T.C.
SELÇUK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BULANIK KÜMELEME KULLANILARAK BENZER BELGE ARANMASI

Rıdvan SARAÇOĞLU

DOKTORA TEZİ

ELEKTRİK-ELEKTRONİK MÜHENDİSLİĞİ ANABİLİM DALI

T.C.
SELÇUKÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BULANIK KÜMELEME KULLANILARAK BENZER BELGE ARANMASI

Rıdvan SARAÇOĞLU

DOKTORA TEZİ
ELEKTRİK-ELEKTRONİK MÜHENDİSLİĞİ ANABİLİM DALI

Bu tez 15.08.2007 tarihinde aşağıdaki jüri tarafından oybirliği/oyçokluğu ile kabul edilmiştir.

Prof. Dr. Novruz ALLAHVERDİ
(Danışman)

Prof. Dr. Ahmet ARSLAN
(Üye)

Prof. Dr. İnan GÜLER
(Üye)

Yrd. Doç. Dr. Salih GÜNEŞ
(Üye)

Yrd. Doç. Dr. Mehmet ÇUNKAŞ
(Üye)

ÖZET

Doktora Tezi

BULANIK KÜMELEME KULLANILARAK BENZER BELGE ARANMASI

Rıdvan SARAÇOĞLU

Selçuk Üniversitesi Fen Bilimleri Enstitüsü

Elektrik-Elektronik Mühendisliği Anabilim Dalı

Danışman: Prof. Dr. Novruz ALLAHVERDİ

2007, 128 sayfa

Jüri: Prof. Dr. Novruz ALLAHVERDİ

Prof. Dr. Ahmet ARSLAN

Prof. Dr. İnan GÜLER

Yrd. Doç. Dr. Salih GÜNEŞ

Yrd. Doç. Dr. Mehmet ÇUNKAŞ

Günümüzde teknolojinin gelişmesi ile birlikte her geçen gün büyük miktarlarda veriler ortaya çıkmaya ve depolanmaya başlanmıştır. Bu verilerden faydalanmanın yolu ise onların verimli bir şekilde organize edilmesi ve yararlı bilgilere dönüştürülmesinden geçmektedir. Bunu amaçlayan veri madenciliğinin bir çeşidi ise metinsel veriler üzerinde çalışan metin madenciliğidir. Metinsel belgelerin kullanışlı bir şekilde organize edilmesi, işlenmesi ve faydalı bilgiler çıkarılması gibi amaçları yerine getirmek için gerekenlerin başında metin sınıflandırıcısı, metinsel belge arama mekanizmaları vb. araçlar gelmektedir.

Bir metinsel belge arama işlemini iki farklı yaklaşımla ele almak mümkündür. Bunlardan biri geniş bir alandaki belgeler üzerinde anahtar kelime seçilmesine dayalı olarak arama yapmaktır (internet arama motorları gibi). Bir diğeri ise daha dar bir

alanda metinin tüm kelimelerini kullanmak suretiyle daha ayrıntılı bir arama yapmaktır (bir kütüphanedeki kitaplar üzerinde yapılacak arama gibi). Bu çalışmada ele alınan konu ise bulanık kümeleme ve metinlerin tüm kelimelerini kullanarak bir arama yaklaşımı ortaya koymaktır. Bu yaklaşım; önişleme, kümeleme/sınıflandırma ve benzerlik ölçümü olmak üzere üç temel aşamadan oluşmaktadır.

Bu çalışmada önişleme aşaması ile ilgili olarak terim ağırlıklandırma yöntemleri üzerinde durulmuştur. Bulanık kümeleme kullanıldığından dolayı mevcut terim ağırlıklandırma yöntemlerinin bulanık kümeleme ile birlikte kullanımları incelenmiş ve performansları karşılaştırılmıştır. En iyi performansı gösteren yöntem belirlenerek daha sonraki aşamalarda bu yöntem kullanılmıştır.

Benzerlik ölçümü aşaması için ise mevcut benzerlik ölçümlerinin önerilen arama yaklaşımındaki performansları incelenmiştir. Yine bu aşama için verinin boyutuna dayalı yeni bir benzerlik ölçümü önerilmiştir. Bu önerilen yeni benzerlik ölçümünün süre ve verimlilik açılarından önceki yöntemlere göre daha iyi olduğu görülmüştür.

Son olarak, bir test belgesinin birden fazla kategoriye ait olması şeklinde özetlenebilecek olan çoklu kategori problemi ele alınmıştır. Bu problemin çözümü için önerilen arama yaklaşımının kümeleme/sınıflandırma aşaması geliştirilmeye çalışılmıştır. Bu amaçla hangi belgelerin birden fazla kategoriye ait olduklarını tespit etmek için mevcut sınıflandırma yöntemi probleme adapte edilmiştir. Ayrıca, kategorilerin arasında bir ilişki matrisi oluşturularak, bir belge birden fazla kategoriye ait ise bunların hangi kategoriler oldukları tespit edilmeye çalışılmıştır. Önceki çalışmalarda pek yer verilmemiş olan bu çoklu kategori probleminde önemli ölçüde bir başarı sağlanmıştır.

Anahtar Kelimeler – Benzer belge arama, bulanık kümeleme, bulanık benzerlik sınıflandırması, terim ağırlıklandırma, benzerlik ölçümü, çoklu kategori problemi

ABSTRACT

PhD Thesis

SEARCHING FOR SIMILAR DOCUMENTS USING FUZZY CLUSTERING

Rıdvan SARAÇOĞLU

Selçuk University

Graduate School of Natural and Applied Sciences

Department of Electrical-Electronics Engineering

Supervisor: Prof. Dr. Novruz ALLAHVERDİ

2007, 128 pages

Jury: Prof. Dr. Novruz ALLAHVERDİ

Prof. Dr. Ahmet ARSLAN

Prof. Dr. İnan GÜLER

Asist. Prof. Dr. Salih GÜNEŞ

Asist. Prof. Dr. Mehmet ÇUNKAŞ

Nowadays, large amount of data has started to arise and stored by development of technology. The way of benefitting these data are to organize them efficiently and convert them to useful information. A kind of data mining that aims this is text minig which works over textual data. The first of necessities for implementing the aims like being organized textual documents usefully, being processed them and extracted useful information are text classifier, textual document search mechanisms and tools like them.

It is possible to discuss a textual document search operation with two diffrent approaches. One of them is to perform a search that bases on selection of a keyword in a large area (like internet search engines). The other is to perform a more detailed search by using all the words of text (a search that will be performed on the books in a library). The subject that is discusses in this study is to produce a search approach by using fuzzy clustering and all the words of text. This approach consists of three main stages like pre-processing, clustering/classification and similarity measurement.

In this study, term weighting methods have been emphasized related to pre-processing stage. Because of using fuzzy clustering, the usage of existing term weighting methods with fuzzy clustering has been investigated and their performances have been compared.The method which shows the best performance has been determined and this method has been used in the following stages.

For similarity measurement stage, the performances of existing similarity measurements in suggested search approach, have been investigated. Still for this stage, a new similarity measurement that bases on the size of data has been suggested. It is seen that this new similarity method that is suggested, is better than previous methods in terms of time and efficiency.

As last, multiple category problems that can be summarized as a test document belonging to more that one category, has been discussed. Clustering/classification stage of the suggested search approach for solution of this problem has been tried to develop. For this aim, existing classification method has been adapted to the problem to determine which documents belong to more than one category. Besides, the categories have been tried to determine by being formed a relation matrix, if a document belongs to more than one category. In this multiple category problem that is not seen in the previous studies, a great amount of achievement has been obtained.

Keywords – Searching similiar document, Fuzzy clustering, Fuzzy similarity classification, Term weighting, Similarity measurement, Multiple category problem

TEŞEKKÜR

Bu çalışmada bana yol gösteren ve her türlü bilimsel katkıyı sağlayan değerli hocam ve danışmanım Sayın Prof. Dr. Novruz ALLAHVERDİ'ye teşekkür ederim. Tez aşamam boyunca bana her türlü desteği sağlayan Tez İzleme Komitesi üyeleri S.Ü. Bilgisayar Mühendisliği Bölüm Başkanı Sayın Prof. Dr. Ahmet ARSLAN'a ve S.Ü. Elektrik-Elektronik Mühendisliği Bölümü Öğretim Üyesi Sayın Yrd. Doç. Dr. Salih GÜNEŞ'e teşekkür ederim.Ayrıca çalışmam esnasında büyük katkıları bulunan mesai arkadaşım Öğr. Gör. Kemal TÜTÜNCÜ'ye de teşekkür ederim.

Bu çalışma boyunca her türlü sabrı, hoşgörüyü ve fedakârlığı gösteren eşim Esra SARAÇOĞLU'na, ailemize ve can dostlarım Aşkın, Bülent, Erdal, Ferman, Orhan ve Salim'e (ABEFORS) en derin şükranlarımı sunarım.

Rıdvan SARAÇOĞLU

Konya, 2007

İÇİNDEKİLER

SİMGELER ve KISALTMALAR

D	Eğitim veri seti
T	Terim kümesi
C	Kategori kümesi
d_i	i numaralı belge
t_j	j numaralı terim
w_{ij}	i numaralı belgedeki j numaralı terimin ağırlığı
c_k	k numaralı kategori
R	Terim-Kategori Matrisi
MC	Çoklu Kategori Matrisi
$\mu_R(t_i,c_j)$	i numaralı terimin j numaralı kategoriye üyelik derecesi

KDH	Kendini Düzenleyen Harita (Self-Organizing Map)
DVM	Destek Vektör Makinesi (Support Vector Machine)
TF	Terim Sıklığı (Term Frequency)
TFIDF	Terim Sıklığı Ters Belge Sıklığı (Term Frequency Inverse Document Frequency)
kNN	k En Yakın Komşu (k Nearest Neighbour)
FSC	Bulanık Benzerlik Sınıflandırması (Fuzzy Similarity Classification)

1. GİRİŞ

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

İşlenmesi gereken bu büyük miktardaki verinin de yine büyük bir kısmını metinsel veriler oluşturmaktadır. Buna en önemli örnek günümüzde sayıları milyarları aşan Web sayfaları verilebilir. Bir başka örnek ise elektronik yayınlar veya elektronik ortama aktarılmış yayınlardır. Günümüzde süreli yayınlanan çok sayıdaki bilimsel dergi veya bilgisayar ortamına aktarılmış birçok kitap yine metinsel verilere örnek olarak verilebilir. Veri madenciliğinin bir çeşidi olan metin madenciliği de büyük miktardaki bu yapısız veya yarı yapılı metinsel veri üzerinden bilgi kazanımını veya bu verilerin verimli yönetimini amaçlamaktadır.

1.1. Benzer Belge Aranması Kavramı ve Temelleri

Bu çalışmanın temelini teşkil eden benzer belge aranması problemi metin madenciliğinin temel problemlerden birisidir. Bu problem, bir belge koleksiyonu içerisinden elimizdeki belgeye benzeyen belgelerin tespit edilmesi şeklinde özetlenebilir. Bu benzeri aranan belge bilimsel bir araştırma makalesi olabileceği gibi bir haber metni de olabilir.

Benzer belge aranması birçok kullanım alanı bulmaktadır. Artık hayatımızda önemli bir yer tutan internet arama motorları bu kullanıma güzel bir örnektir. Verimli, hızlı ve en önemlisi; kullanışlı ve faydalı sonuçlar getiren bir arama aracı, günlük hayatımızda önemli bir eksikliği dolduracaktır.

Benzer belge aranması soru cevaplandırma sistemlerinin ve bu alandaki çalışmalarında önemle üzerinde durduğu bir konudur. Bu konuda varılmak istenen son nokta, soru cevaplandırma sisteminin insana yakın bir şekilde cevap üretme yeteneğine sahip olmasıdır.

Bu arama işlemi bir kütüphanede araştırma veya arama yaparken de kullanılabilir. Mevcut veritabanı sorgularının ötesine geçerek kütüphanedeki tüm kitapların içindekiler bölümleri veya özetleri üzerinden daha detaylı ve zeki bir arama yapabilmek oldukça önemli bir konudur.

Benzer belge aranması problemi, içerisinde iki önemli unsuru ihtiva eder; bunlar metinsel verinin sınıflandırılması veya kümelenmesi ile benzerlik ölçümü kavramlarıdır.

Genellikle, önceki araştırmaların konusu daha iyi bir sınıflandırma yöntemi geliştirmek olmuştur. Birçoğu makine öğrenme yöntemlerine dayanan mevcut bu sınıflandırma yöntemleri içinde aşağıdakiler sayılabilir:

- Karar Ağaçları (Apte ve ark., 1998),
- Bayesian (Sahami ve ark., 1998; Denoyer ve Gallinari, 2004),
- Yapay Sinir Ağları (Ruiz ve Srinivasan, 2002),
- *k*-En Yakın Komşu (Weng ve Lin, 2003; Masand ve ark. 1992),
- Destek Vektör Makinesi (DVM) (Joachims, 1997; Dumais ve ark., 1998; Mine ve ark. 2002),
- Kendini Düzenleyen Harita (KDH) (Klose ve ark., 2000; Gunther 2001; Yanga ve Leeb, 2004),
- Bulanık Mantık (Widyantoro ve Yen, 2000; Miyamoto, 2001).

Öte yandan verimli benzerlik ölçümlerinin ortaya konması da benzer belge aramanın bir diğer önemli çalışma sahasını oluşturmaktadır (Zhang ve Rasmussen 2001; Egghe ve Michel, 2002).

1.2. Belge Benzerliğinde Karşılaşılan Problemler

Benzer belge aranmasındaki en önemli iki yaklaşım, ileride detaylarıyla da bahsedileceği gibi, anahtar kelimeler yardımıyla indeksler oluşturulması ve metinin tamamı üzerinden bir benzerlik ölçümü yapılması yöntemleridir.

Anahtar kelime kullanılması yaklaşımında sistemde daha az kelime bulunacağından daha hızlı bir arama yapılabilecektir. Ancak sistemin performansını doğrudan ve önemli bir ölçüde etkileyecek bu anahtar kelimelerin seçimi ise çok büyük bir problem olarak ortaya çıkmaktadır.

Diğer yaklaşım ise metin içindeki tüm kelimelerin benzerlik ölçümüne dâhil edilmesidir. Bu yaklaşımın önemli bir avantajı, aranılan metin ile aday metinler arasında yapılan karşılaştırmada, tüm metinin kullanılıyor olmasıdır. Bu yaklaşım önceki yaklaşımdaki gibi alana özel kelimelerin çıkarılıp indeks oluşturulmasına oranla daha kapsamlı ve daha alandan bağımsız bir yaklaşımdır.

Bu yaklaşımda karşımıza çıkan en önemli problem, karşılaştırma işlemi için metinlerin içerdiği tüm kelimeler kullanıldığından dolayı ortaya çıkan işlem fazlalığıdır. Çünkü çok fazla sayıda belge içeren bir koleksiyonun içerdiği kelime sayısı da çok fazla sayıda olacaktır. Tüm bu kelimelerin terimlere dönüştürülüp terim-belge ilişkisinin kurulması, bu verilerin hem saklanması hem de işlenme süreci ile ilgili oldukça önemli bir sorun olacaktır.

1.3. Çalışmanın Amacı ve Önemi

Bu çalışmadaki temel amaç, büyük miktardaki metinsel belge için benzer belgelerin, bulanık kümeleme tekniği kullanarak metin madenciliği yolu ile aranmasıdır.

Belge benzerliği konusunda, belge koleksiyonu üzerinde terim-belge ve belge-kategori arasında bağlantı kurularak çözüm aranabilir. Belgeler, sayıları on binleri aşan terimler cinsinden değil, sadece onlarca olan kategoriler cinsinden ifade edileceklerdir. Böyle bir nitelik azaltma işlemi sayesinde veri boyutunun büyüklüğünden kaynaklanan problemin aşılması sağlanmaktadır. Bu ise çalışmanın bir diğer amacı olan arama esnasındaki işlem yoğunluğunun azaltılması konusunun çözümünü sağlamaktadır.

Ayrıca benzer belgeler aranırken birden fazla kategoriye ait olma faktörü ile arama veriminin artırılmasına çalışılmıştır. Bu sayede birden fazla kategoriye sahip belgeler arasındaki önceden ortaya çıkmayabilecek benzerlikler bulunabilecektir. Önceden göz ardı edilen belgeler yeni yaklaşımla tespit edilebilecektir.

Bu çalışmanın bir diğer özelliği ise önerilen arama yaklaşımı içerisinde bulanık mantığın kullanılmış olmasıdır. Bir belgenin aynı anda birden fazla kategoriye aitliği gibi problemler, bulanık bir yaklaşımın kullanılmasına sebep olmuştur. Mevcut bulanık yöntemler incelenmiş, arama işleminin verimliliğini artırmak için bu yöntemler geliştirilmeye çalışılmıştır.

Çalışmanın önemi kısaca şu şekilde özetlenebilir: Benzer belgelerin aranmasının metinsel veriden bilgi çıkarımında önemli bir rol oynaması, ayrıca belgelerin yönetimi için de temel bir işlem olarak göze çarpması bizi bu konuda yeni ve verimli yöntemler geliştirmeye yönlendirmektedir. Bu alanda kullanılacak etkili bir yöntemle, meydana gelen bilgi patlamasına karşı çözüm geliştirilebilir ve çeşitli alanlarda faydalanılabilir. Örnek olarak, daha etkin bir biçimde, ilgilenilen Web sayfalarının bulunması (Pazzani ve ark. 1996), elektronik posta mesajlarının filtrelenmesi (Sahami ve ark. 1998), Internet haberlerinin filtrelenmesi (Lang 1995)

sayılabilir. Yine bu kapsamda, metin halinde özet bilgisi olan kitap veya makaleler üzerinde daha etkin aramalar yapılabileceği gibi, bir müşteri servisi için verimli bir otomatik soru cevaplandırma mekanizması da oluşturulabilir. Yukarıda bahsedilen konular düşünüldüğünde, etkili bir benzer belge arama tekniği büyük önem kazanmaktadır.

Bu alanda önceki çalışmalarda ise daha çok kümeleme ve sınıflandırma metotları üzerine çalışılmıştır. Bu konuda, bulanık benzerlik yardımıyla metin sınıflandırma (Widyantoro ve Yen 2000), bulanık çoklu küme ve bulanık kümeleme (Miyamoto 2001, Mizutani ve Miyamoto 2003), bulanık ilişkisel kümeleme (Krishnapuram ve ark. 2001) gibi çalışmalar mevcuttur. Belge benzerliği ile ilgili olarak ise çoklu kavram ve kavramın dokuman içindeki dağılımı ile ilgili çalışmalar (Weng ve Lin 2003), belge benzerliğinde verimli bir kümeleme yöntemi araştırılması konulu çalışmalar (Sitarama ve ark. 2004) mevcuttur. Bir belgenin birden çok kategoriye ait olması durumunun da göz önünde bulundurulduğu aramalar üzerinde ise şimdiye kadar çalışılmamıştır.

1.4. Literatür Araştırması

Metin madenciliği, bilgi çıkarımı ve benzer belge arama ile ilgili bugüne kadar yapılmış pek çok çalışma mevcuttur. Aşağıda bu çalışmaların bazılarına kısaca yer verilmiştir.

Veri madenciliğine paralel olarak gelişen metin madenciliği oldukça önemli bir kullanım alanı bulmaktadır. Bu alanlar ekonomiden sağlığa kadar geniş bir yelpazeyi kapsamaktadır. Ancak daha çok Web teknolojileri ile ilgili çalışmalar göze çarpmaktadır.

Mine ve ark. (2002) çalışmalarında uluslararası konferansların konuları hakkında ilişkiler bulan bir metin madenciliği sistemi önermişlerdir. DVM yöntemi

kullanılarak konferans konuları çıkartılmış ve kümeleme yapılmıştır. Sistem için grafik kullanıcı arayüzü de bulunmaktadır.

Yanga ve Leeb (2004) çalışmalarında Web sayfalarının organizasyonu ve hiyerarşik olarak düzenlenmesini amaçlamışlardır. Önerdikleri yöntem KDH (Self-Organizing Map – SOM) metoduna dayalıdır ve işlem sürecinde insan müdahalesine ihtiyaç duymamayı amaçlamıştır.

Weng ve Liu (2004) araştırmalarında gelen elektronik postaları cevaplandırmakla görevli müşteri servisi personelinin yükünü hafifletmek için kalıp öneri elektronik postalarına metin sınıflandırması uygulamışlardır. Bu sisteme çoklu kavram yöntemini ve kavramlar arası ilişkileri bütünleştirmeye çalışmışlardır. Ek olarak birden fazla soru içerebilen elektronik postalar için de uygun cevap kalıpları oluşturmaya çalışmışlardır. Kümeleme işlemi için düşük-ağırlık belge kümeleme yöntemi kullanılmıştır.

Amasyalı ve Yıldırım (2004) ise çalışmalarında gazetelerin web sayfalarındaki Türkçe haber metinlerini otomatik olarak sınıflandırmaya çalışmışlardır. Metinsel verilerdeki boyut azaltma işlemi için Bilgi Kazanım (Information Gain - IG) Ölçümleri ve Temel Bileşenler Analizi (Principal Components Analysis - PCA) kullanmışlardır.

Porrata ve ark. (2007) haber metinlerinin konularının tespiti için bir sistem önermişlerdir. Tespit edilecek bu konular hiyerarşik bir yapıdadır. Çalışmaları yeni bir hiyerarşik kümeleme yöntemi de içermektedir. Ayrıca Testor teorisine dayalı yeni bir özet oluşturma yöntemi de önermişlerdir.

Meziane ve Rezgui (2004) çalışmalarında belge organizasyonu, saklanması ve bilgi çıkarımı için içerik benzerliklerine dayalı bir yöntem önermişlerdir. Bu yöntem bilgi çıkarımı belge indekslenmesi, terimlerin çıkarımı ve indekslenmesine dayanmaktadır.

Bao ve ark. (2003) çalışmalarında metinsel veriden özellik çıkarımı için yeni bir yöntem olan semantik sıralı modeli önermişlerdir. Bu model kelime mesafesine, kelime yoğunluğuna ve semantik sıra kavramlarına dayalı bir modeldir. Bu modeli metin madenciliğinde sıkça kullanılan vektör uzayı modeli ve göreli frekans modeli ile karşılaştırmışlardır.

Jing ve ark. (2002) çalışmalarında bir özellik seçme metodu önermişlerdir. Kategori sonuçlarından faydalanarak metotlarının kesinliğini hesaplamışlardır. Performansı artıran bu yeni TFIDF tabanlı özellik seçme yaklaşımlarının analizini yapmışlardır. Veri ön işlemesinin önemi üzerinde durmuşlardır.

Gurusamy ve ark. (2002) çalışmalarında yorum isteği belge serileri (RFC) için metinden bilgi keşfi önermişlerdir. Bu belgelerdeki yapısal ve yapısal olmayan verilerden metin madenciliği için bir sistem sunmuşlardır. Bilgi çıkarımıyla belgelerin kümelenmesi sağlanmış ve bu sayede arama uzaylarının azaltılmasına çalışılmıştır.

Gunther (2001) çalışmasında Bayes sınıflandırma teknikleri, etkileşimli (interaktif) ilişki arama ve hiyerarşiye dayalı KDH birlikte kullanmıştır. Başlangıç denetlemesi ve gereksiz kullanıcı etkileşimini azaltmak amacıyla ilişkisel kural teorisi ile KDH birlikte önerilmiştir. Metin belgelerden elde edilen kök kelime kümeleri, ilişkisel kural çıkarılmasında kullanılmıştır.

Bhuyan ve ark. (1991) çalışmalarında kullanıcı eksenli kümelemeye dayalı bir bilgi çıkarım sistemi önerilmişlerdir. Kümeleme yapılırken belgeler arasındaki benzerlik için kullanıcıların algısı dikkate alınmıştır. Ayrıca kümelemenin verimliliğini artırmak için bir en uygun şekle sokma fonksiyonu geliştirilmiştir.

Sağlık ve biyoloji alanında da metin madenciliği uygulamaları mevcuttur. Bunlara örnek olarak ise şu çalışmalar verilebilir.

Feldman ve ark. (2003) ise çalışmalarında yine biyomedikal literatürüne metin madenciliği uygulanması üzerinde yoğunlaşmışlardır. Biyolojik karmaşıklığın anlaşılabilmesi için genler, proteinler, ilaçlar ve hastalıklar arasında ilişkiler bulmaya çalışmışlardır.

Abulaish ve De (2007) çalışmalarında ontoloji tabanlı bir biyolojik bilgi çıkarımı ve soru cevaplandırma sistemi oluşturmuşlardır. Bunu ise bulanık mantık ve doğal dil işleme yöntemlerini kullanarak konuyla ilgili metinsel koleksiyondan metin madenciliği teknikleriyle elde etmişlerdir.

Metin sınıflandırması da metin madenciliğinde önemli bir araştırma sahası olarak göze çarpmaktadır. Aşağıda örnek olarak verilen çalışmalar metin sınıflandırması üzerine odaklanılan çalışmalardır.

Bayer ve ark. (1998) çalışmalarında alandan ve dilden bağımsız bir metin sınıflandırması üzerinde durmuşlardır. Metni istatistiksel olarak incelemişlerdir ve nitelik azaltma için lineer dönüşüm (linear transformation) kullanmışlardır.

Zhang ve Oles (2001) çalışmalarında metinlerin kategorize edilmesi problemine istatistiksel ve matematiksel açıdan yaklaşmışlardır. Doğrusal sınıflandırma yöntemlerini metin sınıflandırılması üzerinde odaklaşarak karşılaştırmışlardır. Bu yöntemleri istatistiksel ve matematiksel olarak incelemişlerdir.

Dhillon ve ark. (2001) çalışmalarında büyük miktardaki belgenin makul bir sürede kümelenmesi üzerinde durmuşlardır. Bunun için verimli bir hafıza yönetimi ve çoklu yollu bir ön işleme planı önermişlerdir. Ayrıca veri kümesindeki boşluk problemini çözmek için hızlı bir küresel *k*-ortalamalar algoritması önermişlerdir. Özet metinleri üzerinde yapılan deneysel sonuçlar verimli olmuştur. Ayrıca belge sayısının artışıyla işlem süresinin doğrusal olarak arttığı gösterilmiştir.

Ruiz ve Srinivasan (2002) çalışmalarında metin sınıflandırması için düz yapay sinir ağı ile hiyerarşiye dayanı yapay sinir ağı modelini karşılaştırmışlar ve hiyerarşiye dayalı olan yapının performansı artırdığını göstermişlerdir.

Kou ve Gardarin (2002) çalışmalarında belge sınıflandırması için terim-kategori ve kategori-belge özelliklerini incelemişlerdir. Sınıflandırmada *k* yakın komşu (*k*NN) yöntemini kullanmışlardır. Sınıflandırmanın kalitesini artırmak için terim ilişkisi faktörünü eklemişlerdir. Terim ilişkisi hesaplamasında ε benzerlik modeli önerilmiştir. Deneysel sonuçlar ε benzerlik modelinin performansı artırdığını göstermiştir.

Hotho ve ark. (2003) çalışmalarında metin kümelemesi için arka plandaki bilginin kullanılmasını önermişlerdir ve biçimsel (formal) kavram analizi uygulamışlardır. Bölümlemesel (partitional) kümeleme yöntemiyle problem

boyutunu azaltmaya çalışmış ve sonuçların anlaşılmasını kolaylaştırmayı hedeflemişlerdir.

Denoyer ve Gallinari (2004) çalışmalarında yapılı belgelrin sınıflandırması üzerinde durmuşlardır. Bayesian ağlarına dayalı olarak içerik ve yapının her ikisinide kapsayan bir model geliştirmişlerdir.

Li ve ark. (2006) çalışmalarında kaba küme ve durum tabanlı (rough set-based case-based) bir metin sınıflandırma yöntemi önermişlerdir. Terim azalma için kaba küme yöntemi, belge azaltma için ise durum tabanlı bir yöntem kullanmışlardır.

Metin madenciliğinin bir diğer uygulama alanı ise doğal dil işleme ile ilgili çalışmalardır. Bu çalışmalara örnek olarak aşağıdakiler verilebilir.

Cooper ve ark. (2002) çalışmalarında bilgi çıkarımı için hızlı bir belge benzerliği belirleme yöntemi önermişlerdir. Hızlı bir ifade (phrase) tanıyıcı sistem kullanarak her bir belgedeki en önemli terimlerin listesini belirlemişlerdir. Benzerlik şartı iki belgenin belirli bir eşik değerini aşacak oranda aynı kelimelere sahip olmalarıdır.

Perin ve Petry (2003) çalışmalarında tam metinden bilgi çıkarımı probleminde sözlüksel içerikler arasında olan ilişkilerin rolünü incelemişlerdir. Bilgi çıkarımı için, metnin yapısı hakkında bilgi veren metnin içindeki terimlerin göreli mesafelerini kullanmışlardır. Metnin içindeki terimleri, ilgili içeriksel üniteler halinde incelemişlerdir. Deneysel çalışmalarını psikiyatrik raporlar üzerinde yapmışlarıdır.

Ko ve ark. (2004) metin özetlenmesi yöntemlerini kullanarak cümleler için bir önem ölçümü önermişlerdir. Belgeleri cümleler ve önem değerlerinin bir vektörü olarak gösterişlerdir. Sınıflandırıcı olarak ise naive Bayes, Rocchio, *k*NN ve DVM kullanmışlardır.

Amasyalı ve Diri (2005) çalışmalarında Türkçe için doğal dille çalışan bir soru cevaplama sistemi gerçekleştirmişlerdir. Sistem öncelikle kullanıcısının doğal dille sorduğu sorusunu arama motoru sorgusuna çevirmekte ve arama motorunun sonuç sayfasından ya da bağlantılarındaki sayfalardan olası cevap cümlelerini

seçmektedir. Olası cevap cümlelerini çeşitli kıstaslara göre puanlandırıp en yüksek puanı alan ilk beş cümle kullanıcıya iletilmektedir.

Son yıllarda yaygınlaşan bulanık mantık kullanımı, metin madenciliği çalışmalarında da çokça göze çarpmaktadır. Aşağıda bununla ilgili bazı çalışmalara kısaca yer verilmiştir.

Delgado (1995) çalışmasında bulanık kümeleme işleminin geçerliliği üzerinde durmuştur. Bulanık kümelemeden önce uygun bir başlangıç yapısı seçmek için hiyerarşi ye dayalı bir kümeleme analizi yapmaktadır. Bu sayede benzer sonuçlara daha az yineleme ile varılmaktadır.

Widyantoro ve Yen (2000) çalışmalarında metin sınıflandırma problemi için bir bulanık benzerlik yaklaşımı önermişlerdir. Test aşamasında birden fazla bulanık birleşim ve kesişim operatörü denemişler ve bazı özel hallerde bulanık benzerlik yaklaşımının çok iyi sonuçlar verdiğini göstermişlerdir.

Miyamoto (2001) çalışmasında bulanık çoklu kümelere dayalı bir bulanık kümeleme metodu geliştirmiştir. Terim-belge matrisinin kayıtları çoklu küme değerleri olarak varsayılmıştır. Bulanık çoklu kümeye dayalı iki adet başkalık ölçütü önerilmiştir. Bulanık *c*-ortalamalar metodundaki küme merkezlerinin odaklanmasında bu iki ölçüt kullanılmıştır.

