YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

TÜRKÇE BELGELERİN ANLAM TABANLI YÖNTEMLERLE MADENCİLİĞİ

Bilg. Yük. Müh. Ahmet GÜVEN

FBE Bilgisayar Mühendisliği Anabilim Dalı Bilgisayar Mühendisliği Programında Hazırlanan

DOKTORA TEZİ

Tez Savunma Tarihi : 8 Şubat 2007
Tez Danışmanı : Prof. Dr. Oya KALIPSIZ (YTÜ)
Jüri Üyeleri : Prof. Dr. Eşref ADALI (İTÜ)
: Prof. Dr. A. Coşkun SÖNMEZ (YTÜ)
: Prof. Dr. Ümit KOCABIÇAK (SAÜ)
: Doç. Dr. Selim AKYOKUŞ (DOĞUŞ)

İSTANBUL, 2007

YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

TÜRKÇE BELGELERİN ANLAM TABANLI YÖNTEMLERLE MADENCİLİĞİ

Bilg. Yük. Müh. Ahmet GÜVEN

FBE Bilgisayar Mühendisliği Anabilim Dalı Bilgisayar Mühendisliği Programında Hazırlanan

DOKTORA TEZİ

Tez Savunma Tarihi : 8 Şubat 2007
Tez Danışmanı : Prof. Dr. Oya KALIPSIZ (YTÜ)
Jüri Üyeleri : Prof. Dr. Eşref ADALI (İTÜ)
: Prof. Dr. A. Coşkun SÖNMEZ (YTÜ)
: Prof. Dr. Ümit KOCABIÇAK (SAÜ)
: Doç. Dr. Selim AKYOKUŞ (DOĞUŞ)

İSTANBUL, 2007

İÇİNDEKİLER **Sayfa**

SİMGE LİSTESİ iv

KISALTMA LİSTESİ v

ŞEKİL LİSTESİ vi

TABLO LİSTESİ vii

ÖNSÖZ viii

ÖZET ix

ABSTRACT xi

1. VERİ MADENCİLİĞİ 1

1.1 Veri Madenciliğinin Kökenleri 3
1.2 Veri Madenciliğinin Gelişimi 5
1.2.1 Veritabanlarında Veri Madenciliğinden Veri Keşfine 5
1.2.1.1 Temel Tanımlar 8
1.2.1.2 VÜK süreci 9
1.3 Veri Madenciliği Yöntemleri 11
1.4 Veri Madenciliğinin Genişlemesi 12
1.5 Belge Madenciliğine Geçiş 14

2. METİNLERİN ANALİZİ İLE BİLGİ OLUŞTURMA 16

2.1 Metin Veri Madenciliğinin Gelişimi **16**
2.2 Metin Veri Madenciliği Yöntemleri **17**
2.2.1 Metin Sorgulama **17**
2.2.2 Metin Madenciliği İle Sorgulama Sürecini İyileştirme **19**
2.2.2.1 Sorgulama İyileştirme **19**
2.2.2.2 Doğal Dil İle Sorgulama **20**
2.2.2.3 Belgeleri Demetleme **20**
2.2.2.4 Belgelerin sınıflara ayrıştırılması **20**
2.2.2.5 Belgeleri Özetleme **21**
2.2.2.6 Bilgi Çıkarma (Information Extraction) **21**
2.2.2.7 Doğal Dil İşleme **22**
2.2.2.8 Anlam Notları ve Sözlükler **22**
2.2.2.9 Kavram Haritaları (Ontolojiler) **23**
2.2.2.10 Bilgi Keşfi (Knowledge Discovery) **23**
2.2.3 Metin Sınıflama **24**
2.2.3.1 Tekli yada Çoklu Sınıflama **25**
2.2.3.2 Net Sınıflama & Derecelendirilmiş Sınıflama **25**
2.2.3.3 Metin Sınıflama ve Gizli Anlambilimsel Dizinleme (GAD) **26**
2.2.3.4 Metin Sınıflayıcıların Sınıflama Biçimi **26**
2.2.3.5 Eşik Değerini Belirleme **27**

2.2.3.6 Olasılık Temelli Sınıflayıcılar (Probabilistic Classifiers) 28
2.2.3.7 Sembolik Algoritma Temelli Sınıflayıcılar 29
2.2.3.8 Karar Ağaçları 29
2.2.3.9 Tümevarım Kuralları 30
2.2.3.10 Regresyon Yöntemleri 31
2.2.3.11 Online Yöntemler 31
2.2.3.12 Rocchio Yöntemi 33
2.2.3.13 Yapay Sinir Ağları 33
2.2.3.14 Örnek Tabanlı Sınıflayıcılar: 35
2.2.3.15 Destek Vektör Makinesi (Support Vector Machine) 36
2.2.3.16 Sınıflayıcılar Komitesi 37
2.2.3.17 Metin Sınıflama Yöntemlerinin Etkinliğinin Ölçülmesi 37
2.3 Metin Veri Madenciliğinin Genişlemesi 39
2.4 Doğal Dil İşleme ve Belge Veri Madenciliği 40
2.4.1 Biçimbirimsel Çözümleme 40
2.4.2 Sözdizimi Çözümlemesi 42

3. BELGE MADENCİLİĞİ ALTYAPISI 44

3.1 Vektör Uzay Modeli 46
3.2 Vektör Uzay Modelinin Matematiksel Anlamı 48
3.2.1 Lineer Cebir Teorisi 49
3.2.1.1 Vektörler 49
3.2.1.2 Vektör Uzayı 50
3.2.1.3 Vektör Uzayında Lineer Bağımsızlık ve Lineer Bağımlılık 50
3.2.1.4 Temel ve Koordinatlar 51
3.2.1.5 Ortogonallik 51
3.2.1.6 Normalleştirme 51
3.2.1.7 Kosinüs Açısının Özellikleri 52
3.3 Vektör Uzayında işlem 53
3.3.1 Vektör Uzay Modelinin Kısıtları 55
3.3.2 Vektör Uzay Modelini Geliştirme – Temel Bileşen Analizi 55
3.3.3 QR Faktörlerine Ayırma 57
3.3.3.1 QR - Kosinüs Açısının Hesaplanması 61
3.3.4 Düşük Rank ile Yakınsamanın Doğruluğu 62
3.3.5 Tekil Değer Ayrıştırması - Singular Value Decomposition (SVD) 64
3.3.5.1 TDA - Kosinüs Açısının Hesaplanması 69
3.4 Terim Ağırlığı Belirleme Yöntemleri 70

4. İKİNCİ NESİL BELGE MADENCİLİĞİ 75

4.1 Gizli Anlambilimsel Dizinleme 76
4.2 GAD Çalışma Metodolojisi 78
4.3 GAD Yönteminin ve TDA Tekniğinin Benzetimi 80
4.4 GAD Yöntemi ile Belge Madenciliği 84
4.4.1 Belge Sorgulama 85
4.4.2 Belge Kümeleme 85
4.4.2.1 Toplayıcı Yöntemler 87
4.4.2.2 Kümeleme Sonuçlarının Değerlendirilmesi 89
4.4.2.3 GAD ile Belge Kümeleme 90

5. N-GRAM DESTEKLİ GAD 93

5.1 N-gram Yöntemi 94
5.2 GAD ile N-gramın Birleştirilmesi 94
5.3 Konuyla İlgili Çalışmalar 95

6. GAD ve N-GRAM DESTEKLİ GAD İLE MADENCİLİK 97

6.1 Türkçe'nin Zenginliği ve Zorluğu 98
6.2 Türkçe Doğal Dil Projeleri 99
6.2.1 Türkçe Sözcük Veritabanı Projesi 100
6.2.2 Türkçe Kavramsal Sözlük 100
6.2.3 Sözlüksüz Köke ulaşma 102
6.2.4 Zemberek 102
6.3 Zemberek ile Kelimelerin En Yalın Halini Bulma 104
6.4 Uygulamada Kullanılan Belge Setleri 105
6.4.1 Türkçe Belge Seti 106
6.4.2 İngilizce Belge Seti 106
6.5 Atılabilir Kelimeler Listesi 107
6.5.1 Türkçe Atılabilir Kelimeler Listesi: 107
6.5.2 İngilize Atılabilir Kelimeler Listesi: 107
6.6 Uygulama 107
6.7 GAD ve n-gram destekli GAD ile sınıflama 108

7. BULGULAR 109

7.1 Kümeleme 109
7.1.1 Türkçe Belgelerin Kümelenmesi 109
7.1.2 İngilizce Belgelerin Kümelenmesi 112
7.2 Sorgular 113
7.2.1 Türkçe Belgelerin Sorgulanması 115
7.2.2 İngilizce Belgelerin Sorgulanması 117
7.3 Sınıflama 119

8. SONUÇ 123

KAYNAKLAR 126

İNTERNET KAYNAKLARI 132

EK KAYNAKLAR 133

EKLER 135

Ek 1 Türkçe Atılabilir Kelimeler Listesi 135
Ek 2 İngilizce Atılabilir Kelimeler Listesi 136
Ek 3 Sınıf Tanımları 2 138

ÖZGEÇMİŞ 147

SİMGE LİSTESİ

$\|\| A \|\|_2$	: Öklid Matris Normu
A	: Terim x belge Matrisi
Ă	: A Matrisinin Düşük Rank Vektör Uzayındaki Sunumu
A^T	: Matris Transpozu
e	: Birim Matris
e_j	: Birim Matrisin j. Kolonu
tdf	: Ters Belge Frekansı
tf.tdf	: Terim Frekansı Ters Belge Frekansı
V	: Vektör
w	: Terim Ağırlığı

KISALTMA LİSTESİ

a	:	Augmented
d	:	Belge Frekansı
DVM	:	Destek Vektör Makinesi
GAD	:	Gizli Anlambilimsel Dizinleme
idf	:	Inverse Document Frequency
l	:	Logaritma
LSI	:	Latent Semantic Indexing
n	:	Natural
OLAP	:	Online Analytical Processing
PCA	:	Principal Component Analysis
SDD	:	Sınıf Durum Değeri
SQL	:	Structured Query Language
SVD	:	Singular Value Decomposition
SVM	:	Support Vector Machine
T	:	Terim Frekansı
TBA	:	Temel Bileşen Analizi
TDA	:	Tekil Değer Ayrıştırma
Tfidf	:	Term frequency Inverse Document Frequency
Tr	:	Training Set
VÜK	:	Veritabanlarından Üstbilgi Keşfi -Knowladge Discovery in Databases
w	:	Weight

ŞEKİL LİSTESİ

		Sayfa
Şekil 1-1	Veri Madenciliği Konusuna Etki Eden Dsiplinler	4
Şekil 1-2	Veritabanlarından Üstbilgi Keşfi	10
Şekil 1-3	Veri Madenciliğinin Genişlemesi	13
Şekil 2-1	Yapay Sinir Ağları tabanlı Bir Metin Sınıflama Sistemi	33
Şekil 2-2	Yapay Sinir Ağının İç Yapısı	33
Şekil 2-3	Destek Vektör Makinesi ile Sınıflama	36
Şekil 3-1	Vektör Uzayı	49
Şekil 3-2	Koordinat Sistemi	52
Şekil 4-1	Üç Boyutlu Uzayda Müşteri Tercihleri	80
Şekil 4-2	İki Boyutlu Uzayda Kitaplar	83
Şekil 4-3	Hiyerarşik Kümeleme Yöntemleri	87
Şekil 4-4	Tekil Bağ ile Hiyerarşik Toplayıcı Kümeleme	88
Şekil 4-5	Tekil Bağ ile Hiyerarşik Toplayıcı Kümeleme	89
Şekil 4-6	Tekil Değer Ayrıştırmasının Anlamı	91
Şekil 6-1	Zemberek Kelime Çözümleme	103
Şekil 7-1	Türkçe Belge Seti İçin Küme Sayısına Bağlı Entropi Değişim Grafiği	110
Şekil 7-2	Türkçe Belge Seti İçin Küme Sayısına Bağlı F-Measure Değişim Grafiği	111
Şekil 7-3	Türkçe Belge Seti İçin Küme Sayısına Bağlı Entropi Değişim Grafiği	113
Şekil 7-4	Türkçe Belge Seti İçin Küme Sayısına Bağlı F-Measure Değişim Grafiği	114

TABLO LİSTESİ

		Sayfa
Tablo 1-1	2003 Uluslar Arası Veri Madenciliği Konferansı Konu Listesi	15
Tablo 2-1	İhtimal Tablosu	38
Tablo 4 -1	Porterstemmer İle İngilizce Kelimelerin En Yalın Hallerini Elde Etme	79
Tablo 4-2	Kümeleme Teknikleri Kullanım Alanları	86
Tablo 6-1	Gözlükçü Kelimesinin Türkçe Kavramsal Sözlükteki Karşılığı	101
Tablo 7-1	Karşılaştırma İçin Kullanılan Algoritmalar	109
Tablo 7-2	Türkçe Belge Seti İçin Entropi Ölçümleri	110
Tablo 7-3	Türkçe Belge Seti İçin F-Measure Ölçümleri	111
Tablo 7-4	İngilizce Belge Seti İçin Entropi Ölçümleri	112
Tablo 7-5	İngilizce Belge Seti İçin F-Measure Ölçümleri	112
Tablo 7-6	"Bilgi Yönetiminin Pazarlanması" Sorgusu	115
Tablo 7-7	"Kurumsal Kaynak Planlama" Sorgusu	116
Tablo 7-8	"Turizm Sektöründe İnsan Kaynakları Yönetimi" Sorgusu	117
Tablo 7-9	"Oil Company Acquisition" Sorgusunun Sonuçları	118
Tablo 7-10	"Oil İndustry Consumer Price " Sorgusunun Sonuçları	119
Tablo 7-11	Sınıflama İşlemi İçin Önceden Hazırlanan Sınıflar	120
Tablo 7-12	Sınıf Tanımlaması 1	120
Tablo 7-13	1. Sınıf Tanımına Göre Sınıflama	121
Tablo 7-14	2. Sınıf Tanımına Göre Sınıflama	121
Tablo 7-15	1. Sınıf Tanımına Göre Sınıflamanın Etkinliği	122
Tablo 7-16	2. Sınıf Tanımına Göre Sınıflamanın Etkinliği	122

ÖNSÖZ

Günümüzde bilgi, insan gücünden, maddi kaynaklardan daha üstün bir zenginlik faktörüdür. Bilgisayar teknolojisinin getirdiği bilgi işleme, saklama ve üretme kolaylıkları, internet teknolojisinin getirdiği iletişim olanaklarıyla birleşince, bilgi giderek çoğalmakta ve aynı oranda doğru bilgiye ulaşım zorlaşmaktadır. Bilgi genellikle belge bazlı ortamlarda tutulmaktadır. Bu belgelerin internet ve bilgisayar teknolojisi ile toplumda yaygınlaşması, farklı bilgiler arasındaki ilişkilerin ortaya çıkarılmasını sağlarken, ortaya çıkarılan birikimin paylaşılması da güçleşmektedir.

Bilgiye erişim alanında İngilizce başta olmak üzere, pek çok yabancı dille yazılmış kaynağı hızlı bir şekilde amaca yönelik incelemeyi ve taramayı sağlayan teknikler geliştirilmiştir. Türkçe kaynakların taranması ve incelenmesi ile ilgili ise az sayıda ve son dönemde yürütülen çalışmalar bulunmaktadır.

Bu konuda bir çalışmaya beni sevk eden ve çalışmalarım esnasında yardımlarını esirgemeyen değerli hocalarım Prof. Dr. Oya KALIPSIZ'a, Prof. Dr. Eşref ADALI'ya ve Doç. Dr. Selim AKYOKUŞ'a teşekkürü borç biliyorum. Kıymetli eşim Gülşah başta olmak üzere, canım oğlum Eren, ailem ve dostlarım, bu çalışma nedeniyle kendilerine karşı olan sorumluluklarımı aksatmama rağmen beni kucaklayan kişilere layık olmaya çalışacağıma söz veriyorum.

Şubat 2007

Ahmet GÜVEN

ÖZET

Bilgisayar sistemlerinin ilk uygulama alanları veri toplama ve raporlama üzerinedir. Veri saklama kapasitelerinin ve bu verileri işleyecek bilgisayar işlemci gücünün artması ile daha fazla veriyi saklama ve inceleme imkanı doğmuştur. (Fayyad vd, 1996a). Böylece daha önce verilerden elde edilemeyen ilişkilerin, desenlerin ortaya çıkarılması mümkün hale gelmiştir. Geleneksel sorgulama yöntemlerinden farklı olan bu yöntemler veri madenciliği adı altında toplanmıştır.Veri madenciliği, verilerin içerisindeki desenlerin, ilişkilerin, değişimlerin, düzensizliklerin, kuralların ve istatistiksel olarak önemli olan yapıların yarı otomatik olarak keşfedilmesidir (Hand vd., 2001).

Belge bazlı veri yığınları içinden doğru belgelerin bulunması, belgelerin birbirleri arasındaki ilişkilerin sorgulaması işlemleri için veri madenciliği alanındaki teknikler birebir uygulanabilir değildir. Bu nedenle belge madenciliği yapmak için farklı yöntemler geliştirilmiş ve bu alan metin madenciliği, belge madenciliği, yarı yapısal veri madenciliği gibi isimler altında toplanmıştır.

Belge madenciliği çalışmalarında amaç belge içeriğinin, bir insan tarafından okunmuşçasına bilgisayar ortamında belirlenmesini içerir. Bu durumda belgelerin hangi dilde yazıldığı önem kazanmaktadır. Bu yönü itibariyle doğal dil işleme alanı ile belge madenciliği arasında sıkı bir ilişki doğmuştur. Hem belge madenciliği hem de doğal dil işleme çalışmaları uzun yıllardan beri İngilizce başta olmak üzere farklı diller üzerinde yapılmıştır.

Türkçe doğal dil işleme çalışmalarının somut sonuçları yeni yeni elde edilmekte ve henüz net olarak araştırmacılar arasında paylaşılmış değildir. Bu nedenle doğal dil işleme tekniklerini içinde taşıyan bir Türkçe belge madenciliği çalışması yapmak, özellikle bu tez çalışmasının temellerinin atıldığı 2004 yılı içinde pek anlamlı ve mümkün olmamıştır.

Bu tez çalışması Türkçe belgeler üzerinde belge madenciliği yapmak amacıyla, Gizli Anlambilimsel Dizinleme (GAD) yöntemini kullanmakta ve kelimelere uygulanan n-gram yaklaşımını bu yöntemle birleştirmektedir.

Belge madenciliği çalışmalarının uluslararası çapta değerlendirilebilmesi için, her belge madenciliği yöntemi ile kullanılabilecek standart belge kümeleri geliştirilmiştir. Bu konuda Türkçe yapılan çalışmalar olmakla birlikte, standart kabul edilmiş bir derlem ya da belge kümesi henüz bulunmamaktadır. Türkçe belge madenciliği için ortaya attığımız yöntemi test edebilmek için 2000 yılından bu yada yayınlanan iş dünyası dergilerinden elde edilen makalelerden bir belge kümesi oluşturulmuş ve bu küme üzerinde sorgulama ve demetleme teknikleri kullanılarak testler yapılmıştır. Sorgulama testlerinde geleneksel GAD yöntemi, önerdiğimiz n-gram destekli GAD yönteminden geri kalmıştır. Benzer şekilde n-gram destekli GAD ile yapılan demetleme işlemi, geleneksel GAD yöntemini geride bırakmıştır.

Önerdiğimiz yöntem, Türkçe belgelerin madenciliği için kullanılmıştır. Bu amaçla bir Türkçe belge kümesi oluşturulmuştur. Ancak bu belge kümesi, uluslararası standart belge kümeleri gibi Türkçe için kabul edilmiş bir standart değildir. Bu nedenle elde edilen neticelerin değerlendirilmesinde, belge kümesinin yanlılığı gibi bir sebebe dayalı subjektiflikler olduğu iddia edilebilir. Bunu ortadan kaldırmak için aynı yöntem uluslararası kabul görmüş standart İngilizce Reuters21578 belge kümesine uygulanmıştır. Türkçe belge kümesinde elde edilen

sonuçlara paralel olarak Reuters21578 belge kümesi üzerinde yapılan sorgulama ve demetleme işlemleri başarılı neticeler vermiştir.

Anahtar Kelimeler : Veri Madenciliği, Belge İşleme, Bilgi Çıkarımı, Bilgi Erişim, Bilgi Arama

ABSTRACT

One of the first application areas of computer systems had been data collection and reporting. As data storage capacity and the computing power increased, it became possible to store and query larger amounts of data (Fayyad, 1996a). Based on the developments, it is now possible to find out relations and patterns in the data that were not easy to discover before. These techniques are known to be data mining techniques and are different than conventional techniques.Data mining is the analysis of (potentially large) data sets aimed at finding unsuspected relationships, patterns, rules, uncertainties and statistically important structures which are of interest or value to the database owners (Hand vd., 2001).

Data mining techniques are not directly suitable for analyzing document typed data like querying documents, searching for the relationships between documents. For specific document analyzing needs, techniques different than data mining techniques have been developed and this new discipline is known as text mining, document mining, semi-structured data mining.

The objective of document mining is to discover the content of documents by computers as if it is read by human beings. When this is the case, the language of the document becomes important. From this point of view there had been many studies under natural language processing field for decades.

For Turkish language, natural language studies are quite new and very few of them has resulted with useful outcomes. Moreover not all them are shared among researchers. Based on this fact, a document mining study for mining Turkish documents using NLP techniques was out of question, especially by 2004 when this study has started.

The study we are to present in this theses aims to mine Turkish documents by using Latent Semantics Indexing (LSI) technique and to develop LSI, we are proposing to enhance LSI by combining it with n-gram approach.

In order to compare and evaluate different document mining techniques on the international scale, a standart document set had been developed for a specific group of techniques like clustering, questing answering etc. This point is also a missing point for Turkish language. We had to develop our own document set for the study and collected articles from the business magazines that were published in Turkey from year 2000 to 2006. This document set was used to test our technique by doing document querying and clustering. For both document querying and clustering, the test results has shown that, n-gram based LSI technique outperformed conventional LSI.

To overcome the assertions that the document set we had collected is not a standard one as a result of which the test results may not show the reality, we tested our technique on the internationally accepted English document set known as Reuters21578. Parallel to Turkish tests, English document test also showed the same results both for document querying and clustering.

Keywords : Data Mining, Tezt Categorization, Text Retrieval, Information Retrieval, Querying

xiii

1. VERİ MADENCİLİĞİ

Temel olarak dört alanda meydana gelen gelişmeler veri kaynaklarını bulma, onları sorgulama, sorgulanan verileri derleme ve verilerden bilgi üretmek için analiz yapma konusunda yeni fırsatlar ve araştırma konuları doğmasına neden olmuş ve olmaktadır (Zaiane, 2001):

1. **Veri işleme ve saklama** alanındaki teknolojik gelişmeler gün geçtikçe artmakta ve daha fazla veriyi daha kısa sürelerde işlememize olanak sağlamaktadır.
2. **Bilgisayar kullanımı** üçüncü dünya ülkeleri de dahil, yıllar içerisinde artmakta ve gün geçtikçe daha fazla kişi daha fazla dijital ortamda çalışmaya başlayarak daha fazla dijital veri üretmektedir.
3. **İletişim teknolojileri ve internet** gibi altyapılar hızla tüm dünyayı sarmakta, yer ve zamandan bağımsız bir yaşam şekli gelişmektedir.
4. İnsanlar daha hızlı ve doğru karar almak için **veriye dayalı bir araştırma, inceleme ve muhakeme kültürünü** benimsemektedir.

Bu gelişmeler belirli bir tarihi süreç içinde cereyan etmiştir. Bilgisayarların ilk kullanımı kısa süreli amaçlara yönelik verilerin saklanması ve raporlanması olmuştur. Bu ihtiyaç halen günümüzde de önemini korumaktadır. Bu gereksinimin hızlı ve etkin bir şekilde karşılanabilmesi için ilişkisel veritabanları ve yapılandırılmış sorgu dili (SQL) geliştirilmiştir. Zaman içinde daha geniş tarihsel süreci kapsayan veriler üzerinde çalışmak mümkün hale gelmiştir. Bu dönemde geçmiş veriler arasındaki ilişkileri ortaya çıkaran ve bu sayede geleceğe dönük tahminler yapılabilmesini sağlayan teknikler yaygınlaşmaya başlamıştır. Bu dönemin başında kullanılan yöntemler geleneksel veritabanları ve yapılandırılmış sorgu dili temelli olmakla birlikte, ihtiyacı tam olarak karşılayamadığı için zamanla tamamen bu amaca yönelik veri depolama teknolojileri ve sorgulama teknolojileri ortaya çıkmaya başlamıştır. Veri ambarları, çevrimiçi analitik işleme (OLAP) yöntemleri ve araçları bu dönemin ihtiyaçlarını karşılamak üzere geliştirilmiştir.

Zamanla tarihsel süreçte biriken verilerin yanında daha önce toplanmayan verilerin de toplanmaya başlaması ile veri yığınları büyümeye başlamıştır. Örneğin eskiden süper markette basit anlamda satış verisi toplayan kasa üzerinden müşterinin o anda satın almış

olduğu malların toplamı verisi toplanırken, günümüzde kasa yerine kullanılan satış noktası terminalleri sayesinde müşterinin yaptığı alışveriş işlemine ait bütün detay veriler toplanabilmektedir. Binlerce müşteriye ait binlerce ürün satış verisi sayesinde her ürünün zaman içindeki hareketlerine ve müşterilerin hangi ürünleri, ne kadar miktarlarda, ne zaman, hangi ürünlerle birlikte aldığı verilerine ulaşmak ve analiz etmek mümkün olabilmiştir. Benzer şekilde ticari, tıp, askeri, iletişim, vb. birçok alanda bilgisayar sistemlerinin gelişmesi ve yaygınlaşmasına bağlı olarak ortaya çıkan veri hacminin yaklaşık olarak her yirmi ayda iki katına çıktığı tahmin edilmektedir (Frawley, 1991). Bir başka çarpıcı örnek ile verilerin ne kadar hızlı toplandığını ve işlemesinin imkansız bir noktaya geldiğini görmek için NASA'ya bakmak yeterlidir (Fayyad, 2000). NASA'nın kullandığı uyduların sadece birinden, bir günde terabayt'larca veri gelmektedir.

Bütün bu gelişmeler neticesinde büyük veri yığınlarından, daha önce elde etmenin mümkün olmadığı ve keşfedilmeyi bekleyen bilgileri açığa çıkarma işi altında bir disiplin olarak veri madenciliği doğmuştur.

Veri bilgisayar ortamında depolanma biçimlerine göre üç farklı türdedir (Baschab, 2003):

1- yapısal veri
2- yarı yapısal veri
3- yapısal olmayan veri.

Yapısal veri, veritabanı ve veri ambarlarında tutulan ve SQL, OLAP sorgulama yöntemleri sorgulanabilen veri türünü ifade eder. Yarı yapısal veriler ise belgelerdir. Belgelerin kim tarafından, hangi konuda, ne zaman yazıldığı gibi bazı yapısal kısımları olmakla birlikte, bir belgenin içeriğinin tam olarak anlaşılması ancak bir insan tarafından okunması ile ortaya çıkarılabilir. Hangi konuda yazılmış ve konunun hangi bölümünü, hangi bölümlerle yada başka disiplinlerle kıyaslamaktadır gibi içeriği oluşturan bir detay bulunmaktadır. Yapısal olmayan veri ise ses ve görüntü gibi akan veridir. Günümüzde güvenlik videoları, uydu gözlem istasyonları gibi kanallardan büyük miktarlarda görüntü verisi toplanmaktadır. Pek çok ticari kuruluş, çağrı merkezlerinde müşteri ile yaptıkları görüşmelerin ses kayıtlarını tutarak önemli ölçüde ses verisi biriktirmektedir.

Her veri türünün saklanma ve erişimi için özel geliştirilmiş ortamlar vardır. Veri tabanlarında ve veri ambarlarında yapısal veriler dururken, belgeler için belge yönetim sistemleri yada

içerik yönetim sistemleri, ses ve görüntü verileri için ise optik diskotek (jukebox) gibi özel donanım ve yazılım çözümleri mevcuttur.

Veri madenciliği, büyük veri yığınları içinde bilgi keşfi süreci olduğundan, her veri türü içinde bilgi keşfi amacına yönelik uygulanabilecek genel yaklaşımlar bulunmaktadır. Bunlar kısaca (Hand vd., 2001)

1- Sınıflandırma : Veriler daha önceden belirlenmiş sınıflara dağıtılır.
2- Demetleme : Veriler, kendi içindeki niteliklere göre veri madenciliği algoritması tarafından otomatik oluşturulan kümelere dağıtılır.
3- İlişkilendirme : veriler arasındaki neden-sonuç ilişkisi ortaya çıkarma.
4- Zaman serisi analizi : veri içinde dönemsellik gibi örüntüleri çıkarma.

Tezin konusu belge madenciliği olmakla birlikte, belge tipinde bir veri yığını içinde bilgi keşfi yapmak için kullanılacak tekniklerin veri madenciliği disiplininden gelmesi ve işin temelini oluşturması nedeniyle veri madenciliği konusunu anlamak önemlidir. Bundan sonraki bölümlerde veri madenciliğinin kökenleri, diğer disiplinlerle arasındaki ilişkiler, veri madenciliğinde uygulanan temel süreçler, temel veri madenciliği yöntemleri ve veri madenciliği altında gelişen araştırma konuları aktarılacaktır.

