KÜTÜPHANE OTOMASYON SİSTEMİ VE DOKÜMANLARIN OTOMATİK KATEGORİLENDİRİLMESİ

2013

YÜKSEK LİSANS TEZİ

BİLGİSAYAR MÜHENDİSLİĞİ

Selim ÖZDEM

KÜTÜPHANE OTOMASYON SİSTEMİ VE DOKÜMANLARIN OTOMATİK KATEGORİLENDİRİLMESİ

Selim ÖZDEM

Karabük Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalında
Yüksek Lisans Tezi
Olarak Hazırlanmıştır

KARABÜK
Mayıs 2013

Selim ÖZDEM tarafından hazırlanan "KÜTÜPHANE OTOMASYON SİSTEMİ VE DOKÜMANLARIN OTOMATİK KATEGORİLENDİRİLMESİ" başlıklı bu tezin Yüksek Lisans Tezi olarak uygun olduğunu onaylarım.

Yrd. Doç. Dr. İlhami Muharrem ORAK .....

Tez Danışmanı, Bilgisayar Mühendisliği Anabilim Dalı

Bu çalışma, jürimiz tarafından oy birliği ile Bilgisayar Mühendisliği Anabilim Dalında Yüksek Lisans tezi olarak kabul edilmiştir. 29/05/2013

Ünvanı, Adı SOYADI (Kurumu)	İmzası
Başkan : Doç. Dr. İsmail Rakıp KARAŞ (KBÜ)	.....
Üye : Yrd. Doç. Dr. İlhami Muharrem ORAK (KBÜ)	.....
Üye : Yrd. Doç. Dr. Muharrem DÜĞENCİ (KBÜ)	.....

04./07./2013

KBÜ Fen Bilimleri Enstitüsü Yönetim Kurulu, bu tez ile, Yüksek Lisans derecesini onamıştır.

Prof. Dr. Nizamettin KAHRAMAN .....

Fen Bilimleri Enstitüsü Müdürü

“Bu tezdeki tüm bilgilerin akademik kurallara ve etik ilkelere uygun olarak elde edildiğini ve sunulduğunu; ayrıca bu kuralların ve ilkelerin gerektirdiği şekilde, bu çalışmadan kaynaklanmayan bütün atıfları yaptığımı beyan ederim.”

Selim ÖZDEM

ÖZET

Yüksek Lisans Tezi

KÜTÜPHANE OTOMASYON SİSTEMİ VE DOKÜMANLARIN OTOMATİK KATEGORİLENDİRMESİ

Selim ÖZDEM

Karabük Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı:
Yrd. Doç. Dr. İlhami Muharrem ORAK
Mayıs 2013, 93 sayfa

Üniversite kütüphaneleri, bilimsel düşüncelerin pratiğe aktarıldığı, araştırmaların yürütüldüğü üniversitelerin, en önemli birimlerindendir. Bu birimlerin temel amacı bağlı bulunduğu üniversitelerde ve yakın çevresinde bilimsel araştırma ve geliştirme çalışmalarını desteklemek, gelişip ilerleyen, yeni yayınlarla zenginleşen bilimin son verilerini izlemektir. Kütüphanelerin hemen hemen tamamı kütüphanecilik işlem ve hizmetlerini bir bilgisayar yazılımına - otomasyon programına - dayalı olarak gerçekleştirmektedir.

Kütüphane otomasyon sistemi, kütüphanelerin yükünü hafifletmek ve tanımlı kullanıcıların kütüphaneyi daha aktif kullanabilmesi amacıyla geliştirilen bir otomasyondur. Bu yazılım web tabanlı programlama teknikleri kullanılarak kodlanmıştır. Geniş ve ayrıntılı katalog tarama ile az anahtar kelime kullanarak doğru ve çok sayıda bilgi ve belgeye ulaşılması imkânı sağlanmış olacaktır.

Otomasyon sistemi ile kütüphanelerde bulunan tüm kitaplar, tezler ve yayınların geniş bilgilerinin veritabanında bulunması hedeflenmektedir. Özellikle basılı yayınların içindekiler, önsöz, özet gibi bilgilerinin Optical Character Recognition (OCR) Türkçe ifade ile Optik Karakter Tanıma (OKT) ile metne çevrilip veritabanına kaydedilmesi ile kullanıcıların doğru bilgiye en kolay şekilde ulaşmaları hedeflenmektedir. Kütüphane otomasyonu sayesinde "İçindekiler", "Önsöz" ve "Özet" gibi kısımlara ait bilgilerin yazılım tarafından yorumlanarak bazı bölümlerinin otomatik olarak kategorilendirilmesi ve kataloglanması hedeflenmiştir.

Bu çalışma sayesinde kullanıcılar web ortamından giriş yaptığı anahtar kelimelere en uygun yayın, kitap ya da teze ulaşabilecek ve dokümanların kullanılabilirlik durumlarını görebileceklerdir. Kütüphane personeli ise sisteme girdikleri dokümanları otomatik olarak kategorilendirebileceklerdir.

Çalışmada Bilgisayar Bilimleri, Matematik- Geometri bilimleri, Eğitim Bilimleri, Kişisel Gelişim ve İktisat alanlarında dokümanların kategorilendirilmesi sağlanmıştır

Uygulamada programlama dili olarak Visual Studio C# Asp.NET, veritabanı olarak MsSQL Server 2008 kullanıldı. Yazılım üç katmanlı mimari ile kodlanarak modüler bir yapıya kavuşturuldu.

Anahtar Sözcükler : Kütüphane otomasyonu, optik karakter tanıma, veri madenciliği, metin madenciliği, doküman sınıflandırma.

Bilim Kodu : 902.1.014

ABSTRACT

M. Sc. Thesis

LIBRARY AUTOMATION SYSTEM AND CATEGORIZATION OF DOCUMENTS ARE AUTOMATICALLY

Selim ÖZDEM

Karabük University
Graduate School of Natural and Applied Sciences
Department of Computer Engineering

Thesis Advisor:
Assist. Prof. Dr. İlhami Muharrem ORAK
May 2013, 93 pages

University libraries are the important units in which scientific ideas are turned into practicals and researchs area made. Their purpose are to support scientific research and developments in universities which they belong as well as in surrounding region. They also fulfill tracking of scientific researches developping and advancing every day with new publications. Operations and services in almost all libraries take place by using software, automation program.

Smart Library Automation system is an automation system which was developed for to ease the burden on libraries and make member users of the library use the library more active. This software is coded by using the web based programming techniques. This automation is aimed to enable large and detailed catalog search by using few keywords to reach more precise information and documents. It is also aimed to achieve availability of wide information of the books, theses and

publications in the libraries. Especially, informations from “table of contents”, “preface” and “abstract” sections of printed publications, are converted to text with an OCR program and saved to the database so that users can access the accurate information easily. By means of Smart Library Automation System, informations from “Table of contents”, “Preface” and “Abstract” interpreted by software and some parts of these informations are cataloged automaticly.

This software enables the user to search through key words to Access to most related materials, books or journal and check their availibility in the library. Librarians can do automatic categorization for documents that they introduced to system.

In this study, categorization of Computer Science, Mathematics-Geometry sciences, Educational Sciences, Personal Development and Economics documents are provided.

In this study, Visual C#, Asp.NET as programming language and MsSQL Server 2008 Server as database are used. Software is developed in modular system by designing system in three layers architecture.

Key Words **:** Library automation, optical character recognition, data mining, text mining, document classification.

Science Code **:** 902.1.014

TEŞEKKÜR

Bu tez çalışmasının planlanmasında, araştırılmasında, yürütülmesinde ve oluşumunda ilgi ve desteğini esirgemeyen, engin bilgi ve tecrübelerinden yararlandığım, yönlendirme ve bilgilendirmeleriyle çalışmamı bilimsel temeller ışığında şekillendiren sayın hocam Yrd. Doç. Dr. İlhami Muharrem ORAK'a sonsuz teşekkürlerimi sunarım.

Tez çalışmam boyunca yardımlarını esirgemeyen Yrd. Doç. Dr. Z. Nalan YILMAZ, Öğr. Gör. Hayrettin YILMAZ, Öğr. Gör. Ebubekir SEYYARER ve Öğr. Gör. Taner UÇKAN'a teşekkür ederim.

Üzerimdeki emeklerini hiçbir zaman ödeyemeyeceğim aileme ve manevi desteğini eksik etmeyen sevgili eşim Esra ÖZDEM'e tüm kalbimle teşekkür ederim.

İÇİNDEKİLER

Sayfa

Sayfa

Sayfa

Sayfa

ŞEKİLLER DİZİNİ

Sayfa

Sayfa

ÇİZELGELER DİZİNİ

Sayfa

Sayfa

SİMGELER VE KISALTMALAR DİZİNİ

SİMGELER

i	: sayaç değişkeni
d	: Öklid ölçümlerinde vektörel mesafe
idf	: ters doküman değeri
$idfw_i$	: w_i kelimesinin ters doküman değeri
n	: kelime sayısı
tf	: sözcük frekansı
tfw_i	: w_i kelimesinin sözcük frekans ağırlığı
tf-idf	: sözcük frekansı-ters doküman değeri
tf-$idfw_i$	: w_i sözcüğünün sözcük frekansı-ters doküman ağırlığı
w_i	: i indisli sözcük
A	: A dokümanından oluşturulan vektör
B	: B dokümanından oluşturulan vektör
D	: D dokümanından oluşturulan vektör
S_C	: vektörlerin skaler çarpım sonucu
q	: test dokümanının vektörel hali

KISALTMALAR

KDD : Knowledge Discovery in Databases (Veritabanlarında Bilgi Keşfi)
k-NN : *k*-Nearest Neighbor (*k*-En Yakın Komşu)
SQL : Structured Query Language (Yapısal Sorgulama Dili)
SVM : Support Vector Machine (Destekçi Vektör Makinesi)
OKT : Optik Karakter Tanıma
OCR : Optical Character Recognition (Optik Karakter Tanıma)
OPAC : Online Public Accessible Catalog (Çevrimiçi Genel Erişilebilir Katalog)
MODI : Microsoft Office Document Imaging
LAN : Local Area Network (Yerel Alan Ağı)
WAN : Wide Area Network (Geniş Alan Ağı)
Dpi : Dots per inch (inç başına düşen nokta sayısı)
XML : Extensible Markup Language (Genişletilebilir İşaretleme Dili)
SIP2 : Standard Interchange Protocol Version 2 (Standart Takas Protokolü)
NCIP : NISO Circulation Interchange Protocol (NISO Dolaşım Takas Protokolü)

BÖLÜM 1

GİRİŞ

Bilgiye duyulan ihtiyacın artması ve iletişim teknolojisindeki gelişmelerden kütüphanelerde etkilenmiştir. Bilgisayarların kütüphanecilikte kullanımı kütüphane otomasyonunu doğururken, internet kullanımını, kaynakların ve hizmetlerin daha çok kişiye sunulmasını sağlamıştır (Koç, 1999).

Dijital ortamdaki hızlı değişime ayak uydurabilmek için, üretilen her türlü bilginin hızlıca toplanması gerekmektedir. Zaman ve maliyet açısından daha verimli sonuçlar alabilmek için planlama, programlama ve kontrol metotlarının yanı sıra birtakım yardımcı makinelerin kullanılması, geliştirilmesi gerekmektedir ki bu yardımcı makinaların en önemlisi şüphesiz bilgisayarlardır. Gelişmiş ülkelerde 1960'lı yıllarda başlayan otomasyon çalışmaları 1980'li yıllarda başarılı sonuçlar vermiş ve kütüphanelerde kullanılması da kaçınılmaz olmuştur (İlhan, 1988).

Bilgisayarların kütüphanelerde kullanılmaya başlanması yaklaşık olarak 1960'lı yıllarda gerçekleşmiştir. 1961'de Southern Illinois Üniversitesi 1963'te Toronto Üniversitesi ve Kongre kütüphanelerinde ilk denemeler yapılmıştır. Bilgisayarların kütüphanelerde kullanılmalarının nedenleri ve gereklerinin bu yıllarda tartışılmaya başlanmıştır.

Ülkemizdeki otomasyon sistemi ise 1980'li yıllarda kullanılmaya başlamış olup, internetin de yaygınlaşması ile birçok otomasyon daha fazla kullanıcıya ulaşması amacıyla web ortamına taşınmıştır. Kütüphane sistemlerinin gelişimi Çizelge 1.1'deki gibi gösterilebilir (Haravu, 2009).

Çizelge 1.1. Kütüphane sistemlerinin gelişim süreçleri.

Birinci nesil sistemler (1950- 1960)	Standartlar henüz kullanılmamaktaydı. Kullanıcı erişiminden çok kütüphanede ağırlama üzerinde duruluyordu. Kütüphane yönetimiyle sağlayıcı ilişkisi azdı. Çoğunlukla ana-gövde bilgisayarlar ve toplu sistemler uygulandı.
Orta nesil sistemler (1960-1970)	Machine-Readeble Cataloging (MARC) standardı kullanılmaya başlandı. Verilerin değişimi, kataloglamanın merkezileştirilmesi ve katalog kartlarının dağıtımı üzerinde duruldu. Çeşitli modülleri içeren sistemler geliştirildi. Birinci nesil bütünleşik kütüphane yönetim sistemleri üretildi, bu sistemler tek şubeli kütüphaneleri hedeflemekteydi. Tescilli arka uç tasarımlar (örneğin, düz dosyalar) yaygındı ve çoğunlukla mini-bilgisayar tabanlı, karakter tabanlı ara yüzler; bazı sistemlerde ise henüz kendi üretimi yazılımlar kullanılmaktaydı.
İnternet öncesi nesil (1970-1990)	Local area network (LAN) ve Wide area network (WAN) ile ağlar üzerinden kütüphanelerin iletişimi sağlandı. Interaktif uygulamalar Graphical User Interface (GUI)'lerle geliştirildi ve birinci nesil Online Public Accessible Catalog (OPAC)'lar oluşturuldu. Structured Query Language (SQL) tabanlı sistemlere geçişler başladı.
İnternet nesli (Web 1.0) (1990-2000)	İlk olarak OPAC web üzerine taşındı; ancak diğer modüller ise henüz yerel düzeydeydi. Visual basic ve Visual C++ ile zengin GUI son kullanım araçları oluşturuldu. 90'larda güvenilir internet bağlantısı hem ucuzladı hem de yaygınlaştı, webde veri depolama ve işlemler için yeni istemci sunucusu sistemleri kullanılabilir hale geldi. Linux gibi açık kaynak kodlu işletim sistemleri ortaya çıktı. Tarama sistemleri SQL tabanlıydı.
Web 2.0 çağı 2000 sonrası	Yazılım için web platformu önemli bir seçenek oldu. Geliştirme felsefesi değişti. bitmiş ürün yerine çalışan işlem ve sık güncelleme kavramı yerleşti. Protokoller ve Application Programming Interface (API) aracılığıyla elde edilen bilgilerin tekrar kullanımı, daha fazla ortak çalışma, Real Simple Syndication (RSS) uygulamaları ve farklı keşif uygulamaları geliştirildi. Monolitik yapıdaki kütüphane yönetimi sistemi ve OPAC'lar çok sesli sistemlere dönüştürüldü.

Bugün, tüm üniversite kütüphanelerinde, bilgi kaynaklarının dokümantasyonlarına ait verilerin kayıt, sınıflandırma, ayıklama, hesaplama, özetleme, depolama, güncelleme, çoğaltma ve iletme işlemleri bilgisayar ve iletişim araçlarının, buna

bağlı olarak otomasyon kavramının gerektirdiği niteliklere uygun olarak gerçekleştirilmektedir (Özüsağlam vd. 2009). Son yıllarda kütüphaneciliğimizin gündeminde ki en önemli konu otomasyondur. Bunun nedeninin otomasyonun kütüphanelerimizin hızla değişen ve gelişen sorunlarının çözümüne sağlayacağı katkılar olduğu söylenebilir.

Web kullanıcıları, üniversite kütüphane web sitesi içinde aradıkları bilgiye olabildiğince hızlı ve doğru olarak ulaşabilmelidir. Kullanıcıların aradıkları kaynak ya da bilgiye kolaylıkla ulaşabilmeleri, web sitesinin etkin biçimde çalıştığının bir göstergesidir (Kurulgan ve Bayram, 2006).

Kütüphanelerin ve bilgi merkezlerinin işlemlerini ve hizmetlerini daha etkin, hızlı ve doğru bir biçimde gerçekleştirmek amacıyla söz konusu işlem ve hizmetlerde bilgisayar ve uzak iletişim teknolojisi başta olmak üzere bilgi teknolojisinin tüm ürünlerinden yararlanılması kütüphane otomasyonu olarak tanımlanabilir.

İnsanlığın üretim ilişkileri ve bu ilişkilerin ekonomiye, kültüre, teknolojiye, kısacası yaşamın her alanına yansıması, elektroniğin temellerinin atılması ile birlikte, otomasyon kavramının da ortaya çıkmasını sağlamıştır. Bilgisayarların kullanılmasıyla yaşanan bilgisayarlaşma, otomatikleşme veya otomasyon kavramı, düzgün çalışan her kütüphanede yapılması gerekli tekrarlayıcı faaliyetler için ayrılmış personel zamanını azaltmak amacıyla kullanılmaktadır ve literatürde kabul görmüş tanımlarından bazıları şöyledir;

Otomasyon; kütüphanelerde sürekli gerçekleştirilen, doküman sağlama, kataloglama, ödünç verme, süreli yayınların denetimi ve danışma hizmetlerinin bilişim sistemlerine dayalı olarak gerçekleştirilmesi, kütüphane yönetiminde bilgisayar kullanımı ve uzak iletişim teknolojisi başta olmak üzere enformasyon teknolojisinin tüm ürünlerinden yararlanmaktır.

Otomasyon; kendi başına çalışan bir sistem değildir, insan gerektirir. Sorunları kendi başına çözmez sadece yardımcı olur, haber verir. Üretim maliyetlerini düşürür, işletme kolaylığı ve konfor sağlar, ürün kalitesini arttırır. Süreklilik ve

standardizasyon sağlar. Ancak, iş disiplini ve organizasyonu, işletim ve bakım planlaması, üretim birimleri arasında iletişim ile personel eğitimi ve adaptasyonu gerektirir. Tüm bunların kusursuz işleyebilmesi için de; insan inisiyatifini en aza indiren, olası insan hatalarını karşılayabilen, alternatif iletişim yöntemi sunabilen, otomatik yedekleme sistemine sahip, işletim parametreleri kolaylıkla değiştirilebilen, diğer sistemler ile kolaylıkla entegrasyonu sağlanabilen, maksimum düzeyde güvenlik önlemlerine sahip, üretim ve işletim ile ilgili sürekli raporlar üretebilen bir yazılım gereklidir.

Kütüphanelerin problemlerine çözüm getiren ve en çok kullanılan ticari amaçlı yazılımlar aşağıda belirtilmiştir;

1. Koha,
2. Milas,
3. Yordam,
4. Symphony.

Günümüzde Kütüphane otomasyonu olarak yazılımlardan biri Koha'dır. Koha ilk olarak 1999 yılında Yeni Zelanda'da Horowhenua Kütüphanesi için geliştirilmeye başlanmış açık kodlu bir kütüphane yazılımıdır. Dünya çapında çeşitli boyutlarda 600'den fazla kütüphanede kullanılmaktadır. Linux işletim sistemi üzerinde Mysql ve PostgreSQL veri tabanı kullanılarak ve Perl, Php yazılım diliyle geliştirilmiş Apache web sunucu üzerinde çalışmaktadır. Yakın Doğu Üniversitesi tarafından ülkemizdeki bin yüz on iki halk kütüphanesinde kullanılmak üzere geliştirilmektedir.

Kullanılan otomasyonlardan bir diğeri ise MİLAS'dır. Web tabanlıdır ve Php yazılım dili ile geliştirilmiştir. Veritabanı olarak MySQL ve PostgreSQL kullanmaktadır. 2003 yılından itibaren 76 kütüphanede aktif olarak kullanılmaktadır.

Ülkemizdeki Üniversitelerin % 75'inin kullandığı bir diğer Kütüphane otomasyonu Yordam'dır. Yordam Kütüphane otomasyonu Asp ve Php programlama dillerinden geliştirilmekte olup veri tabanı olarak MySQL kullanmaktadır.

Symphony ise iki farklı firmanın (Sirsi ve Dynix) bir araya gelerek oluşturduğu SirsiDynix firması tarafından üretilen kütüphane otomasyon sistemidir. Türkiye'de kullanıcı sayısı hızla artmaktadır. Symphony tam belgeli API ve Web hizmetleri sunmaktadır. SIP2, NCIP standartlarına uyumludur. Dolaşım için web tabanlı personel istemcisine sahiptir. Güçlü Java istemci süreçleri düzenleyerek, bu tür self-servis istasyonları ve malzeme siparişi arabirimleri gibi zaman kazandıran arabirimlerini desteklemektedir. Program mobil uygulamalara elverişlidir. 600'ün üzerinde hazır raporlamaya olanak sunmakta ve XML çıktı seçenekleriyle kullanıcılarına hizmet vermektedir. Symphony, Ankara'da bulunan ofisi ile Türkçe olarak teknik destek sunmaktadır. Literatürde kütüphane otomasyonu konusunda yapılan çalışmalar şöyledir;

Takçı ve Soğukpınar (2002) çalışmalarında kütüphane sitesi web günlüklerine dayalı olarak kütüphane kullanıcılarının erişim örüntüleri bulunmaya çalışılmıştır. Bu çalışma yapılırken istatistiksel yöntemler kullanılmıştır.

Çakmak ve Özel (2010) dokümanların kataloglanması ile ilgili bir çalışma yaparak çevrimiçi kataloglama ve sosyal kataloglamanın önemine dikkat çekmişlerdir. Bu konularda yapılan çalışmalara değinmişlerdir.

Witten et al. (2000) yaptıkları açık kaynak kodlu kütüphane yazılımı ile giriş ve özet bilgilerini sisteme yükleyerek, bu bölümlerde de arama yapılmasını sağlamışlardır. Ayrıca sistemlerinde metin madenciliği ile dokümanlar için metadatalar oluşturmuşlardır.

Bayram ve Çetinkaya (2008), kütüphane otomasyonuna farklı bir açıdan bakıp, kataloglanan, kategorilendirilen dokümanların kütüphane içerisinde raf yerini belirleyen, istenilen dokümanı rafa yerleştiren ve raftan alan yürüyen bant sistemi ile getiren bir sistem geliştirmişlerdir. Kütüphane otomasyonunu zorunlu kılan gerekçeler kısaca şöyle sıralanabilir.

1. Bilginin öneminin anlaşılması,
2. Üretilen bilgiye olan bağımlılığın artışı,

3. Belge ve bilgilerin geometrik bir hızla artışı,
4. Bilgi kaynaklarının çeşitliliği,
5. Kütüphanelerdeki yoğun emek gerektiren işlemler,
6. Rutin kütüphanecilik işlemleri,
7. Personelin çalışma zamanının kısıtlı olması,
8. Bilgi hizmetlerinin hızla verilme gereksinimi,
9. Bilgi yönetimi için gerekli teknolojilerin gelişmiş olması ve maliyetlerin giderek düşmesi (Yılmaz ve Aslan, 1992).

Kütüphane otomasyonu en temel anlamda aşağıda belirtilen modüllerden oluşmaktadır.

1. Tanımlama Modülü
2. Kataloglama - Kategorilendirme Modülü
3. Katalog Tarama Modülü
4. Ödünç Verme Modülü
5. Güvenlik Modülü
6. İstatistik Modülü

Kütüphane otomasyonunun etkin bir şekilde kullanılması için kataloglama ve kategorilendirme modüllerinin etkin olması gerekmektedir. Sisteme eklenecek olan dokümanların telif hakları da dikkate alınarak girilebilecek tüm bilgilerinin sisteme eklenmesi gerekmektedir. Bu şekilde kullanıcı katalog taraması yaparken, tam olarak ulaşmak istediği dokümanlara ulaşması sağlanır.

Günümüz kütüphane otomasyon sistemlerinde dokümanlar ilgili personel tarafından kataloglanır. Personelin dokümanı kataloglarken yapacağı bir hata o dokümana ulaşmayı engelleyebilir ya da doküman ilgisi olmayan bir aramada kullanıcıların karşısına çıkabilir. Bu durumu engellemek için dokümanın temel bilgileri sisteme girilirken, arama bölümünde kullanılacak kategorilendirme bilgilerinin sistemde otomatik olarak oluşturulması hedeflenmiştir.

Dokümanların otomatik olarak kataloglama, kategorilendirme işlemi veri madenciliği alanıyla yakından ilgilidir.

Veri madenciliği, önceden bilinmeyen ve potansiyel olarak faydalı olabilecek, veri içindeki gizli bilgilerin çıkarılmasıdır (Frawley and Matheus, 1991). Veri madenciliği yapısal veriler üzerinde çalışır. Fakat metin dosyaları yapısal olmayan verilerdir. Bu tür verilerin işlenebilmesi için yapısal hale dönüştürülmesi gerekir. Metin madenciliği bu problemlere çözüm olarak sunulan, metin formatındaki verileri kullanarak içerisindeki bilgileri gün ışığına çıkaran ve özellikle 2000'li yıllardan sonra ilginin giderek arttığı önemli bir alandır (Konchady, 2006).

Metin madenciliği, özel amaçlar için metinden bazı bilgiler çıkarmak adına, metnin analiz edilmesi işlemidir. Bu analiz işlemlerinden bir tanesi de tez çalışmasının konusu olan dokümanların kategorilendirilmesi işleminin basamağı sınıflandırmadır. Metin sınıflandırma, önceden belirlenmiş sınıflara dokümanların atanması işlemidir (Mitchell, 1997).

Veri madenciliği, eldeki verilerden üstü kapalı, çok net olmayan, önceden bilinmeyen ancak potansiyel olarak kullanışlı bilginin çıkarılmasıdır. Bu da; kümeleme, veri özetleme, değişikliklerin analizi, sapmaların tespiti gibi belirli sayıda teknik yaklaşımları içerir.

Literatürde veri madenciliği ile metin sınıflandırma ve doküman sınıflandırma ile ilgili, Takçı ve Soğukpınar (2002) Gebze Yüksek Teknoloji Enstitüsündeki kütüphane kullanıcılarına daha iyi hizmet vermek amacıyla, kullanıcıların kütüphane otomasyonundaki kullanım örüntülerini keşfetme üzerinde çalışmalar yapmışlardır. Kullanıcıların, web sayfalarını gezinirken bıraktıkları izlerden veriler toplayarak bu verilerden anlamlı sonuçlar çıkarmaya çalışmışlardır.

Amasyalı and Diri (2006) yaptıkları çalışmada; yazarı bilinmeyen bir dokümanın önceden belirlenmiş 18 farklı yazar içerisinden hangisine ait olabileceği sorusunun cevabı aramışlardır. Naive Bayes, SVM, C 4.5 ve Random Forest sınıflandırıcıları

seçilerek 10 ayrı özellik vektörü, 10'lu çapraz geçerlilik ile çalıştırılmış ve oldukça başarılı sonuçlar elde etmişlerdir.

Aşlıyan ve Günel (2010), metin içerikli Türkçe dokümanların sınıflandırılmasıyla ilgili çalışmalarında, En Yakın Komşu ve *k* En Yakın Komşu (*k*-NN) metotlarını kullanarak dokümanların 5 farklı kategoriye göre sınıflandırılmasını gerçekleştiren sistem tasarlamışlar ve gerçekleştirmişlerdir. En Yakın Komşu metodu, *k*-NN metoduna göre daha başarılı olduğunu tespit etmişler ve bütün sınıflar için % 88,4 oranında başarı elde etmişlerdir.

Yang and Liu (1999) yaptıkları çalışmada; *k*-NN, NaiveBayes ve SVM (Support Vector Machines) yöntemleri kullanılarak metin sınıflandırma performanslarının karşılaştırmışlardır. Çalışma sonucuna göre SVM ve *k*-NN'in Naive Bayes'e göre daha başarılı sınıflandırma yaptığı görülmüştür.

Erol ve Gülseçen (2009) Naive Bayes frekans ağırlıklandırma yöntemi kullanarak haber sınıflandırma üzerinde çalışmışlardır. Sınıflandırma sayesinde arama motorlarının ürettiği sonuçların çok daha başarılı olacağına dikkat çekmişlerdir.

Sayın Tonta vd. tarafından Hacettepe Üniversitesi, Bilgi ve Belge yönetimi bölümünde Yüksek Lisans, Doktora ve Sanatta Yeterlik Tezlerinin Dijitalleştirilmesi konulu bir proje çalışması yapılmıştır. Bu çalışmada basılı tezler bilgisayar ortamına aktarılmıştır. Tezlerin içindekiler, önsöz, giriş bölümleri arama motorunun alt yapısını oluşturmuştur. Yazılım konusunda çeşitli seçenekler değerlendirilmiş, ticari bir veri tabanı yönetim sistemi kullanmak ya da tezlere özel bir yazılım geliştirmek yerine, yabancı ülkelerde bu amaçla kullanılan açık kaynak kodlu yazılımların değerlendirilmesine karar verilip, DSpace yazılımını kullanmışlardır (http://yunus. hacettepe.edu.tr/~tonta/yayinlar/02-G-064-elektronik-tez-projesi-sonuc-raporu.pdf, 2013).