Mizutani ve Miyamoto (2003) çalışmalarında bulanık çoklu küme modeline dayalı bir bulanık kümeleme yöntemi önermişlerdir. Terim belge matrisinin kayıtlarını çoklu küme olarak almışlar ve bulanık çoklu küme üzerinden bir başkalık ölçütü önermişlerdir. Bu ölçütü kullanmak suretiyle bir bulanık *c*-ortalamalar yöntemi geliştirmişlerdir. Bu yöntemi DVMnde yüksek boyutlu bir uzayda doğrusal olmayan dönüşüm (transformation) üzerinde çalıştırmışlardır. Gerçek belge üzerine de bir örnek yer almaktadır.

Subasic (2001) çalışmasında metin belgelerin etki içeriğini analiz etmek için doğal dil işleme ve bulanık mantık tekniklerini birleştirmiştir. Deneysel çalışmalar soncunda etki kümesi ve etki içeriğinin insan yargıları arasında iyi bir uygunluk görülmüştür.

Krishnapuram (2001) çalışmasında ilişkisel verinin bulanık kümelemesi için yeni bir bulanık *c*-mediods algoritması göstermiştir. Amaca ait fonksiyon veri kümesinden *c* temsilci nesnelerinin seçimine dayalı olup buna göre her bir küme içindeki toplam bulanık başkalığı en aza indirmeye çalışmaktadır. Bu algoritmayla ilgili uygulamaları Web madenciliği ile ilgili Web belgeyi kümelemesi ile ilgili yapmıştır.

Qiu (2002) çalışmasında yoğunluk ve mesafeye dayalı yeni bir bulanık kümeleme yöntemi önermiştir. Küme sayısını otomatik kendi belirlemektedir. İris ve diyabet verileri üzerinde yapılan deneysel çalışmalar bu metodun yüksek tanıma oranına sahip olduğunu göstermiştir.

Latiri (2003) çalışmasında metin madenciliğinde, terimler arasında bulanık ilişkisel kural çıkarımı için bulanık Galois bağlantıları kullanmıştır. Terimler arasındaki bu bulanık ilişkisel kuralları kullanarak sorgu genişletilmesi yapılmıştır. Bunun amacı sorgu/belge arasındaki uyumsuzlukları azaltmak suretiyle Bilgi Çıkarımı (IR) verimini artırmaktır.

Benzer belge aranması üzerinde odaklanmış çalışmalara örnek olarak ise aşağıdakiler verilebilir.

Weng ve Lin (2003) çalışmalarında benzer belge aranması için sınıflandırma konusu üzerine odaklaşmanın yanı sıra kavram ve kavramın dağılımı faktörleri üzerinde de durmuşlardır. Benzerlik aranması için çoklu kavram mekanizması önermişler ayrıca benzerlik kalitesini artırmak için kavramın belge içindeki dağılımı faktörünü incelemişlerdir. Kümeleme işlemi için ise *k*NN yöntemini kullanmışlardır. Deneysel sonuçlar, önerdikleri tekniğin geleneksel yaklaşımlardan daha verimli olduğunu göstermiştir.

Atlam ve ark. (2003) çalışmalarında belge benzerliği için bütün metin içeriği yerine alan ilişkisel terimlerini kullanmayı önermişlerdir. Bu terimlere dayalı alan ilişkisel benzerlik ölçümü tanımlamışlardır. Kullanılan bu alanlar hiyerarşiye dayalı bir yapıda bulunmaktadır.

Zhang ve Rasmussen (2001) çalışmalarında mesafe ve açıya dayalı iki farklı benzerlik ölçümünü karşılaştırmışlardır. Bunlara bağlı olarak yeni bir benzerlik ölçümü önermişlerdir.

Önceki çalışmalarda belgelerin birden fazla kategoriye ait olma durumlarının ele alınmadığı görülmektedir.

1.5 Tezin Organizasyonu

Bulanık kümeleme kullanılarak benzer belge aranması üzerine odaklanmış olan bu tez çalışmasının ana hatları aşağıdaki gibidir:

Birinci bölümde, tez çalışmasının konusuna genel bir bakış açısı verilmeye çalışılmıştır. Çalışmanın amacı, çerçevesi ve mevcut çalışmalara göre yerine kısaca değinilmiştir.

İkinci bölümde, metin madenciliği üzerinde durulmuştur. Tezin temelini oluşturan benzer belge aranmasının metin madenciliğindeki yeri vurgulanmaya çalışılmıştır.

Üçüncü bölümde, bulanık mantık kullanılarak oluşturulan bir benzer belge arama yaklaşımı ayrıntılı bir biçimde ortaya konmuştur. Bu yaklaşımın temel bileşenlerine değinilerek daha sonraki bölümlerde kullanılacak olan bir arama sistemi ortaya konmuştur.

Dördüncü bölümde, önerilen arama yaklaşımı için en uygun terim ağırlıklandırma yöntemi araştırılmıştır. Buradan elde edilen sonuç, diğer bölümler için de önem taşıyan bir temel taşı özelliğindedir.

Beşinci bölümde, benzerlik ölçümleri üzerinde durulmuştur. Mevcut benzerlik ölçümleri karşılaştırılarak verinin boyutuna dayalı yeni bir benzerlik ölçümü önerilmiştir. Önerilen benzerlik ölçümünün önceki ölçümlere olan üstünlüklerine değinilmiştir.

Altıncı bölümde, çoklu kategori problemi ele alınmıştır. Önerilen arama yaklaşımı bu amaca yönelik olarak geliştirilmiştir. Bu gelişimi sağlamak üzere, mevcut bulanık sınıflandırma yöntemi probleme adapte edilmiştir. Ayrıca yeni bir kategori tespit yöntemi ortaya konmuştur.

Yedinci bölümde, anahtar kelime tabanlı arama (manüel olarak seçilen anahtar kelimelere göre arama yaklaşımı) ve belgenin içerdiği tüm kelimelerin kullanıldığı arama (tez çalışmasında önerilen arama yaklaşımı) kullanılarak benzer belge aranması yöntemlerinin karşılaştırıldığı bir uygulamaya yer verilmiştir.

Sekizinci bölümde ise tezle ilgili genel sonuç ve önerilere yer verilmiştir.

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

Ek kısmında ise yedinci bölümde bahsedilen benzer belge arama uygulaması ile ilgili örneklere ayrıntılı bir şekilde yer verilmiştir.

2. METİN MADENCİLİĞİ VE BULANIK MANTIK

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

Metinsel verilerin yer aldığı veri tabanları kısaca metin veya belge veri tabanları olarak adlandırılır (Mitra ve Acharya, 2003). Bu veri tabanlarında, kitapların elektronik yayınları, dijital kütüphaneler, elektronik e-postalar, elektronik medya (ortam), teknik veya ticari (mesleki) belgeler, raporlar, araştırma makaleleri, internetteki web sayfaları, html vb. şekilde geniş miktarda kullanılabilir bilgi vardır. Belge koleksiyonu olarak da adlandırılan bunlar gibi metin veri tabanlarından bilgi çıkarımına yardım amacıyla, son zamanlarda veri madenciliği yöntemlerinin özel tipleri geliştirilmiştir. Veri madenciliğinin metin ile uğraşan bu alanı genelde metin madenciliği olarak bilinir. Metin madenciliğinin bir başka tanımı ise, yarı yapılı veya yapısız metinsel verilerden özel veri madenciliği yöntemleri ile yeni bilgi keşfidir.

Halen bilimsel araştırmalar içinde hızlı bir gelişim sürecinde olan metin madenciliği, veri madenciliği yöntemlerine ek olarak birçok çoklu disiplinli bilimsel tekniği de kullanır. Bu teknikler; algı elde etmek, anlamak ve yorumlamak ve metin veritabanlarında her yere dağılmış kullanılabilir metinsel verilerin büyük miktarından otomatik olarak bilgi çıkarmak için kullanılır. Metin madenciliği yöntemlerinin işlevselliği esasen metin analizi tekniklerinin sonuçlarına dayandırılmıştır.

Metin madenciliğini son zamanlarda etkileyen diğer alanlardan bazıları; string (dizgi) eşleme, metin arama, yapay zekâ, makine öğrenmesi, bilgi çıkarımı, doğal dil işleme, istatistik, bilgi teorisi, esnek hesaplamadır.

2.1. Metin Analizi ve Erişimi

Önceleri metin analizi; doğal dil işleme ve bilgi çıkarımındaki çalışmaların bir alanı olmuştur. Internet arama tekniklerinin büyük çoğunluğu metin tabanlı olduğundan dolayı Internet'in gelişimiyle birlikte metin analizi de önem kazanmıştır.

Genellikle metinsel veriler yarı yapılıdır ve insanlarca okunmak ve yorumlanmak için kolaydır. Metin analizi teknikleri genellikle şu amaçlar için kullanılır:

- bir metinden anlamlı anahtar özniteliklerini çıkarmak
- metinsel belgeleri anlamsal içeriklerine dayanarak sınıflandırmak
- belgeleri indekslemek
- metinsel belgelerin veya büyük koleksiyonunun özetini çıkarmak
- büyük belge koleksiyonunu verimli şekilde düzenlemek
- otomatik arama işleminin etkinliğini geliştirmek
- geniş veri tabanlarındaki eş belgeleri saptamak.

Otomatik metin çıkarım sistemlerinde (Full-Text Retrieval Systems), belgelerin otomatik indekslemesi çoğu kez, belgelerde görülen ortak kelimeler ve cümlelerin istatistiksel analizine dayalı olarak yapılır. Otomatikleşmiş metinsel belgelerin indekslemesi için basit bir metot aşağıdaki adımlarla tanımlanabilir (Mitra ve Acharya, 2003):

- Belge koleksiyonunda, her bir belgedeki eşsiz kelimeleri bulun.
- Belge koleksiyonunda her bir belge için bu eşsiz kelimelerin görülme sıklığını hesaplayın.

- Her bir kelimenin toplam görülme sıklığını, koleksiyondaki tüm belgelerin bir tarafından öbür tarafına hesaplayın.
- Kelimeleri, koleksiyonda görülme sıklıklarına göre artan sırada sıralayın.
- Çok yüksek görülme sıklığına sahip olan kelimeleri, bu sıralanmış listeden kaldırın.
- Düşük görülme sıklığına sahip olan kelimeleri, bu sıralanmış listeden kaldırın.
- Kalan kelimeleri metin koleksiyonu için indeks olarak kullanın.

Otomatik metin erişimi ve sınıflandırma metotları ilk olarak 1960'ların başında görülmüştür ve günümüzde halen önemli araştırma bir konusudur (Vasifov, 2001). Bu metotlar genelde aşağıdaki uygulamalarda kullanılır:

- Otomatik belge indeksleme: Her bir belge kendi içeriğini tanımlayan bir yada daha fazla, anahtar kelime yada anahtar deyime atanmıştır. Bu kelime veya deyimler, denetimli sözlük olarak adlandırılan kelimelerin bir sonlu kümesine ait olan ve çoğu kez hiyerarşik bir eş anlamlılar sözlüğünü içeren kelime gruplarıdır.
- Belge filtreleme: Belgelerin dinamik koleksiyonu için kullanılır. Filtreleme sistemi, alıcıyı ilgilendirmeyen belgeler için teslimatı bloke etmeyi amaçlamaktadır.
- Internet uygulaması: YAHOO, Infoseek gibi ticari hiyerarşik katalogları oluşturulmasında Web sayfaları kategorilerine sınıflandırılmıştır. Web sayfaları kategorilerine göre organize edilirse, bir kullanıcının belgeyi kolaylıkla bulması mümkün olacaktır.

2.2. Metinlerin Matematik Modeli

Metin verisi aslında iki temel birimin bir bileşimi olarak düşünülebilir. Bunlar belge ve terimdir. Genel olarak bir belge, bir metnin yapılı ya da yarı yapılı bir parçasıdır. Örneğin, bu tez bir metin belgesidir ve birtakım bölümler şeklinde yapılandırılmıştır. Her bir bölüm parçalardan, her bir parça birtakım alt parçalardan ve paragraflardan vb. oluşabilir. Benzer şekilde bir elektronik posta, bir belge olarak düşünülebilir çünkü belirli bir biçimde, bir mesaj başlığı, konu ve mesaj içeriğini kapsar. Uygulamada mevcut olan bu şekilde birçok belge vardır. Diğer bazı örnekler; kaynak kodları, Web sayfaları, elektronik çizelgeler, telefon rehberi vb. olabilir. Bir terim; belgede bulunan bir kelime, kelime grubu veya bir cümledir. Terimler, string (dizgi) eşleme algoritmalarından herhangi biri kullanılarak, belgeden seçilebilir.

Terim ve belge için bu tanımı kullanılarak, bir metinsel belge modellenebilir. Bir D belgeler koleksiyonunu ve bir T terimler kümesini sırası ile aşağıdaki gibi düşünelim:

$$D = \{d_1, d_2, d_3, ..., d_N\} \tag{2.1}$$

$$T = \{t_1, t_2, t_3, ..., t_M\} \tag{2.2}$$

Her bir d_i belgesi, M boyutlu R^M uzayında aşağıdaki gibi bir vektör olarak modellenebilir:

$$d_i = (w_{i,1}, w_{i,2}, ..., w_{i,M}) \qquad i = 1...N \tag{2.3}$$

Her bir $w_{i,j}$ girişi, t_j terimiyle d_i belgesinin birliktelik ölçüsünü gösterir. $v_{i,j}$ değeri, d_i belgesini, t_j terimini içermiyorsa sıfırdır, aksi halde sıfıra eşit değildir.

Basit olarak iki değerli gösteriminde (boolean), t_j terimi d_i belgesinde görünürse $w_{i,j} = 1$ olur. Ancak, bu ölçüm metin çıkarımında çok başarılı bulunmamıştır. $w_{i,j}$ ilişkisinin daha yaygın ve pratik bir ölçümü, t_j teriminin d_i belgesinde görülme sayısı olarak basitçe tanımlanan terim sıklığıdır. Bu yaklaşımı kullanarak metin, şekil 2.1'de tarif edildiği gibi, "belge - terim sıklığı matrisi" olarak modellenir.

Bir örnek verecek olursak, Şekil 2.1'de beş belge ve beş terim kümesi için belge - terim sıklığı matrisini göstermede 5×5'lik bir dizi verilmiş olsun. Seçilen terimlerin ise aşağıdaki kelimeler olduğunu varsayalım.

- t_1 = aslan
- t_2 = kuş
- t_3 = çiçek
- t_4 = orman
- t_5 = biyoloji

	t_1	t_2	t_3	t_4	t_5
d_1	10	8	1	0	2
d_2	5	9	4	3	1
d_3	0	15	10	2	3
d_4	22	0	0	6	5
d_5	41	18	5	2	0

Şekil 2.1. Beş belge ve beş terim için belge – terim sıklığı matrisi

İkinci belgeyi (d_2 belgesini) temsil eden vektör, matrisin ikinci satırını oluşturan ve sırasıyla, 'aslan' teriminin beş defa, 'kuş' teriminin dokuz defa, 'çiçek'

teriminin dört defa, ‘orman’ teriminin üç defa, ‘biyoloji’ teriminin bir defa görüldüğü (5, 9, 4, 3, 1) vektörüdür.

2.2.1. Vektör Uzay Modeli

Yukarıda bahsedildiği gibi terimler ve belgeler arasında ilişki kurulabilmesi için bir model oluşturulması gerekmektedir. Bu amaçla tasarlanan genelleştirilmiş model ise Vektör Uzay Model (Vector Space Model) olarak bilinmektedir. Bu model; belgeler, sorgular veya kavramların terimler cinsinden birer vektör şeklindeki gösterimleri olarak özetlenebilir (Raghavan ve Wong, 1999). Bu model ilk olarak bilgi erişimi çalışmaları için ortaya konmuş ve yaygın bir şekilde kullanıla gelmiştir (Salton ve McGill, 1983).

Bilgi erişimi çalışmalarındaki amaçlardan biri, mevcut belgeler ile anahtar kelime şeklindeki sorguları arasında benzerlik kıyaslaması yapabilmektir. Bu amaç için ise artık standartlaşmış olan vektör uzay modeli kullanılmaktadır. Bu modelin oluşturulma işlemi kısaca özetlenecek olursa aşağıdaki adımlardan oluşmaktadır.

- Her bir belgenin bulundurduğu terimler (kelimeler) ve sıklıkları belirlenir.
- Sistemdeki tüm belgelerdeki tüm terimlerden bir sözlük oluşturulur.
- Her bir belge içerdiği terimlere göre bir vektör formuna getirilir.
- Tüm belgelerin vektörleri bir araya getirilerek dizi şeklinde tutulur.

Buna örnek olarak aşağıdaki metinler ve bunların vektör formları verilebilir.

A belgesi:

- “Bir kedi ve bir köpek”

bir	kedi	köpek	ve
2	1	1	1

B belgesi:

- “Bir aslan”

aslan	bir
1	1

Her iki belgede bulunan tüm kelimelerden aşağıdaki sözlük oluşturulur.

- aslan, bir, kedi, köpek, ve

Belgeler bu sözlüğe göre aşağıda görülen vektörlere dönüştürülürler.

A belgesi:

- “Bir kedi ve bir köpek”

aslan	bir	kedi	köpek	ve
0	2	1	1	1

Vektör = (0, 2, 1, 1, 1)

B belgesi:

- “Bir aslan”

aslan	bir	kedi	köpek	ve
1	1	0	0	0

Vektör = (1,1, 0, 0, 0)

Elimizdeki iki belgelik bu koleksiyon yanı sıra karşılaştırmak için Tablo 2.1’deki sorgular olsun.

Tablo 2.1. Örnek sorgular

Sorgu adı	Sorgu metni	Vektör formu
X	köpek	(0, 0, 0, 1, 0)
Y	aslan	(1, 0, 0, 0, 0)
Z	köpek ve aslan	(1, 0, 0, 1, 1)

Artık aynı vektör formunda olan belge ve sorgular arasında benzerlik karşılaştırması yapılabilir. Bu benzerlik ölçümü için çeşitli yöntemler bulunmaktadır. En temel ve metin madenciliğinde geniş kullanım alanı olan kosinüs benzerliğidir. Bu benzerlik ölçümü kullanılarak örnek bir karşılaştırma aşağıdaki gibi olacaktır. Çok boyutlu veri (vektör) için kosinüs benzerliği aradaki açıyı temel almaktadır. Her iki vektörün karşılıklı boyutlarındaki değerleri çarpılarak toplanır. Bunlar her bir vektörün boylarının çarpımına bölünür. Bu hesaplama aşağıda görüldüğü gibi A belgesi ve X sorgusu üzerinde örnek olarak yapılmıştır.

A belgesi = (0, 2, 1, 1, 1)

X sorgusu = (0, 0, 0, 1, 0)

$$Benzerlik = \frac{0.0 + 2.0 + 1.0 + 1.1 + 1.0}{\sqrt{0^2 + 2^2 + 1^2 + 1^2 + 1^2}\sqrt{0^2 + 0^2 + 0^2 + 1^2 + 0^2}}$$

$$= \frac{1}{\sqrt{7}\sqrt{1}} = \frac{1}{2.646} = 0.378$$

B belgesi ve X sorgusu arasında hesaplanacak olursa bu benzerlik değerinin 0 olduğu görülecektir.

Bir belge, yukarıda açıklandığı gibi belge – terim sıklığı matrisi gösterimi veya vektör uzay modeli kullanılarak modellendiğinde, metin içindeki kelimelerin birbirlerine bağlı sıralaması kaybolur. Bu sebeple, bir metindeki bir cümle yapısı için gramer gibi metin oluşumunun sözdizimsel bilgisi yok olacaktır. Buna rağmen, sorgu

işleme (query processing), belgelerin karşılaştırılması, belge analizi vb. metin veya belge çıkarımı uygulamalarında, vektör uzay modeli oldukça verimli sonuçlar vermektedir (Mitra ve Acharya, 2003).

2.3. Anahtar Kelime Tabanlı Arama

Metin koleksiyonlarında arama işlemi, geleneksel ilişkisel veri tabanı yönetim sistemlerinde uygulanan arama tekniklerinden farklıdır. Metin koleksiyonları madenciliğinin temel bir yolu, anahtar kelime tabanlı arama yöntemini uygulamaktır. Bu basit yaklaşımda, belgeler metin verisinin imzası olarak düşünülebilecek bir anahtar kelimeler kümesi ile birlikte, stringler olarak kabul edilir ve buna göre indekslenir. Bir anahtar kelime bir metin dosyasının içinde kesin uyum veya yaklaşık uyumu gerektiren string (dizgi) eşleme tekniklerini kullanarak aranabilir. Metin içinde bulunan string (dizgi) - eşlenmiş anahtar kelimeler veya örnekler daha sonra belgeleri indekslemek için kullanılır. Belge, anahtar kelimelerce tanımlandıktan sonra, geleneksel veri madenciliği teknikleri (sınıflandırma, kümeleme, kural çıkarımı vb.), metin veri tabanlarındaki belgelerin koleksiyon karakteristiklerine bağlı olarak, önemli bir ölçüde başarıyla uygulanabilir.

Anahtar kelimelerin anlamsal değerini dikkate almayan böyle basit bir yaklaşımda iki esas problem vardır. Dikkate değer bu iki problem, uzun süre doğal dil işleme alanın problemleri olarak ele alınmış olan eşanlamlılık (synonymy) ve çok anlamlılık (polysemy)tır. Kullanıcı tarafından sağlanan bir anahtar kelime, bu kelime ile ilgili belge çok fazla iken, belgede hiçbir şekilde görülmeyebilir çünkü doğal bir dilde aynı şey çoğu kez başka yollarda tanımlanabilir. Örneğin, belge tam olarak ‘kadın’ kelimesinin herhangi bir örneğini içermiyor fakat sıkça ‘hanım’ kelimesini içeriyorken anahtar kelime ‘kadın’ olabilir. Bu bir eşanlamlılık (synonymy) problemi olarak bilinir. Bu problem, belgeyi filtreleyerek sınırlanabilir; öyle ki, benzer anlamlı kelimeler kuralsal seçilen bir sembolik kelime (token word) ile değiştirilir. İngilizce

için örnek verecek olursak, ‘automobile’, ‘vehicle’ ve ‘vehicular’ kelimeleri basit olarak ‘car’ kelimesi ile değiştirilebilir. Benzer şekilde, ‘is’, ‘are’, ‘am’, ‘were’, ‘was’, ‘been’, ‘being’ kelimeleri bir belgede görüldüğünde, ‘be’ kelimesi ile değiştirilebilir.

Aynı kelimenin faklı içeriklerde farklı anlamlara sahip olması da mümkündür. Örneğin ‘mining’ kelimesi ‘data mining’ bağlamında, ‘coal mining’ ile karşılaştırıldığında farlı anlamlara sahiptir. Buna çok anlamlılık (polysemy) problemi denir. Bu nedenle bu problemleri çözmek için diğer yapay zekâ alanları ile birleşen doğal dil işlemenin başarısı uzun dönemde, büyük etkiye sahip olacaktır.

2.4. Benzerlik Tabanlı Arama

Benzerlik tabanlı aramada belirli anahtar kelimeler seçilmeden belgenin içerdiği tüm kelimeler kullanılarak model oluşturulur. Bu model yardımıyla, iki belgenin benzerliğini bulmak için bir mesafe ölçüsü uygulanır. En basit yaklaşım ise, iki belgeyle ilgili olan iki vektör arasındaki Öklid uzaklığını bulmaktır. Benzerlik tabanlı yaklaşım ilk olarak bilgi erişiminde sorgularla belgelerin eşleştirilmesinde kullanılmıştır. Bir başka deyişle sorgulara benzeyen belgelerin bulunması amaçlanmıştır. Örneğin D belge koleksiyonu ve T terim kümesi sırasıyla Formül 2.1 ve 2.2’deki gibi olsun;

Bu belge koleksiyonundaki belgelerden bir d_q sorgu belgesine benzerlerinin bulunması istenirse, ilk önce terim kümesinin tüm terimler için d_q sorgu belgesinin sıklık vektörü oluşturulur.

$$d_q = (w_{q,1}, w_{q,2}, w_{q,3}, ..., w_{q,M}) \qquad (2.4)$$

D veri kümesindeki d_i sorgu belgesi ve d_q sorgu belgesi arasındaki Öklid mesafesi:

$$Öklid(d_q, d_i) = \sqrt{\left[\sum_{j=1}^{M} (w_{q,j} - w_{i,j})^2\right]} \qquad (2.5)$$

Burada *m* vektörün boyutudur.

İki belge arasındaki benzerliği bulmak için başka mesafe veya benzerlik ölçümleri de uygulanabilir. Örneğin iki belgenin karşılaştırılmasında, iki vektörün kosinüs ölçümleri oldukça etkili olmaktadır (Mitra ve Acharya, 2003). d_i ve d_j vektörlerinin kosinüsü aşağıdaki gibi hesaplanabilir:

$$\cos(d_i, d_j) = \frac{\sum_{k=1}^{M} \left[w_{i,k} * w_{j,k}\right]}{\sqrt{\left[\sum_{k=1}^{M} w_{i,k}^2 \sum_{k=1}^{M} w_{j,k}^2\right]}} \qquad (2.6)$$

Burada *m* vektörün boyutudur.

Yukarıda açıklandığı gibi uzaklık ölçülerinin sayısal değerlerini kullanılarak bir belge koleksiyonunda belgeler arasındaki benzerlikler bulunabilir. Bu benzerlikler ve diğer metin madenciliği tekniklerinin uygulanması yoluyla belgeler üzerinde benzerlik tabanlı indeksler geliştirilebilir.

Sorgulamalar yani sorgunun kendisini, birtakım terimlerle oluşturulmuş bir belge olarak farz ederek aynı terim tabanlı gösterimle ifade edilebilir. Sonuç olarak, yukarıdaki prensipleri, bir belge sorgu eşleme içinde kullanabiliriz. Sorgu, kendi içinde görülen terimlere uyan ağırlıkların bir vektörü olarak ifade edilir ve sorguda mevcut olmayan bu terimler için ağırlıklar tamamıyla sıfır olur. En basit şeklinde, vektör sorguda mevcut olan terimler için bir, diğerleri için sıfır ağırlığını içerebilir. Belge koleksiyonundaki belgelere uyan vektörlerden, daha sonra vektörün uzaklığı ölçülür.

Oldukça verimli ve iyi benzerlik ölçülerine rağmen bahsedilen yaklaşımın sayısal gereksinimleri çok yüksektir. Kullanılabilir metin belge koleksiyonlarının çoğunda, terim kümesindeki terimlerin sayısı 50.000'den fazla olabilir ve belge koleksiyonundaki belgelerin sayısı da çok fazla olabilir. Bu durumda vektör uzay modeli matrisinin boyutu çok yüksek ve işlenmesi zor olacaktır. Bu yüksek boyutluluğun yanı sıra matrisin çok seyrek olması da önemli bir özelliğidir. Ayrıca bu durum ise belgede terimlerin tanımlanmasını zorlaştırır.

2.5. Bulanık Mantık

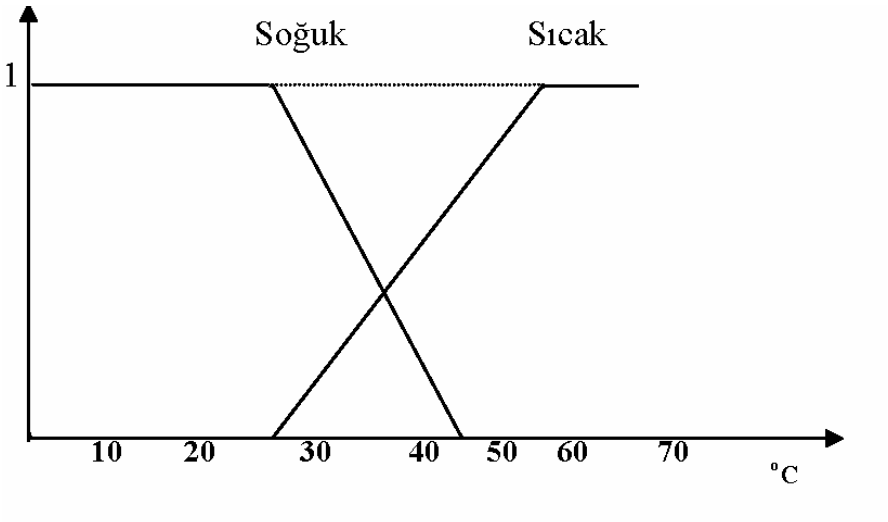
1965 yılında Lotfi A. Zadeh yayınlanan bir makalesinde doğru ve yanlış arasındaki sonsuz farklı değerleri [0,1] aralığındaki sayılarla ifade etmiş, ilk kez bu makalesinde sembolik ifadelerin makinelere aktarılmasının matematiksel bir temele dayandığından bahsetmiştir. Zadeh bu çalışmasında insan düşüncesinin büyük çoğunluğunun bulanık olduğunu, kesin olmadığını belirtmiştir. Bu yüzden 0 ve 1 ile temsil edilen klasik mantık bu düşünce işlemini yeterli bir şekilde ifade edememektedir. Zadeh daha sonra [0,1] aralığındaki sayılarla ifade ettiği teorisini "Bulanık Mantık" adlı çalışmasında tanımlamıştır. Zadeh'in çalışmalarının ardından uygulamalar artarak devam etmiş; gelişmiş bilgisayarların kullanımı ile uygulama alanları genişlemiştir.

Bulanık mantık, bir şey hakkında yargı ortaya atarken, aynı anda, bu yargıyı oluştururken dayandığı matematiksel sınıflandırmaların ne kadar içinde, ne kadar dışında olduğundan bahseder. Verinin ne kadar o yargı kümesine ait, ne kadar ait olmadığı bilgisine dayanarak o veriye yeni bir tanım getirir.

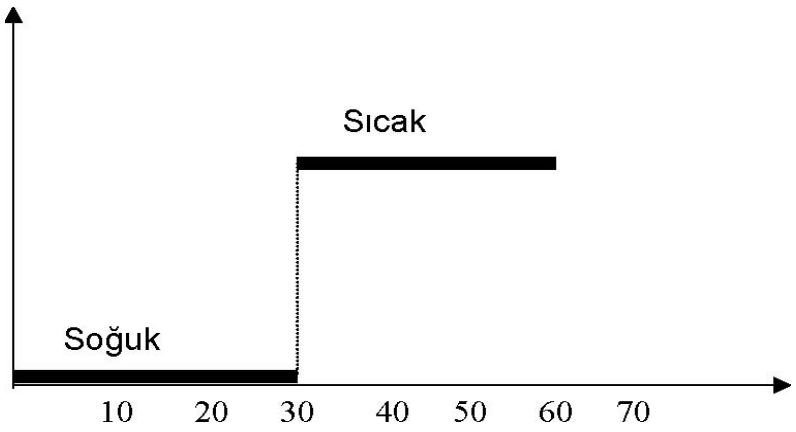
Bulanık mantık, klasik mantığın aksine iki seviyeli değil, çok seviyeli işlemleri kullanmaktadır. Bulanık mantık yaklaşımı, makinelere insanların özel verilerini işleyebilme ve onların deneyimlerinden ve önsezilerinden yararlanarak çalışabilme yeteneği verir. Klasik matematiksel yöntemlerle karmaşık sistemleri

modellemek ve kontrol etmek zordur, çünkü veriler tam olmalıdır. Bulanık mantık kişiyi bu zorunluluktan kurtarır ve daha niteliksel bir tanımlama olanağı sağlar. Bir kişi için 37,5 yaşında demektense sadece orta yaşlı demek birçok uygulama için yeterli bir veridir. Böylece azımsanamayacak ölçüde bir bilgi indirgenmesi söz konusu olacak ve matematiksel bir tanımlama yerine daha kolay anlaşılabilen niteliksel bir tanımlama yapılabilecektir.

