$$U_A(x) = \{^{1, x \in A}_{0, \text{x} \notin \text{A}}$$

1965 yılında Lütfü Askerzade daha esnek üyelik değerlerinin mümkün olduğu bulanık kümeler teorisini ortaya atmıştır. Bulanık kümeler teorisinde çok değişik üyelik değerleri bulunabilmektedir. Bu üyelik değerleri 0 ile 1 aralığında değişebilmektedir. Bu karakteristik fonksiyonla şekil 3.13'te gösterilmiştir (Tsoukalas, 1997).

$$\mu_A(x) : X \rightarrow [0,1]$$

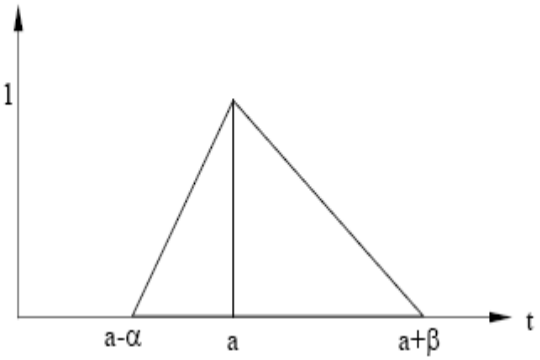
$$X = \{\frac{\mu_A(x_1)}{x_1} + \frac{\mu_A(x_1)}{x_1} + \ldots\} = \{\sum \frac{\mu_A(x_i)}{x_i}\}$$



Şekil 3.13 Isı değerlerinin bulanık ve normal kümelerle gösterimi, üyelik fonksiyonları

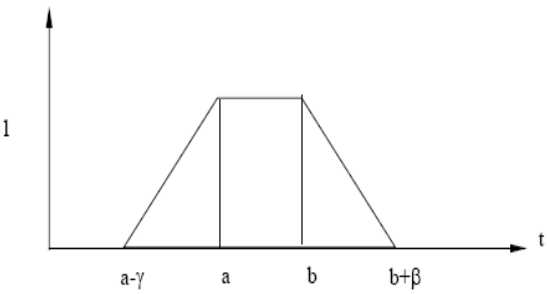
Literatürde kullanılan üyelik fonksiyonları: üçgen üyelik fonksiyonu, yamuk üyelik fonksiyonu, Gauss üyelik fonksiyonu, genelleştirilmiş çan üyelik fonksiyonu ve sigmoidal üyelik fonksiyonudur. Bunlardan en yaygın olarak kullanılanı ve bize kolaylık sağlayan üçgen üyelik fonksiyonu ve yamuk üyelik fonksiyonudur. Üçgen üyelik fonksiyonları ve yamuk üyelik fonksiyonları sırasıyla Şekil 3.15 ve Şekil 3.16'da gösterilmiştir (Wang, 1997).

$$A(r) = \begin{cases} 1-(a-t)/\gamma \mid a-\gamma \le t \le a \\ 1-(t-a)/\beta \mid a \le t \le a+\beta \\ 0 \end{cases}$$



Şekil 3.15 Üçgen üyelik fonksiyonu

$$A(r) = \begin{cases} 1-(a-t)/\gamma \mid a-\gamma \le t \le a \\ 1 \mid a \le t \le b \\ 1-(t-a)/\beta \mid a \le t \le a+\beta \\ 0 \end{cases}$$



Şekil 3.16 Yamuk üyelik fonksiyonu

3.1.4.4 Bulanıklaştırma İşlemi

Genel olarak, klasik küme şeklinde beliren değişim aralıklarının bulanıklaştırılması, bulanık küme, mantık ve sistem işlemleri için gereklidir. Bunun için, bir aralıkta bulunabilecek öğelerin hepsinin, 1' e eşit üyelik derecesine sahip olacak yerde, 0 ile 1 arasında değişik değerlere sahip olması düşünülür. Bazı

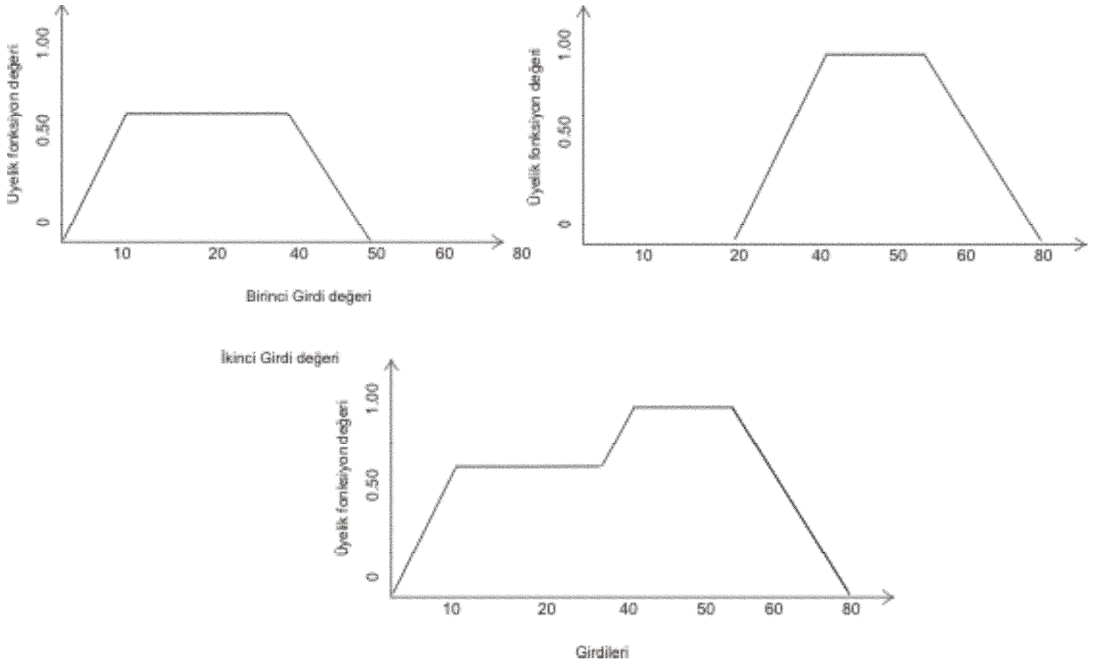
öğelerin belirsizlik içerdikleri kabul edilir. Bu belirsizliğin sayısal olmayan durumlardan kaynaklanması halinde bulanıklıktan söz edilir.

Klasik küme işlemlerinde bir girdi eğer üyelik fonksiyonu söz konusu ise ya 1 değerini yada 0 değerini alır. Bulanık üyelik fonksiyonları ile bulanıklaştırılan girdi değeri artık üye olması dorumunda bile koşulsuz 1 değerini almak yerine üyelik fonksiyonunda aldığı değer ile sisteme girer.

3.1.4.5 Durulaştırma İşlemi

Bulanık sistemlerin çalışabilmesi için girdiler üyelik fonksiyonları ile bulanıklaştırılarak sisteme verilirler sistemin içinde gereken işlemler gerçekleşir ve sonuçlar bulanık kümeler şeklinde dışa verilir. Ancak bu formattaki çıktı mühendislik çalışmalarında kullanılamaz. Bulanık sistem çıktıları gerçek değerler (sayılar) olarak alınmaları lazım. Bu sorunu çözen işlemlerine durulaştırma (defuzzification) işlemleri adı verilir.

Durulaştırma işlem çeşitleri çoktur, ancak burada detaya inmeden hepsinin temelinde yatan fikri açıklayalım.



Şekil 3.17 Durulaştırma işlemin genel temsili

Şekil 3.17 de görülen örnekte 2 tane bulanık girdimiz var. Her iki girdi denk geldikleri üyelik fonksiyonlarda kestikleri noktaların bileşimi ile birer alan oluşturmaktadırlar. Atla görülen grafikte bu alanların bileşimi görülmekte. Bu bileşik alanın değerlendirilmesiyle durulaştırma işlemi gerçekleşir. Örneğin bu alanın ağırlık noktası, ya da entegrali, yada alanı oluşturan doğrunun tepe noktası bizim için gerçek durulaşmış değer olabilir.

3.1.4.6 Temel İşlemler

Ayrışma (Veya) : v(a V b) = en-büyük (v(a), v(b))
Birleşme (Ve) : v(a b) = en-küçük (v(a), v(b))
Olumsuzlama (not) : v(~a) = 1 - v(a)

İki değerli mantıklarda ‘değilleme’, ‘karşıt anlamlı olma’ya karşılık gelir. Bulanık sistemlerde ‘doğru değil’ şeklindeki bir ifade, ‘yanlış’ anlamına gelmeyebilir. Bazı durumlarda ‘doğru değil’i, ‘doğru’ya ‘yanlış’ın olduğundan daha yakın olarak algılamak daha anlamlı olabilir.

Kolayca görülebileceği gibi, değer kümesi, [0, 1] yerine {0, 1} alındığında bu işlemlerden klasik mantıklardaki sonuçlar elde edilecektir.

En-büyük ve en-küçük fonksiyonlarının kullanımının uygunluğu 1973’te Bellman ve Giertz tarafından gösterilmiştir. Fung ve Fu ise 1975’te en-büyük ve en-küçük’ün tek olası işlemler olabileceğini bulmuştur. Matematiksel olarak doğrulanmasının yanında, en-büyük ve en-küçük fonksiyonlarının etkisi, insan nedensellemesinin nasıl olduğunu da ifade ediyor görünmektedir. n tane, derecelendirilmiş doğruluk değerlerine sahip önerme olsun. Her hangi bir kimsenin bunları kullanarak akıl yürüteceğini varsayın. Bunların hepsi ‘veya’ bağlacıyla bağlı olduğunda, doğruluk durumuna olabildiğince yakın olmak isteyecek, ve bu yüzden bu önermeler gurubunun ortak doğruluk değeri olarak, önermeler içinde doğruluk değeri en yüksek olanınkini seçecektir. Bunların hepsi ‘ve’ bağlacıyla bağlı olduğundaysa, en kötü durumu bilmek isteyecektir, bu yüzden bu önermeler gurubunun ortak doğruluk değeri olarak, önermeler içinde doğruluk değeri en düşük olanınkini seçecektir.

Diğer mantık teorilerinde geçerli olan işlemler, bulanık mantık için de geçerlidir. Bulanık mantığı, diğer mantık sistemlerinden ayıran önemli özelliklerden birisi, ‘dışlanmış orta kanunu’ ve ‘çelişmezlik ilkesi’ olarak adlandırılan, ve

v(a V ~a) = Doğru , ve v(a ~a) = Yanlış

şeklinde ifade edilen, diğer mantık sistemleri için oldukça önemli olan, hatta temel kural denebilecek, iki özelliğin, bulanık mantık için geçerli olmamasıdır. Bulanık mantıkta

v(a V ~a) != Doğru , ve v(a ~a) != Yanlış

olur. Burada ‘!=‘, 'eşit değildir' demektir. Bunu sözlü olarak şöyle ifade edebiliriz; bulanık mantıkta ‘bir önerme ya doğrudur ya da yanlıştır’ diyemezsiniz, aynı zamanda ‘bir önerme aynı zamanda hem doğru hem yanlış olamaz’ da diyemezsiniz. Bu, doğruluğun çok değerli oluşundan ve bu çerçevede ‘V ve ‘ bağlaçlarına yüklenen anlamdan kaynaklanmaktadır.

Bulanıklık, bir önermeyle (a), ‘değili’ (~a) arasındaki belirsizlikten kaynaklanır. Eğer v(a)’yı kesin olarak bilmiyorsak, v(~a)’yı de kesin olarak bilmiyoruz demektir. Bu belirsizlik, çelişmezlik ilkesini ihlal edip ‘v(a ~a) != Yanlış’ olmasına, aynı zamanda dışlanmış orta kuralını ihlal edip ‘v(a V ~a) != Doğru’ olmasına yol açar.

3.1.5 Terim Seçme ve Ağırlıklandırma Şemaları

3.1.5.1 Terim Seçme

Metin sınıflama ya da kategorileştirme işlemi kategorisi beli olmayan dokümanın önceden belirlenmiş kategoriliden birine otomatik atama işlemidir. Çevrim içi yani Internet sayesinde ulaşılabilen metin formatında dokümanların ve bilgilerin günden güne artışıyla, verimli doküman endeksleme ve kategorileştirme tekniklerine de ihtiyaç artmakta. Geçmiş senelerde bu problemi çözecek bir dizi çalışma yapılmıştır: regresiyon (gerileme) modelleri (N. Fuhr ve ark, 1991), (Y. Yang ve W.J. Wilbur 1996), en yakın komşuluk sınıflaması (R.H. Creecy ve ark. 1992), (Y.Yang 1994), Beyes olasılık yaklaşımları (K. Tzeras ve S. Hartman 1993) (D.D. Lewis ve M. Ringuette, 1994), karar ağaçları sınıflayıcılar (D.D. Lewis ve M. Ringuette 1994), yapay sinir ağları (E. Wiener 1995).

Metin sınıflama problemlerinin en önemli özellikleri aynı zamanda zorlukları metinin özellik uzayının çok boyutluluğudur. Metnin doğal özellik uzayı tekil terimlerden (kelimeler) den oluşur ve bunların sayısı binlerce olabilirler. Bu sayıdaki özelik bir öğrenme algoritması için idare edilemeyecek bir sayıdır. Burada en uygun çözüm sınıflama işlemi doğruluğundan ödün vermeden özellik uzayının boyut küçültülmesini gerçekleştirmek. Yani terim sayısını azalmak. Aynı zamanda bu işlemi yapacak bir otomatik yöntem geliştirmek olmalıdır çözümün bir başka amacı.

Otomatik özellik seçme metotları ile amaçlanan doküman kümesine göre bilgi taşıyıcı olmayan yani önemi olmayan terimlerin özellik uzayından elemek.

3.1.5.1.1 Terim Seçme Metotları

Bu bölümde, terim seçme metotların kullandıkları beş farklı terim eleme kriteri incelenecektir, doküman frekansı, enformasyon kazancı, müşterek enformasyon, χ^2 istatistiği ve terim gücü.

3.1.5.1.2 Doküman frekansı (DF)

Doküman frekansı bir terimin (kelimenin) bulunduğu doküman sayısıdır (Yiming Yang ve Jan O. Pedersen 1997). Terim eleme için doküman frekans eşik değeri seçilir ve her bir terim için doküman frekansı hesaplandıktan sonra eşik değerini aşamayan terimler elenir. Burada kullanılan basit bir varsayımdır: doküman kümesinde az tekrarlanan bir terim değerli enformasyon taşımamaktadır.