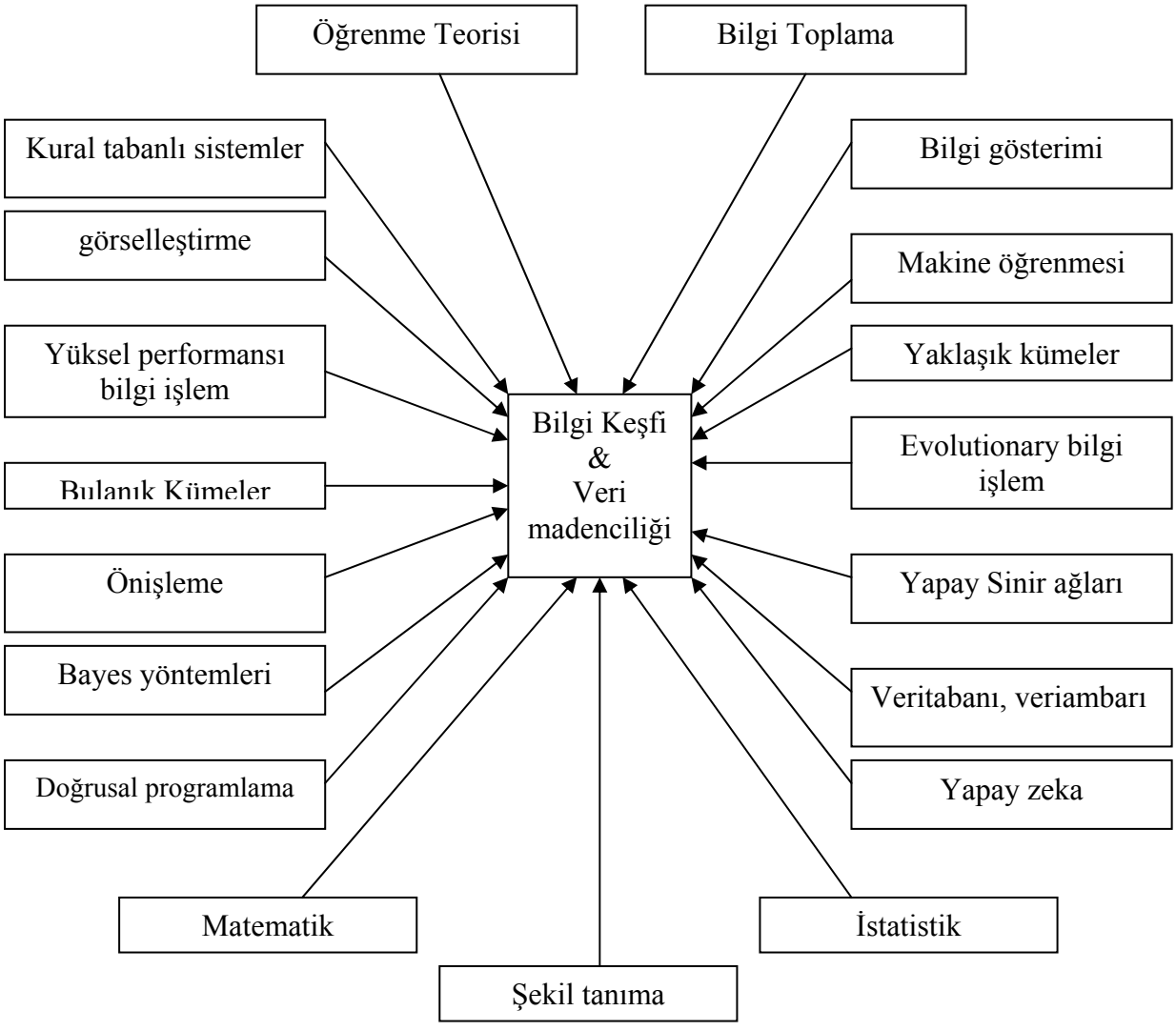
1.1 Veri Madenciliğinin Kökenleri

Veriye dayalı analiz pozitif bilimlerin gelişimi ile ortaya çıkmıştır. Bilimsel araştırmaların yapıldığı tüm disiplinlerin temelinde bilimsel araştırma süreci ve bilimsel bilgi edinme süreci olarak veriye dayalı analiz kullanır. Bilimsel araştırma sürecinin adımları şunlardır [8] :

1. Karşılaşılan sorunun belirlenmesi
2. Gereksinim duyulan verinin belirlenmesi
3. Veri kaynaklarının saptanması
4. Verinin derlenmesi için kullanılacak tekniklerin kararlaştırılması
5. Verinin derlenmesi ve işlenmesi
6. Çözümleme sonuçlarının yorumlanması
7. Sonuçların karar vericilere iletilmesi

Bu sürece bağlı olarak araştırmalarını yürüten araştırıcılar, araştırma konularıyla ilgili verileri, değişik kaynaklardan toplayıp, bunları düzenleyip daha sonra bu veriler arasındaki ilişkileri inceleyerek tezler oluşturmuşlar ya da oluşturdukları tezleri ispatlamaya çalışmışlardır.

Bu süreç, veri madenciliği yada ilk zamanlarda kullanılan ismiyle, bilgi keşfi süreciyle örtüşmektedir. Ancak veri madenciliğini özel yapan bilgisayar sistemlerinde ve internet'te meydana gelen gelişmelerin etkileridir. Geçmişte dünyanın değişik coğrafyalarında, çoğunlukla basılı formda bulunan verileri aramak, bulmak, bir araya getirmek ve derlemek için insana dayalı ciddi bir zaman ve emek gerekirdi. Bugün bilgisayar kullanımının yaygınlaşması ile basılı formlardaki her verinin bir dijital karşılığı bulunmaktadır. Dijital formdaki verinin sorgulanması, bulunması ve taşınması internet gibi altyapılar sayesinde saatler mertebesine düşmüştür ve toplanan verileri saklamak ve oluşturulan veri yığınları üzerinde işlem yapmak için yeterli teknoloji hazırdır.



Şekil 1-1 Veri Madenciliği Konusuna Etki Eden Disiplinler

Görüldüğü üzere bugün ayrı bir araştırma konusu haline gelen veri madenciliği çalışması uzun yıllardan beri pek çok disiplin içinde, kısıtlı amaçlar için kullanılmıştır ama birkaç farkla.

- bu disiplinlerdeki araştırmaların kapsamında gereken verinin miktarı büyük değildir
- veriler çoğunlukla sınırlı konuda ve az boyutludur.

Bu farka rağmen veri madenciliği çalışmalarında kullanılan teknikler ve algoritmalar bu disiplinler tarafından bulunmuş ve kullanılmıştır. Şekil 1-1 'de veri madenciliği konusuna etki eden disiplinler gösterilmiştir (Bernstein vd., 2002) (Faloutsos vd., 1997).

Başlangıç itibariyle veritabanlarındaki veriler üzerinde yürütülen çalışmalar zamanla veritabanında tutulmayan verileri de kapsayacak şekilde genişlemiştir.

1.2 Veri Madenciliğinin Gelişimi

1.2.1 Veritabanlarında Veri Madenciliğinden Veri Keşfine

Veri içindeki yararlı kalıpları (örüntüler) bulma çalışmaları pek çok farklı inceleme alanı altında uzun yıllardan beri farklı isimlerle yapılmaktadır:

- Veri madenciliği
- Üst bilgi Çıkarma (Knowledge Extraction)
- Bilgi keşfi (Information Discovery)
- Bilgi Hasatı (Information Harvesting)
- Veri arkeologluğu (Data Archaelogy)
- Veri Örüntü İşleme (Data Pattern Processing)

İngilizce'de veri için, taşıdığı anlam ve değere göre üç yaygın kelime vardır. Bunlar anlam ve değer sırasına göre küçükten büyüğe "data", "information" ve "knowledge" kelimeleridir. Bu kelimelere karşılık olarak Türkçe'de "veri", "bilgi" ve "üstbilgi" kelimeleri kullanılacaktır.

Veri madenciliği daha ziyade istatikçiler, veri analistleri, ve yönetim bilişim sistemleri alanında çalışanlar tarafından kullanılıyordu. Bunların dışında veritabanı konusunda çalışanlar

için de bu terim bilinen bir terimdi. Veritabanlarındaki verilerin incelenmesi ile üstbilgi keşfi işlemi ilk kez 1989 yılında "veritabanlarından üstbilgi keşfi-VÜK" (VÜK- Knowledge Discovery in Databases) isimli bir çalışma grubu oluşturularak konuşulmaya başlandı (Fayyad vd., 1996a) (Fayyad vd., 1996b) (Fayyad vd., 1996c). Bu çalışma grubu, üstbilgi keşfinin, veritabanlarındaki inceleme işleminin nihayi sonucu olması gerektiğini vurgulayarak sonraki çalışmalara yön verecek şekilde konuyu araştırmacıların dikkatine sunmuş oldu.

Verilerden üstbilgi oluşturma yapay zeka ve onun bir alt kolu olan makine öğrenmesi (machine learning) alanlarında da araştırılan bir konuydu. Hatta 1989'daki ilk çalışma grubundan önce yapay zeka ve makine öğrenmesi alanında çalışanlar, öğrenen bir sistem kurmak için çalışmalar yapıyordu ve bu öğrenen sistemin çalışma prensipleri kendisine verilen verilerden yola çıkarak ilişkiler yakalama alanında yoğunlaşıyordu (Fayyad vd., 1996a).

VÜK adımlarından biri olan veri madenciliği büyük ölçüde istatistik, makine öğrenmesi ve örüntü bulma disiplinlerinde yaygın olarak kullanılan tekniklere dayanmaktadır. Bu açıdan bakılarak, VÜK ve diğer disiplinler arasındaki farkın, kapsamı olduğu görülmektedir. VÜK diğer disiplinlerdeki teknikleri kullanır ama veriden üstbilgi elde etme sürecinin tamamı ile ilgilenirken diğer disiplinler sadece VÜK'ün veri madenciliği adımında kullanılan yöntemleri kullanır (Fayyad vd., 1996a)(Fayyad vd., 1996c).

VÜK anlaşılır örüntüler bulunmasıyla ilgilenir, bu nedenle örneğin yapay sinir ağları yerine karar ağacı tekniği tercih edilir çünkü yapar sinir ağlarının ürettiği üstbilginin nasıl üretildiği yapay sinir ağının içinde gizlidir ve incelenemez. VÜK' ün diğer bir özelliği büyük veri kümeleri üzerinde çalışmaya odaklanması ve bu veri kümelerindeki verilerin eksik, bozuk olmak gibi taşıdığı kirliliği dikkate almasıdır. Diğer bir kritik konu da bir veri kümesinin üzerinde uzun süre çalışıldığında doğru olmayan kalıplar bulunabileceği gerçeğidir ve VÜK bu gibi yanlışlıklara düşülmesini engelleyici prensipler üzerine kurulmuştur (Fayyad vd., 1996a).

VÜK konusundaki çalışmaları sürükleyen diğer bir konu da veritabanlarıdır. Veritabanı alanında yapılan çalışmalar ve ortaya çıkarılan teknikler ve yöntemler VÜK çalışmaları için zemin hazırlamış ve hazırlamaktadır. Örneğin veri ambarı oluşturma yöntemi ile operasyonel işlemlerden elde edilen verilerin temizlenmesi (yanlışların düzeltilmesi, eksikliklerin

giderilmesi vs) ve depolanması kavramı ortaya çıkmıştır. Bu sayede bu verilen çevrimiçi analizi mümkün olmuş, bu verilerden bilgi çıkarmak ve karar destek sistemlerinin hayata geçirilmesi sağlanmıştır. Veri ambarı ortamı VÜK sürecinde kullanılabilecek bir veri temizleme ve veriye erişim ortamı sağlamış olur. Biraz daha detay incelersek, veri temizleme, organizasyonların sürekli artan verilerini tek bir sistemde tutmak isteme felsefesi gereği verilerin tek bir sisteme çevrimi ve eksik ve kirli verilerin düzeltilmesi ile ilgilenir. Veriye erişim ise birleştirilmiş ve tanımlanmış veriye erişim yöntemlerini kullanarak, eski ve yeni verilerden oluşan büyük veri kümelerine hızlı ve kolay erişimle ilgilenir. Veriler temizlenip, erişilebilecekleri formatta hazır hale getirildiğinde, bu kaynaktan nasıl yararlanılacağı önem kazanır. Bu noktada veri analizi gündeme gelir. Veri ambarlarındaki verinin analizinde kullanılan en popüler yöntem "Çevrimiçi Analitik İşleme" (OLAP – Online analytical Processing) yöntemidir. Veritabanlarındaki verilere erişim için kullanılan SQL'den (Structured Query Language) farklı olarak OLAP çok boyutlu veri analizini mümkün kılmaktadır. Özellikle özet bilgiler ve çok değişkenli ortamlarda veri kırılganlıklarını ortaya çıkarabilen bir yaklaşım sunar. OLAP araçları ile VÜK arasındaki fark ise, OLAP araçlarının veri analizi sürecini basitleştirmek ve etkileşimli hale getirmek gibi bir amacı varken, VÜK bu süreci olabildiğince otomatize etmeye çalışmaktadır, yani VÜK, OLAP'a göre bir adım öndedir (Fayyad vd., 1996c) (Zaki vd., 2003).

VÜK veriden yararlı üstbilgi oluşturma sürecidir. Bu süreç içinde birden fazla adım vardır ve veri madenciliği bu adımlardan biridir. Veri madenciliği kelimesi veriden örüntüler çıkartmak için belirli algoritmaların uygulanmasını ifade eder. Sadece veri madenciliği adımını uygulayarak veriden çıkarılan örüntüler çoğunlukla geçersiz ve anlamsız olmaktadır. Bu çıkarımları anlamlı hale getirmek için VÜK içinde veri hazırlama, veri seçme, veri temizleme, tecrübe yada birikimin bütünleştirilmesi ve sonuçların yorumlanması gibi adımlar bulunmaktadır (Fayyad vd., 1996a).

VÜK sürecinde verinin nasıl saklandığı, ona nasıl erişildiği, algoritmaların veri büyüklüğüne karşı nasıl ölçeklendiği ve verimli çalıştığı, sonuçların nasıl yorumlanabileceği ve görselleştirilebileceği ve bilgisayar-insan etkileşiminin nasıl başarılı bir şekilde modellenip desteklenebileceği de incelenmektedir.

Özetlemek gerekirse, VÜK aslında çok disiplinli bir yaklaşımdır. Aşağıda ismi verilen pek çok disiplinin kesişiminde yer alır (Fayyad vd., 1996a).

- Makine Öğrenmesi (Machine Learning)
- Örüntü Bulma (Pattern Recognition)
- Veritabanları (Databases)
- Yapay Zeka (Artificial Intelligence)
- Uzman Sistemler için Üstbilgi Öğrenme (Knowledge Acquisition for Expert systems)
- Veri Görselleştirme (Data visualisation)
- Yüksek Performanslı Bilgi İşlem (High Performance Computing)

Bu disiplinlerin ortak hedefi, alt seviye veriye dayanarak üst seviye üstbilgiyi elde etmektir.

1.2.1.1 Temel Tanımlar

VÜK, veriden, geçerli, potansiyel olarak yararlı, daha önceden bilinmeyen, gizli ve neticede anlaşılabilir örüntülerin bulunması için kullanılan detaylı bir süreçtir (Fayyad vd., 1996a).

Bu tanımdaki veri gerçekler kümesini, örneğin veritabanındaki değişik durumlardan herbirini ifade eder. Örüntü kelimesi ise verinin bir alt kümesini tanımlayan veya bu alt kümeye uygulanabilir bir modelin, bir çeşit gösterim dili ile ifadesi için kullanılmıştır. Başka bir değişle, veriden bir örüntü çıkarmak demek, veriye bir model uydurma, veriden bir yapı çıkarma veya bir veri kümesinin alt seviyede tanımlanması demektir.

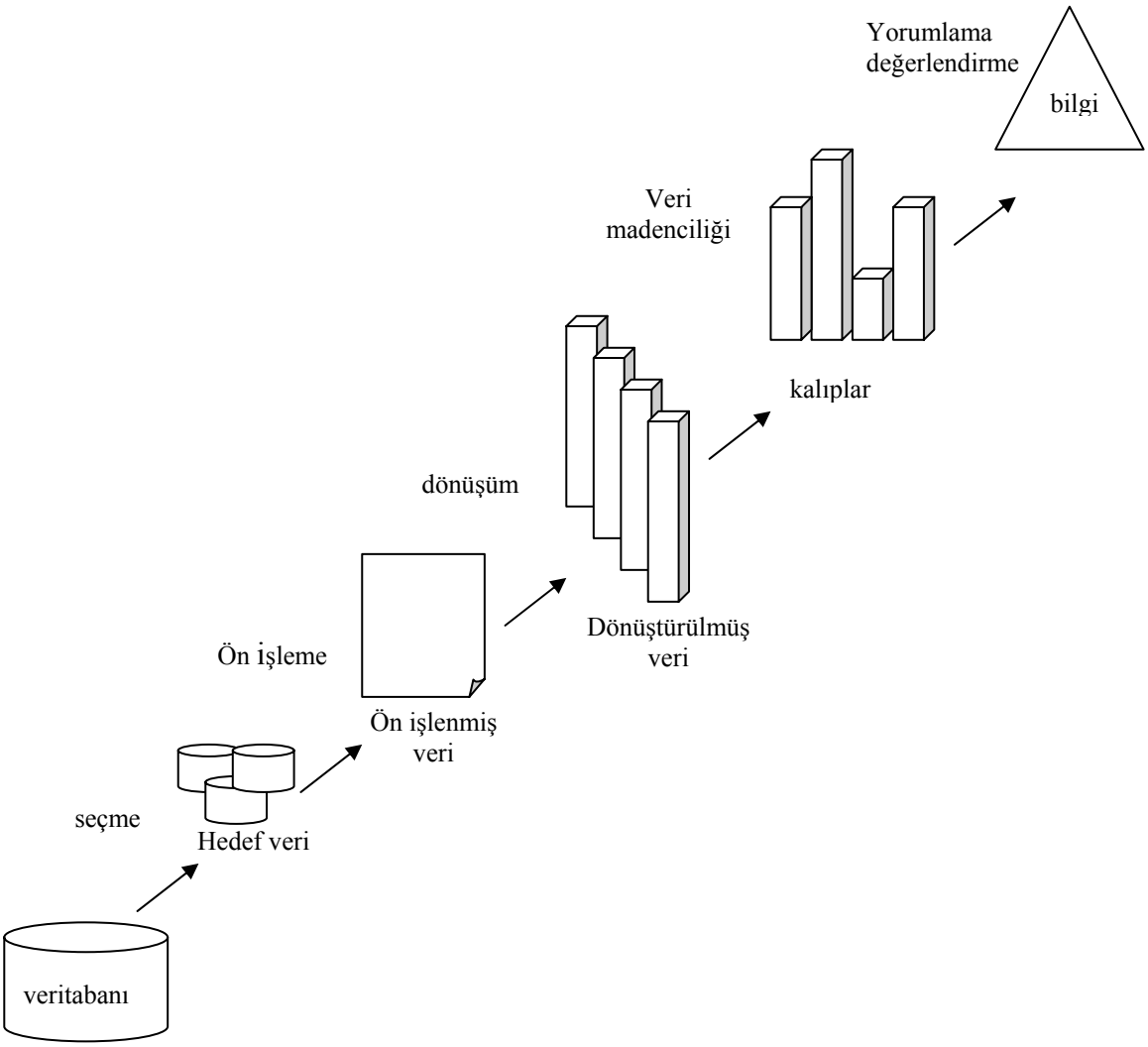
VÜK'nin birçok adımdan oluşan süreç özelliği vardır. Veri madenciliği adımında veriye veri analizinin uygulanması, veriden en önemli örüntülerin önem sırasında göre çıkarılması ve bu çıkarımı yapacak kabul edilebilir performansa sahip algoritmaların üretilmesi işlemleri yer alır. Veriden çok sayıda örüntü çıkarılabilir. Bir örüntünün değerini ifade etmek için kullanılabilecek ölçümler, yararlılık, anlaşılabilirliktir. Bir başka ölçüm 1995 yılında Silberschatz tarafından önerilmiştir (Hand vd., 2001) : "ilginçlik". Buna göre örüntülerin önem sırasını belirleyebilmek için bir ilginçlik fonksiyonu tesbit edilebilir. Bu fonksiyonun içinde kullanılan bir eşik değeri, örüntüleri ilginç olup olmamalarına göre ayırt ederken, ilginçlik değeri yüksek olanlar da derecelerine göre sıralanmış olurlar.

1.2.1.2 VÜK süreci

Şekil 1-2'den de görüldüğü gibi VÜK süreci kullanıcı tarafından bir çok kararın verildiği, çeşitli adımlardan oluşan, etkileşimli ve döngüsel bir süreçtir (Fayyad vd., 1996a).

1. Adım : İlk aşamada üzerinde çalışılacak konunun anlaşılması, bu konuyla ilgili mevcut bilgilerin elde edilmesi ve hedef açısından bu sürecin amacının ne olduğunun belirlenmesine ihtiyaç vardır. VÜK süreci ile veriden pek çok örüntü çıkarmak mümkündür. Örüntülerin değerlendirilmesi ve gereksiz çalışmayı engellemesi açısından bu adımda alınacak kararlar önem taşır.
2. Adım – Seçme : Üzerinde araştırma ve analiz yapılacak veri kümesinin, veri örneklerinin ve/veya değişkenler alt kümesinin tespitinin yapıldığı adımdır. Eğer veritabanındaki tüm veriler ve değişkenler (kolonlar) kullanılırsa hem yanlış kalıplara ulaşmak riski vardır hem de gereksiz veriler daha büyük işlem gücü gerektirir.
3. Adım - Veri Temizleme ve Ön İşleme : Bu adımda verideki kirlilik giderilir. Bu kirliliğin giderilmesi için kullanılacak model yada işlemler, eksik sahaların, zaman serisi verisinin ve verideki zamana bağlı değişimlerin nasıl ele alınacağı stratejileri tespit edilir.
4. Adım - Verinin daraltılması ve veri projeksiyonu : Bu aşamada amaca yönelik olarak veriyi ifade eden yararlı özelliklerin tespiti yapılır. Verinin boyutsal olarak küçültülmesi ve dönüştürme yöntemleri sayesinde işlenecek değişken sayısı (kolon) azaltılır, verinin kolay işlenmesini sağlayacak veri ifade şekilleri bulunur.
5. Adım – Veri madenciliği yönteminin tespiti : ilk adımda belirlenen VÜK amacına uygun veri madenciliği yöntemi yada yöntemleri tespit edilir. Örnek yöntemler : özetleme, sınıflandırma, eğri uydurma (regresyon), demetleme vs.
6. Adım – Model ve hipotez seçimi : Bu adımda detaylı analiz yapılarak veriye uyacak model ve amaca uygun hipotez seçimi yapılır. Kullanılacak veri madenciliği algoritması yada algoritmaları, örüntü bulmak için kullanılacak arama yöntemi belirlenir. Uygun model ve parametrelere (örneğin kategorik veri tabanlı modeller, amaca yönelik olarak değerlendirildiğinde gerçel sayı tabanlı vektörlere göre daha kullanışlı olabilirler), ulaşılmak istenen hedefe uygun yöntemlere (örneğin amaç modelin neticesini değil, kendisini anlamak olabilir) karar verilir.

7. Adım – Örüntü bulma : Bu adımda eğri uydurma (regresyon), sınıflama kuralları, sınıflama ağaçları, demetleme gibi veri madenciliği tekniklerinin uygulanması ile veri içinde örüntüler aranır.
8. Adım – Örüntü yorumlama : Bir önceki aşamada bulunan örüntülerin yorumlanması bu adımda gerçekleştirilir. Bu yorumlama çalışması VÜK sürecinin bu aşamaya kadar olan aşamalarından birine geri dönerek tekrar sürecin işletilmesini de içerir. Bu sayede yorum güçlendirilmeye çalışılır. Bu adım ayrıca çıkarılan örüntülerin ve modellerin görselleştirilmesini veya örüntünün ortaya çıkarılmasında etkin olan değişkenlerin ve verilerin görselleştirilmesini de kapsayabilir.
9. Adım – Bilginin kullanılması : Elde edilen bilgi doğrudan yararlı olabilir ya da yüksek seviyede bir bilginin elde edilmesi için kullanılabilir veya sadece dokümante edilerek saklanabilir. Eğer daha önceden benzer amaçla yapılmış çalışmalar varsa, yeni bilgi ile bunlar karşılaştırılarak bilginin geçerliliği artırılabilir ya da sorgulanabilir.



Şekil 1-2 Veritabanlarından Üstbilgi Keşfi

1.3 Veri Madenciliği Yöntemleri

Birçok veri madenciliği tekniği makine öğrenmesi, örüntü bulma ve istatistik disiplinlerinde denenmiş ve test edilmiş tekniklere dayanır. Bunlar sınıflama, demetleme, eğri uydurma (regresyon) gibi tekniklerdir. Uzman ya da yeni başlamış pek çok veri analizcisi için bu teknikleri hayata geçirmek için kullanılan algoritmaların sayısı şaşırtıcı derece de yüksektir. Aslında bu yöntemlerin bir çoğu temelde birkaç tekniğe dayanır. Bunlar polinomlar, splines, çekirdek ve temel fonksiyonlar, aşık-boolean fonksiyonları gibi tekniklerdir. Genelde benzer teknikleri baz aldıkları için, algoritmaları değerlendirmede modelin veriye ne kadar uyduğu, uygun modelin aranması için kullanılan arama yönteminin karmaşıklığı gibi kriterler kullanılır (Hand vd., 2001).

Bu amaçlara ulaşmak için kullanılan belirli yöntemler (Hand vd., 2001) :

- Sınıflandırma (classification) : Daha önceden belirlenmiş sınıflara veriyi yerleştirmek için kullanılacak fonksiyonun öğrenilmesi işlemini kapsar.
- Eğri Uydurma (regression) : Gerçek hayatta iki değişken arasındaki ilişki fonksiyonunun öğrenilmesini ve bu sayede bir değişkenin değerine bakarak diğerinin değerinin bulunmasını – tahmin edilmesini – kapsar
- Demetleme (clustering) : Verilen bir veri kümesini tanımlamak için sonlu sayıda sınıfa ya da demete bölmeyi kapsar. sınıflar birbirinden ayrı ve detaylı veya hiyerarşik ya da örtüşen niteliklerde olabilir. Demetleme işlemi esnasında veri tabanındaki değişkenlerin (sahaların) olasılık yoğunluk tahminin (probability density function) yapılması önem taşımaktadır. Bu fonksiyonun kestiriminde kullanılan teknikler de vardır. Bu teknikler olasılık dağılımı kestirimi (probability density estimation) adıyla tanımlanırlar.
- Özetleme (summerization) : veritabanındaki verinin tamamını ya da bir alt kümesini topluca tanımlayabilecek ifadeyi bulmak için kullanılan yöntemdir. Bu yöntemin en basit hali, tüm değişkenlerin ortalama ve standart sapmalarının tablolaştırılmasıdır. Bunun dışında daha detaylı yöntemler de vardır:
    - Özet kurallar türetme
    - Çok değişkenli görselleştirme teknikleri

- Değişkenler arası fonksiyonel ilişkilerin keşfi vs.

Özetleme teknikleri özellikle etkileşimli, araştırma amaçlı (exploratory) veri analizinde ve otomatik rapor türetmede tercih edilir.

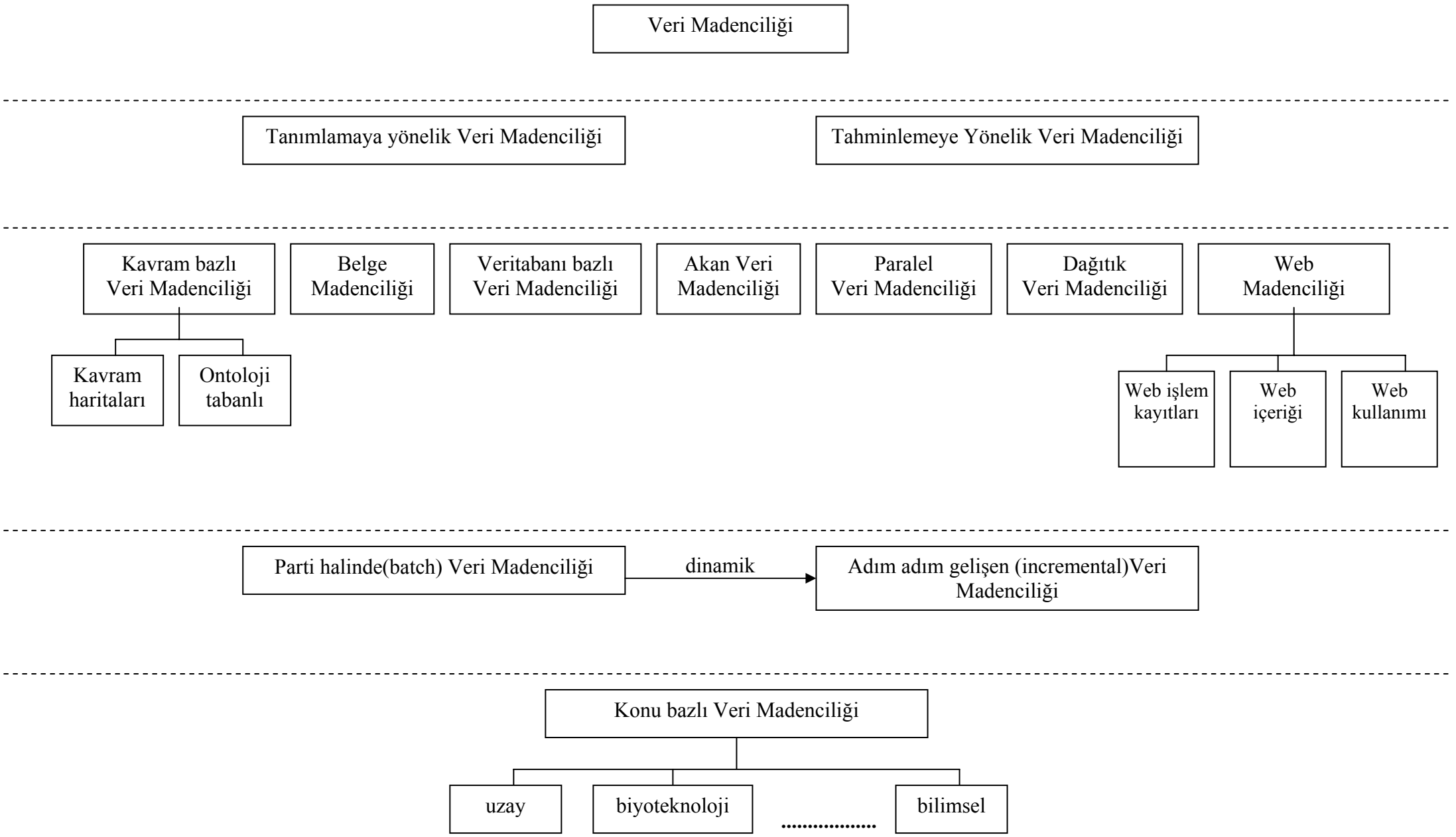
- Bağımlılık Modelleme (Dependency Modelling) : Veritabanındaki değişkenler arasında anlamlı bağımlılıklar arayan bir yöntemdir. Bu amaçla oluşturulan modelleri iki seviyede gruplamak mümkündür:
    - Yapısal Seviye : Genelde grafik olarak gösterilen modelin yapısal seviyesi, yerel olarak hangi değişkenlerin birbirine bağımlı olduğunu gösterir.
    - Nicel Seviye : değişkenler arasındaki bağımlılığın kuvvetliliğini bir nümerik ölçeğe dayanarak ifade eden modellerdir.
- Değişim ve Sapma Bulma (Change and Deviation Detection) : veritabanındaki veriler tarihçe olarak ele alındığı ve bu değişkenlerin gelecekteki olası değişim ve sapmalarının tahmin edilmeye çalışıldığı yöntemdir.