Modern teknolojinin kullandığı kodlama biçimi olan 0 ve 1 mantığına karşın bulanık mantık, 0 ile 1 arasındaki değerlerin varlığından bahseder. Klasik mantık 30 $C^{0,}$ yi "sıcak" kümesinin sınırı olarak kabul ediyorsa, 29,9 C^0 'yi sıcak olarak kabul etme hakkını kaybeder. Aradaki bu küçük fiziksel fark, klasik mantık için hayati anlam ifade etmektedir. Çünkü bu değerin üyelik kümesi değişmiştir artık. 30 C^0, 29,9 C^0 olmakla, "sıcak" olmayı reddetmiş ve bunun sonunda "sıcak" olma kümesinden dışlanmıştır. Fiziksel dünyada 30 C^0yi sıcak kabul edilirken, 29,9 C^0 'nin sıcak olmadığı iddia edilmez. Oysaki bulanık mantık bu tür keskin sınırları kaldırarak, 29,9 C^0 'yi neredeyse (1'e yakın bir değerle) sıcak olarak kabul eder. Klasik mantık için "soğuk" ya da "sıcak" olma vardır. Oysaki Bulanık Mantık "soğuk-sıcak" gibi kavramların yanında, "az soğuk", "çok sıcak", biraz sıcak" gibi söylemleri de kabullenir ve bunları matematiksel olarak tanımlamaya çalışır (Sıramkaya 2006).



Şekil 2.2. Bulanık Mantık modeli



Şekil 2.3. Klasik Mantık Modeli

Dilsel değişken "sıcak" veya "soğuk" gibi kelimeler ve ifadelerle tanımlanabilen değişkenlerdir. Bir dilsel değişkenin değerleri bulanık kümeleri ile ifade edilir. Örneğin oda sıcaklığı dilsel değişken için "sıcak" ve "soğuk" ifadelerini alabilir. Bu iki ifadenin her biri ayrı bulanık kümeler ile modellenir. Bununla ilgili bir örnek Şekil 2.2 ve Şekil 2.3'te görülmektedir.

Bulanık mantık ve bulanık yaklaşımı daha iyi açıklamak için ilerleyen kısımda kesin ve bulanık yaklaşımla ele alınan iki kümeleme yaklaşımına yer verilmiştir.

2.5.1 K-ortalamalar

K-ortalamalar (MacQueen, 1967) kümeleme problemi çözümü için ortaya atılan en basit danışmansız eğitim algoritmalarından birisidir. Yöntem veriyi önceden belirlenmiş bir sayıda (*k* adet) kümeye sınıflandırmak için basit ve kolay bir yol takip eder. Ana fikir, her biri bir kümeyi ifade eden *k* tane merkez belirlemekten ibarettir. Farklı yerleşimler farklı sonuçlar ortaya çıkaracağından bu merkezler akılcı bir yolla yerleştirilmeleri gerekmektedir. Bu yüzden en iyi çözüm ise bu merkezlerin mümkün

olduğunca birbirlerinden uzağa yerleştirilmesidir. Bir sonraki adım ise her veri noktasının en yakın olduğu merkezle ilişkilendirilmesi olacaktır. İlişkilendirilmemiş veri noktası kalmadığında ilk adım tamamlanmış ve bir gruplama yapılmıştır. Bu noktada ise bir önceki adımda yapılan kümelemeye bağlı olarak küme merkezlerinin yeniden belirlenmesi gerekmektedir. Böylece *k* tane yeni küme merkezi oluşacaktır. Tekrar her bir veri noktası bu yeni merkezlere göre en yakın olan ile ilişkilendirilecektir. Böylece bir döngü oluştuğu görülebilir. Bu döngü esnasında *k* tane merkez adım adım yer değiştirecektir. Bu yer değiştirme durduğu adımda ise döngü sonlandırır.

Sonuç olarak bu algoritma bir karesel hata fonksiyonu şeklindeki amaç fonksiyonunu (*objective function*) minimumlaştırır. Bu amaç fonksiyonu aşağıda görülmektedir.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \qquad (2.7)$$

Burada $\left\| x_i^{(j)} - c_j \right\|^2$ ifadesi $x_i^{(j)}$ veri noktası ile c_j küme merkezi arasındaki seçilmiş bir mesafe ölçümüdür. Bu ölçüm her bir veri noktası ile ilişkili olduğu merkez arasında yapılır.

Algoritma aşağıdaki adımlardan oluşmaktadır.

1. Kümelenecek nesnelerden oluşan uzaya *k* tane nokta yerleştirilir. Bu noktalar kümeleri ifade eder.
2. Her bir nesne en yakındaki merkeze atanır.
3. Bütün nesneler atandıktan sonra *k* tane merkezin yeri yeniden belirlenir.
4. 2. ve 3. adımlar *k* tane merkezin yeri değişmeyene kadar tekrar edilir.

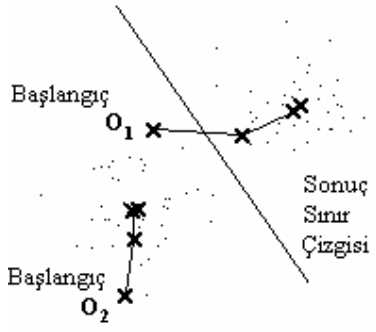
Prosedürün her zaman sonlanmasına karşı bu algoritma genel amaç fonksiyonuna karşılık gelen en iyi çözümü bulmada yeterli değildir. Algoritma, küme merkezlerinin rasgele olan başlangıç değerlerine duyarlıdır. Bu etkiyi azaltmak için defalarca çalıştırılması gerekebilir.

K-ortalamalar birçok probleme adapte edilebilir.

Bir Örnek:

x_1, x_2, …, x_n n tane örnek özellik vektörü olsun ve $k<n$ olmak üzere k tane kümeye ayrılsın. O_i ise i kümesinin merkezi olsun. Kümeler iyi bir şekilde ayrılırsa, minimum mesafe sınıflandırması uygulanabilir. Yani yeni bir noktanın tüm merkezlere olan mesafeleri karşılaştırılarak minimum olan kümeye ait olduğu söylenebilir. Bir x özellik vektörü için $\|x-O_i\|$ değeri k mesafe içinde minimum ise x vektörü i kümesindedir.

Aşağıdaki şekilde örnek olarak iki küme için küme merkezi O_1 ve O_2 nin hareketi gösterilmiştir.



Şekil 2.3. Örnek *k*-ortalamalar kümelemesi (k=2)

2.5.2 Bulanık *c*-ortalamalar

Bulanık c-ortalamalar, verinin bir kısmının birden fazla kümeye ait olmasına izin veren bir kümeleme yöntemidir. Bu yöntem 1973'te Dunn tarafından ortaya konmuş ve 1981'de Bezdek tarafından geliştirilmiştir ve desen tanımada (pattern recognition) sıkça kullanılmaktadır. Bu yöntem aşağıdaki amaç fonksiyonunun minimumlaştırılmasına dayalıdır (Bezdek, 1981).

$$J = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij} \left\| x_i - c_j \right\|^2 \qquad , 1 \le m \le \infty \tag{2.8}$$

Burada m 1'den büyük herhangi bir reel sayıdır. x_i i numaralı ve d boyutlu veridir, c_j yine d boyutlu küme merkezidir, u_{ij} ise x_i verisinin j kümesine üyelik derecesidir. ||*|| herhangi bir veri ile küme merkezi arasındaki benzerliğin bir ölçümüdür. u_{ij} üyeliklerinin ve c_j küme merkezlerinin aşağıdaki gibi güncellenmesi yoluyla yukarıdaki amaç fonksiyonu tekrarlı bir şekilde en iyi şekle getirilir ve bu sayede bulanık bölümleme sağlanmış olur.

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|}\right)^{\frac{2}{m-1}}} \tag{2.9}$$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m \cdot x_i}{\sum_{i=1}^{N} u_{ij}^m} \tag{2.10}$$

Ayrıca aşağıdaki şart sağlandığında döngü duracaktır:

$$\max_{ij} \left\{ \left| u_{ij}^{(k+1)} - u_{ij}^{(k)} \right| \right\} < \varepsilon \qquad (2.11)$$

Burada k tekrar sayısı ve ε 0 ile 1 arasında değer alan bir sonlandırma ölçütüdür. Bu sayede J_m yerel bir minimuma yakınsar.

Yöntem aşağıdaki adımlardan oluşmaktadır.

1. $U=[u_{ij}]$ matrisinin başlangıç değerlerini atanır, $U^{(0)}$
2. k numaralı adımda: $U^{(k)}$ ile $C^{(k)}=[c_j]$ küme merkezleri formül 2.10'e göre hesaplanır.
3. $U^{(k)}$, $U^{(k+1)}$ olarak Formül 2.9'a göre güncellenir.
4. Eğer || $U^{(k+1)}$ - $U^{(k)}$||< ε ise durur, değil ise 2. adıma döner.

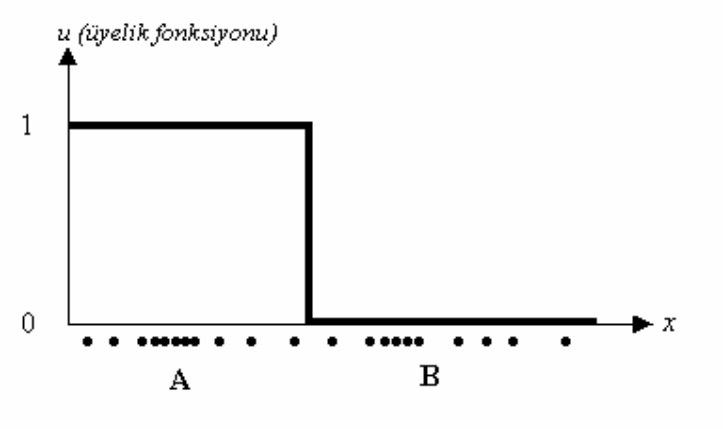
Veriler yöntemin bulanık yaklaşımını ortaya koyan üyelik fonksiyonunun ortalaması vasıtasıyla küme merkezleri belirlenir. Bunun için 0 ile 1 arasında değer alan ve her bir verilerin her bir küme merkezine olan üyelik derecesini belirleyen bir U üyelik fonksiyonu bulunmaktadır.

Tek boyutlu bir veri üzerinde konuyu açıklanacak olursa aşağıdaki gibi bir veri seti olsun.

Şekil 2.4. Tek boyutlu örnek bir veri seti

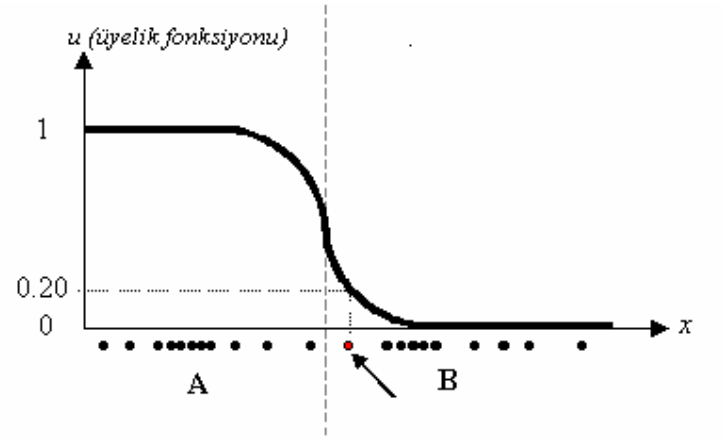
Şekil 2.4'e bakıldığında verinin iki merkezde yoğunlaştığı görülebilir. Bunlar 'A' ve 'B' kümeleri olarak adlandırılsın. İlk yaklaşımda (k-ortalamalar)

her bir veri sadece bir küme merkezi ile ilişkilendirilecek üyelik fonksiyonu aşağıdaki gibi olacaktır:



Şekil 2.5. Kesin üyelik eğrisi

Bulanık *c*-ortalamalar yönteminde ise her bir nokta sadece bir kümeye ait değildir. Bazı noktalar belirli bir üyelik değeri ile birkaç kümeye ait olabilmektedir.



Şekil 2.6. Bulanık üyelik eğrisi

Şekil 2.6'da A kümesinden çok B kümesine ait olan bir veri noktası ok ile işaret edilmiştir. Bu noktanın A kümesine aitlik derecesinin 0.2 olduğu görülmektedir. Grafik gösterim yerine matris şeklinde gösterilecek olursa *U* üyelik fonksiyonu matrisleri aşağıdaki gibi olacaktır.

$$U_{Mc} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ . & .. \\ 0 & 1 \end{bmatrix} \qquad U_{Mc} = \begin{bmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \\ 0.6 & 0.4 \\ .. & .. \\ 0.9 & 0.1 \end{bmatrix}$$

(a) (b)

Şekil 2.7. Kesin ve bulanık üyelik matrisleri

Şekil 2.7'de ise *k*-ortalamalar (a) ve bulanık *c*-ortalamalar (b) durumları için üyelik matrisi görülmektedir. Matristeki satır ve sütun sayısı kaç veri noktası ve kaç küme olduğuna bağlıdır. Yukarıdaki şekilde küme sayısının C=2 ve veri noktası sayısının N olduğu görülmektedir. Matrisin elemanlarının genelleştirilmiş şekli ise $u_{i,j}$ şeklinde olacaktır.

Bu matrisin bazı özelliklerine ise aşağıda yer verilmiştir.

- $u_{ij} \in [0,1] \qquad \forall i, j$
- $\sum_{k=1}^{C} u_{ik} = 1 \qquad \forall i$
- $0 < \sum_{i=1}^{N} u_{ij} < N \quad \forall N$

2.6. Bölüm Sonuçları

Günümüzde önemli bir araştırma alanı olan metin madenciliği metinsel belgelerle ilgili çeşitli ihtiyaçları karşılamayı amaçlamaktadır. Bu ihtiyaçlara ana başlıklar altında aşağıdaki gibi sıralanabilir:

- metinsel belgelerden bilgi çıkarımı,
- belgelerin organizasyonu,
- otomatik metin sınıflandırılması
- benzer belge aranması

Bu bölümde öncelikle genel olarak metin madenciliği kavramı üzerinde durulmuştur. Temel amaçları ve bu amaçlara yönelik yaklaşımlarına değinilmiştir. Ayrıca benzer belge aranmasının metin madenciliği içerisindeki yeri vurgulanmıştır. Daha sonra bulanık mantıkla ilgili genel bir bilgi verilmeye çalışılmıştır. Bulanık yaklaşımın daha iyi anlaşılabilmesi için kesin ve bulanık kümelemeye yöntemlerine birer örnek verilmiştir. İlerleyen bölümlerde ise benzer belge aranması işlemi için bulanık kümeleme konusu üzerinde odaklanılmıştır.

3. BULANIK KÜMELEME KULLANILARAK BENZER BELGE ARANMASI PROBLEMİ

3.1. Genel Tanımı ve Önemi

Bulanık mantık ve buna bağlı yöntemler günümüzde sıkça kullanılmaya başlanmıştır. Günümüzdeki mevcut birçok sınıflandırma yönteminin yanı sıra, metin madenciliği alanında bulanık kümeleme ve bulanık benzerliğe dayalı sınıflandırma ön plana çıkmaktadır.

Bu çalışmanın en önemli unsurlarından birisi benzer belge aranmasında bulanık kümeleme kullanımıdır. Benzer belge aranmasının temel yapı taşlarından birisi olan kümeleme/sınıflandırma aşamasında bulanık kümelemenin kullanılması önerilmiştir. Bulanık kümelemenin tercih edilmesinin en önemli sebebi ise metinlerin sınıflandırılırken birden fazla sınıfa veya kategoriye ait olanlarının da bulunmasıdır. Bu ise aynı anda birden fazla kümeye üye olma özelliğini düşündürmektedir. Bu yüzden kesin (crisp) üyelik yerine bulanık (fuzzy) üyeliklerin kullanılması üzerinde durulmuştur.

Burada ise bir benzer arama sisteminin genel yapısı, temel bileşenleri ve bu bileşenlere ait gerekli bilgiler verilecektir.

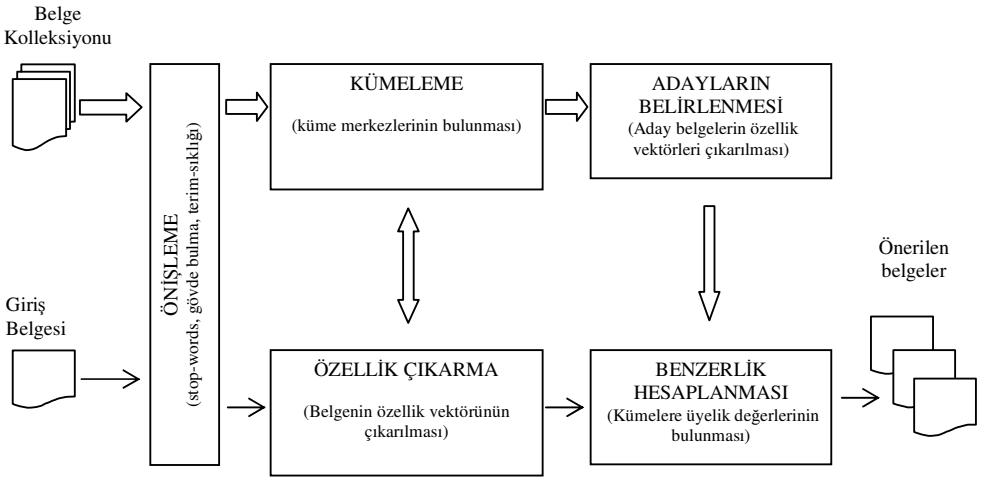
3.2. Temel Aşamaları

Her metin madenciliği işleminde olduğu gibi benzer belge aranırken de yapılacak işlemler bazı temel safhalardan oluşur. Bunlar en genel olarak:

- önişleme safhası
- sınıflandırma veya kümeleme safhası
- benzerlik ölçümü safhasıdır.

Benzeri aranan belge sırasıyla önişleme ve sınıflandırma/kümeleme safhalarından geçerek benzerlik karşılaştırması yapılabilecek forma dönüşür. Bu aşamada olan iki belge birbiri ile benzerliğinin hesaplanabileceği vektörler haline dönüşmüşlerdir. Son safhada ise benzerlik değerinin tespit edilmesi yer almaktadır.

Bu kısımda özetlenen bu arama yaklaşımın mimarisi Şekil 3.1'de görülmektedir. Şekilden de görüleceği gibi üzerinde arama yapılacak belgeler önceden sistem tarafından ön işlenmiş, sınıflandırılmış ve üzerinde arama yapılabilecek vektör formlarına dönüştürülmüşlerdir. Benzeri aranacak bir belge geldiğinde ise yine aynı önişleme ve sınıflandırma aşamalarında geçerek benzerlik hesaplanması yapılabilecek şekilde vektör formuna dönüşecektir. Son olarak aday belgeler ile benzerlik hesaplanması yapılarak en benzer belgeler bulunmuş olacaktır.



Şekil 3.1. Sistem Mimarisi

3.3. Ön İşleme Yöntemleri

Belgeler, arama sisteminin hem oluşturulmasında hem de kullanılmasında, bir başka deyişle sistemin hem eğitim hem de test aşamasında bazı önişleme işlemlerine tabi tutulurlar. Bunlar belgenin sisteme hazırlanması anlamına gelmektedir. Bu hazırlık işlemlerin bir kısmı yöntem ve araca bağlı olmayan, standart işlemlerdir. Bunlara örnek olarak; eğer belge bir HTML sayfası ise HTML taglarından, scriptlerden arındırılması, bir bilimsel makale ise yazar adı, yayımcısı gibi bilgilerden arındırılması verilebilir. Daha sonra bu belgeler içerdikleri kelimelere parçalanacaktır. Sıradaki işlem bu kelimelerin terimlere dönüştürülmesidir.

Terim çıkarımında Murata ve ark. (2000) göre 4 yöntem mevcuttur (Weng and Lin, 2003). Bunlar, sadece en kısa terimleri kullanmak, bütün terim örüntülerini kullanmak, bir kafes (lattic) kullanmak ve aşağı ağırlıklandırma (down-weighting) metodudur. Bu çalışmada işlemi basitleştirmek amacıyla bu dört yöntemden ilk yöntem seçilmiştir, ilerleyen kısımlardaki açıklanan işlemler buna dayanmaktadır.

3.3.1. Stopword Temizleme

Kelimelerin terimlere dönüştürülmesinin ilk aşaması durma kelime listesinin (stopword list) ayıklanmasıdır. Bir metnin içindeki bağlaç, zamir gibi kavramı ilgilendirmeyen kelimelere durma kelimeleri denir. Aşağıdaki tabloda İngilizce için Türkçe için verilen bazı durma kelimeleri gösterilmiştir.

Tablo 3.1. İngilizce ve Türkçe bazı stopword kelimeleri

İngilizce			**Türkçe**	
a	They		a	iki
all	third		ama	ise
and	though		bana	kez
can	thus		bin	mi
did	toward		bir	milyon
else	two		bu	nasıl
few	whereas		çünkü	ne
from	which		da	niye
gone	whom		elli	sen
himself	will		en	şey
perhaps	with		gibi	şundan
somebody	would		hepsi	tüm
tell	yet		her	veya
all	you		hiç	yani
the	zero		için	zira

Önceden oluşturulan durma kelime listesine göre işlenen belgeler bu kelimelerden ayıklanacaktır.

3.3.2. Gövde Bulma

Kelimelerin terimlere dönüştürülmesinde sonraki aşama ise gövde bulma işlemidir. Bu işlem, dilde yer alan ön ve son eklerin kelimelerden atılarak gövde hallerinin bulunması olarak özetlenebilir. Yapılan birçok araştırmaya göre gövde

bulma işlemi, bilgi çıkarımının veya metin madenciliğinin verimini artırmaktadır. Bu sayede dilsel eklerden dolayı farklı kelime gibi gözüken fakat aynı gövdeden oluşan terimler bulunabilecektir. Örnek olarak; “computation”, “computing” ve “computed” kelimeleri “compute” gövdesinden, “kullanılmıştır”, “kullanmak” ve “kullanılması” kelimeleri ise “kullan” gövdesinden oluşmaktadır.

Gövde bulmayla ilgili işlemler dille bağlantılıdır. Fakat genel olarak gövde bulma ile ilgili yaklaşımlar şu şekilde sıralanabilir:

- Eklerin kaldırılması
- Tabloya bakma
- N-Gram yöntemi

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

Tabloya bakma yönteminde ise önceden oluşturulmuş bir tabloda muhtemel tüm kelimeler ve bunlara karşılık gelen gövde halleri tutulmaktadır. Bu tablo üzerinden arama yapılarak bir kelimenin gövde hali bulunmuş olur. Tablo 2.2’de örnek bazı kelimeler ve gövde şekilleri verilmiştir.

Tablo 3.2. Örnek bazı kelimeler ve gövde halleri

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

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

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

Yine aynı kelime için trigramlar ise şunlardır:

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

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

Kelimeler arasındaki ilişki ise aşağıdaki formüle göre hesaplanır:

$$S = \frac{2C}{A + B} \qquad (3.1)$$

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

3.3.3. Terim Ağırlıklandırma

Kelimeler terimlere dönüştükten sonra belgenin terimler ile ifade edilmesi anlamına gelen sunum şeklinin belirlenmesi gerekmektedir. Bu terim ağırlıklandırması olarak ta adlandırılır. Bu saye de her bir belge içerdiği kelimelere göre bir vektör formunda yazılabilir. Bu belge vektörü genel olarak aşağıdaki formatta olacaktır:

$$d = (w_1, \ldots, w_i, \ldots, w_{|T|}) \tag{3.2}$$

Burada w_i, d belgesindeki i numaralı terimin ağırlık değerini, T ise terim kümesini göstermektedir ve $|T|$ ise toplam benzersiz terim sayısıdır.

Bu w_i değerlerinin belirlenmesinde farklı yaklaşımlar mevcuttur. Bunlara örnek olarak terimin belgedeki görülme sıklığı verilebilir. Ayrıca terimin belge koleksiyonunda toplam kaç belgede görüldüğü bilgisinin de eklendiği yaklaşımlar da mevcuttur. Bu yaklaşımların ayrıntıları ve bulanık kümeleme kullanılarak benzer belge aranması işlemine etkileri Bölüm 3'te incelenecektir.

3.4. Bulanık Kümeleme

Bulanık kümeleme kullanılarak benzer belge aranmasını diğer benzer belge arama yaklaşımlarından ayıran en önemli parçası bulanık kümelemedir. Burada kullanılan bulanık kümeleme Bulanık Benzerlik Sınıflandırması (Fuzzy Similarity Classification - FSC) tabanlı bir yöntemdir. FSC, eğitim veri setini kullanarak sınıf (kategori) merkezi vektörlerini bulur. Yeni bir belge bu merkez vektörleri ile

kıyaslanmak suretiyle sınıflandırılır. Bu çalışmadaki yaklaşımda ise yeni belgenin mevcut sınıflara aitlik derecelerinden oluşan vektör belgenin nitelik vektörü olarak ele alınmıştır. Bu vektörün elemanlarından maksimum olan ait olunan sınıfı gösterir. Ancak bu vektör tüm sınıflara olan aitlik değerlerini barındırdığından ve bu değerler benzerlik ölçümünde de işleme katılacağından, işlem kümeleme yapısındadır. Bu sebeplerden dolayı bu yaklaşım bulanık kümeleme şeklinde adlandırılmıştır.

İlerleyen bölümde FSC ile ilgili ayrıntılı bilgiye yer verilmiştir.

3.4.1. Bulanık benzerlik sınıflandırması (FSC)

Bu çalışmada da kullanılan bu yöntem, danışmanlı ve bulanık benzerliğe dayalı bir sınıflandırma yöntemidir (Widyantoro ve Yen, 2000). Bu yöntemin temelini oluşturan şey, terimler ile kategoriler arasında oluşturulacak bulanık bir ilişkidir.

$T=\{t_1,t_2,..,t_m\}$ terimlerinin $C=\{c_1,c_2,..,c_k\}$ kategorileri ile olan ilişkilerini belirlemek buradaki temel problemdir. Çözüm için kullanılan yöntemde terim-kategori ilişkisi aşağıdaki gibi tanımlanır (Miyamoto, 1990).

Verilen D eğitim kümesi yardımı ile R ($R=T$x$C \rightarrow [0,1]$) terim-kategori matrisi oluşturulacaktır. Bu matrisin her bir elemanı uygun gelen terimin uygun gelen kategoriye olan bulanık üyelik derecesini gösterecektir. t_i i numaralı terim ve c_j j numaralı kategori olmak üzere $\mu_R(t_i,c_j)$ üyelik derecesini göstermektedir.

Bu üyelikleri belirleyecek olan eğitim kümesi ise;

$$D=\{(d_1,c(d_1)),(d_2,c(d_2)),..,(d_n,c(d_n))\} \qquad (3.3)$$

olup, n belge içerir ve d_i i numaralı belgeyi, $c(d_i)$ ise i numaralı belgenin ait olduğu kategorileri gösterir. Eğitim kümesindeki her bir belge ise terim-ağırlık çiftinin bir kümesi şeklindedir.

$$d=((t_1,w_1), (t_2,w_2),.., (t_m,w_m)) \qquad (3.4)$$

burada w_i değeri t_i terim ağırlığını gösterir. Bunlara bağlı olarak $\mu_R(t_i,c_j)$ üyelik değerleri aşağıdaki gibi hesaplanır.

Bütün eğitim belgeleri ait oldukları kategorilere göre gruplara ayrılırlar. Birden fazla kategoriye ait belgeler ise ait oldukları her kategorinin grubu içinde yer alacaklardır. Her bir grup içindeki terimlerin ağırlık değerleri toplanacaktır. t_i terimin c_j kategori grubu içindeki toplam ağırlığı, aynı terimin tüm eğitim kümesindeki toplam ağırlığına bölünerek $dist(t_i,c_j)$ değeri bununmuş olacaktır. Eğer bu terim sadece bir kategoride geçiyorsa $dist(t_i,c_j)$ değerinin *1.0* olacağı açık bir şekilde görülebilir. Yine ne kadar fazla kategoride bulunduysa $dist(t_i,c_j)$ değeri o kadar düşük olacaktır. Terimin nihai $\mu_R(t_i,c_j)$ değerinin bulunması için ise her bir terim için bulunan $dist(t_i,c_j)$ değeri aralarındaki maksimum olana bölünür (Widyantoro ve Yen, 2000).

$$\mu_R(t_i,c_j) = \frac{dist(t_i,c_j)}{\max_{q,r}(t_q,c_r)} \qquad (3.5)$$

$$dist(t_i,c_j) = \frac{\sum_{w_i \in d_k \wedge d_k \in D \wedge c(d_k)=c_j} w_i}{\sum_{w_i \in d_k \wedge d_k \in D} w_i} \qquad (3.6)$$

Böylece terim ve kategori arasında bir bulanık ilişki kurulmuş olur. Buradaki her bir kategorideki terimler ve üyelik dereceleri o kategorinin küme merkezini göstermiş olur.

Sonraki aşamada ise hangi kategoriye ait olduğu araştırılan test belgesi bu kategorilerin küme merkezleri ile karşılaştırılarak, belgenin hangi kategoriye ait olduğu belirlenir. Bu aşamada test belgesi ile kategori merkezleri arasında bir benzerlik ölçümüne ihtiyaç duyulmaktadır.

Test belgesini aşağıdaki gibi olduğunu kabul edersek;

$$d= ((t_1, \mu_d(t_1)), (t_2, \mu_d(t_2)),.., (t_m, \mu_d(t_m))) \quad (3.7)$$

$\mu_d(t_i)$, d ye ait olan t_i nin üyelik derecesini gösterir. $\mu_d(t_i)$ için, verimli bir yaklaşım olan, belge içinde olan terimler için 1 değeri, olmayanlar için 0 değeri alınması yöntemi kullanılır.

Verilen bir R (TxC) terim kategori vektörü için, d belgesi ile c_j kategori merkezi arasındaki benzerlik aşağıdaki gibi hesaplanır.

$$sim(d, c_j) = \frac{\sum_{t \in d} \mu_R(t, c_j) \otimes \mu_d(t)}{\sum_{t \in d} \mu_R(t, c_j) \oplus \mu_d(t)} \quad (3.8)$$

$\otimes$, $\oplus$ sırasıyla bulanık kesişim ve birleşim operatörleridir. Bunlar için farklı formüller mevcuttur. Bunlardan en önemlileri Tablo 3.3'te görülmektedir (Widyantoro ve Yen, 2000).

d belgesinin her bir küme merkezine olan bulanık benzerliği bulunduktan sonra bunlardan en büyüğü, test belgesinin ait olduğu kümeyi belirtmektedir.

Yukarıda ayrıntıları verilen sınıflandırma yöntemi yardımıyla benzer belge aranması için iki temel işlev yerine getirilmiş olur.

- Sınıflandırma sonucunda benzerlik aranması yapılacak aday belgelerin bulunması
- Benzerlik karşılaştırması için özellik vektörünün oluşturulması

Tablo 3.3. Çeşitli t-norm ve t-conorm çiftleri

	t-norm(x,y)	**t-conorm(x,y)**
Einstein	$\frac{xy}{2-(x+y-xy)}$	$\frac{xy}{1+xy}$
Cebri (Algebraic)	xy	$x+y-xy$
Min-Max	min(x,y)	max(x,y)
Hamacher	$\frac{xy}{x+y-xy}$	$\frac{x+y-2xy}{1-xy}$
Sınırlandırılmış (Bounded)	max(0,x+y-1)	min(1,x+y)

Bir sonraki aşamada ise özellik vektörleri bazında bir benzerlik hesaplanması olacaktır. Bu hesaplama benzeri aranan belge ve kümeleme sonucu tespit edilen aday belgeler arasında olacaktır.