3.1.5.1.3 Enformasyon Kazancı

Makine öğrenmede sık kullanılan bir özellik seçme kıstasıdır (J.R. Quinlan 1984) (Tom Mitchell 1996). Enformasyon kazancı, terimin doküman içinde mevcut olup olmamasını kullanarak kategori tahmininde kullanılacak enformasyon bit sayısını ölçerek hesaplanmaktadır.

$$\begin{aligned} G(t) = & -\sum_{i=1}^{m} P_{\mathrm{r}}(c_i)\log P_{\mathrm{r}}(c_i) \\ & + P_{\mathrm{r}}(t)\sum_{i=1}^{m} P_{\mathrm{r}}(c_i \mid t)\log P_{\mathrm{r}}(c_i \mid t) \\ & + P_{\mathrm{r}}(\vec{t})\sum_{i=1}^{m} P_{\mathrm{r}}(c_i \mid \vec{t})\log P_{\mathrm{r}}(c_i \mid \vec{t}) \end{aligned}$$

Enformasyon kazancına dayanan terim eleme metodu, enformasyon kazancı belirli bir eşik değerden daha düşük terimlerin elenmesi ile gerçekleşmektedir.

3.1.5.1.4 Müşterek enformasyon (Mutual information)

Bu kriter kelime ilişkilerini modelleyen istatistikisel diller ve bunarlın uygulamalarında geniş kullanımı vardır. (R. Fano 1961),(Kenneth Ward Church ve Patrick Hanks 1989). İki sütunlu bir tablo olduğunu farz edelim burada bir sütün t terimleri bir sütün de c kategorileri temsil etsin. O zaman: A sayısı t'nin ve c'nin ayni anda var oldukları sayısı, B t'nin c siz bulunma sayısı, C ise c'nin t siz tekrarlanma sayısı ve N de toplam doküman şayisi olur. T terimi ve c kriteri için müşterek enformasyon kriterinin hesaplanması alttaki formüler ile yapılır.

$$I(t,c) = \log \frac{P_r(t \Lambda c)}{P_r(t) \times P_r(c)}$$

$$I(t,c) \approx \log \frac{A \times N}{(A+C) \times (A+B)}$$

3.1.5.1.5 χ^2 istatistiği (CHI)

Bu ölçü t terim ve c kategori arasındaki bağımsızlığı verir ve alttaki formül ile hesaplanır:

$$\chi^2(t,c) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)}$$

Burada A t nin ve cinin birlikte bulunma sayısı, B t'nin c siz tekrarlanma sayısı, C c nin t siz tekrarlanma sayısı, D t nin ve c nin birlikte bulunmama sayısı ve N de toplam dokuman sayısıdır. Genel bakıldığında χ^2 değeri bir önceki anlatılan kriterden farklı olmadığı düşünülebilir. χ^2 Değeri normaliz edilmiş yani doküman ve terim sayısından etkilenmeyen bir değerdir.

3.1.5.1.6 Terim güçü

Wilbur ve Sirotkin (1992) tarafından Metin madenciliğinde kelime tabanı küçültme için tanıtılan ve sonralarda Yang ve Wilbur (1996) ile metin sınıflama işlemlerine uyarlanan bir terim eleme kriteridir. Bu kriter, terimin benzer dokümanlar içinde tekrarlanma sayısının terim için bir önem taşıdığı varsayımına dayanarak geliştirilmiştir. Doküman kümesi içinde birbirlerine benzeyen dokümanlar tespit edilerek başlanır işleme. Terim gücü ise bir terimin bu iki benzer dokümandan ilkinde koşulu bulunma ihtimaline eşittir,

$$s(t) = P_r(t \in y \mid t \in x).$$

3.1.5.2 Terim Ağırlıklandırma

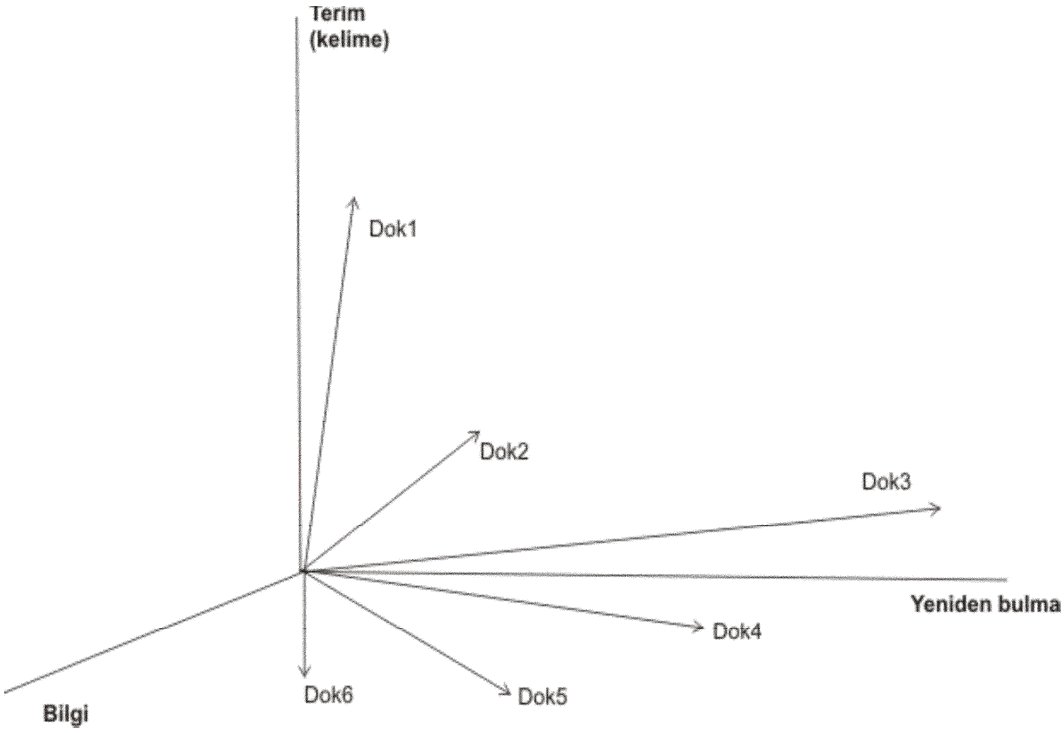
Internet'in dünya çapında gerçekleşen gelişimi ile web site, elektronik posta, haber gurupları, literatür veritabanları gibi dijital formda bulunan doküman sayısında da patlama gerçekleşti. Bu büyük sayıdaki dokümanların analizi için Metin madenciliği teknikleri kullanılır.

Metin madenciliği görevinin en önemli adımlardan biri metni oluşturan kelimelerin bir ağırlıklandırma şeması ile doğru değerlere endekslenmesidir (M.A.Andrade ve A.Valencia 1998). Burada doğru ağırlıklandırma'dan kastedilen dokuman profilini bozmadan sayısal (ağrılık değerleri) bir temsil gerçekleştirmektir. Terim ağırlıklandırma şemaları metin madenciliği prosesinin başarısında çok kritik bir yer alır (E. Leopold ve J.Kinderman 2002),(M .Lan, ve ark. 2005). Bu konuda bir çok araştırma yapılmıştır ve geliştirilen şemalar geniş kullanım bulmuşlardır. En bilindik şemalar: ikili ağırlık, terim frekansı(TF) ve terim frekansı ters dokuman frekansıdır (TFIDF) (G.Salton ve C.Buckley 1988). Bu şemalar yanı sıra tamamen istatistik dağılım analizler temeli şemalar da mevcuttur CTD (B. How ve K.Narayanan 2005), TFRF (M .Lan ve ark. 2005), SQRT-IGFF ve SQRT-IGFS (E.Chisholm ve T. Kolda, 1999) .

Terim ağırlıklandırma şemaların çoğu temelinde istatistiksel veriler ile bir ağırlıklandırma gerçekleştirmektedir. Terimin bulunduğu dokumandaki frekansı TF olsun ve bu terimin bulunduğu dokuman dışında kaç tane dokümanda tekrarlandığı bilgisi de DF olsun. TF ve DF hemen hemen bütün şemaların temelinde yatan bilgilerdir, değişen sadece bunları kullanan formülerdir.

3.1.5.2.1 İlgili Araştırmalar

Son yılarda sınıflama, kümeleme, kategorileştirme gibi metin madenciliği metotlarında on işlem olarak kullanılan birçok terim ağırlıklandırma şeması geliştirmekte. Bu şemaların aralarındaki görünürde büyük farklara rağmen amaçları; dokümanı, terimler (kelime) uzayında bir vektör ile temsil edebilmektir. şekil 3.18.



Şekil 3.18 Kelime uzayında doküman vektörlerin gösterimi

3.1.5.2.2 İkili ağırlık metodu

Kelimenin bir dokümanda bulunması (1) yada bulunmamasını temel alan bir şema olan İkili ağırlık şeması alta görünen formüle göre hesaplanır :

$$w_{i,j} \begin{cases} 1 \text{ eğer i terimi j dokümanında bulunuyor ise} \\ 0 \text{ diğer durumlarda} \end{cases}$$

3.1.5.2.3 Terim frekans (TF) metodu

Bu metot ile bir terimin ağırlığı çok basit bir mantıkla i teriminin j dokümanında tekrarlanma sayısına eşitlenir.

$$w_{i,j} = tf_{i,j}$$

3.1.5.2.4 Klasik TFIDF metodu

TFIDF kısaltması; terim frekansı, ters doküman frekansı anlamında kullanılır. Bu şema iki önemli elemanı bileştirmekte: t terimin d dokümanında tekrarlanma sayısı yani dokümana göre yerel bir ağırlık verisini (terim frekansı); ve aynı terimin küresel bir ağırlık değeri olan doküman setindeki dokümanların içindeki dağılım frekansı (doküman frekansı). Bir terimin beli bir dokümana göre önemi yani ağırlık değeri TF in yüksekliğine ve DF in düşüklüğüne bağlıdır. Tf in yüksek çıkması terimin o dokumana göre önemli olduğunu, DF in düşük çıkması ise bu terimin diğer dokümanlarda az tekrarlandığını yani sadece söz konusu doküman için önemli olduğunu ifade eder.

Otomatik sınıflama, kümeleme ve benzeri metin madenciliği işlemlerinde genellikle önişlem olarak TFIDF ağırlıklandırma şeması kullanılmaktadır. TFIDF in bir ağırlıklandırma şemasında kullanılması alttaki formül ile gerçekleşir:

$$w_{i,j} = tf_{i,j} \times \log(\frac{N}{df_i}).$$

Bu formülde tf, i terimin j dokümanındaki frekansı, N toplam doküman sayısı ve df de i terimini içeren doküman sayısı. (G.Salton ve C.Buckley, 1988) (Y. Liu ve ark. 2004) …

3.1.5.2.5 CDT metodu

TFIDF ağırlıklandırma şemasını geliştirerek yeni bir şema ortaya koyan araştırma (B. How ve K.Narayanan, 2006) da sunulmuştur. TFIDF in tek

doküman üstünden değil de bir kategoriye uygulanması ile ortaya yeni bir ağırlıklandırma şeması çıkmıştır CTF, kategori-temeli frekans.

$$CTD_{t,c} = CTF_{t,c} \times IDF_{t,c} \times ICF_t$$

$$ICF = \log \frac{|C|}{CF_t}$$

3.1.5.2.6 TFRF metodu

Yerel terim frekansını ve yeni tanımlanan ilgi frekans bilgisini yeni bir şemada birleştiren araştırma (M.Lan ve ark. 2006) bu yeni şemaya TFRF adını vermiştir. Bu şema ile ağırlık hesaplanması için

$$w_{i,j} = tf_{i,j} \times rf$$

Formülü kullanılmakta. Burada tf bilindiği üzere i teriminin j dokümanındaki frekansıdır. Rf kısmı ise alta bulunan formül ile tanımlanır:

$$rf = \log(1 + \frac{n_i}{n_{i-}})$$

burada n_i, i terimin bulunduğu doküman sayısı, ve n_{i-} de i terimini içeren ancak kümede bulunmayan doküman sayısıdır.

3.1.5.3 Var Olan Metotlar Üstüne Bir Eleştiri

Yukarıda anlatılan bütün metotların eksikleri var. En büyük eksikleri basit istatistiksel yöntemleri temel almaları ve genellikle sadece bir dokümanla ilgili olmaları. Bu tez çalışmamda terimlerin ağırlık hesaplanması için yeni bir şema öneriyorum, ilerleyen bölümlerde bu şemanın detayları verilecek.

3.2 – Metot

3.2.1 Edebiyat eserlerini web verilerine dayanarak sınıflandırma

Nesneleri özelliklerine göre ayrı sınıflara ayırma işlemine sınıflandırma denir. Sınıflandırma algoritmaları, sınıflanacak nesnelerin özelliklerini öğrenme esnasında inceleyerek bir sınıflandırma gerçekleştirirler. Sınıflanacak nesne özelliklerinin değerlerini içeren veri kaynaklarına eğitim örnekleri denilmektedir. Web tabanlı sınıflandırma modelleri normal sınıflandırma uygulamalarından bu noktada ayrılmaktadırlar. Web tabanlı sınıflandırma modelleri sınıflanacak nesnelerin özellik değerlerinin web sitelerinden elde edilebileceğini var sayarak geliştirilmişlerdir.

Web tabanlı sınıflandırma uygulamalarından biri müzik sanatçılarının yaptıkları müzik tarzlarının internet aracılığı ile öğrenilmesi ve sınıflanması uygulamasıdır (Peter Knees ve ark. 2003). Bu problem, yıllarca sanatçıların müzik parçalarının, ses işleme modelleri ve filtreleri ile incelenerek, çözülmeye çalışılmıştır ve ulaşılan doğruluk %80leri geçmemiştir. Yukarıda atıfta bulunulan makalede ise %90 lara varan bir doğrulukla müzik sanatçıları yaptıkları müzik tarzlarına göre sınıflanmışlardır. Bu modelin başarısı biraz da eski yöntemlerle toplanan bilgilerin kullanılmasa bağlıdır. Başlangıç parametrelerin değerleri bu bilgilere göre seçilmişlerdir.

Veri toplama işlemi bu uygulamaları diğer benzer uygulamalardan farklı kılan işlemdir. Web tabanlı veri ile sınıflandırma, denilmesinin sebebi de burada ortaya çıkmaktadır. Sınıflanacak nesneler hakkında veri web sitelerinden elde edilmektedir. Burada web sitelerinin bulunması değişik yöntemler ile gerçekleştirilmektedir, ancak en geneli web arama motorlarının kullanılmasıdır.

Uygulamalar birbirlerinden kullandıkları arama motorları, bunlara gönderdikleri arama sorgusu ve gelen sonuçları değerlendirme yöntemlerine göre farklılık göstermektedirler. Örneğin, yukarıda atıfta bulunulan müzik sanatçılarını sınıflayan uygulamada arama motoru olarak google ve yahoo, arama sorgusu

olarak “sanatçı + müzik tarzı” , gelen sonuçların da değerlendirilmesi için ilk ellisi seçilmiştir. Bu değerlerin seçimi söz konusu çalışmada önceki araştırmalara atıfta bulunarak açıklanmıştır. Ancak temel teşkil edecek araştırmaların olmaması durumunda buna benzer modellerde sıkıntılar yaşanabileceği ortadadır. Bu problemin çözümü için ileriki bölümlerde yeni bir model tanıtılmaktadır.