1.4 Veri Madenciliğinin Genişlemesi

Veri madenciliği yöntemlerinin değişik alanlara, farklı amaçlarla, farklı veri tiplerine uygulanacak şekilde, farklı yaklaşımlarla uygulanması sonucu kullanımının artması, bu alanın genişlemesini sağlamıştır. Şekil 1-3'de bu gelişimin getirdiği genişlemeyi görmek mümkündür (Bernstien vd., 2002) (Zaki vd., 2003) (Faloutsos vd., 1997) (Pena vd. 2001) (Cannataro vd., 2002) (Berkan vd., 2002) (Honavar vd., 2001).

Bu haritadan da anlaşılacağı gibi veri madenciliği çalışmalarını birden fazla katmanda algılamak doğru olacaktır. Bütün kullanım alanlarında iki amaç mutlaka bulunmaktadır (Hand vd., 2001).

1. Tanımlama : incelenen veri kümesinin tamamı için ya da bu veri kümesindeki bir alt küme için geçerli bir ilişkinin tanımlanmasını sağlayan tekniklerdir ve genelde daha önceden bilinmeyen ilişkiler üzerinde durduklarından önemlidirler.
2. Tahmin etme : Eldeki veri kümesinin incelenmesiyle, veri kümesinin kapsamındaki konuyla ilgili gelecekte neler olabileceğini tahmin etmeye yarayan tekniklerdir.



Şekil 1-3 Veri Madenciliğinin Genişlemesi

1.5 Belge Madenciliğine Geçiş

Veri madenciliğinin ilk uygulama alanları doğası gereği büyük veri kümeleridir. Örneğin uzaydan uyduların gönderdiği fotoğrafların incelenmesinde. O kadar çok fotoğraf gelmektedir ki bunları insan gücü ile incelemek olanaksızdır. Şekil tanıma, sınıflandırma ve kümeleme gibi teknikler bu alanda çok yaygın kullanılmaktadır. Meteorolojik veriler, gen araştırmalarında oluşturulan veriler aynı şekilde doğal olarak bilgisayar teknoloji kullanılmadan işlenemeyecek büyüklüktedirler. Konuya özel veri madenciliği çalışmalarının doğmasına neden olan bu gelişmeler göstermiştir ki veri madenciliğinden istifade edebilmek için konuya özel detayları içeren algoritmaların kullanımı gerekmektedir (Hand vd., 2001).

Veri madenciliği çalışmalarını verinin nerede olduğu, biçimi ve niteliği gibi etkenler de etkilemektedir. Özellikle ilişkisel veritabanlarındaki verilerin analizi için kullanılan teknikler artık web sayfalarının, yarı yapısal tabir edilen belgelerin analizi gibi alanlarda kullanılmakta ve değişik yaklaşımların doğmasını sağlamaktadır.

Veri yığınları niteliklerine göre üç grupta sınıflandırılır (Baschab, 2003).

- Yapısal
- yarı yapısal
- Yapısal olmayan

İstatistiklere göre üretilmiş verilerin %80'i yarı yapısal niteliktedir. Yarı yapısal verilerden kasıt içerisinde metin, resim, grafik vs olan belgelerdir. Belgeler pek çok bilgi işçisi tarafından belirli konulara özel olarak üretilen ve çoğunlukla birkaç kişi tarafından incelenmiş ve kişilere ait bilgisayarda muhafaza edilen verilerdir. Pek çok farklı formatta olabilir. Sade metin, Adobe Acrobat, MS Word, HTML, XML vs. ve internet üzerinden http veya ftp protokolü vasıtasıyla erişilebilir durumda olabilirler. Bu tip verilerin büyük çoğunluğu sık sık değişebilir. Bu sınıftaki verileri analiz etmek çok değerli ilişkilerin yakalanmasını sağlayabilir. Bu yöndeki gelişimlerin bir kanıtı olarak 2003 yılında yapılan uluslar arası verimadenciliği konferansının kapsamı Tablo 1-1'den incelenebilir.

Tablo 1-1 2003 Uluslar Arası Verimadenciliği Konferansı Konu Listesi

Metin verimadenciliği – Uluslar arası veri madenciliği konferansı 2003		
Algoritmalar	Bayesian Models	Bayes Modelleri
	Concept Decomposition	Kavram Ayrıştırma
	Orthogonal Decompostiton	Ortogonal Ayrıştırma
	Probabilistic Models	Olasılık Modelleri
	Vector Space Models	Vektör Uzay Modelleri
	Latent Semantic Indexing	Gizli Anlambilimsel Dizinleme
	Graph-based Models	Grafik tabanlı Modeller
	Text Streaming Models	Metin Serisi Modelleri
Uygulamalar	Clustering	Demetleme
	Factor Analysis	Faktör Analizi
	Visualization Techniques	Görselleştirme Teknikleri
	Metadata Generation	Yardımcı veri Üretme
	Text Classification	Metin Sınıflama
	Text Purification	Metin Netleştirme
	Text Segmentation	Metin Bölümleme
	Text Summarization	Metin Özetleme
	Query Structures	Sorgu Yapıları
	Trend Detection	Eğilim Bulma
	Distributed Storage and Retrieval	Dağıtık depolama ve Erişim

2. METİNLERİN ANALİZİ İLE BİLGİ OLUŞTURMA

Özellikle Internet ve kişisel bilgisayarların yaygınlaşmasına bağlı olarak, gittikçe büyüyen hacme sahip belge yığınları oluşmaktadır. Önemli bilgiler bu yığınlar içinde kaybolup gittiğinden, değerli bilgilere ulaşmak için belgelerin içeriğinin bilgisayar ortamında otomatik olarak belirlenmesi ve buna uygun sorgulanabilmesi ihtiyacı kendini hissettirmektedir.

Bilgisayarlı sistemlerden önce, belgeler içindeki bilgiye erişim için elle indeksleme sistemleri kullanılıyordu. 1996-2001 yılları arasında yayınlanan 164.000 periyodik yayın olduğu düşünülürse, elle indekslemenin ne denli yavaş, zor ve yetersiz olduğunu anlaşılabilir (Fayyad vd., 1993). Şu an Internet'te 2 milyardan fazla web sayfası bulunmaktadır (Fayyad vd., 1995). Bu kadar çok belgenin elle indekslenmesinin ve bu indekslerin güncellenmesinin zorluğunun yanı sıra, elle yapılan indeksleme uzmanların subjektif yorumlarını da içerdiğinden, aranılan bilgilere ulaşmayı kolaylaştıramamanın ötesinde, yanıltıcı sonuçlara da neden olabilmektedir.

Giderek artan belge yığınlarının faydalı bilgiye dönüştürülmesini sağlamak için bilgiye ulaşım (information retrieval) ve bilgi çıkarımı (information extraction) olarak iki alanda yöntemler geliştirilmektedir. Bu yöntemler yarı-yapısal verilerin madenciliği, belge madenciliği yada metin madenciliği adı altında incelenmektedir.

2.1 Metin Veri Madenciliğinin Gelişimi

Pek çok araştırma alanında ve günlük hayatın içinde üretilen bilgiler ağırlıklı olarak metin belgeler şeklinde oluşturulur, bu belgeler taraflar arasında gönderilir, farklı birikimlere sahip kişiler tarafından güncellenir ve belirli amaçlar için değişik ortamlarda saklanır. Örneğin gazetelerdeki yazılar, makaleler birer belgedir. Aynı şekilde bilimsel çalışmalar neticesinde binlerce makale üretilmektedir. Her konuda, günlük hayatın içinde yer alan dergiler birer belge kümesidir. Konuşmalar, raporlar hep belgelerdir.

Giderek artan yoğunluğu, karmaşıklığı ve detayı ile belgeler çok kıymetli bilgiler içermesine rağmen çoğu kez bunlara ulaşılmaz ve onlardan yararlanılmaz. Gelenekselleşmiş yöntemleri

içeren bilgiye ulaşma (information retrieval) (metin sorgulama (text searching)) yöntemleri bir yere kadar belge yığınlarından faydalı ve gerekli bilgileri bulmaya yardımcı olsalar da aslen gerekli detay ve özel bilgilere bu yöntemler ile ulaşmak zordur hatta bazı durumlarda mümkün değildir. Oysa pek çok açıdan belgeler içindeki bilgilere, ilişkilere ulaşmak son derece önemlidir. Örneğin biyolojik veri yığınları hızla artmaktadır. İnsan gen haritası çalışmaları esnasında sayısız belge üretilmektedir. Bugün bir hastalık için ilaç bulmaya çalışan bir araştırmacının, kendisinden önce yapılmış tüm çalışmaları, olabildiğince hızlı bir şekilde incelemesi gerekir. Bu inceleme sürecinde belgelerin içeriğine, konusuna, içinde geçen kavramlara ve bu kavramların diğer belgelerde geçen farklı kavramlarla ilişkisine ulaşması gerekir.

Metin madenciliği çalışmaları bilgiye ulaşma ve bilgi çıkarma tekniklerinin gelişmesine paralel ilerlemektedir. Tezin bu bölümünde, bu noktadan sonra temel metin madenciliği yöntemleri, birbiriyle ilişkileri ve metin madenciliğinin gelişimi anlatılarak, günümüzde metin madenciliğinin geldiği nokta anlatılacaktır. Daha sonra bu alanda cevaplanması gereken mevcut sorunlardan birine yönelik, tez çalışmasının da içeriğini oluşturan konu paylaşılacak ve bu konu için önerilen çözüm ve elde edilen sonuçlar ortaya konacaktır.

2.2 Metin Veri Madenciliği Yöntemleri

2.2.1 Metin Sorgulama

Metin sorgulama bugün de en çok başvurulan metin madenciliği yöntemlerindendir. Özellikle internet kullanımındaki artışa paralel olarak metin sorgulama günlük yaşamın bir parçası haline gelmiştir.

Metin sorgulama, kullanıcının aradığı konudaki belgelerde konuyla ilgili bulunma olasılığı yüksek anahtar kelimeleri belirleyerek, içinde bu kelimelerin geçtiği belgeleri istemesi prensibine dayanır. En yaygın metin sorgulama araçları internet arama motorlarıdır. Bunun dışında kurumlara yada belirli bir topluluğa özel belge yönetim sistemleri de aynı görevi görmektedir (Hand vd., 2001).

Metin sorgulama işleminin yapılması için metin sorgulama araçları önce tüm belgeleri içerdikleri kelimelere ayırır ve bu kelimelere göre indeksler. Değişik metin arama yöntemleri, kullanıcıların anahtar kelimeler arasındaki ilişkileri ifade etmesini sağlayacak farklı yaklaşımlar içerir. Örneğin ikili (boolean) arama yöntemleri anahtar kelimeleri VE, VEYA ve DEĞİL (AND, OR ve NOT) gibi operatörlerle birleştirmeye izin verir. Bunun yanında serbest metin arama yöntemleri doğal dile yakın sorular sorulmasını sağlar. Neticede kullanıcıya aradığı belgelerin bir listesi döner. Genelde bu listede belgeler arama amacına uygunluklarına göre bir sırada kullanıcıya sunulur. Bu sıralamayı oluşturmak için kullanılan değişik yaklaşımlar vardır. Bunların en yaygın olanı anahtar kelimenin bir belgede kaç değişik yerde geçtiğidir. Farklı ortamlarda farklı sıralama yaklaşımların kullanıldığı görülebilir. Örneğin belirli bir arama kriterine göre getirilen web sayfalarının sıralaması, içerdikleri hiperbağların niteliğine bakılarak yapılabilir. Kendisine en çok referans edilen sayfa en uygun sayfa şeklinde değerlendirilir (Hand vd., 2001).

Metin sorgulama yöntemleri çok başarılı değildir, kabaca bir netice döndürürler. Bu yüzden arama amacına uygun görünen pek çok belge, sorgulama neticesi olarak gelir. Bu yöntemlerin kendi aralarında performans değerlemelerini yapmak için bazı kavramlar geliştirilmiştir. Bu sayede bir yöntemin ne kadar başarılı olup olmadı ölçülebilmektedir. Birinci kavram "çağırma" (recall), belge setindeki uygun belgelerden ne kadarının sorgu neticesi olarak döndürüldüğünü ölçer. İkinci kavram "Duyarlılık" (Precision) ise dönen belgelerin ne kadarının uygun belge olduğunu gösterir. Bu kavramların ölçülebilir hale getirilmesi için iki kavram daha bulunmaktadır. İlki, bir sorgu neticesinde gelmesi gereken belge kümesini ifade eder ve "uygun" (relevant) olarak tanımlanır. Diğeri de sorgu neticesinde gelen belge kümesini ifade eder ve "gelen" (retrieved) olarak tanımlanır (Chakrabarti S., 2003). Bunlar kullanılarak

$$hassasiyet = \frac{\|[uygun] \cap [gelen]\|}{\|[gelen]\|}. \qquad (2.1)$$

$$cagirma = \frac{\|[uygun] \cap [gelen]\|}{\|[uygun]\|} \qquad (2.2)$$

Bulanık operatörler (fuzzy operators) kullanımına imkan tanıyan sorgulama teknikleri sayesinde aranan anahtar kelimelerin dil içindeki dil kurallarına uyma zorunluluğu nedeniyle aldığı farklı formlar bir bütün olarak değerlendirilebilir. Bu sayede arama amacına uygun daha fazla belgeye ulaşılır. Yani "çağırma" artar (Manning vd., 1999).

Yakınlık operatörleri (proximity operators) ile anahtar kelimelerin birbirleri arasındaki ilişkileri de sorgulamak mümkündür. Örneğin anahtar kelimelerin aynı cümle içinde geçmesi şartı verilebilir. Bu operatörlerin kullanımı ile de sorgulama sonucu gelen belgelerin arama amacına uygunluğu artırılmış olur. Yani Duyarlılık artar.

Metin sorgulama, belge madenciliğinin çok önemli ve en geniş alanlarından biridir ve üzerinde sürekli geliştirmeler yapılmaktadır. Yukarıda bahsedilen yöntemlerin yanında sorgulama neticesini iyileştiren yaklaşımlar bulunmaktadır. Bunlar metin madenciliğinde sorgulama sürecini iyileştirme başlığı altında incelenecektir.

2.2.2 Metin Madenciliği İle Sorgulama Sürecini İyileştirme

Kullanıcının sorgulama şeklinden bağımsız olarak belgelerin içeriğinin sistematik analizini yapan ve bu sayede sorgulama sonuçlarını sınıflara ayıran, belge içeriklerini özetleyen, kullanıcıya gerektiğinde sorgulamasını yönlendirmesine yardımcı olan yöntemler vardır.

2.2.2.1 Sorgulama İyileştirme

Genelde sorgulama işleminin en ciddi eksiği, kullanıcının sorguladığı anahtar kelimelerin, sorgulanan belgelerde eşanlamlı kelimelerle yer almasıdır. Kullanıcıya doğru anahtar kelimeleri girmesini sağlayacak bir yöntem olarak sorgu iyileştirme geliştirilmiştir. Bu yöntem kullanıcının sorgu sonucu olarak aldığı belgelerden kendisi için yararlı olanı seçip, ona benzer belgeleri istemesine prensibine dayanır. Bu yolla kullanıcının ilk belirttiği anahtar kelimelerin dışında, kendisinin seçtiği belgedeki anahtar kelimelerin tamamı arama kriteri haline gelir.

2.2.2.2 Doğal Dil İle Sorgulama

"Semantic sorgulama" olarak da bilinen bu yöntem, kullanıcıların sorgulamalarını konuşma diline yakın ifadelerle oluşturmalarını sağlar. Bu yöntem sorgu ifadesini sözdizimsel ve anlamsal olarak inceler. Yani belgeler sadece içerdikleri kelimeler itibariyle değil, ayrıca daha üst seviye kelime grupları, kelimeler arasındaki ilişkiler bazında da indekslenir [7] (Kiryakov vd., 2004).

2.2.2.3 Belgeleri Demetleme

Büyük belge yığınlarını organize etmek için kullanılan demetleme teknikleri, sorgulama sonucu olarak dönen belgelerin da demetlenmesinde kullanılır. Demet içindeki belgeler içerik olarak benzer belgelerdir ve belirli anahtar kelimelere göre demetlenirler. Belgeler sadece anahtar kelimelere göre değil konularına göre de demetlenebilir. Konuları belgeler içindeki kelimelerin kullanım sıklıklarına göre belirlenen kelimeler ile ifade eder hale geldikten sonra belgeler konu bazlı demetlenerek, her demeti ifade eden anahtar kelimelerle görsel olarak sunulabilirler. Bu görsel gösterimde demetler birbirlerine yakınlıkları da görülebilecek şekilde aralarında oklarla ifade edilebilirler. Bu yaklaşım IBM'in Text Knowledge Miner ürününde de kullanılmıştır (Lerman 1999) [21].

2.2.2.4 Belgelerin sınıflara ayrıştırılması

Büyük belge yığınlarının uzmanlar tarafından belirlenmiş sınıf haritalarına yerleştirilmesi de sorgulama ve belge arama işlemini kolaylaştırmakta ve etkinliğini artırmaktadır. Daha önceden belirlenen sınıflara belgelerin dağıtılması işlemi, literatürde sınıflandırma olarak da bilinmektedir. Sınıflarla çalışma sürecinin ilk aşaması, konularında uzman kişilerce belirli konular için sınıf haritalarının oluşturulmasıdır. Belgelerin ilgili oldukları sınıflara dağıtılmasında iki yöntem kullanılır. Birinci yöntem, Yahoo arama motoru gibi sistemlerde uygulanan, internet sitelerini inceleyip bunları daha önceden belirlenen sınıf haritalarında ilgili sınıflara yerleştiren uzman grubunun izlediği yöntemdir. İkinci yöntem bu süreci otomatize eden yaklaşımları tanımlar. Otomatik sınıflama işleminin başarısı için öncelikle belirlenen her sınıf için, sınıfın anlamını ifade etme gücü olan örnek belgeler tespit edilir. Bu belgeler eğitim seti olarak da bilinir.

Sınıflandırma araçları bu örnek belgelerden, ilgili olduğu sınıfı ifade edebileceği anahtar terimleri çıkarır ve daha sonra kendisine verilen bir belgeyi doğru sınıfa atar. Makine öğrenmesi, yapay sinir ağları, kural tabanlı sistemler sınıflandırma çalışmalarında kullanılan yaygın tekniklerdir (Lerman, 1999) (Hand vd., 2001).

2.2.2.5 Belgeleri Özetleme

Belge özetlemenin amacı bir belgenin amacını anlatan kısa bir özetinin otomatik olarak oluşturulmasıdır. Etkin bir özetleme sistemi, kullanıcıların arama sonucu olarak elde ettikleri belgelerin özetlerine bakarak, tüm belgeyi inceleme zorunluluğu olmadan doğru belgeye ulaşıp ulaşamadıklarını belirleyebilmeleridir. Değişik seviyelerde özetleme oluşturmak mümkündür. Örneğin sadece anahtar kelimeleri içeren bir özet oluşturulabilir. Yada anahtar kelimeleri içeren cümlelerin seçilmesiyle bir özet oluşturulabilir. Daha ileri seviye özetleme teknikleri anahtar kelimeleri içeren cümleleri alıp, anlamlı bir özet oluşturmak için yeniden düzenleyebilirler. Bunun için doğal dil işleme çalışmalarından yararlanırlar (Hand vd., 2001).

2.2.2.6 Bilgi Çıkarma (Information Extraction)

Bilgi çıkarma yöntemleri metin içindeki unsurları, varlıkları otomatik olarak çıkarır ve bunlar arasındaki ilişkileri ortaya koyar. Metin içindeki cümleler ve paragraflar, içerdikleri önermelerle varlıklara ait bilgiler taşır. Bilgi çıkarma teknikleri bu önermelere bağlı olarak belgeyi oluşturan varlıkları ve bu varlıklar arasındaki ilişkileri çıkarırlar. Örneğin biyoenformasyon araştırmalarında gen yada protein isimleri, proteinlerin birbirleriyle ilişkileri çok önemlidir. Bu alanda yayınlanmış büyük belge kümeleri içinden gen ve protein isimlerini bulacak, bunların fonksiyonlarını çıkararak, aralarındaki ilişkileri ortaya çıkaracak yöntemler somut bulgular elde etmek adına büyük katkı sağlamaktadır ve ancak bilgisayar ortamında hayata geçirilebilirler (Fayyad vd., 2001) (Fayyad vd., 1995).

2.2.2.7 Doğal Dil İşleme

Terimler (gerçekler, kavramlar, varlıklar) ve bunlar arasındaki ilişkilerin görsel olarak gösterimi de eldeki belge yığını hakkında yorum yapmayı sağlayabilecek bir yöntemdir. Buna yönelik çalışmalar vardır. Terimler arasındaki ilişkilerin otomatik olarak çıkarılmasını sağlayan yöntemler vardır. Dilbilim kurallarının kullanımı ile cümleleri oluşturan kelimeler ve paragrafları oluşturan cümleler özerk terimlerine ayrıştırılıp, birbirleriyle dilbilimsel ilişkileri ortaya çıkarılabilir. Ya da terimlerin bir arada olma durumlarının istatistiksel değerlendirmesi yapılabilir. Bu sayede eldeki belge yığınına dair bir resim elde edilerek, araştırmalar yönlendirilebilir (Manning vd., 1999).

2.2.2.8 Anlam Notları ve Sözlükler

Belirli bir konuda ortak dil oluşturmak insanlar arasındaki iletişimi sağlamak için bile oldukça önemlidir. Benzer şekilde insanların oluşturduğu belgelerin bilgisayar sistemleri tarafından anlaşılması için de bilgisayar sistemlerinin o konudaki ortak dili bilmesi büyük avantaj sağlar. Bu ortak dil uzmanlar tarafından hazırlanmıştır ve belirli bir konudaki terimler, terimlerin özellikleri ve terimler arası ilişkileri gösterirler. Bu ortak dil iki şekilde kullanılmaktadır. Birinci yöntemde bir belgenin ilgili olduğu konuyla ilişkisi, içinde geçen kelimeler ve kelime gruplarının, daha önceden belirlenmiş ortak dil ile ne oranda kesiştiği ile belirlenir. İkinci ve daha etkin yaklaşımda, belgenin oluşturulması esnasında belge ilgili olduğu ortak dil kullanılarak hazırlanır (Fensel, 2001).

Birinci yaklaşıma örnek olarak sağlık sektöründe kullanılan MESH (Medical Source Headings) sözlüğü verilebilir. Bu sözlük uzmanlar tarafından kontrollü olarak geliştirilen bir sözlüktür. Sağlıkla ilgili oluşturulmuş belgeler bu sözlük vasıtasıyla kolayca deşifre edilerek, içeriği ortaya çıkarılabilir. Bu sözlük ikinci yaklaşım için de kullanılabilir. Oluşturulacak bir belgenin içeriği, bu sözlükteki terimler kullanılarak şekillendirilir. Bu sayede belgenin içeriğinin belirlenmesi çok daha kolay olur (Bernstein vd., 2002).

2.2.2.9 Kavram Haritaları (Ontolojiler)

İnsanoğlu dünyayı anlamak için sürekli modellere başvurur. Örneğin günümüzde kullanılan pek çok taşıt belirli bir modelin içinde yerleştirilmiştir. Buna göre taşıt dendiğinde kimileri motoru olan, insan ve yük taşımaya yarayan, tekerlekleri olan şeklinde bir üst kavram yaratıp, bunu daha alt dallarına binek taşıtları, yük taşıtları, binek taşıtlarını da üstü açık ve üstü kapalı, iki kapılı veya dört kapılı, arazi veya şehiriçi taşıtları şeklinde ifade edebilir. Daha üst seviye bakan birisi için taşıt kavramı insan veya yük taşımaya yarar genel tanımlamasının altında hava, su ve kara taşıtları olarak ve bunun altında kendi dalları şeklinde yapılandırılabilir.

Kavramsal haritaların da bir standardı oluşturulabilir. Her kavram haritası bir kavramsal çerçeve sunar yani bir konuyu belirli sınırlar içinde tanımlar. Bu çalışmalarda kavramlara karşılık gelen terimler, terimler arasındaki ilişkiler ve ilişkinin anlamı, eşanlamlı terimler, kullanım şekilleri gibi detaylı bilgile yer alır. Kavram haritaları yaygın olarak ontoloji olarak bilinir. Ontolojileri oluştururken kullanılan iki yöntem vardır. Birisici sınıflandırma ve alt sınıflara ayırma, ikincisi ise bütün-parça ilişkileri şeklinde gösterimdir. Ontolojiler bilgiye erişme ve bilgi çıkarma amaçlı çalışmalarda etkin olarak kullanılabilirler. Ontoloji kullanımı ile belgeler içindeki terimler ontolojideki karşılık terimlerine dönüştürülebilir ve belgeler kolayca demetlenebilir, sınıflandırılabilir. Bir başka kullanım alanı belgelere otomatik olarak ontolojik terimler kullanılarak notlar düşülebilmesidir (Fensel, 2001).

Bu kullanım kolaylıkları yanında ontolojilerin sürekli güncel tutulmaları gibi sorunları vardır. Ayrıca otomatik yapılan işlemlerin tamamı istenen neticeleri vermeyebilir ve bazen de olsa insan müdahalesine ihtiyaç duyulabilir (Fensel, 2001).

2.2.2.10 Bilgi Keşfi (Knowledge Discovery)

Bilgiye erişim yöntemlerine nazaran daha etkin sonuçlar elde edilmesini sağlayan bilgi çıkarma tekniklerinin avantajı belge içindeki içeriğin anlamını ön plana çıkaran terimlerin ve terimler arası ilişkilerin bulunmasında yatar. Ancak bazen belgelerin incelenmesindeki amaç daha önceden fark edilmemiş gerçeklerin ve ilişkilerin ortaya çıkarılmasıdır. Bu aşamada devreye bilgi keşfi teknikleri girer. Bilgi keşfi için kullanılan yöntemler metin içeriklerini derler, birbiri ile

entegre eder ve başka kaynaklardan elde edilen sonuçlarla birleştirerek üst seviye bir anlam ve ilişki kümesi oluşturmaya çalışır. Özellikle konuya bağlı olarak terimler ve terimler arası ilişkilerin üzerine de çıkılır ve konuya özel yapılar ve fonksiyonlara bağlı bir ilişki kümesi oluşturulur. Bu amaçla geliştirilen sistemlerin sadece belgeleri değil veritabanlarındaki verileri de kullanması gerekir.

Terimler ve terimler arasındaki ilişkileri gösteren görsel haritalar da yeni bilgilerin keşfinde kullanılabilir. Özellikle anlamsal ve mantıksal ilişkilerin dışında, birlikte bulunmaya bağlı olarak terimler arasında isimlendirilemeyen ilişkiler ortaya çıkarılabilir. Bu ilişkilerin incelenmesi ile henüz tespit edilmemiş ilişkilere ulaşmak mümkündür. Bu yönde yürütülen çalışmalarda örneğin sağlıklı beslenme ve hastalıklarla ilgili belgelerin birlikte incelenmesi ile aslında belgelerde bahsedilmeyen ama var olan ilişkiler yakalanmıştır (Fayyad vd., 2001) (Fayyad vd., 1995).

2.2.3 Metin Sınıflama

İçerik bazlı belge yönetimi işi (genel olarak bilgi alma – information retriaval olarak tanımlanır) belgelere ulaşımda esnekliği amaçlamaktadır. Metin sınıflama çalışması bu amaca ulaşmak için kullanılan bir adımdır ve konuşma dili ile yazılmış metinleri anlamlarına göre daha önceden belirlenmiş sınıflara ayırmaya çalışır. Günümüzde metin sınıflama, kontrollü bir kelime haznesine bağlı olarak belgeleri indeksleme, belgeleri filtreleme, otomatik olarak metadata (data hakkında soyut bilgi) oluşturma, web sayfalarını otomatik olarak hiyerarşik düzenlemeye tabi tutma gibi pratik olarak uygulanan pek çok alanda görmek mümkündür (Fabrizio, 2002).