3.5. Benzerlik Ölçümü

Bu kısımda, özellik vektörleri şekline dönüşen belgelerin birbirleri ile karşılaştırılıp benzerliklerinin tespit edilmesi söz konusudur. Bu karşılaştırma işlemi için çeşitli benzerlik ölçümleri mevcuttur. Bunlardan başlıca olanları arasında kosinüs, zar benzerliği ve Minkowski metrik sayılabilir. Bu ölçümlerle ilgili ayrıntılı bilgi ve performans karşılaştırması Bölüm 5'te yer almaktadır.

3.6. Benzerlik Aramasının Sonuçlandırılması

Benzer belge aramada son işlem safhası olarak, benzeri aranan belge ve aday belgeler arasındaki benzerlik değeri hesaplanır. Bu hesaplamadan sonra büyükten

küçüğe doğru sıralanan belgeler sonuç olarak bulunan belgelerdir. Ancak benzerlik değerleri bulunan tüm bu aday belgelerin önerilmesi mümkün değildir. Bu yüzden en çok benzeyen belirli sayıdaki belgelerin seçilmesi gerekir. Bu noktada bir eşik değeri belirlenerek bu değeri aşan belgeler önerilmesi sağlanır.

3.7. Sınıflandırma Performansı Karşılaştırma

Metin madenciliği araştırmalarının önemli bir kısmını sınıflandırma ile ilgili çalışmalar oluşturur. Bu çalışmalardaki sınıflandırma yöntemleri ile ilgili bir kıyaslama ölçütü olarak en sık karşılaşılan yöntem Precision-Recall'dur.

3.7.1. Precision-Recall

Sıkça kullanılan bu *precision* ve *recall* ölçütleri aşağıda verilmiştir (Kowalski, 1997).

$$precision = \frac{dogru\_olarak\_bulunan\_kategoriler}{bulunan\_kategorilerin\_toplamı} \qquad (3.9)$$

$$recall = \frac{dogru\_olarak\_bulunan\_kategoriler}{dogru\_kategorilerin\_toplamı} \qquad (3.10)$$

Buradaki *dogru_olarak_bulunan_kategoriler* pozitif örnek sayısını, *bulunan_kategorilerin_toplamı* çıkarım sonundaki toplam örnek sayısı, *dogru_kategorilerin_toplamı* ise test verilerindeki o kategoriye ait örnek sayısıdır.

Precision doğru olarak önerilen belgelerin önerilen belgelere oranıdır. Recall ise doğru olarak önerilen belgelerin, tüm (önerilmesi gereken) test verilerine olan oranıdır.

3.7.2. F-ölçüsü

Bunlara ek olarak bu her iki ölçümün de aynı anda iyi olmasını amaçlayan F-Ölçüsü kullanılmaktadır (Larsen ve Aone, 1999). F-Ölçüsü aşağıdaki gibi hesaplanır:

$$F = \frac{2PR}{P+R} \tag{3.11}$$

burada P precision, R ise Recall'dır.

3.8. Bölüm Sonuçları

Bu bölümde genel olarak bulanık kümeleme kullanılan bir benzer belge arama yaklaşımı ortaya konmuştur. Bu arama yaklaşımını oluşturan temel aşamalar genel olarak ele alınmıştır. İlerleyen bölümlerde bu aşamalar hakkında daha ayrıntılı bilgiler verilecektir. Ayrıca bu aşamalardaki yöntemlerin geliştirilmesi

amaçlanmıştır. Bu geliştirme ile ilgili önerilen yeni yöntemlere de yine ilerleyen bölümlerde yer verilecektir.

Ayrıca bu bölümde açıklanan arama yaklaşımın önemli bir özelliği ise içerisinde bulanık kümeleme kullanılmasıdır. Bulanık mantık günümüzde birçok alanda olduğu gibi metin madenciliği ve bir dalı olarak benzer belge aranmasında da kullanılmaktadır. İşte bu bölümde yer verilmiş olan konulardan birisi de bu kullanımdır. Burada çeşitli bulanık yöntemlere değinilmiş ve bu yöntemlerden FSC üzerinde durulmuştur. Önerilen arama yaklaşımı ve FSC yönteminin birlikte kullanımı incelenmiştir.

4. METİN SINIFLANDIRMASINDA METİNSEL BELGELERİN SUNUM YÖNTEMLERİNİN KARŞILAŞTIRMASI VE BENZERLİK ÖLÇÜMLERİ

4.1. Giriş

Metin madenciliğinde kullanılan metin sınıflandırması metinsel bir belgenin önceden belirlenmiş bir veya daha fazla sınıfa atanması problemidir. Sınıflandırmadaki önemli konulardan biri de üzerinde metin madenciliği yöntemlerinin uygulanacağı belgenin sunumunudur. Yani belgenin terimler cinsinden hangi prensibe göre ifade edileceğini belirlemektir. Bu işlem terim ağırlıklandırma olarak ta adlandırılmaktadır.

Bu çalışmada terim ağırlıklandırma işlemi için birkaç alternatif yöntem karşılaştırılmıştır. Belgedeki terimlerin ağırlık derecelerini belirlemek için Terim Sıklığı (Term Frequency - TF), normalize edilmiş Terim Sıklığı (normalized Term Frequency - nTF), Terim Sıklığı Ters Belge Sıklığı (Term Frequency Inverse Document Frequency - TFIDF) ve normalize edilmiş Terim Sıklığı Ters Belge Sıklığı (normalized Term Frequency Inverse Document Frequency - nTFIDF) yöntemleri uygulanmıştır.

Bu bölümün deneysel çalışma kısmında metinsel belge koleksiyonu olarak Reuter-21578 dağıtım 1.0 kullanılmıştır. Bu çalışmanın sonucu olarak bulanık benzerlik için TF sunumunun diğer yaklaşımlardan daha iyi sonuç verdiği görülmüştür. İlerleyen kısımda ise çalışmaya altyapı oluşturan yöntemlerden bahsedilecektir.

4.2. Araştırma Altyapısı

Bu çalışmada kullanılan sınıflandırma yöntemi FSCdir (Widyantoro ve Yen, 2000). Bölüm 3.4.1 de ayrıntıları verilmiş olan bu yöntemin temelini, terimler ile kategoriler arasında kurulacak bulanık bir ilişki oluşturmaktadır. D bir eğitim veri kümesi ve T belge içinde geçen tüm terimlerin kümesi olmak üzere, bu eğitim kümesindeki her bir belge terim-ağırlık çiftinin bir kümesi şeklindedir: $d=\{(t_1,w_1), (t_2,w_2),.., (t_{|T|},w_{|T|})\}$ burada w_i değeri t_i terim ağırlığını gösterir. $\mu_R(t_i,c_j)$ üyelik değerleri bunlara bağlı olarak formül 3.5 ve formül 3.6 ile hesaplanır. Bu değerler ayrıca küme merkezi vektörleri anlamına da gelmektedirler. Daha sonra test belgesi bu vektörlerle formül 3.8 yardımıyla bulanık benzerliğe tabi tutulur. Bu işlem sonrasında test belgesinin hangi kategori merkezine daha yakın olduğu bulunmuş olur. Bu kategori ait olunan kategori olarak kabul edilir.

Yöntemin yukarıda açıklanan bu yapısından dolayı w_i değerlerinin sınıflandırmada çok önemli bir yere ve etkiye sahip olduğu açıktır.

4.3. Terim Ağırlıklandırma Yöntemleri

Yukarıda açıklandığı üzere, t_i terimin belge içindeki w_i ağırlığı sınıflandırmanın temel yapı taşını oluşturmaktadır. Terim ağılıklarının belirlenmesinde farklı yöntemler mevcuttur. Bu çalışmada kullanılan ve sınıflandırma üzerindeki performansları karşılaştırılan yöntemler aşağıda sıralanmıştır.

4.3.1. Terim sıklığı

Ağırlık belirleme yöntemlerinin en basiti TF olarak bilinen terimin belge içindeki bulunma sayısıdır.

$$w_i = tf_i \tag{4.1}$$

w_i, d belgesinde t_i teriminin ağırlığı;

tf_i, d belgesinde t_i teriminin sıklığı;

Bu yöntem bulanık kümelemede kullanılırken yukarıda da açıklandığı üzere, her bir t_i teriminin her bir c_j kategori grubu içindeki toplam ağırlığı, aynı terimin tüm eğitim kümesindeki toplam ağırlığına bölünerek $dist(t_i,c_j)$ değeri bununmuş olacaktır.

4.3.2. Ağırlıklı terim sıklığı

Ağırlık belirlemedeki bir başka yöntem ise, TF değerlerinin karesel ağılıkların toplamının karekköküne bölünmesi yoluyla elde edilen, nTF ağılıklarıdır. Verilen bir w ağırlığının normalleştirme fonksiyonu ise aşağıdaki gibidir.

$$nw = \frac{w}{\sqrt{\sum_{k=1}^{|T|} w_k^2}} \tag{4.2}$$

Burada |*T*|, *T* terim kümesindeki terim sayısını göstermektedir.

Bu yöntemde her bir t_i teriminin her bir c_j kategori grubu içindeki toplam ağırlığının, aynı terimin tüm eğitim kümesindeki toplam ağırlığına bölünmesine gerek kalmamaktadır. $dist(t_i,c_j)$ değeri olarak bulunan bu normalleştirilmiş değerler alınmaktadır.

4.3.3. Terim sıklığı ters belge sıklığı

Üçüncü yöntem olarak metin madenciliğinde sıkça kullanılan TFIDF ağılıkları farklı bir alternatifi oluşturmaktadır. Terim sıklığı tf_i, t_i teriminin belgedeki görülme sıkılığı olsun. Belge sıklığı df_i ise t_i terimin en az bir kere görüldüğü belge sayısı olsun. Ters belge sıklığı idf_i, df_i kullanılarak aşağıdaki gibi hesaplanır.

$$idf_i = \log\left(\frac{|D|}{df_i}\right) \tag{4.3}$$

Burada |D|, *D* eğitim veri setindeki toplam belge sayısıdır. Eğer terim fazla sayıda belgede görülmüş ise bu değer düşük, az sayıda belgede görülmüş ise bu değer yüksek olacaktır. w_i ise en son olarak aşağıdaki gibi bulunur.

$$w_i = tf_i \cdot idf_i \tag{4.4}$$

4.3.4. Ağırlıklı terim sıklığı ters belge sıklığı

Terim ağırlıklandırma ile ilgili olarak dördüncü yöntem, TFIDF değerlerinin yukarıda yer alan formül 4.2 yardımıyla elde edilen, nTFIDF ağılıklardır.

4.4. Karşılaştırma metodu

Bu çalışmanın bu kısmındaki temel amaç, belgedeki terim ağırlıklandırma yöntemlerinin bulanık sınıflandırma üzerindeki etkisini araştırmaktır. Bu amaca yönelik olarak FSC kullanılan bir yaklaşım mevcuttur. Bu sınıflandırma yaklaşımı dört farklı terim ağırlıklandırma yöntemi ile denenmek suretiyle yaklaşımların performansları karşılaştırılmıştır. Bu karşılaştırma için kullanılan yöntem ise Precision-recall yöntemidir.

4.5. Metinsel Sunum Yöntemleri İçin Deneysel Sonuçlar ve Analizi

Deneylerde kullanılan belge koleksiyonu olarak metin madenciliği araştırmalarında sıkça kullanılan Reuters-21578 dağıtım 1.0 seçilmiştir. Bu koleksiyon 135 in üzerinde konuya sahip 21578 belge içermektedir. Mevcut olan bu konulardan bazıları çok az sayıda belgede bulunmaktadır. Bu yüzden en çok belgenin dâhil olduğu 10 konu seçilerek bu çalışmada kullanılmıştır. 8595 adet seçilen belgeden, 6456 adedi eğitim verisi olarak, 2139 adedi ise test verisi olarak kullanılmıştır.

Geleneksel olarak metin madenciliği çalışmalarında ön işleme mevcuttur. Bu çalışma için seçilen eğitim verileri de bir ön işlemeye tabi tutulmuştur. Bu ön işlemede öncelikle bu belgelerden stop-words kelimelerin çıkarılması yer almaktadır. Burada kullanılan stop-words listesi 350 kelimeye sahiptir. Daha sonra ön işleme işlemi olarak gövdeleme işlemi (stemming) gelmektedir. Bu işlem için diğer çalışmalarda da yaygın bir şekilde kullanılan Porter Stemmer algoritması seçilmiştir. Bu aşamadan sonra bulanık benzerliğe dayalı olarak bir sınıflandırma işlemi yapılmıştır.

Eğitim sürecinde 10 konuya ait bir belge seti kullanılmıştır. Ancak eğitim verilerinde birden fazla konusu olan belgeler mevcut olduğu gibi test verileri için de aynı durum söz konusudur. Fakat sistemin test aşamasında ait olunan konu olarak en yüksek değere sahip olan bir konu seçilmiştir. Test işlemindeki bu 10 konudan bir kısmı sürekli bir başka konu ile beraber gözlemlenmektedir ancak yalnız başına gözlemlenmemektedir. Test verilerinin bu yapısından dolayı test sonuçlarında 10 konudan 6 konu test sonuçlarının %99 gibi büyük bir kısmını oluşturmaktadır. Bu durumda çeşitli sunum yöntemleri arasında bir performans karşılaştırması yapabilmek için sınıflandırma işleminin eğitim safhasında 10 konu kullanılmasına karşı, test esnasındaki değerlendirme altı konu üzerinden yapılmıştır.

Uygulama programlarının hazırlanması için MatLab 7.0 yazılım paketi kullanılmıştır.

Terim ağırlıklandırma yöntemleri arasındaki performans ölçümü, diğer sınıflandırma uygulamalarının performans ölçümlerinde de sıkça kullanılan Precision-Recall yöntemi ile yerine getirilmiştir.

Tablo 4.1. Sonuçlar

	Precision	**Recall**	**F-Ölçüsü**
TF	0.8619	0.7403	0.7713
nTFIDF	0.8599	0.7393	0.7697
TFIDF	0.8559	0.6339	0.6824
nTF	0.8712	0.4974	0.5278

Dört farklı sunum yönteminin deneysel olarak karşılaştırılmasından elde edilen sonuçlar Tablo 4.1'de görülmektedir. Burada yer alan değerler her bir yöntem için (TF, nTF, TFIDF, nTFIDF), 6 konunun her birine ait değerlerin aritmetik ortalamalarıdır.

Normalize edilmiş TF yöntemi (nTF) precison değeri olarak en yüksek olmasına karşı recall değeri oldukça düşüktür. TF yöntemi ise hem recall hem de F-Ölçüsü değeri olarak en iyi sonucu vermiştir. Bu sonuçlara göre bulanık benzerliğe dayalı mevcut sınıflandırma sistemi için en uygun belge sunum yöntemi TF yöntemidir.

4.6. Benzerlik Ölçümü Probleminin Tanımı ve Çerçevesi

İki sayısal değer veya değer dizisinin birbirleri ile olan benzerlik ilişkisinin bir fonksiyonu olarak tanımlanabilecek olan benzerlik ölçümü, birçok alanda karşımıza çıkan önemli bir gereksinimdir. Özellikle tanıma ve sınıflandırma yöntemlerinin özünde ya bir benzerlik ölçümü (similarity measure) veya bir benzemezlik/mesafe ölçümü (disimilarity/distance measure) yer alır.

4.7. Benzer Belge Aranmasında Benzerlik Ölçümünün Yeri

Genel bir benzer belge arama yaklaşımı verilen Bölüm 3'te yer aldığı üzere benzerlik ölçümü arama sisteminin üç temel unsurundan biridir. Özellik vektörü formuna dönüştürülen belgelerin karşılaştırılmasında verimli bir benzerlik ölçümü yaklaşımın performansını doğrudan etkileyecektir.

Genellikle arama mekanizmalarında anahtar kelimelere dayalı indeks yapısı kullanılır. Büyük miktardaki koleksiyonlar üzerinde bu sayede hızlı arama yapılabilir. Bu çalışmada da yer aldığı üzere, eğer arama işlemi için belgelerin içerdikleri tüm kelimeler kullanılacak ise arama uzayının boyutu büyük önem kazanmaktadır. Etkin bir arama yapabilmek için, arama işlemini hızlandıracak bir yapı gerekmektedir. Bu çalışmanın amacı bulanık kümeleme kullanılarak oluşturulan bir benzer belge aranması yaklaşımı için hızlı ve verimli bir benzerlik ölçümü ortaya koymaktır. Bunu sağlamak için geleneksel benzerlik ölçümleri yerine verinin boyutuna dayalı yeni bir benzerlik ölçümü önerilmiştir.

4.8. Benzerlik Ölçümü Yöntemleri

Benzer belge aramada sık kullanılan benzerlik ölçümlerinden en önemlileri kosinüs, zar benzerlik ölçümleri ve Minkowski metriktir. İlerleyen kısımlarda bunlarla ilgili daha detaylı bilgiler verilmiştir.

4.8.1. Kosinüs Benzerliği

Metin madenciliğinin yanı sıra birçok alanda da kullanılan kosinüs benzerliği, iki nokta arasındaki açıya dayalıdır. Aşağıdaki gibi hesaplanmaktadır.

$$CSim(x_i, x_j) = \frac{\sum_{k=1}^{m} x_{ik} x_{jk}}{\sqrt{\sum_{i=1}^{m} (x_{ik})^2 \sum_{i=1}^{m} (x_{jk})^2}} \qquad (4.5)$$

Burada *x* özellik vektörü, *m* ise verinin boyutudur.

4.8.2. Zar Benzerliği

Bir başka benzerlik ölçümü ise zar benzerliğidir. Bu benzerliğin formülü aşağıda verilmiştir.

$$DSim(x_i, x_j) = \frac{2\sum_{k=1}^{m} x_{ik} x_{jk}}{\sum_{i=1}^{m} (x_{ik})^2 + \sum_{i=1}^{m} (x_{jk})^2} \tag{4.6}$$

4.8.3. Minkowski Metrik

Yukarıdaki ölçümlere ek olarak, çok boyutlu veriler için, çokça kullanılan bir başka benzerlik ölçümü ise Minkowski metriktir ve aşağıda gösterilmiştir (Ichino and Yaguchi, 1994):

$$Minkowski_p(x_i, x_j) = \left(\sum_{k=1}^{m} \left| x_{ik} - x_{jk} \right|^p \right)^{\frac{1}{p}} \tag{4.7}$$

Burada *x* özellik vektörü, *m* ise verinin boyutudur.

Minkowski metrikte, p=2 özel durumu için formül *Euclidean* mesafesini ifade ederken, p=1 için ise *Manhattan* mesafesini ifade etmektedir. Ne var ki, her hangi bir uygulama için benzerlik ölçümü seçiminde rehberlik edecek genel bir kural bulunmamaktadır.

4.9. Sistem Mimarisi

Benzer belge arama sistemlerinde amaç giriş belgesine en çok benzeyen belgeleri bulmaktır. Giriş belgesi geleneksel soru cevaplandırma sistemlerinde bir soruya karşılık gelir (Clarke ve ark., 2000; Elworthy, 2000). Bu çalışmada ise giriş belgesi aranacak belgelerle aynı biçime sahip bir belgedir.

Sistem öncelikle giriş belgesini bir ön işleme safhasından geçirir. Belgenin içerdiği tüm terimlerle oluşturulacak, belgenin özellik vektörü elde edilir. Daha sonra kümeleme işlemine tabi tutulan giriş belgesi için ait olduğu küme bulunur. Bu kümedeki belgeler aday belgelerdir. Bu aday belgelerin kümelere aitlik derecelerini içeren özellik vektörleri ile giriş belgesinin özellik vektörü karşılaştırılarak benzerlikler tayin edilir. Bu karşılaştırma için veri boyutuna dayalı yeni bir benzerlik ölçümü önerilmiştir.

Sistemin genel yapısı Şekil 3.1'de görüldüğü gibidir. Ön işleme safhasında amaç terimleri ve belgelere göre bulunma sıklıklarını belirlemektir. Bu yöntemde belge kelime bazında parçalanır ve terim olarak bu kelimeler alınır. Daha sonra bilgi çıkarımında önem taşımayan kelimelerin (stop-words) ayıklanması işlemi yapılır. Bu kelimeler gövdeleme (stemming) işlemine tabi tutulur. En son elde edilen kelimeler terimler olarak kabul edilir. Her bir belgedeki terimleri bulunma sıklığı hesaplanır. Önceden mevcut belge koleksiyonu tarafından eğitilmiş bir bulanık kümeleme sistemi mevcuttur. Terimlerin bulunma sıklığı hesaplanmış olan giriş belgesi, bu kümeleme sistemini yardımıyla aday belgelere yönlendirilir. Bu sayede ilk iki aşama (ön işleme ve kümeleme) tamamlanmış olur.

Bir sonraki aşamada ise amaç aday belgeleri benzerliklerine göre sıralamaktır. Belirlenen aday belgeler ile giriş belgesi veri boyutuna dayalı bir benzerlik ölçümüne tabi tutulur. Böylece benzerlik değerleri bulunmuş olur. Burada seçilecek bir eşik değerini aşan aday belgeler giriş belgesine benzer belgeler olarak önerilir.

4.10. Önerilen Benzerlik Ölçümü

Burada önerilen veri boyutuna dayalı benzerlik ölçümü (Boyut Kök Benzerliği, Dimension Root Similarity) aşağıdaki gibidir.

$$DRSim(x_i, x_j) = \left(\frac{\sum_{k=1}^{m} \left| x_{ik} - x_{jk} \right|^2}{m} \right)^{\frac{1}{m}} \qquad (4.8)$$

Burada x_i ve x_j karşılaştırılan belgelerin özellik vektörleri, m. ise özellik vektörünün boyutudur.

Sonuç olarak, benzeri aranan belge mevcut koleksiyonun tamamıyla karşılaştırılmak yerine, önceden hazır olan kümeleme çerçevesinde belirli bir aday grubuna yönlendirilir. Ve daha sonra yine binlerce olabilen terim seviyesindeki bir karşılaştırma yerine, önceden belirlenmiş sınırlı sayıda kümelere olan aitlik dereceleri seviyesinde benzerlik karşılaştırmasına tabi tutulur. Bu sayede tüm terimlerden elde edilmiş özellik vektörü kullanılarak kısa sürede benzerlik tespiti yapılmış olacaktır.

4.11. Benzerlik Ölçümleri İçin Deneysel Sonuçlar ve Analizi

Çalışmanın bu bölümündeki uygulama kısmında iki farklı belge koleksiyonu kullanılmıştır. Bunlardan ilki metin madenciliğinde sıkça kullanılan Reuters-21578 dağıtım 1.0'dır. Bu koleksiyon 135 konuya ait 21578 belgeden oluşmaktadır. Bu konulardan bazıları çok az sayıda belge içermektedir. Bu yüzden, 135 konu arasından en çok belge kapsayan 10 konu seçilmiştir. Bu durumda seçilen 10 konuya ait toplam 8595 belge bulunmaktadır ve bunlardan 6456 adedi eğitim verisi geriye kalanı ise test verisi olarak kullanılmıştır.

İkinci belge koleksiyonu 3 farklı kategoriye ait toplam 4020 özet içermektedir. Bunlardan bilgisayar bilimleri ile ilgili "computer collection" 1587 özet, tıpla ilgili "Medlars collection" 1033 özet, aerodinamik ile ilgili "Cranfiled collection" ise 1400 özet içermektedir. Mevcut 4020 belgelik bu koleksiyondan 3015 adet özet eğitim verisi olarak geriye kalanı ise test verisi olarak kullanılmıştır.

Uygulama programlarının hazırlanması için MatLab 7.0 yazılım paketi kullanılmıştır.

Belgeler öncelikle bir ön işleme safhasından geçirilmiştir. Bu safha da ilk olarak 350 kelimelik stop-words listesi bu belgelerden ayıklanmıştır. Kelimelerin gövdeleme (stemming) işlemi için ise yine literatürde sıkça kullanılan Porter Stemmer algoritması seçilmiştir (Jones ve Willett, 1997). Bunun sonucunda kelimeler terimlere dönüştürülmüş olmaktadır. Böylece belgeler içerdikleri bu terimlere göre kümelenmiştir.

Seçilen eğitim verileri formül 3.5 ve 3.6'daki bulanık benzerlik metodu ile kümeleme işlemine tabi tutulmuştur.

Benzerlik ölçümlerinin (formül 4.5, 4.6, 4.7 ve 4.8) karşılaştırılması için aşağıdaki yöntem uygulanmıştır. Aynı kümeye ait rasgele seçilen 100 belge çiftinden bir koleksiyon oluşturulmuştur. Bu koleksiyona ait çeşitli benzerlik ölçümleri için ortalama benzerlik değerleri bulunmuştur. Benzer şekilde başka bir koleksiyon ise farklı kümeye ait rasgele seçilen 100 belge çifti ile oluşturulmuştur. Yine aynı

işlemler bu koleksiyona da uygulanmıştır. Benzerlik ölçümlerini karşılaştırabilmek için, her bir benzerlik ölçümüne ait "aynı küme koleksiyonu" ve "farklı küme koleksiyonu" değerleri birbirine oranlanmıştır. Ayrıca zaman karşılaştırması yapabilmek için ise, her bir benzerlik ölçümü için her iki koleksiyonun karşılaştırılmasına (toplam 200 karşılaştırma) harcanan süre alınmıştır. Yukarıda bahsedilen yöntem 10 kez tekrar edilerek ortalama sonuçlar alınmıştır. Deneysel sonuçlar göstermiştir ki, önerilen yeni benzerlik ölçümü mevcutlardan daha iyi sonuç vermiştir. Elde edilen ortalama benzerlik değerleri Tablo 4.2 ve Tablo 4.3'de görülmektedir.

Karşılaştırılan Benzerlik ölçümleri ise şunlardır:

- Boyut Kök Benzerliği (Dimension Root Similarity-DRSim)
- Zar Benzerliği (Dice Similarity-DSim)
- Kosinüs Benzerliği (Cosine Similarity-CSim)
- Manhattan Benzerliği (Manhattan Similarity-MSim)
- Öklid Benzerliği (Euclidean Similarity-ESim)
- *p* parametresinin boyut değeri alındığı *Minkowski* Benzerliği (MDSim)
- *p* parametresinin 20 alındığı *Minkowski* Benzerliği (M20Sim)
- *p* parametresinin 50 alındığı *Minkowski* Benzerliği (M50Sim).

DRSim benzerlik ölçümü diğer benzerlik ölçümlerinden ve geleneksel Kosinüs benzerliğinden daha verimlidir. Bu verimlilik Tablo 4.2 ve Tablo 4.3 teki "Ayrıştırma Oranı" sütununda da açık bir biçimde görülmektedir. Bu sütunda, aynı kümeye ait belge çiftleri ile farklı kümeye ait belge çiftleri arasındaki toplam benzerlik değerlerinin oranları gösterilmiştir.

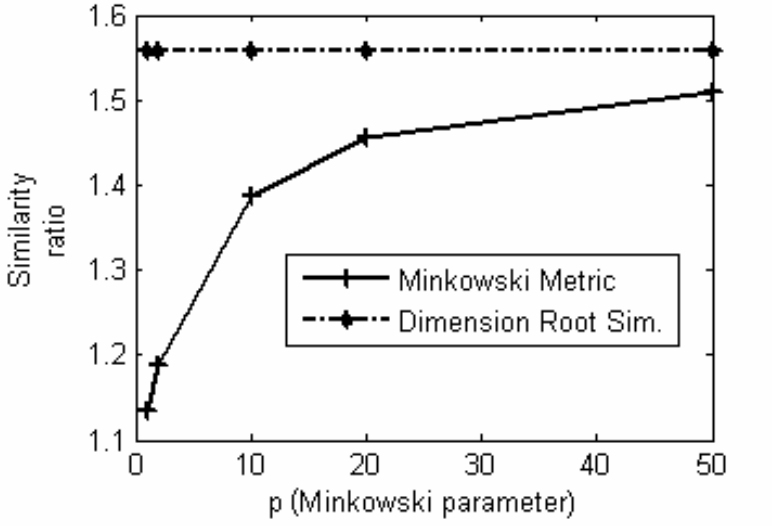
Tablo 4.2. Router Koleksiyonuna ait deneysel sonuçlar

	Aynı Küme	**Farklı Küme**	**Ayrıştırma oranı**	**Toplam Süre (ms)**
DRSim	0.405599	0.260450	1.5573	7.673
DSim	0.952949	0.734766	1.2969	10.407
CSim	0.955275	0.739114	1.2925	10.736
MSim	0.929763	0.819675	1.1343	6.986
Esim	0.911628	0.767758	1.1876	7.455
MD(10)Sim	0.856086	0.617108	1.3873	12.891
M20Sim	0.840893	0.577592	1.4559	13.202
M50Sim	0.830160	0.549523	1.5107	13.735

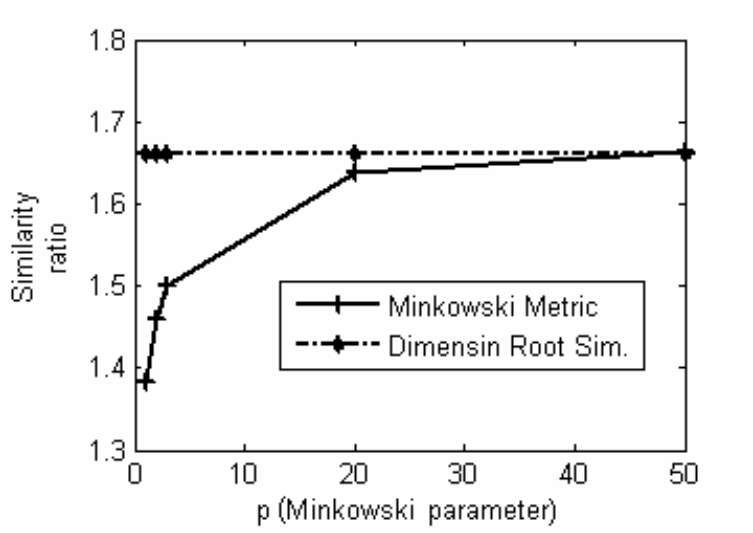
Tablo 4.3. İkinci Koleksiyona ait deneysel sonuçlar

	Aynı Küme	**Farklı Küme**	**Ayrıştırma oranı**	**Toplam Süre (ms)**
DRSim	0.784886	0.472734	1.6603	7.080
DSim	0.979118	0.766672	1.2771	7.469
CSim	0.980513	0.768626	1.2757	7.023
MSim	0.907535	0.655913	1.3836	7.078
Esim	0.895316	0.613923	1.4583	7.220
MD(3)Sim	0.887064	0.590788	1.5014	7.938
M20Sim	0.860155	0.524840	1.6389	8.734
M50Sim	0.851423	0.512209	1.6623	9.718

Önerilen benzerlik ölçümünün verimlilik karşılaştırması Şekil 4.1 ve Şekil 4.2'de görülmektedir. Şekil 4.2'de de görülebileceği gibi, *p* parametresinin daha büyük değeri daha verimli bir sonuç ortaya koymaktadır (*p* parametresinin değerinin artışı benzerlik verimini artırmaktadır).