Karaca (2012) tez çalışmasında internet gazetelerindeki Ekonomi, Spor, Sağlık, Eğitim ve Yaşam gibi 5 farklı kategorideki köşe yazılarının sınıflandırılmasını sağlamıştır. Bu çalışmasında Bit, *tf*, *idf* ve *tf-idf* ağırlıklandırma, 37 sınıf özellik

vektörü ve farklı sınıflandırma algoritmaları kullanılarak 105 sınıflandırma gerçekleştirmiştir. Uygulama sonuçlarına göre *tf-idf* ağırlıklandırma, kosinüs benzerliği ve *k=7* değeri kullanılarak yapılan sınıflandırmalarda 4 farklı sınıf özellik vektöründe bütün yazıların sınıflandırılması % 100,00 doğrulukla gerçekleştirmiştir. 2 farklı sınıf özellik vektöründe, *tf-idf* ağırlıklandırmanın Multi-Nominal modelle uygulandığı sınıflandırmalarda da % 100,00 başarı elde etmiştir.

Bu tez çalışmasının amacı, Kütüphaneler için arama motoru güçlü ve dokümanları otomatik olarak kategorilendirebilen bir otomasyon sistemi geliştirilmesidir. Basılı dokümanların sisteme veri girişi sağlanırken, önceden taranan içindekiler, önsöz, giriş gibi bilgilerinin Optik Karakter Tanıma (OKT) yöntemi ile veritabanına metin olarak kaydedilmesi amaçlanmıştır. İçindekiler, önsöz, giriş gibi bilgiler içerisinde de arama yapabilecek bir sistem düşünülmüştür. Bilgisayar, Matematik–Geometri, Eğitim, Kişisel Gelişim, İktisat gibi 5 farklı alandaki dokümanların sistemde sınıflandırılması amaçlanmıştır. Bu işlemleri gerçekleştirmek için Visual C# ile Visual Basic 2010 ortamında, veritabanı olarak MsSQL olan web tabanlı bir yazılım geliştirilmiştir.

Bu tez çalışması yedi bölümden oluşmaktadır. İkinci bölümde kütüphane otomasyonu ve kütüphane otomasyonunu oluşturan unsurlardan bahsedilmiştir. Üçüncü bölümde ise bu tezdeki uygulamanın da bir parçası olan OKT konusuna değinilmiştir. Dördüncü bölümde dokümanların sınıflandırılması için kullanılan veri ve metin madenciliği anlatılmış olup beşinci bölümde ise doküman sınıflandırma işlemine değinilmiştir. Altıncı bölümde uygulama, yedinci bölüm ise sonuç ve öneriler kısmını oluşturmaktadır.

BÖLÜM 2

KÜTÜPHANE OTOMASYONU

Gelişen web teknolojileri ile beraber kütüphanelerin ve kütüphanecilerin de rolleri değişmektedir. Kullanıcılarının bilgiye erişme davranışlarını temel alarak yapılan çalışmaların hedefi, etkin bir kütüphane kullanıcısına sahip olmak ve onların en kısa yoldan aradıkları bilgiye ulaşmalarını sağlamaktır.

Hizmet kalitelerini artırma çabasında içindeki kütüphanelerimiz e-yayınların sayısındaki artış ve konsorsiyumlar sayesinde koleksiyonlarını zenginleştirmektedir. Bundan sonraki adım ise bu kaynaklardan kullanıcılarını en etkin ve etkili şekilde yararlandırmalarıdır. Her zaman bilişim teknolojilerinin en hızlı uygulayıcısı olan kütüphaneler, gelişen yeni web teknolojileri sayesinde sahip oldukları elektronik ve basılı kaynaklara ulaşımı daha kolay hale getirmenin yollarını keşfetmekte, kendilerinin ayrılmaz bir parçası olan kullanıcılarını geliştiren bu dinamik uygulamalarla sistemin gelişmesinde paydaş hale getirmektedir.

Kütüphanelerimizin zaman ve mekân problemi olmadan tüm kaynaklarını hedef kitlelerine ulaştırması günümüz teknolojileriyle teorik olarak hiç zor değildir. Fakat kütüphanelerin teknolojiyi kullanması ve bu teknolojinin sürdürülebilir olması için arka planda ciddi bir ekip ile çalışması gerekmektedir.

Kullanıcıların, kütüphane kaynaklarına her yerden ulaşabilmesi, kuşkusuz kaynakların internet ortamına uygun olarak kataloglanıp, erişeme açılmasıyla olur. Bu işlem de kütüphanelerin etkin web sayfalarının oluşturulmasıyla gerçekleşebilir.

Bir web sitesinin oluşturulmasında ve yönetiminde amaç, kullanıcılara yararlı bir hizmet sağlamak ve kuruluşun hizmetlerini duyurmak için uygun bir pencere açmaktır (Al ve Bahşişoğlu, 2000).

Bugün kütüphane ve bilgi merkezlerinin bilgisayarlar ve internet aracılığı ile gerçekleştirdikleri hizmetleri şu başlıklar altında listelemek mümkündür.

1. Bilgi ve/veya belgenin kütüphanede var olup olmadığını izleme işleminde,
2. Sipariş (abone ve/veya satın alma) kütüklerinin oluşturulması ve siparişlerin izlenmesinde,
3. Belgelerin bibliyografik ve içerik tanımlarının, kataloglama, sınıflandırma, dizinleme, özet çıkarma, katalog kayıtlarının hazırlanması ve çoğaltılması işlemlerinde,
4. Kütüphane materyalinin fiziksel olarak kullanıma hazırlanmasında,
5. Ödünç verme (dolaşım) işlemlerinin yürütülmesinde,
6. Materyalin saklanması ve korunmasında,
7. Yönetimle ilgili faaliyetlerde.

Yukarda belirtilen hizmetler ve daha fazlası kütüphane otomasyonu sayesinde yapılabilir.

2.1. DÜNYADA KÜTÜPHANE OTOMASYONU

Ülkemizde olduğu gibi dünyanın birçok kütüphanesinde Online Computer Library Center (OCLC) kurumunun geliştirmiş olduğu OCLC Connexion adlı web tabanlı program kullanılmaktadır. Bu program dünyada 10 000'in üzerinde kütüphanenin kullanmakta olduğu bir sistemdir.

OCLC Kataloglama işlemlerinin uluslararası standartlarda ve kaliteli bir şekilde yapılmasını sağlamaktadır. OCLC Connexion bir standartlaşma ve ortaklaşa kataloglamayı sağlaması sayesinde kütüphane kaynaklarına tüm dünyadan tek bir arayüzden (WorldCat) erişilebilme olanağı sağlamaktadır.

Bu ağa dahil olan bir kütüphanenin kaynaklarına hem dünyanın en geniş ve büyük kataloğu olan WorldCat'ten, hem de kütüphanenin kendi kataloğundan kesintisiz olarak erişilmeye başlanır.

2.2. KÜTÜPHANE OTOMASYONU MODÜLLERİ

Kütüphane otomasyonu en temel anlamda aşağıda belirtilen modüllerden oluşmaktadır.

1. Tanımlama Modülü,
2. Kataloglama - Kategorilendirme Modülü,
3. Katalog Tarama Modülü,
4. Ödünç Verme Modülü,
5. Güvenlik Modülü,
6. İstatistik Modülü.

2.2.1. Tanımlama Modülü

Bu bölümünde yöneticiler, kullanıcılar, personeller, sağlayıcılar, kategoriler, unvanlar gibi tanımlamaların yapılması sağlanır. Tanımlama modülünde yapılan işlemler programın kullanıcı dostu olmasını sağlar. Bu bölümde tanımlanacak veriler diğer modüllerde kullanılırlar.

Özellikle kataloglama işlemlerinin sağlıklı yapılabilmesi için bazı bilgilerin standart olarak girilmesi gerekmektedir. Tanımlama modülünde bu standartların girilmesi sağlanır.

2.2.2. Kataloglama ve Kategorilendirme Modülü

Kütüphaneler, dokümanları kullanıcıların hizmetine anlamlı ve ulaşılabilir bir biçimde sunmak için belirli kurallar ve sistemler yardımıyla listeleme yaparlar.

Kataloglama, bir dermedeki bilgi kaynaklarının belirli kurallar doğrultusunda bibliyografik olarak tanımlanıp; yazar, konu, eser adı gibi erişim ögelerinin sağlanması ve bunların kütüphane kataloglarında sunulması işlemlerini içermektedir (Keenan ve Johnston, 2000).

İyi bir kütüphane hizmeti vermek, iyi düzenlenmiş katalog bilgilerine sahip olmayı zorunlu kılar. Anglo Amerikan Kataloglama Kuralları kitabının editörü olan ünlü katalogcu Gorman (2002), bu konu hakkında ki görüşlerini şu şekilde dile getirmektedir;

"Kataloglama kütüphaneciliğin temelidir. Sanıyorum ki bütün alanlardaki iyi kütüphaneciler kataloglama konusunda bilgi sahibidir. Danışma kütüphanecisi sadece bilginin nasıl düzenlendiği konusunda değil aynı zamanda kataloglama ve sınıflamanın temel yapısını ve bilginin genelden özele doğru düzenini de takip eder. Derme geliştirme ve yönetme ile ilgili kütüphaneciler sorumluluklarını sınıflamanın konu guruplarına ve konu başlıklarına göre yerine getirirler. Bir çocuk kütüphanecisi çocuğa iyi bir kitap önerebilmek için kataloglama tecrübesinden yararlanarak konuya ya da yaşa göre gruplandırmaları düşünecektir".

Bloomberg and Evans (1989) kataloglamanın en temel amacının, kullanıcılara kütüphanelerin sahip oldukları dokümanları göstermek ve bu dokümanların raftaki yerlerini bildirmek olduğunu belirtmişlerdir.

Kütüphaneler kendi standartlarına göre katalog kayıtlarını tutarken 1961 yılında Kütüphane Dernek ve Kurumları Milletlerarası Federasyonu (International Federation of Library Associations and Institutions, IFLA)'nın gerçekleştirdiği toplantıda kayıtlarının standartlaştırılması hakkında "Paris Bildirileri" olarak anılan bildiri yayınlandı. Bu bildiri ile kataloglamada standartlaşmanın temelleri atıldığı varsayılmaktadır.

Günümüzde, dünyada en çok kullanılan kataloglama standardı olarak, Amerika Birleşik Devletleri (ABD) kongre kütüphanesinin geliştirdiği, Machine-Readeble Cataloging (MARC) formatıdır. Bu standart sayesinde kütüphaneler arasında katalog kaydı paylaşmak mümkün olmuştur.

2.2.3. Katalog Tarama Modülü

Kütüphane otomasyonlarının etkinliği, kullanıcılarının ulaşmak istedikleri dokümana, bilgiye, az anahtar kelime kullanarak ya da daha az menü kullanarak hızlı bir şekilde ulaşabilmeleriyle ölçülür. Sayın Küçük (1999), kullanıcıların eğilimi bilgi kaynaklarından ziyade, bilginin kendisine erişim isteği ile değişmekte olduğunu vurgulamıştır.

Bilgiye ulaşmaktaki hızın önemi giderek artmaktadır. Günümüz kütüphane otomasyonlarında katalog tarama bölümleri basit, gelişmiş ve detaylı tarama olarak farklı kategorilerde kullanıcılarına hizmet vermektedir. Katalog tarama modülü bir otomasyonda, kullanıcının bilgiye ulaşmadaki en önemli aracıdır.

2.2.4. Ödünç Verme (Dolaşım) Modülü

Kütüphanenin ve kütüphane otomasyon sisteminin kullanıcı ile en etkileşimli olduğu modül ödünç verme modülüdür. Bu modül ile kütüphane kaynakları kullanıcının hizmetine sunulmaktadır.

Katalog bilgisi ile okuyucu bilgilerini birlikte kullanarak ödünç verme ve iade işlemlerinin izlenmesi, gözden geçirilmesi, mevcutların ve okuyucu üzerinde olanların gösterilmesi, para cezalarının hesaplanması, hatırlatma ve uyarı notlarının üretilmesi, koleksiyonun kullanım oranına ilişkin istatistiklerin tutulması işlemlerinin tümüne ödünç verme işlemleri denir (Erol, 1990).

Ödünç verme işlemi, kullanıcılara verilen çok önemli bir hizmet olduğu gibi iyi yönetilemediği takdirde kütüphaneleri de zor durumda bırakabilecek bir uygulamadır. Ödünç verme sisteminin sorunsuz bir şekilde işletilebilmesi için aşağıda bahsedilenlerin otomasyon sisteminde olması gerekmektedir.

1. Kullanıcının (okuyucunun) üyelik kaydı,
2. Kullanıcının materyal ödünç alma durumunun kontrolü (üzerinde başka bir materyal ya da ceza durumunun olup olmadığının kontrolü),

3. Kullanıcıya kaynağın ödünç verilmesi,
4. Süre bitiminde kaynağın iadesinin kabulü,
5. Kullanıcının kaynağı zamanında teslim edip etmediğinin kontrolü,
6. Zamanında teslim edilmeyen kaynaklar için ceza uygulanması.

2.2.5. Güvenlik Modülü

Yazılım geliştirme sürecinin en önemli adımlarından biri de güvenliktir. Yazılımların yaygın olarak kullanılmaya başlandığı ilk yıllarda kaliteli ve olgun yazılım üretmek, son yıllarda ise özellikle güvenli yazılım geliştirmek için çok sayıda model ve çerçeve üzerinde çalışılmıştır. Bu durumun en büyük tetikleyicisi son yıllarda güvenlik açıklıklarının artmasıdır. Yazılım alanında uygulama sayısı gün geçtikçe artmaktadır. Bu artışla paralel olarak uygulamaların güvenlik problemleri de artmaktadır.

2.2.5.1. Yazılım Güvenliği

Yazılım güvenliği kavramı ile ilgili yapılan en önemli yanlış, onu sadece kodun güvenliği ve ek olarak da yetkilendirme güvenliği olarak algılamaktır. Hâlbuki yazılım güvenliği kavramını "güvenilir bilişim" (trusted computing) kavramı ile yakından ilişkilendirmek gerekmektedir.

Şöyle ki, Trusted Computing Group (TCG) tarafından konmuş olan güvenilir bilişim kavramı 4 sacayağı üzerinde durmaktadır. Bunları şöyle sıralayabiliriz;

1. Gizlilik,
2. Bütünlük,
3. Erişilebilirlik,
4. Kurtarılabilirlik.

İşte ancak bu dördünün ve bunların ima ettiği alt unsurların tam olarak sağlandığından emin isek bir "yazılım güvenliği" söz konusu olabilir.

1. Şifreleme (kriptografi),
2. Yetkilendirme,
3. Erişim kontrol listeleri,
4. Bütünlük kontrol yöntemleri,
5. Veritabanı hash'leri,
6. Kompartıman içinde çalıştırma,
7. Veri girdi kontrolleri,
8. Bellek kontrolleri,
9. Kimlik kontrolleri,
10. Kullanılabilirlik,
11. Kurtarılabilir veri saklama yöntemleri,
12. Güvenli yazılım kodlama teknikleri.

Bu anlamda yazılım güvenlik denetimi yalnızca yazılım kod denetimi olarak anlaşılamaz. Yazılım ile ilişkili her nesne bir tehdit unsuru adledilir ve buna göre güvenli hale getirilir ya da hareket alanı kısıtlanır.

Web tabanlı yazılımlar diğer yazılımlara göre güvenlik açısından çok daha büyük tehlike altındadırlar zira yazılım ve kaynaklar erişime kısmen de olsa açıktırlar.

2.2.6. İstatistik - Raporlama Modülü

İstatistik ve raporlama modülleri bir yazılımın çıktılarıdır. Çıktılar ise yazılımın ve yazılımı kullanan birimin faaliyetlerini belgeler. Alınan raporlar sayesinde kullanıcılar ileriye yönelik kestirimlerde bulunabilirler.

BÖLÜM 3

OPTİK KARAKTER TANIMA (OKT) SİSTEMLERİ

İmge tanıma teknolojilerinin alanlarından birisi olan OKT sistemi kâğıda basılı metinlerin ya da bir kamera görüntüsünün taranıp otomatik olarak bilgisayar ortamına aktarılması kullanıcıların zaman sarfiyatını azaltmak açısından çok önemlidir. OKT, bilgisayar aracılığıyla, basılı metinlerin okunarak dijital ortama taşınması işlemidir. Bu işlem basılı metinler üzerinde çalışan birimler için vazgeçilmez bir teknolojidir. Doküman işleme işlemleri ile çalışırken maliyetleri düşürmek ve en yüksek düzeyde verim elde etmek için OKT teknolojilerinden faydalanmak kaçınılmaz olmuştur.

OKT teknolojisi sıklıkla aşağıdaki alanlarda kullanılmaktadır;

1. Kütüphanelerde,
2. Masa üstü yayıncılıkta,
3. Adliyelerde,
4. Vergi daireleri ve tahsilatta,
5. Personel kayıt yönetiminde,
6. Nüfus sayımı formlarının işlenmesinde,
7. Emeklik fonu işlemlerinde,
8. Sipariş işlemlerinde,
9. Hastanelerde.

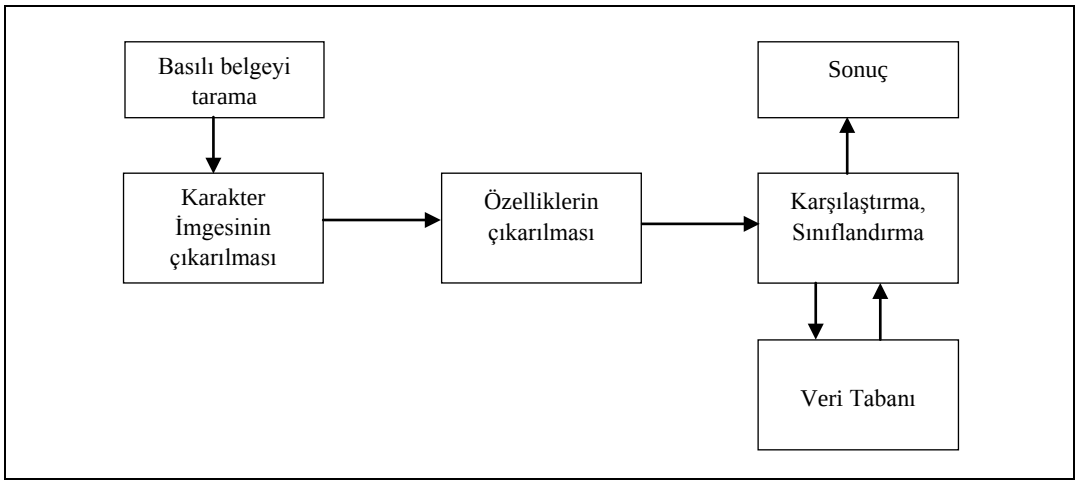
Literatürde karakter tanıma teknolojilerinin sağladığı kolaylıklardan bahsedilmektedir. Örneğin; mektupların üstlerindeki adreslerin tanınıp posta koduna göre otomatik olarak ayrıştırılması, bankalara yollanan çeklerin otomatik olarak tanınıp gerekli hesap işlemlerinin elektronik ortamlarda gerçekleştirilmesi gibi işlemler karakter tanıma teknolojileri ile gerçekleştirilmektedir (Şekerci, 2007).

Karakter tanımanın tarihi, bilgisayar tarihinden daha eskidir. İlk zamanlar tüm OKT sistemleri analog tabanlı yapılmaktaydı. Bunun nedeni o zamanlar analog/donanım teknolojisinin, dijital/yazılım teknolojisine göre daha çok gelişme göstermesiydi. 1962 senesinde, RCA, ilk olarak elektron tüp üzerine 91 kanallı, İngiliz ve Rus alfabesinin tüm harflerini tanıyabilen karmaşık bir OKT sistemi geliştirdi. Ama ticari amaçlı sistem geliştirilmedi (Parker, 1997).

OKT ya da OCR teknolojisi görme engelliler için de çok önemli bir yere sahiptir. Ray Kurzweil tarafından bilgisayar şirketi, görme engelliler için bir sistem geliştirmiştir. Sistem günümüz tarayıcılarından ve yazıyı sese dönüştüren bir teknolojiden oluşmaktadır.

3.1. OPTİK KARAKTER TANIMA SİSTEMİNİN GENEL YAPISI

OKT sistemleri, genel olarak basılı belgeyi tarar, taranan belgede bulunan metindeki karakter imgelerini ayırtır ve önceden belli olan karakter şekilleri ile karşılaştırma yaparak bu imgenin hangi karaktere ait olduğunu tespit eder. Sistem, önceden belli olan karakter imgelerini kendi veritabanında tutar. Optik Karakter tanıma sistemlerinin genel yapısı Şekil 3.1'de verilmiştir.



Şekil 3.1. OKT sisteminin genel yapısı.

Şekil 3.1'de Karakter imgesi modülünde, metnin taranması ve karakterlerin imgelerinin ayırtılarak sistem girişine verilmesi işlemleri yapılır. Sonraki adımda

karakter imgesinin şekilsel özellikleri çıkartılır ve veri tabanında olan karakter imgelerinin şekilsel özellikleri ile karşılaştırılarak hangi karakter olduğuna karar verilir. Her karakter, farklı ölçülerde ve şekillerde gösterilebilir. Karakterlerin karşılaştırılması, önceden sisteme tanımlı olan ve veri tabanında bulunan karakter imgelerine göre sınıflandırılması ile gerçekleşir. Bu işlemi yapan modüle sınıflandırıcı denir. Sistemin çıkışı, tanınan karakterin kodudur (Musayev, 2004).

Bu tez çalışmasında karakter tanıma yöntemi olarak, Microsoft Office'in Microsoft Office Document Imaging (MODI) *dll*'i kullanılmıştır. MODI.dll'i Visual Studio C# da hazırlanan uygulamada referans olarak eklenmiştir. Literatürde MODI.dll'in başarı sonuçlarına yönelik bir bulguya rastlanılmamıştır. Bölüm 6.3'te bu çalışma sonucunda elde edilen veriler paylaşılmıştır.

BÖLÜM 4

VERİ VE METİN MADENCİLİĞİ

4.1. VERİ MADENCİLİĞİ

Verilerin dijital ortamda saklanmaya başlanması ile birlikte, yeryüzündeki bilgi miktarının her yirmi ayda bir kendini iki katına çıkardığı günümüzde veri tabanlarının sayısı da benzer, hatta daha yüksek bir oranda artmaktadır. Yüksek kapasiteli işlem yapabilme gücünün ucuzlamasının bir sonucu olarak, veri saklama hem daha kolay olmuş, hem de verinin kendisi de ucuzlamıştır (Vahaplar ve İnceoğlu, 2001).

Veri madenciliği, büyük miktarlardaki verilerden fayda sağlayıcı bilgileri ortaya çıkararak, veriye anlam kazandırma işlemidir (Han and Kamber, 2006).

Veri madenciliği, diğer bir adla veritabanında bilgi keşfi; çok büyük veri hacimleri arasında tutulan, anlamı daha önce keşfedilmemiş potansiyel olarak faydalı ve anlaşılır bilgilerin çıkarıldığı ve arka planda veritabanı yönetim sistemleri, istatistik, yapay zekâ, makine öğrenme, paralel ve dağıtık işlemlerin bulunduğu veri analiz tekniklerine, veri madenciliği adı verilir (Berry and Linoff, 2000).

Fayyad et al. (1996) veri madenciliğini; kabul edilebilir etkinlik sınırlarına sahip bilgisayar teknikleriyle, veri üzerinden olağandışı örüntü ve model sıralamaları üreten süreç olarak tanımlanan, veritabanlarından bilgi keşfi sürecinin bir adımı olarak tanımlamışlardır.

Kütüphanelerin dijital ortama taşınma ve kaynaklarını internet üzerinden erişime açma çalışmalarında veri madenciliği ve elde edilen verilerin sınıflandırılmasının önemi her geçen gün artmaktadır.

4.1.1. Veri Madenciliğinde Bilgiyi Elde Etme (Keşfetme) Süreci

Büyük veritabanlarında ilginç ve değerli olan bilgiyi algılamak ve erişmek oldukça zordur. Veritabanında bilgi keşif sürecinin aşamaları (Knowledge Discovery in Databases) bu değerli, önceden bilinmeyen, kullanılabilir olan bilgiye belirli metotlar uygulayarak tanımlamada çok büyük rol oynamaktadır.

Veri madenciliği süreci aşağıdaki aşamalardan oluşur;

1. Uygulama alanının incelenmesi,
2. Amaca uygun veri kümesi oluşturma,
3. Veri temizleme,
4. Veri bütünleştirme,
5. Veri indirgeme,
6. Veri dönüştürme,
7. Veri madenciliği tekniği seçme,
8. Veri madenciliği algoritmasını uygulama,
9. Sonuçları değerlendirme.

4.1.1.1. Uygulama Alanının İncelenmesi

Öncelikle konuyla ilgili bilgi ve uygulama amaçların belirlenmesi gerekmektedir. Bu tez çalışmasında üniversite kütüphaneleri incelenmiştir. Kütüphanelerde kullanılan yazılımlar ve bu yazılımlarda dokümantasyon ve kataloglama işlemlerinin nasıl yapıldığı bu işlemlerde veri madenciliği alanında yapılan çalışmalar incelenmiştir.

4.1.1.2. Amaca Uygun Veri Kümesi Oluşturma

Analiz edilecek verinin hangi veritabanında yapılacağını belirterek, veri seçme ya da keşif edilecek alt veri örnekleri oluşturma. Bu tez çalışmasında web tabanlı kütüphane otomasyonu geliştirilerek bu otomasyonda dokümanların veri girişleri sağlanmıştır. Bu dokümanların tutulduğu veritabanında ilgili tablolardan veriler alınmıştır.

4.1.1.3. Veri Temizleme

Veri tabanında yer alan tutarsız ve hatalı verilere gürültü denir. Verilerdeki gürültüyü temizlemek için; eksik değer içeren kayıtlar atılabilir, kayıp değerlerin yerine sabit bir değer atanabilir, diğer verilerin ortalaması hesaplanarak kayıp veriler yerine bu değer yazılabilir, verilere uygun bir tahmin (karar ağacı, regresyon) yapılarak eksik veri yerine kullanılabilir.

4.1.1.4. Veri Bütünleştirme

Farklı veri tabanlarından ya da veri kaynaklarından elde edilen verilerin birlikte değerlendirmeye alınabilmesi için farklı türdeki verilerin tek türe dönüştürülmesi işlemidir. Bunun en yaygın örneği cinsiyette görülmektedir. Çok fazla tipte tutulabilen bir veri olup, bir veri tabanında 0/1 olarak tutulurken diğer veri tabanında E/K veya Erkek/Kadın şeklinde tutulabilir. Bilginin keşfinde başarı verinin uyumuna da bağlı olmaktadır.

4.1.1.5. Veri İndirgeme

Veri madenciliği uygulamalarında çözümlemeden elde edilecek sonucun değişmeyeceğine inanılıyorsa veri sayısı ya da değişkenlerin sayısı azaltılabilir. Veri indirgeme yöntemleri; veri sıkıştırma, örnekleme, genelleme, birleştirme veya veri küpü, boyut indirgemedir.

4.1.1.6. Veri Dönüştürme

Verinin kullanılacak modele göre içeriğini koruyarak şeklinin dönüştürülmesi işlemidir. Dönüştürme işlemi kullanılacak modele uygun biçimde yapılmalıdır. Çünkü verinin gösterilmesinde kullanılacak model ve algoritma önemli bir rol oynamaktadır.

Değişkenlerin ortalama ve varyansları birbirlerinden önemli ölçüde farklı olduğu takdirde, büyük ortalama ve varyansa sahip değişkenlerin diğerleri üzerindeki baskısı

daha fazla olur ve onların rollerini önemli ölçüde azaltır. Bu yüzden veri üzerinde normalizasyon işlemi yapılmalıdır.

4.1.1.7. Veri Madenciliği Tekniği Seçme

Sınıflandırma (classify), bağlantı kuralları (association rules), kümeleme (clustering) gibi tekniklerden hangisi ya da hangilerinin kullanılacağının seçimi yapılır. Bu tez çalışmasında sınıflandırma tekniği kullanılmıştır.

4.1.1.8. Veri Madenciliği Algoritmasını Uygulama

Veri hazır hale getirildikten sonra konuyla ilgili veri madenciliği algoritması ya da algoritmaları uygulanır. Seçilecek olan algoritma konu ve alana göre değişebilir.

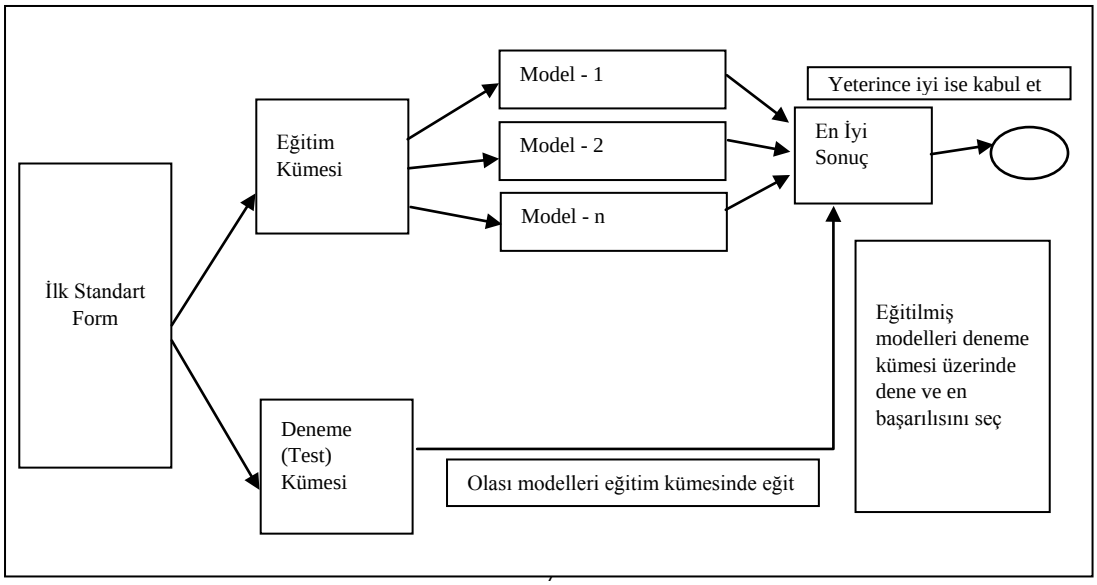
4.1.1.9. Sonuçları Değerlendirme

Uygulanan algoritma sonucunda elde edilen değerler raporlanır. Farklı algoritmalar ile elde edilen sonuçlar varsa bu sonuçlar karşılaştırılır ve nihai sonuca varılır.