Metin sınıflama çalışması D alanındaki belgelere C sınıfları bazında bir mantıksal değer atama işlemine indirgenebilir. Yani her

$$(d_j, c_i) \in D \times C$$

ikilisi için E (Evet) veya H (Hayır) değerleri belirlenir. Bir (d_j, c_i) için T değeri d_j belgesinin c_i sınıfında bulunduğunu, F değeri ise bulunmadığını gösterir. Değer atama işlemini bir fonksiyon olarak gösterirsek ve bu fonksiyonu Φ ile ifade edersek

$$\Phi : DxC \rightarrow \{F, E\}$$

Bu fonksiyonu bir kural, hipotez veya model olarak adlandırmak da mümkündür.

Metin sınıflama çalışmasını incelerken iki varsayımda bulunuyoruz:

- sınıflar sadece sembolik etiket görevi görürler, anlamları ile ilgili bilgi içermezler
- sınıflandırma için kullanılan bilgi dış etkilere bağlı değildir, yani belgenin yaratılma tarihi, tipi, kaynağı yoktur, sadece içeriği bilgisi vardır.

Bu varsayımlar, üretilecek metin sınıflama çözümünün genel geçerlilik taşımasını sağlar. Bu varsayımlara bağlı olarak sınıflandırma yapmak bile subjektif kalabilir çünkü gerçek hayatta bir konuda uzman iki kişinin bile bir belgeyi farklı iki sınıfa koydukları rastlanan bir durumdur (Fabrizio, 2002).

2.2.3.1 Tekli yada Çoklu Sınıflama

Metin sınıflama çalışmasında tercih gerektiren bir durum, herhangi bir $d_j \in D$ belgesinin sadece ve sadece bir sınıfa ait olabileceği ile birden fazla sınıfa de ait olabileceğinin kararının verilmesidir. Sadece bir sınıfa ait olabilecekse bu sisteme tekil-etiketli (single-label) yada örtüşmeyen sınıflı (nonoverlapping categories) denir. Bir belgenin birden fazla sınıfa ait olabileceği sistem ise çoklu-etiketli (multi-label) yada örtüşen sınıflı (overlapping categories) adı verilir.

Tekil-etiketli sistemler içinde özel bir sistem vardır. Bu iki sınıflı bir sistemdir. Bir d_j belgesi c_i sınıfına aittir ya da bu sınıfın tersi $\acute{c}_i$ ye aittir. İki etiketli sistem kolaylıkla diğer sistemlerin uygulanmasında da kullanılabilecek olması nedeniyle üzerinde çalışmaya değer bir yaklaşımdır. Zira çoklu-etiketli bir sistem n sınıflı ise, n adet iki sınıflı sistem gibi çalıştırılabilir. Burada yapılan varsayım, sınıfların birbirleri üzerinde etkisinin olmamasıdır. Yani bir belge c_1 sınıfına ait değilse c_2 ait olma durumu söz konusu olmamalıdır (Yang, 1999).

2.2.3.2 Net Sınıflama & Derecelendirilmiş Sınıflama

Metinleri sınıflandırırken bir $d_j \in D$ belgesinin bir sınıfa ait olduğu, diğerlerine ait olmadığı gibi net bir karar alınabilir. Buna alternatif olarak $d_j \in D$ belgesinin ait olabileceği sınıflara aitliği

derecelendirilerek de ifade edilebilir. Derecelendirme işlemi kurgulanan sistemin sınıf bazlı ya da belge bazlı olmasına bağlı olarak belgeler yada sınıflar üzerine uygulanır.

Metin sınıflama uygulamaları otomatik ve yarı otomatik olarak ikiye ayrılabilir. Yarı otomatik sınıflama sistemleri bir uzmanın gözetiminde, gerektiğinde uzmanın sınıflandırmaya müdahale ettiği bir etkileşim içerir. Otomatik sınıflama sistemleri ise bir uzman gereksinimi olmadan çalışan sistemlerdir. Ne tip bir sistem kurgulanacağı, eğitim seti adı verilen, sistemin sınıflandırmayı öğrenmek için kullanacağı örneklem uzayının içeriğine bağlıdır. Eğer yeterli içerik varsa otomatik sınıflandırma yaptırılır (Yang, 1999).

2.2.3.3 Metin Sınıflama ve Gizli Anlambilimsel Dizinleme (GAD)

GAD yöntemi 1990 yılında ortaya atılmıştır ve eşanlamlı, çok anlamlı yada yakın anlamlı kelimelerin belgeleri temsil eden terimler olması nedeniyle doğan problemleri gidermeyi amaçlamıştır. Bu yaklaşımda belgeler birer vektör olarak ele alınır ve belgelerda bulunan terimlerin bir arada bulunma kalıplarına bakılarak daha küçük boyutlu vektörler haline getirilir. Burada kullanılan benzerlik fonksiyonu, orijinal belge vektörlerinden oluşan matrise Tekil Değer Ayrıştırması uygulanmasıyla elde edilir. Geleneksel metin sınıflama konusu dahilinde benzerlik fonksiyonu eğitim setinden elde edilir ve daha sonra hem eğitim setine hem de test setine uygulanır.

Yapılan deneyler GAD yönteminin etkinliğini ortaya koymaktadır. Örneğin Schütze 1995'de GAD ile üç farklı teknik kullanan ki-kare yöntemlerini karşılaştırmıştır – kullanılan ki-kare teknikleri doğrusal ayrık analiz (lineer discriminant analysis), mantıksal regresyon (logical regression), ve yapay sinir ağlarıdır. Bu deneyler GAD yönteminin hepsinden daha etkili olduğunu göstermiştir (Schütze vd., 1995).

2.2.3.4 Metin Sınıflayıcıların Sınıflama Biçimi

Sınıflamanın net (otomatik) ve derecelendirilmiş (yarı otomatik) olarak iki grupta olduğundan bahsedilmişti. Derecelendirilmiş sınıflama yöntemi sınıf durum değeri (SDD) (categorization

status value :CSV) adı ile ifade edilen ve her d_j belgesi için [0,1] arasında bir değer atayan tanım içerir. Bu değer $d_j \in c_i$ derecesidir.

$$SDD_i : D \rightarrow [0,1]$$

Belgeler SDD_i değerlerine göre derecelendirilirler. Bu şekilde yapılan derecelendirilmeye belge-dereceli metin sınıflama denir. Sınıf-dereceli metin sınıflama ise her dj belgesinin değişik sınıflar için SDD_i değerlerine bakılarak tasarlanır (Fabrizio, 2002).

Kullanılan öğrenme yöntemine bağlı olarak SDD_i fonksiyonun anlamı farklılaşabilir. Örneğin Naive Bayes yaklaşımında $SDD_i(d_j)$ ile ifade edilen fonksiyon aslen bir olasılık fonksiyonuyken, GAD ve Rocchio yöntemlerinde $SDD_i(d_j)$, |T| boyutlu uzayda vektörler arasındaki uzaklık(benzerlik) fonksiyonudur.

Net sınıflama tanımı $SDD_i : D \rightarrow [E,H]$'dur. Ancak gerçekte bu ifade

$$SDD_i : D \rightarrow [0,1] \text{ dir}$$

ve bir τ_i eşik değerine bağlı olarak E ve H değerleri verilir. Yani

$$E : SDD_i(d_j) >= \tau_i \ , \ H : T : SDD_i(d_j) < \tau_i$$

2.2.3.5 Eşik Değerini Belirleme

Eşik değerini belirlemek için iki yol vardır. Analitik olarak veya deneysel olarak hesaplama. Eşik değerinin önemi sınıflama sisteminin etkin çalışmasına olan katkısındandır. Bu nedenle doğru eşik değeri sınıflama sisteminin görevini ne kadar iyi yaptığı ile değerlendirilir. Sistemin görevini ne kadar doğru yaptığını bulmak için kullanılan fonksiyona etkililik (effectiveness) fonksiyonu denir. Etkililik fonksiyonu da birkaç farklı yöntem ile tanımlanabilir.Örneğin belgeleri tasnif ettikten sonra en doğru tasnif edilenlerin toplam tasnif edilen belge sayısına oranı alınarak bir olasılık fonksiyonu tasarlanabilir. Ya da belgelerin yanlış tasnifi sonucunda ne gibi kayıplar olacağına bakılabilir ve bu kayıplar ekonomik birer değer olarak tesbit edilebilir. Bu ikinci yöntem için daha iyi bir anlatım bir örnekle yapılabilir; eğer sistem gelen elektronik postaları sınıflıyorsa ve içeriye gereksiz bir mail girmesine izin veriyorsa ortaya çıkan ekonomik kayıp, gelen bir fatura içerikli elektronik postanın çöp sınıfına atılmasından çok çok daha azdır. Bu gibi

ikinci sınıfa giren etkililik fonksiyonlarının kullanacağı eşik değerleri analitik olarak hesaplanır (Lewis, 1995a).

Bu gibi teorik sonuçlar baştan bilinmiyorsa, eşik değeri deneysel olarak hesaplanır. Değişik τ_i değerleri doğrulama seti üzerinde denenir ve etkililiği maksimum yapan eşik değeri seçilir. Bu yaklaşıma "SDD eşik değeri bulma" adı verilir. Genelde farklı c_i sınıfları için farklı τ_i değerleri seçilir

"SDD eşik değeri bulma" yönteminden farklı deneysel bir yöntem de "orantısal eşik değeri bulma"dır. Bu yöntem sistemin doğrulama seti ile test seti için tasnif ettiği belgelerin birbiriyle yüzdesel olarak tutarlı olmasını sağlamayı hedefleyen hesaplamalarla çalışır.

Bir başka yöntem ve aslında ilk kullanılan eşik değeri yöntemi tüm sınıflar için sabit bir değeri bulmaya çalışır. Eşik değeri bulma yöntemleri uygulamanın kendisine de bağlı olduğundan hangisinin diğerinden daha iyi olduğu net olarak söylenemez (Iwayama vd., 1995).

2.2.3.6 Olasılık Temelli Sınıflayıcılar (Probabilistic Classifiers)

Sınıflamanın $SDD_i(d_j)$ değerleri ile yapıldığından bahsedilmişti. Olasılık temelli sınıflayıcılar için $SDD_i(d_j)$ fonksiyonu $P(c_i \mid \vec{d}_j)$ terimleri ile tanımlanır. $P(c_i \mid \vec{d}_j)$ $\vec{d}_j$ vektörü ile gösterilen belgesin c_i sınıfına ait olma olasılığıdır ve $\vec{d}_j$

$\vec{d}_j = (w_{1j}, ...,w_{|T|j})$, $P(c_i \mid \vec{d}_j)$ de Bayes teoreminden

$$P(c_i \mid \vec{d}_j) = \frac{P(c_i)\, P(\vec{d}_j \mid c_i)}{P(\vec{d}_j)} \qquad (2,3)$$

Burada $P(\vec{d}_j)$, rastgele seçilen bir belgesin d_j olma olasılığını, $P(c_i)$ ise rastgele seçilen bir belgesin c_i sınıfına ait olma olasılığını gösterir.

$P(c_i \mid \vec{d}_j)$ 'nin hesaplanması zordur çünkü ile $\vec{d}_j$ vektörlerinin sayısı çok fazladır. Bunu aşmak için şu varsayımda bulunmak mümkündür. Bir belge vektörünün herhangi iki koordinatı, rastgele

sayılar olarak ele alınırsa, istatistiksel olarak birbirinden bağımsız olduğu görülür. Bunu matematiksel olarak ifade edersek

$$P(c_i \mid \vec{d}_j) = \prod_{k=1}^{|T|} P(w_{kj} \mid c_i) \qquad (2,4)$$

Bu yaklaşımı kullanan olasılık temelli sınıflandırıcılara Naive Bayes sınıflandırıcıları denir (Lewis, 1998).

2.2.3.7 Sembolik Algoritma Temelli Sınıflayıcılar

Olasılık temelli yöntemler nicel yöntemlerdir. Bu yönleri nedeniyle insanlar tarafından yorumlanması zor yöntemlerdir. Sembolik olarak sınıflandırma yapan ve kolayca anlaşılır yöntemler geliştirilerek bu sorun giderilmiştir. Bunların en önemli iki tanesi

- Karar ağaçları (Decision Tree)
- Tümevarım kuralları (inductive rules)

2.2.3.8 Karar Ağaçları

Karar ağaçları bir ağaç yapısına sahiptir, bu ağacın dallanma noktaları terimler ile işaretlenir. Her dal terimin test belgeleri üzerindeki ağırlığı ile (w) işaretlenir. Ağacın yaprakları da sınıflar ile işaretlenmiştir. Bu gibi bir sınıflayıcı aldığı bir dj test belgesini tekrarlı olarak, bu belgesin $\vec{d}_j$ vektöründeki ağırlık değerlerini terimlerinkiyle karşılaştırır ve bu işlem ağacın bir yaprağına ulaşıncaya devam eder. Ulaşılan yaprak test belgesinin ait olduğu sınıfı gösterir (Mitchell, 1996).

Bu yaklaşımı kullanan ve standartlaşmış paket uygulamalar bulunmaktadır. Bunların en yaygınları

- C4.5 (Cohen vd., 1998)
- ID3 (Fuhr 1991)
- C5 (Li vd., 1998)

Hemen hemen tüm ağaç yapısı oluşturan yöntemler “böl ve yönet” taktiğini kullanır. Önce tüm test belgelerinin c_i ye mi yoksa c_i’ mi ait olduğuna bakılır. Öyle değilse eğitim seti içinden bir t_k

terim alınır ve eğitim setindeki belgeler bu terime bakılarak ayrı bir sınıfa alınır. Bu sınıflar hep bir alt sınıf olarak detaylandırılır. Görüleceği gibi buradaki kritik karar t_k teriminin seçimidir. Bu seçim uzmanlık gerektiren bir bilgi ister. Bunu bulmak için farklı yöntemler vardır (information gain, entropy gibi) (Fabrizio, 2002).

Ağaç yapısı eğitim setine aşırı uyumlu hale gelme sorunu yaşatır. Bunu engellemek için ağacı spesifik belgelere göre uygun hale getiren dalların budanmasını sağlayan teknikler kullanılır.

2.2.3.9 Tümevarım Kuralları

Tümevarım kuralları ile öğrenen bir sistem her c_i sınıfı için DNF (disjunctive normal form) formatında bir kural içerir. Bir kuralın tam formatı:

If (*DNF formülü*) then (*sınıf*)

şeklindedir. Örneğin ingilizcede isimlerin başına getirilen öneklerden Mr., Mrs. Ve Miss öneklerinin kullanımının kuralını yazacak olursak

If (erkek = True) then onek=Mr.

If (erkek = False and bayan=True and evli=true) then onek=Mrs.

If (erkek = False and bayan=True and evli=False) then onek=Miss.

DNF biçiminin aslında terimler artıkça aşağıdaki gibi olması gerekir:

If (x=y) or
  (x=z and x!=t) or
  (....)
then *sınıf*

Eğitim setindeki tüm belgeler için kurallar çıkarıldıktan sonra minimum DNF formülü ile çalışan kurallar sırasıyla seçilirler. Bu açıdan bakıldığında karar ağaçları yukarıdan aşağıda bir yöntem izlerken, tümevarım kuralları aşağıdan yukarıya yöntemini kullanır. Tüm kurallar çıktıktan sonra bazı kuralların bazı belgeler için aşırı uyumlu hale geldiği görülür. Kural düzenleme algoritması kuralları budayarak, bazılarını bir araya getirerek tüm eğitim seti belgelerini tasnif edecek son haline ulaştırır (Fabrizio, 2002).

Burada anlatılan yöntemler önerme mantığını (propositional logic) kullanırlar. First Order Logic kullanan yaklaşımlarda olmuştur. Ancak Cohen 1995'de yaptığı testlerde iki kural oluşturma yöntemi arasında büyük bir fark olmadığını tesbit etmiştir (Cohen, 1995a).

2.2.3.10 Regresyon Yöntemleri

Regresyon gerçek değerler ile çalışan bir fonksiyondur, sınıflandırmada olduğu gibi (E,H) değerleri ile çalışmaz. Bu Φ' fonksiyonu eğitim seti belgelerini tasnif eden Φ fonksiyonunun bir yaklaşık versiyonudur. En küçük kareler yöntemini bir versiyonu olan "doğrusal en küçük kareler ile uyum" (Linear Least Square Fit) yöntemi metin sınıflamana uygulanmıştır (Yang vd., 1994).
Bu yöntem her d_j belgesine iki vektör atar.

- Bir girdi vektörü: |T| ağırlıklı terimden oluşan I (d_j)
- Bir çıktı vektörü : |C| ağırlık değeri, her sınıf için O (d_j)

Sınıflama problemi verilen bir d_j belgesi için, verilen girdi vektöründen çıktı vektörünü oluşturmaya indirgenmiş olur. Daha açık bir ifade ile öyle bir |C|x|T| matrisi(M) bulunmalıdır ki M I (d_j) = O (d_j) olsun.

2.2.3.11 Online Yöntemler

Bir c_i sınıfı için doğrusal sınıflayıcı
$\vec{c}_i = (w_{1i},.....,w_{|T|i})$ vektörüdür. Bu vektör belge vektörlerinin de bulunduğu |T| uzayındadır. $\vec{d}_j = (w_{1j}, ...,w_{|T|j})$. Buradan

$$SDD_i (d_j) = \sum_{k=1}^{|T|} w_{ki}\, w_{kj} \qquad (2,5)$$

bulunur. Aynı uzaydaki vektörlerin arasındaki açının kosinüsü lineer cebir teoremleri gereği aşağıdaki gibidir:

$$\text{Cos}(\alpha) = \frac{\sum_{k=1}^{|T|} w_{ki}\, w_{kj}}{\sqrt{\sum_{k=1}^{|T|} w^2_{ki}}\ \sqrt{\sum_{k=1}^{|T|} w^2_{kj}}} \qquad (2,6)$$

Bu değer iki vektörün benzerliği fonksiyonu olarak da kullanılabilir. Benzerlik fonksiyonu

$$S(c_i, d_j) = \cos(\alpha)$$ şeklinde gösterilir.

Doğrusal sınıflama yöntemi ile öğrenme yöntemlerini iki ana grupta toplamak mümkündür. Toplu işlem (batch) yöntemler ve çevrimiçi (online) yöntemler. Toplu işlem yöntemleri eğitim setinin tamamını bir kerede işler. Bu yöntemlerden metin sınıflama konusunda en etkili olan Rocchio'dur ve ileride anlatılacaktır. Ayrıca doğrusal ayırım analizi (linear discriminant analysis) terimler arasındaki rastgele beğımlılıkları bularak çalışır. Terimler arasındaki bu rastgele bağımlılıklar sınıfların kovaryans matrisleri üzerine bina edilir (Fabrizio, 2002).

Çevrimiçi yöntemler ilk eğitim setini işledikten sonra sınıflandırma yapabilir hale gelir. Her yeni belgesin incelenmesi ile sistem iyileştirilir. Bu yöntemler özellikle eğitim seti tam olarak hazırlanamamış alanlara uygulanırlar. Aynı şekilde sınıfların anlamlarının zaman içinde değiştiği alanlar için de uygundurlar. Basit bir çevrimiçi yöntem olan perceptron algoritması şöyle çalışır:

"c_i sınıfı için sınıflandırıcı tüm ağırlıklara (w_{ki}) sabit bir ilk değer atar. Sistem bir $\vec{d}_j$ vektörü ile ifade edilen d_j belgesini sınıflandırır. Eğer netice doğruysa bir şey yapılmaz, yanlışsa sınıflandırıcının sınıflarına atanmış ağırlık değerleri değiştirilir. Eğer d_j, c_i sınıfına girecek ise (pozitif örnek ise) belgesi sınıflandırmak için aktif olan terimlerin w_{ki} değerleri bir $\alpha > 0$ kadar artırılır. α'ya öğrenme oranı adı verilir. Belge negatif örnek ise w_{ki} değerleri bir α kadar azaltılır. Sistem tekrarlı olarak çalıştıktan sonra düşük w_{ki} değerlerine sahip terimlerin belgesin sınıflandırılmasına katkısı olmadığı neticesi ortaya çıkar. Bu terimler sınıflandırma terimleri dışına çıkarılarak boyut daraltma işlemi yapılmış olur."

Perceptron algoritması α değerini toplama işlemi ile kullanır. Çarpma ile kullanılan başka bir versiyonu "Positive Winnow" olarak bilinir (Dagan vd., 1997). Dagan "Positive Winnow" yöntemini "Balanced Winnow" yöntemi ile geliştirmiştir. "Balanced Winnow" yöntemi iki α değeri ve iki ağırlık değeri kullanır.

2.2.3.12 Rocchio Yöntemi

Bazı doğrusal sınıflayıcılar sınıflar için prototip belgeler oluştururlar. Bunların faydası sistemin anlaşılmasını kolaylaştırmasıdır. Rochio yöntemi bir sınıflayıcı oluşturur, öyleki

$$\vec{c}_i = (w_{1i},.....,w_{|T|i})$$

ile ifade edilen her sınıf için ağırlık değerleri aşağıdaki gibi hesaplanır

$$w_{ki} = \beta \sum_{\{d_j \in POS_i\}} \frac{w_{kj}}{|POS_i|} - \gamma \sum_{\{d_j \in NEG_i\}} \frac{w_{kj}}{|NEG_i|} \qquad (2.7)$$

Bu formulasyonda

w_{kj} t_k teriminin d_j belgesindeki ağırlığı,

$POS_i = \{d_j \in Tr \mid \Phi'(d_j,c_i) = E\}$ yani c_i sınıfı için pozitif örnekler

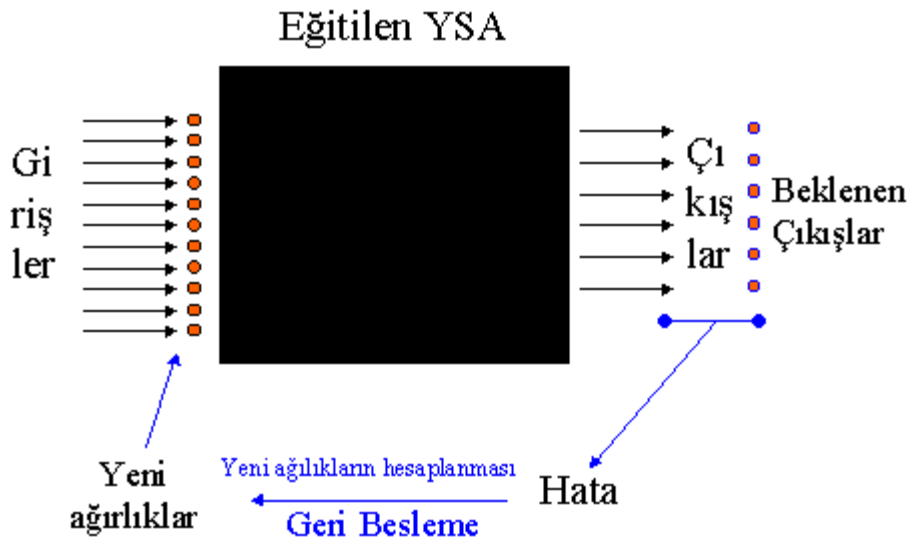
$NEG_i = \{d_j \in Tr \mid \Phi'(d_j,c_i) = H\}$ yani c_i sınıfı için negatif örnekler

β ve γ değerleri negatif ve pozitif örneklerin önem derecesini ayarlamak için kullanılan parametrelerdir. Örneğin $\beta=1$ ve $\gamma=0$ ise c_i sınıfının prototip belgesi pozitif örneklerden türetilir.

Bu yöntem uygulaması kolay ve etkili bir yöntemdir. Ancak bu yöntem ile sınıflandırma yöntemi belirlenen bir sınıfa ait belgeler birbirinden ayrık demetler içindeyse, sınıflayıcı belgelerin çoğunu sınıflayamayabilir. Örneğin Spor sınıfı altında boks, futbol gibi konulara ait belgelerin sınıflandırılması mümkün olmayabilir. Çünkü örneğin bir c_i sınıfına ait pozitif belgeler baz alınarak ($\beta > \gamma$) oluşturulan prototip pozitif belgelerin merkezini gösterir ve yeni gelen bir belge bu merkeze olan uzaklığına bakılarak sınıflandırılır. Doğrusal bir yaklaşım olduğu için yapay sinir ağları gibi yöntemlerin sağladığı esnekliği sağlamaz (Fabrizio, 2002).

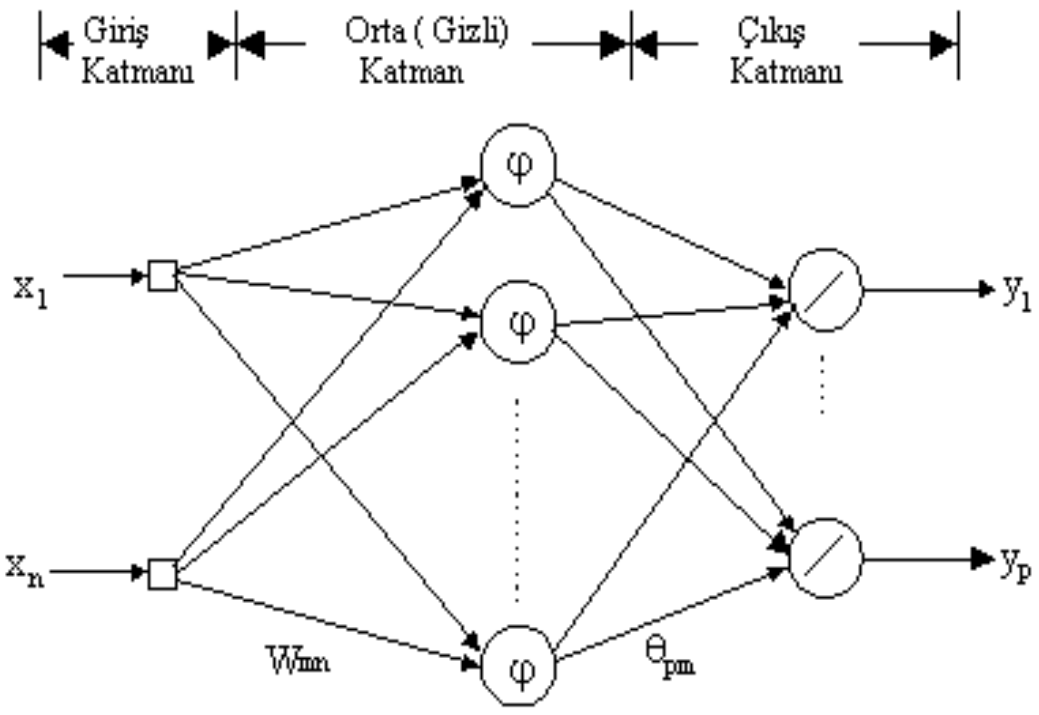
2.2.3.13 Yapay Sinir Ağları

Yapay Sinir Ağları tabanlı bir metin sınıflama sistemi Şekil 2-1 deki gibidir.



Şekil 2-1 Yapay Sinir Ağları tabanlı bir metin sınıflama sistemi

Girdiler terimler, çıktılar ise sınıflardır. Yapay sinir ağlarının iç yapısı ise Şekil 2.2'deki gibidir.



Şekil 2-2 Yapay Sinir Ağının İç Yapısı

Buradaki her w_{mn} bağımlılık ilişkilerini gösterir. Bir d_j test belgesini doğru sınıfa aktarmak için, bu belgesin w_{kj} ağırlıkları yapay sinir ağına girdi olarak verilir. Bu girdiler ağ içinde ilerler ve bir çıktı oluşur. Bu çıktı belgesin ait olduğu sınıfı gösterir. Eğer belge doğru sınıflandırılmamışsa hata geri besleme yöntemi ile ağa bildirilir ve ağ içindeki w_{mn} ağırlıklarının doğru değerlere ayarlanması sağlanır (Fabrizio, 2002).

2.2.3.14 Örnek Tabanlı Sınıflayıcılar:

Bu gruptaki yöntemler tembel yöntemler olarak bilinir çünkü sınıflar için ayrı tanımlayıcı tanımlar yoktur. Belgeler, eğitim setindeki belgelere olan benzerliklerine göre sınıflandırılırlar. En çok bilinen örneği "k en yakın komşu" k-NN (k nearest neigbour) yöntemidir (Yang 1994).

k-NN yöntemi bir d_j test belgesinin c_i sınıfına ait olup olmadığını bulmak için kendisine benzeyen k tane eğitim seti içinden belgesin c_i sınıfına ait olup olmadığına bakar. Burada kritik konu d_j test belgesine benzerliğin nasıl hesaplanacağıdır. Bu fonksiyon RSV (d_j,d_z) ile gösterilirse amacı d_j test belgesi ile d_z eğitim belgesi arasındaki benzerliği ölçmektir. Bu fonksiyonu uygularken olasılık temelli, vektör temelli veya herhangi bir yaklaşım kullanılabilir.