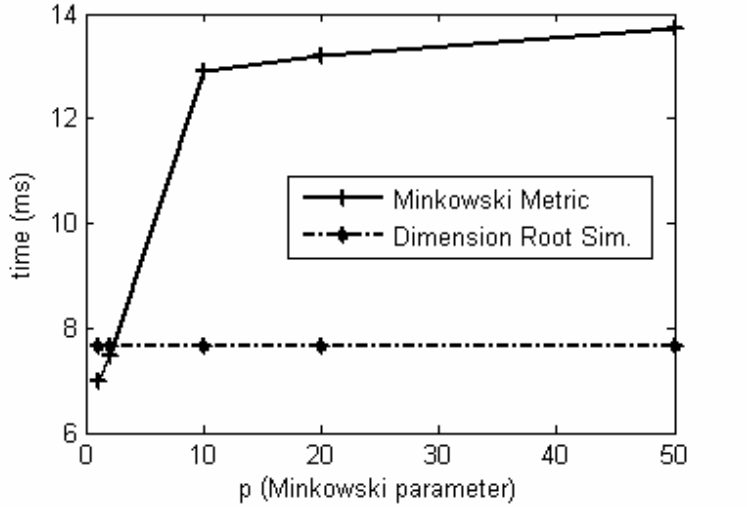


(a)

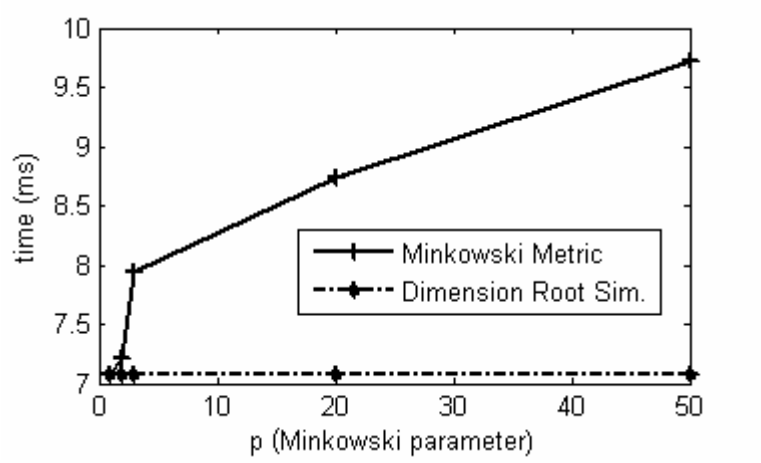


(b)

Şekil 4.1. (a) Reuter Koleksiyonu (10 kategori); (b) İkinci Koleksiyon (3 kategori). Minkowski Metrik ve DRSim arasındaki benzerlik oranları



(a)



(b)

Şekil 4.2. (a) Reuter Koleksiyonu (10 kategori); (b) İkinci Collection (3 kategori). Minkowski Metrik ve DRSim arasındaki zaman karşılaştırması

DRSim ve *p* değerinin 50 alındığı Minkowski metrik için benzerlik oranı sonuçları birbirine oldukça yakın çıkmıştır. Fakat Minkowski metrikte *p* parametresi koleksiyondaki kategori sayısı alındığında (Reuter için 10, ikinci koleksiyon için 3), önerilen DRSim daha iyi sonuç vermektedir. Bu sonuç Reuter içi %12 ve ikinci koleksiyon için ise %10 daha iyidir.

Şekil 5.2'de ise zaman karşılaştırması gösterilmiştir. Buradan da açık bir şekilde görülebileceği gibi, Minkowski metrikte *p* parametresinin 2 den büyük tüm değerleri için DRSim daha iyi sonuç vermektedir. Yine Minkowski metrikte *p* parametresi her bir koleksiyondaki kategori sayısı alındığında, DRSim, Reuter için %68 ve ikinci koleksiyon için ise %12 daha hızlıdır (Saraçoğlu ve ark, 2007).

Minkowski metrikte *p* parametresi 50 değerine kadar seçilmiştir. Önerilen yeni benzerlik ölçümü *p* değerinin 1-50 aralığında olduğu değerden daha iyi sonuç vermiştir.

4.12. Bölüm Sonuçları

Belge sınıflandırma metin madenciliğinde temel bir işlemdir. Önceki metin madenciliği çalışmalarında çoğunlukla bu konu üzerinde durulmuştur. Bundan dolayı sınıflandırma veriminin artırılması, metin madenciliği uygulamalarının veriminin artırılması açısından çok önemlidir.

Bu bölümün ilk kısmında FSC yaklaşımı için yaklaşım performansının en iyi olduğu terim ağırlıklandırma yöntemi araştırılmıştır. Bu kapsamda yapılmış olan bir program yardımıyla karşılaştırılan terim ağırlıklandırma yöntemlerinden TF (Term Frequency) yönteminin diğerlerine göre daha iyi sonuç verdiği gözlemlenmiştir.

Bu sonuç benzer belge aranması konusu için de önem arz etmektedir. Çünkü terim ağırlıklandırma benzer belge aranması sisteminin üç temel bileşeninden biri olan ön işlemenin bir parçasıdır. Bu yüzden bulanık mantık kullanarak benzer belge aranması için hangi terim ağırlıklandırma yönteminin daha başarılı olduğu ortaya konmuş olmaktadır.

Bu bölümün ikinci kısmında ise benzer belge aramada benzerlik ölçümü üzerine odaklanılmıştır. Mevcut benzerlik ölçümlerine karşılaştırılmalı olarak yer verildikten sonra veri boyutuna dayalı yeni bir benzerlik ölçümü tanımlanmıştır.

Deneysel sonuçlara göre, önerilen benzerlik ölçümü (DRSim) geleneksel benzerlik ölçümlerinden daha iyi performans göstermiştir.

Bu çalışmada, bir belge birden fazla kategoriye ait ise bu durum eğitim aşamasında göz önünde bulundurulmaktadır. Ancak çalışmanın buraya kadarki kısmında, test aşaması için belgelerin sadece bir tek kategoriye ait olabilecekleri düşünülmektedir. Ve sadece bu kategorideki belgeler ile benzerlik karşılaştırılması yapılmaktadır. İlerleyen bölümde, belgelerin test aşamasında da birden fazla kategoriye ait olabilecekleri göz önüne alınmıştır. Bu sayede daha kapsamlı bir arama yapabilme amaçlanmıştır.

5. BELGELERİN BİRDEN FAZLA KATEGORİYE AİT OLMA PROBLEMİ

5.1. Çoklu Kategori Kavramı

Çoklu kategori durumu, bir belgenin birden fazla alan veya konu ile ilgili olması şeklinde tanılanabilir. Çok disiplinli bir bilimsel araştırma makalesi bu tür belgelere güzel bir örnektir. Mesela, hastalık tespitinde YSA kullanımı ile ilgili bir makale, hem yapay zekâ konusu hem de tıpla ilgilidir. Bu belgenin kategorize edilmesinde belge her iki konuya da dâhil edilebilir. Çoklu kategori durumuna günlük haber metinlerinde de rastlamak mümkündür. Örneğin bir futbolcunun transfer haberi hem spor ile ilgili hem de finans ile ilgili bir haber niteliği taşıyabilir. Bu şekilde birçok belge sadece tek bir konu veya alanla ilgili olmak yerine aslında birden fazla alan veya konu ile ilgilidir. Bu alanların kendi aralarında bir öncelikleri olabilir, ancak bu durum belgenin birden fazla kategoriye ait olduğu gerçeğini değiştirmemektedir. Belgenin ilgili olduğu konulardan en öncelikli konunun ait olunan konu olarak seçilmesinin bir bilgi ve özellik kaybına yol açacağı kesindir.

5.2. Benzer Belge Aranmasındaki Çoklu Kategori Probleminin Yeri

Önceki çalışmalarda eğitim verisindeki belgelerden bazıları birden fazla kategoriye ait olabilmektedir. Bu durum göz önüne alınarak eğitim sürecinde böyle bir belge, birden fazla küme merkezi hesaplanmasında kullanılmaktadır. Ancak kümeleme veya sınıflandırma oluştuktan sonra, test verisi için birden fazla

kategoriye üyelik ihtimali göz ardı edilmektedir. Test verisi için ait olduğu kategori tek bir kategori olarak kabul edilmektedir. Bu yüzden, bir belge için aday kategorilerden sadece yüksek değere ait kategori nihai sonuç olarak alınmaktadır. Bu açıdan, çoklu kategori durumu benzer belge arama veya metin madenciliği sistemlerinde ele alınmamıştır.

Özellikle benzer belge aranması işleminde yukarıda açıklanan eksiklikten kaçınmak önemlidir. Bu noktada iki soru karşımıza çıkmaktadır. Birincisi bir belgenin bir kümeye mi yoksa birden fazla kümeye mi ait olduğudur. İkinci soru ise bir belge birden fazla kümeye ait ise bu kümelerin hangileri olduğudur. Bu yüzden bu çalışmada, kümeleme sonucunda sadece tek bir küme belirlemek yerine birden fazla küme belirlemek amaçlanmıştır. Çalışmada bu belirtilen probleme, bulanık kümeleme yardımıyla bir çözüm aranmıştır.

Çoklu kategori probleminin ele alınması beraberinde sınıflandırma ve kümeleme yöntemlerinin kullanılmasını ve geliştirilmesini gündeme getirmektedir. İlerleyen kısımda sınıflandırma ve kümeleme konusu üzerinde durulmuştur.

5.3. Metin Madenciliğindeki Bazı Sınıflandırma/Kümeleme Yöntemleri

Genel olarak kümeleme, bilgi noktalarını farklı homojen sınıflara veya kümelere bölerek, aynı küme içindeki öğelerin mümkün olduğunca birbirine benzer ve farklı küme içindeki öğelerin mümkün olduğunca birbirinden ayrı olmasını sağlar. Sınıflandırma veya kümeleme yöntemleri çok değişik şekilde kategorize edilebilirler. Bu kategorize yaklaşımlarından en önemlileri aşağıda verilmiştir.

Öncelikle sınıflandırma veya kümeleme yöntemleri için temelde iki öğrenme yaklaşımı vardır. Bunlar gözetimli öğrenme ve gözetimsiz öğrenmedir. Gözetimli öğrenmede, önceden sınıfların bilinmesi ve bu sınıflara ait verilerden oluşan bir öğrenme kümesi gerekir. Gözetimsiz öğrenmede ise öğrenme kümesindeki verilerin sınıfların önceden bilinmesine ihtiyaç yoktur. Gözetimli öğrenme içeren yöntemlere

örnek olarak *k*-NN (*k*- en yakın komşu), naive Bayes (NB), DVM, Rocchio Algoritması, FSC verilebilir. Gözetimsiz öğrenme içeren yöntemlere örnek olarak ise *k*-Ortalamalar (*k*-Means), FCM (Bulanık *c* ortalamalar) verilebilir.

Kümeleme yöntemleri bölümlemeli ve hiyerarşik olarak da kategorize edilmektedir. Bölümlemeli yöntemlerde, belirli bir sayıda küme seçilir ve veriler kümeler arasında geçiş yaparak en iyi durum bulunmaya çalışılır (*k*-Ortalamalar, FCM vs.). Hiyerarşik yöntemlerde ise küme sayısı olarak 2 den başlanır kümeler bölünür ve küme sayısı artarak en iyi kümelemeye ulaşılır (divisive). Veya küme sayısı olarak veri sayısından başlanır (her bir veri bir kümeyi ifade eder) kümeler birleşir ve küme sayısı azalarak en iyi kümelemeye ulaşılır (agglomerative).

Bir başka açıdan ise yine bu yöntemler katı-kesin (hard-crisp) ve esnek-bulanık (soft-fuzzy) olarak kategorize edilirler. Bunların arasındaki temel fark ise her bir verinin kümelere üyelik değerleridir. Katı olarak adlandırdığımız yöntemlerde eğitim safhasında bir veri sadece bir sınıf/kümeye ait olur iken, bulanık olarak adlandırdığımız yöntemlerde ise bir veri farklı üyelik dereceleri ile birkaç sınıf veya kümeye ait olabilmektedir.

Metin madenciliği çalışmalarında veri boyutu oldukça büyüktür. Çünkü belge koleksiyonu içinde geçen tüm terimlerden bir sözlük oluşturulur ve her bir belge, bu sözlük ile (belge koleksiyonundaki tüm terimler cinsinden) ifade edilir. Bu ise veri boyutunun on binlerce olabileceği anlamına gelmektedir. Çalışma alanının bu özelliğinde dolayı bu çalışmada sınıflandırma/kümeleme yöntemleri farklı bir açıdan kategorize edilmiştir. Bu ise yöntemlerin nitelik azaltmaya ihtiyaç duyup duymamalarıdır. Yöntemler bu özelliklerine göre kategorize edilmiştir.

Yöntemlerin bu tür kategorize edilmesinin sebeplerinden birisi ise sınıflandırma/kümeleme işlemi öncesinde bir nitelik azalma işleminin veride özellik kaybına yol açmasıdır. Çünkü nitelik azaltma işleminde ya bazı nitelikler tamamen atılacak (genetik algoritma, IG gibi) yada tüm nitelikleri temsil edecek şekilde daha az sayıda nitelik oluşturulacaktır (PCA gibi).

Bir diğer sebebi ise veri boyutunun çok büyük olduğu durumlarda (örneğin 20000) istatistiksel nitelik azaltma yöntemlerinin (örneğin PCA) uygulama güçlüğüdür.

Veride nitelik azaltma işlemine ihtiyaç duymayan ve metin madenciliğinde sıkça kullanılan sınıflandırma/kümeleme yöntemlerine örnek olarak Rocchio algoritması, FSC ve NB yaklaşımı verilebilir. Bu yöntemlerle ilgili ayrıntılı bilgiye 4. bölümde yer verilmiştir.

Nitelik azaltmaya ihtiyaç duyan yöntemlere örnek olarak ise *k*NN, Yapay Sinir Ağları (YSA), *k*-ortalamalar, FCM, DVM verilebilir. Bu yöntemlerin nitelik azaltılmadan uygulanması oldukça güçtür. Örneğin 20000 girişli bir YSA tasarımı ve eğitilmesi oldukça yavaş olacaktır.

Bu çalışmanın yapısına uygun olan ayrıca, yukarıda sayılan sebeplerden dolayı nitelik azaltmaya ihtiyaç duymayan sınıflandırma yöntemlerinin en önemlileri ve bunlarla ilgili ayrıntılar aşağıda verilmiştir.

5.3.1. Rocchio algoritması

Metin sınıflandırmasının amacı, belgeleri önceden belirlenmiş sınıflara ayırmaktır. Rocchio algoritması metin madenciliğinde en popüler ve geniş bir uygulama alanı bulan danışmanlı bir öğrenme yöntemidir (Rocchio, 1971). Bu algoritmanın en önemli parçasını TFIDF oluşturmaktadır (Salton ve Buckley, 1988). İlk olarak bilgi çıkarımı için geliştirilen yöntem, belgelerin sınıflara üyeliklerini belirlemede bir karar kuralı oluşturmayı amaçlar. Avantajı ise bu kuralı oluştururken bir eşik değerine bağlı olmamasıdır. Bu yöntem daha sonra metin sınıflandırmasına da adapte edilmiştir. Bu yöntem aşağıdaki gibi tanımlanabilir (Joachims, 1997).

D eğitim belge setini, *C* kategorileri ve *T* terim kümesini göstermek üzere, her bir *d* belgesinin belirli bir kategoriye atanmış olduğu kabul edilir. Her bir *d* belgesi, tüm terimlerin ağırlık değerlerinin bir vektörü şeklinde sunulur.

$$d = \left(w_1, \ldots, w_i, \ldots, w_{|T|} \right) \qquad (6.1)$$

Burada w_i, belgedeki i numaralı terimin ağırlık değerini göstermektedir. w_i değeri ise Bölüm 4.3.3'te ayrıntılı bir şekilde açıklanan TFIDF yöntemi ile hesaplanır. Hatırlanacağı üzere, bir terim belgede sıkça geçiyorsa önemi artmaktadır. Fakat aynı zamanda çok sayıda belgede de bulunuyorsa bu terimin önemi azalacaktır.

Bu yöntemde, öğrenme sonucunda her bir C_j sınıfını göstermek üzere bir c_j vektörü oluşturulmaktadır. Bunun için bahsedilen sınıfa ait normalize edilmiş pozitif örnekler ile yine bu sınıfa ait olmayan normalize edilmiş negatif örnekler kullanılır. C_j vektörü bu örneklerin ağırlıklandırılmış farklarından aşağıdaki gibi hesaplanır.

$$c_j = \alpha \frac{1}{\left|C_j\right|} \sum_{d \in C_j} \frac{d}{\|d\|} - \beta \frac{1}{\left|D - C_j\right|} \sum_{d \in D - C_j} \frac{d}{\|d\|} \qquad (6.2)$$

Burada α ve β, sırasıyla pozitif ve negatif örneklerin göreceli etkilerinin ayarlanmasını yapan parametrelerdir. Bunların önerilen değerleri α=16, β=4'dür (Buckley et al., 1994). C_j j sınıfına ait eğitim belgelerini göstermektedir. $\|d\|$ ise d vektörüne ait öklit mesafesidir (Euclidian length). Bunlara ek olarak Rocchio algoritması c_j vektöründeki negatif bileşenlere 0 değeri atanmasını gerektirir.

Sınıf vektörlerinin seti tamamlandıktan sonra her bir sınıfa ait bir vektör mevcuttur ve bu vektörler öğrenmiş modeli oluşturmaktadır. Yeni bir belge d' (test belgesi) bu model yardımıyla sınıflandırılabilir. d' belgesi da yukarıda belirtildiği gibi bir vektör formundadır. Bu belgenin ait olduğu sınıfı bulmak için belge ile her biri sınıf vektörünün kosinüsü hesaplanır. En yüksek kosinüs değerine sahip olan vektör ait olunan sınıfın vektörü olacaktır.

$$H_{TFIDF}\left(d'\right) = \arg\max \cos\left(c_j, d'\right) \qquad (6.3)$$

argmax f(x), f(x) fonksiyonunu maksimum yapan x değerini döndürmektedir. *H_TFIDF* ise algoritmanın d' belgesini atandığı sınıfı göstermektedir.

5.3.2. Naive Bayes

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

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

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

$$H_{BAYES}(d') = \arg\max_{c_j \in C} \Pr(c_j \mid d') \qquad (6.4)$$

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

$$\Pr(c_j \mid d') = \frac{\Pr(c_j) \cdot \prod_{i=1}^{|d'|} \Pr(w_i, c_j)}{\sum_{c' \in C} \Pr(c') \cdot \prod_{i=1}^{|d'|} \Pr(w_i, c')} \tag{6.5}$$

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

$$\Pr(c_j) = \frac{|c_j|}{\sum_{c' \in C} |c'|} = \frac{|c_j|}{|D|} \tag{6.6}$$

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

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

$$\Pr(w_i, c_j) = \frac{1 + TF(w_i, c_j)}{\sum_{w' \in d' \in C_j} TF(w', c_j)} \tag{6.7}$$

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

Son olarak Formül 6.5 aşağıdaki şekle dönüşmüş olur:

$$H_{BAYES}(d') = \arg\max_{c_j \in C} \frac{\Pr(c_j) \cdot \prod_{i=1}^{|d'|} \Pr(w_i, c_j)}{\sum_{c' \in C} \Pr(c') \cdot \prod_{i=1}^{|d'|} \Pr(w_i, c')} \tag{6.8}$$

5.3.3. Sınıflandırma Yöntemlerinin Karşılaştırma Kriterleri

Bir sınıflandırma yönteminin performansını belirlemek için çeşitli yöntemler mevcuttur. Bu yöntemlerin en bilinenlerinden biri ve aynı zamanda bu bölümde kullanılan yöntem ise Precision-Recall yöntemidir. Bu yöntemin ayrıntılarına Bölüm 2.7'de yer verilmiştir.

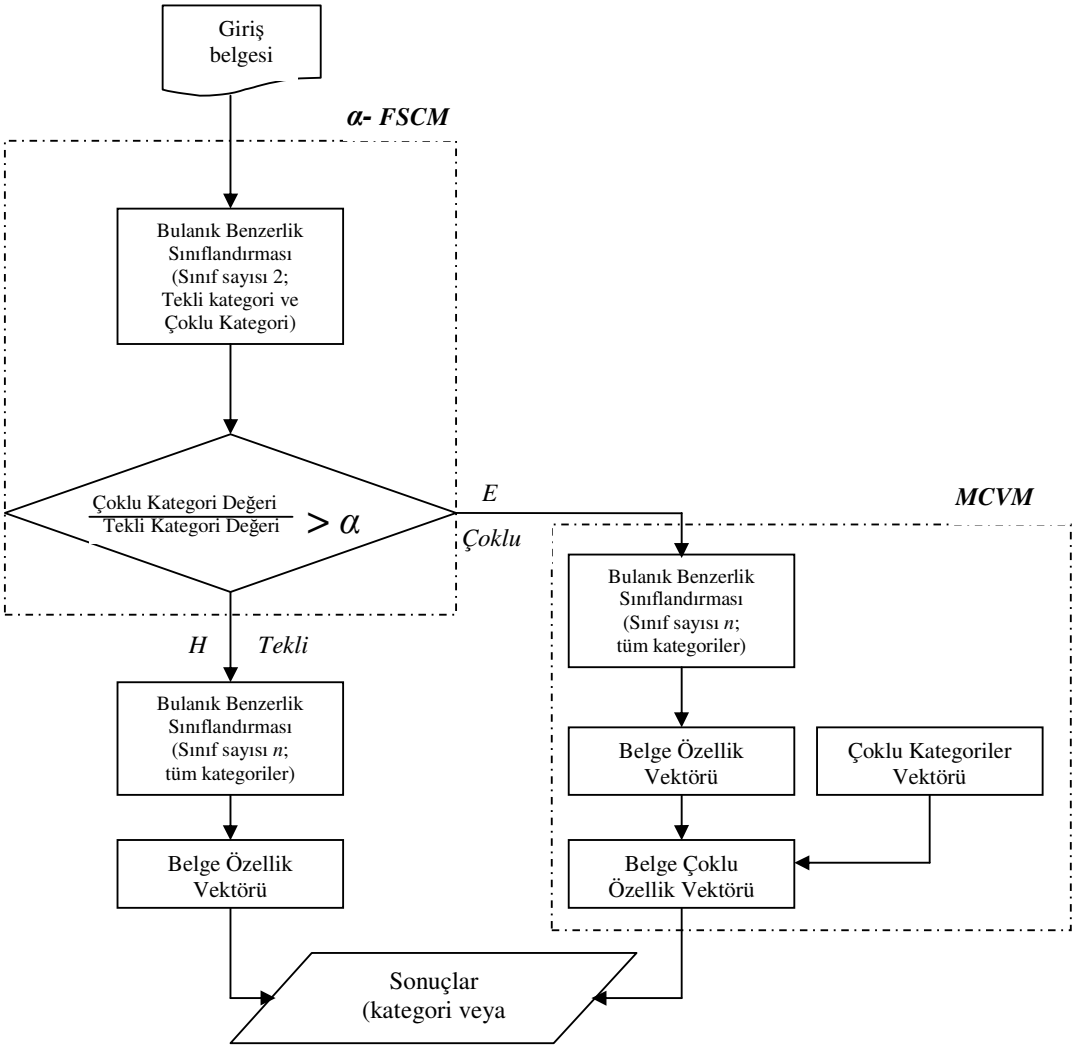
5.4. Genel Bir Benzer Belge Arama Sistemi

Genel bir benzer belge arama sistemi Şekil 3.1'de görülen bir yapıya sahiptir. Ön işleme safhasında amaç terimleri ve belgelere göre bulunma sıklıklarını belirlemektir. Bu yüzden belge kelime bazında parçalanır ve bu kelimeler terim olarak adlandırılır. Daha sonra bilgi çıkarımında önem taşımayan kelimelerin (stop-words) ayıklanması işlemi yapılır. Bu kelimeler gövdeleme (stemming) işlemine tabi tutulur. En son elde edilen kelimeler terimler olarak kabul edilir. Her bir belgedeki terimlerin bulunma sıklığı hesaplanır. Önceden mevcut belge koleksiyonu tarafından eğitilmiş bir bulanık kümeleme sistemi mevcuttur. Bu kümeleme işlemine tabi tutulan giriş belgesi için ait olduğu küme veya kümeler tespit edilir. Bu küme veya kümelerdeki belgeler aday belgelerdir.

Bu aday belgelerin kategori özellik vektörleri ile giriş belgesinin kategori özellik vektörü karşılaştırılarak benzerlik değerleri hesaplanır. Bu karşılaştırma için veri boyutuna dayalı bir benzerlik ölçümü kullanılmıştır. Böylece benzerlik değerleri bulunmuş olur. Burada seçilecek bir eşik değerini aşan aday belgeler giriş belgesine benzer belgeler olarak önerilir.

Sonuç olarak, benzeri aranan belge mevcut koleksiyonun tamamıyla karşılaştırılmak yerine, önceden hazır olan kümeleme çerçevesinde belirli bir aday

grubuna yönlendirilir. Daha sonra yüzlerce olabilen terim seviyesindeki bir karşılaştırma yapılmaz. Bunun yerine, önceden belirlenmiş sınırlı sayıda kümelere olan aitlik dereceleri seviyesinde benzerlik karşılaştırmasına tabi tutulur. Bu sayede tüm terimlerden elde edilmiş özellik vektörü kullanılarak kısa sürede benzerlik tespiti yapılmış olacaktır.



Şekil 5.1. Kategorilerin Belirlenmesinin Akış Şeması

5.5. Çoklu Kategori Probleminin Çözümü İçin Önerilen Yaklaşım

Benzer belge aramada çoklu kategori probleminin ele alındığı bu kısımda çoklu kategori problemi, bir test belgesinin birden fazla kategoriye ait olması durumunda bu kategorilerin tespiti şeklinde özetlenebilir.

Bu çoklu kategori problemi iki kısımda ele alınabilir. Birincisi hangi belgelerin birden fazla kategoriye ait olduğunun tespitidir. İkincisi ise, bir belge birden fazla kategoriye ait olduğu durumda bunların hangi kategoriler olduğunun tespitidir. Bu iki kısım birbirinden bağımsız olarak düşünülebilir. Bu iki kısım için önerilen iki yöntem sırasıyla *α-threshold Fuzzy Similarity Classification Method* (*α-FSCM*) ve *Multiple Categories Vector Method* (*MCVM*) dir. Şekil 5.1'de verilen akış şemasında bu iki kısım için önerilen iki çözüm yöntemi *α*–FSCM ve MCVM görülmektedir. Bu iki yöntem ilerleyen bölümlerde ayrıntılı bir şekilde açıklanmaktadır.

5.5.1. *α*-FSCM

Çalışmanın bu kısmının odak noktası kümeleme işlemidir. Bundan dolayı mevcut bulanık benzerlik ölçümüne dayalı kümeleme sistemi geliştirilmeye ve bu sayede çoklu kategori probleminin çözümüne çalışılmıştır

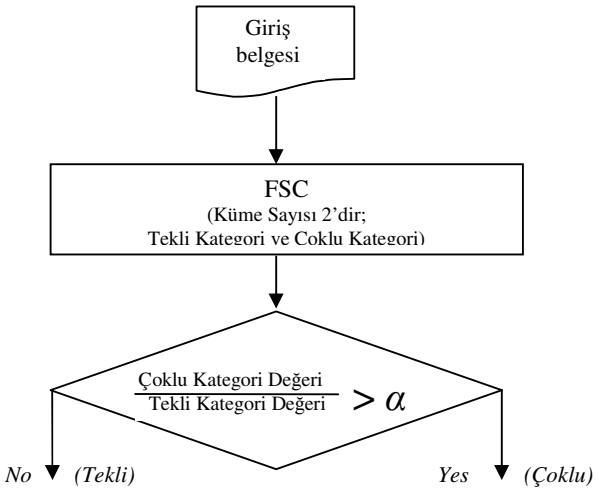
Hangi belgelerin birden fazla kategoriye ait olduğu probleminin çözümü için önerilen *α*-FSCM, bölüm 3.4.1'de bahsedilen FSC yi temel almaktadır. Yöntem açıklaması ise aşağıda yer almaktadır.

FSC geleneksel olarak kategori sayısı adedince küme merkezi vektörlerinin hesaplanmasını içerir. Bu hesaplama için eğitim verisi kullanılır. Bu küme merkezi vektörleri kümelemeyi ifade eder. Her kategori bir kümeye karşılık gelmektedir. Test

belgesi ise tüm bu küme merkezi vektörleri ile bulanık benzerliğe tabi tutulur. Bu sayede test belgesinin kümelere aitlik değeri bulunur. Küme ve kategoriler bire bir eşleştiği için test belgesinin kategorilere aitlik değerleri de bulunmuş olur.

Birden fazla kategoriye ait belgelerin tespiti problemi için önerilen yöntemin temelinde ise yukarıdakine benzer bir yaklaşım mevcuttur. Sadece iki kümenin olduğu (tekli ve çoklu kategori kümeleri) bir küme merkezi hesaplama (kümeleme) yapılacaktır. Önceki yaklaşımdan farklı olarak küme merkezleri doğrudan iki kategori ile (tekli kategori sınıfı - çoklu kategori sınıfı) eşleşmemektedir. Bu küme vektörleri yardımıyla hesaplanan tekli kategori küme değeri ve çoklu kategori küme değeri ek işlemlerden geçtikten sonra kategoriyi belirleyecektir.

Bu tekli-çoklu kategori sınıflandırmasını, klasik bulanık benzerlik ölçümü yöntemi ile çözülmesi mümkün olmamaktadır. Bunun sebebi, belge koleksiyonunun iki sınıfa dağılması fakat bu dağılımdan sınıf ayrımının doğrudan ve açık bir şekilde ortaya çıkarılamamasıdır. Ayrıca çoklu kategori sınıfına ait belgeler tüm koleksiyonun küçük bir kısmını oluşturmaktadır. Bu ise klasik FSC'nin sınıfları açıkça ayırt etmesini güçleştirmektedir. Yapılan uygulamalardan elde edilen sonuçlar da bunu doğrular niteliktedir.



Şekil 5.2. α- FSCM'nin Akış Şeması

Yukarıda sayılan sebeplerden dolayı klasik FSC yaklaşımı geliştirilerek çoklu kategori problemine adapte edilmiştir. Bunu sağlayan ek işlemler ve yaklaşımın genel yapısı Şekil 5.2'de görülmektedir. Klasik FSC den elde edilen çoklu kategori küme değeri ve tekli kategori küme değeri oranlanmaktadır. Önceden belirlenen bir α eşik değeri ile karşılaştırılmaktadır. Bu eşiği aşanlar çoklu kategori sınıfına, eşiğin altında kalanlar tekli kategori sınıfına atanmaktadır.

5.5.2. α değerinin belirlenmesi

Bu yöntemde eşik değerinin belirlenmesi önem kazanmaktadır. Bu eşik için önerdiğimiz genelleştirilmiş yöntem ise aşağıda açıklanmıştır.

D bir eğitim veri seti, *M* ve *S, D* nin ayrık birer alt kümesi olsun (*M* birden fazla kategoriye ait olan belgelerin seti, *S* ise sadece bir kategoriye ait olan belgelerin seti).

$$\begin{aligned} M \cup S &= D \\ M \cap S &= \varnothing \end{aligned} \qquad (6.9)$$

|*M*| ve |*S*| sırasıyla *M ve S* belge setinin içerdiği belge sayısını gösterir.