4.1.2. Veri Madenciliği Metodolojisi

Bir veri madenciliği çalışmasında kullanılan metodoloji Şekil 4.1'de verilmiştir. Standart form içinde verilen veri, eğitim ve deneme olmak üzere ikiye ayrılır. Her uygulamada kullanılabilecek birden çok teknik vardır ve önceden hangisinin en başarılı olacağını kestirmek olası değildir. Bu yüzden öğrenme kümesi üzerinde L değişik teknik kullanılarak n tane model oluşturulur. Sonra bu n model deneme kümesi üzerinde denenerek en başarılı olanı, yani deneme kümesi üzerindeki tahmin başarısı en yüksek olanı seçilir.

Şekil 4.1'de kullanılan metodoloji de elde edilen en iyi sonuç başarılıysa kullanılır, değilse başa dönerek çalışma tekrarlanır (Çankırı vd., 2009).



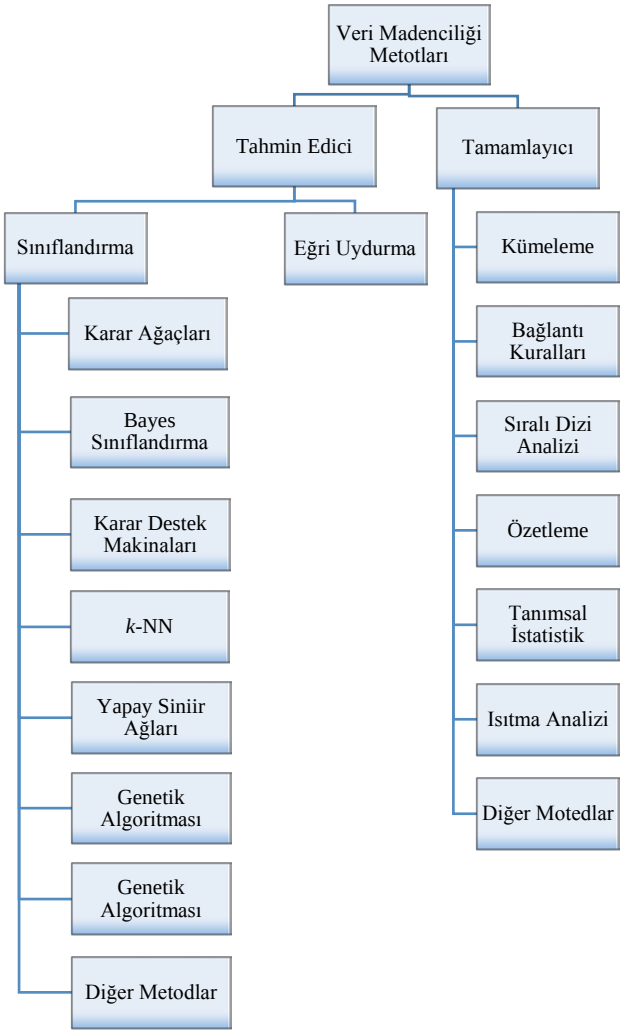
Şekil 4.1. Veri madenciliği çalışmasında kullanılan metodoloji.

4.1.3. Veri Madenciliği Metodları

Veri madenciliğinde kullanılan modeller, tahmin edici (Predictive) ve tanımlayıcı (Descriptive) olmak üzere iki ana başlık altında incelenmektedir (Zhong and Zhou, 1999).

Tahmin edici modellerde, sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçlan bilinmeyen veri kümeleri için sonuç değerlerin tahmin edilmesi amaçlanmaktadır (Özekes, 2003).

Tanımlayıcı modellerde ise karar vermeye rehberlik etmede kullanılabilecek mevcut verilerdeki örüntülerin tanımlanması sağlanmaktadır (Özekes, 2003). Şekil 4.2'de veri madenciliği metodları gösterilmektedir.



Şekil 4.2. Veri madenciliği metodları (Han and Kamber, 2006).

Bu tez çalışmasında tahmin edici model kullanılarak örnek bir kütüphanede kayıtlı dokümanların sınıfı belirlenmiştir. Dokümanlar sınıflandırılırken *k*-NN algoritması kullanılmıştır.

4.1.4. Veri Madenciliğinde Kullanılan Teknikler

Veri madenciliği teknikleri işlevlerine göre üç grupta toplanır (Akbulut, 2006);

1. Sınıflandırma,
2. Kümeleme,
3. Birliktelik kuralları ve sıralı örüntüler.

4.1.4.1. Sınıflandırma

Sınıflandırma veri madenciliğinin en çok kullanıldığı alandır. Var olan veri tabanının bir kısmı eğitim olarak kullanılarak sınıflandırma kuralları oluşturulur. Bu kurallar yardımıyla yeni bir durum ortaya çıktığında sistemin nasıl karar verileceği belirlenir.

Sınıflandırma modeli üç aşamadan oluşmaktadır (Bilekdemir, 2010);

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

Denetimli öğrenmede; öğreniciye nesneler ve nesnelerin özellikleri ve yine bu nesnelerin tanımlanmış, gelecek aşamalarda tahmini istenecek olan değişkenler verilmektedir. Denetimsiz öğrenmede nesnelerin özellikleri bilinirken, tahmin için kullanılacak olan nesnelerin isimleri verilmemektedir (Silahtaroğlu, 2008). Ayrıca denetimsiz öğrenmede herhangi bir organize olmadan, yöntem kendi yolunu bulabilmektedir.

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

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

Veri madenciliği, kayıtlı olan veriler arasındaki gizli ilişkilerin keşfedilmesi ve karar alma sürecinde kullanılan bir uygulamadır. Veritabanlarında bulunan gizli

örüntülerin çıkarılması için bir takım veri madenciliği teknik ve algoritmaları kullanılmaktadır. Model seçimi başlığında anlatılan modellerin kullandığı başlıca teknik ve algoritmalar aşağıda açıklanmaktadır.

Temel sınıflama algoritmaları aşağıdadır;

1. *k*-NN,
2. Naive Bayes,
3. Karar Ağaçları,
4. Yapay Sinir Ağları,
5. Genetik Algoritma,
6. Regresyon Analizi.

k-En Yakın Komşu (*k*-Nearest Neighbor, *k*-NN) Algoritması

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

k-NN algoritması, sınıflandırma problemini çözen denetimli öğrenme (sınıflandırma için öğrenme kümesi kullanır) algoritmalarından biridir. Sınıflandırma, yeni bir nesnenin özniteliklerini inceleme ve bu nesneyi önceden tanımlanmış bir sınıfa atamaktır. Burada önemli olan, her bir sınıfın özelliklerinin önceden net bir şekilde belirlenmiş olmasıdır.

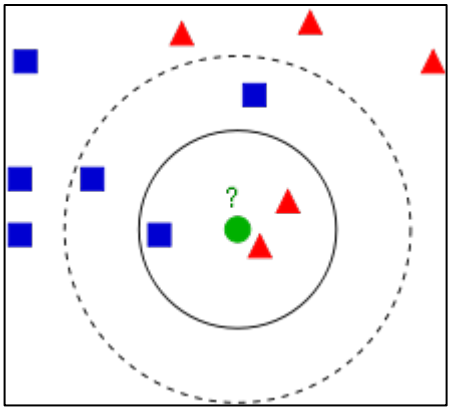
Dasarathy'e (1991) göre *k*-NN algoritması ile sınıflandırma, önceden belirlenmiş *k* değerine göre uzaklıkları hesaplanmış eğitim dokümanları içerisinden en yakın *k* dokümandaki en yüksek frekansa sahip sınıfa göre test dokümanının sınıfını belirleme işlemidir.

k-NN algoritmasında, sınıflandırılacak niteliğin, veritabanında eğitim kümesindeki her bir niteliğe olan uzaklığı ölçülür. En çok kullanılan mesafe ölçüsü öklid uzaklığı olsa da, başka herhangi bir ölçü de kullanılabilir. Tanımlanacak olan niteliğin, veri tabanında eğitim kümesindeki kendisine en çok benzeyen, niteliğin sınıfından sayılır ve bu şekilde tanımlanır (Türkoğlu, 1996).

k-NN algoritması, daha matematiksel olarak şöyle ifade edilebilir; *n* boyutlu R^n alanı içindeki bütün noktaları duruma uyduğunu varsayar bütün örnekler *n* boyutlu bir uzayda bir noktaya karşı düşürülür. Verilen noktaya göre en yakın komşular Öklid mesafesi kullanılarak hesaplanır. *a1(x),a2(x)….,an(x)* örnek noktalar olarak tanımlanır. İki örnek arasındaki uzaklık *d(xi,xj)* olmak üzere eşitlik 4.1'deki gibi tanımlanır (Mitchell, 1997).

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n}(a_r(x_i) - a_r(x_j))^2} \tag{4.1}$$

Bütün eğitim dokümanları ile test dokümanlarının uzaklıkları tek tek hesaplanır ve belirlenecek *k* değerine göre sınıflandırma sonucuna karar verilir. Şekil 4.3'te görüldüğü gibi *k=3* alındığında dokümanın sınıfı üçgen olacakken *k=5* alındığında kare olmuştur. *k* değerinin yüksek seçilmesi benzemeyen dokümanların işleme dahil edilmesine, düşük seçilmesi benzeyenlerin dahil edilmemesine neden olur.



Şekil 4.3. *k*-NN sınıflandırma örneği.

Naive Bayes Algoritması

Naive Bayes Kolay uygulanabilir olduğu kadar üstün performansıyla da metin sınıflandırma çalışmalarında en çok kullanılan metotlardan biri haline gelmiştir. Metotta önce tüm eğitim verisindeki metinlerde kullanılan kelimelerden bir sözlük oluşturulur. Daha sonra her bir kelimenin her bir sınıftaki tekrar sayıları (frekansı) bulunur. Sınıflandırılması istenen yeni bir metin önceden geldiğinde oluşturulan sözlükte var olan kelimelerin her bir sınıftaki frekansları bulunur. Bir metnin *C* sınıfına dâhil olma olasılığı *C* sınıfının eğitim setindeki oranıyla, metnin içindeki her bir kelimenin *C* sınıfına ait olma olasılıkları çarpılarak bulunur (Amasyalı ve Yıldırım, 2004).

Karar Ağaçları

Yapay sinir ağlarında veriden bir fonksiyon öğrenildikten sonra bu fonksiyonun insanlar tarafından anlaşılabilecek bir kural olarak yorumlanması zordur. Karar ağaçlarında, ağaç oluşturulduktan sonra kökten yaprağa doğru inilerek kurallar (IF-THEN rules) yazılabilir (Mitchell, 1997). Bu şekilde kural çıkarma (rule extraction), veri madenciliği çalışmasının sonucunun doğrulanmasını sağlar. Bu kurallar uygulama konusunda uzman bir kişiye gösterilerek sonucun anlamlı olup olmadığı denetlenebilir. Sonradan başka bir teknik kullanılacak bile olsa, karar ağacı ile önce bir kısa çalışma yapmak, önemli değişkenler ve yaklaşık kurallar konusunda bize bilgi verir ve önerilir.

Yapay Sinir Ağları

1980'lerden sonra yaygınlaşan yapay sinir ağlarında (artificial neural networks) amaç fonksiyon birbirine bağlı basit işlemci ünitelerinden oluşan bir ağ üzerine dağıtılmıştır (Bishop, 1996). Yapay sinir ağlarında kullanılan öğrenme algoritmaları veriden üniteler arasındaki bağlantı ağırlıklarını hesaplar. YSA istatistiksel yöntemler gibi veri hakkında parametrik bir model varsaymaz yani uygulama alanı daha geniştir ve bellek tabanlı yöntemler kadar yüksek işlem ve bellek gerektirmez.

Genetik Algoritmalar

Diğer veri madenciliği algoritmalarını geliştirmek için kullanılan optimizasyon tekniklerdir. Sonuç model veriye uygulanarak gizli kalmış kalıpları ortaya çıkarılmakta ve bu sayede tahminler yapılabilmektedir. Doğrudan postalama, risk analizi ve perakende analizlerinde kullanılabilir.

4.1.4.2. Kümeleme

Kümeleme, verideki benzer kayıtların gruplandırılmasını sağlayan bir tekniktir. Kümelemede, genellikle *k*-ortalamalar algoritması ya da Kohonen şebekesi gibi istatistiksel yöntemler kullanılmaktadır. Hangi yöntem kullanılırsa kullanılsın süreç aynı şekilde isler. Her kayıt var olan kümelerle karşılaştırılır. Bir kayıt kendisine en yakın kümeye atanır ve bu kümeyi tanımlayan değeri değiştirir. Optimum çözüm bulununcaya kadar kayıtlar yeniden atanır ve küme merkezleri ayarlanır (Hui and Jha, 2000).

4.1.4.3. Birliktelik Kuralları ve Ardışık Örüntüler

Birliktelik analizi, bir veri kümesindeki kayıtlar arasındaki bağlantıları arayan denetimsiz veri madenciliği şeklidir. Birliktelik analizi çoğu zaman perakende sektöründe süpermarket müşterilerinin satın alma davranışlarını ortaya koymak için kullanıldığından "pazar sepeti analizi" olarak da adlandırılır (Hui and Jha, 2000).

Birliktelik kurallarına ait bir örnek: "Düşük yağlı peynir ve yağsız süt alan müşteriler % 85 olasılıkla diyet süt alırlar."

Ardışık analiz ise birbiriyle ilişkisi olan ancak birbirini izleyen dönemlerde gerçeklesen ilişkilerin tanımlanmasında kullanılır. Aşağıda ardışık analize ait örnekler yer almaktadır.

1. "Çadır alan müşterilerin % 10'u bir ay içerisinde sırt çantası almaktadır"
2. "A hissesi % 15 artarsa üç gün içinde B hissesi % 60 olasılıkla artacaktır"

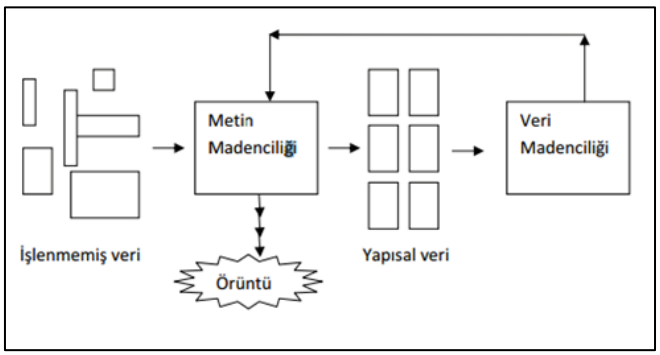
4.2. METİN MADENCİLİĞİ

İnternet kullanımının yaygınlaşması, internet ortamında o denli veri olduğu anlamına gelir. Bu verilerin veri madenciliği teknikleriyle incelenmesi yapısal veri olmadıklarından mümkün değildir. Metin veri madenciliği, metin koleksiyonlarından bilgiye erişen, bireysel metinlerden bilgi çıkaran, veritabanlarından bilgi keşfeden, organizasyonlarda bilgi yönetimini ve veri ile bilginin görselleştirilmesi aşamalarını birleştiren bir mimaridir (Losiewicz et al., 2000).

Bu bölümde metin madenciliği metotları incelenerek yapısal olmayan bir veri türü olan metinlerden bilgi çıkarma konusunda farklı yaklaşımlar incelenecek, çalışma prensipleri açıklanacaktır. Veri madenciliği, eldeki verilerden çok net olmayan, önceden bilinmeyen ancak potansiyel olarak kullanışlı bilginin çıkarılması yaklaşımıdır.

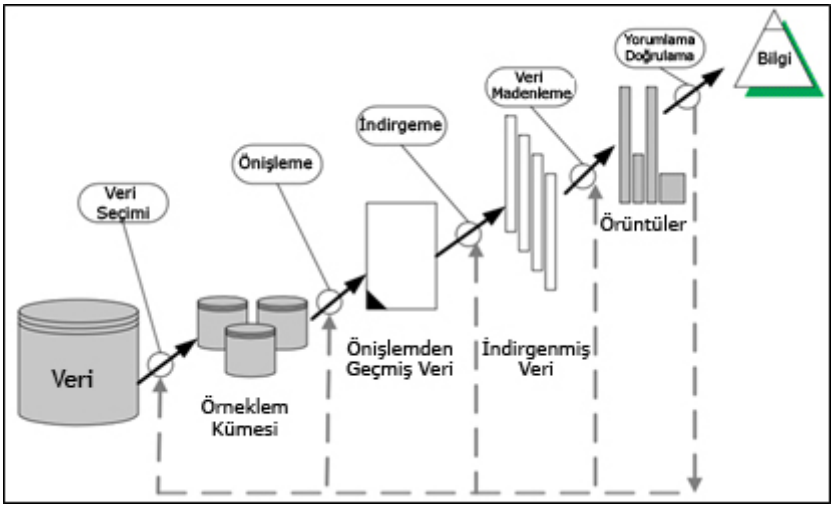
Veri madenciliğinin alt dalı olarak ele alınan metin madenciliği ise yazılmış farklı dokümanlardan yeni, önceden bilinmeyen bilgilerin bilgisayar tarafından otomatik bir şekilde keşfedilmesidir.

Şekil 4.4'te görüldüğü gibi, metin ve veri madenciliği arasında interaktif bir ilişki vardır. Metin madenciliği sonucunda elde edilene yapısal veri, veri madenciliği modellerinde kullanılmakta ve elde edilen sonuçlar daha sonra metnin yapısının incelenmesinde kullanılmaktadır.



Şekil 4.4. Metin madenciliğinde süreçler arasındaki ilişki.

Metin madenciliğini veri madenciliğinden ayıran en büyük fark metin madenciliğinde kalıpların düzgün veritabanlarından çok, doğal dil metinlerinden çıkarılmasıdır (İlhan vd., 2008).



Şekil 4.5. Veritabanında bilgi keşfi süreci.

Şekil 4.5'te verilen Veritabanında bilgi keşfi sürecinde aşağıdaki işlemler gerçekleştirilir.

1. Veri Seçimi (Data Selection) : Bu adım birkaç veri kümesini birleştirerek, sorguya uygun örneklem kümesini elde etmeyi gerektirir.
2. Veri Temizleme ve Önişleme (Data Cleaning & Preprocessing) :Seçilen örneklemde yer alan hatalı tutanakların çıkarıldığı ve eksik nitelik değerlerinin değiştirildiği aşamadır ve keşfedilen bilginin kalitesini artırır.
3. Veri İndirgeme (Data Reduction) : Seçilen örneklemden ilgisiz niteliklerin atıldığı ve tekrarlı tutanakların ayıklandığı adımdır. Bu aşama ile seçilen veri madenciliği sorgusunun çalışma zamanını iyileştirir.
4. Veri Madenciliği (Data Mining) : Verilen bir veri madenciliği sorgusunun (sınıflama, güdümsüz öbekleme, eşleştirme, vb.) işletilmesidir.
5. Değerlendirme (Evaluation) : Keşfedilen bilginin geçerlilik, yenilik, yararlılık ve basitlik kıstaslarına göre değerlendirilmesi aşamasıdır.

Çizelge 4.1. Metin madenciliği işlemleri.

	Metin Önişleme	Metin Dönüşümü	Özellik Seçimi	Veri Madenciliği, Bilgi Keşfi	Yorum, Değerlendirme
METİN	Söz dizimsel, Semantik analiz Sözcük türü etiketleme Kelime anlamı belirginleştirme Ayrıştırma (parsing)	Kelime torbası Kelimeler Kök bulma, Etkisiz kelimeler	Basit hesaplama İstatistik (boyut azaltma, ilişkisiz özellikler)	Sınıflandırma (Danışmanlı) Kümeleme (Danışmansız)	Analiz Sonuçları

Zohar'a (2002) göre metin madenciliği işlemleri Çizelge 4.1'de gösterilmiştir. Veri madenciliğinde analiz edilecek giriş verilerinin belirli bir formata sahip olması ayrıca bozuk veya gereksiz verilerden temizlenmiş olması gerekmektedir. Metin madenciliğinin en büyük sorunu, işleyeceği veri kümesinin yapısal olmamasıdır. Genellikle doğal dil kullanılarak yazılmış dokümanlar üzerinde çalışılan metin madenciliği alanında ön işleme aşaması, veri temizlemenin yanında veriyi uygun formata getirme işlemini de gerçekleştirmektedir (Feldman and Sanger, 2007).

Türkçe yapı bakımından sondan eklemeli dildir. Yapılan çalışmalarda kelimenin kendisi yerine kelime kökü kullanılmıştır. Kelime kökleri ile çalışılmazsa aynı kelimeyi temsil eden farklı çekim eki almış kelimeler, farklı kelimeler gibi değerlendirilir. Böyle olunca hem sözcük hem vektör boyutu artmış olur hem de uygulamalarda yanlış sonuçlar elde edilir. Çizelge 4.2'de aynı köke sahip çekim eki almış kelimeler görülmektedir.

Çizelge 4.2. Aynı köke sahip kelimeler örneği.

Kelime	**Kök**
Gel	Gel
Geldim	Gel
Gelmiştim	Gel
Gelmedim	Gel
Gelemedim	Gel

BÖLÜM 5

DOKÜMAN SINIFLANDIRMA

Doküman sınıflandırmadaki amaç, bir dokümanın özelliklerine bakılarak önceden belirlenmiş belli sayıdaki kategorilerden hangisine dâhil olacağını belirlemektir. Doküman sınıflandırma bilgi alma (information retrieval), bilgi çıkarma (information extraction), doküman indeksleme, doküman filtreleme, otomatik olarak metadata elde etme ve web sayfalarını hiyerarşik olarak düzenleme gibi pek çok alanda önemli bir rol oynamaktadır.

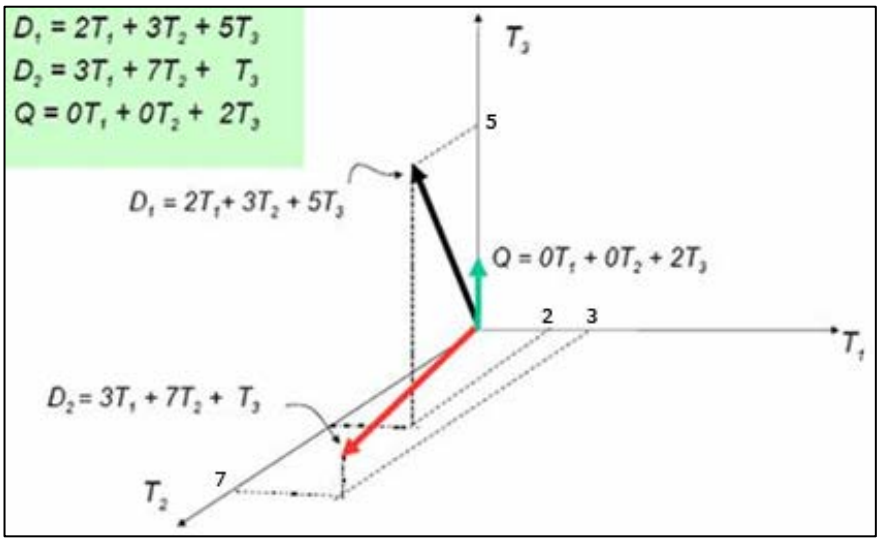
Yaygın olarak kullanılan sınıflandırma yöntemleri; *k*-NN, Naive Bayes, Karar Ağaçları, Maksimum Entropi Modelleri, Bulanık Mantık Teorisi Yaklaşımları, Destek Vektör Makineleri ve Yapay Sinir Ağlarıdır.

Sınıflandırma ile ilgili çalışmalarda *k*-NN ve Naive Bayes'in bit ağırlıklandırma kullanılarak yapılan metin sınıflandırma işleminde, *k*-NN'in kosinüs benzerliği ile birlikte uygulandığında, Naive Bayes'ten daha başarılı olmuştur (Soucy and Mineau, 2001). Bu tez çalışmasında sınıflandırma algoritması olarak *k*-NN kullanılmıştır.

5.1. VEKTÖR UZAY MODELİ

Vektör uzay modeli bilgi çıkarımı, bilgi filtreleme, indeksleme gibi alanlarda kullanılan cebirsel bir modeldir. Doğal dil belgelerinin çok boyutlu uzayda özel bir anlamını simgelemektedir. Vektör uzay modelinde her nesne, vektör yapısında tanımlanmaktadır. Nesnelerin sahip olduğu farklı özellikler, vektör uzayının eksenlerini oluşturmakta ve her nesne sahip olduğu özelliklere göre vektör uzayında belli bir konuma sahip olmaktadır. Metin madenciliğinde analiz edilecek metinlerin özellikleri içinde geçen kelimeler ya da parçalardır. Tüm kelime ya da parçaların toplam kümesi vektör uzayını gösterir. Vektör uzayı içerisinde sahip oldukları

noktaların birleşimi dokümanların vektörünü oluşturur ve vektörler arasında kalan açı dokümanların birbirlerine yakınlığını gösterir. Metin madenciliğinde kelime veya parçalar ağırlıklandırılır ve dokümanlar bu ağırlıkların vektörü tarafından gösterilir. Vektör uzayının boyutları, kelime veya parçalardır.



Şekil 5.1. Kelimelerin vektörel gösterimi.

Dokümanlar Şekil 5.1'de görüldüğü gibi kelimelerin vektörleri olarak ifade edilirler. T'ler aslında kelimeleri ifade etmektedirler (Jun ve Hokuan, 2002).

5.2. AĞIRLIKLANDIRMA YÖNTEMLERİ

Bir dokümanı diğer dokümanlardan ayıran içeriğidir. Dokümanın içeriğini belirleyen sözcüklerin etkisi sayısal değerlerle ifade edilirler. Dokümanların vektörlerinde kullanılacak değerler üç farklı yöntemle bulunabilir. Bu Yöntemler aşağıda listelenmiştir, yöntemler ile ilgili örnekler Çizelge 5.1'deki metinlere göre verilmiştir.

1. Bit,
2. Frekans,
3. *tf-idf.*

Çizelge 5.1. Örnek metinler.

ID	Kelimeler
1	Gribe yakalanan hasta grip olduğunu anlamamıştı. İlacını almamıştı.
2	İlacını aksatanlar hastalığa davetiye çıkarırlar.
3	Yıllık enflasyon oranı bu senede yükselişte
4	Tarımla uğraşanlar bu yıl tarımdan zarar edecekler.
5	Hakemin gözü önünde olmasına rağmen hakem penaltı çalmadı. (Spor)
6	Taraftarlara erken gelen gol ilaç gibi geldi ve taraftarlar golden sonra hiç susmadı. (Spor)

5.2.1. Bit Ağırlıklandırma Yöntemi

Anahtar sözcük sözlüğünde yer alan sözcüklerin metinde yer alıp almadığı gösteren vektörel bir gösterge oluşturulmaktadır. Çizelge 5.2'de örnek metinler için oluşturulmuş olan sözlük ve metinlerin anahtar kelimelere göre bitsel tanımlamaları aşağıda görülmektedir.

Sözlük= {enflasyon, grip, hakem, ilaç, taraftar, tarım}

D1=(0,1,0,1,0,0)
D2=(0,0,0,1,0,0)
D3=(1,0,0,0,0,0)
D4=(0,0,0,0,0,1)
D5=(0,0,1,0,0,0)
D6=(0,0,0,1,1,0)

Bu yöntemde vektörün ağırlıkları dokümanda bulunma veya bulunmamasına göre belirlenir. Yani bir kelimenin bir dokümanda 2 kere bulunması ağırlığını değiştirmeyecektir. Kelimenin dokümanda olması 1, olmaması 0 ile ifade edilmelidir.

5.2.2. Frekansa Göre Ağırlıklandırma Yöntemi

Sözcüklerin metinlerde kaç defa kullanıldığına dayanan bir yöntemdir. Örnek metinler için oluşturulmuş olan sözlük ve metinlerin anahtar kelimelere göre frekans tanımlamaları aşağıda görülmektedir.

Sözlük= {enflasyon, grip, hakem, ilaç, taraftar, tarım}

D1=(0,2,0,1,0,0)
D2=(0,0,0,1,0,0)
D3=(1,0,0,0,0,0)
D4=(0,0,0,0,0,2)
D5=(0,0,2,0,0,0)
D6=(0,0,0,1,2,0)

Bu yöntemle vektörlerin ağırlıkları dokümanda kelimenin bulunma sayısına göre hesaplanır. Örneğin bir kelime bir dokümanda 2 kere bulunuyorsa ağırlığı 2 alınmalıdır.