RSV fonksiyonu kullanılarak belgesin $SDD_i(d_j)$ değeri hesaplanabilir :

$$SDD_i(d_j) = \sum_{d_z \in Tr_k(d_j)} RSV(d_j,d_z) \left\{ \Phi'(d_j,c_i) \right\}$$

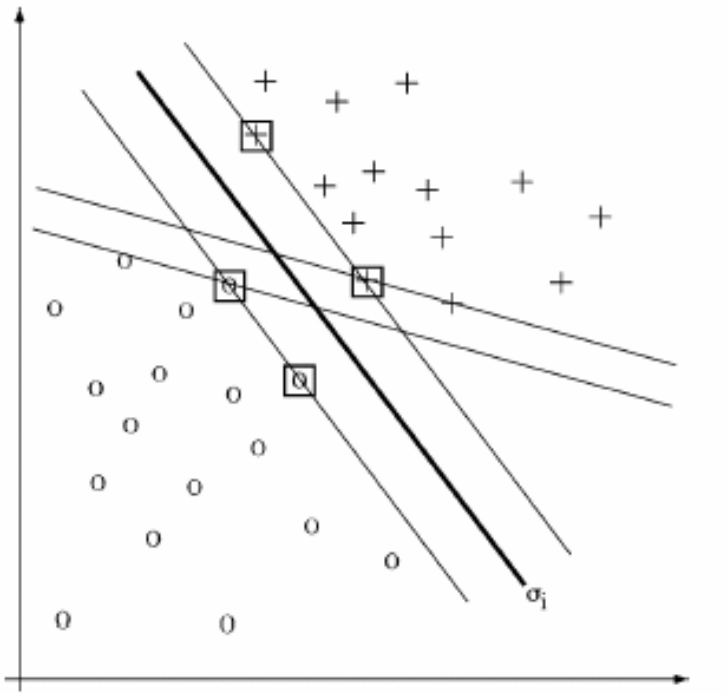
$$\left\{ \Phi'(d_j,c_i) \right\} = \begin{cases} 0 & \Phi'(d_j,c_i) = T \\ 1 & \Phi'(d_j,c_i) = F \end{cases} \qquad (2.8)$$

$Tr_k(d_j)$, RSV (d_j,d_z) değerini maksimum yapan k adet eğitim seti belgesini gösterir.
$SDD_i(d_j)$ değeri sınıflandırma sisteminin etkinliğini gösterir. Bu değeri k değeri etkiler. Larkey ve Croft (1996) ve Yang (1999) yaptıkları testlerde sırasıyla k=20 ve 30<k<=45 değerlerini bulmuştur (Larkey vd., 1996) (Yang, 1999).

k-NN yöntemi doğrusal sınıflayıcılar gibi belge uzayını doğrusal olarak bölmez. K-NN yöntemi çok etkili olmakla birlikte doğrusal sınıflama tekniklerine göre hesaplama açısından bir dezavantaj yaşamaktadır çünkü tüm eğitim seti belgelerinin test seti belgeleri ile aralarındaki benzerliği hesaplamaktadır.

2.2.3.15 Destek Vektör Makinesi (Support Vector Machine)

Bu yöntem |T| boyutlu uzayda eğitim seti belgelerini c_i sınıfları bazında pozitif ve negatif örnekler olarak ikiye böler. Aşağıdaki örnekte görüldüğü gibi + lar pozitif belgeleri o lar negatif belgeleri gösterirse, bunları en iyi ayıran vektör σ_i vektörüdür çünkü pozitif ve negatif örneklere maksimum uzaklığa sahiptir.



Şekil 2-3 Destek Vektör Makinesi ile Sınıflama

Bu yöntemin avantajı karar yüzeylerinin belirlenmesinde kullanılan eğitim seti örneklerinin az olmasıdır ve bu örneklere "destek vektörü" denir. Bu yöntem ilk kez 1999 yılında Joachim tarafından kullanılmıştır (Joachims, 1999). Yöntemin diğer avantajları

- Terim seçme ihtiyacını ortadan kaldırır.

- Aşırı uyumluluğa karşı teknik içermesine gerek yoktur.
- Ölçeklenebilme yeteneği yüksektir.
- Doğrulama seti üzerinde parametre ayarlama gibi bir ihtiyaç yoktur.

Sistemin öğrenme hızı da oldukça yüksektir.

2.2.3.16 Sınıflayıcılar Komitesi

Buraya kadar bir çok sınıflayıcıdan bahsedilmiştir. Birden fazla uzmanın ortak görüşü bir uzmanın görüşünden daha iyidir felsefesi gereği, sınıflayıcılar birlikte kullanılabilirler. Daha doğru bir ifade ile sınıflayıcılar bir belgesi tasnif etmek için eğitim ve test setlerine bağlı olarak çalıştırılırlar ve her birinden birer netice alınır. Alınan bu neticeler bir araya getirilerek daha doğru bir tasnif yapmak mümkündür. Bu şekilde yapılan bir sınıflayıcı organizasyonu sınıflayıcılar komitesi olarak adlandırılır (Scott vd, 1999) (Li vd., 1998).
Sınıflayıcılardan elde edilen neticeleri bir araya getirmek için farklı alternatifler söz konusudur:

- k adet sınıflayıcının neticeleri içinden (k+1)/2 tanesinin neticeleri aynı ise bu netice nihayi netice olarak değerlendirilebilir
- Biliyoruz ki her sınıflayıcının etkililiği ile ilgili bir değer hesaplanabilmektedir. Bu değerler sınıflayıcıların kendi aralarında derecelendirilmelerini sağlar. Nihayi netice her sınıflayıcının ağırlına göre doğrusal bir hesaplama ile oluşturulur.
- Sınıflayıcılar değişik alanlar için çok iyi neticeler verirken, bazıları için iyi neticeler oluşturamazlar. Bu veriden yola çıkarak, ilgili alana bağlı en etkili neticeyi veren sınıflayıcının neticesi nihayi netice olarak kabul edilebilir

2.2.3.17 Metin Sınıflama Yöntemlerinin Etkinliğinin Ölçülmesi

Duyarlılık ve Çağırma (Precision and Recall) başlığı altında iki yöntemle sınıflama işlemi yapan yöntemlerin etkinliği ölçülebilir.

Duyarlılık π ile gösterirsek her c_i sınıfı için bir π_i duyarlılık ölçüsü vardır ve bunun formulasyonu koşullu olasılık olarak aşağıdaki gibidir:

$$P(\Phi'(d_x,c_i) = T \mid \Phi(d_x,c_i) = T)$$

Bunun anlamı eğer rasgele alınan bir d_x belgesi c_i sınıfına yerleştirilmiş ise doğru yerleştirilmiştir. Çağırmayı ise ρ ile gösterirsek her c_i sınıfı için bir ρ_i hatırlama ölçüsü vardır ve bunun formulasyonu koşullu olasılık olarak aşağıdaki gibidir:

$$P(\Phi(d_x,c_i) = T \mid \Phi'(d_x,c_i) = T)$$

Bunun anlamı eğer rasgele alınan bir d_x belgesi c_i sınıfına yerleştirilmesi gerekiyorsa c_i sınıfına yerleştirilir.

Sınıflara bağlı olarak hesaplanan bu değerler tüm sistemi hakkında genel bir değer oluşturmak için ortalamaları alınarak kullanılabilir (Fabrizio, 2002).Bu değerlerin hesaplanması bir test seti baz alınarak c_i sınıfları için ihtimal tablosu yardımı ile yapılabilir:

Tablo 2-1 İhtimal Tablosu

c_i sınıfı		Uzman görüşü	
		Evet	Hayır
Sınıflayıcının sonucu	Evet	Sınıflayıcı pozitif Uzman doğru TP_i	Sınıflayıcı pozitif Uzman yanlış FP_i
	hayır	Sınıflayıcı negatif Uzman yanlış FN_i	Sınıflayıcı negatif Uzman doğru TN_i

Örneğin ihtimal tablosunda uzmanın c_i sınıfına ait olması gerekir dediği belgelerin, sınıflayıcının yanlış sınıflandırdıklarının toplamı FP_i olarak gösterilmiştir. Buradan

$$\pi'_i = \frac{TP_i}{TP_i + FP_i} \tag{2.9}$$

$$\rho'_i = \frac{TP_i}{TP_i + FN_i} \tag{2.10}$$

2.3 Metin Veri Madenciliğinin Genişlemesi

Metin veri madenciliğinin ilk çıkış noktası, internet üzerindeki milyonlarca web sayfası içinden, doğru bilgiye ulaşmayı sağlamak için ortaya çıkmış arama motorlarıdır. Başlangıçta iki tip arama motoru geliştirilmiştir. Bunlardan ilki Yahoo'da olduğu gibi elle yapılan indeksleme yöntemine dayanır. İkinci tip arama motorları ise indeksleme işini otomatik yapma prensibini benimsemiştir. İnternet altyapısının günümüzdeki kadar yüksek bant genişlikleri üzerinden sunulmadığı bu ilk dönemlerde, belgeler, web sayfaları halinde düzenlenmiş, HTML formatına sahip belgelerdir. Zamanla iletişim kapasitelerindeki artışlar, internet üzerinden pdf, doc, ps gibi, kişisel bilgisayar ortamlarında hazırlanan belgelerin paylaşımını mümkün kılmış ve günümüzde ağırlıklı olarak bilgi içeren belgeler artık HTML formatı dışındaki formatlarda tutulur olmuştur.

İnternetin sürekli büyüyen yapısı içinde, aranılan bilgilere erişimde güçlükler ortaya çıkmıştır. Kişisel bilgisayarların da yaygınlaşması ile işe özel çalışma ortamlarında veya ilgi gruplarında, ilgi grubuna veya işe özel bilgiler üretilmeye başlanmıştır. Bilgi ihtiyacının özelleşmeye başlaması üzerine, metin yığınları içinden amaca yönelik en uygun bilgiyi bulmak önem kazanmıştır. Örneğin gen araştırmaları esnasında çok fazla belge üretilmiş ve araştırmacılar daha önce bulunmuş ve tecrübe edilmiş bilgileri, metin veri madenciliğinin gen araştırmaları alanına özelleşmiş teknikleri ile tekrar araştırmadan kullanabilmişlerdir.

Hızlı bilgi artışının bir etkisi de yeni bilgilerin eski bilgileri geçersiz kılmasıdır. Bu nedenle metin veri madenciliği tekniklerinin, yeni bilgileri hızla öğrenebilmesi gereksinimi doğmuştur. Bugün tüm dünyada en fazla kullanılan arama motoru olan Google, bu özelliğini ön plana çıkararak farklılaşmıştır. Öğrenebilen bir metin veri madenciliği yöntemi, belgelerin üretildiği doğal dili tanıması oranında, öğrenme yeteneğini geliştirebilmektedir. Bu nedenle doğal dil işleme çalışmalarının metin madenciliğinde kullanılması ile oldukça etkin araçlar elde edilebilir.

Doğal dil işleme alanının belge madenciliğinin gelişmesi açısından taşıdığı büyük önem nedeniyle, doğal dil işleme konusunda kısa bir açıklama gereklidir ve bir sonraki bölümde bu konu anlatılmıştır.

2.4 Doğal Dil İşleme ve Belge Veri Madenciliği

Belgelerin analizi için, içeriğinin anlamını taşıyan kavramların tespit edilmesi gerekmektedir. Bu kavramlar kelimeler veya kelime grupları ile ifade edilir ve terimler olarak adlandırılır. Belge içindeki terimlerin çıkarılması başlı başına bir konudur ve doğal dil işleme çalışmaları kapsamında incelenen bir alandır. Doğal dil işlemenin belge analizi sürecindeki en önemli faydası terimlerin yani kelimelerin ayrıştırılması, eklerinden arındırılarak anlamını kaybetmeyen en kısa biçimlerine dönüştürülmesidir çünkü aynı anlam için kullanılan kelimeler dilbilgisi kuralları gereği farklı biçimlerde bulunabilir ve bu farklı kullanım biçimleri ortadan kaldırılmadığı takdirde farklı anlam taşıyan terimler gibi işleme alınarak, belgelerin gerçek anlamına ulaşılmasını engelleyebilirler. Doğal dil işleme çalışmaları kapsamında yürütülen girişimler dört ana grup altında toplanabilir [23].

- Biçimbirimsel çözümleme (Morfolojik analiz)
- Sözdizimi çözümlemesi (Sentaktik çözümleme)
- Anlam çözümlemesi (Semantik çözümlemesi)
- Anlam kargaşasının giderilmesi

2.4.1 Biçimbirimsel Çözümleme

Kelimedeki köklerin ve o köklere gelen eklerin görevlerinin ayrıştırılmasını sağlar [23]. Mesela, Türkçe “geldim” kelimesinin çözümlemesi yapıldığında

gel+di+m

şeklinde bir açılım görürüz ve bunu bir ara forma çevirirsek

“gel+fiil+geçmiş+1.tekilşahıs”

şeklinde ifade edebiliriz.

Aynı şekilde “evlerde” kelimesinin açılımı

ev+ler+de

olarak yapılabilir ve

“ev+isim+3.çoğulşahıs +bulunma”

şeklinde bir ara forma dönüştürülebilir.

Ancak, işler her zaman bu kadar kolay olmayabilir. Gerek Türkçe'de, gerekse diğer dillerde ses uyumu kuralları ve çeşitli durumlarda araya giren ya da düşen harfler, kurallarda oynamalar meydana getirebilir. Mesela, "yalıyla" kelimesi

"yalı+yla"

şeklinde açılırken, "sarayla" kelimesi

"saray+la"

şeklinde açılacaktır. Aynı çoğul eki

"ev+ler"

derken "e" harfi ile yazılırken,

"araba+lar"

derken "a" ile yazılacaktır.

Tüm bu ses uyumu kuralları ve eklerin ayrılması morfolojik analiz çerçevesinde gerçekleşir. Bu durumun üstesinden gelebilmek için, ses uyumu kuralları ve ekler ayrı ayrı işleme tabi tutulur. Mesela, çoğul ekini göz önüne alırsak, tüm isimler için yalnızca tek çoğul eki varmış gibi işlem yapılır ve bu çoğul eki genel bir ifadeyle "lAr" olarak somutlaştırılır ve aradaki A harfinin hangi durumlarda a, hangi durumlarda e olarak çözümleneceği ses uyumu kurallarına göre belirlenir. Aynı şekilde, birliktelik eki "ylA" şeklinde ifade edilir ve y harfinin hangi durumda düşüp, hangi durumda kalacağı da gene ses uyumu kurallarına göre belirlenir.

Ses uyumu kuralları, hangi işaretin hangi durumda, ne şekle dönüşeceğini belirler. Mesela, yukarıda örneğini verdiğimiz A harfi için yazılacak bir kural şöyle olabilir: bu harften önceki sesli harfin "a, ı, o, u" harflerinden birisi olması halinde "a", "e, i, ö, ü" harflerinden birisi olması halinde "e" şeklinde yazılır. Buna göre "ev+lAr" ifadesinde A'dan önceki sesli harf "e" olduğu için, A e'ye dönüşecek ve "ev+ler" şeklinde yazılacaktır. Aynı şekilde, "okul+lAr" ifadesindeki A da önceki sesli harf o olduğundan a'ya dönüşecektir. Yine, yukarıdaki birliktelik anlamı veren "ylA" ekinde A harfi bu kurala göre çözümlenirken, y harfi için yazılacak kural ise, bu harfin sessiz harften sonra gelmesi halinde düşeceği ve sesli harften sonra gelmesi halinde korunacağıdır. Bu durumda, "saray+ylA" ifadesinde A harfi, önceki sesli harf a olduğundan a'ya dönüşecek, saray kelimesi de sessiz harfle bittiğinden y harfi düşecektir. Fakat, "yalı+ylA" ifadesinde yalı kelimesi sesli harf ile bittiğinden y harfi korunacak ve yalıyla şeklinde yazılacaktır.

Bazen köklerin ve o kökün arkasından gelebilecek eklerin sırasının belirlenmesi, bir kelime için tek bir sonuç elde etmekte yeterli olmaz. Mesela, “kalem” kelimesi “kalem+isim+3.tekilşahıs” şeklinde açılabilecekken, “bana ait olan kale” anlamında “kale+isim+3.tekilşahıs+iyelik1.tekil” şeklinde de açılabilir. Morfolojik çözümleme yalnızca kelimeler üzerinde işlem yaptığı, daha önce ve sonra gelen kelimelerin etkisini göz önüne almadığı için, iki açılımı da doğru kabul eder ve doğru açılımı bulmayı daha üst seviyelere bırakır.

2.4.2 Sözdizimi Çözümlemesi

Cümleyi oluşturan kelimeler arasındaki ilişkiler, kelime grupları, tamlamalar vs. belirlenir [23]. Mesela,

“benim kitabın”

şeklindeki bir tamlama, biçimbirimsel çözümlemeden başarıyla geçtiği halde, sözdizimi yönünden hatalıdır. Çünkü, benim kelimesinden sonra gelen kelimenin -ım/-im 1. tekil şahıs iyelik ekini alması gerekir. Oysa, bu cümlede 2.tekil şahıs iyelik ekini (-ın) alarak hatalı bir yapıya sebep olmuştur. Benzer şekilde,

“Öğrenci geldik.”

cümlesi de söz dizimi açısından hatalıdır. Çünkü, cümlenin öznesi (öğrenci) 3. tekil şahıs olduğu halde, fiilin şahsı 1. çoğul şahıs olarak söylenmiştir.

.2.4.3 Anlam Çözümlemesi

Anlamlandırma ve anlam kargaşasını çözümleme oldukça zor bir iştir ve yalnızca kelime ya da kelimenin içinde geçtiği metnin çözümlenmesi yeterli olmaz. Buna ek olarak, önceden öğrenilmiş bilgilerin kullanılması da gerekir (23). Mesela, kalem kelimesinin

“yazı yazmakta kullanılan araç” ve

“bana ait olan kale”

şeklinde iki çözümlemesi olabilir.

“Kalemle yazdım”

cümlesini ele aldığımızda, kalem kelimesinin her iki anlama gelmesi, morfolojik ve sözdizim kuralları açısından mümkün olduğu halde,

"bana ait olan kaleyi kullanarak yazı yazdım"

anlamını ifade edemeyeceğini ve çok büyük ihtimalle

"yazı yazmaya yarayan aracı kullanarak yazdım"

anlamına geleceğini söyleyebiliriz. Ancak, bu bilgiyi bu cümleden çıkartamayız. Daha önceden "kalem" ve "yazı yazmak" arasında bir ilişki olduğu öğrenilmiş olmalıdır. Aynı şekilde,

"Kalemi fethetti"

cümlesindeki kelimenin de "yazı yazmaya yarayan alet" değil de, çok büyük ihtimalle "bana ait olan, kalın duvarlar vasıtasıyla korunmaya yarayan yapı" anlamına geleceği sonucuna yine "kale" ve "fethetmek" arasındaki önceden bilinen bilgiyi kullanarak ulaşabiliriz.

Görüldüğü gibi belgelerdeki kelimelerin ayrıştırılması ve eklerinden arındırılması ve daha önceki birikimlerden elde edilen diğer kelimelerle birlikte kullanım şekilleri ve ilişkileri, belgelerin içeriğinin ortaya çıkarılmasında son derece önemli ve bir o kadar da çaba gerektiren bir süreçtir.

Tez konumuz çerçevesinde Türkçe belgelerin madenciliği için bir teknik geliştirilmiştir. Bu teknik içinde doğal dil işleme çalışmalarının bir parçası olan, belgeyi oluşturan kelimelerin ayrıştırılması ve köklerinden arındırılması önemli bir adım olarak yer almıştır.

3. BELGE MADENCİLİĞİ ALTYAPISI

Büyük belge yığınlarından insan becerileri ile elde edilemeyecek bilgilere ulaşmak için belge madenciliği teknikleri geliştirilmiştir ve geliştirilmektedir. Bu teknikler bir önceki bölümde anlatılmıştır. Belge madenciliği ile yapılabilecek bazı somut çalışmaları bir kez daha hatırlarsak (Hand vd., 2001) :

Özetleme : belgelerin içeriğinin kısaca anlatımı, özeti hazırlanabilir. Bu sayede bir belgeyi okumadan, hakkında bilgi sahibi olmak mümkündür. Etkin bir özetleme sistemi, insanlara büyük zaman kazandıracaktır. Özellikle belirli bir uzmanlık alanına yönelik belge yığınının özetleri sayesinde o alanda çalışan uzmanlar hızlı ve ciddi ilerlemeler kaydedebilir.

Sorgulama : belirli bir konuyla ilgili belgelerin bulunması amacıyla bir belge yığını otomatik olarak taranabilir. Özellikle hızlı belge üretilen günümüz koşullarında, belgelerin takibi mümkün olmamaktadır. En güncel belgelerin de yer aldığı bir belge havuzu içinden belirli bir amaca yönelik belgelere erişebilmek çok önemli hale gelmiştir.

Sınıflama : daha önceden bilinen sınıflara, eldeki belgelerin tasnifi, insan müdahalesi gerekmeden yapılabilir. Özellikle aktif çalışma alanlarında, sürekli yeni belgelerin farklı kaynaklar üzerinden üretildiği ortamlarda, çalışmaların seyrini kontrol edebilmek için, sınıflama çalışmalarına ihtiyaç duyulmaktadır.

Gruplama : verilen bir belge yığının, kendi içinde anlamlı bütün oluşturacak şekilde sınıflarına ayrıştırılması ihtiyacı, daha önce hiç üzerinde çalışılmamış belge kümelerinde ortaya çıkmıştır. Bu sayede belge kümesindeki tüm belgeleri tek tek incelemeden, belgeler kendi alt sınıflarına ayrıştırılabilir. Bu yöntem sayesinde de çalışanların hızlı ve etkin olmaları sağlanmış olmaktadır.

Belge madenciliği yöntemlerinin başarısı, belgelerin içeriklerini ne kadar iyi temsil edebildikleri ile doğru orantılıdır. Her belge madenciliği yönteminin, kendisine verilen bir belgeyi tanımlamak amacıyla kullandığı bir tekniği vardır. Tüm yöntemlerin kendilerine has teknikleri olmasına

rağmen, belge madenciliği için kullanılan teknikler iki ana dalda incelenebilir (Chakrabarti S., 2003).

- Eğitmenli sistemler
- Eğitmensiz sistemler

Eğitmenli sistemler, önce bir deney seti ile girdi ile çıktı arasındaki ilişkiyi öğrenen ve daha sonra kendisine verilen girdi için uygun çıktıyı üreten sistemlerdir. Bu gibi sistemlerin başarısı, öğrenme süreçlerinin doğruluğuna bağlıdır. Belge madenciliğinde belirli bir konuyu örnekleyen bir belge kümesi, eğitmenli bir sisteme verilerek, sistemin bu tip belgeleri öğrenmesi sağlanır. Daha sonra sisteme verilen belgeleri, sistem öğrendiği belgelerle karşılaştırarak benzer olup olmadıklarına karar verir. Bu sayede belirli bir alana yönelik üretilen belgeler, diğer belgelerdan ayrıştırılmış olur. Eğitmenli sistemlerin ilk eğitim süreci içermesi, sürekli genişleyerek büyüyen alanlara yönelik belge yığınlarına uygulanmaları zorlaştırmaktadır. Çünkü bu alan ait belgeleri örneklemeye yarayan sabit bir deney belge seti oluşturmak zordur. Bu gibi durumlarda sistemin esnekliğini sağlamak için eğitmensiz sistemler tercih edilir.

Eğitmensiz sistemler, eğitmenli sistemlerdeki, sistemin eğitilmesi için yapılan ön işlemleri içermez. Bu sistemler doğrudan belge seti üzerinde çalışır. Eğitmensiz sistemlerin her türlü belge yığını üzerinde hiçbir ön işlem yapmadan çalıştırılabilmesi, eğitmenli sistemlere göre büyük esneklik sağlar. Eğitmesiz sistemler, genel olarak tüm belgelere uygulanabildiğinden, eğitmenli sistemlerin belirli bir alana odaklanarak sağladığı başarıyı yakalamaktan uzaktırlar. Ancak bu alanda sürekli geliştirmeler yapılmaktadır ve tez kapsamında kullanılacak eğitmensiz sistem yaklaşımı buna bir örnek teşkil etmektedir. Eğitmensiz sistemlerin gelişmesinin önemli bir sebebi, sürekli gelişen bilgisayar işleme kapasiteleri ve veri depolama kapasiteleri sayesinde, tüm belgeleri örnekleyebilecek kadar büyük belge örneklemlerinin bu sistemlere aktarılabilmesi ve işlenebilmesidir.

Bu bölümün ilk konularında tez kapsamında geliştirilen eğitmensiz belge madenciliği sisteminin altyapısını oluşturan unsurlar anlatılacaktır. Daha sonra geliştirilen belge madenciliği sisteminin somut örnekler üzerindeki başarılı sonuçları aktarılacaktır.

3.1 Vektör Uzay Modeli

Belgelerin veri madenciliği algoritmaları ile analizinin otomatik olarak yapılabilmesi için belge içeriğinin bilgisayar üzerinde işlenebilecek, bir yapıya aktarılması gerekmektedir. Belgeler kendilerini oluşturan harfler, kelimeler, cümleler yada paragraflar halinde ifade edilebilirler. Bir belgenin içinde geçen harflerle ifade edilmesi anlamlı değildir. İçinde 100 adet a harfi, 150 adet b harfi, 300 adet c harfi geçen belge gibi bir ifade, belgenin hangi konuda yazılmış olduğunu ifade etmekten uzaktır. Harflere alternatif olarak bir belgenin anlamını ortaya çıkaracak en küçük parçalar kelimeleri oluşturan heceler olabilir. Ancak heceler de harfler gibi dil içindeki sesleri temsil ettiklerinden harflerden çok farklı bir güce sahip değildir. Bir dil içindeki kavramlar, varlıklar, eylemler, durumlar gibi tüm unsurlar kelimelerle ifade edilir.Bu nedenle bir belgeyi ifade edebilecek en küçük yapı taşı o belgeyi oluşturan kelimelerdir (Dumais vd., 1996) (Rehder vd., 1998).

Belge madenciliği uygulamalarında belgeler bilgisayar ortamında birer kelime dizisi olarak modellenir. Matematiksel olarak bu ifadenin karşılığı; her belge bir kelime vektörüdür. Tüm belgeleri ayrı ayrı diziler olarak göstermek mümkün olsa da, bir belge yığının bilgisayar üzerinde bir "kelime x belge" boyutlarına sahip iki boyutlu dizi olarak modellenmesi işlemleri daha kolaylaştırır. Bu nedenle bir belge madenciliği çalışması öncesi, belge yığını içindeki belgeler, bilgisayar ortamında oluşturulan, o belge yığınındaki tüm kelimeleri satırlarında taşıyan iki boyutlu diziye ayrı kolonlar olarak aktarılır (Berry vd., 1999)

Bu iki boyutlu dizi bir matristir. Bu matrise A matrisi dersek, matrisin elemanları (a_{ij}), d_i belgesi içinde geçen k_j kelimesinin belge analizi için taşıdığı önemi ifade eden sayısal bir rakamdır. Bu rakam en basit haliyle kelimenin belge içinde kaç kez geçtiğini gösterir (Manning vd., 1999).

Bir belge yığınının bu şekilde tanımlanmasıyla elde edilen yapı, vektör uzay modeli olarak bilinir.

Basit anlamda yukarıda anlatıldığı şekliyle oluşturulan vektörü uzay modelinde, belge yığınının bilgisayar ortamındaki karşılığı olan A matrisi, büyük belge yığınları söz konusu olduğunda üzerinde işlem yapmayı mümkün kılmayacak boyutlara ulaşabilir. Aynı zamanda bu matrisin

satırlarını oluşturan kelimelerin tamamı, o belgeleri ifade etmek için gerekli değildir. Örneğin dil içinde kullanılan bağlaçların bağımsız kelimeler olarak bir anlamı yoktur. Cümle içinde yer aldıklarında önemli hale gelirler. Ancak burada bahsedilen yaklaşımda her kelime ayrı olarak ele alındığından bağlaç gibi tek başına bir anlam ifade etmeyen kelimeler matrisin boyutlarını gereksiz yere büyütmektedir. Bu kelimelerin matristen çıkarılması bir sorun oluşturmayacak, hatta işlem yapmayı kolaylaştıracaktır.

Kelime X belge matrisindeki kelimelerin, belge yığınındaki belgelerin her birini ne kadar iyi ifade ettiği ile doğru orantılı olarak belge madenciliği uygulamasının da başarısı artmaktadır. Bu noktada üzerinde durulması gereken iki husus ortaya çıkmaktadır.

- İlki her kelimenin her zaman tek başına kullanılmadığı, bazı durumlarda tek başına belge içinde anlama katkı sağladığı, bazı durumlarda ise diğer kelime yada kelime gruplarıyla bir arada kullanılarak belge anlamına etki ettiğidir. Örneğin Amerika Birleşik Devletleri üç kelimenin birlikte ele alınmasını gerektirir. Bu ayrımı yapmak, doğal dil işleme tekniklerinin kullanımı ile mümkündür ancak mevcut Türkçe doğal dil işleme çalışmaları kapsamında bu konu cevaplandırılmış değildir. Bu husus tez çalışması içinde sistematik bir yaklaşımla aşılmaya çalışılmıştır ve tezin özgünlüğünü sağlayan en önemli noktalardan biridir.
- İkinci husus kısaltmaların anlamlı hale getirilmesidir. Örneğin belge içinde geçen T.B.M.M kısaltmasının Türkiye Büyük Millet Meclisi haline dönüştürülmesi, o belge yığını içinde, başka bir belgede geçen Türkiye Büyük Millet Meclisi ifadesi ile aynı seviyeye getirilmesini ve bu iki belgenin birbiri ile ilişkili olabilme durumunun ortaya çıkarılmasını sağlar. Tez kapsamında kullanılan ve daha sonra detaylı olarak açıklanacak Gizli Anlambilimsel Dizinleme yöntemi ile bu sorun ek işlem yapmadan giderilmektedir.

A matrisinde satırları oluşturan kelimelerin, olabildiğince belge yığınındaki belgeleri ifade etmesini sağlayacak diğer bir gereksinim, aynı kelimenin dil içinde kullanılırken dil bilgisi kurallarına göre çekim ekleri, çoğul ekleri, iyelik ekleri gibi eklerle sanki farklı kelime haline getirilmesi konusunun çözümlenmesidir. Bunun çözümü için her dilin kendine has yapısını baz alan yöntemler geliştirilmiştir. Türkçe için doğal dil işleme çalışmalarından yararlanmak gerekmektedir. Bu hususta yapılanlar tezin ileriki bölümlerinde detaylı olarak anlatılacaktır.