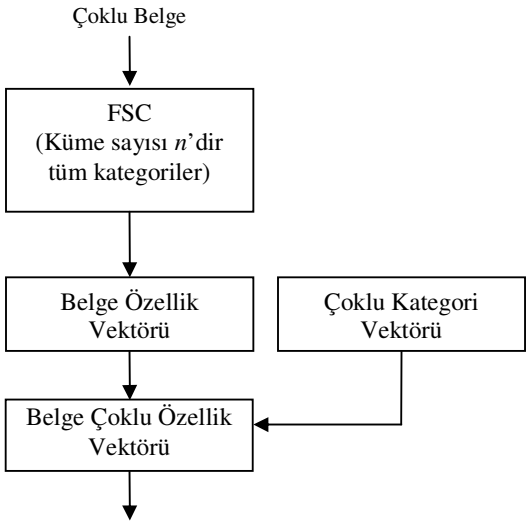
Sınıflandırma için öncelikle klasik FSC uygulanarak *D* belge setindeki her bir *d* belgesi için $sim(d,c_M)$ ve $sim(d,c_S)$ değerleri (çoklu kategori kümesine üyelik değeri ve tekli kategori kümesine üyelik değeri) elde edilir. Bu değerler bir birine oranlanır. Belgeler bu oran değerlerine göre büyükten küçüğe doğru sıralanır. Bu sıralama sonucu ilk |*M*| tane belge, yöntem gereği sınıflandırmamızın sonucunda çoklu kategori sınıfına atanacak belgelerdir. Sıralamadaki |*M*| numaralı belgenin oran değeri ise bizim aradığımız α eşik değeri olarak seçilir. Çünkü eşik değeri bu değer seçildiğinde buna göre belgeler iki parçaya ayrılacaktır. Bu değer ve üzerindeki |*M*| tane belge çoklu kategoriye ait olacaktır. Bu eşik değeri altındaki geriye kalan |*S*|

adet belge ise tekli kategoriye ait olacaktır. Eğitim veri seti üzerinden tespit edilen bu eşik değeri artık test belgeleri için sınıflandırma yapılırken de kullanılacaktır.

5.5.3. MCVM

Çoklu kategori probleminin ikinci kısmı ise bir belge birden fazla kategoriye ait olduğunda bunların hangi kategoriler olduğunun tespiti edilmesidir.

Klasik FSC yönteminde, bir belge için belge-kategori vektöründe maksimum değere sahip olan kategori ait olunan kategori idi. Eğer belge birden fazla kategoriye ait ise, ait olduğu kategoriler belge-kategori vektöründeki sırayla maksimum değere sahip olan kategorilerdir.



Şekil 5.3. MCVM'nin Akış Şeması

Örneğin belge iki kategoriye ait ise, ait olunan ilk kategori maksimum değerli olandır, ikinci kategori ise ilk kategoriden sonra maksimum değere sahip olan kategoridir. Bu klasik bir yaklaşımdır. Fakat klasik yaklaşımın bu çalışmada yapılan uygulamada başarısız sonuçlar verdiği görülmüştür. Bu klasik yaklaşım yerine önerdiğimiz ise MCVM'dir. Blok diyagramı Şekil 5.3'te görülen bu yöntem aşağıda açıklanmaktadır.

$C=\{c_1,c_2,..,c_K\}$ kategorilerinin birbirleri ile olan ilişkilerini belirlemek buradaki temel problemdir. Çözüm için kullanılan yöntemde kategori-kategori ilişkisi aşağıdaki gibi tanımlanır.

Verilen D eğitim veri seti yardımıyla, MC (MC: CxC → [0,1]) çoklu kategori matrisi oluşturulacaktır. Bu matrisin her bir elemanı uygun gelen kategorilerin birbirleri ile olan bulanık birliktelik derecelerini gösterecektir. c_i i numaralı ve c_j j numaralı kategori olmak üzere $\mu_{MC}(c_i,c_j)$ üyelik derecesini göstermektedir. Bu üyelikleri belirleyecek olan eğitim kümesi ise $D=\{(d_1,c(d_1)),(d_2,c(d_2)),\ldots,(d_n,c(d_n))\}$ olup, n belge içerir ve d_i i numaralı belgesi, $c(d_i)$ ise i numaralı belgenin ait olduğu kategorileri gösterir. $\mu_{MC}(c_i,c_j)$ üyelik değerleri aşağıdaki gibi hesaplanır.

$$\mu_{MC}(i,j)=\begin{cases}\dfrac{c_i \sqcap c_j}{\sum_{\substack{k=1\\k\neq j}}^{K} c_k \sqcap c_j} & , \quad if \quad i>j \\ 1 & , \quad if \quad i=j \\ \dfrac{c_i \sqcap c_j}{\sum_{\substack{k=1\\k\neq i}}^{K} c_i \sqcap c_k} & , \quad if \quad i<j \end{cases} \tag{6.10}$$

Buradaki $c_i \sqcap c_j$ ifadesi, aynı anda hem i hem de j kategorisine ait belgelerin sayısını, K ise toplam kategori sayısını göstermektedir.

Aşağıdaki teoremler önerilen yöntemlerin bazı özelliklerini tespit etmektedir.

Teorem 5.1. *Çoklu kategori matrisi MC simetriktir.*

$$\mu_{MC}(i,j) = \mu_{MC}(j,i)$$

İspat: $c_i \square c_j$ ifadesi, aynı anda hem *i* hem de *j* kategorisine ait belgelerin sayısını göstermekte idi, $c_j \square c_i$ ifadesi ise, aynı anda hem *j* hem de *i* kategorisine ait belgelerin sayısını göstereceğinden birbirine eşittir,

$c_i \square c_j = c_j \square c_i$ dir.

$i > j$ ise

$$\mu_{MC}(i,j) = \frac{c_i \square c_j}{\sum_{\substack{k=1 \\ k \neq j}}^{K} c_k \square c_j}$$

$$\mu_{MC}(j,i) = \frac{c_j \square c_i}{\sum_{\substack{k=1 \\ k \neq j}}^{K} c_j \square c_k} = \frac{c_i \square c_j}{\sum_{\substack{k=1 \\ k \neq j}}^{K} c_k \square c_j}$$

olduğundan dolayı

$\mu_{MC}(i,j) = \mu_{MC}(j,i)$ olur.

$i < j$ ise

$$\mu_{MC}(i,j) = \frac{c_i \square c_j}{\sum_{\substack{k=1 \\ k \neq i}}^{K} c_i \square c_k}$$

$$\mu_{MC}(j,i) = \frac{c_j \square c_i}{\sum_{\substack{k=1 \\ k \neq i}}^{K} c_k \square c_i} = \frac{c_i \square c_j}{\sum_{\substack{k=1 \\ k \neq i}}^{K} c_i \square c_k}$$

Bundan dolayı

$\mu_{MC}(i,j) = \mu_{MC}(j,i)$ olur.

Bu ise *MC* matrisinin simetrik olduğunu göstermektedir. Bu özelliğe dayanarak *MC* matrisi aşağıdaki şekilde kolayca hesaplanabilir.

$$\mu_{MC}(i,j) = \begin{cases} \dfrac{c_i \square c_j}{\sum_{\substack{k=1 \\ k \neq j}}^{K} c_k \square c_j} & , \quad if \quad i > j \\ 1 & , \quad if \quad i = j \\ \mu_{MC}(j,i) & , \quad if \quad i < j \end{cases} \qquad (6.11)$$

Daha sonra bu iki ilişki aşağıdaki formül yardımıyla birleştirilerek yeni çoklu belge-küme vektörü elde edilir.

$$Msim(d,c_j) = \frac{sim(d,c_j) + \left(sim(d,c_x) \cdot \mu_{MC}(x,j)\right)}{2} \qquad (6.12)$$

Burada *x ise sim*(d,c_j) değerini maksimum yapan *j* değeridir

Teorem 5.2. $sim(d,c_j)$ değerini maksimum yapan j değeri x ise, $Msim(d,c_j)$ değerini maksimum yapan j değeri de x dir.

$$\arg\max sim\left(d,c_j\right) = \arg\max Msim\left(d,c_j\right) \quad j = 1,2,\ldots,K$$

İspat : $sim\left(d,C\right) = \left[a_1,a_2,\ldots,a_t,\ldots,a_k\right]$ ve $sim\left(d,c_x\right) = a_t$ maksimum olsun.

Eğer $j = x$ ise

$$Msim\left(d,c_j\right) = \frac{sim(d,c_j) + \left(sim(d,c_x)\cdot \mu_{MC}(x,j)\right)}{2}$$

$\mu_{MC}(x,j) = 1$ ve $sim\left(d,c_x\right) = a_t$ olduğunda dolayı

$$Msim\left(d,c_j\right) = \frac{sim(d,c_j) + a_t}{2} = \frac{2a_t}{2} = a_t$$

Eğer $j \neq x$ ise

$sim\left(d,c_x\right) = a_t$ olduğundan

$$Msim\left(d,c_j\right) = \frac{sim(d,c_j) + sim(d,c_x)\cdot \mu_{MC}(x,j)}{2} = \frac{sim(d,c_j) + a_t\cdot \mu_{MC}(x,j)}{2}$$

olur. Ayrıca $sim\left(d,c_j\right) < a_t$ olduğu için

$$Msim\left(d,c_j\right) = \frac{sim(d,c_j) + a_t\cdot \mu_{MC}(x,j)}{2} < \frac{a_t + a_t\cdot \mu_{MC}(x,j)}{2}$$

$\mu_{MC}(x, j) < 1$ olduğundan $a_t \cdot \mu_{MC}(x, j) < a_t \cdot 1$ olacaktır ve son olarak

$$Msim\left(d, c_j\right) < \frac{a_t + a_t \cdot \mu_{MC}(x, j)}{2} < \frac{a_t + a_t}{2} = a_t$$

olur. Bu ise $sim\left(d, C\right)$ vektörünün maksimum elemanı, $Msim\left(d, C\right)$ vektörünün de maksimum elemanıdır. İşlem maksimum elemanı değiştirmemekte ancak diğer elemanları değiştirmektedir. Bu yüzden ait olunan ikinci kategori daha doğru bir şekilde tespit edilmektedir.

Birden fazla kategoriye ait bir belgenin ait olduğu kategoriler bu yöntemle belirlenirken bu yeni çoklu belge-kategori vektörü kullanılacaktır. Örneğin ait olunan kategori iki tane ise, bu kategoriler en büyük değere sahip ilk iki kategoridir.

5.6. Deneysel Sonuçlar ve Analizi

Bu çalışmada, belgelerin birden fazla kategoriye ait olma durumlarının da ele alındığı bir arama yöntemi geliştirmeye çalışılmaktadır. Bu bölümde öncelikle, deneysel çalışmada kullanılan belge koleksiyonu ve deneysel metodoloji açıklanmıştır. Daha sonra elde edilen sonuçların analizlerine yer verilmiştir.

5.6.1. Belge Koleksiyonu

Metin madenciliği araştırmalarında kullanılan metin koleksiyonlarının çoğu hiyerarşik bir yapıya sahiptir. Bunlara örnek The 20 Newsgroups data set (Ken Lang

tarafından düzenlenmiştir.), Industry Sector (Elkan, 2005), Cora dataset (McCallum ve ark., 2000) vs. verilebilir. Bu hiyerarşiden dolayı belgeler birden fazla kategoriye ait değillerdir. Üst seviye kategoriler ve alt seviye kategoriler vardır. Bu da belgelerin kategorilere üyeliklerinin kesin (crisp) bir özellik taşımasına yol açmaktadır. Bu yüzden bu tür metin koleksiyonları çoklu kategori problemi araştırmalarında kullanılmaya uygun değildir.

Bu çalışmada ise çoklu kategori problemi bulanık bir yaklaşım ile ele alınmaktadır. Bu yüzden hiyerarşik belge koleksiyonları bu çalışmanın yapısına uygun değildir. Burada kullanılan metin belge koleksiyonu ise metin madenciliği araştırmalarında sıkça kullanılan Reuters-21578 dağıtım 1.0'dır. Bu koleksiyonun özelliği kategorilerin hiyerarşik olmamasıdır. Koleksiyondaki bazı belgeler aynı anda birden fazla kategoriye ait olduğundan dolayı araştırmanın doğasına uygundur. Bu koleksiyon 135'in üzerinde konuya sahip 21578 belge içermektedir. Mevcut olan 135 konudan bazıları çok az sayıda belgede bulunmaktadır. Bu yüzden bu çalışmada kullanılmak üzere 135 konudan en çok yer alan 10 konu seçilmiştir. Seçilen bu konulara ait 8595 belge mevcut olup, bunlardan 6456 belge eğitim verisi olarak, 2139 belge ise test verisi olarak kullanılmıştır. Eğitim verisi olan belgelerden 622 adedi, test verisi olan belgelerden ise 221 adedi birden fazla konuya aittir.

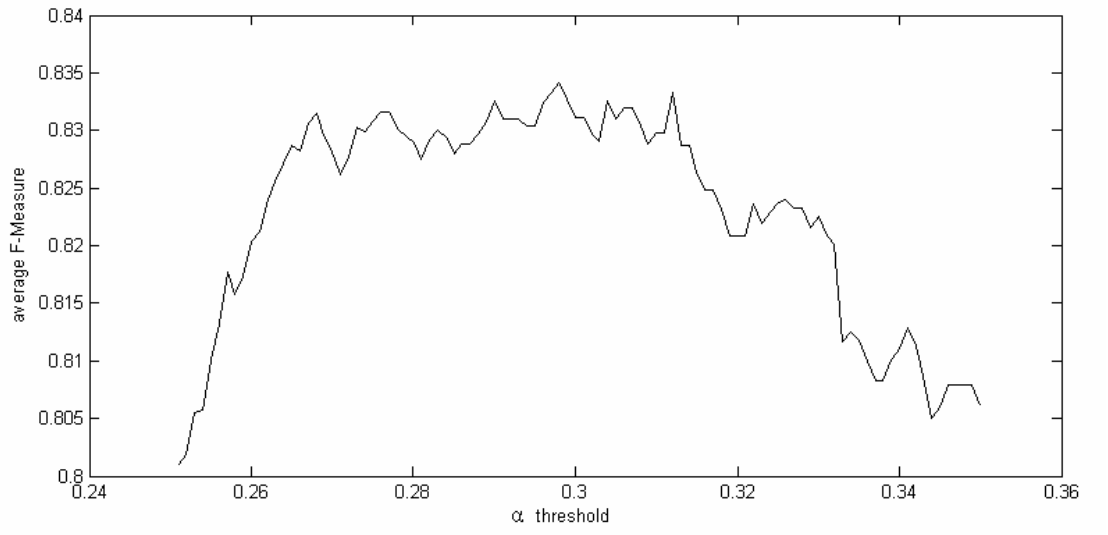
Seçilen eğitim verileri ön işlemeye tabi tutulmuştur. İlk olarak 350 kelimeden oluşan stop-words kelimeleri bu belgelerden çıkarılmıştır. Kelimelerin gövdeleme işlemi için, yaygın bir şekilde kullanılan Porter Stemmer algoritması seçilmiştir (Jones ve Willett, 1997). Bunun sonucunda belgeler içerdikleri kök kelimelere yani terimlere göre kümelenecektir.

5.6.2. α-FSCM Uygulaması

Önerilen yöntemde belirtildiği gibi ilk olarak test belgelerinin tekli-çoklu kategori sınıflarına üyelik dereceleri bulunmuştur. Bu değerlerden çoklu kategori

değeri tekli kategori değerine oranlanmıştır. Burada belgelerin sınıflandırılması için son işlem bir eşik değer belirlenmesidir. Önerilen yöntemdeki eşik değeri hesaplanmadan önce, deneysel olarak eşik değerinin sınıflandırma üzerindeki etkisi incelenmiştir. Bu Şekil 5.4'te görülmektedir. Burada farklı eşik değerleri için ortalama F-Ölçüsü değeri görülmektedir.

Daha sonra ise önerilen yöntemde (α-FSCM) belirtildiği gibi eşik değeri tespit edilmelidir. Hatırlanacağı üzere eğitim verisi 6456 belgeden oluşmaktadır. Bu belgelerin 622 si birden fazla kategoriye aittir. Önerilen yöntem kullanılarak bu eğitim belgelerinin iki sınıfa aitlik değerleri bulunmuştur. Bu değerlerden çoklu küme değeri tekli küme değerine oranlanmıştır ve elde edilen değere göre belgeler sıralanmıştır. Önerilen yönteme göre 623. belgenin oran değeri eşik olarak seçilmiştir. Bu eşik değeri 0.317 olarak tespit edilmiştir. Bu değer test verisinin sınıflandırması için kullanılacaktır.



Şekil 6.4. α eşik değerine göre Ortalama F-Ölçüsü grafiği

Tablo 5.1. Test verisinin eşik değeri sonuçları

α – eşik	Tekli Kategori			Çoklu Kategori			Ortalama F-Ölçüsü
	Recall	Precision	F-Ölçüsü	Recall	Precision	F-Ölçüsü	
0,291	0,95255	0,97181	0,96209	0,76018	0,64865	0,7	0,83104
0,292	0,95255	0,97181	0,96209	0,76018	0,64865	0,7	0,83104
0,293	0,95255	0,97181	0,96209	0,76018	0,64865	0,7	0,83104
0,294	0,95308	0,97131	0,96211	0,75566	0,64981	0,69874	0,83043
0,295	0,95308	0,97131	0,96211	0,75566	0,64981	0,69874	0,83043
0,296	0,95516	0,97085	0,96294	0,75113	0,65873	0,7019	0,83242
0,297	0,95568	0,97087	0,96322	0,75113	0,66135	0,70339	0,8333
0,298	0,9562	0,97088	0,96349	0,75113	0,664	0,70488	0,83419
0,299	0,9562	0,97037	0,96324	0,74661	0,66265	0,70213	0,83268
0,300	0,9562	0,96986	0,96298	0,74208	0,66129	0,69936	0,83117
0,301	0,9562	0,96986	0,96298	0,74208	0,66129	0,69936	0,83117
0,302	0,9562	0,96934	0,96273	0,73756	0,65992	0,69658	0,82966
0,303	0,95673	0,96885	0,96275	0,73303	0,66122	0,69528	0,82901
0,304	0,95881	0,96891	0,96384	0,73303	0,6722	0,7013	0,83257
0,305	0,95881	0,9684	0,96358	0,72851	0,67083	0,69848	0,83103
0,306	0,95933	0,96842	0,96386	0,72851	0,67364	0,7	0,83193
0,307	0,95933	0,96842	0,96386	0,72851	0,67364	0,7	0,83193
0,308	0,95933	0,96791	0,9636	0,72398	0,67227	0,69717	0,83039
0,309	0,95933	0,9674	0,96335	0,71946	0,67089	0,69432	0,82884
0,310	0,95985	0,96742	0,96362	0,71946	0,67373	0,69584	0,82973
0,311	0,95985	0,96742	0,96362	0,71946	0,67373	0,69584	0,82973
0,312	0,96194	0,96749	0,96471	0,71946	0,68534	0,70199	0,83335
0,313	0,96194	0,96597	0,96395	0,70588	0,68122	0,69333	0,82864
0,314	0,96194	0,96597	0,96395	0,70588	0,68122	0,69333	0,82864
0,315	0,96246	0,96498	0,96372	0,69683	0,68142	0,68904	0,82638
0,316	0,96246	0,96447	0,96347	0,69231	0,68	0,6861	0,82478
0,317	0,96246	0,96447	0,96347	0,69231	0,68	0,6861	0,82478
0,318	0,96246	0,96397	0,96321	0,68778	0,67857	0,68315	0,82318
0,319	0,96298	0,96298	0,96298	0,67873	0,67873	0,67873	0,82086
0,320	0,96298	0,96298	0,96298	0,67873	0,67873	0,67873	0,82086

Test verisi üzerinde yapılan eşik değeri ile ilgili deneysel sonuçlardan bazıları Tablo 5.1'de verilmiştir. Burada eşik değeri için 0,291-0,320 değer aralığında elde edilen sınıflandırma sonuçları yer almaktadır. Tablo 5.1'deki değerlerden de anlaşılacağı üzere eşik değerinin tespitinde önerilen yöntem büyük ölçüde deneysel sonuçlarla uyuşmaktadır. Önerilen yöntem ile bulunan eşik değeri 0,317'dir ve bu değer ile elde edilen ortalama F-Ölçüsü 0,82478'dir. Deneysel sonuçlarla ulaşılan en iyi eşik değeri 0.298'dir ve bu eşik ile bulunan ortalama F-Ölçüsü değeri ise 0,83419 olmaktadır.

Deneysel en iyi sonuç ile önerilen yöntemden elde edilen sonuç arasındaki farklılık için şu yorumu yapmak mümkündür. Eğitim verisi ile test verisi incelendiğinde çoklu kümeye ait belgelerin tüm belgelere oranları arasında bir fark göze çarpmaktadır. Eğitim verisinde 6456 belgeden 622 adedi birden fazla kategoriye aittir. 2139 test belgesinden ise 221 adedi birden fazla kategoriye aittir. Eğitim verisinin %9,6 sı çoklu kategoriye sahip iken, test verisinin %10,3 ü çoklu kategoriye sahiptir. Eğitim ve test verisinin çoklu küme oranları ne kadar birbirine yakın ise sonucun o kadar daha iyi olacağı açıktır. Nitekim bu Tablo 5.2'te görülmektedir.

Tablo 5.2'deki "benzer dağılımda" satırı için eşik değeri bulunurken eğitim verisinin de çoklu kategori oranı %10,3 olarak kabul edilmiştir. Bu sonuçlar yukarıdaki iddiayı doğrular niteliktedir.

Tablo 5.2. Örnek Eşik ve ortalama F Ölçüsü değerleri

	α – eşik	**Ortalama F Ölçüsü**
Deneysel en iyi	0,298	0,834
Benzer dağılımda	0,306	0,832
Önerilen yöntem ile	0,317	0,825

Tablo 5.3. Sınıflandırma Performans Karşılaştırması

Yöntem	Tekli Kategori			Çoklu Kategori			Ortalama F-Ölçüsü
	Recall	Precision	F-Ölçüsü	Recall	Precision	F-Ölçüsü	
Rocchio	0,708	0,897	0,791	0,299	0,105	0,155	0,473
Naive Bayes	0,511	0,945	0,663	0,742	0,149	0,248	0,455
α-FSCM	0,962	0,964	0,963	0,692	0,680	0,686	0,824

Son olarak önerilen α-FSCM ile metin madenciliğinde sıkça kullanılan sınıflandırma yöntemlerinden olan Rocchio Algoritması ve naive Bayes yöntemi karşılaştırılmıştır. Bu iki yöntemin tekli-çoklu kategori sınıflandırmasında oldukça yetersiz kaldığı Tablo 5.3'de görülmektedir. Yine bu tabloda, önerilen yöntemin hem Rocchio algoritması hem de naive Bayes yöntemine oranla oldukça başarılı olduğu görülmektedir.

5.6.3. MCVM Uygulaması

Bir sonraki adımda incelediğimiz konu ise çoklu kategoriye ait belgelerin ait oldukları kategorilerin tespitidir. Ancak bu test kısımda çoklu kategori iki adet kategori olarak kabul edilmiştir. Yani bir belge çoklu kategoriye sahip ise bu belgenin ait olduğu kategori sayısının 2 olduğu kabul edilmiştir. Bunların hangileri olduğunun tespiti amaçlanmıştır. Bunun için uygulanan çoklu kategori vektörü yönteminden elde edilen sonuçlar Tablo 5.4'te gösterilmiştir.

Tablo 5.4 oluşturulurken iki farklı eşik değeri temel alınmıştır. Bu değerlere göre önerilen yöntem ile klasik yöntem karşılaştırılmıştır. Bu eşik değerlerinden ilki, α eşik bulanık benzerlik sınıflandırma yönteminden elde ettiğimiz eşik değeridir. İkinci eşik değeri ise yine aynı yöntem ile test verisinin en iyi sınıflandırıldığı (deneysel olarak bulunan) eşik değeridir.

Tablo 5.4. Seçilen α eşik değerlerine göre elde edilen kategori tespiti değerleri

	***α* eşik**	**Toplam**	**CR**	**CM**	**MCVM**	**Artış (%)**
Önerilen yöntem ile	0,317	221	153	29	81	179,31
En iyi sınıflandırma	0,298	221	166	32	89	178,13

Tablo 5.4'de yer alan sütunlarının açıklamaları ise şudur:

Toplam: Test belge koleksiyonundaki birden fazla kategoriye ait olan belgelerin toplam sayısıdır.

CR (Classification Results): Test koleksiyonundaki belgelerden, önerilen *α eşik bulanık benzerlik sınıflandırması yöntemi* uygulanarak doğru olarak tespit edilen belgelerin toplam sayısıdır.

CM (Classical Method): Klasik yöntem ile olduğu kategoriler doğru olarak bulunan belgelerin toplam sayısıdır.

MCVM (Multiple Categories Vector Method): Önerilen çoklu kategori vektörü yöntemi ile ait olduğu kategoriler doğru olarak bulunan belgelerin toplam sayısıdır.

Artış: Klasik yöntemle önerilen yöntemin sonuçları arasındaki artışın yüzdelik değeridir.

Tablo 5.4'te de görüldüğü üzere, ait olunan kategorilerin doğru tespitinde, önerilen yöntem ile klasik yöntem arasında iki kata yaklaşan bir başarı artışı olmuştur.

Belgelerin birden fazla kategoriye aitlik durumunun göz önünde tutulması benzer belge arama sistemleri için önemlidir. Bu önem Tablo 5.5 ve Tablo 5.6 da verilen örnekte açıkça görülmektedir. Burada *A* kategorisi Reuter 21578 belge koleksiyonundaki "money-fx" kategorisini, *B* ise "interest" kategorisini göstermektedir. Örnek olarak bu kategorilere ait 12 belge eğitim verisi içerisinden seçilmiştir. Seçilen bu belgelerde 4'ü sadece *A* kategorisine, 4'ü sadece *B* kategorisine, geriye kalan 4 tanesi ise aynı anda hem A hem de B kategorisine aittir. Eğitim belgeleri içinden seçilen bu belgeler test verisi içinden örnek olarak seçilen belgelerle karşılaştırılmıştır. Burada test verisinden seçilen iki örnek belge için elde

edilen sonuçlar iki ayrı Tablo içerisinde görülmektedir. (9138 numaralı eğitim belgesi için Tablo 5.5, 6452 numaralı test belgesi için Tablo 5.6). Bu iki test verisi belgesi hem A hem de B kategorisine aitlerdir. Aralarındaki fark ise ait olunan kategorilerin öncelikleridir.

Çoklu kategori tespitinin faydası burada açıkça görülmektedir. Örneğin Tablo 5.5 de görüldüğü üzere, eğer 9138 numaralı test belgesi önceki çalışmalarda olduğu gibi sadece B kategorisine ait kabul edilse idi aday belge olarak ya sadece B kümesine ait olanlar veya hem A hem de B kategorisine ait olan belgeler seçilecekti. Hâlbuki sadece A kategorisine ait olup elimizdeki test belgesine büyük oranda benzeyen belgeler göz ardı edilmiş olacaktı (örneğin 21561, 21539). Bu ise bir belgenin birden fazla kategoriye ait olduğu durumlarda tüm bu kategorilere ait belgelerin aday olarak seçilmesi gerektiğini göstermektedir.

Tablo 5.5. Test belgesi ile eğitim belgeleri arasındaki benzerlik

Test belgesinin numarası	**Test belgesinin ilk kategorisi**	**Test belgesinin ikinci kategorisi**	**Eğitim belgesi numarası**	**Benzerlik**	**Eğitim belgesinin kategorisi(leri)**
9138	B	A	15364	0,52275	A-B
			20500	0,47769	A-B
			21561	0,44498	A
			19557	0,43490	B
			19201	0,42534	A-B
			21539	0,42131	A
			19512	0,42084	B
			19061	0,41997	A
			21343	0,41134	A
			19237	0,39752	A-B
			21285	0,36056	B
			19529	0,36037	B

Tablo 5.6. Test belgesi ile eğitim belgeleri arasındaki benzerlik

Test belgesinin numarası	Test belgesinin ilk kategorisi	Test belgesinin ikinci kategorisi	Eğitim belgesi numarası	Benzerlik	Eğitim belgesinin kategorisi(leri)
6452	A	B	19201	0,42279	A-B
			21343	0,42051	A
			19237	0,41775	A-B
			19512	0,41582	B
			21285	0,40890	B
			21508	0,39984	A-B
			20500	0,39711	A-B
			19529	0,38071	B
			15364	0,37444	A-B
			21539	0,36893	A
			21556	0,35933	A
			19557	0,34262	B
			21561	0,34166	A

Özet olarak, yeni yöntem ile benzerlik araştırmasında aday belgelerin sayısı artmış olmaktadır. Bu yüzden benzerlik oranı yüksek olduğu halde tekli kategori yaklaşımında göz ardı edilen belgeler bu yöntem ile ele alınabilecektir.

5.7. Bölüm Sonuçları

Bu çalışmaya konu olan benzer belge aranmasında çoklu kategori problemi önceki çalışmalarda fazla ele alınmamış bir konudur. Problemin çözümü, çoklu kategorili belgelerin tespiti ve bu belgelerin ait oldukları kategorilerin tespiti olmak üzere iki aşamada incelenmiştir. İlk aşamada, birden fazla kategoriye aitlik söz

konusu olduğundan dolayı çoklu kategorili belgelerin tespitinde bulanık mantık tabanlı bir yaklaşım benimsenmiştir. Metin sınıflandırmasında önemli bir yeri olan FSC yöntemi probleme adapte olacak şekilde geliştirilmiştir. Önerilen yeni yöntem (α-FSCM) IR ve metin madenciliği uygulamalarında sıkça kullanılan Rocchio algoritması ve naive Bayes yöntemi ile karşılaştırıldığında önemli ölçüde bir başarı göstermiştir. İkinci aşamayı oluşturan, çoklu kategorili belgelerin ait oldukları kategorilerin tespitinde ise kategorilerin birliktelik sıklığı bilgisinden faydalanmıştır. Bu aşama için önerilen MCVM, bu konu ile ilgili klasik yaklaşıma göre oldukça başarılıdır (Saraçoğlu ve ark., 2008).

Bu çalışma ile benzer belge aranmasında çoklu kategori probleminin önemi görülmüştür. Test belgelerinin sadece tek bir kategoriye ait oldukları kabul edilen önceki yaklaşımlar önemli bir miktardaki aday belgesi göz ardı etmektedir. Önerilen benzerlik arama yaklaşımı bunu gidermeyi amaçlamış ve önemli bir ölçüde de başarmıştır.

Önerilen benzerlik aranması çok aşamalı bir yapıya sahiptir. Bu yüzden her bir aşamasının ayrıca geliştirilmesi mümkün olabilir. İlk aşamasında α parametresinin tespiti buna örnek olarak verilebilir. Yine problemin ikinci aşamasında çoklu kategori matrisi oluşturulmasında farklı parametreler kullanılarak performans artışı araştırılabilir.

6. ÖRNEK BİR BENZER BELGE ARAMA UYGULAMASI

Bu kısımda, artık hayatımızda önemli bir yer tutan arama mekanizmalarına farklı bir yaklaşım ortaya koyan bulanık kümeleme kullanılarak benzer belge aranması yönteminin bir uygulamasına yer verilmiştir. Bu yöntemde, metinden manüel olarak seçilen anahtar kelimelere dayalı bir aramaya alternatif olarak metnin tüm kelimelerinin kullanıldığı bir arama mekanizması ortaya konmaktadır. Bu sayede bir metnin benzerinin bulunması için anahtar kelime seçimi ve manüel olan bu seçim işlemde ortaya çıkabilecek problemlerin ortadan kaldırılması amaçlanmaktadır. Anahtar kelime tabanlı aramaya oranla daha karmaşık bir yapıya sahip olmasına karşı daha etkili bir biçimde benzerlik bulunması sağlanabilecektir.