5.2.3. *tf-idf* Ağırlıklandırma Yöntemi

tf-idf, dokümanları vektör uzay modelinde tanımlayabilmemiz için kullanılan en önemli ağırlıklandırma metotlarından biridir. *tf-idf* ağırlıklandırmasında her bir dokümandaki sözcüklerin frekansı önemlidir. *tf* ağırlıklandırma yönteminde dokümanda daha fazla geçen (*tf* değeri büyük sözcükler) o doküman için daha değerli olmaktadır.

idf ağırlıklandırmada ise tüm dokümanlarda seyrek geçen sözcükler ile ilgili bir ölçü vermektedir. Bu değer tüm eğitim dokümanları ele alınarak hesaplanmaktadır. Bu yüzden eğer bir sözcük dokümanlarda sık geçiyorsa, o doküman için belirleyici olmadığı düşünülebilir. Eğer sözcük dokümanlarda çok sık geçmiyorsa o sözcüğün o doküman için belirleyici özelliği olduğu kabul edilebilir.

tf-idf genel olarak sorgu vektörü ile eğitim dokümanı vektörü arasındaki benzerlik oranını bulmak için kullanılır.

Vektörde kullanılacak ağırlık aşağıdaki gibi hesaplanabilir.

idfi = log(TDs/dfi)

*wi = tfi*Idfi*

i : Kelime ya da parçanın indisi

TDs : Toplam Doküman sayısı

tfi : Dokümanda bulunan kelimenin dokümandaki görülme sıklığı

dfi : Kelimenin ya da parçanın geçtiği doküman sayısı

wi : Kelime ya da parçanın vektörde kullanılacak ağırlığı

Örneğin:

D1 : Her gün işe 8'de gidiyorum.

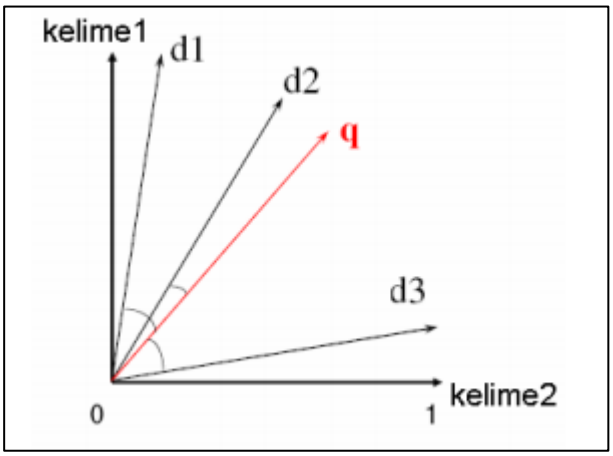
D2 : İş günleri çok yoruluyorum.

Sözcük Tablosu: "iş", "gün", "git" olsun;

D1 vektörü: 1*log(2/2)+1*log(2/2)+1*log(2/1) = 0 + 0 + 0,257

D2 vektörü: 1*log(2/2)+1*log(2/2)+0*log(2/1) = 0 + 0 + 0

Tüm dokümanlar ve yeni sınıflandırılacak doküman vektör uzay modeli kuralları doğrultusunda vektörel olarak ifade edilirler. Her bir boyut aslında kelimelere karşılık gelmektedir.



Şekil 5.2. Vektör uzay modelinde dokümanların gösterimi.

Şekil 5.2'de gösterilen d1, d2 ve d3 eğitim dokümanlarımızdan oluşan vektörler, *q* ise sınıfını bulmak istediğimiz vektördür.

5.3. ANAHTAR SÖZCÜK SEÇİMİ

Dokümanlar arasında kullanımı az olan ve dokümanların ayırt edilebilmesini sağlayacak özellikteki sözcükler, anahtar sözcük olarak nitelendirilmektedirler. Dokümanlar arasında sadece bir dokümanda geçen sözcük veya sözcükler, o doküman için en verimli anahtar sözcükler olarak kabul edilir. Bu anahtar sözcükler kullanılarak yapılan sorgu işlemleri, elemanı oldukları dokümana ulaşmanın en güçlü yolu olacaktır.

Anahtar sözcük seçimi çalışmanın en önemli noktasını oluşturmaktadır. Anahtar sözcük seçimi süreci; ön işleme aşamasındaki anahtar sözcük olamayacak belirli sözcüklerin metinden çıkarılması ile başlamakta, vektör oluşturma aşamasındaki gövde bulma ve aynı gövdeye sahip olan sözcüklerin aynı anahtar sözcük olarak kabul edilmesiyle devam etmektedir.

5.4. BENZERLİK HESAPLAMA

Vektör uzay modelinde kosinüs benzerliği, dokümanlar ve sorgular arasındaki benzerliği hesaplamak için kullanılır. Vektörler arasındaki gerçek açıları hesaplamak yerine gerçek açıların kosinüsleri hesaplanır ve karşılaştırılır. Kosinüs benzerliği, *n* boyutlu iki vektör arasındaki benzerliği iki vektör arasındaki açının cos ile ifade eder ve sıklıkla doküman karşılaştırmasında kullanılmaktadır. *A={a1,a2,a3,……..,an}* ve *B={b1,b2,b3,……..,bn}* vektörlerinin kosinüs benzerlik değeri *s*, eşitlik 5.1'de gösterildiği gibi, *A* ve *B*'nin skaler çarpımının, *A* ve *B*'nin mutlak değerinin çarpımına bölünmesi ile elde edilir. Doküman benzerliğinde *A* ve *B* vektörlerinin özellikleri dokümanlarda geçen word-gramların frekanslarıdır. Bir dokümanda bulunan diğerinde bulunmayan özellik için bulunmayan dokümanda o özellik 0 olarak alınır.

$$s = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2}\sqrt{\sum_{i=1}^{n} (B_i)^2}} \tag{5.1}$$

Eşitlik 5.2'de, vektörlerin skaler çarpımlarını ifade etmektedir. Eğer bir sözcük bir vektörde var diğer vektörde yoksa skaler çarpım sonucu o sözcük için 0'dır.

$$S\_C = \sum_{i=1}^{n} A_i B_i \tag{5.2}$$

Kosinüs benzerliği, sözcüğün vektörde varlığını araştırır. İki vektörde de bulunan sözcük değerlendirmeye alınır.

BÖLÜM 6

UYGULAMA

Günümüzde kütüphane ve bilgi hizmetlerinde, geçmiş dönemlerdeki öngörülen bazı fikirlerinde ötesine geçilmiştir. Külcü farklı araştırmacılara dayandırdığı çalışmasında, 2000'li yıllarda kütüphane ve bilgi hizmetlerinde beklenen gelişmeleri, günümüzden yaklaşık on yıl önce, genel olarak aşağıdaki maddeler halinde derlemiştir (Külcü, 2001);

1. Elektronik bilgiye erişimi sağlama,
2. Bilginin veya kaynağın bireysel olarak "self-servis" sağlanması, büyük boyutlu ve hantal kütüphanelerin ortadan kaldırılması,
3. Uzaktan erişim olanaklarını geliştirilmesi ve uzaktan eğitimi destekleyen kütüphaneler oluşturma,
4. Personel eğitiminde elektronik bilgi kaynakları konusunun ağırlık kazanması,
5. Elektronik ortamda depolanan dokümanların yüksek oranda basımı ve tekrar üretimine yönelik merkezi bir yerleşim olma,
6. Veritabanlarına eklenecek materyaller için kullanıcı profili oluşturma ve hizmet stratejilerinin belirlenmesi,
7. Bilgiye ve belgeye olabildiğince ucuza erişimi sağlamak için hizmet politikalarının belirlenmesi ve uygulanması,
8. Kaynak türleri arasındaki entegrasyon sağlanarak veritabanlarının oluşturulması,
9. Var olan kaynaklardan kullanıcı istek ve ihtiyaçlarına göre bilginin veritabanlarından çekilmesi ve hizmete sunulması.

Bu başlıkların tamamı kütüphane ve bilgi hizmetlerinin yürütülmesinde bugün itibariyle yerine getirilen hizmetlerdir. Ancak kütüphane ve bilgi merkezlerinin geldiği noktada sadece bu yeniliklerle sınırlı kalınmamıştır.

Örneğin, gelişen teknolojilerle birlikte kütüphaneler mobil hizmetler üretmeye başlamışlardır. Böylece zaman ve yer sınırlamaları ortadan kalkmış, aynı zamanda donanım ihtiyaçları (bilgisayar, telefon kablosu, modem vb.) azalarak kullanım yaygınlaşmıştır.

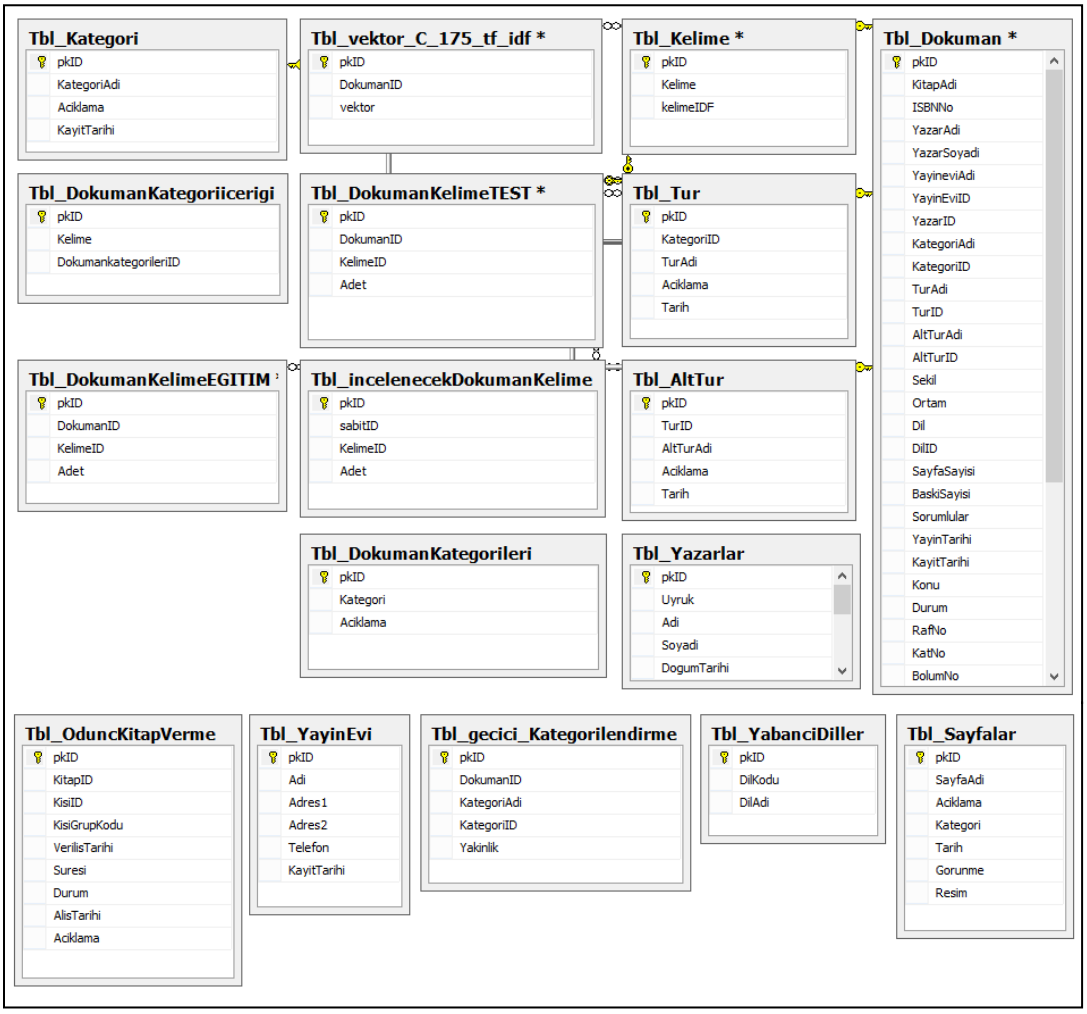
Kütüphaneler sundukları hizmetlerin bütünlüğüne göre, birbirini tamamlayan farklı iş süreçlerinden oluşmaktadır. Süreçler birbirini destekler nitelikte olmalıdır. Bu amaç doğrultusunda bütünleşik kütüphane otomasyon sistemleri geliştirilmiştir. Bütünleşik kütüphane otomasyon sistemi terim olarak bir sistem içinde birçok görevin yerine getirilmesini ifade eder (Takçı ve Soğukpınar, 2001).

Kütüphane otomasyon sistemlerinde belirlenen görevleri yerine getirmek üzere çeşitli modüller oluşturulmaktadır. Bu modüller için beklenen genel özellikler şu şekilde sıralanabilir;

1. Kullanım kolaylığı,
2. Teknik yardım,
3. Yüksek verim,
4. Amaca uygunluk,
5. Memnuniyet,
6. Anlaşılır dil,
7. Kullanıcı dostu (User friendly),
8. Teknik donanım.

6.1. SİSTEM VERİTABANI MODELİ

Çalışmada kullanılan veritabanının yapısı, tablolar ve ilişkileri Şekil 6.1'de verilmiştir.



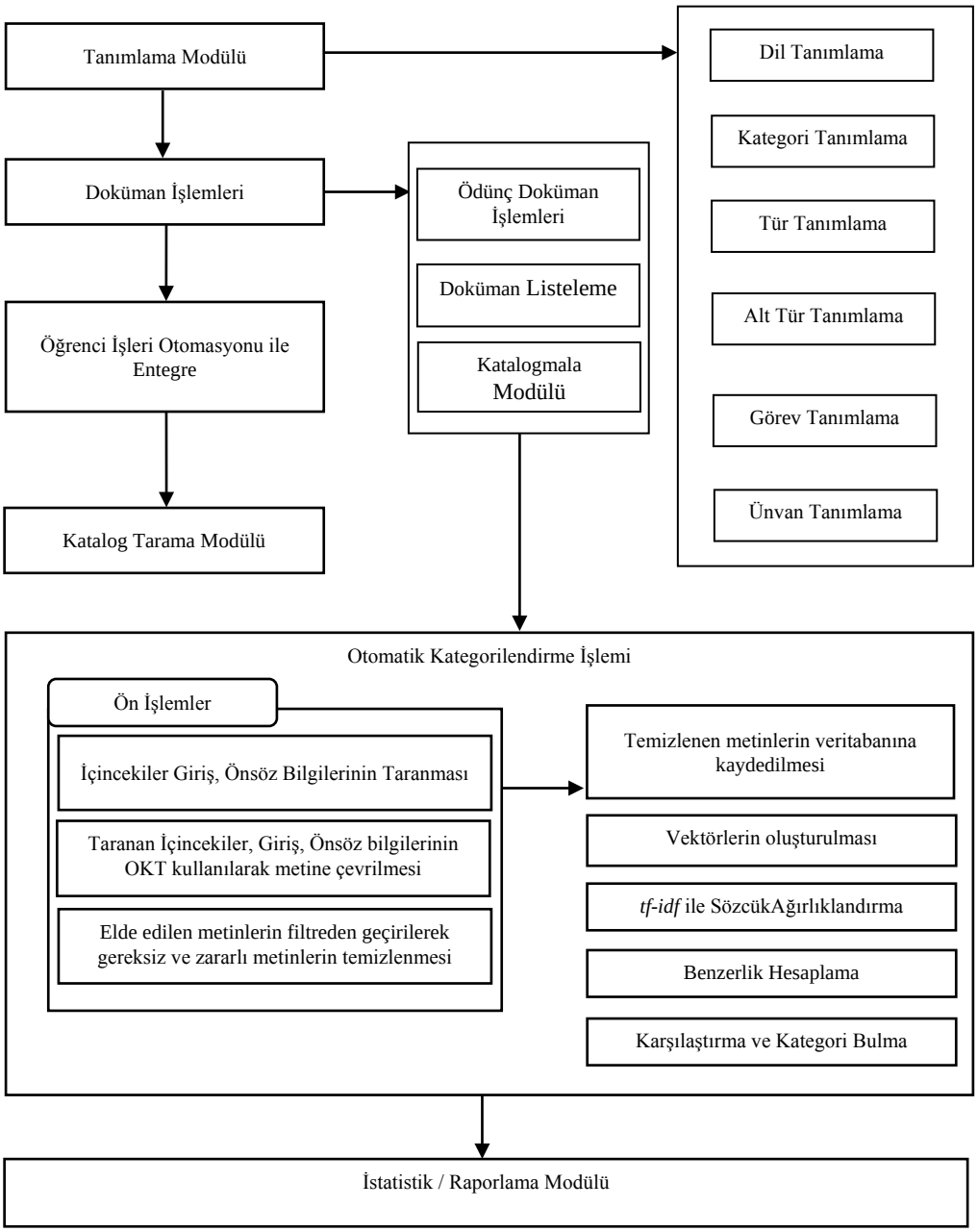
Şekil 6.1. Veritabanı yapısı.

6.2. KÜTÜPHANE OTOMASYONU GENEL YAPISI

Bu tez çalışmasındaki Kütüphane otomasyonu aşağıda belirtilen modüllerden oluşmaktadır. Sistemin yapısı Şekil 6.2'de gösterilmiştir.

1. Tanımlama Modülü,
2. Doküman İşlemleri Modülü,

3. Kataloglama - Kategorilendirme Modülü,
4. Kullanıcı Katalog Tarama Modülü,
5. Yönetici Katalog Tarama – Doküman Güncelleme,
6. Ödünç Verme Modülü,
7. Güvenlik Modülü,
8. İstatistik Modülü.



Şekil 6.2. Sistemin yapısı.

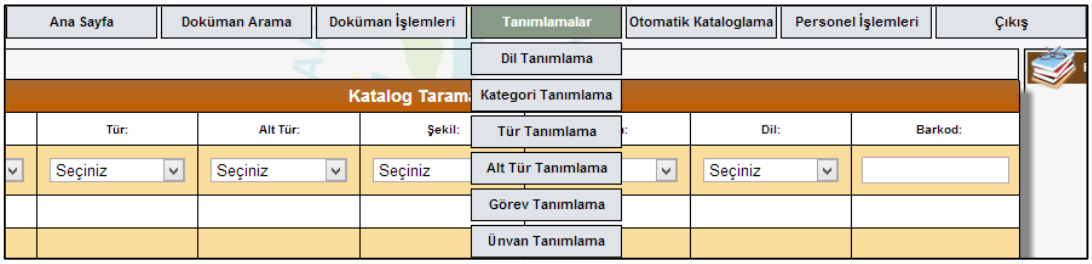
Otomasyon sisteminin giriş ekranı Şekil 6.3'te, giriş yapan kullanıcıya gösterilen sayfa da Şekil 6.4'te gösterilmiştir.

Şekil 6.3. Kütüphane otomasyonu giriş sayfası (Giris.aspx).

Şekil 6 4. Kütüphane otomasyonu ana sayfa *(Anasayfa.aspx)*.

6.2.1. Tanımlama Modülü

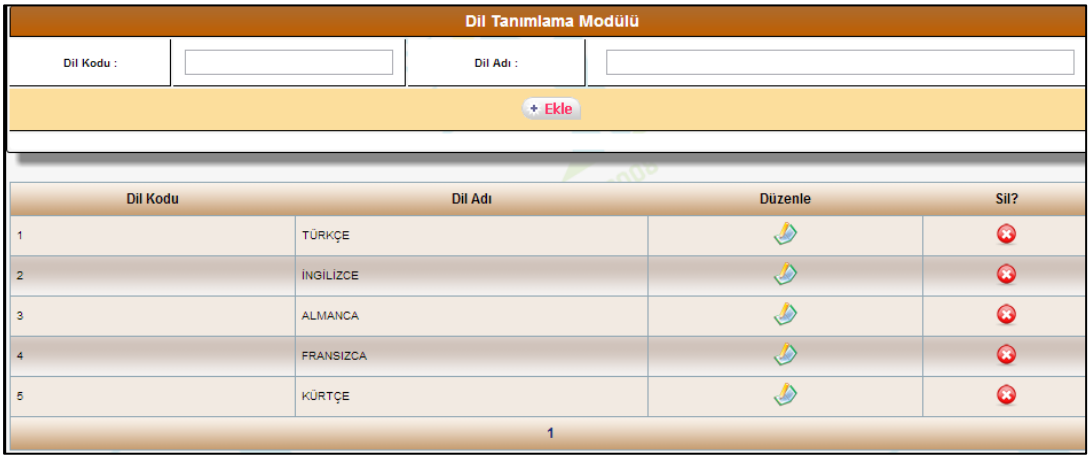
Bu bölümde Kataloglama Modülü için gerekli olan bilgilerin önceden tanımlanması sağlanmaktadır. Tanımlama modülünün menüsü Şekil 6.5'te gösterildiği gibidir.



Şekil 6.5. Tanımlama modülü.

6.2.1.1. Dil Tanımlama

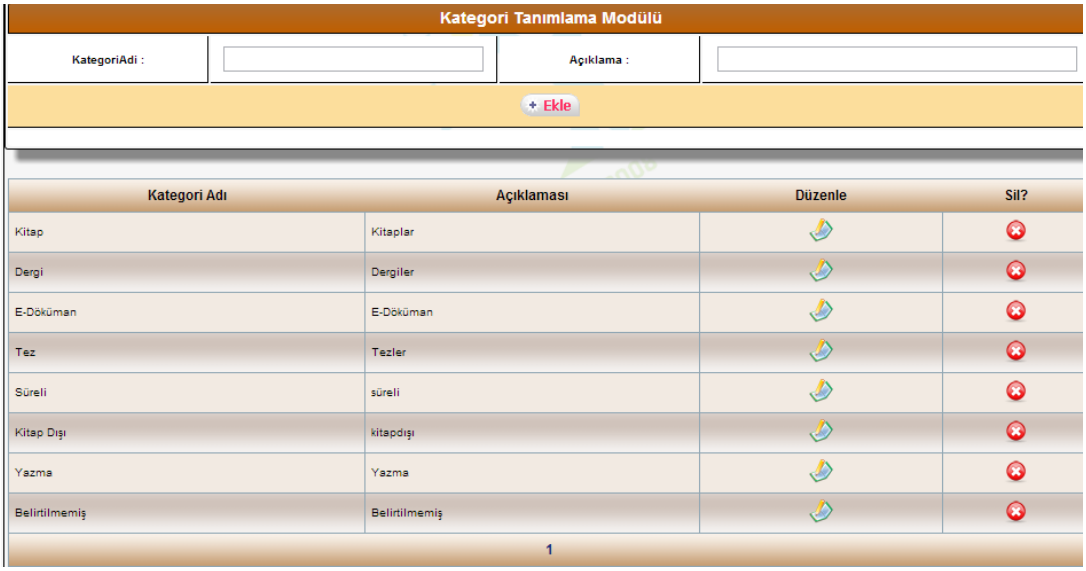
Dokümanları kataloglarken dokümanın dilini seçmek gerekecektir. Seçilecek dillerin veritabanına kaydı Dil Tanımlama modülü ile gerçekleşir. Bu modüldeki bilgiler *Tbl_Yabancidiller* tablosunda tutulmaktadır. Yabancı dil ekleme, silme ve güncelleme sayfası Şekil 6.6'da gösterilmektedir.



Şekil 6.6. Dil tanımlama modülü (Diltanimlama.aspx).

6.2.1.2. Kategori Tanımlama Modülü

Dokümanların otomatik olarak değil elle kataloglanması gerektiğinde kullanılacak olan modüldür. Doküman tiplerinin belirtildiği bölüm olarak da düşünülebilir. Kategori olarak kitap, dergi, e-doküman, tez, süreli yayın vb. belirtilebilir. Bu modüldeki bilgiler *Tbl_Kategori* tablosunda tutulmaktadır. Kategorilerin eklendiği sayfa Şekil 6.7'de gösterildiği gibidir.



Şekil 6.7. Kategori tanımlama modülü (KategoriTanimlama.aspx).

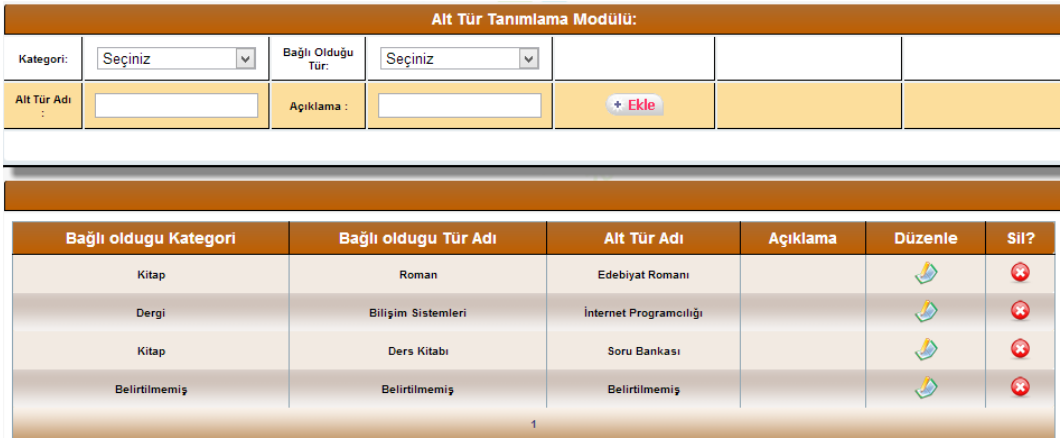
6.2.1.3. Tür Tanımlama Modülü

Kategoriye ait dokümanın türlerinin *Tbl_Tur* tablosunda kaydedildiği modüldür. Bağlı olduğu kategori mutlaka seçilmelidir. Örneğin kitap kategorisinde roman, anı-roman, ders kitabı gibi türler olabilir. Tür tanımlama modülü Şekil 6.8'de gösterilmektedir.

Şekil 6.8. Tür tanımlama modülü (TurTanimlama.aspx).

6.2.1.4. Alt Tür Tanımlama Modülü

Katalog taraması yapılırken, listelenecek doküman sayısından çok, arama kriterlerine uygun kayıtların listelenmesi, kullanıcı için çok önemlidir. Bundan dolayı dokümanları detaylı bir şekilde kataloglamak otomasyonun etkin kullanımını artıracaktır. Örneğin *Kitap* kategorisindeki, *Roman* türündeki bir dokümanın ne tip bir roman olduğu da belirtilirse daha detaylı kataloglanmış olur. Alt Tür Tanımlama modülü bu düşünceyle sisteme eklendi ve bu bilgiler *Tbl_AltTur* tablosunda tutulmaktadır. Şekil 6.9'da Alt tür tanımlama modülü gösterilmektedir.



Şekil 6.9. Alt tür tanımlama modülü (AltTurTanimlama.aspx).

6.2.1.5. Doküman Sınıfı Tanımlama Modülü

Doküman sınıfları Bilgisayar Bilimleri, Matematik – Geometri, Eğitim Bilimleri, Kişisel Gelişim - İletişim, İktisat sınıflarından oluşmaktadır. Çizelge 6.1'de sınıfların hangi durumda kullanıldıkları belirtilmiştir. Sınıflar, sınıflandırılacak yazıya atanması beklenen ve doğru sınıflandırma yapıldığının belirlenmesinde kullanılacak sınıflardır.

Çizelge 6.1. Dokümanların sınıflandırılacağı kategoriler.

SınıfID	Doküman Sınıfları	Açıklama
1	Bilgisayar Bilimleri	Bilgisayar dokümanlarını ifade eder.
2	Matematik – Geometri	Matematik – Geometri dokümanlarını ifade eder.
3	Eğitim Bilimleri	Eğitim Bilimleri dokümanlarını ifade eder.
4	Kişisel Gelişim - İletişim	Kişisel Gelişim - İletişim dokümanlarını ifade eder.
5	İktisat	İktisat dokümanlarını ifade eder.
6	Test	Test Dokümanlarını ifade eder.

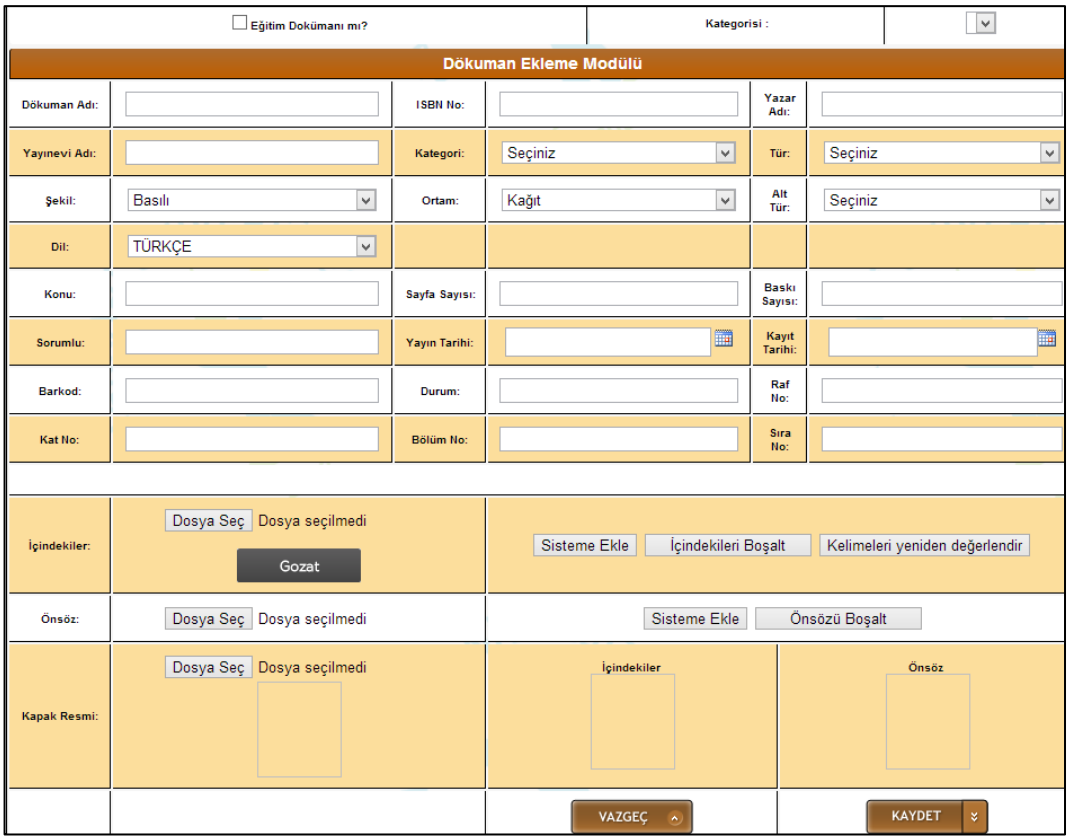
6.2.2. Doküman İşlemleri

Kütüphane kayıtlarına geçecek olan dokümanların kaydedildiği, güncellendiği modüldür. Tezin önemli bölümlerinden olan otomatik kataloglama için gerekli verilerin girildiği bölümdür.