Web tabanlı veri ile sınıflandırma web sitelerinden bilgi keşfetmektir. Web siteleri belirli bir yapı göstermedikleri ve genelde resim, video, metin vs den oluştukları için gürültüye ve istenmeyen verilere sahiptirler. Web tabanlı sınıflandırma yöntemleri ilgilendikleri veriler dışındaki verileri özellikle html kodları, bannerler, reklamlar vs gürültü olarak değerlendirmekteler.

Web sitelerindeki gürültü temizliğini konu alan bir çok çalışma mevcutur. Bazıları html sayfalarını html ağaçlar ile temsil ederek gereksiz verileri eleme yolarını geliştirmişlerdir (Lan Yi ve Bing Liu 2003). Bazı çalışmalarda ise algoritmik bir yaklaşım ile adım adım eleme yaparak sadece ilgilenilen verilere ulaşılmaktadır.
Müzik sanatçılarını sınıflayan çalışmada veri temizliği adına html kodlar, bağlaçlar (stop words), noktalama işlemleri, problemi ilgilendirmeyen kelimeler ve metin dışındaki bütün verilerin eleme işlemi yapılmıştır.

Web tabanlı veri ile sınıflandırma çalışmaları, açısından önişlem süreci veri madenciliğine hazırlıktır. Metin madenciliği açısından bu hazırlık web sitelerin bir temsil modeli ile madenciliğe uygun hale getirmektir. Müzik sanatçılarını sınıflayan uygulamada kelime vektör temsili kullanılmıştır ve benzer çalışmalarda genelde bu model kullanılmaktadır. Doğal dil işleme yöntemleri tavsiye edilen temsil modelleridirler. Kelime vektör modellerini geliştirmeyi amaç edinmiş makaleler de mevcuttur (Fabrizio Sebastiani ve Franca Debole 2004).

Örnek teşkil etmesi açısından atıfta bulunulan çalışmanın içeriğine değinelim: “Müzik sanatçılarının web tabanlı bilgiler ile sınıflanması” (Peter Knees1 ve ark. 2003) isimli makalede müzik sanatçıları google arama motoru ile sorgulanmışlardır. Sorgu, sanatçı ismi + müzik tarzı olarak seçilmiştir. Gelen

sonuçlar genelde sanatçıların fan siteleridir. Gelen sonuçlardan ilk ellisi ele alınmıştır. Sonuç siteleri html taglardan ve bağlaçlardan arındırılmıştır. Her a sanatçısına bağlı dokümanı için, her a sanatçısı ve her t terimi için t terimin tekrarlanma sayısı (frekansı) tfta hesaplanmıştır. Ayrıca t terimin geçtiği doküman sayısı, (frekansı) dfta hesaplamıştır. Bu hesaplanan terim ve doküman frekansları sayesinde sınıflandırma için kullanılacak terimlere birer ağırlık değeri verilmektedir.

$$w_{ta} = \begin{cases} (1+\log_2 \mathrm{tf}_{\mathrm{ta}})\log_2 \dfrac{N}{df_t} \\ 0, \end{cases}$$

Terim frekansı $\mathrm{tf}_{\mathrm{ta}}$ sıfır olduğunda, ağırlık değeri W_{ta} sıfır olarak seçilmektedir. 200 sanatçı için yapılmış aramalar sonucunda 200,000 farklı terim elde edilmiştir. Bu sayıyı azaltmak için ek bir terim eleme işlemi uygulanmıştır. İstatistikte χ^2 kare olarak bilinen dağılım kullanılarak terimler sayısı %70 kadar azaltılmıştır. χ^2 kare testi sayesinde t terimin c sınıfıyla bağımsızlığı test edilmektedir.

$$\chi^2_{tc} = \frac{N(AD - BC)}{(A+B)(A+C)(B+D)(C+D)}$$

Formülde c kategorisindeki t terimini içeren doküman sayısı A , c kategorisinde olmayan t yi içeren doküman sayısı B , c de olan t yi içermeyen C , c de olmayan ve t yi içermeyen doküman sayısı da D ile gösterilmektedir, N ise arama motorundan elde edilen toplam doküman sayısı ile gösterilmektedir. Her terim için χ^2 kare değeri hesaplandıktan sonra her c kategorisi için en yüksek test değerine sahip ilk 100 terim seçilmiştir. Bütün kategoriler (müzik tarzı sınıfı) için seçilmiş olan yüzer tane terim bir araya getirilmiştir.

Terimler seçimi sonucunda her sanatçı terimlerin ağırlık değerlerine göre bir kelime (terim) vektörü ile temsil edilmişlerdir. Burada ağırlık değerleri kosinüs normalizasyonu ile (-1,1) normaliz edilmişlerdir.

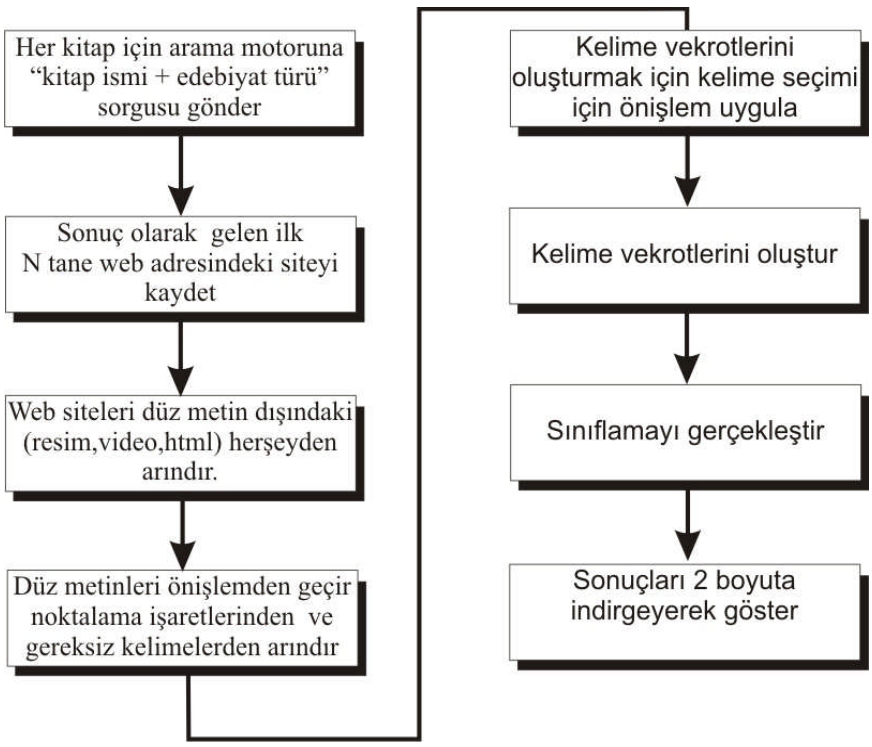
Sanatçıları sınıflandırmak için destek vektör makineleri kullanılmıştır. Sonuçların gösterimi de SOM algoritması ile yapılmıştır. 14 farklı sınıf yani müzik tarzı (pop,rok vs) için 200 sanatçı sorgulanmıştır ve %90 lara varan doğrulukla sınıflandırma gerçekleştirilmiştir

3.2.1.1 Sınıflama modeli ve algoritma

Bu kısımda tezin deneysel çalışması anlatılacak. Tezin amacı edebiyat eserlerinin web verilerine dayanarak sınıflanmasıdır. Önceki bölümlerde teorik bilgiler verildiğinden metnin devamında gerektiğinde bunarla atıfta bulunulacak.

3.2.1.2 Modelin Genel görünümü ve Akış şemaları

Alta görülen akış şemasındaki her adım ilerleyen alt bölümlerde ayrıntılı anlatılacaklar, bu bölümde sadece genel bir bakış ile model tanıtılacaktır.



Şekil 3.19 Çalışmayı temsil eden Akış diyagramı

Bu modelin amacı internette ortak kullanıma açık genel web sitelerindeki bilgilerden yararlanılarak bir sınıflandırma gerçekleştirmektir. Sınıflandırılacak nesneler edebiyat eserleri, sınıflar ise edebiyat türleridir (Roman, Öykü, Masal, Tiyatro, Deneme, bilim kurgu, vs).

Şekil 3.19 de görünen akış diyagramından da anlaşılacağı gibi model önceden belirlenmiş parametrelere göre çalışarak (öğrenerek) bir sınıflandırma problemini çözecektir. Optimal sonuçları üretmesi için uygun önişlem ve madencilik teknikleri yanı sıra parametrelerin doğru seçilmesini de gerektirir. İnternetteki verilerden yararlanarak sınıflandırma gerçekleştiren benzer çalışmalarda bu parametrelin başlangıç değerleri için bazı öneriler bulunabilir ancak bu çalışmaya temel alınabilecek bir çalışma bulanamadığı için bu başlangıç değerleri ancak sistemin gerçekleştirilmesi esnasında tespit edilmiştir.

3.2.1.3 Sınıflanacak Verilerin Elde Edilmesi

Sınıflandırma işlemini gerçekleştirmek için bir veri tabanı (eğitim örnekleri) gereklidir. Sınıflandırma işlemi bir görerek öğrenme yöntemidir. Yada diğer tanımı ile örneklerden öğrenme işlemidir. Dolayısıyla örnekleri temsil edecek veri tabanın oluşturulması için internet kullanılacaktır. Sınıflandırılacak nesneler (kitaplar) hakkında bilgi tabanı web siteleridir. Bu sitelere ulaşmanın yolu Web içerik madenciliği ve Bilgi keşfi yöntemleridir. Ancak günümüzde yaygın olarak kullanılan akıllı arama motorları zaten bu işlemi yaptıkları için onların kullanılması daha mantıklıdır. Bu yüzden istenilen web sitelerinin elde edilmesi arama motorları ile gerçekleştirilecektir. Uygun bir sorgu ile internette kitabın ismi ve türü hakkında bilginin birlikte geçtiği siteler arama motoru ile elde edilmektedir.

Örnek sorgu => “romeo and juliet” + “book” + “review”

İnternet arama motorlarının sonuçları sıralama işlemini birçok kıstasa bağlı olarak yapmaktadırlar. Ancak sıralamada en öne alınan siteler genelde başka sitelerden kendilerine link alan sitelerdir. Bu bizim sınıflandırma işlemi için bir

şey ifade etmez çünkü bizi ilgilendiren sonuç olarak gelen sitelerin içinde sınıflandırmaya yarayacak bilgilerin olmasıdır.

3.2.1.4 Sınıflama için edebiyat eserlerin ve sınıfların seçilmesi

Sınıflama modelin geliştirilmesi için ilk önce sınıflar ve bu sınıflara ait eğitim ve test örnekleri tanımlanmalıdır. Bu çalışmanın amacı edebiyat eserlerinin yani kitapların sınıflanması olduğu için ilk aşama bir test için edebiyat sınıfların ve eserlerin belirlenmesidir. Birinci test altı sınıf ve her sınıfa ait sekizer kitaptan oluşmaktadır.

Alta görünen listeler wikkipendia.com Internet ansiklopedisi ve bir çok başka kaynaktan yararlanılarak elde edilmiştir. Kitapların sınıflaması hakkında bir çok kaynak literatür bulunmasına rağmen bu çalamaya uygun şeklinin bulunması zahmetli bir süreçti. Bu listelerin oluşturulması için seçilen çalışmalar İngilizce konuşma bölgelerinde yapılan ve İngilizce dilinde yazılmış eserlere verilen ödülleri kapsamaktadır. Ödülleri veren kurumlar da İngiliz dili bölgelerindendirler (Amerika, İngiltere , vs). Bu yöntemin seçilme sebebi bu eserler hakkında web sitelerinde daha kolay verimli bilgilere ulaşma fikrine dayanır.

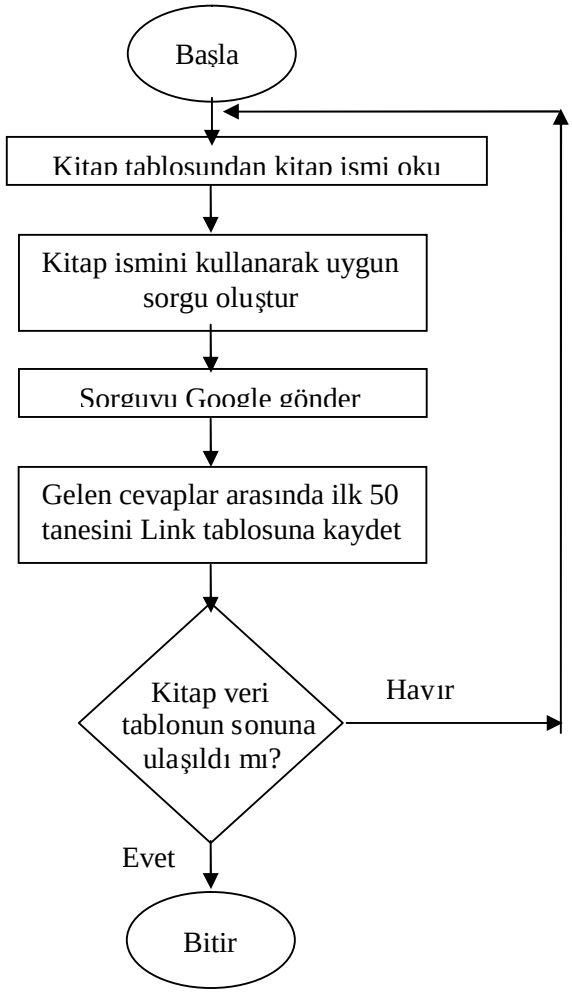
Edebiyat eserleri için seçilmis sınıflar :

1. Çocuk kitapları (Children's Boks)
2. Esrarengiz (Mystery)
3. Bilim kurgu – Fantezi (Sci-fi/Fantasy)
4. Genel edebiyat (General Nonfiction)
5. Edebiyat Kurgu (Literary Fiction)
6. Şiir (Poetry)

Her bir sınıf için seçilen kitaplar son senelerde büyük edebiyat ödülleri kazanmış eserlerdir. Bunların tam listesini ek-A da görülebilir.
İkinci ve daha büyük bir test de bilgisayar kitapların sınıflandığı çalışmadır. Bu test için seçilen sınıflar ve örnekler ek-B de mevcutlar.

3.2.1.5 Edebiyat eserleri hakkında bilgi içeren web sitelerin elde edilmesi

Edebiyat eserleri isimleri tespit edildiğinde, bunlar hakkında bilgi toplayıp sınıflamaya hazır hale getirmek gerekmektedir. Veri tabanın uygun tablosunda kitap isimleri kayıtlı olduğunu farz edersek o zaman sınıflamada kullanılan yazılım bu tablodan okuduğu her kitap ismi için uygun bir sorgu gerçekleştirerek bir arama motoruna gönderir. Arama motorundan cevap olarak gelen ilk 50 link kitap ID sine göre link tablosuna kitap id ve gelen linkler olarak kaydedilir.