Belge yığınını tanımlayacak şekilde doğru kelimelerin seçilmesi ve kelimelerin düzenlenmesi neticesinde, vektör uzay modeli (A matrisi) üzerinde daha kolay ve anlamlı işlem yapmak mümkün olacaktır. Ancak matrisin her elemanına karşılık gelen a_{ij}, (d_i belgesi içinde geçen k_j kelimesinin belge analizi için taşıdığı önemi ifade eden sayısal bir rakam) çok önemlidir. Yukarıda bu rakamın basit olarak d_i belgesi içinde geçen k_j kelimesinin sayısı olabileceği ifade edilmişse de bu şekilde elde edilen sayı ile anlamlı belge madenciliği sonuçları elde etmek mümkün değildir. Örneğin bir kelime tüm belgelerde birer kez geçiyor ise, o kelimenin, o belge yığını içindeki belgeleri temsil etme gücü yok denebilir. Bu konuda da farklı yaklaşımlar ortaya atılmıştır. Bu yaklaşımlar ve tez çalışması içinde kullanılan yaklaşım da bu bölümde anlatılacaktır.

Vektör uzay modeli literatüründe matristeki kelimeler, terimler olarak ifade edilirler. Bu nedenle tez içinde terimler olarak kullanılacaklardır.

Sonuç olarak, elde edilen vektör uzay modeli sayesinde

- belge vektörleri arasındaki geometrik ilişkiye bakarak, belgelerin benzerlikleri ve farklılıkları bulunabilir.
- terim vektörleri geometrik olarak karşılaştırılarak, terimlerin kullanım benzerlikleri ve farklılıkları tesbit edilir.

Bu ölçütler hem belgeleri sorgulamak hem de belgeleri demetlemek amacıyla kullanılacaktır. Vektör uzay modelinin matematik altyapısı ile bilgiler bir sonraki bölümde verilerek, modelin nasıl işlediği anlatılacaktır.

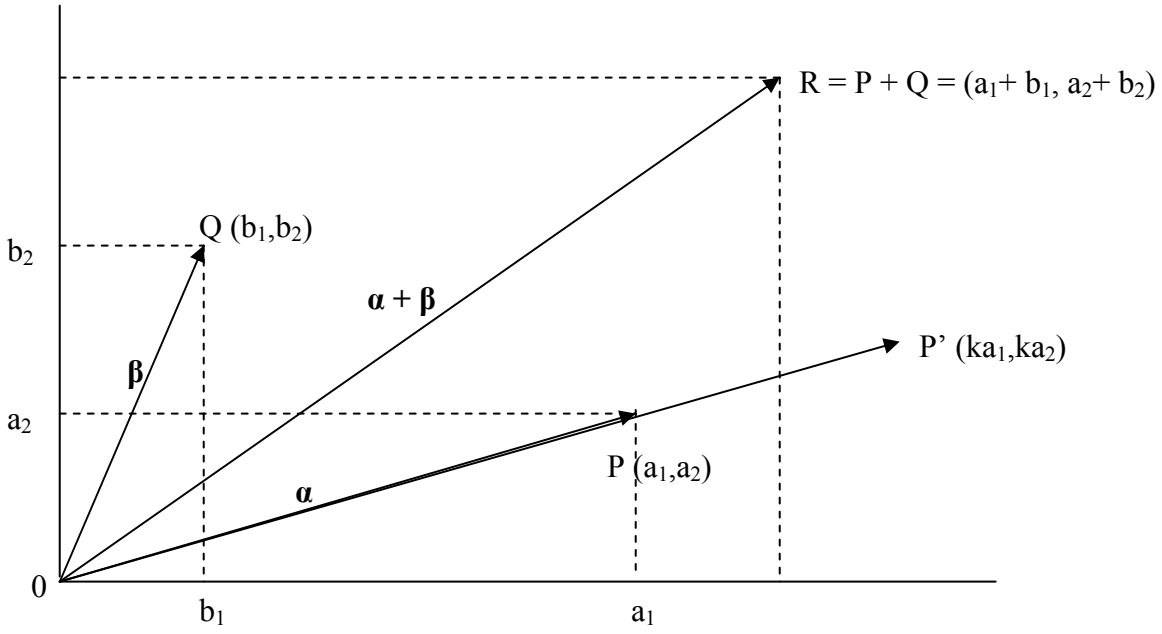
3.2 Vektör Uzay Modelinin Matematiksel Anlamı

Vektör uzayı, bileşenleri vektörler olan bir uzaydır. Vektör uzayı matematiğin lineer cebir alanında incelenir. Bu bölümde vektörler ve vektörlerin oluşturduğu vektör uzayı kavramları üzerinde durulacak, vektör uzayında vektörleri birbiriyle ilişkilerini ifade eden tanımlamalar ve işlemler anlatılacaktır.

3.2.1 Lineer Cebir Teorisi

3.2.1.1 Vektörler

Vektör, büyüklüğü ve yönü olan bir büyüklüktür (Kreyszig, 1999).
Vektörleri ikili koordinat sisteminde aşağıdaki gibi gösteririz-Şekil 3-1:



Şekil 3-1 Vektör Uzayı

α vektörünü

OP şeklinde gösterebiliriz

Büyüklüğü = $\sqrt{a_1^2 + a_2^2}$

Yönünü $\tan\theta = \frac{a_1}{A_2}$, θ x ekseni ile vektör arasındaki açıdır.

Bu bilgiler ışığında görülmektedir ki, vektörle ilgili tüm bilgiler a_1 ve a_2 değerlerinden türetilmektedir. Bu nedenle vektörü $P(a_1,a_2)$ şeklinde ifade etmek mümkündür.

Aynı şekilde **α** vektörünü bir sayıl değer ile çarptığımızda veya başka bir vektör ile topladığımızda da çıkan neticeyi, iki reel sayı ile ifade edebiliriz:

$P' = k\,\alpha = (ka_1, ka_2)$

$R = P + Q = \alpha + \beta = (a_1 + b_1,\ a_2 + b_2)$

Sonuç olarak koordinat sisteminden faydalanarak vektörleri iki reel sayı ile ifade eder hale geldik.

3.2.1.2 Vektör Uzayı

Vektör uzayının matematiksel tanımı, belirli şartları taşıyan vektörlere dayanır. Buna göre :

V vektörler kümesi ve R reel sayılar kümesi ise (Kreyszig, 1999) :

Tüm

$v \in V$ ve $a,b \in R$ için

$a\,(v1+v2) = a\,v1 + a\,v2$

$(a + b)\,v1 = a\,v1 + b\,v1$

$(a\,b)\,v1 = a\,(b\,v1)$

$1\,v_1 = v_1$

$0\,v_1 = 0$ (0 vektörü)

Bu şartları sağlayan vektörlerin oluşturduğu uzaya vektör uzayının yanında, lineer uzay ya da lineer vektör uzayı da denir.

3.2.1.3 Vektör Uzayında Lineer Bağımsızlık ve Lineer Bağımlılık

$a_1 v_1 + a_2 v_2 + \ldots\ldots\ldots\ldots a_n v_n = 0$

yukarıdaki eşitliği sağlayan a_i değerleri varsa v_i vektörleri lineer bağımlıdır denir. Bu eşitlik ancak a_i değerleri 0 için söz konusu ise v_i vektörleri lineer bağımsızdır (Kreyszig, 1999).

Lineer bağımlılığın anlamı, lineer bağımlı vektörlerin her birini diğerleri cinsinden ifade edebilmektir.

$V_i = -\ (1 / a_i)\ (a_1 v_1 + \ldots\ldots\ldots\ldots + a_{i-1} v_{i-1} + a_{i+1} v_{i+1} + \ldots\ldots\ldots\ldots + a_n v_n = 0)$

3.2.1.4 Temel ve Koordinatlar

Vektör Uzayı Temeli

Bir V vektör uzayının temeli, bu vektör uzayını tarayan, birbirinden lineer bağımsız vektörler kümesidir (Kreyszig, 1999). Bu vektörlerin sayısı sonlu ise, V sonlu boyutludur (finite dimensional)

Vektör uzayında koordinatlar

V vektör uzayında her vektör lineer bağımsız vektörlerin (temelin) bir lineer birleşimi olarak ifade edilebilir. Yani bir A vektörü lineer bağımsız a_i vektörleri cinsinden

$$A = i\, a_1 + j\, a_2 + k\, a_3$$

Bu V vektör uzayında A gibi vektörleri, ikili koordinat sisteminde $P(x_1,y_1)$ vektörünü gösterdiğimiz gibi gösterebiliriz (Kreyszig, 1999).

A (i, j, k) gösterimi A vektörünü ifade eder. $A = i\, a_1 + j\, a_2 + k\, a_3$ olduğuna göre diyebiliriz ki, a_1, a_2, a_3 vektörleri koordinatlardır.

3.2.1.5 Ortogonallik

İki vektör birbirine dik ise, vektörler ortogonaldir denir. Bunun matematiksel ifadesi ise, iki vektörün standart iç çarpımı sıfır olduğunda, vektörler diktir şeklindedir. Lineer bağımsız vektörler ortogonaldir (Kreyszig, 1999).

3.2.1.6 Normalleştirme

Vektörlerin büyüklüğü göstermek için farklı ölçütler kullanılır. Bunlardan birisi ve en yaygın olanı vektör normudur.

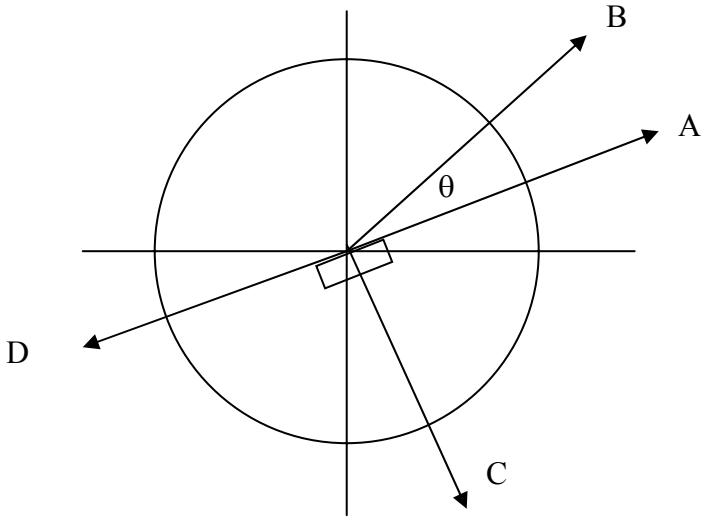
Bir v (x,y) vektörünün normu norm v veya $||\,v\,||$ şeklinde gösterilir:

$$||\,v\,|| = \sqrt{x^2 + y^2}$$

Bir vektörün normunun 1 olması için vektör elemanları değerlerine göre oranlanabilir. Bu durumda vektör normalleştirilmiş olur (Kreyszig, 1999). Normları 1 olan iki vektör ortogonal ise, ortonormal adını alır.

3.2.1.7 Kosinüs Açısının Özellikleri

Koordinat sisteminde oluşturulmuş aşağıdaki gösterim ile (Şekil 3-2) açılar arasındaki ilişkileri incelemek mümkündür (Kreyszig, 1999).



Şekil 3-2 Koordinat Sistemi

Buna göre;

$\cos\theta\ (AB) = \quad > 0$

$\cos\theta\ (AC) = \cos 90 = 0$

$\cos\theta\ (AD) = \cos 180 = -1$

$\cos\theta\ (AA) = \cos 0 = 1$

Belge madenciliğinde, her belgenin birer vektör olarak ifade edildiğini belirtmiştik. İki vektörün birbirine paralel olması, ikisin aynı olduğunu gösterir ki iki belge vektörü de birbirine paralel ise birbiri ile aynıdır. Bu ifadenin matematiksel karşılığı, iki vektörün arasındaki kosinüs açısının

değerinin 1 olmasıdır. Belgelerin birbirine ne kadar yakın olduğu bu bilgiye bağlı ortaya çıkarıldığından kosinüs açısının özellikleri belge madenciliğinde önemlidir.

3.3 Vektör Uzayında işlem

Şimdiye kadar yapılan tanımlamalar ışığında bir belge yığını üzerinde işlem yapabilmek için o belge yığınını bir matematiksel model üzerine aktarıp, belgeler arası ve belgelerin kendi içinde yapmak istediğimiz işlemleri yapabiliriz.

Daha önce de belirtildiği gibi belgeler, anlamlarını kaybetmeyecek en küçük bileşenleri olan kelimelere ayrıştırılabilir. Bir belge yığınındaki tüm belgeler, kelimelerine ayrıştırıldıktan sonra ortaya çıkan kelimeler bir koordinat uzayı gibi ele alınırsa, her belge bu koordinat uzayında bir vektör olacaktır.

Kelimeler, dil bilgisi kurallarına göre, anlamlarını kaybetmeden farklı biçimlerde kullanılabilirler. Bunun sebebi kelimeye gelen iyelik ekleri, çekim ekleri gibi dil bilgisi kurallarıdır. Aynı anlamı taşıyan farklı biçimlere sahip kelimelerin ortak bir biçime dönüştürülmeleri gerekir çünkü vektör uzayında belgeleri ifade etmek için vektör uzayını oluşturan koordinatların, yani kelimelerin birbirine dik olması gerekir. Bu nedenle kelimeler eklerinden ayrıştırılarak anlamını kaybetmeyecekleri en yalın biçimlerine dönüştürülürler. Kelimelerin bu hallerine terim denir.

Bu bilgiler ışığında bir belge yığını içinde d adet belge varsa ve bunlar t adet terim ile ifade ediliyor ise, her belge

$$d_i \text{ x t vektörü}, \quad i = 1,2,3\ldots\ldots..d$$

ile gösterilebilir. Buradan yola çıkarak tüm belgeler tek bir gösterimle, d x t matrisi ile gösterilebilir.

Bu bizim vektör uzayımızdır. Matrisin kolonları belgelerdir. Matris üzerindeki a_{ij} elemanı i. belgede yer alan j. Terimin belge için önem derecesini veren sayısal bir ifadedir. Terimlerin her belge için ayrı ayrı önem derecelerini gösteren sayısal değeri bulmak için kullanılan

yöntemlerden en yaygını, ilgili terimin belge içindeki frekansıdır. Bu konuya daha sonra detaylı olarak değinilecektir.

"d x t" gösterimi kullanarak her bir terim için tüm belgeler içindeki yeri ve anlamı hakkında yorumda bulunmak mümkün değildir ama bu gösterim elimizdeyken belge vektörlerinin birbirine göre geometrik konumlarına bakılarak benzerlikleri yada farklılıkları hakkında bilgi sahibi olabiliriz. Aynı şekilde terim vektörlerinin birbirine göre geometrik konumları incelendiğinde terimlerin kullanım farklılıkları yada benzerlikleri hakkında yorumda bulunmak mümkündür. Fark edildiği üzere bahsedilen işlemlerin matematiksel karşılığı vektör uzayında yapılan açı hesaplamalarına dönüşür.

Vektör uzayında (matris üzerinde) mevcut belgeler ve terimler arası ilişkileri elde etmek mümkün iken, belirli bir konu hakkındaki belge yada belgeleri sorgulamak mümkün değildir. Vektör uzayında birbirine anlam olarak yakın belgelerin, aralarındaki kosinüs açısına bakarak tesbit edilebileceğini belirtmiştik. Sorgulama işleminde de benzer bir gereksinim vardır. Bir sorgu ifadesi ile belgeler arasındaki anlamsal yakınlık bulunmaya çalışılmaktadır. Bu durumda sorgu ifadesini, vektör uzayında bir vektör haline getirip, d x t matrisine ekleyerek istenilen işlemi yapmak mümkündür. Sorgulama işleminin sonucu, vektör uzayında "sorgu vektörü" ne en yakın "belge vektörleri" ni arama işleminin neticesidir.

Geometrik uzayda vektörler arası yakınlığı ölçmek için kullanılan yöntem, vektörler arası kosinüs değerine bakmaktır demiştik. Bu değer ne kadar az ise belgeler birbirlerine o kadar yakındır. Bu yöntemin matematiksel gösterimi :

Belge vektörleri a_i ile gösterilirse,
j=1,2,…….,d olmak üzere
tüm belgelerin q sorgu vektörü ile arasındaki kosinüs açısı aşağıdaki gibi ölçülür

$$\cos\theta_j = \frac{a_i^T\, q}{||a_j||_2 \quad ||q||_2} = \frac{\sum_{i=1}^{t} a_{ij}\, q_i}{\sqrt{\sum_{i=1}^{t} a_{ij}^2}\ \sqrt{\sum_{i=1}^{t} q_i^2}}$$

$||x||_2$ öklid vektör normudur. Norm, vektör büyüklüğü için kullanılan bir büyüklük birimidir ve

$$\sqrt{X^T X}$$

ile hesap edilir.

Belirli bir eşik değerinin üzerinde kalan kosinüs açılarına sahip belgeler sorgu neticesi olarak döner. Pratikte bu değer 0,9'dur.

3.3.1 Vektör Uzay Modelinin Kısıtları

Belge setindeki belgelerin sayısı arttıkça ve belgelerin kendi boyutları büyüdükçe, bu tür belge setlerinin vektör uzay modeli ile modellenmesi sonucu elde edilen "belge X terim" matrisinin boyutları büyür. Bu büyüme sebebi ile bu matris üzerinde belgelerin birbirleri ile olan ilişkilerinin sorgulanması veya bir sorgu ifadesine uygun belgelerin aranması işlemleri daha da zorlaşır. Öte yandan büyüyen matriste, terim sayısı çok olacak ancak bu terimlerin her bir belgede sadece bir kısmı geçtiği için matrisin büyük bir kısmı 0 değerli hücrelerden oluşacaktır. Yani matrisin büyüklüğünün belge setini daha iyi modellemesi ile ilişkisi yoktur (Rosario, 2000).

Matris cebiri içinde, matrislerin orijinal halinin daha az boyutla ifade edilmesini sağlayan yöntemler bulunmaktadır. Bu yöntemler temel olarak orijinal matrisi oluşturan vektörleri oluşturmak için kullanılabilecek ana vektörleri (faktörleri) bulma prensibine dayanır. Bu işleme Temel Bileşen Analizi adı verilir. Bu işlem için doğrusal cebir içinde kullanılan yöntemler

- QR Faktörlerine ayırma
- Tekil Değer Ayrıştırması'dır.

Sırasıyla TBA yaklaşımı ve TBA yapmayı sağlayan QR Faktörlerine ayırma veTekil Değer Ayrıştırma yöntemleri incelenerek Tekil Değer Ayrıştırmasının avantajları ortaya konacaktır. Tekil değer ayrıştırması, tez çalışmasının ana bileşenlerinden biridir.

3.3.2 Vektör Uzay Modelini Geliştirme – Temel Bileşen Analizi

Faktör analizi orijinal bir matrisinin boyutunu, ortogonal faktör uzayı kullanarak küçültmeye yarayan bir tekniktir ve çok boyutlu (değişkenli- multivariate) veriler için kullanılır. Veri içindeki

temel bileşenleri, diğer bir değişle belirgin faktörleri bulup, bu faktörlerle veriyi yeniden modellemeyi hedefler [17] (Suh vd., 2002).

Örneğin s1 = (4, 3, 6), s2 = (2, 3, 2), s3 = (2, 0, 4) vektörlerinin matrisi aşağıdaki gibidir:

$$D = \begin{pmatrix} 4 & 2 & 2 \\ 3 & 3 & 0 \\ 6 & 2 & 4 \end{pmatrix}$$

s1, s2 ve s3 vektörleri 3 boyutludur. Ancak 3 boyutlu uzayı tarayamazlar çünkü bu üç vektörün lineer birleşiminden bir v vektörü oluşturursak, v vektörü aşağıda gösterildiği gibi 2 boyutlu düzlemde bulunur

$$s3 = s1 - s2 \quad \rightarrow \quad v = a.s1 + b.s2 + c.s3$$
$$= a.s1 + b.s2 + c.(s1-s2)$$
$$= (a+c).s1 + (b-c).s2$$

Temel Bileşen Analizi (TBA) görüldüğü gibi, verilen çok boyutlu vektörlerin oluşturduğu uzayı, gerçekte, temel olarak nasıl oluşturulabileceğinin cevabını verir. TBA aynı zamanda "eiegen analizi" olarak da bilinir. Veri matrisi yeni r boyutlu vektörlere dönüştürülür. Bu yeni vektörler, orijinal matrisin kolonlarının taradığı uzayı tarar. Bu yeni vektörler ise iki parçadan oluşan bir biçime sahip modelle ifade edilebilirler. Bu parçalardan biri eigen değerleri, diğeri eigen vektörleridir[20]. Eigen değerleri orijinal veri içinden çıkarılan faktörlerin, orijinal verinin modellenmesinde ne decere etkin olduğunu belirler. Faktörlerin belirlenmesinde kullanılan yöntemin temeli lineer en küçük kareler yöntemine dayanır (Wall vd., 2002).

Matris boyutlarını düşürmek amacıyla faktörlerin hesaplanması için QR faktörlerine ayırma yada Tekil Değer Ayrıştırma (Singular Value Decomposition) yöntemleri de kullanılabilir demiştik. QR faktörlerine ayırma yöntemi bilgiye ulaşma sistemlerinde kullanılmamıştır. Her iki yöntem de matrisleri ortogonal (dik) faktörlerine ayırır ancak TDA, QR yöntemine göre daha etkilidir. QR yönteminde belgelerin kendi aralarındaki benzerlik karşılaştırılabilirken, TDA yönteminde hem belgelerin benzerliği hem de kelimelerin benzerliği ortaya çıkarılabilir (Chouchoulas, 2002).

3.3.3 QR Faktörlerine Ayırma

Vektör uzay modelinde t x d boyutlarında bir matrisle d adet belgeden oluşan bir belge yığınını t adet terim ile ifade ediyorduk. Bu matrisin kolonları üzerinde Gauss yoketme yöntemi kullanarak matrisin boyutunu (rank) düşürmeye çalışarak şunu yapmış oluruz; hangi belge vektörlerini diğer belge vektörlerinin lineer birleşiminden oluşturabileceğimizi bulur ve bunu matristen çıkarak lineer olarak bağımsız belgeleri buluruz. Lineer olarak bağımsız belgeler TBA literatüründe, faktörlerimizi temsil eder. Bir belgenin lineer olarak başka belgelerin birleşiminden oluşmasının pratikteki örneği internet web sayfalarında aynalanmış yani başka web sitelerinden kopyalanmış yada başka belgelerin birleşiminden oluşmuş olduğu durumunda karşımıza çıkar.

QR faktörlerine ayırma yöntemi, eldeki bir matrisin kolon uzayı için bir temel oluşturmayı sağlar. Vektör uzayında "temel" oluşturmak tanımı daha önce yapmıştık. Rank düşürme işleminin ilk adımı "terim X belge matrisi" ndeki belgeler veya terimler arasındaki bağımlılığı ortaya çıkarmaktır. Örneğin r_A rank değerine indirgenmiş "t x d" matrisi için şu ifade kullanılır; d adet kolon yerine matrisin kolon uzayı, r_A adet temel kolon vektörü ile gösterilebilir (Berry vd., 1999).

QR faktörlerine ayırma yöntemi A matrisini iki adet matrisinin çarpımı haline dönüştürür.

$$A = Q\,R$$

Buradaki Q "t x t" boyutlarında ortogonal bir matristir, yani matrisin kolonları geometrik uzayda birbirine diktir. Kolonların birbirine dik olduğunu anlamak için aşağıdaki işlemin geçerli olması gerekir :

$$\sqrt{q_j^T q^i} = 0 \quad \text{i eşit değildir j}$$

Ayrıca Q matrisinin satırları ve kolonları ortonormaldir. Yani birim öklid büyüklüğüne sahiptirler. Bu ifadenin matematiksel karşılığı

kolonlar için, $||q_j||_2 = \sqrt{q_j^T q_j} = 1,$

satırlar için, $X^T X = X X^T = I$ (I birim matristir.)

R matrisi ise üst köşegen bir matristir.

A=QR eşitliği tüm matrisler için geçerlidir. A,Q ve R matrislerinin arasında şöyle bir ilişki vardır.

A'nın kolon vektörleri Q'nun kolon vektörlerinin lineer birleşiminden meydana gelir,

Q'nun kolon sayısının r_A kadarı A'nın kolon uzayının temelini (basic) oluşturur.

Örnek ;

5 adet kitabın isminden oluşan bir belge seti için, 6 adet terim tesbit edilmiş olsun

Belgeler

d1 : Reçete olmadan ekmek nasıl pişirilir.

D2 : Pasta sanatı

d3 : Ürün reçeteleri şirket için çok önemlidir.

D4 : ekmekler, pastalar, kekler ve tartlar için ölçekli pişirme reçeteleri

d5 : En güzel Fransız pastaları reçeteleri

Terimler

t1 : pişir (me, ilir)bak(e,ing)

t2 : reçete (leri)

t3 : ekmek (ler)

t4 : kek (ler)

t5 : pasta (lar, ları)

t6 : tart

Terim belge matrisi, t x d

		d1	d2	d3	d4	d5
	t1	1	0	0	1	0
	t2	1	0	1	1	1
A =	t3	1	0	0	1	0
	t4	0	0	0	1	0
	t5	0	1	0	1	1
	t6	0	0	0	1	0

Bu matrisin her kolonu bir belgeyi tanımlar. Belgeler içerdikleri terimlerin göreceli ağırlıklarına göre tanımlanmalıdır. Biz ise sadece terim frekanslarını kullandığımızdan her kolonun birim uzunluğu 1 olacak şekilde (öklid uzunluğu) matrisi yeniden düzenleyerek terimlerin göreceli ağırlıklarını kullanmış oluruz :

	0,5774	0	0	0,4082	0
	0,5774	0	1,00	0,4082	0,7071
A =	0,5774	0	0	0,4082	0
	0	0	0	0,4082	0
	0	1,00	0	0,4082	0,7071
	0	0	0	0,4082	0

A = QR faktörlerine ayırma işlemi uygulanırsa

	-0.5774	0	-0.4082	-0.0000	-0.7071	0.0000
	-0.5774	0	0.8165	-0.0000	-0.0000	0.0000
Q =	-0.5774	0	-0.4082	0.0000	0.7071	-0.0000
	0	0	0	-0.7071	0.0000	-0.7071
	0	-1.00	0	0	0	0
	0	0	0	-0.7071	0.0000	0.7071

R =

-1,0001	0	-0,5774	-0,707	-0,40824
0	-1	0	-0,4082	-0,7071
0	0	0,8165	0	0,5773
0	0	0	-0,5774	0
0	0	0	0	0
0	0	0	0	0

A = QR için R'nin son iki satırının sıfır olduğu görünmektedir. Buradan yola çıkarak A = QR ifadesini aşağıdaki gibi yeniden yazabiliriz

$$A = (Q_A\ Q_{1/A}) \begin{bmatrix} R_A \\ 0 \end{bmatrix} = Q_A R_A$$

Bu ifade Q matrisinin son iki kolonunun A matrisini oluşturmada etkisi olmadığını göstermektedir. Bu özellik QR faktörlerine ayırma yönteminin bir özelliğidir. Bu özelliğin diğer bir anlamı da, belge setinin anlamsal içeriğinin terim x belge matrisinin kolon uzayının temeli(basic) ile (Q_A) tam olarak kapsandığıdır (Berry vd., 1999).

Bu belgeler içinden bir belgeyi arama işleminde bir sorgu ifadesi kullanılır. Bu sorgu ifadesi de bir belge gibi ele alınarak, belge vektörleri ile sorgu vektörü arasındaki uzaklık hesaplanır ve en yakın belgeler bulunur. Bu yakınlık için belge ile sorgu vektörü arasındaki kosinüs açısından faydalanılabilir. Bu açıya θ dersek, q sorgu vektörü, a_j j. belge, Q_A kolon uzayının temeli ve

$$a_j = Q_A r_j$$

(r_j, R matrisi-nin j. kolonu) olmak üzere, j = 1....d 'ye kadar (Berry vd., 1999)

$$\cos\theta_j = \frac{a_j^T q}{||a_j||_2\ ||q||_2} = \frac{(Q_A r_j)^T q}{||Q_A r_j||_2\ ||q||_2} = \frac{r_j^T (Q_A^T q)}{||r_j||_2\ ||q||_2}$$

$$||Q_A r_j||_2 = \sqrt{(Q_A r_j)^T (Q_A r_j)} = \sqrt{r_j^T Q_A^T Q_A r_j}, \quad Q_A^T Q_A = I$$

$$= \sqrt{r_j^T I\, r_j}$$

$$= ||\, r_j \,||_2$$

3.3.3.1 QR - Kosinüs Açısının Hesaplanması

Bir ortogonal Q matrisi için aşağıdaki eşitlik doğrudur.

$$Q^T Q = I = (Q_A\, Q_{1/A})^T (Q_A\, Q_{1/A}) = Q_A\, Q_A^T + Q_{1/A}\, Q_{1/A})^T$$

q sorgu vektörünü q = I q olarak yazarsak ve I yerine yukarıdaki eşitliği koyarsak

$$q = [Q_A\, Q_A^T + Q_{1/A}\, Q_{1/A})^T]\, q = Q_A\, Q_A^T q + Q_{1/A}\, Q_{1/A})^T q$$

$$\Rightarrow q_A = Q_A\, Q_A^T q \;, \quad q_{1/A} = Q_{1/A}\, Q_{1/A})^T q \text{ dersek}$$

$$q = q_A + q_{1/A}$$

Bu şekilde q vektörü kendi iki bileşenine bölünmüş olur. q_A, A'nın kolon uzayındaki (Q_A'daki) bileşen, $q_{1/A}$ A'nın kolon uzayının tamamlayıcı uzayındaki ($Q_{1/A}$'daki) bileşendir. Bizim A matrisi için temel kolon uzayı olarak elde ettiğimiz Q_A uzayında q sorgu vektörünün karşılığını bulmadan, sorgu vektörü ile belge vektörlerini karşılaştırmak, doğru sonuç vermez. Bu nedenle q sorgu vektörü $q_A + q_{1/A}$ bileşenlerine ayrıştırılmıştır ve q'nun q_A bileşeni Q_A uzayında q'nun ortogonal izdüşümü (orthogonal projection) olarak adlandırılır. Bu bileşen q vektörünün A'nın kolon uzayındaki en yakın karşılığıdır. $q_{1/A}$ bileşeni ise $Q_{1/A}$ uzayındaki, sorgulama işlemine etki etmeyen kısımdır. Belgelerle karşılaştırılacak q sorgu vektörünün asıl bileşeni kosinüs hesaplamasındaki yerine yerleştirilirse (Berry vd., 1999) :

$$\cos\theta_j = \frac{a_j^T\, q}{||a_j||_2\, ||q||_2} = \frac{a_j^T q_A + a_j^T q_{1/A}}{||a_j||_2\, ||q||_2} = \frac{a_j^T q_A + a_j^T\, Q_{1/A}\, (Q_{1/A})^T\, q}{||a_j||_2\, ||q||_2}$$

a_j vektörü A 'nın bir kolonu olduğundan, $Q_{1/A}$ kolonlarına diktir, yani $a_j^T\, Q_{1/A}=0$ dır. Bu durumda yukarıdaki formül aşıdaki hale getirilebilir

$$\cos\theta_j = \frac{a_j^T q_A}{||a_j||_2 \, ||q||_2} \qquad (3.1)$$

Burada yapılan q_A dönüşümü ile q sorgu vektörü, A matrisi içinde yapılacak sorgulama için, terim uzayı kullanılarak en uygun hale getirilmektedir.