Resim formatındaki İçindekiler, Önsöz gibi bilgilerinin OKT ile taranıp metin formatında veritabanına aktarıldığı bölüm burada yer almaktadır. Şekil 6.10'da içindekiler, önsöz etiketleri ile ilgili bölümler gösterilmektedir.

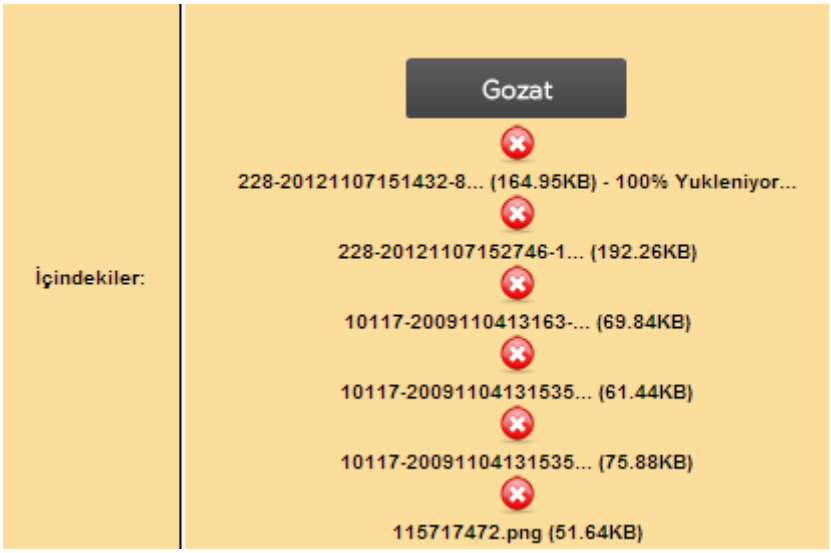
İçindekiler alanında yer alan *Gözat* butonu ile önceden taranmış içindekiler bölümlerini çoklu dosya metodu ile seçilir ve ajax kodu ile sisteme yüklenir.

Doküman eğitim dokümanı ise formun en üstündeki *Eğitim Dokümanı mı?* Checkbox'ı işaretlenir ve dokümanın sınıfı seçilir. Veritabanına yüklenen doküman eğitim dokümanı olarak işaretlenir.



Şekil 6.10. Doküman ekleme modülü (DokumanEkle.aspx).

Sisteme eklenen dokümanlar *Tbl_Dokumanlar* tablosunda tutulur. Şekil 6.11'de *jpg* formatındaki içindekiler dosyasının sisteme aktarıldığı ajax uygulaması görülmektedir.



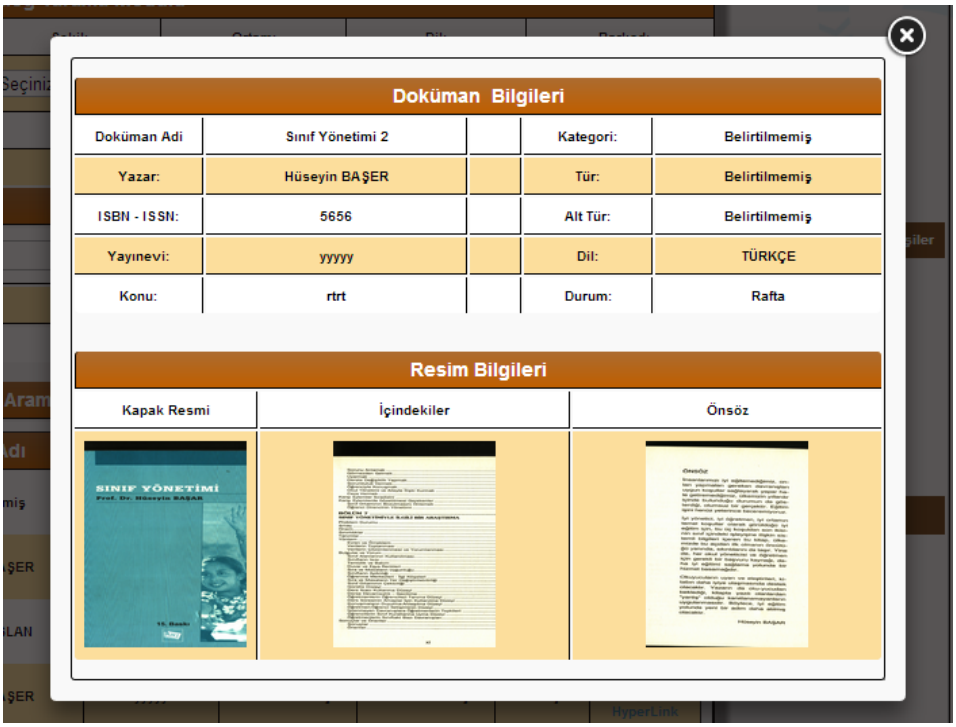
Şekil 6.11. İçindekiler sayfalarının sisteme yüklenmesi.

6.2.3. Kullanıcı Katalog Tarama Modülü

Üye girişi yapmaya gerek kalmadan herkese açık olan bölümdür. Paylaşıma açık, kütüphane dokümanlarını ve durumlarını gösteren modüldür. Bu modül sayesinde kullanıcılar kütüphanede var olan dokümanları görebilirler. Kullanıcı katalog tarama modülü Şekil 6.12'de gösterilmektedir.

Şekil 6.12. Kullanıcı katalog tarama modülü (Ara.aspx).

Son derece gelişmiş olan bu arama modülünde *Eser Adına, Yazara, Önsöz, içindekiler*, bölümlerine kadar arama yapılabilmektedir. Arama sonucunda incelenen doküman Şekil 6.13'te gösterilmektedir.



Şekil 6.13. Katalog taramada incelenen doküman (Ara.aspx).

6.2.4. Yönetici Katalog Tarama Modülü

Dokümanların güncellenmesi, silinmesi işlemlerinin yanında doküman kategorisinin otomatik olarak belirlendiği bölümdür. *Düzenle* butonu ile bir dokümanın güncellenmesi sağlanır. Otomatik kategorilendirme bu modül üzerinde yapılmaktadır ve Şekil 6.14'te gösterilmektedir.

Şekil 6.14. Yönetici katalog tarama modülü.

6.2.5. Ödünç Verme - Dolaşım Modülü

Kullanıcılara dokümanların ödünç verildiği ve takibinin yapıldığı bölümdür. Bu bölüm Bir üniversite de Öğrenci Bilgi Sistemi ile tümleşik olarak çalışmaktadır. Bilgiler *Tbl_OduncDokumanverme* tablosunda tutulur. Bu bölüm Şekil 6.15'te gösterilmektedir.

Şekil 6.15. Ödünç verme dolaşım modülü (OduncDokumanverme.aspx).

6.2.6. Güvenlik Modülü

Web sitelerinin veri güvenliğinin sağlanması için kullanıcı ve yöneticilerin sisteme giriş yaparken bazı kontrollerinin yapılması gerekmektedir. Web sitelerini bekleyen saldırıların bir tanesi de Sql injection saldırılarıdır. Otomasyon sistemini bu tip saldırılardan korumak için Şekil 6.16'da belirtilen kodlar ile güvenlik sağlanmıştır.

Kullanıcı girişinin yapıldığı bölümde, kullanıcı adı, şifre ve güvenlik kodu sorulmaktadır. Bu ve otomasyon sisteminin bütün metin girişi yapıldığı bölümlerde kullanıcının veri olarak girdiği her metin bu class tan geçirilmektedir.

Kullanıcı Adı bölümüne veri girişi yapan kullanıcı aşağıdaki kodlar ile Şekil 6.16'da belirtilen güvenlik class'ından geçerek girilen verinin sistem açısından zararlı olup olmadığı öğrenilir.

```
Public class enj{
Public static string[] karaliste = {"--
","char","nchar","varchar","nvarchar","alter","begin","cast","create","cursor","declare","delete","drop","end
","exec","execute","fetch","insert","kill","open","join","union","select","sys","sysobjects","syscolumns","w
here","modify","rename","<script>","dbo","sysdatabases","collate","having","group","xp_","table","update
","truncate","rolback"};
Public string inj(string parametre) {
for (int i = 0; i < karaliste.Length; i++){
if ((parametre.IndexOf(karaliste[i]) >= 0)) {
parametre = parametre.Replace(karaliste[i].ToString(), "");} }
return parametre.Replace("'", "''"). Replace("<", "&lt;").Replace(">", "&gt;");
}
}
```

Şekil 6.16. Veri girişlerinin kontrol edildiği class (enj.cs).

```
enj inj = new enj();
```
(6.1)

```
inj.inj(txt_KullaniciAdi.Text)
```
(6.2)

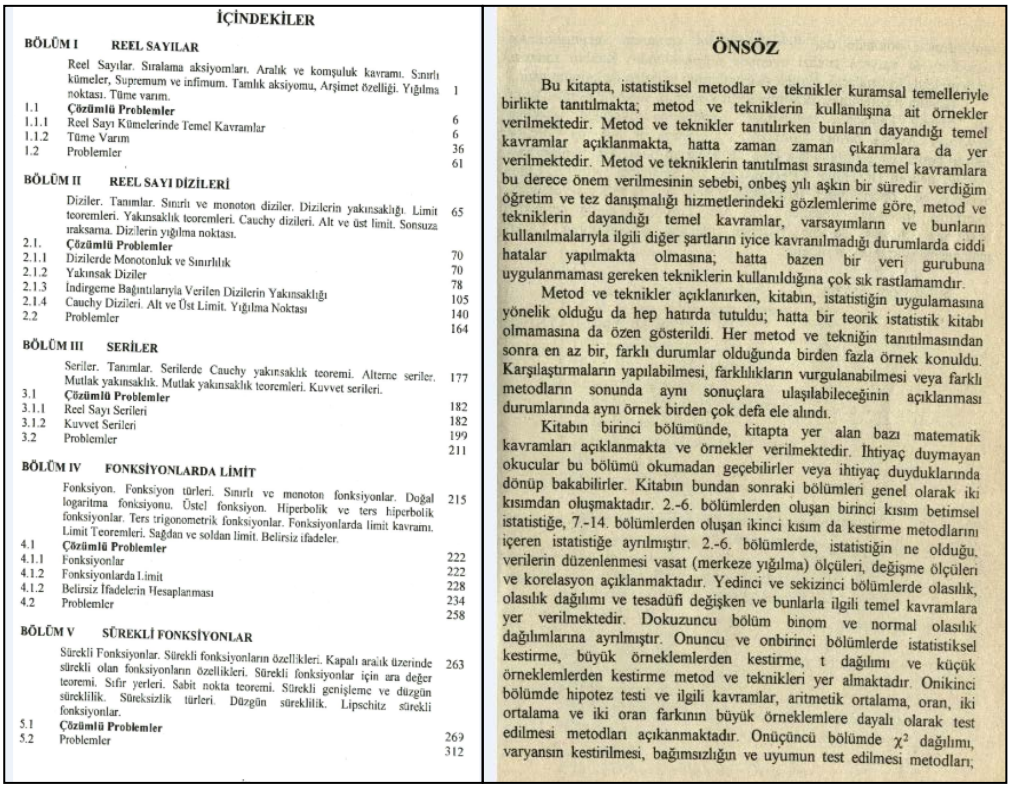
Yukardaki eşitlik 6.1'de kod ile enj.cs sınıfından inj adında bir nesne oluşturulup eşitlik 6.2'de kod ile *txt_KullaniciAdi* alanına girilen veri enj.cs dosyasına gönderilir, zararlı kodlar varsa temizlenir, değiştirilir ve geri gönderilir.

6.2.7. İstatistik Modülü

Dokümanların sağlanmasından, ödünç verilmesine kadar birçok raporlamanın alındığı bölümdür. Raporlar bir otomasyonun çıktılarıdır. Bu çıktılara bakılarak otomasyonun ve kütüphanenin fonksiyonlarının nasıl çalıştığı gözlemlenir.

6.3. DOKÜMANLARIN İÇİNDEKİLER VE ÖNSÖZ SAYFALARININ OKT İLE METNE ÇEVRİLMESİ

Tarama yöntemi ile elde edilen içindekiler ve önsöz sayfaları Şekil 6.17'de gösterilmektedir.



İÇİNDEKİLER

ÖNSÖZ

Bu kitapta, istatistiksel metodlar ve teknikler kuramsal temelleriyle birlikte tanıtılmakta; metod ve tekniklerin kullanılışına ait örnekler verilmektedir. Metod ve teknikler tanıtılırken bunların dayandığı temel kavramlar açıklanmakta, hatta zaman zaman çıkarımlara da yer verilmektedir. Metod ve tekniklerin tanıtılması sırasında temel kavramlara bu derece önem verilmesinin sebebi, onbeş yılı aşkın bir süredir verdiğim öğretim ve tez danışmalığı hizmetlerindeki gözlemlerime göre, metod ve tekniklerin dayandığı temel kavramlar, varsayımların ve bunların kullanılmalarıyla ilgili diğer şartların iyice kavranılmadığı durumlarda ciddi hatalar yapılmakta olmasına; hatta bazen bir veri gurubuna uygulanmaması gereken tekniklerin kullanıldığına çok sık rastlamamdır.

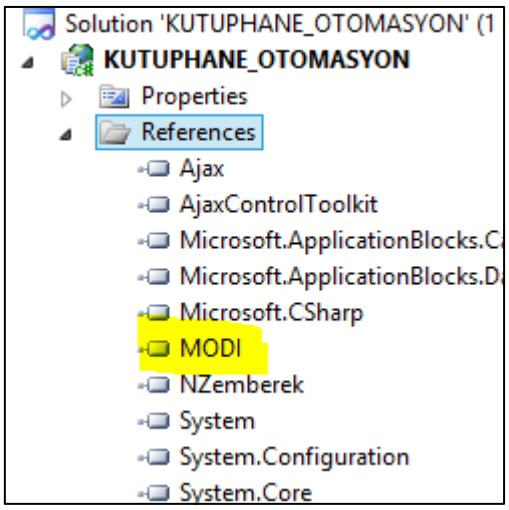
Metod ve teknikler açıklanırken, kitabın, istatistiğin uygulamasına yönelik olduğu da hep hatırda tutuldu; hatta bir teorik istatistik kitabı olmamasına da özen gösterildi. Her metod ve tekniğin tanıtılmasından sonra en az bir, farklı durumlar olduğunda birden fazla örnek konuldu. Karşılaştırmaların yapılabilmesi, farklılıkların vurgulanabilmesi veya farklı metodların sonunda aynı sonuçlara ulaşılabileceğinin açıklanması durumlarında aynı örnek birden çok defa ele alındı.

Kitabın birinci bölümünde, kitapta yer alan bazı matematik kavramları açıklanmakta ve örnekler verilmektedir. İhtiyaç duymayan okucular bu bölümü okumadan geçebilirler veya ihtiyaç duyduklarında dönüp bakabilirler. Kitabın bundan sonraki bölümleri genel olarak iki kısımdan oluşmaktadır. 2.-6. bölümlerden oluşan birinci kısım betimsel istatistiğe, 7.-14. bölümlerden oluşan ikinci kısım da kestirme metodlarını içeren istatistiğe ayrılmıştır. 2.-6. bölümlerde, istatistiğin ne olduğu, verilerin düzenlenmesi vasat (merkeze yığılma) ölçüleri, değişme ölçüleri ve korelasyon açıklanmaktadır. Yedinci ve sekizinci bölümlerde olasılık, olasılık dağılımı ve tesadüfi değişken ve bunlarla ilgili temel kavramlara yer verilmektedir. Dokuzuncu bölüm binom ve normal olasılık dağılımlarına ayrılmıştır. Onuncu ve onbirinci bölümlerde istatistiksel kestirme, büyük örneklemlerden kestirme, t dağılımı ve küçük örneklemlerden kestirme metod ve teknikleri yer almaktadır. Onikinci bölümde hipotez testi ve ilgili kavramlar, aritmetik ortalama, oran, iki ortalama ve iki oran farkının büyük örneklemlere dayalı olarak test edilmesi metodları açıkanmaktadır. Onüçüncü bölümde χ^2 dağılımı, varyansın kestirilmesi, bağımsızlığın ve uyumun test edilmesi metodları;

Şekil 6.17. Taranan içindekiler ve önsöz sayfaları.

Bölüm 3.1'de bahsedildiği üzere MODI.dll'i kullanılarak Otomasyon içerisinde OKT modülü yazılmıştır. Bu kısımda, alınan resim dosyalarından elde edilen metinlerin Bölüm 6.4.2'de bahsedilen önişlemlerden geçirilerek veritabanına kaydı

yapılmaktadır. MODI.dll referansı uygulamada Şekil 6.18'de gösterildiği gibi kullanılmıştır.



Şekil 6.18. MODI.dll'inin referansı.

Fileupload nesnesi ile alınan resim dosyaları Şekil 6.19'da verilen *resimdenoku* fonksiyonuna gönderilmiştir.

```
Public string resimdenoku (string dosyayolu, string dosyaadi){
System.IO.FileStream fstream;
System.IO.StreamWriter swriter;
System.IO.StreamReader sreader;
try{
MODI.Document mdoc = new
MODI.Document();mdoc.Create(HttpContext.Current.Server.MapPath("~/Sayfalar/icindekiler/" + dosyayolu));
//mdoc.OCR(MODI.MiLANGUAGES.miLANG_ENGLISH, true, true);
mdoc.OCR(MODI.MiLANGUAGES.miLANG_TURKISH, true, true);
MODI.Image mimg = mdoc.Images[0] as MODI.Image;
fstream = new System.IO.FileStream(HttpContext.Current.Server.MapPath("~/selim.txt"),
System.IO.FileMode.Append);
swriter = new System.IO.StreamWriter(fstream);
okunanmetin = filitre.inj(mimg.Layout.Text.ToLower());
ViewState["icindekiler"] = filitre.inj(mimg.Layout.Text.ToLower());
swriter.Write(mimg.Layout.Text);
fstream = new System.IO.FileStream(HttpContext.Current.Server.MapPath("~/selim.txt"), System.IO.FileMode.Open,
System.IO.FileAccess.Read);
sreader = new System.IO.StreamReader(fstream);
Session["metinicerigi"] = sreader.ReadToEnd();
fstream.Close();
ScriptManager.RegisterStartupScript(Page, typeof(string), Guid.NewGuid().ToString(), "alert('İçindekiler Sayfası
Hafızaya Alındı');", true);
}
catch{
ScriptManager.RegisterStartupScript(Page, typeof(string), Guid.NewGuid().ToString(), "alert('Hata oluştu');", true);
}
return dosyayolu}
```

Şekil 6.19. Resimdenoku fonksiyonu.

resimdenoku fonksiyonunda okunan veri enj.cs classından geçirilerek *session, viewstatelere* aktarılır ve veri tabanına kaydedilir.

MODI.dll kullanılarak 5 kategoride yatay ve dikey çözünürlüğü 300 dpi (dots per inch), bit derinliği 24 olan toplam 25 doküman taranmıştır. Uygulamadaki OKT sistemi ile elde edilen metinlerin başarım oranları Çizelge 6.2'de gösterilmiştir.

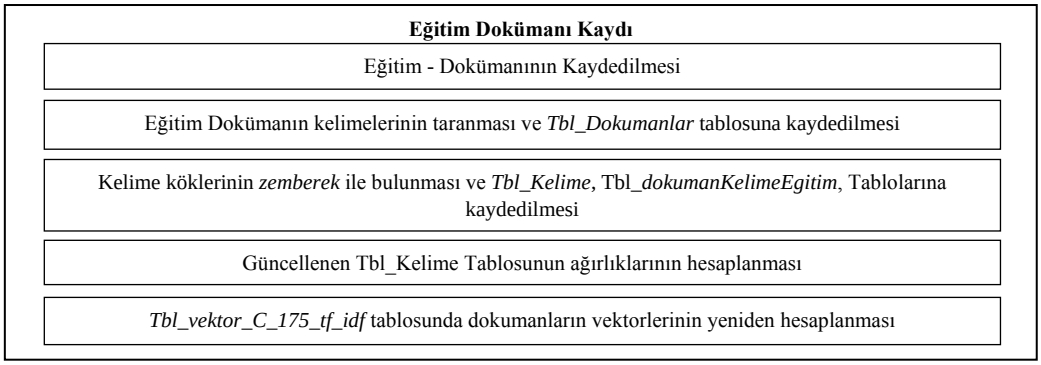
Çizelge 6.2. MODI.dll ile elde edilen kelime sayıları.

	Bilgisayar Bilimleri	**Matematik – Geometri**	**Eğitim Bilimleri**	**Kişisel Gelişim - İletişim**	**İktisat**
Basılı Dokümanlardaki kelime sayısı	2,370	1,600	3,250	3,306	3,170
OKT ile tanımlanan Kelime sayısı	2,256	1,299	3,131	3,268	3,051
Başarı oranı (%)	95	81,18	96,3	98,8	96,2

Çizelge 6.2'den anlaşılacağı üzere en yüksek başarı % 98,8 oranında Kişisel Gelişim – İletişim kategorisinde elde edilmiştir. Bu oran taranan doküman tipine ve sayısına göre değişkenlik göstermektedir.

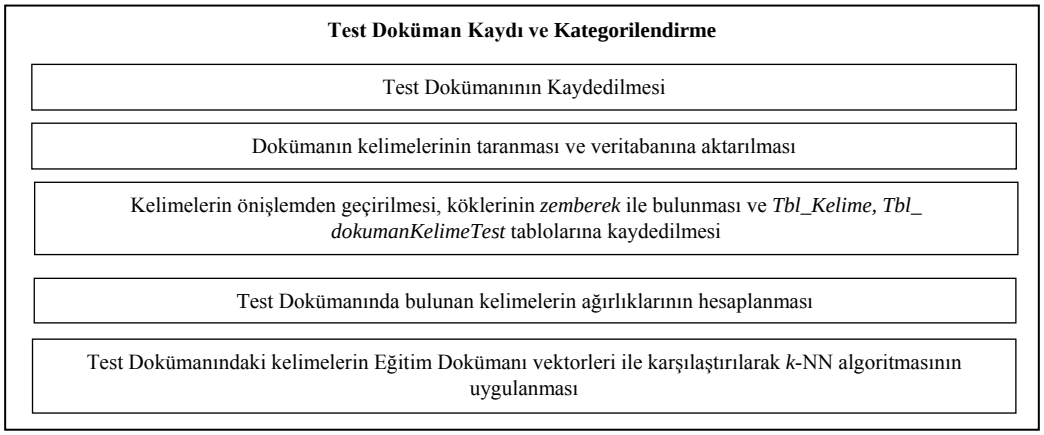
6.4. DOKÜMANLARIN HAZIRLANMASI

Kategorilendirme bölümü iki aşamadan oluşur. İlk aşamada, sistemde kategorilendirmek için tanımlı eğitim dokümanlarının bulunması gereklidir. Eğitim dokümanlarının kayıt aşamaları Şekil 6.20'de gösterilmektedir.



Şekil 6.20. Eğitim dokümanın kaydının yapılması.

İkinci aşamada ise eğitim dokümanları ile eğitilen sisteme test dokümanları girilerek kategorilendirilmesi sağlanır. Bu işlemler Şekil 6.21'de gösterilmektedir.



Şekil 6.21. Test dokümanı kaydı ve kategorilendirme adımları.

6.4.1. Dokümanlarının Kaydedilmesi

Dokümanlar *DokumanEkle.aspx* sayfasında kaydedilmektedir. Eğitim dokümanları kaydedilirken Dokümanın Kategorisinin seçilmesi gerekmektedir. Bu tez çalışmasında 5 farklı kategorideki dokümanların sınıflandırılması amaçlanmıştır. Bu kategoriler Şekil 6.22'de gösterilmektedir.

Results | Messages

	pkID	Kategori	Aciklama
1	1	Bilgisayar Bilimleri	Bilgisayar Bilimleri
2	2	Matematik-Geometri	Matematik-Geometri
3	3	Eğitim Bilimleri	Eğitim
4	4	Kişisel Gelişim	Kişisel Gelişim
5	5	İktisat	İktisat

Şekil 6.22. Eğitim dokümanlarının kategorileri (Tbl_DokumanKategorileri).

DokumanEkle.aspx sayfasında bir belge eğitim dokümanı olarak kaydedilirken kategori bölümünde Şekil 6.22'deki liste gelir. Buradan belgenin hangi eğitim kategorisine ait olduğu seçilir. Bu işlem Şekil 6.23'te gösterilmiştir.

Şekil 6.23. Eğitim dokümanın kategori seçimi (DokumanEkle.aspx).

Dokümanların, veritabanında *Tbl_Dokumanlar* tablosunda tutulduğunu daha önce belirtilmişti. Doküman, Eğitim dokümanı ise *Tbl_Dokumanlar* tablosunda *Egitim* kolonu *evet* olarak kaydedilir ve aynı tabloda *DokumanKategoriEgitimID* alanı *Tbl_DokumanKategorileri* tablosundaki *pkID* kolonundaki ID numarası yazılır. Şekil 6.24'te veritabanına kaydedilmiş doküman gösterilmektedir.

icindekilerresim	onsozresim	Egitim	DokumanEgitimKategoriID
PTDC0004.JPG	onsoz2342013235756.jpg	evet	2
icindekiler2442013102617.jpg	onsoz2442013001004.jpg	evet	2
PTDC0003.JPG		hayir	0
PTDC0004.JPG		hayir	0
PTDC0012.JPG	onsoz2442013103558.jpg	evet	2
icindekiler2442013103948.jpg	onsoz2442013104012.jpg	evet	2
PTDC0022.JPG	onsoz2442013104227.jpg	evet	2

Şekil 6.24. Dokümanların veritabanındaki kaydı (Tbl_Dokumanlar).

6.4.2. Doküman Metinlerinin Ön İşlemden Geçirilmesi

Metin madenciliğinde ön işlem, gereksiz ya da sistem için zararlı kelimelerin arındırılıp, kelime köklerinin bulunması ile tamamlanır.

Aynı zamanda boyut azaltma işlemi de olan ön işlem aşamasının amacı, yazının sınıflandırılması işleminde sınıflandırıcının karar vermesine yardımcı olmayan, temizlenmediğinde yanlış sonuçlar çıkmasına neden olabilecek verilerin yazıdan temizlenmesi ve kelime köklerinin elde edilmesidir.

6.4.2.1. İçeriğin Karakterlerden Temizlenmesi

Çift tırnak ("), virgül (,), tire (-), yıldız (?, *) gibi karakterler temizlenir. SQL komutlarında tırnak (') metin ayıracı olarak kullanıldığında yapılacak sorgularda problem yaşanmaması için (') karakteri (' ') ile değiştirilir. Ayrıca büyük (>) işareti html kodu (>) koduna, (<) işareti de html kodu (<) koduna dönüştürülür. Bu durum Şekil 6.16' da gösterilmektedir.

6.4.2.2. İçeriğin Gereksiz Kelimelerden Temizlenmesi

```
Public classfiltre{
public string tumcumle = "";
string karaliste2 = "ol ve VE ki Kİ hem HEM hemde HEMDE de DE da DA İLE ile AMA ama FAKAT fakat LAKİN
lakin YALNIZ yalnız NE ne YA ya PEK pek TE te TA ta ÇOK çok BU bu gibi GİBİ için İÇİN bölüm BÖLÜM .BÖLÜM
.bölüm .konu .KONU içindekiler İÇİNDEKİLER kaynakça KAYNAKÇA KAYNAKCA KAYNAK iş İŞ İŞTE işte acaba
,ACABA. Ötürü, ÖTÜRÜ";
public string inj(string parametre){
string[] kelimeler = parametre.Split(' '); // ayırıcı parametre boşluk
foreach (string kelime in kelimeler) {
if ((karaliste2.IndexOf(kelime) >= 0)){}
else{
tumcumle += " " + kelime.ToString();} }
return tumcumle.Replace("'", "''").Replace("<", "&lt;").Replace(">", "&gt;"); } }
```

Şekil 6.25. Gereksiz kelimelerin temizlendiği class (filtre.cs).

Türkçede tek başına anlamı bulunmayan, anlama herhangi bir etkisi olmayan edat, (gibi, için vb.) bağlaç (ile, ama vb.) gibi gereksiz kelimelerin (stop words) içerikten temizlenmesi gerekir. Çünkü bu kelimelerin anlama herhangi bir katkısı olmadığı gibi temizlenmediğinde de yanlış sonuçlar elde edilmesine neden olur. Bu kelimeler sisteme tanıtılmıştır ve bu kelimeler bulunursa içeriğe dahil edilmez. *filtre* class'ı, gelen metinden Şekil 6.25'te gösterilen kelimeleri temizler ve filtrelenmiş metni elde eder.