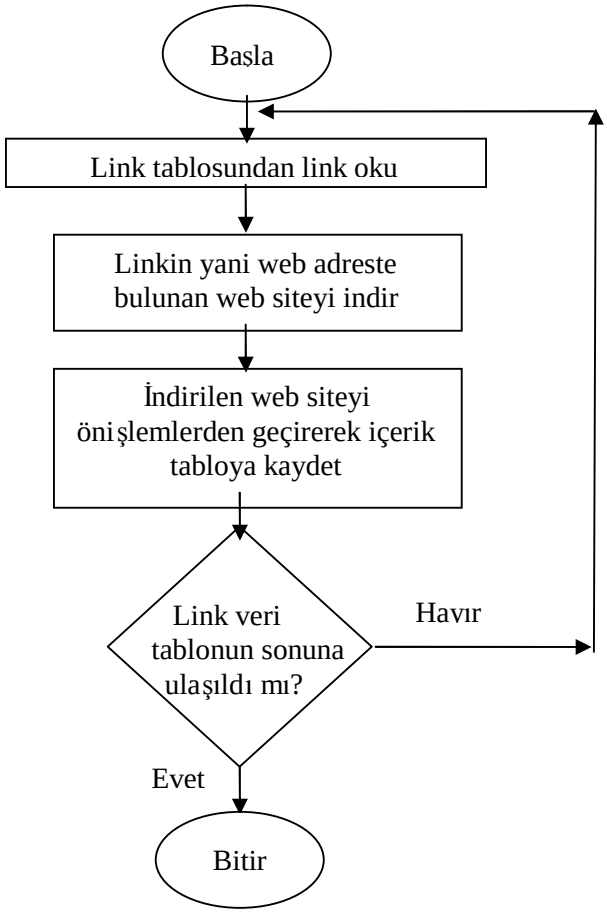


Şekil 3.20 Web site adreslerin elde edileme algoritması

Algoritmada uygun arama sorgusu olarak kastedilen internet arama motoruna gönderilecek arama cümlesinin oluşturulmasıdır. Arama motorları anahtar kelimelerine göre çalışırlar ve birçok denemeden sonra bu çalışma için en

uygun arama cümlesi **,“ kitap ismi ” + book + review -amazon.com –pdf** , olduğu görülmüştür. Burada ilk parametre kitap ismidir, ve program tarafından veri tabanından okunarak elde edilmektedir. Ayrıca eklenen book ve review kelimeleri ile sonuçların filtrelenmesi amaçlanmıştır. Eksi amazon.com ve pdf parametreleri ile amazon.com web sitesi ve alt domainler ve pdf tipinde dosyaları sonuçlar arasına getirilmesi engellenmiştir. Amazon sitesinden sınıflamayı yanıltan ve devamlı tekrarlanan cümle yapıları geldiğinden, linkleri sonuçlardan elenmiştir.

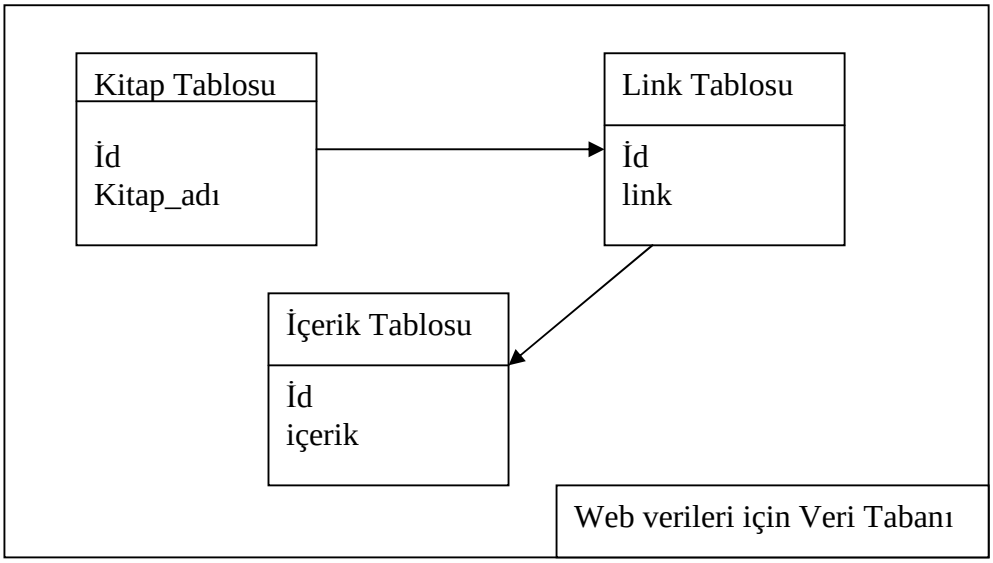
Her bir kitap için 50 web site adresi elde edildikten ve bu adreslerin link tablosuna uygun kitap referansı ile kaydedildikten sonra artık web siteleri yerel veri tabanına kaydedilmeleri gerekmektedir. Bu işlemin özetlenmiş hali şekil 3.21 de görülmekte.



Şekil 3.21 Web sitelerin elde edileme algoritması

4.2.4 Yardımcı veri tabanı tasarımı

Deneylerin gerçekleştirildiği yazılım bir veri tabanı bağlantısı sayesinde gereken verileri internetten toplamaktadır. Verilerin toplandığı alanda bir MS Access veri tabanıdır. Veri tabanı tasarımı alta görülmektedir.



Veri tabanı üç tablodan oluşmaktadır. Kitap tablosu, Link tablosu ve İçerik tablosu. Kitap tablosu indeks alanı olan id, ve kitap isimlerini içeren Kitap_adı alanından oluşmaktadır. Bu tablo sınıflama modelin gerçekleşmesi için gerek eğitim ve test örneklerini oluşturacağımız kitap örneklerini kaydetmek için kullanılmaktadır.

Kitap tablosuna kitap isimleri kaydedilmiştir, kullanılan index ise iki dijitlik bir numaradır ilk numara 0-5 arası yazılır ve sınıfı temsil eder ikinci dijit de 0-7 arası seçilir ve kitap sayısını temsil eder. Tablonun birkaç kaydı alttaki şekilde görülmekte.

id	Kitap_ismi
00	The Hello, Goodbye Window
01	Kitten's First Full Moon

Örnekte 00 sıfırcı sınıfa yani çocuk kitaplarına ait sıfır numarali eserdir. 63 koduna sahip kitap Şiir sınıfına ait 4 numaralı kitaptır.

İkinci tablo Link tablosu kitap tablosuyla bağlıdır. Tablonun amacı her bir kitap için arama motorundan elde edilecek linkleri kaydetmektir. Üçüncü tablo da içerik tablosudur, ve yardımcı programın link tablosunda okuduğu linklerde bulunan web sitelerin içeriğinin indirip kaydedildiği tablodur.

Deneylerin yapılması için bu çalışma kapsamında geliştirilen yazılım veri toplama işlemlerini yukarıdaki sıraya göre yapmaktadır. Girdi olarak okunan kitap isimlerine göre google web servislerini kullanarak kitapla ilgili web siteler veri tabana indirilirler. Bu işlemler doğal olarak uzun sürmekteler.
İlk teste altı sınıf ve her bir sınıfa ait sekiz kitapla, 46 elemanlı bir set oluşmuştur. 46 kitap için 50 şer web site yani 2800 tane web dokümanı tespit edilip veri tabanına kaydedilmesi gereklidir. İkinci teste ise yerel veri tabanına 10 sınıf, her sinifa ait 14 kitap ve toplam 10x14x50 = 7000 web dokümanı kaydedilmekte.

Burada önemli bir nokta web dokümanların yazılım tarafından sadece HTML kodları olarak veri tabanına transfer edilmeleridir. İlk testin sonucunda oluşan veri tabanı boyutu yaklaşık 70 MB dır.

3.2.1.7 Gürültü Temizliği ve Önişlemler

İnternet aracılığı ile elde edilen sonuçlar birer web site oldukları için bir sonraki adımda bu web siteler bir önişlemden geçmektedirler.

Web siteler sadece düz metinden oluşmazlar, içlerinde resimler, video görüntüler, reklâmlar, banerler vs. bulunmaktadır. Metin dışındaki veriler bizim için bu çalışma acısında gürültü sayılmaktadır. Gürültüden siteyi arındırma işlemi veri temizleme işlemidir ve birçok yöntem geliştirilmiştir.

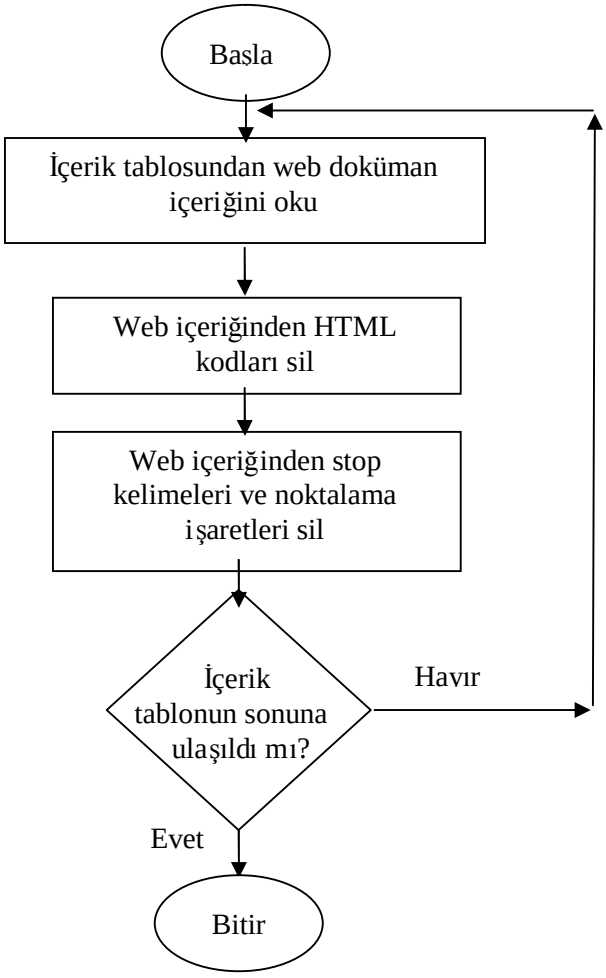
Genel olarak gürültü temizliği iki işlemden oluşur:

1) metin dışındaki verilerin elenmesi
2) ve HTML kodlarından arındırılması.

Gürültü temizliği gerçekleştirdikten sonra kalan saf metnin içinde de istenilmeyen veriler vardır. Bunlar noktalama işaretleri (nokta, virgül, soru işareti), İngilizcede stop words olarak tanımlanan kelimeler (ve,veya, evet hayır, vs), konuyu ilgilendirmeyen kelimeler, vs.

Bu çalışmada yapılan önişleme, noktalama işaretlerinin ve stop wordların çıkarılmasından oluşmaktadır. Herhangi bir işlemde veri içindeki gürültü, sonuçların beklenenin çok altında olmasına sebep olmaktadır. Metin madenciliği yapısal olmayan verilerle çalıştığı için gürültüye ve istenmeyen sonuçlara açıktır. Metin madenciliği tekniklerinin başarısı gürültünün azaltılması oranında değişmektedir. Yani daha iyi önişleme metodu sonuçlarının daha iyi olmasına sebep olmaktadır.

Yazılımda önceki adımlarda elde edilen binlerce web site bu önişlemlere tabi tutulurlar. İlk önce HTML formatında olan doküman HTML Parser adını verdiğim nesne tarafından HTML kodlarından arındırılır. Burada html kodların tespit edilmesi için regular expression isimli string işleme metotları kullanılmaktadır. Elde edilen düz metinden de ikinci aşamada stop kelimeler ve noktalama işaretleri çıkartılır. Geriye kalan kelime seti de içerik veri tablosundaki eski yerine yazılır. Şekil 3.22 de bu işlemin algoritması görülmekte.



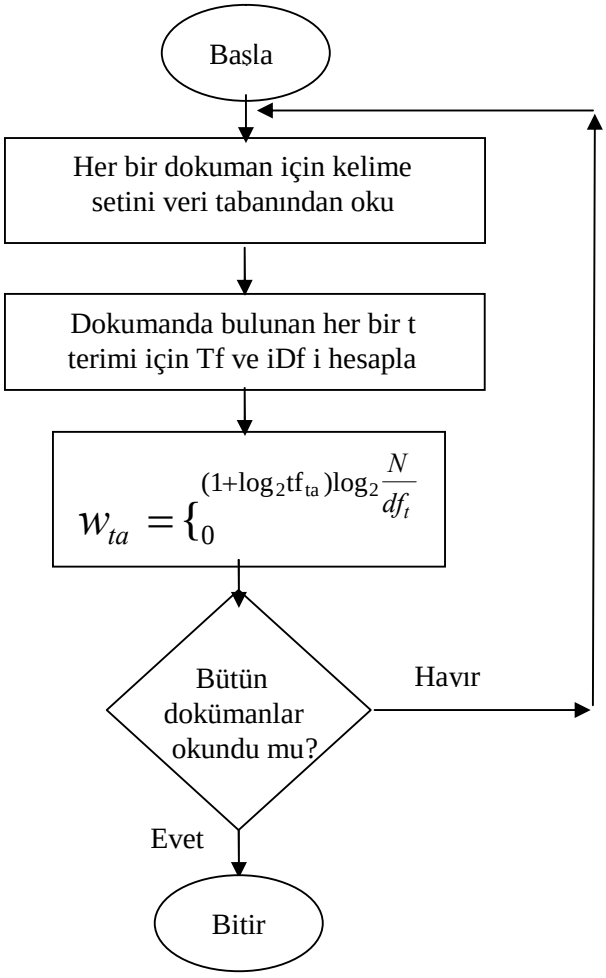
Şekil 3.22 Web sitelerin madenciliğe uygun hale getirilmesi algoritması

3.2.1.8 Kelime Vektörün Oluşturulması

İnternetten, arama motoru ile elde edilen web siteleri kaydedilip, gürültüden arındırılıp önişlemlerden geçirildikten sonra elimizde düz metin kalmaktadır. Ancak bu haliyle metin henüz madencilik için uygun değildir. Bir sonraki aşamada metin bir temsil modeli ile madencilik için uygun hale getirilmelidir. En basit ve sıkça kullanılan metin temsil modeli kelime-vektör temsilidir (başka bir ismiyle kelime-çuvalı bag-of-words representation) . Fikir gayet basittir metinden kelimeler sıra ve yapı gözetmeksizin bir çuvala atarcasına toplanmaktadır. Her bir kelime için kelime-vektöründe o kelimenin metin içinde tekrarlanma sayısı tutulur. Bu model en basit olmasına rağmen uzun metinlerde en iyi temsil yöntemidir. Kelime vektörü modelinin yetersiz kaldığı problemlerde model geliştirilmiştir; mesela tek kelime yerine kelimeler ikişer ya da üçer olarak gruplanmış ve o şekilde kelime-vektörüne yerleştirilmişlerdir. Bazı uygulamalarda bu temsilerin ötesinde metnin cümle yapısına göre temsil modelleri geliştirilmiştir. Çalışmamızda kelime-vektörü temsili yeterli olacaktır. Yetmemesi durumunda hybrid bir yönteme başvurulacak.