3.3.4 Düşük Rank ile Yakınsamanın Doğruluğu

QR faktörlerine ayırma yöntemi ile A matrisinin daha az boyutlu bir karşılığı oluşturuluştur. Elde edilen Q ve R matrislerinin A matrisini ne kadar yakınsayabildiği önemlidir. Daha düşük boyutlu matrislerle çalışmanın getirdiği performans avantajları sunan bu yakınsamanın çok daha önemli bir katkısı da vardır. A matrisi faktörlerine ayrıştırıldığından elde edilen Q, R matrisleri A matrisi içinde bulunan kirlilikleri de temizlemektedir. Bu kirlikler, örneğin bir belge yığını içindeki belgelerin farklı kişiler tarafından indekslenirken farklı terimlerle indekslemesi olabilir (Johnston R., 2001).

Daha düşük boyutlu bir matrisle A matrisinin ifade edilmesi durumunda, A matrisinin içeriğinin ne kadarının kaybedildiğini yada içeriğindeki kirliliklerin ne kadarının giderildiğini ölçmek için A matrisini içindeki belirsizlikleri de kapsayacak şekilde A+E matrisi şeklinde tanımlamak mümkündür. E matrisi, belirsizlik matrisi olup, A matrisi elemanlarının eksik yada fazlalıklarını taşır. Burada vurgulanan argüman, belge yığınını ifade etmek için A matrisinin, olası birçok alternatiften birisi olduğudur (Berry vd., 1999) (Berry vd., 1995).

A matrisini daha düşük boyutta ifade eden Q ve R matrislerinin yanında daha da düşük boyutlu yakınsamaları oluşturulabilir. Doğrudan A'nın boyutlarını düşürmek için Q,R matrislerinin boyutları düşürülebilir. Örneğin R matrisin ele alırsak, R matrisi bir üst üçgen matristir ve rank'i A ile aynıdır. Yani R'nin rank'ini düşürmekle A'nın da rank'ini düşürmüş oluruz. R matrisini seçmemizin nedeni, R matrisinin rankini hesaplamanın daha kolay olmasıdır; sıfır olmayan satırların toplamıdır. Kolona göre eksenleme (column pivoting) yöntemi ile R matrisini büyük değerler sol üst köşeye, küçük değerler sağ alt köşeye gelecek şekilde düzenlemek mümkündür.

Bu şekilde R matrisini düzenledikten sonra R matrisini, küçük parçayı ayıkmayı sağlayacak şekilde parçalama mümkündür (Berry vd., 1999):

R =

-1,0001	0	-0,5774	-0,707	-0,40824
0	-1	0	-0,4082	-0,7071
0	0	0,8165	0	0,5773
0	0	0	-0,5774	0
0	0	0	0	0
0	0	0	0	0

$$= \begin{pmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{pmatrix}$$

R'nin R_{11} parçası ile QR yani A'nın daha düşük boyutlu yakınsaması oluşturulabilir. Bu şekilde elde edilen yakınsamanın getiri ve götürüsünün ölçümü için matris boyutunu hesaplamak gerekecektir. Vektör uzunluklarını ölçmekte kullanılan Öklid normu matrislere de uygulanabilir ve matrisler için kullanıldığında Frobenius matris normu adını alır. Frobenius matris normu aşağıdaki gibi hesaplanır:

$$||A||_F = \sqrt{\sum_{i=1}^{t} \sum_{j=1}^{d} a_{ij}^2} = \sqrt{\text{Trace}(A^T A)}$$

Frobenius matris normu, matris izi (Trace) cinsinden de ifade edilebilir

$$||A||_F = \sqrt{\text{Trace}(A^T A)} = \sqrt{\text{Trace}(A A^T)}$$

Toplam R matrisi içinde R_{22}'nin payının ne olduğunu ve bu şekilde A matrisi yerine kullanılacak QR yakınsamasında R_{22}'nin yer almaması durumunda ne kadarlık bir kayıp olacağı ölçülebilir.

$$\frac{||R_{22}||_F}{||R||_F} = \frac{0.5774}{2.2361} = 0.2582$$ oranının küçüklüğünden yola çıkarak

R_{22}, R matrisinin küçük bir bölümüdür diyebiliriz ve değerini sıfır (0) kabul edebiliriz. Bu durumda R, onun bir yakınsaması olan Ř olur. A+E = QR eşitliği de

A+E = Q Ř haline gelir. Buradan E belirsizlik matrisi E = (A+E) – A ‘ dan

$$= Q \begin{bmatrix} R_{11} & R_{12} \\ 0 & 0 \end{bmatrix} - Q \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix} = Q \begin{bmatrix} 0 & 0 \\ 0 & -R_{22} \end{bmatrix}$$

Dikkat edilirse $||E||_F = ||R_{22}||_F$ dir. Bu bilgi, E belirsizlik matrisinin A matrisine yaptığı katkı ile R_{22} matrisinin R matrisine yaptığı katkı benzer ve aynıdır şeklinde bir anlam oluşturmamızı sağlar. Zaten A matrisi yerine, A+E matrisini kullanmaya karar verdiğimizden, bunun karşılı olan QŘ matrisini kullanmayı da seçmiş oluruz. Bu kullanımın pratikte kabul görüp görmeyeceği, bir sorgu vektörü ile QŘ matrisinin karşılaştırılması ile mümkündür.

3.3.5 Tekil Değer Ayrıştırması - Singular Value Decomposition (SVD)

Bir önceki bölümde vektör uzay modeline QR faktörlerine ayırma yöntemi uygulayarak, bir belge setinin içeriğini indekslemek için kullanılan terim x belge matrisinin kolon uzayında boyut indirimi yapılabileceği gösterdik. Kolon uzayı, belgeleri ifade eder ve bu uzaydaki belge vektörleri arasındaki ilişkiyi ortaya koymak için bir yöntem sunar. Oysa terim x belge matrisinin satır uzayı da vardır. Satır uzayı, terimleri ifade eder ve bu uzayda terimler arasındaki ilişkileri bulmak da mümkündür (Berry vd., 1999) (Berry vd., 1995).

SVD yöntemi hem kolon, hem de satır uzayında indeksleme yapma ve her iki uzayda da boyut indirgeme yapma imkanı veren bir tekniktir. QR faktörlerine ayırma yöntemindeki gibi, SVD ile de bir A terim x belge matrisi S,V ve D matrislerinden oluşan faktörlerine ayrılır :

$$A = S\ V\ D^T$$

S : t x t ortogonal bir matris olup kolonlarında A’nın sol tekil vektörlerini (left singular vectors) taşır.

D : d x d ortogonal bir matris olup kolonlarında A’nın sağ tekil vektörlerini (right singular vectors) taşır.

V : t x d köşegen bir matris olup, köşegen elemanları $\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \ldots\ldots\ldots \geq \sigma_{min(t,d)}$ olacak şekilde A'nın tekil değerlerini (singular values) taşır.

Terim sayısı t, belge sayısı d'den büyük bir A matrisi için

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots\ldots & a_{1d} \\ a_{21} & a_{22} & & \\ . & & & \\ . & & & \\ a_{t1} & \ldots\ldots & & a_{td} \end{pmatrix} = \begin{pmatrix} s_{11} & s_{12} & \ldots\ldots & s_{1r} \\ s_{21} & s_{22} & & \\ . & & & \\ . & & & \\ s_{t1} & \ldots\ldots & & s_{tr} \end{pmatrix} \begin{pmatrix} v_{11} & & & \\ & v_{22} & & \\ & & . & \\ & & & v_{rr} \end{pmatrix} \begin{pmatrix} d_{11} & d_{12} & \ldots\ldots & d_{1d} \\ d_{21} & & & \\ . & & & \\ d_{r1} & \ldots\ldots & & d_{rd} \end{pmatrix}$$

A matrisinin rank'i sıfır (0) olmayan tekil değerlerinin (v matrisinin sıfır(0) olmayan köşegenlerinin) sayısına eşittir ve r_A olarak isimlendirilirse

$$||A||_F = ||SVD^T||_F = ||VD^T||_F = ||V||_F = \sqrt{\Sigma_{j=1}^{rA} \sigma_j^2}$$

QR faktörlerine ayırma yöntemine benzer şekilde, S matrisinin ilk r_A kolonu A matrisinin temel(basic) kolon uzayını, D^T matrisinin ilk r_A satırı da A matrisinin temel satır uzayını oluşturur. Yine QR faktörlerine ayırma yönteminde A matrisi için daha düşük rank'e sahip bir karşılık bulunduğu gibi, SVD yönteminde de r_A rank'inden daha düşük bir rank'e sahip ve A matrisinin yerini alabilecek bir matris aşağıdaki gibi oluşturulabilir

k ($k < r_A$)boyutunda rank'e sahip bir A_k matrisi elde etmek için V matrisinin en büyük k adet köşegeni dışındaki köşegen değerleri sıfır (0) yapılır ve SVD çarpımı hesaplanır.
A matrisinin A_k yakınsaması ile arasındaki fark minimum olmalıdır çünkü A_k, A yerine kullanılır. Gerçekte de bu durum doğrudur. Eckart ve Young tarafından ortaya konan teoreme göre, bu iki matris arasındaki fark, A'nın tekil değerlerine bağlıdır:

$$||A - A_k||_F = \sqrt{\sigma_{k+1}^2 + \ldots + \sigma_{rA}^2}$$

$$A_k = S_k V_k D_k^T$$

S : t x k bir matris olup S'nin ilk k kolonundan oluşur.

D : d x k bir matris olup D'nin ilk k kolonundan oluşur.

V : k x k bir matris olup A'nın en büyük ilk k tekil değerinden oluşur.

QR faktörlerine ayırma örnek çalışmasını, SVD için yaparsak :

Belgeler

d1 : Reçete olmadan ekmek nasıl pişirilir.

d2 : Pasta sanatı

d3 : Ürün reçeteleri şirket için çok önemlidir.

d4 : ekmekler, pastalar, kekler ve tartlar için ölçekli pişirme reçeteleri

d5 : En güzel Fransız pastaları reçeteleri

Terimler

t1 : pişir (me, ilir) bak(e,ing)

t2 : reçete (leri)

t3 : ekmek (ler)

t4 : kek (ler)

t5 : pasta (lar, ları)

t6 : tart

Terim belge matrisi, t x d

A =

1	0	0	1	0
1	0	1	1	1
1	0	0	1	0
0	0	0	1	0
0	1	0	1	1
0	0	0	1	0

Bu matrisin her kolonu bir belgeyi tanımlar. Belgeler içerdikleri terimlerin göreceli ağırlıklarına göre tanımlanmalıdır. Biz ise sadece terim frekanslarını kullandığımızdan her kolonun birim uzunluğu 1 olacak şekilde (öklid uzunluğu) matrisi yeniden düzenleyerek terimlerin göreceli ağırlıklarını kullanmış oluruz :

A =

0,5774	0	0	0,4082	0
0,5774	0	1,00	0,4082	0,7071
0,5774	0	0	0,4082	0
0	0	0	0,4082	0
0	1,00	0	0,4082	0,7071
0	0	0	0,4082	0

A = SVD faktörlerine ayırma işlemi uygulanırsa

S =

-0.2670	0.2567	-0.5308	0.2847	-0.7071	-0.0000
-0.7479	0.3980	0.5249	-0.0816	0.0000	-0.0000
-0.2670	0.2567	-0.5308	0.2847	0.7071	0.0000
-0.1182	0.0127	-0.2774	-0.6395	-0.0000	-0.7071
-0.5197	-0.8423	-0.0839	0.1158	0.0000	0.0000
-0.1182	0.0127	-0.2774	-0.6395	-0.0000	0.7071

V =

16.950	0	0	0	0
0	11.158	0	0	0
0	0	0.8403	0	0
0	0	0	0.4195	0
0	0	0	0	0.0000
0	0	0	0	0

D =

-0.4367	0.4717	-0.3688	0.6714	0
-0.3066	-0.7549	-0.0998	0.2760	0.5000
-0.4413	0.3567	0.6247	-0.1944	0.5000
-0.4908	0.0346	-0.5710	-0.6572	0
-0.5288	-0.2815	0.3711	0.0577	-0.7071

$A = S\ V\ D^T$ matrisinin rank'i dörtdür (4) ve V matrisinin son iki satırı sıfır olduğundan A matrisinin kolon uzayının temeli, D matrisinin ilk dört kolonudur.

Rank'i 4 olan A matrisinin, daha düşük rank'li bir yakınsamasının A matrisine ne kadar benzediğini bulmak için, A matrisinin boyutu ile düşük rank'li yakınsamasının boyutunu karşılaştırabiliriz:

$$\frac{||A-A_3||_F}{||A||_F} = \frac{\sigma_4}{} = \frac{0.4195}{2.2361} = 0.1876$$

$$\frac{||A-A_2||_F}{||A||_F} = \frac{}{} = \frac{0{,}939162}{2.2361} = 0.4200$$

Bu hesaplamaların anlamı şudur ; A matrisini A_3 ile ifade edersek A matrisinde %19'luk bir değişim, A_2 ile ifade edersek %42'lik bir değişim yapmış oluruz. Bu hesaplamaların yararı, işlem performansını artırmak için daha düşük rank'li bir A matrisi karşılığı kullanırken, belge yığını içeriğinden ne kadar taviz verildiğinin ölçülmesidir.

Aynı yöntemle QR ve SVD yöntemlerinin karşılaştırılması da yapılabilir. QR ve SVD ayrıştırma yöntemleri uygulanarak elde edilen rank'i 3 olan A matrisi karşılıkları aşağıdadır.

Ă (QR) =

0.5774	0	0	0.4082	0
0.5774	0	1	0.4082	0.7071
0.5774	0	0	0.4082	0
0	0	0	0	0
0	1	0	0.4082	0.7071
0	0	0	0	0

A_3(SVD) =

0.49722	-0.032965	0.02322	0.48668	-0.0068903
0.60037	0.0094447	0.99335	0.38571	0.70907
0.49722	-0.032965	0.02322	0.48668	-0.0068903
0.18011	0.074047	-0.052159	0.23191	0.015477
-0.032614	0.98659	0.0094447	0.44012	0.7043
0.18011	0.074047	-0.052159	0.23191	0.015477

Ă(QR) kullanıldığında değişiliklik %42, A_3(SVD) kullanıldığında değişiklik %19 olur. İlginç olan, Ă(QR) matrisinin görsel olarak A'ya çok benzemesine rağmen daha farklı olmasıdır. A_3(SVD) matrisindeki negatif değerler A_k matrisinin A matrisi elemanlarının lineer birleşiminden meydana geldiğini gösterir. Şu unutulmamalıdır ki belge yığınının içeriği belge vektörlerinin elemanlarının değerlerinden değil, belge vektörlerinin geometrik ilişkisinden meydana gelir.

3.3.5.1 TDA - Kosinüs Açısının Hesaplanması

Veritabanı içeriğini A_k matrisi ile indeksleyip, bu içerik içinde bir q sorgu ile arama yapıldığında, arama neticesi olarak q sorgusuna yakın içerikteki belgeler döner. Bu yakınlığın belirlenmesi için, daha önce QR faktörlerine ayırma yöntemi incelenirken anlatılan, vektör uzayında belge ve sorgu vektörlerindeki kosinüs açısına bakılır. Arama işlemi sonucu olarak belirli eşik değeri aşan vektörler döner.

Bir dxd birim matrisin j. kolonuna e_j dersek, A_k matrisinin j. kolonunu $A_k e_j$ (matris çarpımı) şeklinde gösterebiliriz (Berry vd., 1999).

$$A_k = S_k V_k D_k^T$$

$$V_k = V_k^T$$

$$\cos\theta_j = \frac{(A_k e_j)^T q}{\| A_k e_j \|_2 \, \| q \|_2} = \frac{(S_k V_k D_k^T e_j)^T q}{\| S_k V_k D_k^T e_j \|_2 \, \| q \|_2} = \frac{e_j^T D_k V_k (S_k q)^T}{\| V_k D_k^T e_j \|_2 \, \| q \|_2}$$

$j = 1 \ldots . d$ için $s_j = V_k D_k^T e_j$ dersek, formülü daha basite indirgeyebiliriz

$$\cos\theta_j = \frac{s_j^T (S_k^T q)^T}{\| s_j \|_2 \, \| q \|_2} \qquad (3.2)$$

Bu formül sayesinde A_k matrisi hesaplanmadan kosinüs açısını hesaplama mümkün hale gelmektedir. Ayrıca $\| s_j \|_2$ hesaplamasını bir kereye mahsus yapmak yeterli olacağından işlem performansı kazanılmış olur.

3.4 Terim Ağırlığı Belirleme Yöntemleri

Bir belge yığınının t x d matrisi haline getirilip, belgeler ve terimler arasındaki ilişkileri ortaya çıkarmak için bu matrisin tekrar nasıl düzenleneceğini ve bu matris üzerinde belgeleri sorgulamak amacıyla nasıl bir hesaplama yapılacağını gördük. t x d matrisi bu süreçteki ana girdi olarak çok önemlidir ve bu matrisin elemanlarının değerleri, yapılan hesaplamaları doğrudan etkilemektedir. Bu aynı zamanda matrisin satırlarını oluşturan her terimin, bir belgenin anlamını ifade etmedeki gücünü de belirtir. Daha önce, matris elemanlarının (a_{ij}) sayısal değerinin, d_i belgesi içinde geçen k_j teriminin frekansı olarak kısaca kullanılabileceğini söylemiştik. Bu yaklaşım çok kaba bir yaklaşımdır ve büyük belge yığınları için uygulandığında anlamsız sonuçlar verir. Daha anlamlı sonuçlara ulaşmayı sağlayacak bir yöntem ortaya çıkarmak için,

terim ağırlık değerinin aşağıdaki unsurları içinde barındırması gereklidir (Chakrabarti S., 2003) (Manning vd., 1999).

T terim frekansı (tf_{ij}) : t_i teriminin d_j belgesi içinde geçme sayısı
D Belge frekansı (df_i): : t_i teriminin içinde bulunduran belge sayısı
Y Yığın frekansı (cf_i): : t_i teriminin belge yığını içinde kaç kere geçtiği.

Bu değerlere bağlı olarak farklı terim ağırlıkları hesaplamak mümkündür. Örneğin terim ağırlığı olarak

- $\sqrt{T}$
- $1 + \log(T)$

kullanılabilir.

Bu kullanımlarda T değerinin düşürülmesi daha anlamlı bir rakamsal sonuç oluşturur. Çünkü bir terimin belgede 3 kez geçmesi ile 33 geç geçmesi durumlarında, terimin belge içeriğine katkısı 11 kat oranında artmamaktadır. Ancak bir terimin bir belge için anlamı sadece T değerinin o belgedeki sayısı üzerinden hesaplanamaz. D değeri kullanılarak terimin belge için ne kadar geçerli olduğu bilgisine bir katkı yapılabilir. Örneğin belgeler içinde homojen olarak dağılan terimlerin bilgi verme derecesi düşüktür. Bir terimin her belgede aynı T değerine sahip olması, onun belgeleri birbirinden farklılaştırmak için kullanılamayacağını gösterir. Daha geçerli bir terim ağırlığı için terim frekansı ile belge frekansı tek bir değer oluşturmak için birleştirilebilir.

$$\text{Ağırlık}(i,j) = \begin{cases} (1+\log(tf_{ij}))\log\dfrac{N}{df_i} & tf_{ij} >= 1 \\ 0 & tf_{ij} = 0 \end{cases} \qquad (3.3)$$

N toplam belge sayısı

$$\log\frac{N}{df_i} = \log N - \log df_i$$

bu hesaplama sonucu; eğer bir terim belgelerin tamamında bulunuyorsa, $df_i = N$'dir.

$$\log \frac{N}{df_i} = \log N - \log N = 0$$

Yani bu terimin belge yığını için belge yığınındaki belgeleri işlemeye yarayacak bir katkısı yoktur. Benzer şekilde terim, belge yığını içinde az geçiyorsa, katkısı logaritmik olarak artacaktır.

Literatürde belge frekansının bu şekilde kullanımı, ters belge frekansı (tdf) (inverse documant frequency-idf) olarak bilinir. Ters belge frekansının farklı biçimlerde ölçümlenme alternatifleri de vardır. Bu nedenle yukarıdaki kullanım şekli "terim frekansı.ters belge frekansı" (tf.tdf) (term frequency.inverse document frequency-tf.idf) olarak bilinir.

Terim ağırlığı belirleme üzerinde çok fazla çalışma bulunmaktadır. Bu çalışmalar, bir belge yığınını, onu ifade etme gücü en yüksek belge vektörü uzayı ile oluşturmak için yapılmıştır. Her yöntem belirli belge yığınlarında test edilmiştir. Sonuçta ortaya atılan tüm terim ağırlığı belirleme yaklaşımları, birbirinden ayırmak için, her yöntemin kullandığı tekniği de açıklayabilecek bir isimlendirme yöntemi geliştirilmiştir. Önce kullanılan tekniklerin bileşenleri belirlenmiştir. Buna göre her yöntem temelde üç teknik kullanır.

1. tf değerinin hesaplanması
2. df değerinin hesaplanması
3. normalizasyon yapılıp yapılmadığı

tf değerinin hesaplanmasında kullanılan teknikler ise şunlardır :

n	:	natural (doğal)	$tf_{t,d}$
l	:	logaritma	$1+\log(tf_{t,d})$
a	:	augmented (birleştirilmiş)	$0,5 \frac{0,5\, tf_{t,d}}{\max_t (tf_{t,d})}$

df değerinin hesaplanmasında kullanılan teknikler de:

n	:	natural (doğal)	df_i
t	:		$\log \frac{N}{df_t}$

normalizasyon hesaplama teknikleri ise:

n	:	not (yok)	normalleştirme yapılmamıştır
c	:	cosinus (kosinüs)	$\frac{1}{\sqrt{w_1^2 + w_2^2 + \ldots\ldots + w_n^2}}$

Bu isimlendirme yaklaşımı çerçevesinde tf.tdf yöntemi ltn olarak adlandırılır. Yani

l	:	tf hesaplarken logaritmasını kullanır
t	:	df hesaplarken logaritmasını kullanır
n	:	normalleştirme yapmaz.

Bunun dışında olasılık temelli ağırlık belirleme yöntemleri de vardır. Bu yöntemler belirli bir örneklem içinde kelimelerin örneklemi oluşturan belgelerin içindeki oranlarından yararlanarak, belge yığını içindeki oranını bulma prensibine dayanır. Poisson dağılımı bu amaç için kullanılan yöntemlerden biridir.

Terimlerin ağırlıklarının tesbiti için en yaygın kullanılan yöntem tfidf fonksiyonunu kullanmaktır (Salton ve Buckley, 1988).

$$Tfidf\,(t_k, d_j) = \#(t_k, d_j) \log \frac{|Tr|}{\#_{Tr}(t_k)} \qquad (3.4)$$

Bu formülde $\#(t_k, d_j)$ ifadesi t_k teriminin d_j belgesinde bulunma sayısını, $\#_{Tr}(t_k)$ ifadesi ise t_k teriminin içinde geçtiği belge sayısını gösterir. Tr eğitim seti kümesini gösterir. Bu formülün anlamı

- bir terim bir belgenin içinde ne kadar çok varsa, o belgeyi o kadar iyi temsil eder.

- bir terim ne kadar çok belgede geçiyorsa, o terimin belgeleri yırt etme gücü o kadar azdır.

Bu formül terimlerin belge içindeki pozisyonlarını ve sözdizimsel özelliklerini dikkate almaz. Terimlerin ağırlıklarının [0,1] aralığında yer almasını ve belgelerin eşit uzunlukta vektörler ile ifadesi sağlamak için tfidf fonksiyonu kosinüs normalizasyonuna tabi tutulur :

$$W_{kj} = \frac{\text{tfidf}(t_k, d_j)}{\sqrt{\sum_{S=1}^{|T|} (\text{tfidf}(t_k, d_j))^2}} \qquad (3.5)$$

Bu fonksiyonun özelleştirilmiş halleri değişik uygulamalarda kullanılmakla beraber, bu fonksiyonun kullanılamayacağı durumlar da vardır. Örneğin Tr'nin daha önceden tesbit edilemediği zamanlarda tfidf (t_k, d_j) değeri bulunamaz.

Tez çalışmasında, terim ağırlığı hesaplama yöntemi olarak tf.tdf yöntemi kullanılmıştır.

4. İKİNCİ NESİL BELGE MADENCİLİĞİ

Vektör uzay modeli üzerine bina edilen belge madenciliği yöntemleri ile başarılı analiz çalışmaları yapılabilmektedir. Geleneksel yöntemler olarak tanımlanabilecek bu yöntemler büyük belge yığınları içinde, o zamana kadar insan emeği ile yapılması mümkün olmayan işlemler yapılmasını sağlamıştır.

Geleneksel bilgiye erişim sistemleri, belgeleri, içindeki kelimelerle; belgelerin arasındaki ilişkileri de bu kelimelerin tüm belge yığını içindeki istatistiki kullanım örüntüleri ile ifade etmektedirler Kelimeleri baz alan bu sistemler otomatik indeksleme yapabildiklerinden, belgelerin madenciliği açısından büyük kolaylıklar sağlamaktadır. Ancak bu sistemler belirli kısıtlara da sahip olmuşlardır.

Örneğin geleneksel bilgiye ulaşma yöntemleri içinde en yaygın kullanılan ve oldukça başarılı sonuçlar üreten Google arama motoru üzerinde "müşteri ilişkileri yönetimi" gibi bir sorgulama yapıldığında, bu konuyla ilişkili belgeler sorunsuz bir şekilde sorgu sonucu olarak gelmektedir. Ancak, arama yapılan belge sayısı artıkça, sorgu neticesinde gelen belge sayısı binlerle ifade edilir hale gelebilmektedir. Bu durum, asıl sorgulanan belgeler sorgu sonucu olarak gelse bile, gelen binlerce belge arasından çıkarılması gibi ikinci bir sorgulama ihtiyacı ortaya çıkarmaktadır. Zaten çoğunlukla sorgulama ihtiyaçları da verilen "müşteri ilişkileri yönetimi" örneğinde olduğu gibi ana konularla ilgili olmamaktadır. Daha detay analiz gerektiren sorgulamalara ihtiyaç duyulmaktadır. Örneğin "müşteri ilişkileri yönetiminin işletme verimliliğine etkisi" gibi bir konunun sorgulanması ihtiyacı doğmaktadır.

Geleneksel yöntemlerin kısıtlarını sistematik olarak incelersek ;

1- Yazarların belgeleri oluştururken kullandığı kelimelerle, bu belgeler içinde arama yapmak isteyen kişilerin sorgu ifadelerinde kullandıkları kelimelerin birbirinden farklı olabilmektedir. Bunun nedeni kelimelerin çokanlamlı ve/veya eşanlamlı olabilmesidir (Foltz vd., 1998). Benzer şekilde farklı kültürlerde veya farklı disiplinlerde aynı kelimelerin farklı kavramları ifade edecek şekilde kullanılması durumu da bu kısıtın başka bir boyutudur. Bu nedenle, kelimeler her zaman aynı anlamı taşımayabilir. Örneğin yüz kelimesi yüzmek fiili, 100 sayısı veya insan yüzü olarak kullanılmış olabilir. Bu

çokanlamlı kelime örneğinin tam tersi olarak aynı varlığı, durumu yada eylemi ifade eden eşanlamlı kelimeler de söz konusudur. Doktor, hekim kelimeleri eşanlamlıdır.

2- Kelimeler tek başlarına kullanıldıklarında belirli bir anlam taşırken başka kelimelerle yan yana kullanıldıklarında bambaşka anlamlar taşırlar. Örneğin "at" kelimesi binek hayvanı olarak bir anlam taşırken "at arabası" şeklinde kullanıldığında bir canlıyı değil, cansız bir varlığı ifade eder.