6.4.2.3. Ön Elemeden Geçen Kelimelerin Zemberek ile Kelime Gövdelerini Bulma

Türkçe, Ural-Altay dil grubuna giren bir dildir. Sözcük yapısı ve üretimi açısından Türkçe sondan eklemeli bir dil olduğundan farklı iki kelime benzer anlamları taşıyabilmektedir. Bu projede çekim ekleri farklılıklarının benzerlikte etkisinin

olmaması gerektiği düşüncesi ile kelimelerin sözlük tablosunda gövde halleri ile saklanması ve karşılaştırmaların kelimelerin gövdeleri arasında olması gerektiği düşünülmüştür.

Buna abartılı bir örnek olarak "gözlemlenemeyeceklerindendir" sözcüğünü verebiliriz. Kökü “gözlem” olan bu sözcüğün biçim birimleri şu şekilde gösterilebilir:

gözlem+le+ne+meye+cek+ler+in+den+dir

Literatürde bu probleme çözüm olarak sadece kelime köklerinin kullanılması önerilmektedir (Torunoğlu vd., 2011). Bu sayede hem aynı anlama işaret eden kelimelerin birleştirilmesi (böylelikle metinler arası benzerliğin daha iyi ifade edilmesi sağlanmakta), hem de özellik boyutunun azaltılmasıyla işlem karmaşıklığının azaltılması sağlanmaktadır. Kelimelerin köklerinin bulunması için Zemberek kütüphanesi kullanılmıştır

Bu çalışmada zemberek projesinin çalışma prensibi değiştirilmeden, .NET ortamı için hazırlanan NZemberek.dll dosyası kullanılarak kelime gövdelerine ulaşılmaktadır.

OKT ile elde edilen, kelimeler, NZemberek kütüphanesi yardımıyla kök ve eklerine ayrılır. Ek listesinden çekim ekleri kaldırılarak kelimenin kökü ve yeni ek listesi ile yeni kelime üretilir. Bu şekilde ayrıştırılan kelimenin gövdesi (kök) bulunur ve veri tabanında kelime tablosunda yer almayan kelimeler *Tbl_kelime* tablosuna eklenir.

Örneğin “kalemlikte” kelimesi zemberek aracılığı ile çözümlenir. Zemberekten çözümleme sonrası dönen cevap şu şekildedir;

kalem-lik-te (isim_kök, yapım_eki_bulunmalı, bulunma_de)

Gelen cevaba göre ekler yapım eki ve çekim eki olarak tasnif edilir, çekim ekleri kelimeden çıkarılarak kelime köklerine ulaşılmış olur.

Zemberek kütüphanesi birçok akademik çalışmada kullanılmıştır. Zembereğin yazarları tarafından yapılan testlerde Çizelge 6.3'teki sonuçlar elde edilmiştir.

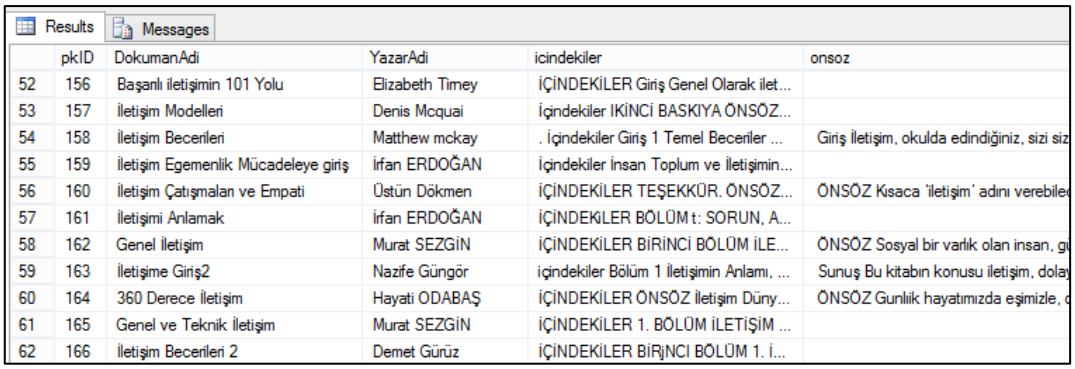
Çizelge 6.3. Zemberek kelime istatistikleri.

Toplam kelime sayısı	5 160 619
Toplam isim sayısı	5 278 505
Toplam sıfat sayısı	269 701
Toplam fiil sayısı	1 414 014
Toplam sayı sayısı	232 832
Toplam kok sayısı	14 531
Toplam isim kök sayısı	11 755
Toplam sıfat kök sayısı	570
Toplam fiil kök sayısı	2 068

Tablodan da görüldüğü gibi 5 milyon Türkçe kelime için 14 531 kök kullanılmıştır. Bu rakam kullanılan metinlerin çeşitliliği ve miktarı arttıkça daha da büyüyecektir, ancak Zemberek yaklaşık 22 000 kök, 6 000 isim ve bir kaç bin özel isim köklerini tanır.

6.4.3. Kelimelerin Sisteme Kaydedilmesi

Dokümanları sisteme kaydederken dokuman içerisindeki kelimeler bir diziye aktarılır. Elde edilen kelimeler Şekil 6.25'te belirtilen İçeriğin gereksiz kelimelerden temizlenmesi aşaması gerçekleşir ve *Tbl_dokuman* tablosunda *icindekiler*, *onsoz* alanlarına kaydedilir. Bir dokümanda İçindekiler haricindeki bilgiler (önsöz, giriş, özet) *onsoz* alanına kaydedilir. Şekil 6.26'da kaydedilen dokümanlara ait *icindekiler* ve *onsoz* alanlarındaki kelime bilgileri gösterilmektedir.

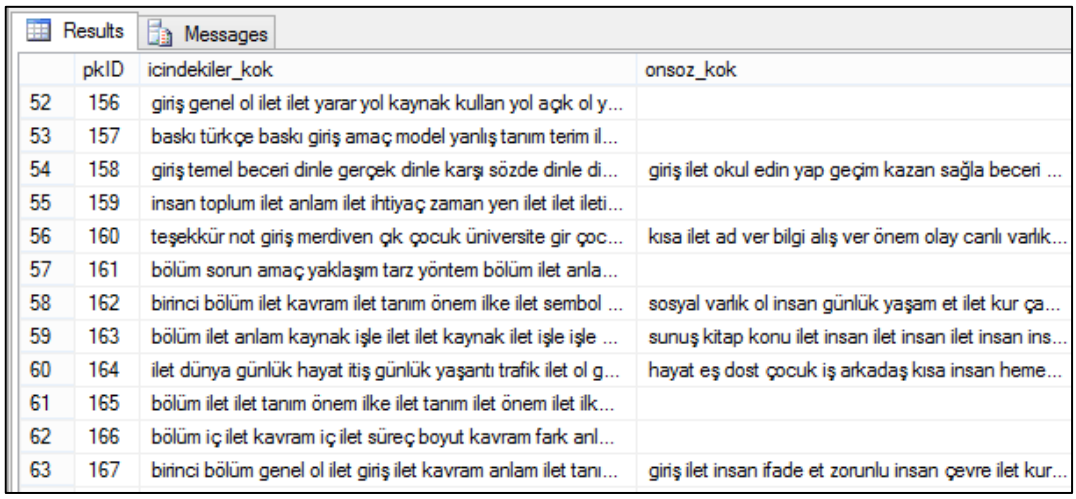


Results | Messages

	pkID	DokumanAdi	YazarAdi	icindekiler	onsoz
52	156	Başarılı iletişimin 101 Yolu	Elizabeth Timey	İÇİNDEKİLER Giriş Genel Olarak ilet...	
53	157	İletişim Modelleri	Denis Mcquai	İçindekiler IKİNCİ BASKIYA ÖNSÖZ...	
54	158	İletişim Becerileri	Matthew mckay	. İçindekiler Giriş 1 Temel Beceriler ...	Giriş İletişim, okulda edindiğiniz, sizi siz
55	159	İletişim Egemenlik Mücadeleye giriş	İrfan ERDOĞAN	İçindekiler İnsan Toplum ve İletişimin...	
56	160	İletişim Çatışmaları ve Empati	Üstün Dökmen	İÇİNDEKİLER TEŞEKKÜR. ÖNSÖZ...	ÖNSÖZ Kısaca 'iletişim' adını verebile
57	161	İletişimi Anlamak	İrfan ERDOĞAN	İÇİNDEKİLER BÖLÜM t: SORUN, A...	
58	162	Genel İletişim	Murat SEZGİN	İÇİNDEKİLER BİRİNCİ BÖLÜM İLE...	ÖNSÖZ Sosyal bir varlık olan insan, g
59	163	İletişime Giriş2	Nazife Güngör	içindekiler Bölüm 1 İletişimin Anlamı, ...	Sunuş Bu kitabın konusu iletişim, dola
60	164	360 Derece İletişim	Hayati ODABAŞ	İÇİNDEKİLER ÖNSÖZ İletişim Düny...	ÖNSÖZ Gunlilik hayatımızda eşimizle,
61	165	Genel ve Teknik İletişim	Murat SEZGİN	İÇİNDEKİLER 1. BÖLÜM İLETİŞİM ...	
62	166	İletişim Becerileri 2	Demet Gürüz	İÇİNDEKİLER BİRİNCİ BÖLÜM 1. İ...	

Şekil 6.26. İçindekiler ve önsöz bilgilerinin *Tbl_Dokuman* tablosundaki görünümü.

Ara.aspx sayfasında katalog tarama yaparken *icindekiler ve onsoz* alanlarından da arama yapılabilir. Bu kelimeler metin ön işlem aşamalarından geçirildikten sonra *Tbl_Dokuman* tablosunda *icindekiler_kok* ve *onsoz_kok* alanlarına gereksiz ve zararlı kelimelerden arınmış kelime kökleri kaydedilir. Zemberek kütüphanesi kullanılarak elde edilen kelime köklerinin alanları Şekil 6.27'de gösterilmektedir.



Results | Messages

	pkID	icindekiler_kok	onsoz_kok
52	156	giriş genel ol ilet ilet yarar yol kaynak kullan yol açık ol y...	
53	157	baskı türkçe baskı giriş amaç model yanlış tanım terim il...	
54	158	giriş temel beceri dinle gerçek dinle karşı sözde dinle di...	giriş ilet okul edin yap geçim kazan sağla beceri ...
55	159	insan toplum ilet anlam ilet ihtiyaç zaman yen ilet ilet ileti...	
56	160	teşekkür not giriş merdiven çık çocuk üniversite gir çoc...	kısa ilet ad ver bilgi alış ver önem olay canlı varlık...
57	161	bölüm sorun amaç yaklaşım tarz yöntem bölüm ilet anla...	
58	162	birinci bölüm ilet kavram ilet tanım önem ilke ilet sembol ...	sosyal varlık ol insan günlük yaşam et ilet kur ça...
59	163	bölüm ilet anlam kaynak işle ilet ilet kaynak ilet işle işle ...	sunuş kitap konu ilet insan ilet insan ilet insan ins...
60	164	ilet dünya günlük hayat itiş günlük yaşantı trafik ilet ol g...	hayat eş dost çocuk iş arkadaş kısa insan heme...
61	165	bölüm ilet ilet tanım önem ilke ilet tanım ilet önem ilet ilk...	
62	166	bölüm iç ilet kavram iç ilet süreç boyut kavram fark anl...	
63	167	birinci bölüm genel ol ilet giriş ilet kavram anlam ilet tanı...	giriş ilet insan ifade et zorunlu insan çevre ilet kur...

Şekil 6.27. Kelime köklerinin *Tbl_Dokuman* tablosundaki görünümü.

Tbl_Dokuman tablosunda hem test dokümanlarının bilgileri hem de eğitim dokümanlarının bilgileri bulunmaktadır. Tüm dokümanlara ait kelimeler *Tbl_Kelime* tablosunda yoksa bu tabloya kaydedilir. Doküman kaydetme işlemi bittikten sonra Uygulamada *Otomatik Kategorilendir -> Kelime idf* menüsünden kelimelerin *idf* değerleri hesaplattırılır. Bu işlem sonunda her kelimenin bir sayısal değeri olup,

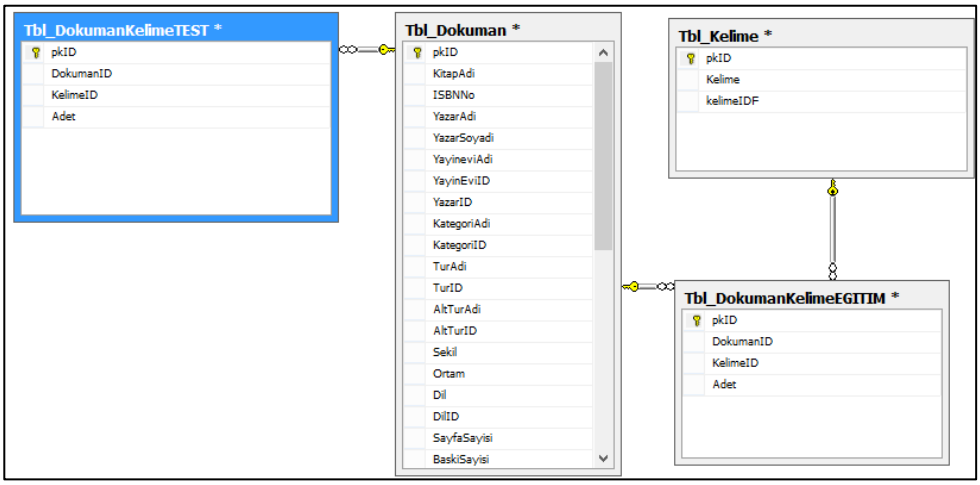
kelimelerin sayısallaştırma işlemi gerçekleştirilir. Şekil 6.28'de bazı kelimelerin hesaplanan *idf* değerleri gösterilmektedir.

Results | Messages

	pkID	Kelime	kelimeIDF
25	580	sistem	0,477121254719662
26	581	başarı	0,602059991327962
27	582	aday	1,20411998265592
28	583	zihin	1,30102999566398
29	584	gücün	1,61278385671974
30	585	yan	0,778151250383644
31	586	sıra	0,778151250383644
32	587	yeter	1,11394335230684
33	588	düzey	0,903089986991944
34	589	bilgi	0,477121254719662
35	590	beceri	0,602059991327962
36	591	sahip	0,778151250383644
37	592	gerek	0,477121254719662

Şekil 6.28. *Tbl_Kelime* tablosunda kelime kökleri ve ağırlıkları.

Doküman kayıt esnasında her kelime Tbl_kelime tablosuna aktarıldıktan sonra dokümanın tipine göre işlemler devam eder. Eğer kaydedilen doküman eğitim dokümanı ise diziye aktarılan kelimeler Tbl_DokumanKelimeEGITIM tablosuna da kaydedilir. Doküman test dokümanı ise diziye aktarılan kelimeler Tbl_DokumanKelimeTEST tablosuna kaydedilir. Bu tablolarda hangi sözcük hangi dokümanda kaç kez geçmiş bunun bilgisi tutulmaktadır. Uygulamada kelimeler ilgili dokümana göre üç farklı tabloda tutulmaktadır ve Şekil 6.29'da bu tablolar gösterilmektedir.



Şekil 6.29. Doküman ve kelime tabloları arasındaki ilişki.

6.4.4. Kelimelerin Ağırlıklandırılması

Dokümanlar vektörler ile ifade edilirler. Doküman vektörü, dokümanı oluşturan kelimelerden oluşur. Vektörler oluşurken kelimelerin vektörde hangi değerle temsil edileceği kelime ağırlıklandırma ile hesaplanır. Bu tez çalışmasında *tf-idf* ağırlıklandırma yöntemi kullanılmıştır. Kelime Ağırlıklandırma menüsü Şekil 6.30'da gösterilmiştir.

Şekil 6.30. Kelime-*idf* hesaplama menüsü.

6.4.5. Doküman Vektörünü Oluşturacak Kelime Seçimi

Dokümanı hangi sözcüklerin temsil edeceğinin belirlendiği bölümdür. Bu işlem sayesinde vektör boyutları azalacak ve vektör dizisi içerisinde çalışmak daha hızlı olacaktır.

idf değerinde sözcüklerin geçtiği eğitim dokümanı sayısı dikkate alınır. Eğer bir sözcük, az sayıda eğitim dokümanında geçiyorsa yüksek, çok sayıda dokümanda geçiyorsa düşük *idf* değerli sözcük şeklinde değerlendirilir ve yüksek *idf* değerli sözcükler seçilir.

Şekil 6.31'de dokümanlarda geçen kelimelerin *DokumanID, KelimeID* ve *Adet* bilgileri *Tbl_DokumanKelimeEGITIM* tablosunda tutulmaktadır. Bu tablodaki *Adet* alanı kullanılarak, özellik seçimi sonucu seçilecek kelimeler elde edilir. Bu tez çalışmasında kategorileri belirleyecek, dokümanlarda en çok geçen ilk 175 *(C_175)* kelime seçilmiştir. Literatürde kelime sayısı az seçildiğinde sınıflandırma başarısının düştüğü çok fazla seçildiğinde ise hem başarının hem de performansın düştüğü

gözlemlenmiştir. Şekil 6.32'de veritabanından her eğitim kategorisi için en çok geçen 175 kelimeyi çeken kod bölümü gösterilmektedir.

Results | Messages

	pkID	DokumanID	KelimeID	Adet
1	1158	105	650	8
2	1368	113	650	5
3	1506	117	650	2
4	1690	108	650	11
5	2647	139	650	2
6	3620	147	650	4
7	5435	163	650	2
8	6657	116	650	1

Şekil 6.31. Kelimelerin dokümanlar üzerindeki dağılım tablosu.

```
Select KelimeID from (SELECT top 175 * FROM (SELECT KelimeID, COUNT(dke.Adet) as toplamAdet FROM
Tbl_DokumanKelimeEGITIM dke
LEFT JOIN Tbl_Dokuman ON dke.DokumanID = Tbl_Dokuman.pkID GROUP BY KelimeID,
DokumanEgitimKategoriID , Egitim HAVING DokumanEgitimKategoriID=1and egitim='evet' ) t
order by toplamAdet desc) S1
UNION …
```

Şekil 6.32. Bütün sınıflarda, en fazla dokümanda geçen 175 kelimeyi alıp sınıf özellik vektörünü oluşturan kod bölümü.

6.4.6. Vektörlerin Oluşturulması

Özellik seçimi sonucu elde edilen sözcüklerle Tbl_DokumanKelimeTEST tablosu, her bir yazı için ayrı ayrı olmak üzere eşleştirilir. Bu şekilde sözlükteki kelimelerden yazı içerisinde geçenler bulunur, ağırlıklandırması yapılır, vektör oluşturulur ve veritabanına kaydedilir. Vektör elemanlarının birbirinden ayrılması için önişlem aşamasında karakter değişimi için de kullanılan "·" karakteri kullanılır.

Yeni bir eğitim dokümanı eklendiğinde veya silindiğinde özellik seçimi sonucu meydana gelen sözlük değişebileceğinden sınıf özellik vektörü ve dolayısıyla doküman vektörü oluşturma işlemi yinelenmelidir. Eğitim dokümanları işlemlerinin son adımı olan vektörlerinin oluşturulmasıyla yazılar sınıflandırma işleminde kullanılmak üzere hazırlanmış olmaktadırlar.

Uygulamada *Otomatik Kategorilendir -> Count 175 tf-idf Vektör Oluştur* menüsünden dokümanların vektörleri otomatik olarak oluşturulması sağlanmıştır. Dokümanların vektörleri veritabanında *Tbl_vektor_C_175_tf_idf* tablosunda tutulur.

Sınıf özellik vektörü ve dokümanı oluşturan kelimeler eşleştirilerek sözcüklerin varlığı, yazı içerisinde araştırılır ve "·" karakteri ile ağırlıklar birleştirilir. *Tbl_vektor_C_175_tf_idf* tablosundaki, *DokumanID* alanına, Dokumanın *pkID* değerini; *vektor* alanına, birleştirilmiş ağırlıkları kaydeder. Burada *C* Count anlamındadır. Kelimelerin vektörel olarak ifadesi Şekil 6.33'te gösterilmektedir.

DokumanID	vektor
104	0·0·7,15681882079495·12,8822738774309·3,61235994796776·1,43136376415899·8,91154681845472·5,248333
105	4,21441993929573·0·3,38039216005703·2,79588001734408·2,40823996531185·1,20411998265592·1,90848501
108	10,4139268515823·6,02059991327962·0·1,43136376415899·10,0195463491129·3,31132995230379·15,7450014
109	0·6,24835611094938·0·1,20411998265592·0·0·0·0·0·0·0·0·0·0·0·0·0·0·0·0·0·0·0·0·0·0·0·0·0·0·0·0·0·0·0·0·0·0·0·0·0·0·0
110	11,4509101132719·1,80617997398389·7,15681882079495·1,80617997398389·5,5697167615342·4,21441993925
111	1,90848501887865·3,01029995663981·0,954242509439324·1,39794000867204·8,33114148126584·0·0,9542425

Şekil 6.33. *C_175_tf_idf* sınıf özellik vektörü uygulanarak elde edilen doküman vektörleri.

6.5. DOKÜMAN KATEGORİLERİNİN OTOMATİK BULUNMASI

Kategorilendirmenin otomatik olarak yapılabilmesi için öncelikle sistemin eğitilmesi gerekmektedir. Bu işlem Bölüm 6.4'te bahsedildiği gibi eğitim dokümanları ile yapılmaktadır. Her kategori için eğitim dokümanlarının sayısı ve her kategorinin ifade edildiği toplam kelime sayısı sınıflandırma için çok önemlidir. Uygulamada tüm kategoriler için eşit sayıda eğitim dokümanı düşünülmüş fakat elde edilen sonuçlara bakıldığında en az kelime ile Matematik – Geometri kategorisinin ifade edildiği tespit edilmiştir. Bu durum sınıflandırmayı olumsuz etkilemiş ve doküman sayısından ziyade her kategorinin birbirine yakın sayıda kelime ile ifade edilmesi uygun görülmüştür. Bu tez çalışmasında kullanılan doküman ve kelime sayıları Çizelge 6.4'te gösterilmektedir.

Çizelge 6.4. Eğitim dokümanları ve ifade edildikleri kelime sayıları.

	Bilgisayar Bilimleri	**Matematik – Geometri**	**Eğitim Bilimleri**	**Kişisel Gelişim - İletişim**	**İktisat**
Doküman Sayısı	35	43	22	25	25
Kelime Sayısı	2 125	2 064	2 300	2 500	2 235

Bu çalışmada dokümanların kategorilerinin bulunması için *k*-NN algoritması kullanılmıştır. *k* değeri olarak *(k=3), (k=5), (k=7)* kullanılarak karşılaştırmalar yapılmıştır. *tf-idf* ağırlıklandırması kosinüs benzerliği ile uygulanmıştır.

Benzerlik oranları aynı olan birden fazla sınıf varsa dokümanın sınıfı *Benzer* olarak belirlenmiştir. *k*-NN bazı uygulamalarında, test dokümanıyla herhangi bir eğitim dokümanı arasında benzerlik bulunamamıştır. Bu tür dokümanların sınıfı *Kategorilendirilmemiş* olarak belirlenmiştir.

Kategorilendirme işlemi ekranına, sisteme yönetici olarak giriş yaptıktan sonra "*Doküman İşlemleri* " menüsünden, "*Doküman Ara* " linkine tıklayarak ulaşılabilir. Buradan da hangi dokümanın kategorisi bulunmak isteniyorsa "*Otomatik Kategorilendir*" linki ile kategorilendirme başlatılır.

6.6. UYGULAMA SONUÇLARI VE DEĞERLENDİRİLMESİ

Kullanılan sınıf özellik vektörü *C_175* ve sınıflandırma algoritması *k*-NN ile farklı *k* sabitleri uygulanarak elde edilen sonuçlar Şekil 6.34'te gösterilmektedir. *k=7* seçildiğinde Matematik Seti (test) adlı dokümanın kategorilendirme işlemleri Şekil 6.34'te gösterilmiştir.

Matematik Seti (Test)

Doküman Kategorileri	Dokümanın Kategorilere olan yakınlığı
Bilgisayar	% 0
Matematik	% 71,43
Eğitim Bilimleri	% 0
Kişisel Gelişim	% 28,57
İktisat	% 0

(a)

DokumanID	KategoriAdi	KategoriID	Yakinlik
156	Kişisel Gelişim	4	0,31312017730203
117	Matematik	2	0,291763682486781
157	Kişisel Gelişim	4	0,268041372762226
177	Matematik	2	0,256567371306669
116	Matematik	2	0,248285682415404
189	Matematik	2	0,247960626910196
185	Matematik	2	0,246145297803883

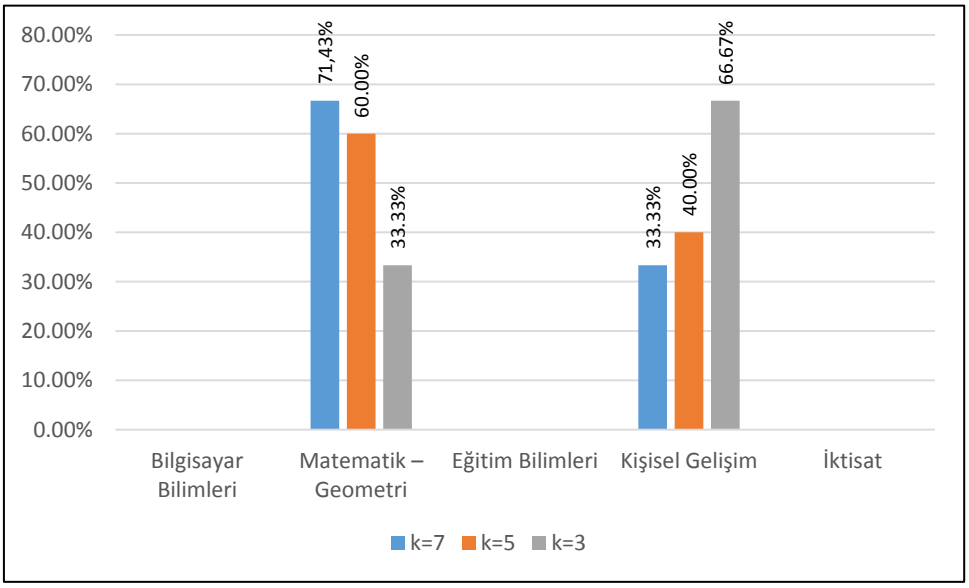
(b)

Şekil 6.34. *k=7* için otomatik kategorilendirme sonuçları a) başarı sonuçları b) yakınlık değerleri.

Çizelge 6.5. Matematik – Geometri kategorisindeki test dokümanının otomatik kategorilendirme sonuçları.

		Başarı Oranları				
Doküman Adı	***k* değeri**	**Bilgisayar Bilimleri(%)**	**Matematik – Geometri (%)**	**Eğitim Bilimleri(%)**	**Kişisel Gelişim – İletişim (%)**	**İktisat(%)**
Matematik Seti (Test)	7	0,00	71,43	0,00	28,57	0,00
Matematik Seti (Test)	5	0,00	60,00	0,00	40, 00	0,00
Matematik Seti (Test)	3	0,00	33,33	0,00	66,67	0,00

k-NN'nin *(k=7)* değerinin *tf-idf* ağırlıklandırma kullanılarak yapılan sınıflandırmalarında başarı kosinüs benzerliği için belirlenen sınıf özellik vektöründe % 71,43'dür.



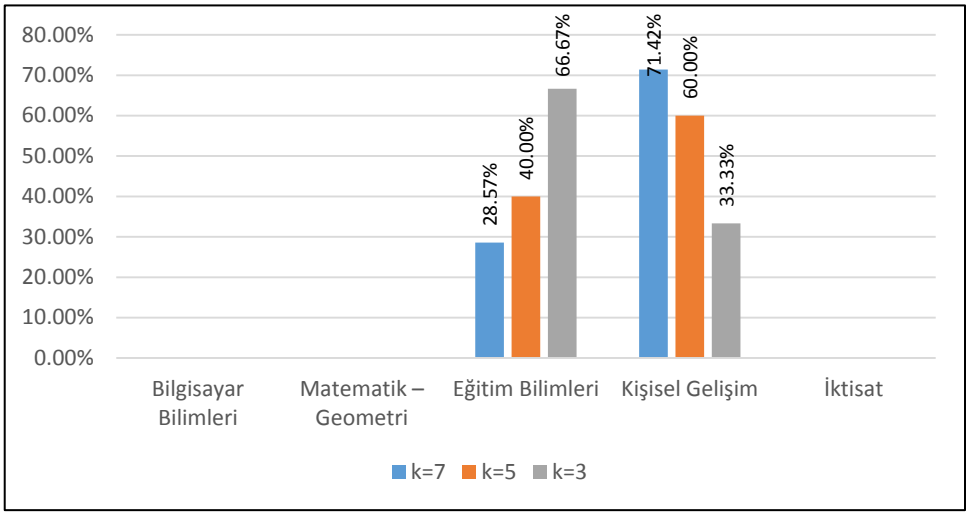
Şekil 6.35. Matematik Seti (test) dokümanı için *farklı k* değerleri sonuçları.

Çizelge 6.5'te Matematik Seti (Test) dokümanı sisteme içindekiler ve önsöz sayfaları taranarak eğitim dokümanları ile karşılaştırılıp *k=7* alınarak % 71,43 oranıyla Matematik-Geometri sınıfından olduğu tespit edilmiştir. *k=3* alındığında ise Matematik-Geometri sınıfından olan doküman Kişisel Gelişim - İletişim dokümanı olarak tanımlanmıştır.