Kelime- vektörü temsil modelinin tez çalışmasına uygulanmasını genel hatlarıyla açıklayalım. Her bir kitap için web sayfalarından elde edilen düz metinler birleştirildikten sonra içlerinde geçen kelime sayısı binlerce hatta on binlerce olabilir. Bu sayıyı azaltmanın yöntemleri mevcuttur. Genelde 3 harften kısa kelimeler atılır, yada konuyla ilgi olup olmadıkları incelenir. Bu çalışmadaki kelimeler dokümanlar içinde bulunma frekanslarına göre sıralanarak, en yüksek frekansa sahip ilk n tanesi seçilerek kelime-vektör temsili oluşturulmaktadır.

Sınıflama işlemine başlamadan önce son adimdir kelime vektör oluşturulması. Teorik esasları önceki bölümlerde verilen bu işlem burada sadece uygulama açısından incelenecek. Önceki adımlardan geriye her bir kitap için kalan birer kelime seti, bu adımda uygun işlemler ile ağırlıkları hesaplanıp birer sayısal değere dönüştürülmeleri gerek. Şekil 3.22 de bu adımda uygun yazılım tarafından gerçekleştirilen algoritma görülmekte



Şekil 3.22 Terim seçme ve ağırlıklandırma şeması algoritması

Burada ağirlık değeri hesabında kullanılan yöntem TFIDF olarak bilinen ve yaygın kullanılan yöntemdir.

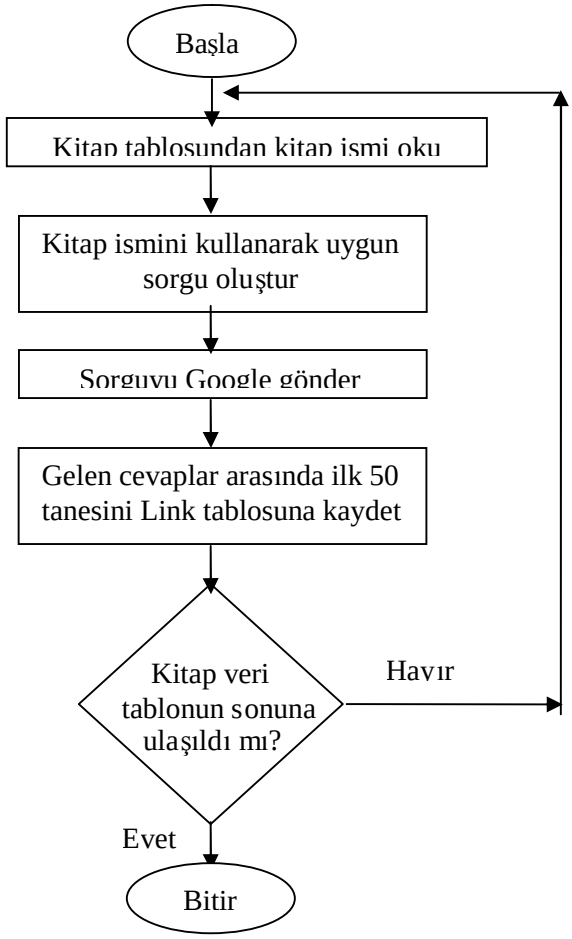
$$w_{ta} = \{_{0,}^{(1+\log_2 tf_{ta})\log_2 \frac{N}{df_t}}$$

Bu çalışmada terim seçme şeması olarak sunulan Bulanık terim seçme yönteminden elde edilen çıkış değeri yukarıdaki formülde dft yerine konulabilir. Bu da terim ağırlıklandırma da bir yeniliktir, ancak verimi hakkında henüz yeteri kadar bilgi edinilmemiştir.

Daha önce de bir çok yerde vurgulandığı gibi dokümanların kelime seti oluşturulduğunda çok sayıda kelime kalmaktadır. Bu çalışmada ilk aşama kelime

yani terim filtreleme olarak 4 harften kısa kelimelerin ve harf ile başlamayan kelimelerin elenmesidir. Ancak bu işlem de kelime sayısının istenilen sınırlarda olmasını sağlayamamaktadır. Teorik bilgiler arasında bu durumlar için geliştirilmiş terim eleme şemaları anlatılmaktadır. Bu çalışmada terim elem için çi kare metodu ve bu çalışma kapsamında önerilen yeni Bulanık terim seçme şeması kullanılmış ve sonuçları test edilmiştir. Önerilen yeni şema daha sonraki bölümde daha ayrıntılı incelenecek. Çi kare testi teorik bilgilerde ayrıntılı açıklanmıştır. Uygulamada kullanılma şekli ise her bir terim için çi ağırlıklandırma ile bir değer üretilir ve en yüksek değere sahip ilk 100 terim dokümanı tarif edecek terimler olarak seçilmekteler.

Terim seçme işlemini gerçekleştiren modülün kullandığı algoritma şekil 3.23 da görülmekte. Bu algoritmada kullanılacak terim seçme şeması belirlenmeden genel bi algoritma tanıtılmakta. Yazılımda seçilen şema parametre olarak verilir ve seçime göre bir çıkış elde edilmektedir.



Şekil 3.23 Terim seçme ve ağırlıklandırma şeması algoritması

Bu aşamada yazılım bir önceki adımdan seçilmiş terim eleme yönteminden gelen terimler ve bunarlın ağırlık değerlerinden 46 kitap için sayısal değerler içeren 46 tane vektör oluşturmakta ve bunları bir dosyaya kaydetmektedir. Bu dosya sonradan Matlab yazılımına aktarılarak sınıflama modeli kurulmaktadır.

3.2.1.9 Sınıflama İşlemi

Kelime vektörleri oluşturulduktan sonra metin madenciliği teknikleri uygulanarak çalışmanın amacı yerine getirilmektedir. Sınıflanacak nesneler (kitaplar) sınıflara (edebiyat türleri) otomatik atama işlemi bir metin madencilik sınıflama algoritması olan destek vektör makineler ile yapılamaktadır. Bu yöntem sınıfların önceden bilinmesi durumunda kullanıldığından ve özellikle de çok boyutlu yapılarla iyi performans gösteren bir sınıflandırma yöntemi olduğu için seçilmiştir.

Destek vektör makineleri bir öğrenme metodudur. Çekirdek tabanlı doğrusal olmayan sınıflandırıcıların sinyal işleme, yapay öğrenme ve VM alanındaki pratik problemlerde iyi sonuçlar verdiği bulunmuştur. Matlab yazılımı için geliştirilmiş açık kodlu ve birçok makalede en iyi SVM implementasyonu olarak tavsiye edilen LİBSVM toolbox u sınıflama işlemlerinde kullanılmaktadır. (http://www.csie.ntu.edu.tw/˜cjlin/libsvm).

Önceki aşamalarda kıtaların birer ağırlık değerler vektörleri ile temsileri yapılmış ve bu veriler bir dosyaya kaydedilmiştir. Artık bu verilerin bir kısmının eğitim verisi olarak kullanılarak LİBSVM aracı ile bir sınıflama modeli oluşturulur. Sonunda da model rasgele seçilmiş kelime vektörleri ile test edilmiştir.

4.7 Deneyler için geliştirilen ve kullanılan yazılımlar ve araçlar.

Edebiyat eserlerinin web verilerine dayanarak sınıflandırma süreci üç temel aşamadan oluşmaktadır:

1. Ön hazırlık aşaması, analiz surecini kapsamaktadır. Sadece bu aşama otomatik değildir ve yazılım kullanılmadan uygulanmaktadır.
2. Sınıflama için verilerin Internet ten toplanması, önişlemler ile sınıflamaya uygun hale getirilmesi aşaması, tamamen otomatik ve özel olarak bu deneyler için geliştirilmiş yazılım ile yapılamaktadır.
3. Terim seçme ve ağırlıklandırma aşaması, tamamen otomatik ve aynı yazılım ile gerçekleştirilmektedir.
4. Sınıflama modelin oluşturulması ve test edilmesi, Matlab ve LİBSVM isimli sınıflama aracı ile otomatik yapılmaktadır.

Matlab yazılımı için geliştirilmiş LİBSVM bir SVM sınıflama aracıdır. Bu araç seti verilerin bir dosyadan Matlaba aktarılmasını, aktarılan verilerden sınıflama modelinin kurulmasını ve test edilmesini sağlamaktadır.

Sınıflamanın gerçekleştirilmesi için gereken verilerin internete tespit edilip, yerel veri tabana aktarılması, gereken önişlemlerin uygulanması, metinin terim seçme ve eleme şemalarının uygulanması işlemlerini bir yazılım aracılığı ile yapılmaktadır. Yazılım deneysel çalışmalar için, Microsoft un. NET 2.0 ve VS2005 yazılım geliştirme ortamında geliştirilmiştir. Yukarıda anlatılan çalışma sürecin her adımı için uygun sınıflar yazılarak tam modüler ve para metrik bir yazılım geliştirilmiştir. Web den gereken site adreslerinin tespit edilmesi için google web servisleri kullanılmaktadır. Web site içeriklerin veri tabana kaydedilmesi için html protokolünü gerçekleştiren uygun C# yazılım dili sınıfları kullanılarak veri tabanına sadece html kodlar aktarılmaktadır. Web site içerikleri Html Parser adı verilen bir sınıf ile Html kodlardan arındırılmakta. Bulanık mantık terim seçme ve ağırlıklandırma şeması ise Matlab Fuzzy Toolbox ile gerçekleştirilmiştir.

3.2.2 Bulanık Terim Seçme Şeması (B-TSŞ)

3.2.2.1 Genel bilgi

Terim seçme ve ağırlıklandırma şemaları hakkında teorik bilgi bu metnin 3.1.5 numaralı bölümde ayrıntıyla verilmektedir. Bunun yanında mevcut olan şemalar hakkında eleştiri metnin üçüncü bolümün sonunda verilmektedir.

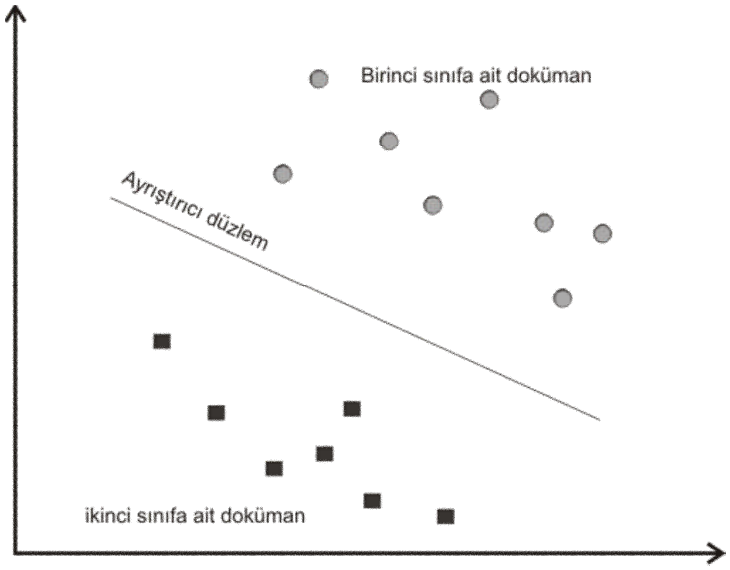
Tez kapsamında yapılan sınıflama görevi için ön işlem olarak kullanılan terim seçme şemaları araştırılmıştır. Birçoğunun probleme özgü değil de istatistik yöntemlerinin özelik seçme işlemine uyarlanmış halidir. Buna karşın değişik araştırmalar ideal terim seçme şemasını söyle tanıtmaktadır: İdeal terim seçme şeması ya da algoritması probleme özgü kurulmuş olandır. Bu algoritma sınıflamayı en verimli sonuca yaklaştıracak terimlerin bir alt kümesini seçip uygun bir ağırlıklandırma gerçekleştirmelidir. Metin madenciliğinde terim seçme ve ağırlıklandırma algoritmaları dokümanların birbirlerinden farklılıklarını ortaya çıkaracak terimleri seçmeyi amaçlarlar.

Bulanık mantık sistemlerin en büyük avantajı bilirkişilerin öznel bilgilerini kullanabilmeleridir. Bilirkişilerden alınacak bilgiler ile giriş çıkış üyelik fonksiyonları ve kural tablosu oluşturulur. Bu parametrelere göre bulanık sistemler bir çıkış vermekteler.

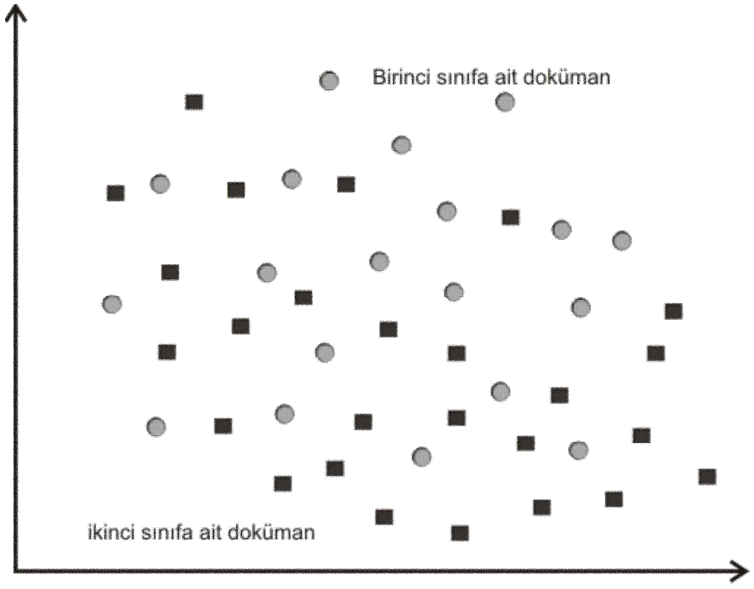
Bulanık sistemlerin terim seçme da kullanılabilir, sonucuna ulaşabiliriz bu noktada. Terim ağırlıklandırma için ise dolaylı kullanılabilirler. Bulanık sistemlerde üyelik fonksiyonların probleme göre uyarlamakla terim seçme şemaların probleme özgü olmaları sağlanır.

Bu çalışmada metin madenciliği yöntemlerinden olan sınıflama tekniği kullanılmıştır. Her bir dokümanı tarif eden birer terim kümesi oluşturulmuştur. Sınıflamanın başarılı olması için bu terim kümelerini oluşturan terimler, dokümanları bir birinden ayrıştıracak ve özeliklerini yansıtacak şekilde seçilmeleri lazım.

Dokumanlar genelde yüzlerce farklı terimden oluşurlar. Bu da doküman terim uzayında gösterilecek iki doküman setinin binlerce elemanı oluşturmasına sebep olmaktadır. Şekil 3.24 de ideal ve Şekil 3.25 de gerçek doküman – terim uzayındaki terim dağılımını görebilmekteyiz.