6.1. Veri Kümesi

Bu uygulama kısmında kullanılan veri kümesi, ScienceDirect bilimsel yayın veritabanı tarafından taranan 2007 yılında yayınlanmış makale özetlerinden seçilmiştir. Örnek bir makale özeti Şekil 7.1'de görülmektedir.

Bu çalışmada uygulama amacı ile kullanılan özetler, 4 ayrı konudan seçilmiş toplam 200 özetten oluşmaktadır. Seçilen konular ve bu konulara göre belge dağılımı Tablo 7.1'de görülmektedir.

Bu özetlerden 160 adedi eğitim verisi olarak kullanılmış geriye kalan ise test verisi olarak kullanılmıştır. Eğitim verisinden kasıt, üzerinde arama yapılacak verilerdir.

<Journal> Applied Soft Computing

<Title> Word segmentation of handwritten text using supervised classification techniques

<Author> Yi Sun, Timothy S. Butler, Alex Shafarenko, Rod Adams, Martin Loomes and Neil Davey

<Abstract> Recent work on extracting features of gaps in handwritten text allows a classification of these gaps into inter-word and intra-word classes using suitable classification techniques. In this paper, we first analyze the features of the gaps using mutual information. We then investigate the underlying data distribution by using visualization methods. These suggest that a complicated structure exists, which makes them difficult to be separated into two distinct classes. We apply five different supervised classification algorithms from the machine learning field on both the original dataset and a dataset with the best features selected using mutual information. Moreover, we improve the classification result with the aid of a set of feature variables of strokes preceding and following each gap. The classifiers are compared by employing McNemar's test. We find that SVMs and MLPs outperform the other classifiers and that preprocessing to select features works well. The best classification result attained suggests that the technique we employ is particularly suitable for digital ink manipulation at the level of words.

<Keywords> Handwriting; Supervised classification; Mutual information; McNemar's test

Şekil 7.1. Örnek bir makale özeti

Tablo 7.1. ScienceDirect veri tabanından seçilen veri kümesi

Konu	**Belge No Aralığı**	**Eğitim Veri Sayısı**	**Test Veri Sayısı**	**Toplam Veri Sayısı**
Bilgisayar Bilimleri	1001-1050	40	10	50
Ekonomi	2001-2050	40	10	50
Ziraat	3001-3050	40	10	50
Sosyal ve Beşeri Bilimler	4001-4050	40	10	50
Toplam		160	40	200

Eğitim verileri üzerinde arama mekanizması oluşturulmuştur. Daha sonra test verisi olarak ayrılan özetlere benzer olan belgeler aranacaktır. Bu arama işlemi eğitim verileri üzerinden olmaktadır.

6.2. Arama Mekanizmasının Oluşturulması

Bu uygulama da temel olarak iki arama yaklaşımının kıyaslanması amaçlanmıştır. Bunlardan ilki klasik *anahtar kelime tabanlı* aramadır. Bu yaklaşımda belgeler için manüel olarak belirlenen anahtar kelimeler kullanılarak benzerlik tespiti yapılmaktadır. İkinci yaklaşım ise belgedeki tüm kelimelerin ele alındığı, *bulanık kümeleme kullanılarak benzer belge aranması* yaklaşımıdır.

6.2.1. Anahtar kelime yaklaşımı

Bu yaklaşımın başarısını belirleyen en önemli özellik anahtar kelime seçimidir. Manüel olarak yapılan bu seçim belgeyi en iyi biçimde tespit edecek kelime veya kelime gruplarının seçilmesidir. Bu veri kümesinde kullanılan veriler bilimsel makale özetleri olduğu için bu makalelerin zaten mevcut olan anahtar kelime kısımları bu uygulamada aynen kullanılmıştır. Bu anahtar kelimeler veri tabanına kaydedilmiştir. Benzeri aranan makale özetinin anahtar kelimeleri bu veri tabanında sorgulanarak bir benzerlik sıralaması yapılmıştır. Bu sıralama da doğal olarak en çok anahtar kelimesi eşleşen özetten en az eşleşen özete doğru yapılmıştır.

Daha sonra bu arama genişletilerek tüm özet metni üzerine yayılmıştır. Yani benzeri aranan makalenin anahtar kelimeleri, aday makalelerin özet metinlerinin

tümü üzerinde aranmıştır. Bu sonuçlara göre sıralama yapılmıştır. Benzerlik derecesini belirlemek için ise karşılaştırmaya giren anahtar kelime sayısı ve eşleşen/içerilen anahtar kelime sayısı arasındaki ilişki temel alınmıştır. Örneğin benzeri aranan makale özetine ait 5 anahtar kelime varsa ve bu 5 kelimeden 3 tanesi aday belgede (veya anahtar kelime kısmında) geçiyorsa benzerlik %60 olarak ifade edilmektedir.

6.2.2. Bulanık kümeleme kullanılarak benzer belge aranması yaklaşımı

Arama işlemine bütün kelimelerin dâhil edildiği, bulanık kümeleme kullanılarak benzer belge aranması için arama mekanizması oluşturulurken ilk adım önişlemedir. Bu safhada öncelikle belgeler kelimelere parçalanmaktadır. Daha sonra bu kelimelerin içerisinden anlam için önemsiz olanları (stopwords) çıkartılmaktadır. Bir sonraki işlem olarak ise bu kelimeler gövdeleme işlemine tabi tutulmaktadır. Kelimelerin bu son şekli "Terim" olarak da adlandırılmaktadır. Son olarak terim tekrarları sayılmaktadır.

Bu uygulamada 350 kelimelik bir stopwords listesi kullanılmıştır. Gövdeleme yöntemi için ise Porter Stemmer algoritması kullanılmıştır. Bu işlemler sonucunda koleksiyondaki birbirinden farklı kelime (terim) sayısı 2804 olarak tespit edilmiştir. Buradan da açıkça anlaşılacağı üzere bu belge koleksiyonundaki her belge 2804 adet sayı ile ifade edilecektir. Bu bir vektör olarak düşünüldüğünde ise birçok elemanı "0" değerinden oluşacaktır. Çünkü yüzlerle ifade edilebilen kelime sayısına sahip olan bir belge (bu uygulamada 200-300 kelimeden oluşan bir bilimsel makale özeti) ancak binlerle ifade edilebilen kelime sayısına sahip bir sözlük üzerinden (bu uygulamada 2804) ifade edilebilir. Bu ise belgenin matematiksel ifadesi anlamına gelen özellik vektörünün büyük oranda boş olduğunu göstermektedir.

Tablo 7.2. Sınıflandırma sonuçları

Konular	Precision	Recall	F-Ölçüsü
Bilgisayar Bilimleri	1	1	1
Ekonomi	1	1	1
Ziraat	1	1	1
Sosyal ve Beşeri Bilimler	1	1	1

Bir sonraki safha ise bu aşırı büyük olan boyutu azaltmayı amaçlayan (bir başka değişle özellik çıkarımı) sınıflandırma/kümeleme safhasıdır. Bu safha da öncelikle bir bulanık sınıflandırma uygulanmaktadır. Bunun sonucunda elde edilen her bir sınıfa üyelik değerleri, özellik vektörü olarak kullanılmaktadır. Sınıflandırma işleminde kullanılan yöntem FSC yöntemidir. Eğitim ve test verileri için bu sınıflandırma işleminden elde edilen sonuçlar Tablo 7.2'de verilmiştir. Buradan da anlaşılacağı üzere eğitim ve test verileri için %100 bir sınıflandırma değerine ulaşılmıştır.

Eğitim verilerinin (üzerinde benzerlik araması yapılan belgelerin) her biri için 4 sınıfa aitlik değerleri belirlendikten sonra bu değerler özellik vektörü olarak kullanılmaktadır. Benzeri aranacak bir belge (test belgesi) yine bu sınıflandırma içlemlerine tabi tutularak 4 sınıf için aitlik dereceleri belirlenir ve bu değerler özellik vektörünü ifade eder. Bu yapılan işleme, benzeri aranan belge için bir özellik çıkarımı olarak bakılabilir. Bu özellik çıkarımından sonra benzeri aranan bu belge ile (sisteme önceden girilmiş olan) mevcut aday belgeler benzerlik ölçümüne tabi tutulmaktadır. Bunun sonucunda mevcut belgelerin her birinin test belgesine olan benzerlikleri tespit edilmiş olmaktadır.

6.3. Arama İşleminin Gerçekleştirilmesi

Yukarıda yapıları ile ilgili ayrıntılı bilgiler verilen iki yaklaşımın test edilme aşaması, yeni bir belgenin benzerinin aranması anlamına gelmektedir.

Yeni bir belgenin benzerleri (bu uygulama için belge, bilimsel makale özeti anlamına gelmektedir) anahtar kelime yaklaşımı ile aranırken sadece anahtar kelimeler (bu uygulama için makalenin anahtar kelime kısmı anlamına gelmektedir) kullanılır. Bu arama işlemi ise iki farklı şekilde yapılmaktadır. Bunlardan ilki önceden kayıtlı belgelerin sadece anahtar kelime alanları üzerinden bir karşılaştırma yapılır. İkinci olarak ise tüm metin üzerinde bir karşılaştırma yapılır. Bulunan anahtar kelime sayısının aranan anahtar kelime sayısına oranı yardımıyla sonuç tespit edilmiş olur.

Yeni bir belgenin benzerleri bulanık kümeleme kullanılarak aranırken ise tüm belge metni kullanılmaktadır. Yeni belge önişleme ve sınıflandırma/kümeleme aşamalarından geçtikten sonra özellik vektörü çıkartılmaktadır. Daha sonra mevcut belgeler ile benzerlik tespiti yapılmaktadır. En yüksek benzerlik değerine sahip belgelerden istenilen adedi listelenmektedir. Bu uygulamada kullanılan benzerlik ölçümü beşinci bölümde önerilen boyut kök benzerliğidir.

Bu uygulamadaki benzerlik aramaları ile ilgili örnekler ayrıntılı bir biçimde ek kısmında verilmiştir.

6.4. Örnek Arama Uygulamaları

Bu kısımda hazırlanan arama sistemi üzerinde yapılmış bazı arama uygulamalarına yer verilmiştir. Bu örnek uygulamalar, dört ayrı konudan seçilen ikişer belgenin benzerlerinin aranması şeklindedir.

İlk iki örnek "Bilgisayar Bilimleri" konusuna ait belgelerden seçilen benzerlik arama sonuçlarıdır. Seçilen bu belgeler (1005 ve 1045 numaralı belgeler) için benzer belge araması sonuçları Tablo 7.3 ve Tablo 7.4'te görülmektedir. Tablo 7.3'te anahtar kelime yaklaşımı ile bulunan belgelerden 2031 ve 4002 numaralı belgeler, farklı bir konuya ait olduklarından dolayı (sırasıyla "Ekonomi" ve "Sosyal ve Beşeri Bilimler" konuları) Bulanık Kümeleme yaklaşımı ile bulunamamıştır. Bu ise birden fazla kategoriye aitlik durumunun ele alınması ihtiyacını ortaya koymaktadır.

Tablo 7.3. 1005 numaralı belge için arama sonuçları

Benzeri Aranan Belge no	Anahtar Kelime kullanılarak bulunan		Bulanık kümeleme kullanılarak bulunan	
	Belge no	Benzerlik oranı	Belge no	Benzerlik değeri
1005	1017	%25	1024	0.743
	1037	%25	1007	0.733
	2031	%25	1017	0.726
	4002	%25	1003	0.724
			1016	0.721
			1032	0.717
			1034	0.706
			1031	0.704
			1038	0.703
			1013	0.703

Tablo 7.4. 1045 numaralı belge için arama sonuçları

Benzeri Aranan Belge no	Anahtar Kelime kullanılarak bulunan		Bulanık kümeleme kullanılarak bulunan	
	Belge no	Benzerlik oranı	Belge no	Benzerlik Değeri
1045	1049	%25	1049	0.705
	3006	%25	1032	0.692
			1013	0.690
			1038	0.687
			1047	0.682
			1019	0.680
			1008	0.675
			1031	0.674
			1034	0.674
			1003	0.674

Üçüncü ve dördüncü örneklerde ise Ekonomi alanından seçilen iki belgenin (2035 ve 2050 numaralı belgelerin) benzerinin aranması ile ilgili sonuçlara yer verilmiştir. Bu sonuçlar Tablo 7.5 ve Tablo 7.6'da gösterilmiştir. Tablo 7.5'te anahtar kelime yaklaşımı ile bulunan 2046 numaralı belge, Bulanık Kümeleme yaklaşımı ile bulunamamaktadır. Bu durum, belgenin içerdiği tüm kelimelerin aramaya dahil edildiğinde anahtar kelime yaklaşımdan farklı sonuçlar ortaya çıkabildiğini göstermektedir.

Tablo 7.5. 2035 numaralı belge için arama sonuçları

Benzeri Aranan Belge no	Anahtar Kelime kullanılarak bulunan		Bulanık kümeleme kullanılarak bulunan	
	Belge no	Benzerlik oranı	Belge no	Benzerlik Değeri
2035	2046	%20	2038	0.700
			2044	0.693
			2043	0.690
			2013	0.689
			2047	0.687
			2034	0.674
			2033	0.673
			2016	0.670
			2036	0.669
			2029	0.668

Tablo 7.6. 2050 numaralı belge için arama sonuçları

Benzeri Aranan Belge no	Anahtar Kelime kullanılarak bulunan		Bulanık kümeleme kullanılarak bulunan	
	Belge no	Benzerlik oranı	Belge no	Benzerlik Değeri
2050	2009	%33	2009	0.771
	2008	%33	2036	0.755
	3029	%33	2047	0.751
			2012	0.726
			2037	0.710
			2016	0.707
			2043	0.704
			2021	0.700
			2013	0.699
			2008	0.690

Beş ve altıncı örneklerde benzeri aranan belgeler ise Ziraat konusuna ait belgelerdir. 3015 ve 3025 numaralı bu belgeler için arama sonuçları Tablo 7.7 ve Tablo 7.8'de görülmektedir. Tablo 7.8'den anlaşılacağı üzere, anahtar kelime tabanlı yaklaşımla 3025 numaralı belgeye benzer belge bulunamamıştır. Bütün kelimelerin aramaya dâhil edildiği bulanık kümeleme yaklaşımında bulunan belgelerden ilk 10'u listelenmiştir.

Tablo 7.7. 3015 numaralı belge için arama sonuçları

Benzeri Aranan Belge no	Anahtar Kelime kullanılarak bulunan		Bulanık kümeleme kullanılarak bulunan	
	Belge no	Benzerlik oranı	Belge no	Benzerlik Değeri
3015	3008	%20	3039	0.788
	3009	%20	3008	0.766
	3038	%20	3023	0.760
			3019	0.741
			3038	0.732
			3022	0.726
			3018	0.723
			3049	0.716
			3011	0.714
			3031	0.708

Tablo 7.8. 3025 numaralı belge için arama sonuçları

Benzeri Aranan Belge no	Anahtar Kelime kullanılarak bulunan		Bulanık kümeleme kullanılarak bulunan	
	Belge no	Benzerlik oranı	Belge no	Benzerlik Değeri
3025	Bulunamamıştır		3039	0.634
			3031	0.613
			3032	0.612
			3002	0.608
			3044	0.605
			3023	0.600
			3028	0.595
			3029	0.595
			3014	0.593
			3004	0.589

Son olarak yedinci ve sekizinci örnekler ise Sosyal ve Beşeri Bilimler konusuna ait belgelerden seçilmişlerdir. Bu belgelere (4010 ve 4020 numaralı belgelere) ait arama sonuçlarına ise Tablo 7.9 ve Tablo 7.10'da yer verilmiştir.

Tablo 7.9. 4010 numaralı belge için arama sonuçları

Benzeri Aranan Belge no	Anahtar Kelime kullanılarak bulunan		Bulanık kümeleme kullanılarak bulunan	
	Belge no	Benzerlik oranı	Belge no	Benzerlik Değeri
4010	Bulunamamıştır		4026	0.767
			4012	0.764
			4006	0.763
			4013	0.733
			4028	0.733
			4044	0.732
			4004	0.727
			4036	0.715
			4017	0.712
			4008	0.704

Tablo 7.10. 4020 numaralı belge için arama sonuçları

Benzeri Aranan Belge no	Anahtar Kelime kullanılarak bulunan		Bulanık kümeleme kullanılarak bulunan	
	Belge no	Benzerlik oranı	Belge no	Benzerlik Değeri
4020	4019	%50	4019	0.843
	3007	%25	4006	0.835
	3031	%25	4004	0.832
	2003	%25	4021	0.822
	4021	%25	4026	0.819
	4022	%25	4008	0.816
	4024	%25	4036	0.808
			4032	0.805
			4024	0.793
			4043	0.792

6.5. Bölüm Sonuçları

Bu bölümde bulanık kümeleme kullanılarak benzer belge aranması klasik anahtar kelime tabanlı arama ile karşılaştırılmıştır. Uygulama alanı olarak bilimsel makaleler seçilmiştir. Bunun en önemli sebebi ise bu belgelerin içinde yazarı tarafından manüel olarak belirlenmiş anahtar kelimeler kısmının bulunmasıdır. Tez çalışmasında önerilen arama yaklaşımın en önemli amaçlarından birisinin anahtar kelime seçiminden ortaya çıkan problemlerin aşılması olduğu düşünüldüğünde bu belgelerin seçilmesinin çok uygun olduğu görülmektedir.

Bu yeni arama yaklaşımı sayesinde hatalı veya eksik anahtar kelime seçiminden kaynaklanan problemler en aza indirilebilecektir. Çünkü önerilen yöntemde belgedeki kelimelerin tümü kullanılmaktadır. Yapılan uygulamalar da (örnek arama işlemleri) bu durumu doğrular niteliktedir.

Tablo 7.11-14'te sırasıyla Bilgisayar Bilimleri, Ekonomi, Ziraat ve Sosyal ve Beşeri Bilimler ile ilgili toplu test sonuçlarına yer verilmiştir. Bu tablolardaki "Arama Türü" sütününda geçen bazı kısaltmalar ve anlamları aşağıda açıklandığı gibidir:

AKY1 (Anahtar Kelime Yaklaşımı-1): Benzeri aranan belgenin anahtar kelimeleri ile aday belgelerin sadece anahtar kelimeleri karşılaştırılarak elde edilen sonuçlardır.

AKY2 (Anahtar Kelime Yaklaşımı-2): Benzeri aranan belgenin anahtar kelimeleri ile aday belgenin tüm metni karşılaştırılarak elde edilen sonuçlardır. Yani aday belgenin tüm metni (tüm kelimeleri) anahtar kelime olarak kabul edilerek yapılan karşılaştırma sonuçlarıdır.

BBA (Bulanık Kümeleme Yaklaşımı): Bulanık kümeleme kullanılarak yapılan benzer belge araması (önerilen arama yaklaşımı) sonucunda elde edilen sonuçlardır.

Tablo 7.11. Bilgisayar Bilimleri alanı için toplu test sonuçları

Aranan Belge	Arama Türü	Bulunan Belgeler									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1005	AKY1	1017	1037								
	AKY2	1017	1037	2031	4002						
	BKY	1024	1007	1017	1003	1016	1032	1034	1031	1038	1013
1010	AKY1	1008									
	AKY2	1008	1043								
	BKY	1008	1034	1024	1019	1007	1032	1023	1037	1038	1031
1015	AKY1	Bulunamamıştır									
	AKY2	4017									
	BKY	1008	1034	1019	1032	1047	1023	1013	1038	1024	1007
1020	AKY1	Bulunamamıştır									
	AKY2	Bulunamamıştır									
	BKY	1008	1032	1024	1007	1034	1023	1038	1019	1031	1047
1025	AKY1	Bulunamamıştır									
	AKY2	Bulunamamıştır									
	BKY	1024	1032	1007	1008	1034	1038	1016	1013	1019	1031
1030	AKY1	Bulunamamıştır									
	AKY2	Bulunamamıştır									
	BKY	1003	1032	1013	1038	1007	1024	1016	1006	1031	1017
1035	AKY1	Bulunamamıştır									
	AKY2	Bulunamamıştır									
	BKY	1032	1038	1008	1023	1047	1007	1013	1031	1024	1019
1040	AKY1	1019									
	AKY2	1019									
	BKY	1019	1047	1023	1008	1032	1038	1013	1031	1034	1049
1045	AKY1	1049	3006								
	AKY2	1049	3006								
	BKY	1049	1032	1013	1038	1047	1019	1008	1031	1034	1003
1050	AKY1	Bulunamamıştır									
	AKY2	Bulunamamıştır									
	BKY	1024	1007	1032	1008	1038	1031	1016	1034	1023	1013

Bu tablolardaki sonuçlar ile ilgili aşağıdaki yorumlar yapılabilir:

1. Anahtar kelime yaklaşımı ve bulanık kümeleme yaklaşımları büyük oranda birbirleri ile tutarlı sonuçlar vermişlerdir. Anahtar kelimeye göre bulunan en benzer belgeler bulanık kümeleme yaklaşımı ile elde edilen belgelerin ilk beşi içerisinde genellikle yer almaktadır. Buna örnek olarak 1005, 1010, 1040 ve 1045 numaralı belgeler verilebilir. Diğer belgeler için anahtar kelime yaklaşımı ile herhangi bir benzer belge bulunamamıştır.

Tablo 7.12. Ekonomi alanı için toplu test sonuçları

Aranan Belge	Arama Türü	Bulunan Belgeler									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
2005	AKY1	Bulunamamıştır									
	AKY2	Bulunamamıştır									
	BKY	2036	2037	2012	2047	2042	2021	2009	2043	2008	2019
2010	AKY1	Bulunamamıştır									
	AKY2	Bulunamamıştır									
	BKY	2047	2043	2016	2027	2036	2017	2029	2008	2026	2013
2015	AKY1	Bulunamamıştır									
	AKY2	Bulunamamıştır									
	BKY	2047	2036	2043	2012	2042	2021	2016	2038	2013	2018
2020	AKY1	Bulunamamıştır									
	AKY2	Bulunamamıştır									
	BKY	2012	2042	2021	2036	2037	2009	2019	2047	2024	2038
2025	AKY1	Bulunamamıştır									
	AKY2	Bulunamamıştır									
	BKY	2043	2047	2016	2036	2029	2017	2027	2013	2038	2018
2030	AKY1	Bulunamamıştır									
	AKY2	Bulunamamıştır									
	BKY	2043	2047	2038	2036	2013	2044	2016	2012	2029	2021
2035	AKY1	Bulunamamıştır									
	AKY2	2046									
	BKY	2038	2044	2043	2013	2047	2034	2033	2016	2036	2029
2040	AKY1	Bulunamamıştır									
	AKY2	Bulunamamıştır									
	BKY	2012	2036	2037	2047	2042	2021	2009	2019	2043	2016
2045	AKY1	Bulunamamıştır									
	AKY2	Bulunamamıştır									
	BKY	2036	2047	2012	2043	2042	2038	2021	2013	2016	2009
2050	AKY1	2009									
	AKY2	2008	2009	3029							
	BKY	2009	2036	2047	2012	2037	2016	2043	2021	2013	2008

2. Aranan anahtar kelimeler, aday belgenin metninin içinde geçtiği halde (AKY2 satırında gösterilmektedir) anahtar kelime olarak seçilmediğinden (AKY1 satırında gösterilmektedir) bu aday belge "benzer belge" olarak kabul edilmemektedir. Ancak bulanık kümeleme yaklaşımı ile bu aday belgeler benzer belge olarak tespit edilebilmektedir. Buna örnek olarak 2050, 3015, 3020, 3045, 4020 ve 4040 numaralı belgeler verilebilir.

Tablo 7.13. Ziraat alanı için toplu test sonuçları

Aranan Belge	Arama Türü	Bulunan Belgeler									
		1	2	3	4	5	6	7	8	9	10
3005	AKY1	Bulunamamıştır									
	AKY2	Bulunamamıştır									
	BKY	3039	3032	3031	3016	3033	3004	3047	3029	3013	3027
3010	AKY1	Bulunamamıştır									
	AKY2	Bulunamamıştır									
	BKY	3032	3039	3031	3004	3029	3002	3033	3028	3016	3047
3015	AKY1	3038									
	AKY2	3008	3009	3038							
	BKY	3039	3008	3023	3019	3038	3022	3018	3049	3011	3031
3020	AKY1	3008	3014	3018	3021	3022					
	AKY2	3008	3014	3018	3019	3021	3022				
	BKY	3039	3022	3021	3018	3008	3019	3032	3023	3016	3031
3025	AKY1	Bulunamamıştır									
	AKY2	Bulunamamıştır									
	BKY	3039	3031	3032	3002	3044	3023	3028	3029	3014	3004
3030	AKY1	3028	3029								
	AKY2	3028	3029								
	BKY	3029	3039	3028	3032	3004	3002	3016	3033	3044	3047
3035	AKY1	3024									
	AKY2	3024									
	BKY	3024	3039	3032	3031	3002	3029	3004	3033	3028	3044
3040	AKY1	Bulunamamıştır									
	AKY2	3007	3036	3049							
	BKY	3032	3028	3002	3004	3029	3031	3039	3033	3044	3016
3045	AKY1	3026	3038	3046	3047						
	AKY2	3009	3026	3036	3038	3044	3046	3047			
	BKY	3024	3031	3032	3002	3029	3044	3039	3033	3016	3026
3050	AKY1	Bulunamamıştır									
	AKY2	Bulunamamıştır									
	BKY	3039	3023	3031	3032	3022	3014	3002	3008	3019	3024

3. Bu koleksiyondaki belgelerin tamamı tek bir konuya ait olarak kabul edilmiştir. Ancak test sonuçları, bazı belgelerin anahtar kelimelerinin başka konuya ait belgelerin metinlerinde ve hatta anahtar kelime kısımlarında geçtiğini göstermştir. Bu da daha iyi bir benzerlik araması için bir belgenin birden fazla konuya ait olması durumunun göz ardı edilmemesi gerektiğini göstermektedir. Yukarıdaki duruma örnek olarak 1005, 1015, 1045, 2050, 4020 ve 4045 numaralı belgeler örnek olarak verilebilir.

Tablo 7.14. Sosyal ve Beşeri Bilimler alanı için toplu test sonuçları

Aranan Belge	Arama Türü	Bulunan Belgeler									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
4005	AKY1	4041									
	AKY2	4041									
	BKY	4012	4026	4002	4006	4044	4013	4028	4048	4004	4017
4010	AKY1	Bulunamamıştır									
	AKY2	Bulunamamıştır									
	BKY	4026	4012	4006	4013	4028	4044	4004	4036	4017	4008
4015	AKY1	Bulunamamıştır									
	AKY2	Bulunamamıştır									
	BKY	4002	4012	4017	4013	4006	4044	4018	4028	4008	4026
4020	AKY1	3007	3031								
	AKY2	4019	2003	3007	3031	4021	4022	4024			
	BKY	4019	4006	4004	4021	4026	4008	4036	4032	4024	4043
4025	AKY1	Bulunamamıştır									
	AKY2	Bulunamamıştır									
	BKY	4026	4019	4046	4004	4049	4006	4007	4031	4042	4012
4030	AKY1	Bulunamamıştır									
	AKY2	Bulunamamıştır									
	BKY	4006	4012	4026	4002	4013	4004	4028	4044	4008	4019
4035	AKY1	4038	4039	4042							
	AKY2	4038	4039	4042							
	BKY	4042	4012	4002	4026	4044	4016	4017	4006	4048	4028
4040	AKY1	Bulunamamıştır									
	AKY2	4047	4049								
	BKY	4049	4026	4012	4047	4048	4042	4006	4019	4004	4046
4045	AKY1	4026									
	AKY2	3007	4026								
	BKY	4026	4012	4006	4044	4028	4004	4017	4013	4042	4048
4050	AKY1	Bulunamamıştır									
	AKY2	Bulunamamıştır									
	BKY	4002	4012	4017	4048	4026	4006	4044	4042	4013	4049

4. SONUÇ ve ÖNERİLER

Hızla artan miktarlarda olan metinsel belgelerin etkili bir şekilde organize edilmesi ve bu metinlerden faydalanılması gerekmektedir. Bu amaca yönelik olarak yapılan metin madenciliği araştırmalarının önemli bir parçası ise benzer belge aranmasıdır. Bu çalışmada benzer belge aranması konusu bulanık mantık kullanılarak ele alınmıştır.

Çalışmanın önemli bir özelliği, bir belgenin benzerinin aranması işlemi için anahtar kelime yerine tüm kelimelerin kullanılmasıdır. Klasik anahtar kelime yaklaşımında, bir metin için manüel olarak belirlenmiş anahtar kelimeler mevcuttur. Arama işlemi veya metnin indekslenmesinde bu kelimeler kullanılır. Ancak bu kelimelerin uygun bir şekilde seçilmesi yani metni iyi ifade edebilecek şekilde seçilmesi çok önemlidir. Bu anahtar kelime seçimindeki eksiklik veya hatalar tüm arama performansını etkileyecektir. Çalışmada bu sorundan sakınmak amacıyla metnin tüm kelimelerinin kullanıldığı bir yaklaşım benimsenmiştir.

Bir belgenin benzerlerinin bulunması işlemi için öncelikle bu belgelerin iyi bir şekilde organize edilmesi ve verimli bir arama alt yapısı oluşturulması gerekmektedir. Çalışmada öncelikle bu amaca yönelik olarak etkili bir arama altyapısı geliştirilmeye çalışılmıştır. Oluşturulan bu altyapının en önemli özelliklerinden biri içerisinde bulanık mantık kullanılmış olmasıdır.

Bilindiği üzere bazı metinsel belgeler içerikleri gereği tek bir konu ile ilgili olmayabilirler. Bir başka deyişle bazı belgeler birden fazla konu ile ilgilidirler. Bu durumda ise belgelerin sınıflandırılması veya kümelenmesi işlemlerinde kesin bir yaklaşım yerine bulanık bir yaklaşım benimsenmesi problemin doğasına daha uygun olacaktır. Bulanık mantığın metin madenciliğindeki kullanım şekilleri göz önünde bulundurularak, bulanık yaklaşım için bulanık benzerlik sınıflandırmasına kullanılmıştır. Bu yöntemi temel alan bir arama iskeleti oluşturulmuştur.

Bu çalışmada elde edilen sonuçların en önemlileri aşağıda sıralanmıştır:

1- Benzer belge araması; önişleme, sınıflandırma/kümeleme ve benzerlik ölçümü olarak üç aşamada ele alınmıştır. Ön işleme aşaması için metinlerin sunumu aşamasında terim ağırlıklandırma üzerinde durulmuştur. Burada karşılaştırılan terim ağırlıklandırma yöntemleri; terim sıklığı, normalize edilmiş terim sıklığı, terim sıklığı – ters belge sıklığı ve normalize edilmiş terim sıklığı – ters belge sıklığıdır. Sonuç olarak terim sıklığı yönteminin arama yaklaşımına en uygun yöntem olduğu tespit edilmiştir.