Şekil 6.35'te *k=7, k=5* ve *k=3* değerleri için elde edilen sonuçlar grafiksel olarak gösterilmiştir. Diğer kategoriler için elde edilen sonuçlar aşağıdaki çizelge ve şekillerde gösterilmektedir.

Çizelge 6.6. Kişisel Gelişim - İletişim kategorisindeki test dokümanının otomatik kategorilendirme sonuçları.

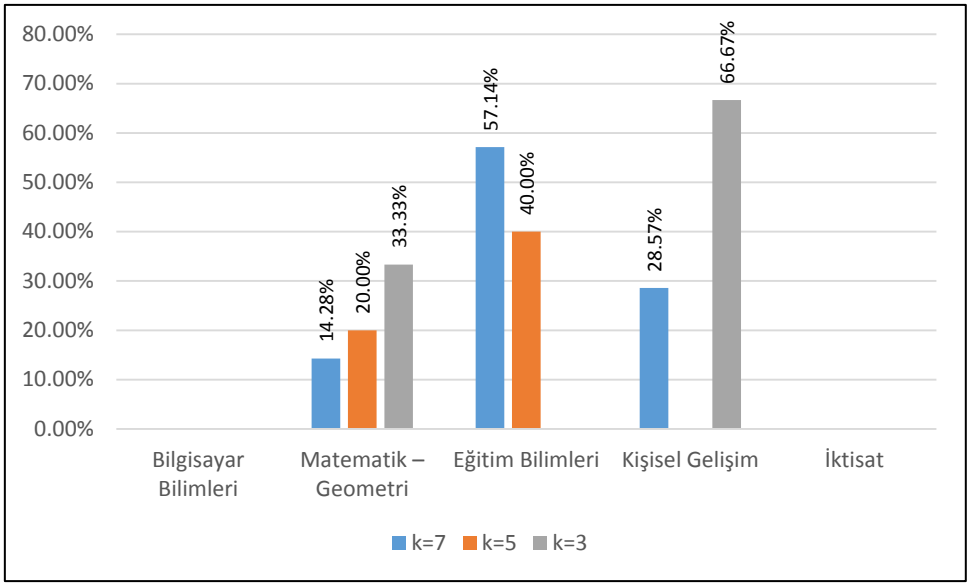
		Başarı Oranları				
Doküman Adı	***k* değeri**	**Bilgisayar Bilimleri(%)**	**Matematik – Geometri(%)**	**Eğitim Bilimleri(%)**	**Kişisel Gelişim - İletişim(%)**	**İktisat(%)**
Kişisel Gelişim İletişim (Test)	7	0,00	0,00	28,57	71,42	0,00
Kişisel Gelişim İletişim (Test)	5	0,00	0,00	40,00	60,00	0,00
Kişisel Gelişim İletişim (Test)	3	0,00	0,00	66,67	33,33	0,00



Şekil 6.36. Kişisel Gelişim - İletişim (test) dokümanı için farklı *k* değerleri sonuçları.

Çizelge 6.7. Eğitim Bilimleri kategorisindeki test dokümanının otomatik kategorilendirme sonuçları.

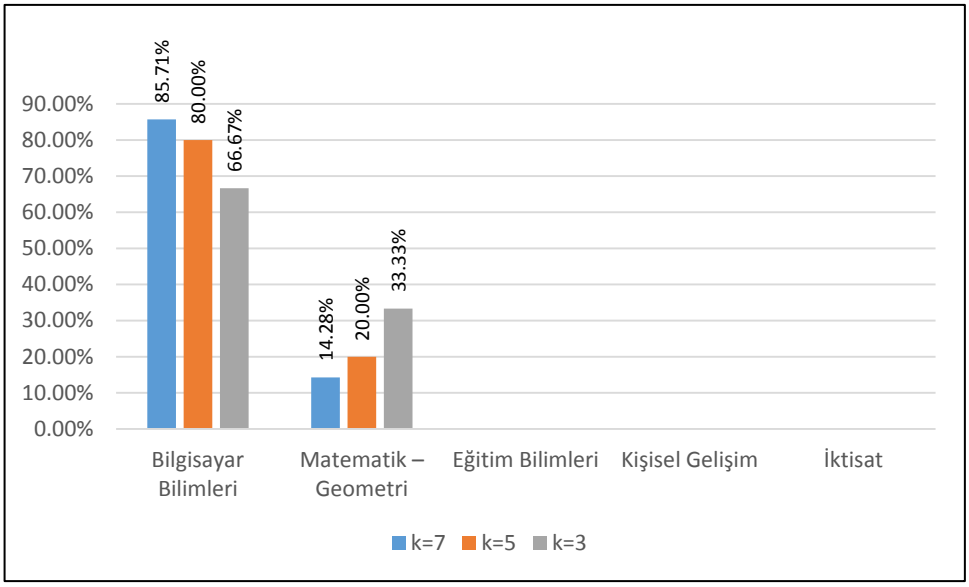
		Başarı Oranı				
Doküman Adı	***k* değeri**	**Bilgisayar Bilimleri (%)**	**Matematik – Geometri (%)**	**Eğitim Bilimleri (%)**	**Kişisel Gelişim - İletişim (%)**	**İktisat (%)**
Eğitim Bilimleri (Test)	7	0,00	14,28	57,14	28,57	0,00
Eğitim Bilimleri (Test)	5	0,00	20,00	40,00	40, 00	0,00
Eğitim Bilimleri (Test)	3	0,00	33,33	0,00	66,67	0,00



Şekil 6.37. Eğitim Bilimleri (test) dokümanı için farklı *k* değerleri sonuçları.

Çizelge 6.8. Bilgisayar bilimleri kategorisindeki test dokümanının otomatik kategorilendirme sonuçları.

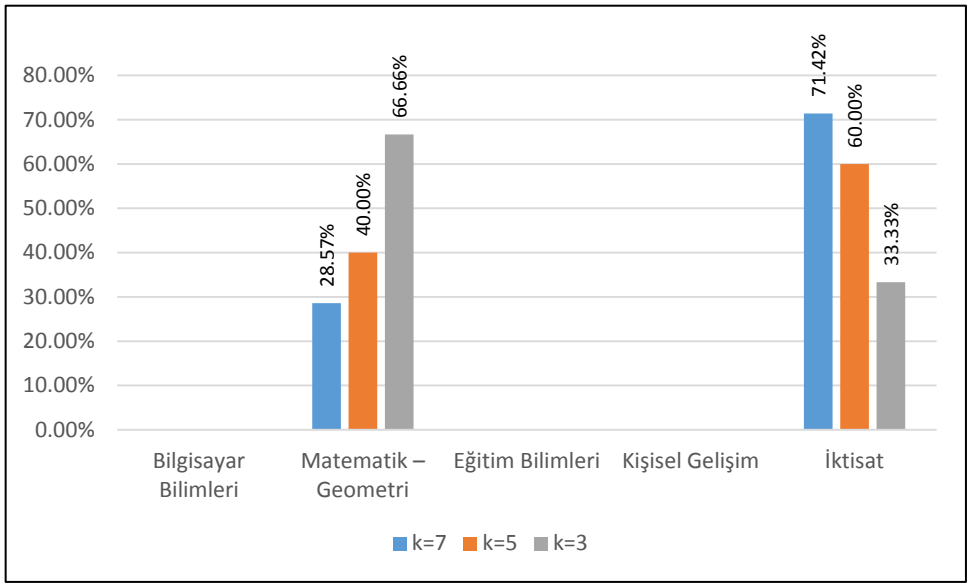
		Başarı Oranı				
Doküman Adı	***k* değeri**	**Bilgisayar Bilimleri (%)**	**Matematik – Geometri (%)**	**Eğitim Bilimleri (%)**	**Kişisel Gelişim - İletişim (%)**	**İktisat (%)**
Bilgisayar Bilimleri (Test)	7	85,71	14,28	0,00	0,00	0,00
Bilgisayar Bilimleri (Test)	5	80,00	20,00	0,00	0, 00	0,00
Bilgisayar Bilimleri (Test)	3	66,67	33,33	0,00	0,00	0,00



Şekil 6.38. Bilgisayar bilimleri (test) dokümanı için farklı *k* değerleri sonuçları.

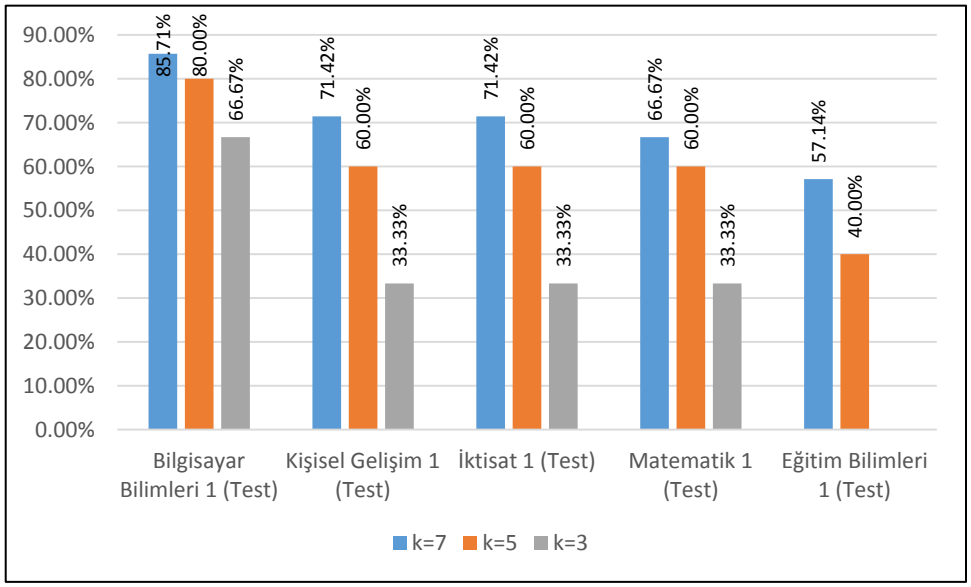
Çizelge 6.9. İktisat kategorisindeki test dokümanının otomatik kategorilendirme sonuçları.

		Başarı Oranları				
Doküman Adı	***k* değeri**	**Bilgisayar Bilimleri (%)**	**Matematik – Geometri (%)**	**Eğitim Bilimleri (%)**	**Kişisel Gelişim - İletişim (%)**	**İktisat (%)**
İktisat (Test)	7	0,00	28,57	0,00	0,00	71,42
İktisat (Test)	5	0,00	40,00	0,00	0, 00	60,00
İktisat (Test)	3	0,00	66,66	0,00	0,00	33,33



Şekil 6.39. İktisat (test) dokümanı için farklı *k* değerleri sonuçları.

5 farklı test dokümanı için *k=3*, *k=5* ve *k=7* değerleri için elde edilen sonuçlar Şekil 6.40'ta gösterilmektedir.



Şekil 6.40. Farklı *k* değerleri için 5 test dokümanının karşılaştırılması.

Şekil 6.40'ta gösterilen verilere bakarak en başarılı sonuçlar *k=7* seçildiğinde elde edilmişlerdir. *k=3* seçildiğinde ise yanlış sonuçlar alındığı gözlenmektedir. Uygulamada *k=7* seçildiğinde sınıflandırma işleminde en yüksek değerler Bilgisayar Bilimleri sınıfında alınmıştır. Şekil 6.36'ya göre *k=3* seçildiğinde Kişisel Gelişim -

İletişim dokümanının Eğitim Bilimleri dokümanı olarak sınıflandırıldığı görülmektedir. Bu durum Kişisel Gelişim - İletişim ve Eğitim Bilimleri alanlarındaki eğitim dokümanlarında kullanılan ortak kelimelerin çok olduğu düşünülebilir. Benzer durum Şekil 6.38'de Matematik ve Bilgisayar bilimleri alanlarında da görülmektedir.

Çizelge 6.10. Matematik – Geometri sınıfı *k=7* için sonuçlar.

		Dokümanın Tüm kategorilere En yüksek Yakınlıkları				
Test Dokümanı Adı	***k*-NN *k*=7 İçin Başarı Oranı (%)**	**Bilgisayar Bilimleri**	**Matematik – Geometri**	**Eğitim Bilimleri**	**Kişisel Gelişim - İletişim**	**İktisat**
Temel Matematik	100,00	0,000	0,128	0,000	0,000	0,000
Adım Adım matematik	85,70	0,110	0,162	0,000	0,000	0,000
12. Sınıf Matematik	85,70	0,000	0,065	0,000	0,000	0,042
11. Sınıf Matematik	85,70	0,000	0,126	0,000	0,000	0,085
6. Sınıf Matematik	85,70	0,000	0,125	0,120	0,000	0,000
Meraklısına Matematik	85,70	0,045	0,890	0,000	0,000	0,000
Çözümlü Matematik Soru Bankası	71,42	0,111	0,114	0,000	0,000	0,000
7. Sınıf Matematik	71,42	0,040	0,090	0,000	0,000	0,000
Meraklısına ilköğretim matematik	57,14	0,000	0,113	0,045	0,040	0,000
İstatistik Metotlar ve Uygulamalar	42,80	0,119	0,117	0,000	0,000	0,040
Ortalama	77,13					

Uygulama sonuçlarının daha net analiz edilmesi amacıyla her kategori için 10'ar test dokümanı sisteme aktarılmış ve *k=7, k=5, k=3* değerleri verilerek kategorilendirilmiştir. 5 kategori için elde edilen sonuçlar çizelgeler halinde gösterilmektedir.

Çizelge 6.11. Matematik – Geometri sınıfı *k=5* için sonuçlar.

		Dokümanın Tüm kategorilere En yüksek Yakınlıkları				
Test Dokümanı Adı	***k*-NN *k=5* İçin Başarı Oranı** (%)	**Bilgisayar Bilimleri**	**Matematik – Geometri**	**Eğitim Bilimleri**	**Kişisel Gelişim - İletişim**	**İktisat**
Temel Matematik	100,00	0,000	0,128	0,000	0,000	0,000
Adım Adım matematik	80,00	0,110	0,162	0,000	0,000	0,000
12. Sınıf Matematik	80,00	0,000	0,065	0,000	0,000	0,042
11. Sınıf Matematik	80,00	0,000	0,126	0,000	0,000	0,085
6. Sınıf Matematik	80,00	0,000	0,125	0,120	0,000	0,000
Meraklısına Matematik	80,00	0,045	0,890	0,000	0,000	0,000
Çözümlü Matematik Soru Bankası	60,00	0,111	0,114	0,000	0,000	0,000
Meraklısına ilköğretim matematik	60,00	0,000	0,113	0,045	0,040	0,000
7. Sınıf Matematik	60,00	0,040	0,090	0,000	0,000	0,000
İstatistik Metotlar ve Uygulamalar	14,29	0,119	0,117	0,000	0,000	0,040
Ortalama	69,43					

Çizelge 6.11'de *İstatistik metotlar ve uygulamalar* adlı doküman sınıflandırma işlemi sonucunda Bilgisayar kategorisi olarak yanlış sınıflandırılmıştır.

Çizelge 6.12. Matematik – Geometri sınıfı *k=3* için sonuçlar.

		Dokümanın Tüm kategorilere En yüksek Yakınlıkları				
Test Dokümanı Adı	***k*-NN *k=3* İçin Başarı Oranı** (%)	**Bilgisayar Bilimleri**	**Matematik – Geometri**	**Eğitim Bilimleri**	**Kişisel Gelişim - İletişim**	**İktisat**
12. Sınıf Matematik	100,00	0,000	0,065	0,000	0,000	0,000
Temel Matematik	100,00	0,000	0,128	0,000	0,000	0,000
Adım Adım matematik	80,00	0,110	0,162	0,000	0,000	0,000
Çözümlü Matematik Soru Bankası	66,66	0,111	0,114	0,000	0,000	0,000
11. Sınıf Matematik	66,66	0,000	0,126	0,000	0,000	0,085
6. Sınıf Matematik	66,66	0,000	0,125	0,120	0,000	0,000
Meraklısına ilköğretim matematik	66,66	0,000	0,113	0,045	0,000	0,000
7. Sınıf Matematik	66,66	0,040	0,090	0,000	0,000	0,000
Meraklısına Matematik	66,66	0,045	0,890	0,000	0,000	0,000
İstatistik Metotlar ve Uygulamalar	0,00	0,119	0,117	0,000	0,000	0,000
Ortalama Başarı Oranı	68,00					

Matematik – Geometri sınıfı için Çizelge 6.10, Çizelge 6.11 ve Çizelge 6.12'de 10 test dokümanı için sırasıyla *k=7*, *k=5* ve *k=3* için sınıflandırma işlemi sonuçları gösterilmiştir. Bu sınıflandırmaların ortalama değerleri Çizelge 6.13'te gösterilmektedir.

Çizelge 6.13. Matematik – Geometri sınıfının 10 farklı dokümanda farklı *k* değerleri için ortalamalar.

Özellik Seçimi	Count 175 *tf-idf* ağırlıklandırma		
Doküman Sınıfı	Matematik - Geometri		
	k=7	*k=5*	*k=3*
Sınıflandırmada Ortalama Başarı Yüzdesi (%)	77,13	69,43	68,00
Sınıflandırmada Genel Ortalama Başarı Yüzdesi (%)	71,52		

Çizelge 6.13'te Matematik – Geometri kategorisinde en yüksek başarı % 77,13 ile *k=7* seçildiğinde elde edilmiştir.

Çizelge 6.14. Bilgisayar Bilimleri sınıfı *k=7* için sonuçlar.

		Dokümanın Tüm kategorilere En yüksek Yakınlıkları				
Test Dokümanı Adı	***k*-NN *k=7* İçin Başarı Oranı** (%)	**Bilgisayar Bilimleri**	**Matematik – Geometri**	**Eğitim Bilimleri**	**Kişisel Gelişim - İletişim**	**İktisat**
Borland Delphi 7	57,14	0,123	0,112	0,084	0,000	0,000
Visual Basic .NET	71,42	0,096	0,000	0,000	0,094	0,000
Visual Studio 2010	85,70	0,116	0,000	0,000	0,000	0,014
Turbo C	85,70	0,089	0,060	0,000	0,000	0,000
Borland C++	85,70	0,110	0,086	0,060	0,000	0,000
Mikroişlemciler	71,42	0,113	0,140	0,080	0,000	0,000
Bilgisayar Donanımı	57,14	0,114	0,068	0,000	0,113	0,000
Kim Korkar Bilgisayardan	100,00	0,117	0,000	0,000	0,000	0,000
Python Programlama	71,42	0,110	0,112	0,000	0,000	0,075
Ağ Temelleri	85,70	0,080	0,084	0,000	0,000	0,000
Ortalama Başarı Oranı	77,13					

Çizelge 6.15. Bilgisayar Bilimleri sınıfı *k=5* için sonuçlar.

		Dokümanın Tüm kategorilere En yüksek Yakınlıkları				
Test Dokümanı Adı	***k*-NN *k=5* İçin Başarı Oranı** (%)	**Bilgisayar Bilimleri**	**Matematik – Geometri**	**Eğitim Bilimleri**	**Kişisel Gelişim - İletişim**	**İktisat**
Borland Delphi 7	40,00	0,123	0,112	0,084	0,000	0,000
Visual Basic .NET	80,00	0,096	0,000	0,000	0,094	0,000
Visual Studio 2010	80,00	0,116	0,000	0,000	0,000	0,014
Turbo C	80,00	0,089	0,060	0,000	0,000	0,000
Borland C++	40,00	0,110	0,086	0,060	0,000	0,000
Mikroişlemciler	60,00	0,113	0,140	0,080	0,000	0,000
Bilgisayar Donanımı	60,00	0,114	0,068	0,000	0,113	0,000
Kim Korkar Bilgisayardan	100,00	0,117	0,000	0,000	0,000	0,000
Python Programlama	60,00	0,110	0,112	0,000	0,000	0,075
Ağ Temelleri	80,00	0,080	0,084	0,000	0,000	0,000
Ortalama Başarı Oranı	68,00					

Çizelge 6.16. Bilgisayar Bilimleri sınıfı *k=3* için sonuçlar.

		Dokümanın Tüm Kategorilere En yüksek Yakınlıkları				
Test Dokümanı Adı	***k*-NN *k=3* İçin Başarı Oranı** (%)	**Bilgisayar Bilimleri**	**Matematik – Geometri**	**Eğitim Bilimleri**	**Kişisel Gelişim - İletişim**	**İktisat**
Borland Delphi 7	33,33	0,123	0,112	0,084	0,000	0,000
Visual Basic .NET	66,66	0,096	0,000	0,000	0,094	0,000
Visual Studio 2010	66,66	0,116	0,000	0,000	0,000	0,014
Turbo C	66,66	0,089	0,060	0,000	0,000	0,000
Borland C++	66,66	0,110	0,086	0,060	0,000	0,000
Mikroişlemciler	33,33	0,113	0,140	0,080	0,000	0,000
Bilgisayar Donanımı	66,66	0,000	0,068	0,000	0,113	0,000
Kim Korkar Bilgisayardan	33,33	0,117	0,000	0,000	0,000	0,000
Python Programlama	66,66	0,110	0,112	0,000	0,000	0,000
Ağ Temelleri	100,00	0,000	0,084	0,000	0,000	0,000
Ortalama Başarı Oranı	60,00					

Bilgisayar Bilimleri sınıfı için Çizelge 6.14, Çizelge 6.15 ve Çizelge 6.16'da 10 dokümanın sırasıyla *k=7, k=5* ve *k=3* için sınıflandırma işlemleri sonuçları gösterilmiştir. Bu sınıflandırmaların ortalama değerleri Çizelge 6.17'de gösterilmektedir.

Çizelge 6.17. Bilgisayar Bilimleri sınıfının 10 farklı dokümanda farklı *k* değerleri için ortalamalar.

Özellik Seçimi	Count 175 *tf-idf* ağırlıklandırma		
Doküman Sınıfı	Bilgisayar Bilimleri		
	k=7	*k=5*	*k=3*
Sınıflandırmada Ortalama Başarı Yüzdesi (%)	77,13	68,00	60,00
Sınıflandırmada Genel Ortalama Başarı Yüzdesi (%)	68,37		

Çizelge 6.17'de Bilgisayar Bilimleri kategorisinde en yüksek başarı % 77,13 ile *k=7* seçildiğinde elde edilmiştir. Bu kategoride ortalama sınıflandırma başarısı % 68,37 olarak belirlenmiştir.

Çizelge 6.18. Eğitim Bilimleri sınıfı *k=7* için sonuçlar

		Dokümanın Tüm Kategorilere En yüksek Yakınlıkları				
Test Dokümanı Adı	***k*-NN *k=7* İçin Başarı Oranı** (%)	**Bilgisayar Bilimleri**	**Matematik – Geometri**	**Eğitim Bilimleri**	**Kişisel Gelişim - İletişim**	**İktisat**
Sınıf Yönetimi	85,70	0,000	0,000	0,160	0,000	0,156
Eğitim Bilimlerine Giriş	85,70	0,000	0,000	0,140	0,141	0,000
Etkili Sınıf Yönetimi	71,42	0,093	0,000	0,125	0,120	0,000
Öğrenme Sanatı	85,70	0,000	0,000	0,122	0,000	0,098
Eğitimde Mükemmellik Anlayışı	57,14	0,211	0,000	0,082	0,085	0,084
Gelişim ve Öğrenme Sanatı	42,85	0,040	0,000	0,072	0,057	0,050
Öğrenen Okul	85,70	0,000	0,000	0,163	0,000	0,142
Birleştirilmiş Sınıflarda Öğretim	57,14	0,420	0,000	0,146	0,114	0,087
Gelişim Öğrenme ve Öğretim	57,14	0,163	0,000	0,172	0,086	0,000
Eğitim Psikolojisi	85,70	0,000	0,000	0,127	0,120	0,000
Ortalama Başarı Oranı	71,42					

Çizelge 6.19. Eğitim Bilimleri sınıfı *k=5* için sonuçlar.

		Dokümanın Tüm Kategorilere En yüksek Yakınlıkları				
Test Dokümanı Adı	***k*-NN *k=5* İçin Başarı Oranı** (%)	**Bilgisayar Bilimleri**	**Matematik – Geometri**	**Eğitim Bilimleri**	**Kişisel Gelişim - İletişim**	**İktisat**
Sınıf Yönetimi	100,00	0,000	0,000	0,160	0,000	0,156
Eğitim Bilimlerine Giriş	80,00	0,000	0,000	0,140	0,141	0,000
Etkili Sınıf Yönetimi	60,00	0,093	0,000	0,125	0,120	0,000
Öğrenme Sanatı	80,00	0,000	0,000	0,122	0,000	0,098
Eğitimde Mükemmellik Anlayışı	40,00	0,211	0,000	0,082	0,085	0,084
Gelişim ve Öğrenme Sanatı	20,00	0,040	0,000	0,072	0,057	0,050
Öğrenen Okul	80,00	0,000	0,000	0,163	0,000	0,142
Birleştirilmiş Sınıflarda Öğretim	40,00	0,420	0,000	0,146	0,114	0,087
Gelişim Öğrenme ve Öğretim	40,00	0,163	0,000	0,172	0,086	0,000
Eğitim Psikolojisi	80,00	0,000	0,000	0,127	0,120	0,000
Ortalama Başarı Oranı	62,00					

Çizelge 6.20. Eğitim Bilimleri sınıfı *k=3* için sonuçlar.

		Dokümanın Tüm Kategorilere En yüksek Yakınlıkları				
Test Dokümanı Adı	***k*-NN *k=3* İçin Başarı Oranı** (%)	**Bilgisayar Bilimleri**	**Matematik – Geometri**	**Eğitim Bilimleri**	**Kişisel Gelişim - İletişim**	**İktisat**
Sınıf Yönetimi	100,00	0,000	0,000	0,160	0,000	0,156
Eğitim Bilimlerine Giriş	66,66	0,000	0,000	0,140	0,141	0,000
Etkili Sınıf Yönetimi	66,66	0,000	0,000	0,125	0,120	0,000
Öğrenme Sanatı	100,00	0,000	0,000	0,122	0,000	0,000
Eğitimde Mükemmellik Anlayışı	33,33	0,000	0,000	0,082	0,085	0,084
Gelişim ve Öğrenme Sanatı	33,33	0,000	0,000	0,072	0,057	0,050
Öğrenen Okul	66,66	0,000	0,000	0,163	0,000	0,142
Birleştirilmiş Sınıflarda Öğretim	33,33	0,000	0,000	0,146	0,114	0,087
Gelişim Öğrenme ve Öğretim	66,66	0,000	0,000	0,172	0,086	0,000
Eğitim Psikolojisi	66,66	0,000	0,000	0,127	0,120	0,000
Ortalama Başarı Oranı	63,33					

Eğitim Bilimleri sınıfı için Çizelge 6.18, Çizelge 6.19 ve Çizelge 6.20'de 10 dokümanın sırasıyla *k=7, k=5* ve *k=3* için sınıflandırma işlemleri sonuçları

gösterilmiştir. Bu sınıflandırmaların ortalama değerleri Çizelge 6.21'de gösterilmektedir.

Çizelge 6.21. Eğitim Bilimleri sınıfının 10 farklı dokümanda farklı *k* değerleri için ortalamalar.

Özellik Seçimi	Count 175 *tf-idf* ağırlıklandırma		
Doküman Sınıfı	Eğitim Bilimleri		
	k=7	*k=5*	*k=3*
Sınıflandırmada Ortalama Başarı Yüzdesi (%)	71,42	62,00	63,33
Sınıflandırmada Genel Ortalama Başarı Yüzdesi (%)	65,58		

Çizelge 6.21'de Eğitim Bilimleri kategorisinde en yüksek başarı % 71,42 ile *k=7* seçildiğinde elde edilmiştir. *k=3* seçildiğinde başarı oranı *k=5*'e göre daha fazla olduğu görülmüştür.

Çizelge 6.22. Kişisel Gelişim – İletişim sınıfı *k=7* için sonuçlar.

		Dokümanın Tüm Kategorilere En yüksek Yakınlıkları				
Test Dokümanı Adı	***k*-NN *k=7* İçin Başarı Oranı** (%)	**Bilgisayar Bilimleri**	**Matematik – Geometri**	**Eğitim Bilimleri**	**Kişisel Gelişim - İletişim**	**İktisat**
Pazarlama İletişimi Yönetimi	85,70	0,000	0,000	0,000	0,173	0,172
Öteki Kuram	85,70	0,000	0,000	0,000	0,145	0,144
Temel Konuşma Teknikleri	100,00	0,000	0,000	0,000	0,142	0,000
Popüler Kültür ve iletişim	71,42	0,162	0,000	0,000	0,165	0,164
İletişim Nedir	57,14	0,124	0,000	0,126	0,125	0,119
İletişim ve Önemi	85,70	0,000	0,173	0,000	0,175	0,000
Kendini Ateşle	71,42	0,140	0,000	0,145	0,147	0,000
Dinleme Becerisi	85,70	0,000	0,000	0,140	0,138	0,000
İletişim Modelleri	85,70	0,000	0,000	0,000	0,179	0,172
Kişisel Gelişim	100,00	0,000	0,000	0,000	0,171	0,000
Ortalama Başarı Oranı	82,85					

Çizelge 6.23. Kişisel Gelişim – İletişim Sınıfı *k=5* için sonuçlar.