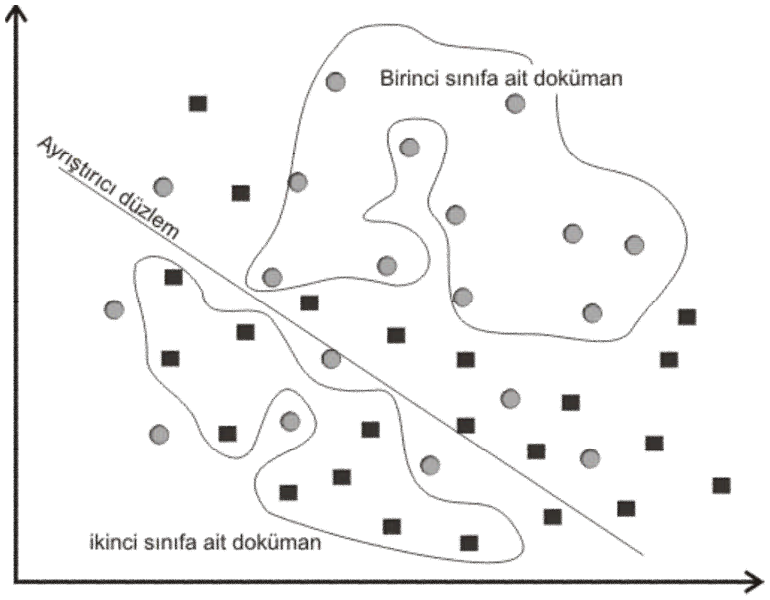
Şekil 3.24 Doküman – terim uzayında İdeal terim dağılımı



Şekil 3.25 Doküman – terim uzayında Gerçek terim dağılımı

Şeklilerden de anlayabileceğimiz gibi dokümanları tarif eden terimler çok sayıdadırlar. Bütün bu terimlerle sınıflamayı gerçekleştirmek imkânsızdır, ayrıştırıcı düzlem bulunamaz. Bir çözüm terimlerin dokümanları ayrıştıracak

şekilde seçmektir. Şekil 3.26 de iki dokümanı ayrıştıracak bir düzlem bulunacak şekilde terim seçimi yapılmıştır.



Şekil 3.26 Terim seçimi

Terimlerden sınıflamayı kolaylaştıracak olanları seçmek terim seçimi ve ağırlıklandırma şemalarının amacıdır. Bu çalışmada değişik terim seçme ağırlıklandırma şemaları uygulandı ve ayrıca bir de yeni bir şema denendi.

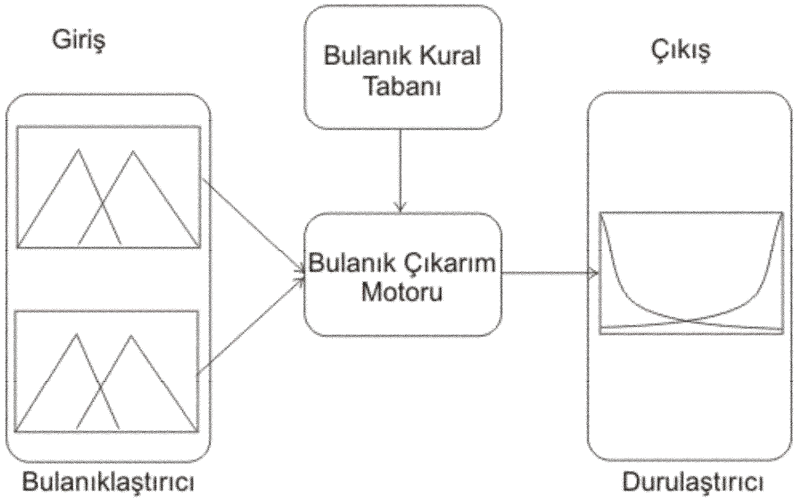
Önerilen şemanın temelinde yatan mantık bir kaç varsayımdan yola çıkılarak bulunmuştur: ”Bir D dokümanında bulunan t terimi, D dokümanında çok tekrarlanıyor ve diğer dokümanlarda az tekrarlanıyor ise t terimi D dokümanı için önemlidir”.

“Bir t terimi bir doküman setinde bulunduğu doküman ve bulunmadığı doküman sayıları karşılaştırıldığında terim ayrıştırıcılığı hakkında bilgi verir, öyle ki bir dokuman seti olsun, bu setin içinde de belli dokuman kategorileri olsun, bir kategori deki D dokumanı içinden seçilen bir terim, bu kategoride doküman frekansı düşük diğer kategorilerdeki toplam dokuman frekansı yüksek ise bu terim ayrıştırıcı bilgi taşır ve seçilmesi lazım ”.

Yukarıda sayılan ve örnekleri çoğaltılabilen varsayımlar ve kurallar ile bir bulanık sistem kurulabileceğinden, bir Bulanık Terim ağırlıklandırma şeması tanımlanıp denendi bu çalışmada.

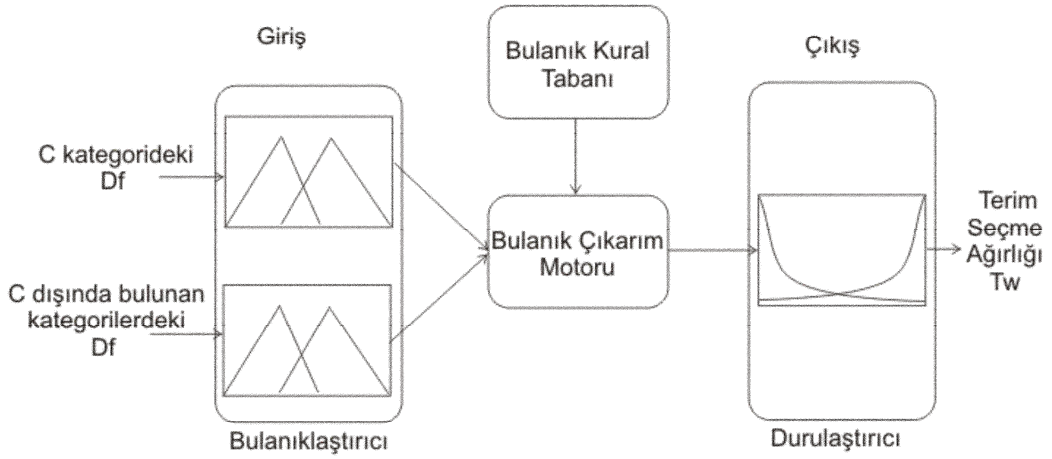
3.2.2.2 Şemanın yapısı

Bulanık sistemler, gerçek giriş ve çıkış değerleri ile çalışırlar. Giriş değerlerini giriş fonksiyonları ile bulanıklaştırırlar, çıkış değerlerini durulaştırıcı adı verilen bir yöntem ile gerçek değere dönüştürler. Şekil 3.27 de B-TAŞ için tasarlanan bulanık sistem verilmekte.



Şekil 3.27 B-TSŞ için Bulanık Sistem

Önerilen bulanık sistemi 2 giriş ve bir çıkışı vardır. Her bir terim için ağırlık değeri bu sistem sayesinde hesaplanacak. Bir t terimi için 2 değer belirlenecek. Birincisi, terimin bulunduğu dokümandaki tekrarlanma sayısı yani terim frekansı, ikinci değer ise t terimin bulunduğu dokuman sayısı yani doküman frekansı. Çıkış değeri ise bulanık sistemden alınacak ve t terimin hesaplanan ağırlık değeri olacak. Şekil 3.27 görünene B-TSŞ şeması Şekil 3.28 de bu yeni bilgilerle tamamlanmaktadır.

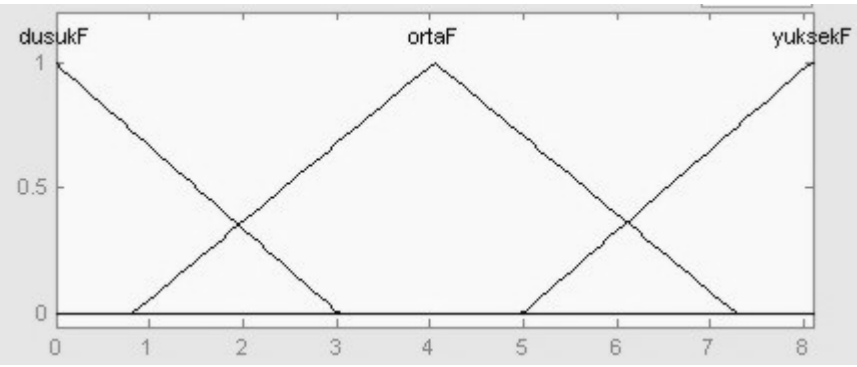


Şekil 3.28 B-TSŞ için Bulanık Sistem

B-TSŞ için bulanık sistemi genel hatları tanıtıldı. Metnin devamında bulanıklaştırma ve durulaştırma işlemlerini gerçekleştirmek amacıyla kullanılan giriş ve çıkış üyelik fonksiyonları incelenecektir.

3.2.2.3 Üyelik fonksiyonlar

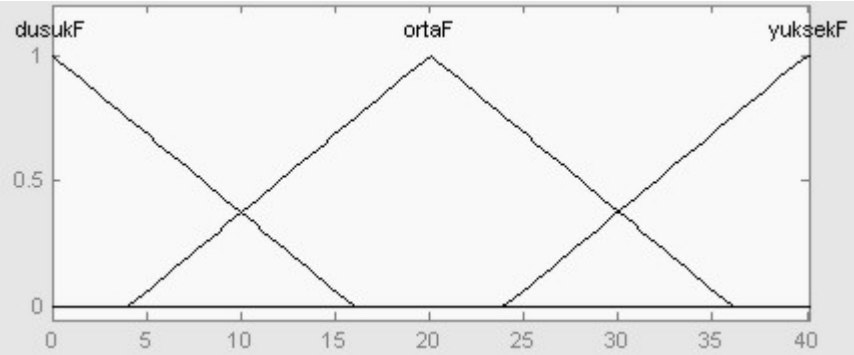
Bulanık sistemler giriş değerlerini bulanıklaştırmak için üyelik fonksiyonlarını kullanırlar. B-TSŞ sisteminde iki giriş verisi tanımlamıştık. İlki kategorideki doküman frekansı, ikincisi ise diğer kategorilerdeki toplam dokunma frekansıdır. Kategori içindeki doküman frekans değerini bulanıklaştıracak giriş Şekil 3.28 da görülmekte.



Şekil 3.28 Kategori içi doküman frekansı

Kategori içi dokuman frekansını bulanıklaştıran ve şekil 3.29 da görünen üyelik fonksiyonlar, düşük frekans orta frekans ve yüksek doküman frekansı olarak tanımlanmışlardır. Şekilde görünen grafik ilk testini yaptığım şemadan alınmıştır, ondan dolayı 8 şayisi en yüksek değer olarak tanımlanmıştır.

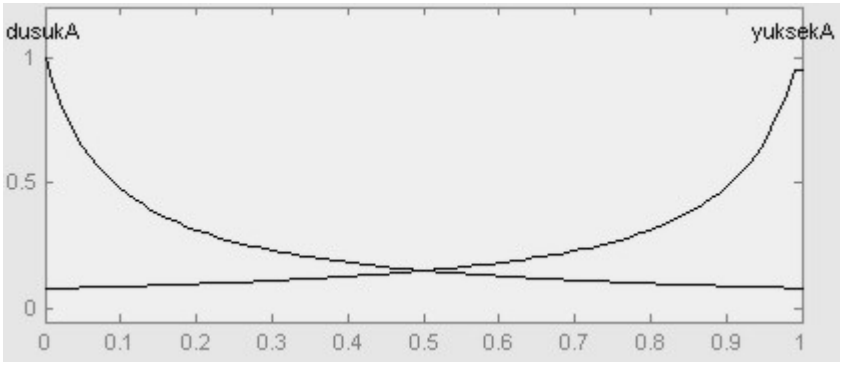
Normal bir uygulamada üyelik fonksiyonların sıfır noktaları probleme özgü tanımlanmalılar. Bu örnekte bir kategoride 8 tane kitap bulunduğundan kategori içi en büyük değer 8 olabileceğinden üyelik fonksiyona bu değer atandı.
Şekilde görüldüğü gibi frekans değeri 0-3 arası ise düşük, 1-7 arası orta ve 5-8 arası yüksek kabul edilecek. Bu değerler bu uygulamada uygun oldukları göründüklerinden bunlar tavsiye edilmekteler. Ancak baksa uygulamalarda bu parametreler uygun sonuç bulunana kadar değiştirilmeleri lazım.



Şekil 3.29 Kategori dışı toplam doküman frekansı

İkinci bulanıklaştırılacak giriş değeri kategori dışı toplam doküman frekansıdır. Bu uygulamada 6 tane sınıf ve her sınıfa ait 8 er kitap belirlenmiştir. C kategorideki t terimi için hesaplanacak bu ikinci değer, söz konusu C kategori dışındaki bütün kategoriler de t terimin bulunduğu doküman sayısıdır. C dışındaki 5 kategoride 8 er kitap 40 değerini vermektedir. Bu yüzden bulanıklaştırıcı üyelik fonksiyonları da en yüksek değer olarak 40 değerini alabilmekte.
Giriş değeri 0-15 arasında ise düşük frekans, 5-35 arası orta frekans ve 25-40 arası da yüksek frekans olarak değerlendirilmektedir.

Çıkış değeri olarak bulanık sitemden alınacak bulanık değer çıkış üyelik fonksiyonu ile elde edilir. Bu değer 0-1 arası bir terim seçme ağırlık değeri olarak tasarlandı.



Şekil 3.30 Terim seçme Ağırlık değeri için üyelik fonksiyonları

Şekilde iki üyelik fonksiyonu görünüyor. Çıkış olarak seçilen fonksiyonların ilki düşük terim seçme ağırlığını temsil eder ve yapısı da sonucun sıfır değerine yakınsayacak şekilde seçilmiştir. Diğer üyelik fonksiyonu ise yüksek terim seçme ağırlığını temsil etmektedir. Çıkış değeri yükselirken ağılık değeri de bire yakınsaması lazım olduğu için fonksiyon da eğriliği ile bu koşulu gerçekleştirmektedir.

1.2.2 Kural tablosu

Bulanık sistemlerin en önemli elemanı kural tablosudur. Bu tablo, giriş değerleri ve bunarlın üyelik fonksiyonlarında aldıkları değerler ile çıkış üyelik fonksiyonlarında haritalanacakları değerler, eğer-ise kuraları ile tanımlandıkları bir tablodur. Bulanık terim seçme sisteminde tanılanan eğer ise karaları listelemeden önce g1 ile kategori içi frekansı yani birinci giriş, g2 ile kategori dışı dokuman frekansı yani ikinci giriş tanımlansın.