2- Benzerlik ölçümü aşamasında mevcut benzerlik ölçümlerinin önerilen arama yaklaşımındaki performansları değerlendirilmiştir. Burada incelenen yöntemler kosinüs, zar benzerlik ölçümleri ve Minkowski metriktir. Bu kısımda ayrıca verinin boyutuna dayalı yeni bir benzerlik ölçümü (DRSim) önerilmiştir. İki farklı veri seti üzerinde yapılan deneysel çalışma sonucunda; önerilen benzerlik ölçümünün kosinüs ve zar benzerliğine yakın zaman değerlerine sahip olduğu ve aynı zamanda Minkowski metrik (p=50 durumu) sonuçlarına yakın bir ayrıştırma oranına ulaştığı görülmüştür.

3- Benzeri aranan metinsel belgelerden bazılarının birden fazla konu (kategori) ile ilgili oldukları bilinmektedir. Bir belgenin birden fazla kategoriye aitliği durumu ve bu durumda ait olunan kategorilerin tespit edilmesi problemi *çoklu kategori problemi* adıyla ele alınmıştır. Bu problemin, birden fazla konu ile ilgili belgelerin tespiti ve birden fazla olan bu konuların hangi konular olduğunun belirlenmesi olarak iki aşama halinde ele alınmasının başarılı bir şekilde uygulanabildiği görülmüştür.

4- Birden fazla konu ile ilgili belgelerin tespiti için öncelikle metin madenciliğinde sıkça kullanılan Rocchio algoritması ve Naive Bayes yöntemi kullanılarak bir sınıflandırma yapılmıştır. Daha sonra mevcut FSC yöntemi daha da geliştirilerek problemin çözümüne uygun hale getirilmiştir. Önerilen bu yöntemin (α-FSCM) en önemli kısmının α eşik değerinin belirlenmesi olduğu görülmüştür. Bu değerin belirlenmesi için ise yine eğitim veri kümesine dayalı bir çözüm kullanılmıştır. Reuter-21578 dağıtım 1.0 metin koleksiyonun üzerinde yapılan deneysel çalışmada önerilen α-FSCM ve bu yöntemdeki α eşik değerinin belirlenmesinin oldukça başarılı olduğu

görülmüştür. Ayrıca önerilen yöntemin, Rocchio algoritması ve Naive Bayes yönteminin her ikisinden de daha yüksek bir sınıflandırma değerine ulaştığı görülmüştür.

5- Birden fazla konuya ait olan belgeler için bu konuların hangi konular olduğunun belirlenmesi için konuların birlikte görülme sıklıklarının kullanıldığı bir yöntem (MCVM) önerilmiştir. Deneysel çalışma yine Reuter-21578 dağıtım 1.0 metin koleksiyonundan seçilen belgeler üzerinde yapılmıştır. Bu belgeler için çoklu kategori tespitinde (2 kategoriye ait olma durumunda) önerilen yöntemin klasik yönteme göre %179 başarı artışı sağladığı görülmüştür.

Özet olarak bu çalışma bir benzer belge aranması işleminde;

- uygun bir terim ağırlıklandırma yönteminin seçilmesinin,
- verimli bir benzerlik ölçümümün kullanılmasının ve
- belgelerin birden fazla kategoriye ait olması durumlarının ele alınmasının oldukça önemli olduğunu ortaya koymuştur.

Birçok alanda kullanılan ve en temel kavramlardan bir olan benzerlik ölçümü metin madenciliğinde de önemli bir yere sahiptir. Bundan dolayı bu çalışmada önerilen verinin boyutuna dayalı benzerlik ölçümü, metin madenciliği ile ilgili diğer çalışmalarda da kullanım alanı bulabilecektir.

Ayrıca, bir belgenin birden çok kategoriye ait olması ile ilgili aramalar üzerine şimdiye kadar çalışılmamış olması bu çalışmanın önemini artırmaktadır. Bu çalışma metin madenciliğinde çoklu kategori problemi ile ilgili gelecek çalışmalar açısından faydalı olabilecektir.

KAYNAKLAR

Abulaish ve De (2007) Abulaish M., De L., Biological Relation Extraction and Query Answering from MEDLINE Abstracts Using Ontology-Based Text Mining, Data & Knowledge Engineering, 61: 228-262, 2007.

Apte ve ark. (1998) Apte C., Damerau P. and Weiss S., Text Mining with Decision Rules and Decision Trees, In Proceedings of the Conference Automated Learning and Discovery, CMU, 1998.

Amasyalı ve Yıldırım (2004) Amasyalı M. F., Yıldırım T., Otomatik Haber Metinleri Sınıflandırma, SIU'04, Kuşadası, 2004.

Amasyalı ve Diri (2005) Amasyalı M. F., Diri B., Bir Soru Cevaplama Sistemi: Baybilmiş, Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, sayfa:37-51, 2005/1, 2005.

Atlam ve ark (2003) Atlam E. S., Fuketa M., Morita K., Aoe J. I., Documents Similarity Measurement Using Field Association Terms, Information Processing and Management , Vol. 39, Issue 6, pp. 809-824, 2003.

Bao ve ark. (2003) Bao J.P., Shen J.y., Liu X.D., Song Q.B., A New Text Feature Extraction Model and Its Application in Document Copy Detection, Proceeding of 2nd International Conference on Machine Learning and Cybernetics, pp. 82-87, 2003.

Bayer ve ark. (1998) Bayer T., Kressel U., Mogg-Schneider H., Renz I., Categorizing Paper Documents: A Generic System for Domain and Language Independent Text Categorization, Computer Vısıon And Image Understandıng, Vol. 70, No. 3, pp. 299-306, 1998.

Bezdek (1981) Bezdek J. C., Pattern Recognition with Fuzzy Objective Function Algoritms, Plenum Press, New York, 1981.

Bhuyan ve ark. (1991) Bhuyan J. N., Deogun J. S., and Raghavan V. V., Cluster-Based Adaptive Information Retrieval, System Sciences, Proceedings of the Twenty-Fourth Annual Hawaii International Conference, Vol. 1, pp. 307-316, 1991.

Burges (1998) Burges J. C., A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery, Vol. 2, No. 2, sayfa 121-167, 1998.

Clarke ve ark. (2000) Clarke C. L. A., Cormack G. V., Kisman D. I. E., Lynam T. R., Question Answering by Passage Selection, The Ninth Text Retrieval Conference, 2000.

Cooper ve ark. (2002) Cooper J.W., Coden A.R., Brown E.W., System Sciences, A Novel Method for Detecting Similar Documents, HICSS. Proceedings of the 35th Annual Hawaii International Conference, pp. 1153-1159, 2002.

Delgado ve ark. (1995) Delgado M., Gomez-Skarmeta A., Vila M.A., Hierarchical Clustering to validate Fuzzy Clustering, Proceedings of the IEEE Int. Conf. on Fuzzy Systems, pp. 1807-1812, 1995.

Denoyer ve Gallinari (2004) Denoyer L. and Gallinari P., Bayesian Network Model for Semi-Structured Document Classification, Information Processing and Management 40, pp. 807-827, 2004.

Dhillon ve ark. (2001) Dhillon I. S., Fan J., Guan Y., Efficient Clustering of Very Large Document Collections, Data Mining for Scientific and Engineering Applications, 2001.

Duda ve Hart (1973) Duda, R. O. and Hart, P. E., Pattern Classification and Scene Analysis, A Wiley-Interscience Publication, 1973.

Dumanis ve ark. (1998) Dumais S., Platt J., Heckerman D., and Sahami M., Inductive Learning Algorithm and Representations for Text Categorization, In Proceedings of the 1998 ACM 7th International Conference on Information and Knowledge Management, 148-155, 1998.

Dunn (1973) Dunn J. C., A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters, Journal of Cybernetics 3: 32-57, 1973.

Eghhe ve Michel (2002) Eghhe L. and Michel C., Strong Similarity Measures for Ordered Sets of Documents in Information Retrieval, Information Processing and Management, 38, pp. 823-848, 2002.

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

Elworthy (2000) Elworthy D., Question Answering Using a Large NLP System, The Ninth Text Retrieval Conference, Gaithersburg, 2000.

Feldman ve ark. (2003) Feldman R., Regev Y., Hurvitz E. and Finkelstein-Landau M., Mining the Biomedical Literature Using Semantic Analysis and Natural Language Processing Techniques, BIOSILICO Vol. 1 No:2, 2003.

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

Gunter ve Chen (2001) Gunther P. and Chen P., A New Approach to Hybrid SOM Implementation for Text Classification, Proceedings of the 10th International IEEE conference on Fuzzy Systems, IEEE CS Press, pp.968-972, 2001.

Gurusamy ve ark. (2002) Gurusamy S., Manjula D., Geetha T.V., Text Mining in 'Request for Comments Document Series', Proceedings of the Language Engineering Conference (LEC'02), 2002.

Hotho ve ark. (2003) Hotho A., Staab S., Stumme G., Text Clustering Based on Background Knowledge, University of Karlsruhe, Institute AIFB, 2003.

Huaizhong ve Gardarin (2002) Huaizhong K., Gardarin G., Similarity Model and Term Association for Document Categorization, Database and Expert Systems Applications, Proceedings. 13th International Workshop, pp. 256-260, 2002.

Jing ve ark. (2002) Jing L.P., Huang H.K., Shi H.B., Improved Feature Selection Approach TFIDF in Text Mining, Proceeding of 1st International Conference on Machine Learning and Cybernetics, pp. 944-946, 2002.

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

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

Kim ve ark. (2000) Kim S., Baek D., Kim S., Rim H., Question Answering Considering Semantic Categories and Co-occurrence Density, The Ninth Text Retrieval Conference, 2000.

Klose ve ark. (2000) Klose A., Nürnberger A., Kruse R., Hartmann G., Richards M., Interactive Text Retrieval Based on Document Similarities, Phys. Chem. Earth (A), Vol. 25, No.8, pp. 649-654, 2000.

Ko ve ark. (2004) Ko Y., Park J., Seo J., Improving Text Categorization Using the Importance of Sentences, Information Processing and Management, 40, pp. 65-79, 2004.

Kou ve Gardarin (2002) Kou H., Gardarin G., Similarity Model and Term Association for Document Categorization, Proceedings of the 13th International Workshop on Database and Expert Systems Applications (DEXA'02) , 2002.

Kowalski (1997) Kowalski G., Information Retrieval Systems: Teory and Implementation Boston: Kluwer Academic Publishers, 1997.

Krishnapuram ve ark. (2001) Krishnapuram R., Joshi A., Nasraoui O., Yi L., Low-Complexity Fuzzy Relational Clustering Algorithms for Web Mining, IEEE Transactional on Fuzzy Systems, Vol. 9, No.4, 2001.

Lang (1995) Lang K., NewsWeeder: Learning to Filter News, In proceedings of the 12th International Conference on Machine learning, pp. 331-339, 1995.

Larsen ve Aone (1999) Larsen B. and Aone C., Fast and Effective Text Mining Using Linear-Time Document Clustering, Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, pp. 16-22, 1999.

Latiri ve ark. (2003) Latiri C. C., Elloumi S., Chevallet J. P., Jaouay A., Extension of Fuzzy Galois Connection for Information Retrieval using a Fuzzy Quantifier, ACS/IEEE International Conference on Computer Systems and Applications (AICCSA'03) , 2003.

Li ve ark. (2006) Li Y., Shiu S. C. K., Pal S. K., Liu J. N. K., A Rough Set-Based Case-Based Reasoner for Text Categorization, International Journal of Approximate Reasoning, 41, pp.229-255, 2006.

Masand ve ark. (1992) Masand B., Linoff G., and Waltz D., Classifying News Stories Using Memory Based Reasoning, In Proceedings of the 15th Annual ACM/SIGIR Conference on Research and Development in Information Retrieval, pp. 59-65, 1992.

MacQueen (1967) MacQueen J.B., Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, 1:281-297, 1967.

Meziane ve Rezgui (2004) Meziane F., Rezgui Y., A Document Management Methodology Based on Similarity Contents, Information Sciences, Vol. 158, pp. 15-36, 2004.

Mine ve ark. (2002) Mine T., Lu S., Amamiya M., Discovering Relationships between Topics of Conferences by Filtering, Extracting and Clustering, Proceedings of the 13th International Workshop on Database and Expert Systems Applications (DEXA'02), 2002.

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

Miyamoto (1990) Miyamoto S., Fuzzy Sets in Information Retrieval and Cluster Analysis, Kluwer Academic Publisher, 1990.

Miyamoto (2001) Miyamoto S., Fuzzy Multisets and Fuzzy Clustering of Documents, In Proc. of the IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2001.

Mizutani ve Miyamoto (2003) Mizutani K. and Miyamoto S., Fuzzy Multiset Model for Information Retrieval and Clustering Using a Kernel Function. ISMIS 2003, pp. 417-421, 2003.

Morton (1999) Morton T. S., Using Coreference in Question Answering, The Eighth Text Retrieval Conference, 1999.

Murata ve ark. (2000) Murata M., Ma Q., Uchimoto K., Ozaku H., Utiyama M., Isahara H., Japanese Probabilistic Information Retrieval Using Location and Category Information, Proceedings of The Fifth International Workshop on Information Retrieval with Asian Language, 2000.

Pazzani ve ark. (1996) Pazzani M., Muramatsu J., Billsos D., Syskill & Webert: Identifying Interesting Web Sites, In Proceedings of the 13th National Conference on Artificial Intelligence, pp 54-61, 1996.

Perin ve Petry (2003) Perrin P. and Petry F. E., Extraction and Representation of Contextual Information for Knowledge Discovery in Texts, Information Sciences 151, pp. 125-152, 2003.

Qiu ve ark. (2002) Qiu X., Tang Y., Meng D., Xu Y., A New Fuzzy Clustering method Based On Distance and Densty, IEEE International Conference on Systems, Man and Cybernetics, Vol. 7, Page(s): 5, 2002.

Ruiz ve Srinivasan (2002) Ruiz M. E. and Srinivasan P., Hierarchical Text Categorization Using Neural Networks, Kluwer Academic Publishers. Information Retrieval, 5, pp. 87-118, 2002.

Sahami ve ark. (1998) Sahami M., Dumais S., Heckerman D., Horvitz E., A Bayesian Approach to Filtering Junk e-mail, AAAI 98 Workshops on Text Categorization, 1998.

Salton (1989) Salton G., Automatic Text Processing: The transformation, analysis, and retrieval of information by computer, Addison-Wesley, 1989.

Salton ve McGill (1983) Salton G. and McGill, M. H., Introduction to Modern Information Retrieval. New York: McGraw-Hill, 1983.

Salton ve Buckley (1988) Salton G. and Buckley, C., Term Weighting Approaches in Automatic Text Retrieval, Information Processing and Management, Vol. 24, No. 5, 513-523, 1988.

Saraçoğlu ve ark. (2007) Saraçoğlu R., Tütüncü K., Allahverdi N., A Fuzzy Clustering Approach For Finding Similar Documents Using A Novel Similarity Measure. Expert Systems with Applications, 33(3): 600-605, 2007.

Saraçoğlu ve ark. (2008) Saraçoğlu R., Tütüncü K., Allahverdi N., A New Approach on Search for Similar Documents with Multiple Categories Using Fuzzy Clustering. Expert Systems with Applications, 36: In Press, 2008.

Sıramkaya (2006) Sıramkaya E., 2006, Veri Madenciliğinde Bulanık Mantık Uygulaması, Yüksek Lisans Tezi, Fen Bilimleri Enstitüsü, Selçuk Üniversitesi, Konya.

Sitarama ve ark. (2004) Sitarama S., Mahadevan U., Abrol M., Efficient cluster representation in similar document search, Proceedings of WWW Conference, 2004.

Subasic ve Huettner (2001) Subasic P. and Huettner A., Affect Analysis of Text Using Fuzzy Semantic Typing, IEEE Transactions on Fuzzy Systems, Vol. 9, No. 4, 2001.

Vasifov (2001) Vasifov Z., Example Based Text Categorization for Turkish Language, Yüksek Lisans Tezi, Fen Bilimleri Enstitüsü, Dokuz Eylül Üniversitesi, İzmir, 2001.

Weng ve Lin (2003)	Weng S. S. And Lin Y. J., A Study on Searching For Similar Documents Based on Multiple Concepts and Distribution of Concepts, Expert Systems with Applications, Volume 25, No. 3, pp. 355-368, 2003.
Weng ve Liu (2004)	Weng S. S. and Liu, C. K., Using Text Classification and Multiple Concepts to Answer e-mails, Expert Systems with Applications, Volume: 26, Issue: 4, pp. 529-543, 2004.
Widyantoro ve Yen (2000)	Widyantoro D. H. and Yen J., A Fuzzy Similarity Approach in Text Classification Task, IEEE, 2000.
Yang (1999)	Yang Y., An Evaluation of Statistical Approaches to Text Categorization, Journal of Information Retrieval, 1(1/2), pp. 67-88, 1999.
Zhang ve Ramussen (2001)	Zhang J. and Rasmussen E. M., Developing a New Similarity Measure from Two Different Perspectives, Information Processing & Management, Volume 37, Issue 2, pp. 279-294, 2001.
Zhang ve Oles (2001)	Zhang T. and Oles F. J., Text Categorization Based on Regularized Linear Classification Methods, Kluwer Academic Publishers. Information Retrieval, 4, pp. 5-31, 2001.

EK. ÖRNEK BENZER BELGE ARAMA UYGULAMASI

Örnek 1:

Burada verilen örnek bilgisayar bilimleri konusuna ait belgelerden seçilen 1005 numaralı belgedir. Bu belgenin metni Şekil Ek.1.'de görülmektedir.

Belge No: 1005
<Journal> Applied Soft Computing **<Title>** Modeling and control of non-linear systems using soft computing techniques **<Author>** M.A. Denaï, F. Palis and A. Zeghbi **<Abstract>** This work is an attempt to illustrate the utility and effectiveness of soft computing approaches in handling the modeling and control of complex systems. Soft computing research is concerned with the integration of artificial intelligent tools (neural networks, fuzzy technology, evolutionary algorithms, …) in a complementary hybrid framework for solving real world problems. There are several approaches to integrate neural networks and fuzzy logic to form a neuro-fuzzy system. The present work will concentrate on the pioneering neuro-fuzzy system, Adaptive Neuro-Fuzzy Inference System (ANFIS). ANFIS is first used to model non-linear knee-joint dynamics from recorded clinical data. The established model is then used to predict the behavior of the underlying system and for the design and evaluation of various intelligent control strategies. **<Keywords>** ANFIS; Neural networks; Modeling; Intelligent control;

Şekil Ek.1. 1005 numaralı belge

Bu belge için benzer belge arama işleminden sonra elde edilen sonuçlar Tablo Ek.1.'de verilmiştir.

Tablo Ek.1. 1005 numaralı belge için arama sonuçları

Benzeri Aranan Belge no	Anahtar Kelime kullanılarak bulunan		Bulanık kümeleme kullanılarak bulunan	
	Belge no	Benzerlik oranı	Belge no	Benzerlik değeri
1005	1017	%25	1024	0.743
	1037	%25	1007	0.733
			1017	0.726
			1003	0.724
			1016	0.721
			1032	0.717
			1034	0.706
			1031	0.704
			1038	0.703
			1013	0.703

Bu belge için yapılan anahtar kelime tabanlı arama sonucunda elde edilen iki belge (1017 ve 1037 numaralı belgeler) Şekil Ek.2 ve Şekil Ek.3'te görülmektedir.

Belge No: 1017

<Journal> Applied Soft Computing

<Title> Machine learning for frequency estimation of power systems

<Author> E.S. Karapidakis

<Abstract> In this paper the application of machine learning techniques for on-line dynamic security assessment of power systems is presented. Decision trees (DT), artificial neural networks (ANN) and entropy networks (EN) are developed and applied on the power system of Crete, the largest Greek island. Comparison of these methods reveals their relative advantages and disadvantages. These methods have been integrated in the dynamic security assessment module of the advanced control system of Crete island, helping to identify the operating conditions and parameters that lead to a less robust operation of the system. The results are considered very satisfactory, both in accuracy that increases the reliability of the method and in computational time, which is a necessity for real time applications.

<Keywords> Power systems; Dynamic security assessment; Machine learning; Decision trees; Entropy trees; Neural networks; energy management systems;

Şekil Ek.2. 1017 numaralı belgenin metni

Belge No: 1037
<Journal> Computer Networks **<Title>** Describing and simulating internet routes **<Author>** Jérémie Leguay, Matthieu Latapy, Timur Friedman, and Kavé Salamatian **<Abstract>** This contribution deals with actual routes followed by packets in the Internet at the ip level. We first propose a set of statistical properties to analyse such routes. We then use the results to suggest and evaluate methods for generating artificial routes suitable for simulation purposes. The proposed approach also leads to insight on various network models. The present work is based on large data sets provided mainly by caida's skitter infrastructure. **<Keywords>** Internet; Routing; Routes; Complex networks; Graphs; Measurement; Modeling;

Şekil Ek.3. 1037 numaralı belgenin metni

Örnek belge için tüm kelimelerin kullanıldığı yaklaşımla yapılan arama işleminden sonra elde edilen sonuç belgeler ise Şekil Ek.4-13'te yer almaktadır.

Belge No: 1024
<Journal> Computer Languages, Systems & Structures **<Title>** Extending movilog for supporting Web services **<Author>** Cristian Mateos, Alejandro Zunino, and Marcelo Campo **<Abstract>** Web Services enable computers to interact and exploit Web-accessible programs without human intervention. Despite researchers agree that mobile agent technology will obtain significant benefits from this line of research, the lack of proper development tools hinder the widespread adoption of mobile agent technology on the Web. This paper describes a novel programming language called WS-Log whose goal is to provide a tight integration between mobile agents and Web Services. Examples and experimental results showing some of the advantages of WS-Log are also reported. **<Keywords>** Mobile agents; Logic programming; Web services; Intelligent agents;

Şekil Ek.4. 1024 numaralı belgenin metni

Belge No: 1007
<Journal> Applied Soft Computing **<Title>** Artificial neural network based prediction of drill flank wear from motor current signals **<Author>** Karali Patra, Surjya K. Pal, and Kingshook Bhattacharyya **<Abstract>** In this work, a multilayer neural network with back propagation algorithm (BPNN) has been applied to predict the average flank wear of a high speed steel (HSS) drill bit for drilling on a mild steel work piece. Root mean square (RMS) value of the spindle motor current, drill diameter, spindle speed and feed-rate are inputs to the network, and drill wear is the output. Drilling experiments have been carried out over a wide range of cutting conditions and the effects of drill wear, cutting conditions (speed, drill diameter, feed-rate) on the spindle motor current have been investigated. The performance of the trained neural network has been tested for new cutting conditions, and found to be in very good agreement to the experimentally determined drill wear values. The accuracy of the prediction of drill wear using neural network is found to be better than that using regression model. **<Keywords>** Drilling; Flank wear; Current sensors; Artificial neural network; Regression model;

Şekil Ek.5. 1024 numaralı belgenin metni

Belge No: 1003

<Journal> Artificial Intelligence

<Title> A representation theorem for minmax regret policies

<Author> Sanjiang Li

<Abstract> Decision making under uncertainty is one of the central tasks of artificial agents. Due to their simplicity and ease of specification, qualitative decision tools are popular in artificial intelligence. Brafman and Tennenholtz [R.I. Brafman, M. Tennenholtz, An axiomatic treatment of three qualitative decision criteria, J. ACM 47 (3) (2000) 452–482] model an agent's uncertain knowledge as her local state, which consists of states of the world that she deems possible. A policy determines for each local state a total preorder of the set of actions, which represents the agent's preference over these actions. It is known that a policy is maximin representable if and only if it is closed under unions and satisfies a certain acyclicity condition.

In this paper we show that the above conditions, although necessary, are insufficient for minmax regret and competitive ratio policies. A complete characterization of these policies is obtained by introducing the best-equally strictness.

<Keywords> Qualitative decision; Policy; maximin; minmax regret; competitive ratio;

Şekil Ek.6. 1003 numaralı belgenin metni

Belge No: 1016

<Journal> Applied Soft Computing

<Title> Word segmentation of handwritten text using supervised classification techniques

<Author> Yi Sun, Timothy S. Butler, Alex Shafarenko, Rod Adams, Martin Loomes and Neil Davey

<Abstract> Recent work on extracting features of gaps in handwritten text allows a classification of these gaps into inter-word and intra-word classes using suitable classification techniques. In this paper, we first analyse the features of the gaps using mutual information. We then investigate the underlying data distribution by using visualisation methods. These suggest that a complicated structure exists, which makes them difficult to be separated into two distinct classes. We apply five different supervised classification algorithms from the machine learning field on both the original dataset and a dataset with the best features selected using mutual information. Moreover, we improve the classification result with the aid of a set of feature variables of strokes preceding and following each gap. The classifiers are compared by employing McNemar's test. We find that SVMs and MLPs outperform the other classifiers and that preprocessing to select features works well. The best classification result attained suggests that the technique we employ is particularly suitable for digital ink manipulation at the level of words.

<Keywords> Handwriting; Supervised classification; Mutual information; McNemar's test;

Şekil Ek.7. 1016 numaralı belgenin metni

Belge No: 1034

<Journal> Computer Networks

<Title> Consistent proportional delay differentiation: A fuzzy control approach

<Author> Jianbin Wei, and Cheng-Zhong Xu,

<Abstract> Proportional delay differentiation (PDD) is an important service model for providing relative differentiated services on the Internet. It aims to maintain pre-specified packet queueing-delay ratios between different classes of traffic at each hop. Existing rate-allocation approaches for PDD services assume the average queueing delay of a class is inversely proportional to its service rate. This assumption is not necessarily valid when the system is not heavily loaded. To provide consistent PDD services under various load conditions, in this paper, we propose a novel rate-allocation approach that applies fuzzy control theory to capture the nonlinear relationship between the queueing delay and the service rate. In the approach, a class's service rate is adjusted according to a set of fuzzy control rules defined over its error (the difference between the target delay ratio and the achieved one), the change of error, and the change of service rate. We prove that the fuzzy control system is stable and the service rate of a class converges to its equilibrium point at steady state. Simulation results demonstrate that, in comparison with other rate-allocation approaches, the fuzzy control approach is able to provide consistent PDD services under wide range load conditions. It is also shown robust under various system conditions, including with multiple classes, changing target delay ratios, changing load conditions, and different traffic patterns.

<Keywords> Quality of service; Proportional delay differentiation; Fuzzy control; Rate allocation;

Şekil Ek.8. 1034 numaralı belgenin metni

Belge No: 1032

<Journal> Computer Networks

<Title> Risk-based attack strategies for mobile ad hoc networks under probabilistic attack modeling framework

<Author> Vasileios Karyotis, and Symeon Papavassiliou

<Abstract> In this paper, we introduce and design a modeling framework that allows for the study and analysis of attack propagation in mobile ad hoc networks. The choice of a statistical approach for the problem is motivated by the dynamic characteristics of the ad hoc topology and the stochastic nature of threat propagation. Based on this probabilistic modeling framework, we study the impact of topology and mobility in the propagation of software threats over ad hoc networks. We design topology control algorithms that indicate how to properly adjust an attacker's transmission radius, according to the measured topological characteristics and availability of its resources, in the process of infecting a network more effectively. Then based on these topology control algorithms we develop different attack strategies that may range from independent attacks to cooperative scenarios in order to increase the negative impact of an attack on the network. Our performance evaluation results demonstrate that the proposed topology control algorithms and respective attack strategies effectively balance the tradeoffs between the potential network damage and the attackers' lifetime, and as a result significantly outperform any other flat and threshold-based approaches.

<Keywords> Attack modeling; Mobile ad hoc networks; Topology control; QoS; Security;

Şekil Ek.9. 1032 numaralı belgenin metni

Belge No: 1031

<Journal> Computer Networks

<Title> Spectrum sharing in IEEE 802.11s wireless mesh networks

<Author> Sebastian Max, Guido R. Hiertz, Erik Weiss, Dee Denteneer, and Bernhard H. Walke

<Abstract> With current amendments, transmission rates of 100 Mb/s and more become possible with IEEE 802.11 WLANs. On the one hand, this allows the end user to change from wired to wireless infrastructure in even more application scenarios; on the other hand interference sensitive modes reduce the maximum range between the mobile station and the access point (AP). To extend the transmission range transparently, relay APs form a mesh network and provide wireless connection over large areas.

Besides path selection, a crucial capability of a wireless mesh network is the ability to share the available spectrum among the participants. In this work, we classify two inherently different MAC protocols according to this ability. The well-known IEEE 802.11 DCF takes the position of a typical CSMA/CA protocol, whereas the Mesh Network Alliance (MNA) represents a distributed, reservation-based approach.

To assess their performance, we follow a dual approach: first we develop a method to compute the capacity bounds of the protocols in the considered scenarios. It helps to estimate the absolute gain of spectrum sharing in wireless mesh networks. Second, the WARP2 simulation engine is used to compare the distributed behaviour of both protocols. This results in a relative evaluation. A final conclusion is drawn by combining the simulation and the theoretical results. It underlines the significant possibilities of the MNA approach and shows future directions for capacity gains.

<Keywords> Wireless mesh network; MAC protocol; IEEE 802.11s; Concurrent transmission; Spectrum sharing;

Şekil Ek.10. 1031 numaralı belgenin metni

Belge No: 1038

<Journal> Computer Networks

<Title> Improving the efficiency of multipath traffic via opportunistic traffic scheduling

<Author> Coskun Cetinkaya

<Abstract> Multipath routing, as defined by OSPF extensions and other protocols, enables a network's traffic to be split among two, or more, possibly disjoint paths. Advantages of multipath vs. unipath routing include load balancing, reduced latency, and improved throughput. However, once the control plane establishes multiple routes, a policy is needed for efficiently splitting traffic among the selected paths. In this paper, we introduce opportunistic multipath scheduling (OMS), a technique for exploiting short-term variations in path quality to minimize delay, while simultaneously ensuring that the splitting rules dictated by the routing protocol are satisfied. We develop a performance model of OMS and derive an asymptotic lower bound on the performance of OMS as a function of path conditions (mean, variance, and Hurst parameter) for self-similar traffic. Finally, we use an extensive simulation-based performance study to evaluate the accuracy of the analytical model, explore the impact of OMS on TCP throughput and performance of real-time traffic, and study the impact of factors such as delayed measurements.

<Keywords> Opportunistic scheduling; Performance analysis; Multipath traffic;

Şekil Ek.11. 1038 numaralı belgenin metni

Belge No: 1013

<Journal> Applied Soft Computing

<Title> A hybrid approach based on neural networks and genetic algorithms for detecting temporal patterns in stock markets

<Author> Hyun-jung Kim, and Kyung-shik Shin,

<Abstract> This study investigates the effectiveness of a hybrid approach based on the artificial neural networks (ANNs) for time series properties, such as the adaptive time delay neural networks (ATNNs) and the time delay neural networks (TDNNs), with the genetic algorithms (GAs) in detecting temporal patterns for stock market prediction tasks. Since ATNN and TDNN use time-delayed links of the network into a multi-layer feed-forward network, the topology of which grows by on layer at every time step, it has one more estimate of the number of time delays in addition to several control variables of the ANN design. To estimate these many aspects of the ATNN and TDNN design, a general method based on trial and error along with various heuristics or statistical techniques is proposed. However, for the reason that determining the number of time delays or network architectural factors in a stand-alone mode does not guarantee the illuminating improvement of the performance for building the ATNN and TDNN model, we apply GAs to support optimization of the number of time delays and network architectural factors simultaneously for the ATNN and TDNN model. The results show that the accuracy of the integrated approach proposed for this study is higher than that of the standard ATNN, TDNN and the recurrent neural network (RNN).

<Keywords> Adaptive time delay neural networks; Time delay neural networks; Genetic algorithms; Time series prediction; Stock market prediction;

Şekil Ek.12. 1013 numaralı belgenin metni