		Dokümanın Tüm Kategorilere En yüksek Yakınlıkları				
Test Dokümanı Adı	***k*-NN *k=5* İçin Başarı Oranı** (%)	**Bilgisayar Bilimleri**	**Matematik – Geometri**	**Eğitim Bilimleri**	**Kişisel Gelişim – İletişim**	**İktisat**
Pazarlama İletişimi Yönetimi	80,00	0,000	0,000	0,000	0,173	0,172
Öteki Kuram	80,00	0,000	0,000	0,000	0,145	0,144
Temel Konuşma Teknikleri	100,00	0,000	0,000	0,000	0,142	0,000
Popüler Kültür ve iletişim	60,00	0,162	0,000	0,000	0,165	0,164
İletişim Nedir	40,00	0,124	0,000	0,126	0,125	0,119
İletişim ve Önemi	100,00	0,000	0,000	0,000	0,175	0,000
Kendini Ateşle	60,00	0,140	0,000	0,145	0,147	0,000
Dinleme Becerisi	80,00	0,000	0,000	0,140	0,138	0,000
İletişim Modelleri	80,00	0,000	0,000	0,000	0,179	0,172
Kişisel Gelişim	100,00	0,000	0,000	0,000	0,171	0,000
Ortalama Başarı Oranı	78,00					

Çizelge 6.23'te Kişisel Gelişim – İletişim kategorisi test dokümanı olan *İletişim Nedir* adlı doküman *k=5* seçildiğinde *Eğitim Bilimleri* dokümanı olarak yanlış sınıflandırılmıştır. *k=7* ve *k=3* seçildiğinde ise doğru sınıflandırılmıştır.

Kişisel Gelişim – İletişim sınıfı için Çizelge 6.22, Çizelge 6.23 ve Çizelge 6.24'te 10 dokümanın sırasıyla *k=7, k=5* ve *k=3* için sınıflandırma işlemleri sonuçları gösterilmiştir. Bu sınıflandırmaların ortalama değerleri Çizelge 6.25'te gösterilmektedir.

Çizelge 6.24. Kişisel Gelişim – İletişim Sınıfı *k=3* için sonuçlar.

		Dokümanın Tüm Kategorilere En yüksek Yakınlıkları				
Test Dokümanı Adı	***k*-NN *k=3* İçin Başarı Oranı** (%)	**Bilgisayar Bilimleri**	**Matematik – Geometri**	**Eğitim Bilimleri**	**Kişisel Gelişim – İletişim**	**İktisat**
Pazarlama İletişimi Yönetimi	100,00	0,000	0,000	0,000	0,173	0,172
Öteki Kuram	66,66	0,000	0,000	0,000	0,145	0,144
Temel Konuşma Teknikleri	100,00	0,000	0,000	0,000	0,142	0,000
Popüler Kültür ve iletişim	66,66	0,000	0,000	0,000	0,165	0,164
İletişim Nedir	33,33	0,124	0,000	0,126	0,125	0,000
İletişim ve Önemi	100,00	0,000	0,000	0,000	0,175	0,000
Kendini Ateşle	66,66	0,000	0,000	0,145	0,147	0,000
Dinleme Becerisi	66,66	0,000	0,000	0,140	0,138	0,000
İletişim Modelleri	66,66	0,000	0,000	0,000	0,179	0,172
Kişisel Gelişim	100,00	0,000	0,000	0,000	0,171	0,000
Ortalama Başarı Oranı	76,66					

Çizelge 6.25. Kişisel Gelişim – İletişim sınıfının 10 farklı dokümanda farklı *k* değerleri için ortalamalar.

Özellik Seçimi	Count 175 *tf-idf* ağırlıklandırma		
Doküman Sınıfı	Kişisel Gelişim – İletişim		
	k=7	*k=5*	*k=3*
Sınıflandırmada Ortalama Başarı Yüzdesi (%)	82,85	78,00	76,66
Sınıflandırmada Genel Ortalama Başarı Yüzdesi (%)	79,17		

Çizelge 6.25'te Kişisel Gelişim – İletişim kategorisinde en yüksek başarı % 82,85 ile *k=7* seçildiğinde elde edilmiştir. Bu uygulamada en yüksek başarı oranı *Kişisel Gelişim – İletişim* kategorisinde elde edilmiştir.

Çizelge 6.26. İktisat sınıfı *k=7* için sonuçlar.

		Dokümanın Tüm Kategorilere En yüksek Yakınlıkları				
Test Dokümanı Adı	***k*-NN *k=7* İçin Başarı Oranı** (%)	**Bilgisayar Bilimleri**	**Matematik – Geometri**	**Eğitim Bilimleri**	**Kişisel Gelişim**	**İktisat**
İktisat Teorisi	85,70	0,000	0,000	0,000	0,260	0,266
Ekonometri Temel Kavramlar	71,42	0,253	0,000	0,244	0,000	0,249
Kalite ve Hayata İzdüşümler	100,00	0,000	0,000	0,000	0,000	0,283
Para Teorisi	85,70	0,000	0,000	0,000	0,270	0,272
Genel Muhasebe	71,42	0,263	0,000	0,000	0,000	0,272
Mikro İktisat	57,14	0,000	0,000	0,000	0,286	0,261
Devlet Bütçesi	71,42	0,000	0,286	0,000	0,000	0,289
Ekonomik Politikalar	85,70	0,360	0,000	0,000	0,000	0,251
Makro İktisat	71,42	0,000	0,291	0,000	0,288	0,000
Hizmetler Ekonomisi	57,14	0,110	0,000	0,096	0,000	0,113
Ortalama Başarı Oranı	75,71					

Çizelge 6.27. İktisat Sınıfı *k=5* için sonuçlar.

		Dokümanın Tüm Kategorilere En yüksek Yakınlıkları				
Test Dokümanı Adı	***k*-NN *k=5* İçin Başarı Oranı** (%)	**Bilgisayar Bilimleri**	**Matematik – Geometri**	**Eğitim Bilimleri**	**Kişisel Gelişim**	**İktisat**
İktisat Teorisi	80,00	0,000	0,000	0,000	0,260	0,266
Ekonometri Temel Kavramlar	80,00	0,253	0,000	0,000	0,000	0,249
Kalite ve Hayata İzdüşümler	100,00	0,000	0,000	0,000	0,000	0,283
Para Teorisi	80,00	0,000	0,000	0,000	0,270	0,272
Genel Muhasebe	80,00	0,263	0,000	0,000	0,000	0,272
Mikro İktisat	40,00	0,000	0,000	0,000	0,286	0,261
Devlet Bütçesi	80,00	0,000	0,286	0,000	0,000	0,289
Ekonomik Politikalar	80,00	0,360	0,000	0,000	0,000	0,251
Makro İktisat	60,00	0,000	0,291	0,000	0,288	0,000
Hizmetler Ekonomisi	40,00	0,110	0,000	0,096	0,000	0,113
Ortalama Başarı Oranı	72,00					

Çizelge 6.28. İktisat sınıfı *k=3* için sonuçlar.

		Dokümanın Tüm Kategorilere En yüksek Yakınlıkları				
Test Dokümanı Adı	***k*-NN *k=3* İçin Başarı Oranı** (%)	**Bilgisayar Bilimleri**	**Matematik – Geometri**	**Eğitim Bilimleri**	**Kişisel Gelişim**	**İktisat**
İktisat Teorisi	100,00	0,000	0,000	0,000	0,000	0,266
Ekonometri Temel Kavramlar	66,66	0,253	0,000	0,000	0,000	0,249
Kalite ve Hayata İzdüşümler	100,00	0,000	0,000	0,000	0,000	0,283
Para Teorisi	66,66	0,000	0,000	0,000	0,270	0,272
Genel Muhasebe	66,66	0,263	0,000	0,000	0,000	0,272
Mikro İktisat	66,66	0,000	0,000	0,000	0,286	0,261
Devlet Bütçesi	66,66	0,000	0,286	0,000	0,000	0,289
Ekonomik Politikalar	66,66	0,360	0,000	0,000	0,000	0,251
Makro İktisat	33,33	0,000	0,291	0,000	0,288	0,000
Hizmetler Ekonomisi	66,66	0,110	0,000	0,000	0,000	0,113
Ortalama Başarı Oranı	70,00					

İktisat sınıfı için Çizelge 6.26, Çizelge 6.27, Çizelge 6.28'de 10 dokümanın sırasıyla *k=7, k=5* ve *k=3* için sınıflandırma işlemleri sonuçları gösterilmiştir. Bu sınıflandırmaların ortalama değerleri Çizelge 6.29'da gösterilmektedir.

Çizelge 6.29. İktisat sınıfının 10 farklı dokümanda farklı *k* değerleri için ortalamalar.

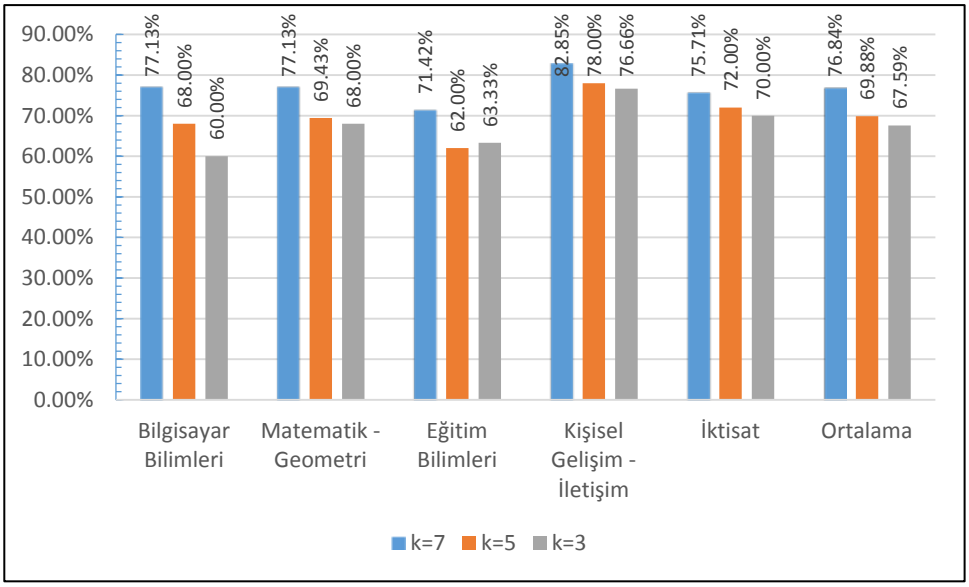
Özellik Seçimi	Count 175 *tf-idf* ağırlıklandırma		
Doküman Sınıfı	İktisat		
	k=7	*k=5*	*k=3*
Sınıflandırmada Ortalama Başarı Yüzdesi (%)	75,71	72,00	70,00
Sınıflandırmada Genel Ortalama Başarı Yüzdesi (%)	72,57		

Çizelge 6.29'da İktisat kategorisinde en yüksek başarı ortalaması % 75,71 ile *k=7* seçildiğinde elde edilmiştir. İktisat kategorisinde tüm *k* değerleri baz alındığında başarı yüzdesi % 72,57 olarak belirlenmiştir.

Çizelge 6.30. 5 kategoride farklı *k* değerleri için sınıflandırma başarı oranları.

Özellik Seçimi	**Count 175 tf-*idf* ağırlıklandırma**			
	k=7	***k=5***	***k=3***	**Kategori bazında *k* değerleri için Ortalama (%)**
Bilgisayar Bilimleri (%)	77,13	68,00	60,00	67,40
Matematik – Geometri (%)	77,13	69,43	68,00	71,52
Eğitim Bilimleri (%)	71,42	62,00	63,33	65,58
Kişisel Gelişim – İletişim (%)	82,85	78,00	76,66	79,17
İktisat (%)	75,71	72,00	70,00	72,57
Genel Ortalama (%)	76,84	69,88	67,59	

Uygulama sonucunda sınıflandırmada en yüksek başarı Kişisel Gelişim – İletişim kategorisinde elde edilmiştir. Bu kategori de *k=7* seçildiğinde % 82,85 oranında bir başarı elde edilmiştir. Elde edilen bu sonuçlar Çizelge 6.30'da gösterilmektedir. Tüm kategoriler değerlendirildiğinde sınıflandırmada *k=7* seçildiğinde başarı oranı en yüksek (% 76,84) çıkmıştır. Bu durumlar grafiksel olarak Şekil 6.41'de gösterilmektedir.



Şekil 6.41. Farklı *k* değerleri için sınıflandırmada başarı oranları grafiksel gösterimi.

BÖLÜM 7

SONUÇ VE ÖNERİLER

Dokümanları, önceden tanımlanmış kategorilere atama işlemine doküman kategorilendirme, doküman sınıflandırma denir. Elle kategorilendirme işleminin yavaş ve pahalı olması ile kategorilendirmede sürekli aynı sonucun alınamama ihtimali, manüel kategorilendirmeyi tercih edilen kategorilendirme olmaktan çıkarmaktadır. Ayrıca veri miktarının büyüklüğü, kategorilendirme işleminin bilgisayar programları aracılığıyla yapılmasını gerekli hale getirmiştir.

Kütüphanelerdeki iş yükünün fazla, personel sayısının yetersiz olduğu ve personel hatasının çok olduğu düşünülürse otomatik kategorilendirme işlemenin ne kadar önemli olduğu anlaşılmaktadır.

Bu çalışmada kütüphanelerde bulunan dokümanların dijital ortama aktarılma işlemine değinilmiştir. Bir dokümanda kapak, içindekiler, önsöz, giriş gibi bölümlerin sisteme aktarılıp OKT metodu ile bu bilgilerin metin olarak veritabanında saklanması sağlanmıştır. Bu verilerin kütüphane dokümanlarını tararken ne kadar faydalı olacağı açıktır. Ayrıca kütüphanelerin kalbini oluşturan kütüphane otomasyonu, OKT sistemi, veri ve metin madenciliği konuları alt başlıklar halinde verilmiş ve web tabanlı bir kütüphane otomasyon sistemi yazılmıştır.

Bu tez çalışmasında eğitim ve test dokümanlarının alınıp metne çevrilip kategorilendirilmesine kadar olan bütün işlemler gerçekleştirilmiştir. İçindekiler, önsöz, giriş, özet sayfalarındaki kelime köklerinin bulunması için zemberek kütüphanesi kullanılarak "*kelime_koklerini_bul.jar*" adında ayrıca bir yazılım geliştirilmiş ve program üzerinden çağrılarak çalıştırılmıştır.

Zemberek kütüphanesinde yer almayan kelimeler olabilir. Zemberek kütüphanesi tarafından çözülemeyen kelimeler veri tabanına eklenmez. Bu durum eğitim ya da test dokümanları için önemli olabilecek kelimelerin gözardı edilmesi demektir ve sistemin performansını düşürecektir.

5 kategoride toplam 150 eğitim dokümanı taranarak sistem eğitilmiştir. Her kategoride 10 doküman test edilerek ortalama % 76,84 doğru kategorilendirme yapılarak büyük başarı elde edilmiştir.

tf-idf ağırlıklandırma, sınıf özellik vektörü ve *k*-NN sınıflandırma algoritması kullanılarak sınıflandırma gerçekleştirilmiştir. *k*-NN algoritmasında, kosinüs benzerliği ile birlikte *k=3, k=5* ve *k=7* değerleri alınarak sınıflandırma gerçekleştirilmiştir. Bilgisayar Bilimleri, Matematik-Geometri, Eğitim Bilimleri, Kişisel Gelişim ve İktisat sınıfları, dokümanların atandığı sınıflardır. Uygulamanın sonuçlarına göre *k*-NN algoritması için en başarılı sonuçlar *k=7* alınarak elde edilmiştir.

Taranan dokümanların metne çevrilmesinde Microsoft firmasının geliştirmiş olduğu MODI eklentisi kullanılmıştır. Bu eklenti dijital baskı ile basılmış dokümanlarda son derece başarılıdır fakat el yazımı dokümanlarda ve düşük kalitede taranmış dokümanlarda başarısı düşüktür. El yazması ve düşük kalitedeki dokümanlar için performansı yüksek bir OKT yazılımı kullanılması gerekmektedir.

OKT yazılımlarının başarısı dokümanın kalitesine, baskı özelliğine, tarama kalitesine bağlıdır. Düşük çözünürlük ile taranmış bir belgeyi metne çevirirken belli başlı sorunlar vardır. Bu sorunlardan uygulama esnasında karşılaşılmıştır. Örneğin "*m*" harfi "rn" harfleri olarak çevrildiği gözlemlenmiştir. Donanım kelimesi Donanırn olarak zemberek kütüphanesine aktarıldığında tanımlanmayan kelime olarak algılanacak ve kökü bulunamadığından veritabanına aktarılmayacaktır. Bu probleme çözüm olarak kelimelerin tahmin edilmesi düşünülebilir.

Bu çalışmada *k*-NN algoritması uygulanırken; örneğin *k=7* için, bir test dokümanının kategorilendirilmesi, kendisine en yakın yedi doküman alınıp bunların içerisinde en

çok hangi kategorideki doküman varsa o kategoriye atanarak yapılmıştır. Farklı bir yaklaşım olarak test dokümanına en yakın yedi dokümanın yakınlıklarının ortalaması hesaplanarak çıkan sonuca en yakın eğitim dokümanının kategorisi test dokümanın kategorisi olarak değerlendirilebilir.

Bu alanda yapılacak çalışmalara yönelik olarak eğitim ve test dokümanlarının sayısı artırılıp, sınıflandırma aşamasında kullanılan yöntemin etkinliği daha iyi gözlemlenebilir. Bu çalışmada kullanılan *k*-NN sınıflandırma algoritmasının yanında farklı algoritmalar kullanılarak daha yüksek başarılar elde edilebilir ve sonuçları kıyaslanabilir. Kullanılan *tf-idf*, dışında sözcük ağırlıklandırma yöntemleri (yapay sinir ağları, destek vektör makineleri) ve diğer özellik seçim teknikleri kullanılarak sınıflandırma işlemleri gerçekleştirilebilir.

KAYNAKLAR

Al, U. ve Bahşişoğlu, H., "Türkiye'deki üniversite kütüphanelerine ait web sitelerinin içerik açısından değerlendirilmesi", ***Bilgi Dünyası,*** 1 (2): 307-329 (2000).

Akbulut, S., "Veri madenciliği teknikleri ile bir kozmetik markanın ayrılan müşteri analizi ve müşteri segmentasyonu", Yüksek Lisans Tezi***, Gazi Üniversitesi, Fen Bilimleri Enstitüsü***, Ankara, 91-95 (2006).

Amasyalı, M. F. and Diri, B., "Automatic Turkish text categorization in terms of author, genre and gender", ***11th International Conference on Applications of Natural Language to Information Systems***, Austria, 221-226 (2006).

Amasyalı, M. F. ve Yıldırım, T., "Otomatik haber metinleri sınıflandırma", ***Signal Processing and Communications Applications (SIU 2004), 2004 IEEE 12th Conference on***, Aydın, 224-226 (2004).

Aşlıyan, R. ve Günel, K., "Metin içerikli Türkçe doküman sınıflandırılması", ***Akademik Bilişim 2010***, Muğla, 529-535 (2010).

Bayram, U. ve Çetinkaya, V., "Kütüphane otomasyonu", ***IV. Otomasyon Sempozyumu***, Samsun, 69-71 (2008).

Berry, M. J. and Linoff, G. S., "Mastering Data Mining, 2nd ed.", ***John Wiley & Sons***, New York, 5-407 (2000).

Bilekdemir, G., "Veri madenciliği tekniklerini kullanarak üretim süresi tahmini ve bir uygulama", Yüksek Lisans Tezi, ***Dokuz Eylül Üniversitesi, Fen Bilimleri Enstitüsü***, İzmir, 55-60 (2010).

Bishop C., "Neural Networks For Pattern Recognition", ***University of Oxford***, Oxford, 188-200 (1996).

Bloomberg, M. and Evans, G., "Kütüphane teknisyenleri için teknik hizmetlere giriş", Çeviri Editörü: Nilüfer Tuncer, ***Türk Kütüphaneciler Derneği***, Ankara, 19-25 (1989).

Çakmak, T. ve Özel, N., "Çevrimiçi kütüphane kataloglarının sosyal ağlarla yeniden yapılandırılması: yazılımlar ve projeler", ***Ünak 2010 Bildirileri***, Samsun, 29-36 (2010).

Çankırı, S., Kartal, E. Yıldırım, K. ve Gülseçen, S., "Organizasyonlarda bilgi yönetimi sürecinde veri madenciliği yaklaşımı", ***Ünak 2009 Bildirileri***, İstanbul, 75-82 (2009).

Dasarathy, B. V., "Nearest-neighbor classification techniques", ***IEEE Computer Society Press***, Los Alamitos, California, 102-120 (1991).

Erol, N., "Orta Doğu Teknik Üniversitesi'nin mevcut ödünç verme sisteminin otomasyon tasarımı ve sistem analizi", Yüksek Lisans Tezi, ***Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü***, Ankara, 20-22 (1990).

Erol, U. ve Gülseçen, S., "Naive bayes kullanarak çevrimiçi haber sınıflandırma", ***26. Bilişim Kurultayı Bildiriler Kitabı***, Ankara, 83-86 (2009).

Fayyad, U., Madigan, D. and Smyth, P., "From data mining to knowledge discovery in databases", ***AI Magazine***, 17 (3): 37-54 (1996).

Feldman, R. and Sanger J., "The text mining handbook advanced approaches in advanced approaches in analyzing unstructured data", ***University of Cambridge***, Cambridge, 85-90 (2007).

Frawley, W. J. and Matheus, C. J., "Knowledge discovery in databases", ***AAAI Press***, California, 13 (3): 1-27 (1991).

Gorman, M., "Why teach cataloguing and classification?", ***Cataloging and Classification Quarterly***, 34 (2): 1-13 (2002).

Han, J. and Kamber, M., "Data mining: concepts and techniques 2nd ed.", ***Morgan Kaufmann Publishers***, San Francisco, 348-350 (2006).

Haravu, L. J., "Emerging initiatives in library managment systems", ***International Conference on Academic Libraries***, Delhi, 239-248 (2009).

Hui, S. and Jha, G., "Application data mining for customer service support", ***Information and Management***, 38 (1): 1-13 (2000).

İlhan, G., "Türkiye'de kütüphane otomasyonu ve sorunları", Yüksek Lisans Tezi, ***Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü***, Ankara, 10-16 (1988).

İlhan, S., Duru, N., Karagöz, Ş. ve Sağır, M., "Metin madenciliği ile soru cevaplama sistemi", ***Eleco 2008***, Bursa, 356-359 (2008).

İnternet: Tonta, Y., Küçük, M. E., Al, U., Alır, G., Ertürk, K. L., Olcay, N. E., Soydal, İ. ve Ünal, Y., "Hacettepe Üniversitesi elektronik tez projesi", **http://yunus.hacettepe.edu.tr/~tonta/yayinlar/02-G-064-elektronik-tez-projesi-sonuc- raporu.pdf** (2013).

İnternet: Ulusal Bilgi Güvenliği Kapısı, "Yazılım Geliştirme Süreçleri ve Iso 27001 Bilgi Güvenliği Yönetim Sistemi - Ulusal Bilgi Güvenliği Kapısı", **http://www.bilgiguvenligi.gov.tr/yazilim-guvenligi/yazilim-gelistirme-surecleri-ve-iso-27001-bilgi-guvenligi-yonetim-sistemi.html** (2013).

Jun, H. and Hokuan, H., "An algorithm for text categorization with svm", ***Tencon '02 Proceedings***, Beijing, 47-50 (2002).

Karaca, M. F., "Metin madenciliği yöntemi ile haber sitelerindeki köşe yazılarının sınıflandırılması", Yüksek Lisans Tezi, ***Karabük Üniversitesi, Fen Bilimleri Enstitüsü***, Karabük, 58-65 (2012).

Keenan, S. and Johnston, C., "Concise dictionary of library and information science", ***DE Gruyter***, London, 65-72 (2000).

Koç, K., "Kütüphanelerde bilgisayarlaşma ve internet kullanımının hizmet verimliliğine katkısı", ***Türk Kütüphaneciliği Dergisi,*** 13 (2): 288-292 (1999).

Konchady, M., "Text Mining Application Programming 1st ed.", ***Charles River Media***, Boston, (2006).

Kurulgan, M. ve Bayram, F., "Üniversite kütüphaneleri web sitelerinin biçim ve içerik analizi: Türkiye'deki uygulamaya ilişkin bir araştırma", ***Hakemli Yazılar Türk Kütüphaneciliği***, 20 (2): 141-172 (2006).

Küçük, M. E., "Kütüphanelerde www kullanımı", ***Türk Kütüphaneciliği***, 13 (3): 267- 275 (1999).

Külcü, Ö., "Toplumsal ve ekonomik değişim sürecinde bilgi ve bilgi hizmetleri", **http://tk.kutuphaneci.org.tr/index.php/tk/article/view/1751/3501** (2011).

Losiewicz, P., Oard, D. W. and Kostoff, R. N., "Textual data mining to support science and technology management", ***Journal of Intelligent Information Systems***, 15 (2): 99-119 (2000).

Mitchell, T., "Machine Learning 1st ed.", ***McGraw-Hill***, New York, 166-168 (1997).

Musayev. E., "Bilgisayar destekli karakter tanıma sistemi tasarımı", Yüksek Lisans Tezi, ***İstanbul Üniversitesi, Fen Bilimleri Enstitüsü***, İstanbul, 69-71 (2004).

Parker, J. R., "Algorithms for Image Processing and Computer Vision 2nd ed.", ***John Wiley and Sons Inc***, New York, 234-544 (1997).

Silahtaroğlu, G., "Kavram ve Algoritmalarıyla Temel Veri Madenciliği 2.Baskı", ***Papatya Yayıncılık***, İstanbul, 26-28 (2008).

Soucy, P. and Mineau, G. W., "A simple *k*-NN algorithm for text categorization", ***Proceedings IEEE International Conference on Data Mining***, California, 647-648 (2001).

Şekerci, M., "Birleşik ve eğik Türkçe el yazısı tanıma sistemi", Yüksek Lisans Tezi, ***Trakya Üniversitesi, Fen Bilimleri Enstitüsü,*** Edirne, 105-110 (2007).

Takçı, H. ve Soğukpınar, İ., "Kütüphane kullanıcılarının erişim örüntülerinin keşfi", ***Bilgi Dünyası***, 3 (1): 12-26 (2002).

Torunoğlu, D., Çakırman, E., Ganiz, M. C., Akyokuş, S. ve Gürbüz, M. Z., "Analysis of preprocessing methods on classification of Turkish texts", ***Innovations in Intelligent Systems and Applications***, İstanbul, 112-117 (2011).

Türkoğlu, İ., "Yapay sinir ağları ile nesne tanıma", Yüksek Lisans Tezi, ***Fırat Üniversitesi, Fen Bilimleri Enstitüsü***, Elazığ, 56-60 (1996).

Özekes, S., "Veri madenciliği modelleri ve uygulama alanları", ***İstanbul Ticaret Üniversitesi Dergisi***, 1 (3): 65-82 (2003).

Özüsağlam, E., Selçuk, M. ve Fen, L., "Aksaray Üniversitesi kütüphane yazılımı seçimi", ***Akademik Bilişim'09 Bildirileri,*** Şanlıurfa, 563-566 (2009).

Vahaplar, A. ve İnceoğlu, M., "Veri madenciliği ve elektronik ticaret", ***Türkiye'de İnternet Konferansları***, İstanbul, 1-3.(2001).

Witten, I. H., Boddie, S. J., Bainbridge, D. and McNab, R. J., "Greenstone: a comprehensive open-source digital library software system", ***In Proceedings of the Fifth ACM Conference on Digital Libraries***, San Antonio, 113-121 (2000).

Yang, Y. and Liu, X., "A re-examination of text categorization methods", ***22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval***, Berkeley, 42-49 (1999).

Yılmaz, A. ve Aslan, H., "Veri tabanına dayalı Türkçe kütüphane otomasyonu yazılımı: Kybele", ***Türk Kütüphaneciliği***, 6 (1): 10-17 (1992).

Zhong, N. and Zhou, L., "Methodologies for knowledge discovery and data mining", ***Third Pacific-Asia Conference***, Beijing, 42-32 (1999).

Zohar, E. Y., "Introduction to text mining, supercomputing" ***Automated Learning Group National Center for Supercomputing Applications***, University of Illinois, Illinois, 35-45 (2002).

ÖZGEÇMİŞ

Selim ÖZDEM 1981 yılında Çorum – Alaca'da doğdu. İlkokulu Samsun'da, ortaokulu Çorum'da tamamladı. 1998 yılında Çorum Endüstri Meslek Lisesi Elektronik Bölümünden mezun oldu. 2001 yılında Çukurova Üniversitesi Kozan Meslek Yüksekokulu, Bilgisayar Programcılığı bölümünü bölüm birinciliği ile tamamladı. 2003 yılında Fırat Üniversitesi Elektronik Öğretmenliği bölümüne başladı 3. Sınıfa geçerken tekrar sınavlara hazırlandı ve 2005 yılında aynı üniversitenin Bilgisayar Öğretmenliği bölümünü kazandı. 2008 yılında Bilgisayar Öğretmeni olarak mezun oldu. 2009 – 2010 eğitim öğretim döneminde Çorum Özel Pınar Kolejinde Bilgisayar Öğretmeni olarak görev yaptı. 2010 yılında Hakkâri Üniversitesi Yüksekova Meslek Yüksekokulunda Öğretim Görevlisi olarak göreve başladı ve halen görevine devam etmektedir.

ADRES BİLGİLERİ

Adres : Hakkâri Üniversitesi
Yüksekova Meslek Yüksekokulu
Merkez/ Hakkâri

Tel : (532) 179 34 51
E-posta : selimozdem@gmail.com