- Eğer g1 = düşükF ve g2 = düşükF ise ağırlık = düşükA
- Eğer g1 = düşükF ve g2 = ortaF ise ağırlık = yüksekA
- Eğer g1 = düşükF ve g2 = yüksekF ise ağırlık = yüksekA
- Eğer g1 = ortaF ve g2 = düşükF ise ağırlık = yüksekA

- Eğer g1 = ortaF ve g2 = ortaF ise ağırlık = düşükA
- Eğer g1 = ortaF ve g2 = yüksekF ise ağırlık = yüksekA
- Eğer g1 = yüksekF ve g2 = düşükF ise ağırlık = yüksekA
- Eğer g1 = yüksekF ve g2 = ortaF ise ağırlık = yüksekA
- Eğer g1 = yüksekF ve g2 = ortaF ise ağırlık = yüksekA
- Eğer g1 = yüksekF ve g2 = yüksekF ise ağırlık = düşükA

Kısaca bu kuraların ne anlama geldiklerine bakacak olursak daha önce terim seçme şemalarında kullanılan fikri bulabiliriz. Bir terim bir kategoride az tekrarlanıyor ancak diğer kategorilerde çok tekrarlanıyor ise bu terim ayrıştırıcı değer taşımaktadır ve terim seçme ağırlığı yüksek olmalıdır. Diğer yandan bir terim kategoride çok tekrarlanıyor ancak diğer kategorilerde de aynen çok tekrarlanıyor ise o zaman bu terim ayrıştırıcı değer taşımamaktadır ve terim seçme ağırlığı olan çıkış değeri de düşük olmasıdır.

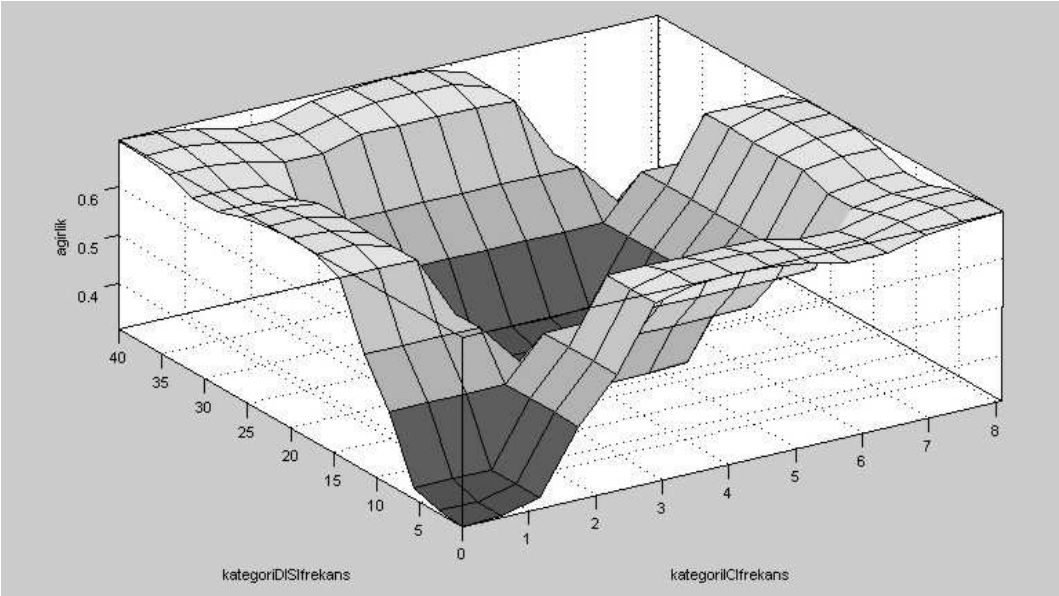
3.2.2.4 Durulaştırma

Bulanık sistemin durulaştırıcı yöntemi için, durulaştırma metotlarından "Ağırlıklı Ortalama Yöntemi" kullanılmıştır. Bu yöntem en çok kabul gören yöntemlerden biridir.

$$Z^{\bullet} = \frac{\sum ü}{ü(n)}$$

Çözüm uzayı

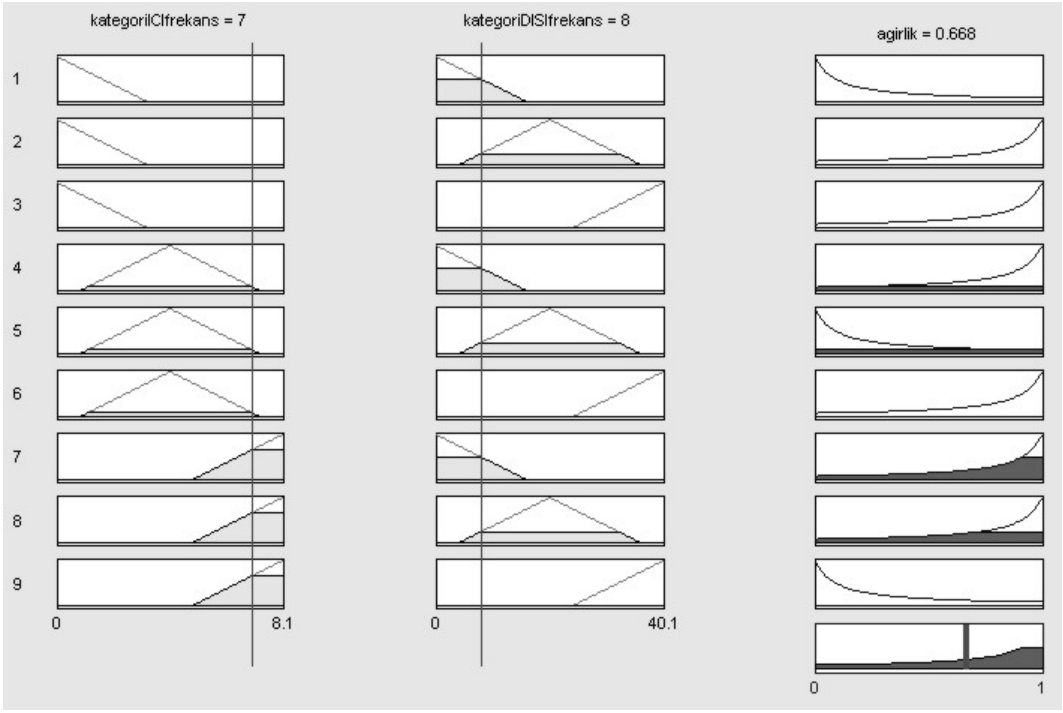
Şekil 3.31 de görünen grafik problem için kurulan deneme bulanık sistemin çözüm uzayını temsil etmektedir.



Şekil 3.31 Bulanık sistemin çözüm uzayı

3.2.2.7 Bulanık sistem için bir örnek

Kurulan bulanık sistemin daha kolay anlaşılması ve bir örnekle test edilmesi için birkaç giriş için elde edilecek çıkışı inceleyelim. İlk örnek için C kategorisinde bulunan D dokümanında t terimi için C kategorisinde bulunduğu dokuman sayısı yani dokuman frekansı 7 olsun, ve diğer kategorilerin toplamında bulunduğu dokuman sayısı da 5 olsun.



Şekil 3.32 Giriş değerlerine göre elde edilecek çıkış değer örneği

Şekil 3.32 da birinci değer 7olarak verildiği görünüyor. Bu değer yüksek frekans üyelik fonksiyonunu ve orta frekans isimli üyelik fonksiyonu kesmektedir. Diğer giriş değeri ise 8 olarak seçilmişti, bu da kategori dışı değişkeninde düşük frekans üyelik fonksiyonu ve orda frekans isimli üyelik fonksiyonu kesmektedir. Durulaştırma işlemi sonrasında elde edilen çıkış değeri 0.668 dır.

Giriş değerleri 7 ve 8 olarak seçilen girişler terimin bulunduğu kategoride çok tekrarlanıyor diğer kategorilerin toplamında az tekrarlandığını temsil etmektedir. Beklenen çıkış değeri orta ve biraz üstünde bir çıkış değeridir, elde edilen değer de 0.668 dir. Girişlerin arasındaki fark büyüdükçe çıkış değeri de yükselmektedir ve tam tersi. Bu da tam istenilen bir ağırlıklandırma dır.

3.2.2.8 Önerilen yeni şemanın avantajları ve dezavantajları

Çalışmada önerilen yeni terim seçme şemasının en büyük avantajı bulanık sisteme dayanması ve probleme özgü uyarlanması kolaylığıdır. Bulanık sistemleri tanımlayan üyelik fonksiyon ve kural tabloların değişimiyle sistemin istenilen sonuçları vermesi sağlanır. Bu da ideal bir terim seçme şemasının ön şartıdır.

Bulanık terim seçme şemasının ağırlıklandırma için kullanılabilmesi de mümkündür. TFIDF isimli ve terim frekansı çarpı ters dokuman frekansı olan şemada ikinci elemanı Bulanık terim seçme çıkış değeri olan ağırlıkla değiştirsen yeni bir terim ağırlıklandırma elde edilir.

$$w_{i,j} = tf_{i,j} x \log(\frac{N}{df_i})$$

Burada dfi elemanını bulanık sistem çıkışı olan terim seçme ağırlığı ile yer değiştirebiliriz. Sonuçta önerilen yeni şemanın formülü :

$$w_{i,j} = tf_{i,j} x \log(\frac{N}{tsA_{i,j}})$$

Deneysel çalışmalarda istenilen sonuçlar elde edilememiştir. Gelecek çalışmalarda bunun nedeni araştırılabilir.

Önerilen yeni terim seçme şemasının en büyük dezavantajı var olan yöntemlerden hesaplanabilirliği daha yüksek ve benzer sonuçlar vermesinde.
Yeni şemanın bu çalışma kapsamında sadece iki defa test edilmesi mümkün olmuştu. Daha büyük ve verimli testlerle bu şemanın gerçek değeri anlaşılabilir

4 SONUÇ

Bu çalışma da web verilerine dayanarak edebiyat eserlerin sınıflaması gerçekleştirilmiştir. Deneysel çalışmalar ile web sitelerde bulanan verilere göre bir sınıflamanın gerçekleştirilmesi mümkün olduğu gösterilmiştir. En iyi sonuçlar "kitap ismi" + book + review –amazon.com –pdf sorgu ile elde edilmiştir. Deneylerden elde edilen bir başka ilginç sonuç ise sadece 2 tane eğitim örneği ile %70 lik başarı gösteren sınıflama modelinin gerçekleştirilebilmesidir.

Terim seçme yöntemlerinden χ^2 kare ile en iyi sonuçlar elde edilmiştir. Ayrıca bu çalışmada geliştirilen bulanık terim seçme şeması ile de χ^2 kareye yakın sonuçlar elde edilerek hibrid sistemlerin de kullanılabileceği gösterilmiştir.

Ancak bu çalışmanın web de bulunan verileri kullanmasından dolayı bazı sınırlamalar da tespit edilmiştir. En büyük sınırlama edebiyat eserlerini tarif edecek web sitelerin elde edilmesi için kullanılan arama motorlarının bulduğu sitelerde problemle ilgili verilerin bulunup bulunmadığını önceden bilememizdir. Edebiyat eserlerini yukarıdaki sorgu ile sorguladığımızda karşımıza gelen siteler genelde çevrimiçi satış yapan web sitelerdir. Bu sitelerin içerisinde kitabi tarif eden bilgiler diğer konu ile ilgili olmayan kelimeler sayısıyla kıyasla çok azdırlar. Örneğin en yüksek ağırlığı olan kelimeler, items, books, order gibi kelimelerdir, sebebi ise bu kelimelerin en çok defa tekrarlanmalarıdır. Bütün bu problem ile ilgisi olmayan kelimeler sınıflamayı yanıltan faktörlerdir. Bu kelimelerin elenmesi için χ^2 kare testi kullanılmış ve ayrıca da bulanık mantık sistemini temel alan yeni bir terim eleme metodu tanıtılmıştır. Burada önemli bir başka sonuç da önerilen yeni terim seçme ve ağırlıklandırma şemanın diğer denenen şemalar ile yakın sonuçlar verebilmesidir.

6 sınıf ve her bir sınıf için 8 er kitaptan oluşan test kümesi ve bir kaç farklı terim eleme ve ağırlıklandırma şeması ile aşağıdaki sonuçlara ulaşıldı. Burada t ve d ile sınıflama modelin kaç tane eğitim ve test örneği ile kurulup test edildiği gösterilmekte.

Terim seçme şeması / test ve eğitim örnek sayısı	t2-d6	t4-d4	t6-d2
tf*idf (500 özellik)	16.6667% (2/12)	20.8333% (5/24)	27.7778% (10/36)
χ^2 kare (100 özellik)	83.3333% (10/12)	87.5% (21/24)	77.7778% (28/36)
Bulanık terim eleme şeması (100 özellik)	75% (9/12)	89.333 (20/24)	72.222 (26/36)

Tablo 4.1 Sınıflama sonuçları.

KAYNAKLAR

Agrawal, R., Imielinski, T. and Swami, A., May 1993, Mining association rules between sets of items in farge databases, In ACM SIGMOD Conf. Management of Data.

Almuallim, H. and Dietterich, T., 1991, Learning with many irrelevant features, In Proceedings of AAI 91, (Menlo Park, CA), AAAI Press, pp. 547-552.

Alpaydın, E., 1999, Zeki veri madenciliği: Ham veriden altın bilgiye ulaşma yöntemleri, www.cmpe.boun.edu.tr/~ethem/files/papers/veri-maden_2k-notlar.doc .

Askerzade, L., 2001. Toward a Perception-based Theory of Probabilistic Reasoning with İmprecise Probabilities**.** Journal of Statistical Planning and İnference 105 (2002) 233-264.

B. How and K.Narayanan, Categorical term descrtiptor: A proposed term weighting shememe for fuature selection, in The 2005 IEEE/WIC/ACM International Conference on Web Intelligence [WI 2005], France, 2005, pp. 313-316

Belen, E. Özgür, Ç. ve Özakar, B., 2003. WALA : Web Erisim Kütük Arastırmacısı. Türkiye Bilisim Dernegi 3. Bilisim Haftası, _stanbul. http://kurultay.tbd.org.tr/kurultay20/Bildiriler/Belgin_Ozakar/bildiri.pdf

C. D. Manning and H. Sch¨utze. Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA, 2001

C. J. van Rijsbergen. A non-classical logic for information retrieval**.** The Computer Journal, 29(6):481–485, 1986.

Chan, K. C. C. and Wong, A. K. C., 1991, A statistical technique for extracting classifıcatory knowledge from databases, In Knowledge Discovery In Databases (G. Piatetsky-Shapiro and W. J. Frawley, eds.), pp. 107-123, Cambridge, MA: AAAI/MIT.

Corinna, C., Drucker, H., Hoover, D. and Vapnik, V., Capacity and complexity control in predicting the spread between harrowing and lending interest rates, In The First International Conference on Knowledge Discovery and Data Mining (U. Fayyad and R. Uthurusamy, eds.), (Montreal, Quebec,Canada), pp. 51-76, aug 1995.

Cross industry standard process for data mining. http://www.crisp-dm.org/, 1999.

D.D. Lewis and M. Ringuette. Comparison of two learning algorithms for text categorization. In Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR'94), 1994.

Deogun, J. S., Raghavan, V. V. and Sever, H., 1995, August, Exploiting upper approximations in the rough set methodology, In The First International Conference on Knowledge Discovery and Data Mining (U. Fayyad and R. Uthurusamy, eds.), (Montreal, Quebec, Canada), pp. 69-74.

Dunham, M.H. 2003. Data Mining Introductory and Advanced Topics. Prentice Hall, New Jersey, 5-19 P, 195-220 P.

E. Leopold and J.Kindermann, Text weighting and ranking with support vector machines. How to represent text in input space? Machine Learning, 46 (2002), pp. 423-444.

E. Wiener, J.O. Pedersen, and A.S. Weigend. A neural network approach to topic spotting. In Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval (SDAIR'95), 1995.

E.Chisholm and T. Kolda, New term weighting formulas for the vector space method in information retrival, Technical report, ORNL/TM-13765, Computer Science and Mathematics Division, Oak Ridge National Laboratory, (1999)

Elder-IV, J. F. and Pregibon, D., 1995, August, A statistical perspective on KDD, In The First International Conference on Knowledge Discovery and Data Mining (U. Fayyad and R. Uthurusamy, eds.), (Montreal, Quebec, Canada), pp. 87-93.

Elmas, Ç., 2003. Bulanık Mantık Denetleyiciler. Seçkin Yayıncılık San. Tic. A.Ş.

Etzioni, O. 1996. The World Wide Web: Quagmire or gold mine. Communications ofthe ACM, 65-68. http://www.cs.washington.edu/homes/etzioni/papers/cacm96.pdf

Fayyad, P. S. U. M., Piatetsky-Shapiro, G. and Uthurusamy, R., 1996a, Advances in knowledge discovery and data mining, Cambridge, MA: MİT Press.

Fayyad, P. S. U. M., Weir, N., Djorgovski, S.G., 2000, Mart 22, Automated analysis of a large-scale sky survey: The SKİ CAT System, http://techreports.ipl.nasa.gov/1993/93-0597.pdf .

Fayyad, P.S.U. M., Piatetsky-Shapiro, G. , 1996b, The KDD process for extracting useful knowledge from volumes of data, Communications Of ACM 39, 11, 27-34.

Frawley, W. J., Piatetsky-Shapiro, G., Matheus, C. J., 1991, Knowledge discovery databases: An overview, In Knowledge Discovery In Databases (G. Piatetsky-Shapiro and W. J. Frawley, eds.), Cambridge, MA: AAAI/M1T pp. 1-27.

G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. Information Processing & Management, 24(5):513–523, 1988.

G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. Communications of the ACM, 18(11):613–620, 1975. (see also TR74-218, Cornell University, NY, USA).

G.Salton and C.Buckley, Term weighting approaches in automatic text retrival, Information Processing and Management, 24 (1988), pp 513-523.

Grzymala-Busse, J. W., 1991, On the unknown attribute values in learning from examples, In Proceedings of Methodologies for Intelligent Systerrzs (Z. W. Ras and M. Zeınankowa, eds.), Lecture Notes in Al, New York: Springer-Verlag, 542, pp. 368-377.

Holsheimer, M. and Siebes, A.P J. M., 1994, Data mining: the search for knowledge in databases.,Technical Report CS-R9406, CWJ, P .0. Bo x 94079, 1090 BG Amsterdam, The Netherlands.

I. H. Witten, A. Moffat, and T. C. Bell. Managing Gigabytes: Compressing and Indexing Documents and Images. Morgan Kaufmann Publishers, San Francisco, 1999.

J. M. G. Hidalgo. Tutorial on text mining and internet content filtering.

J. R. Quinlan. Induction of decision trees. Machine Learning, 1:81–106, 1986.

J.R. Quinlan. Induction of decision trees. Machine Learning, 1(1):81-106, 1986.

J.W. Wilbur and K. Sirotkin. The automatic identication of stop words. J. Inf. ci., 18:45 - 55,1992.

K. Sparck-Jones and P. Willett, editors. Readings in Information Retrieval. Morgan Kaufmann, 1997.

K. Tzeras and S. Hartman. Automatic indexing based on bayesian inference networks. In Proc 16th Ann Int ACM SIGIR Conference on Re-search and Development in Information Retrieval (SIGIR'93), pages 22-34, 1993.

Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information and lexicography. In Proceedings of ACL 27, pages 76-83, Vancouver, Canada, 1989.

Kira, K. and Rendeli, L., 1992, The feature selection problem; Tradational methods and a new algorithm, In Proceedings of AAAI 92, AAAI Press pp. 129-134.

Klir, G.J. and Yuan, B., 1995. Fuzzy Sets and Fuzzy Logic: Theory and Application. Prentice Hall, New Jersey

Koyuncu E., (2004), Yeni Matematiksel Kod; Bulanık Mantık, Teknolojist, ITU IEEE

Lee, S. K., 1992, An extended relational database model for uncertain and imprecise İnformation, In Proceedings Of The 1sth VLDB conference, Vancouver, British Columbia, Canada, pp. 211-218.

Luba, T. and Lasocki, R., 1994, On unknown attribute values in functional dependencies, In Proceedings Of The International Workshop On Rough Sets And Soft Computing, San Jose, CA, pp. 490-497.

M. A. Andrade and A.Valencia, Automatic extarction of keywords from scientific text: Aplication to the knowledge domain of protein families, Bioinformatics,14 (1998), pp. 600-607

M. Hearst. Untangling text data mining. In Proc. of ACL'99 the 37th Annual Meeting of the Association for Computational Linguistics, 1999.

M.Lan, S.-Y. Sung,H.-B. Low, and C.-L. Tan, A comparative stydy on term weighting schemes for text categorization, in International Join Conferece on Neural Network IJCNN, Montreal, Canada, 2005.

Matheus, C. J., Chan, P. K., and Piatetsky-Shapiro, G., 1993, Systems for knowledge discovery in databases, IEEE Trans. On Knowledge And Data Engineering, vol. 5, no. 6, pp. 903-912.

MatLab 6.5, Fuzzy logic toolbox.

Michalski, R. S. and Stepp, R. E., 1983, Learning from observation: Conceptual clustering, In R. S. Michalski, J, G, Oneli C., and Mite T. M., hell editors, Machine Learning: An Artifîcial Intelligence Approach , Vol 1, Morgan Kaufmann, pp 331-363.

N. Fuhr, S. Hartmanna, G. Lustig, M. Schwantner, and K. Tzeras. Air/x - a rule-based multistage indexing systems for large subject _elds. In 606-623, editor, Proceedings of RIAO'91, 1991.

Özakar, B. ve Püskülcü, H., 2002. Web içerik ve web kullanım madenciligi tekniklerinin entegrasyonu ile olusmus bir veri tabanından nasıl yararlanılabilir?.Türkiye'de _nternet Konferansları-VIII. http://inet-tr.org.tr/inetconf8/bildiri/119.doc

Pawlak, Z., Slowinski, K., and Slowinski, R., 1986, Rough classifıcation of patients after highly selective vagotomy for duodenal ulcer, International Journal Of Man-Machine Studies, vol. 24, pp. 413-433.

Peter Knees, Elias Pampalk, Gerhard Widmer, Artist Classification with Web Based Data 1Austrian Research Institute for Artificial Intelligence Freyung 6/6, A-1010 Vienna, Austria 2Department of Medical Cybernetics and Artificial Intelligence Medical University of Vienna, Austria

Quinlan, J. R., 1986, Induction of decision trees, Machine Learning, vol. l, pp. 81- 106.

R. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval. Addison Wesley Longman, 1999.

R. Fano. Transmission of Information. MIT Press, Cambridge, MA, 1961.

R. Feldman and I. Dagan. Kdt - knowledge discovery in texts. In Proc. of the First Int. Conf. on Knowledge Discovery (KDD), pages 112–117, 1995.

R. Gaizauskas. An information extraction perspective on text mining: Tasks, technologies and prototype applications. http://www.itri.bton.ac.uk/projects/euromap/TextMiningEvent/Rob_Gaizauskas.pdf, 2003.

R.H. Creecy, B.M. Masand, S.J. Smith, and D.L. Waltz. Trading mips and memory for knowledge engineering: classifying census returns on the connection machine. Comm. ACM, 35:4863, 1992.

S. Deerwester, S.T. Dumais, G.W. Furnas, and T.K. Landauer. Indexing by latent semantic analysis. Journal of the American Society for Information Sciences, 41:391–407, 1990.

S. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In 7th Int. Conf. on Information and Knowledge Managment, 1998. E. Leopold and J. Kindermann. Text categorization with support vector machines. How to represent texts in input space? Machine Learning, 46:423 – 444, 2002.

S. E. Robertson. The probability ranking principle. Journal of Documentation, 33:294–304, 1977.

Sakiroglu, M. Tug, E. ve Bulun, M., 2003. Web Log Dosyalarından Genetik Algoritma Yöntemiyle Sıralı Erisimlerin Tespit Edilmesi.Türkiye Bilisim Dernegi, 3. Bilisim Haftası, _stanbul. http://kurultay.tbd.org.tr/kurultay20/Bildiriler/Merve_Sakiroglu/bildiri.pdf

Shapiro, G. P. and Matheus, C. J., 1992, Knowledge discovery workbench for exploring business databases, International Journal of Inteldigent Systems, vol. 7, pp. 675-686.

Simoudis, E., October 1996, Reality check for data mining, In IEEE Expert: Intelligent Systems and Their Applications , pages ll(5):26-33.

T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In C. Nedellec and C. Rouveirol, editors, European Conf. on Machine Learning (ECML), 1998.

Takcı, H. ve Sogukpınar, 2002. Erisim Desenleriyle Saldırı Tespiti. Bilgi Teknolojileri Kongresi, Pamukkale Üniversitesi, 6-8 Mayıs 2002, Denizli.

Text mining summit conference brochure. http://www.textminingnews.com/, 2005.

Tom Mitchell. Machine Learning. McCraw Hill, 1996.

Tsoukalas, L.H., 1997. Fuzzy and Neural Approaches in Engineering. John Wiley and Sons, İnc. Publication, Canada.

Tutorial Notes Online: http://ecmlpkdd.cs.helsinki. fi/pdf/hidalgo.pdf, 2002.

U. Nahm and R. Mooney. Text mining with information extraction. In Proceedings of the AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases, 2002.

Vahaplar, A. ve İnceoglu, M.M. 2001. Veri Madenciligi ve Elektronik Ticaret Türkiye'de _nternet Konferansları-VII. Elektronik Poster Bildiri. http://inettr.org.tr/inetconf7/eposter/inceoglu.doc

Wang, L., 1997. A course in fuzzy systems and control. Prantice Hall Publishers, Toronto.

Weiss, S. M. and Kulikowski, C. A., 1991, Computer systems that learn: classifıcation and prediction methods from statistics, Neural Nets, Machine Learning, and Expert Systems , Morgan Kaufman.

Y. Kodratoff. Knowledge discovery in texts: A definition and applications. Lecture Notes in Computer Science, 1609:16–29, 1999.

Y. Liu, B.J. Ciliaxi K.Borges, V. Dasigi, A. Ram, S.B.Navathe and R.DİngledineR.D, Comparasion of two schemes for automatic keyword extraction from MEDLINE for functionla gene clustering, in Proc. Of 2004 IEEE Computational System Bioinformatics Conference (CSB2004), Stanford University, 2004,pp 394-404.

Y. Wilks. Information extraction as a core language technology. In M-T. Pazienza, editor, Information Extraction. Springer, Berlin, 1997.

Y. Yang and W.J. Wilbur. Using corpus statistics to remove redundant words in text categorization. In J Amer Soc Inf Sci, 1996.

Y. Yang. Expert network: E_ective and e_cient learning from human decisions in text categorization and retrieval. In 17th Ann Int ACM SI-GIR Conference on Research and Development in Information Retrieval (SIGIR'94), pages 13-22, 1994.

Yen-Cheng Tseng, Tsung-Ying Sun, (2005), Adaptive Fuzzy Search Algorithm for Improving the Efficiency of Block Motion Estimation.

Zekai Şen, (2001), Bulanık Mantık ve Modelleme İlkeleri,.

Zhong, N. and Ohsuga, S., 1994, Discovering concept clusters by decomposing databases, Data & Knowledge Engineering, vol. 12, pp. 223-244.

EK-1

Deneysel çalışmada sınıflaması gerçekleştirilen edebiyat eserler listesi.

Children's Books

00 2006 The Hello, Goodbye Window
01 2005 Kitten's First Full Moon by Kevin Henkes
02 2006 Criss Cross by Lynne Rae Perkins
03 2005 Kira-Kira by Cynthia Kadohata
04 2006 Remember: The Journey to School Integration by Toni Morrison
05 2005 The First Part Last by Angela Johnson
06 2006 Fat Kid Rules the World by K. L. Going
07 2006 The House of the Scorpion by Nancy Farmer

Mystery

10 2006 Citizen Vince: A Novel by Jess Walter
11 2005 California Girl by T. Jefferson Parker
12 2004 Resurrection Men: An Inspector Rebus Novel by Ian Rankin
13 2003 Winter and Night by S. J. Rozan
14 2003 Fox Evil by Minette Walters
15 2002 The Athenian Murders by Jose Carlos Somoza
16 2001 Sidetracked by Henning Mankell
17 2000 Motherless Brooklyn by Jonathan Lethem

Sci-fi/Fantasy

20 2006 Spin by Robert Charles Wilson
21 2005 Jonathan Strange and Mr. Norrell by Susanna Clarke
22 2004 Paladin of Souls by Lois McMaster Bujold
23 2003 Hominids by Robert J. Sawyer
24 2005 Camouflage by Joe Haldeman
25 2004 Paladin of Souls by Lois McMaster Bujold
26 2003 The Speed of Dark by Elizabeth Moon
27 2002 American Gods by Neil Gaiman

General Nonfiction

30 2003 From the Land of Green Ghosts: A Burmese Odyssey by Pascal Khoo Thwe
31 2002 River Town: Two Years on the Yangtze by Peter Hessler
32 2006 The Worst Hard Time: The Untold Story of Those Who Survived the Great American Dust Bowl by Timothy Egan
33 2005 The Year of Magical Thinking by Joan Didion
34 2005 Plague and Fire: Battling Black Death and the 1900 Burning of Honolulu's Chinatown by James C. Mohr
35 2004 Portland: People, Politics, and Power, 1851-2001 by Jewel Lansing
36 2006 Imperial Reckoning: The Untold Story of Britain's Gulag in Kenya by Caroline Elkins
37 2005 Ghost Wars: The Secret History of the CIA, Afghanistan, and Bin Laden, from the Soviet Invasion to September 10, 2001 by Steve Coll

Literary Fiction

40 2005 The Time in Between by David Bergen
41 2005 Small Island by Andrea Levy
42 2006 March by Geraldine Brooks
43 2005 War Trash by Ha Jin
44 2006 On Beauty by Zadie Smith
45 2006 How I Paid for College: A Novel of Sex, Theft, Friendship, and Musical Theater by Marc Acito
46 2006 The March by E. L. Doctorow
47 2006 Rocks That Float by Kathy B. Steele

Poetry

50 2006 Splay Anthem by Nathaniel Mackey
51 2005 Migration: New & Selected Poems by W. S. Merwin
52 2006 Late Wife: Poems by Claudia Emerson
53 2005 Delights & Shadows by Ted Kooser
54 2005 Refusing Heaven by Jack Gilbert
55 2004 The School Among the Ruins: Poems 2000-2004 by Adrienne Rich
56 2003 Columbarium by Susan Stewart
57 2004 Early Occult Memory Systems of the Lower Midwest by B. H. Fairchild