3- Mevcut sistemler, kelimelerin, belgelerin anlamına kattığı değeri hesaplamak için sadece kelimelerin belgelerde ne sıklıkla kullanıldıklarına bakar, oysa kelimelerin belgelerin anlamını ifade etmede diğerlerinden daha etkili olduğu farklı durumlar da vardır. Örneğin Türkçe cümlelerde fiile yakın kelimeler anlam açısından daha önemlidir. "Ahmet camı bugün kırdı" cümlesinde vurgu zamana odaklanırken aynı kelimelerden oluşan "Ahmet bugün camı kırdı" cümlesi kırılan cama odaklanmaktadır. Eğer kelimelerin sırası "Camı bugün Ahmet kırdı" şeklinde olsaydı, vurgu fiili gerçekleştiren kişi olan Ahmet'e odaklanmış olurdu.

Bu kısıtlardan ilki için – eşanlamlı/çokanlamlı kelimeler – bilgiye erişim yöntemlerinden biri olan Gizli Anlambilimsel Dizinleme (GAD) bir çözüm sunmaktadır. GAD vektör uzay modeline dayandığı ve vektör uzayında boyut indirgeme sağladığı için büyük veri yığınları üzerinde işlem yapmayı da kolaylaştırmaktadır. İkinci ve üçüncü kısıtları gidermek için GAD yöntemi, bu tez çalışmasının da ana konusunu oluşturan bir yöntemle geliştirilmiştir. Bu bölümde önce GAD yöntemi, sonra da bu yöntemin geliştirilmesi için önerilen çözüm anlatılacaktır (Holzman vd., 1994).

4.1 Gizli Anlambilimsel Dizinleme

Bir belge yığınını, vektör uzay modeli ile t x d boyutlu vektör uzayına taşıdıktan sonra, üzerinde nasıl işlem yapacağımızı 3. bölümde gördük. Vektör uzayının boyutları çok büyük olduğunda ortaya çıkan hesaplama yapma problemini de boyutları küçültmek için TBA yöntemleri uygulayarak çözmüştük. Üçüncü bölümde boyut küçültme yöntemi olarak anlatılan tekil değer ayrıştırması (TDA), belge madenciliği çalışmalarına GAD yöntemi sayesinde taşınmıştır. Bu nedenle TDA, GAD belge madenciliği yönteminin esas aldığı faktör analiz tekniğidir (Deerwester vd., 1990).

GAD yöntemi vektör uzay modeline dayandığından belgeleri içerdikleri kelimelere göre değerlendirir. Ancak TDA yöntemini kullandığından belge yığınını bir bütün olarak ele alır. Yani belgeler arası ilişkileri kullanarak, bir A belgesi ile ilişkili B belgesi eğer bir C belgesi ile ilişkili ise, A ve C belgeleri arasındaki ilişkiyi de ortaya çıkarma yeteneğine sahiptir. Bunu yapabilmek için bir belgede yer alan kelimelerin başka hangi belgelerde yer aldığını ve diğer belgelerde bu kelimelerle birlikte kullanılan ortak kelimeleri göz önünde bulundurur. Geleneksel vektör uzay modeline dayalı yöntemlerden farklı olarak belge içinde gizli kalmış ilişkileri de ortaya çıkardığından "gizli anlambilimsel dizinleme" ismini almıştır (Kontostathis, 2002).

GAD yönteminin bir belge yığınını indekslemek için kullandığı teknik, insanların bir belge yığınını inceleyip, belgeler ve aralarındaki ilişki ile ilgili ortaya koydukları neticelere yakın sonuçlar üretmektedir. Bu yönüyle GAD için insan öğrenmesini modelleyen bir yöntem tanımlaması da yapmak mümkündür. GAD yöntemi insan öğrenmesini modelleme yeteneğini belge yığını içindeki belgeleri oluşturan kelimelerin anlamını bilmeden otomatik yaptığı için ayrı bir öneme sahiptir (Landauer vd., 1998) (Landauer vd., 1997) (Weimer-Hasting, 1999).

GAD yöntemi ile indekslenmiş bir belge yığınında yapılan arama işleminde, içeriği ifade etmek için seçilmiş tüm kelimelerin benzerlik değerlerine bakılır ve aranan anahtar kelime veya kelimelerin en yüksek benzerlik değerine sahip olduğu belgeler arama sonucu olarak döner. Herhangi iki belge, ortak kelimelere sahip olmasalar da anlamsal olarak aynı olabileceğinden ve bu ilişki TDA sayesinde yakalandığından, GAD aranan anahtar kelimelerin birebir indekslenmiş belgelerde olup olmadığına bakmaz. Bu sayede daha gerçekçi bir arama işlemi yapılmış olur (Landauer vd., 1998) (Weimer-Hasting, 1999). Örneğin matematik konusunda yazılmış belgelerden oluşan bir belge seti GAD yöntemi ile indekslenmiş olsun. Bu belge setini oluşturan belgelerde matris, lineer cebir, doğrusal cebir kelimeleri yeteri kadar belgede yer alıyorsa, bu kelimelerin anlamsal olarak birbirlerine yakın olduğu sonucuna ulaşılır. Matris anahtar kelimesini içeren belgeleri bulmak için başlatılan bir arama işlemi sonucunda matris kelimesini içermeyen ama lineer cebir ve/veya doğrusal cebir kelimelerini içeren belgeler de döner. Görüldüğü gibi arama mekanizması matematik konusunu bilmediği halde, yeterli sayıda belgeyi inceleyerek matematik konusunda kullanılabilen kelimeleri öğrenerek arama işlemini buna göre yapmaktadır (Quesada vd., 2003a) (Quesada vd., 2003b) Kontostathis (2002).

Yukarıdaki örnekten de görüldüğü gibi GAD yöntemi belgelerin anlamlarını bilmeden anlamlarını aramayı sağlayacak bir yaklaşım sunmaktadır. Bu son derece yararlı bir özelliktir. Bu sayede konu ve dil bağımsız belgeler indekslenebilir ve geleneksel arama yöntemlerinin ötesinde bir fayda sağlayacak şekilde belgelere ulaşım sağlanır.

GAD yöntemi, vektör uzay modeli üzerine bina edildiğinden çalışma prensipleri bellidir. Ancak TDA yöntemi ile vektör uzayını oluşturan vektörler faktörlerine ayrıştırıldıktan sonra oluşturulan vektör uzayı üzerinde işlem yapıldığından, GAD yöntemi ile elde edilen başarılı neticelerin tam olarak nasıl oluştuğu bilinmemektedir. Neticelerin istatistiki dayanakları olduğundan, insanların öğrenme sürecini modellediği iddia edilmektedir (Landauer vd., 1998) (Landauer vd., 1997) (Weimer-Hasting, 1999).

4.2 GAD Çalışma Metodolojisi

GAD yöntemi ile bir belge yığını üzerinde belge madenciliği çalışması yapmak için, öncelikle her belgedeki, dolayısıyla tüm belge yığınındaki kelimeler belirlenir. Bu kelimelerin neler olduğu, yapılacak belge madenciliği çalışmalarının sağlığı açısından son derece önemlidir. Belgeler içinde geçen tüm kelimeler anlam içermez. Örneğin bağlaçlar (ve, veya gibi) bu gruptaki kelimelerdendir. Anlama katkısı olmayan kelimelerin vektör uzayında yer alması, hem vektör uzayının boyutlarını artırıp, işlem maliyetini artırır hem de kirlilik yaratarak belge madenciliği çalışmalarının kalitesini düşürür. Bu nedenle öncelikle belgeleri ifade etme gücü olan kelimeler tespit edilir. Bu işi yapacak algoritma aşağıdaki gibidir.

1. belge seti içindeki tüm kelimelerin bir listesi yapılır
2. bağlaçlar (ve, veya, çünkü, zira), işaret zamirleri (bu, şu,o) ayıklanır
3. çok kullanılan kelimeler ayıklanır
4. zamirler (ben, sen, o) ayıklanır
5. çok kullanılan sıfatlar (geç, büyük, yüksek) ayıklanır
6. tüm belgelerde kullanılan ortak kelimeler ayıklanır
7. sadece bir belgede kullanılan kelimeler ayıklanır

İçeriği ifade eden kelimeler tesbit edildikten sonra, ikinci adım bu kelimelerin eklerinden temizlenerek anlamını kaybetmeyeceği en yalın hallerine dönüştürülmesidir. Dilbilgisi kuralları

gereği, kelimeler cümle içinde kullanılırken biçim değiştirebilirler. Biçim değiştirme iki şekilde olabilir. İngilizce dilinde olduğu gibi çok sınırlı kurallarla ve eklerle yapıldığı durumda, kelimelerin yalın hallerine ulaşmak rutin bir algoritma ile gerçekleştirilebilir. Nitekim İngilizce için yaygın kullanılan Porter Stemmer adlı bir algoritma ile bu işlem kolayca yapılabilmektedir [19].

Tablo 4 -1 Porterstemmer ile İngilizce kelimelerin en yalın hallerini elde etme

Institution : institut Available : avail Engineering : engin	Association : associ Accessable : access Going : go

Tablo 4-1'de görüldüğü gibi Porter Stemmer algoritması İngilizce dili içinde kelimelere gelebilecek standart ekleri ve bu eklerle kelimelerin hangi şekilleri alabileceğinin kurallarını içinde bulundurur ve bir kelimeden ekleri çıkardıktan sonra kelimenin nasıl bir hal alacağını otomatik belirler. Biçim değiştirmenin ikinci şekli Türkçe gibi eklemeli dillerde, kelime köklerine getirilen eklerle yeni kelimeler oluşturulması durumunda ortaya çıkar. Bu durumda kelimelerin anlamını kaybetmeyeceği en yalın haline taşınması için kısa ve doğru bir yol yoktur. Doğal dil çalışmaları kapsamında bir kelime kullanıldığı cümledeki anlamını kaybetmeyeceği en yalın haline dönüştürülebilir ancak bu yönde bir çalışma henüz tamamlanmamıştır.

Örneğin "gözlükçüler " kelimesinin dilbilgisi kurallarına göre ekleri şöyledir

Kök: göz tip:IS

Ekler: ISIM_YALIN_BOS + ISIM_BULUNMA_LIK + ISIM_ILGI_CI + ISIM_COGUL_LER

Oysa "gözlükçüler yeni tasarımlarını paylaşmak için bir araya geldiler" cümlesinde gözlükçüler kelimesinin anlamını yitirmediği en yalın hal "gözlükçü" dür çünkü isin olarak onlar ifade edilmektedir. Geleneksel anlamda ekler ayrıştırıldığında bu kelime "göz" haline gelmekte ve bu haliyle cümledeki anlam kaybolmaktadır. Türkçe gibi eklemeli dillerde bir sözlük yardımı ile kelimelerin eklerine ayrıştırılmaları mümkündür.

Üçüncü adımda hangi belgede hangi kelimelerin olduğunu, hangi kelimelerin hangi belgelerde olduğunu görmeye ve sorgulamaya yarayan bir kelime-belge matrisi oluşturulur. Kelimeler

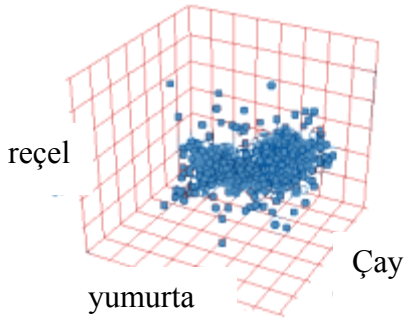
matrisin satırlarını, belgeler de matrisin kolonlarını oluşturur. Eğer bir kelime bir belgede yer alıyorsa matrisin ilgili hücresine x işareti yerleştirilir. Eğer kelime belgede yer almıyorsa ilgili hücre boş tutulur.

Elde ettiğimiz matrisin hücreleri boş ve x değerleri yerine 0 ve 1 değerleri ile doldurulursa matematiksel olarak bilgi içeren bir matris elde edilir. Bu aşamadan sonra GAD yönteminin anahtar adımı olan matrisi yeniden düzenleme (decomposition) adımına gelinir. GAD yönteminde matris, tekil değer ayrıştırması (TDA) tekniği ile yeniden düzenlenir (Yu vd., 2002) (Weimer-Hasting, 1999) (Landauer vd., 1998).

4.3 GAD Yönteminin ve TDA Tekniğinin Benzetimi

TDA tekniğinin, çok boyutlu bir A matrisini daha küçük boyutlu üç matrise dönüştüren, A matrisini oluşturan faktörleri ve bu faktörlerin önem sırasını ortaya koyan bir yöntem olarak inceledik. TDA işlemi sonucunda elde edilen üç matris kullanılarak hem kolon uzayında hem de satır uzayındaki ilişkiler hesaplayabileceğimizi belirtmiştik. Burada tekniğin nasıl işlediğini gösteren bir örnek verilecektir (Berry vd.,1994) (Wolfe vd., 1998).

Bir otel işletmesi sahibi olduğumuzu düşünelim. İnsanların kahvaltıda ne sipariş ettiklerini incelemek istediğimizi farz edelim. (Yu vd., 2002) Bunun için yoğun bir gün içinde insanların kahvaltı siparişlerinde ne kadar çay, reçel ve yumurta istediklerini kaydetmiş olalım. Elde ettiğimiz verileri görsel olarak incelemek için üç boyutlu uzayda Şekil 4-1'deki gibi bir uzay oluşturup her kahvaltı siparişini bu uzayda işaretleyelim.



Şekil 4-1 Üç Boyutlu Uzayda Müşteri Tercihleri

Bu gösterimde grafiğin merkezinden her noktaya çizilecek çizgiler birer vektör olacaktır ve bu vektörlerin herbiri ayrı bir kahvaltı siparişini gösterecektir. Matemetiksel ifadesi ile bu vektörler "reçel,yumurta,çay" uzayında bir vektör olacaktır. Bir vektör tek başına bir kahvaltı siparişinde kaç reçel, kaç yumurta ve kaç çay istendiğini gösterecektir. Tüm vektörler ele alındığında ise müşterilerin kahvaltı tercihlerinin ne yönde olduğu bulunabilecektir. Vektör uzayı olarak da ifade edilebilecek bu gösterim gerçekte üç boyuttan daha fazla boyutta olacaktır. Örneğin eğer kahvaltıda peynir ve zeytin de varsa beş boyutlu uzayda kahvaltı siparişleri gösterilir ve incelenir.

Konuyu belge madenciliğine bir örnek olabilecek benzetimle anlatmak istersek; altı adet kitap adı beş adet kelimeden oluşmuş olsun (Quesada vd, 2003b). Her kitabın isminde geçen kelimelerin karşılığı 1 olarak işaretlenmiş, geçmeyen kalimeler karşılığına 0 konmuştur

	Kitap1	Kitap2	Kitap3	Kitap4	Kitap5	Kitap6
Kozmonot	1	0	1	0	0	0
Astronot	0	1	0	0	0	0
Ay	1	1	0	0	0	0
Araba	1	0	0	1	1	0
kamyon	0	0	0	1	0	1

kelime x kitap matrisi (A), TDA ile faktörlerine ayrıştırılırsa

A =

1	0	1	0	0	0
0	1	0	0	0	0
1	1	0	0	0	0
1	0	0	1	1	0
0	0	0	1	0	1

$$A = SVD^T$$

S =

-0,44	-0,3	0,57	0,58	0,25
-0,13	-0,33	-0,59	0,0	0,73
-0,48	-0,51	-0,37	0,0	-0,61
-0,70	0,35	0,15	-0,58	0,16
-0,26	0,65	-0,41	0,58	-0,09

V =

2,16	0	0	0	0
0	1,59	0	0	0
0	0	1,28	0	0
0	0	0	1,0	0
0	0	0	0	0,39

D^T =

-0,75	-0,28	-0,2	-0,45	-0,33	-0,12
-0,29	-0,53	-0,19	0,63	0,22	0,41
0,28	-0,75	0,45	-0,2	0,12	-0,33
0,0	0,0	0,58	0,0	-0,58	0,58
0,53	0,29	0,63	0,19	0,41	0,22

V matrisi tekil değerlerin en önemli ikisi 2,16 ve 1,59 alınır, diğerleri sıfırlanıp A matrisi SVD^T çarpımı olarak yeniden hesaplanırsa kitap isimleri beş boyuttan iki boyuta taşınmış olur. Sonuçta ortaya çıkan düşük boyutlu A matrisi aşağıdadır

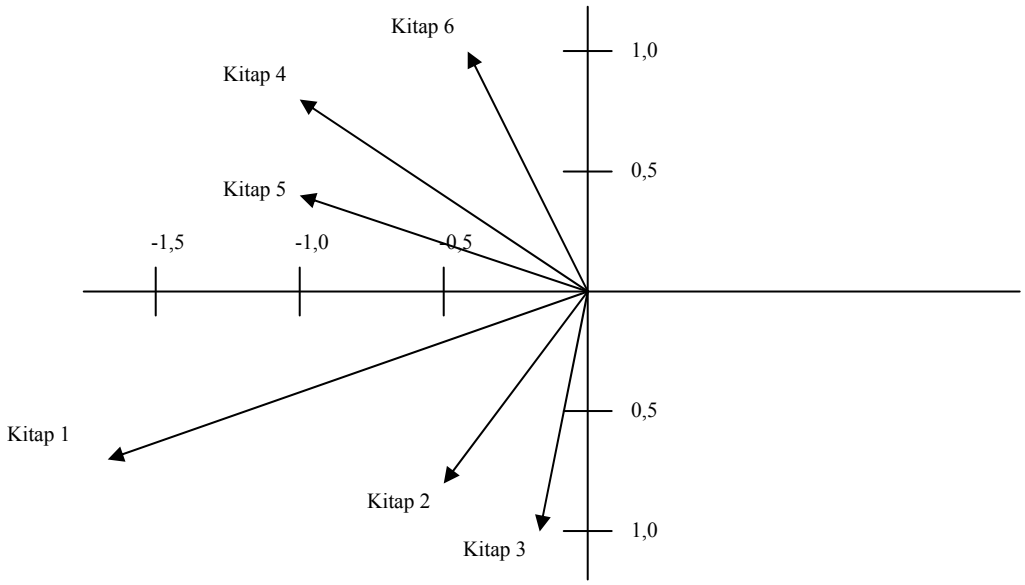
$\check{A}=A_3(SVD)=$

-0,75	-0,28	-0,2	-0,45	-0,33	-0,12
-0,29	-0,53	-0,19	0,63	0,22	0,41

Vektör uzay karşılı ise

	Kitap1	Kitap2	Kitap3	Kitap4	Kitap5	Kitap6
Kavram 1	-1,62	-0,60	-0,441	-0,97	-0,70	-0,26
Kavram 2	-0,46	-0,84	-0,30	1,00	0,35	0,65

Yeni iki boyutlu uzayda kitapların gösterimi Şekil 4-2'de görülmektedir.



Şekil 4-2 İki Boyutlu Uzayda Kitaplar

Görüldüğü gibi, hesaplama yapmadan da kitap1 ve kitap3 arasındaki ilişki ve kitap4 ve kitap5 arasındaki ilişki fark edilmektedir. Ve bu ilişki orijinal uzaydan çok daha düşük bir boyutlu uzayda yakalanabilmektedir.

Belge yığınlarını içerdikleri içeriği ifade eden kelimelerle bir uzayda gösterecek olsak yüzlerce boyutluk bir uzay söz konusu olabilir. Bu uzayda birbirine içerik olarak yakın belgelerin vektörleri arasındaki uzaklığın da düşük olacağı ortadadır (Yu vd., 2002).

TDA, söz konucu çok boyutlu uzayı, işlenebilecek makul daha düşük boyutlu bir uzaya indirger. Bu işlem sonucu anlamsal olarak yakın kelimeler birbirine yakınlaşacaktır. Bu sayede geleneksel

arama yöntemlerinin anahtar kelimeleri birebir belgelerin içinde bulma gereksinimi gibi bir ihtiyaç ortadan kalkmış olacaktır (Weimer-Hasting, 1999) (Landauer vd., 1998).

Boyut sayısını indirmek aslında bilgi kaybetmektir. Bu ilk bakışta kötü bir şey gibi durur. Ancak TDA tekniği ile benzer kelimeler birbirine daha yakınlaştırılırken, ayrı kelimeler daha da uzaklaştırılarak kirlilik ortadan kaldırılır ve daha iyi neticeler elde etmeyi sağlayacak indeksleme yapılmış olur (Weimer-Hasting, 1999) (Landauer vd., 1998).

4.4 GAD Yöntemi ile Belge Madenciliği

Tez çalışmasının konusu Türkçe belgelerin madenciliğidir. Bu çerçevede GAD yöntemi kullanılarak iki belge madenciliği tekniği ile Türkçe belgeler üzerinde işlemler yapılacaktır. Bu teknikler belge sorgulama ve belge kümelemedir. GAD yöntemi ile bu teknikler kullanılarak elde sonuçlar daha sonra, GAD yöntemini geliştirmek için önerilen yöntem ile tekrarlanacak ve aradaki başarım farkları ortaya konacaktır. Öncelikle belge sorgulama ve belge kümeleme teknikleri anlatılacaktır. Sorgulama ile ilgili önceki bölümlerde detaylı bilgi verildiğinden, özellikle kümeleme ile ilgili detaylı bilgi verilecektir (Dunlavy, 2002).

GAD yönteminin belge madenciliği için seçilmesinin önemli bir sebebi de, yöntemin kelimeleri, insanların kavramları işaretlemekte kullandığı gibi kullanması, diğer bir deyişle insan öğrenmesine benzer bir öğrenme şekline sahip olmasıdır. Örneğin GAD algoritması ile 5 milyonda fazla kelime içeren, lise dengi bir öğrencinin bilmesi gereken konuları içeren İngilizce belgelerden oluşan bir yığın GAD yöntemi ile indekslendikten sonra, oluşturulan terim x belge uzayına İngilizce yeterlilik testi uygulanmıştır. Test verilen bir kelimenin anlamına en yakın anlamda kullanılan kelimenin, seçeneklerde verilen dört cümle içinden bulunmasını amaçlamaktadır. Sonuç olarak sistem ortalamada %64'lük bir derece ile testi başarmıştır ve bu dünya genelinde İngilizcesini sertifikalamak isteyen kişilerin ortalamasına eşittir (Fayyad vd ., 1996a).

4.4.1 Belge Sorgulama

Kelime x belge matrisine TDA uygulandıktan sonra elde edilen daraltılmış vektör uzayında, belirli bir konuda bilgi içeren belgeler sorgulanabilir. Sorgu işlemi öncelikle sorgu ifadesinin de bir belge gibi ele alınıp, vektör uzayını oluşturan diğer belgelerin geçtiği işlemlerden geçmesi ve vektör uzayına yerleştirilmesi ile başlar. Daha sonra sorgu vektörü ile diğer tüm belge vektörleri arasındaki mesafe ölçülür ve sorgu vektörüne en yakın belge vektörleri tespit edilerek, sorgu sonucu elde edilir (Weimer-Hasting, 1999) (Landauer vd., 1998).

Vektörler arası mesafe ölçümü için, vektörlerin arasındaki kosinüs açısı hesaplanır. Birbirlerine yakın vektörler 1'e yakın değerler alırken birbirleriyle ilgisiz vektörler 0'a yakın değerler alırlar.

4.4.2 Belge Kümeleme

Kümeleme, nesneleri, benzerliklerine göre gruplayan bir tekniktir. Belge madenciliği bakış açısından, artan belge sayısındaki büyüklük nedeniyle, insanlar tarafından yapılması neredeyse imkansız bir işlem olan, benzer belgelerin gruplanması giderek önem kazanmaktadır.

Kümeleme işleminin kalbi, belgeler arasındaki benzerlik yada farklılıkların derecesini analiz etmek için kullanılabilecek yöntemlerdir. Bu amaçla geliştirilmiş değişik yöntemler bulunmaktadır. Bu yöntemler aşağıdaki gibi sınıflandırılabilir (Chakrabarti S., 2003) (Wang vd., 2006):

- Hiyerarşik kümeleme (hierarchical clustering)
    - Toplayıcı kümeleme (agglomerative clustering)
    - Bölücü kümeleme (divisive clustering)
- Bölümleme kümeleme (Partitioning clustering)
    - K-ortalama kümeleme (k-means clustering)
    - Olasılık temelli kümeleme (probabilistic clustering)

Hiyerarşik kümelemede her uç nokta bağlı olduğu bir üst kümenin alt kümesidir. Yatay kümelemede ise belirli sayıda küme vardır ve kümeler arası ilişki belirli değildir (Szymkowiak vd., 2001).

Kümeleme yapılırken dikkat edilmesi gereken diğer bir durum ise, bir nesnenin sadece ve sadece bir kümeye mi yoksa belirli üyelik dereceleri ile birden fazla kümeye mi ait olduğudur. Hiyerarşik kümelemede katı kümeleme yani bir nesnenin sadece bir kümeye aitliği kullanılır. Farklı kümeleme yaklaşımlarının kullanım alanları Tablo 4-2'de anlatılmıştır (Manning C. D., Schütze H., 1999).

Tez çalışmamız kapsamında kullanılacak kümeleme yöntemi hiyerarşik kümeleme sınıfında yer almaktadır.

Tablo 4-2 Kümeleme teknikleri kullanım alanları

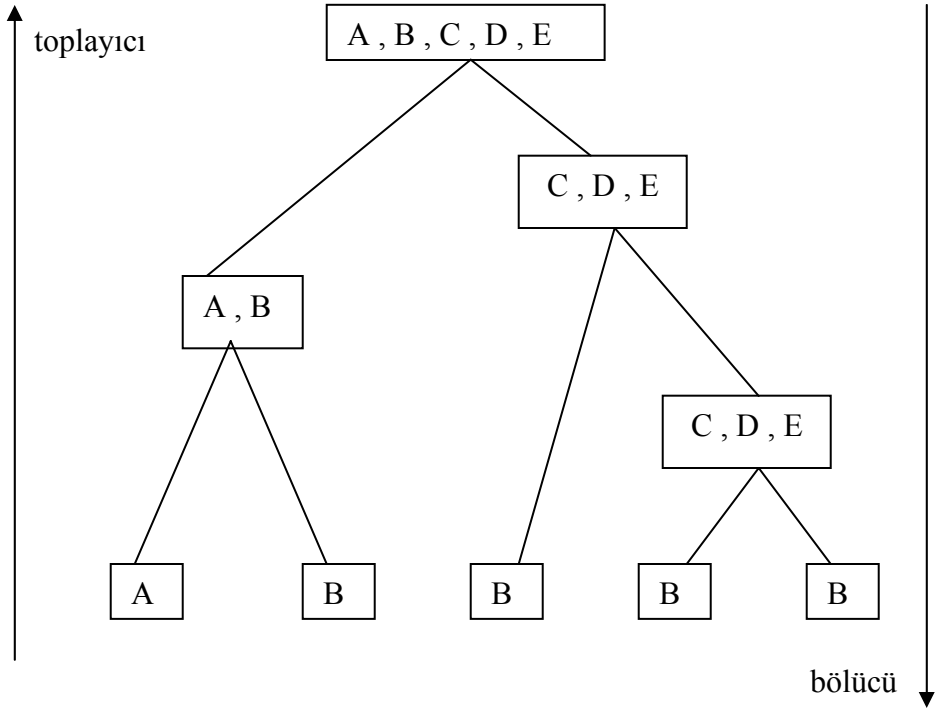
Hiyerarşik kümeleme	**Yatay kümeleme**
Detaylı veri analizi sağlar	Veri kümeleri çok büyük ve işleme verimliliği gerekiyorsa tercih edilir
Yatay kümelemeye göre daha fazla bilgi sağlar	K-means kavram olarak en basit ve yeterli yaklaşımdır, özelikle yeni bir veri kümesi için
Belirli bir en iyi algoritması yoktur.	K-means basit Öklid uzayı ortamı varsayımı ile çalıştığından renkler gibi sadece metinsel değerlere sahip veriler üzerinde kullanılamaz
Yatay kümelemeye göre daha fazla işlem gücü gerektirir.	Bu gibi durumlarda EM algoritması kullanılır

Hiyerarşik kümelemede, belgelerin kümelere ayrıştırılması tek bir işlem adımı içinde gerçekleştirilmez. Belgeler önce tek bir kümeye ve daha sonra da belge sayısı kadar kümeye birer adet dağıtılacak şekilde her bir adımda yeniden kümelenerek, kümelere dağıtılabilirler. Hiyerarşik kümeleme kendi içinde toplayıcı ve bölücü kümeleme olarak ikiye ayrılır.

Toplayıcı kümelemede bütün belgeler önce her biri bir kümede yer alacak şekilde belge sayısı kadar kümeye ayrılır ve daha sonra belgeler arasındaki benzerliklere bakılarak, birbirine en yakın belgeler her adımda tek kümede birleştirilir. Bölücü kümeleme yaklaşımında tüm belgeler tek bir

kümeye yerleştirilir ve kümeleme işleminin her bir adımında birbirinden farklı belgeler ayrı kümelere dağıtılır. Toplayıcı kümeleme yöntemi, bölücü kümelemeye göre daha fazla kullanılmaktadır.

Hiyerarşik kümeleme işleminin sonucu dendrogram ismi verilen iki boyutlu bir çizimle görselleştirilebilir. Dendrogam üzerinde, kümeleme işleminin her adımındaki kümeler ve kümelerin elemanları izlenebilir [24]. Örnek bir dendrogram Şekil 4-3'de gösterilmiştir.



Şekil 4-3 Hiyerarşik Kümeleme Yöntemleri

4.4.2.1 Toplayıcı Yöntemler

Toplayıcı hiyerarşik kümeleme işlemi, her adımda belge yığınının bir parçasını gruplar. İlk önce belge sayısı kadar grup vardır ve her grupta birer adet belge bulunur. Her işlem adımı ile bir önceki parçalardan birbirine en yakın benzerliğe sahip olan iki parça birleştirilir taki en sonunda belge yığını tek bir grupta toplanıncaya kadar.

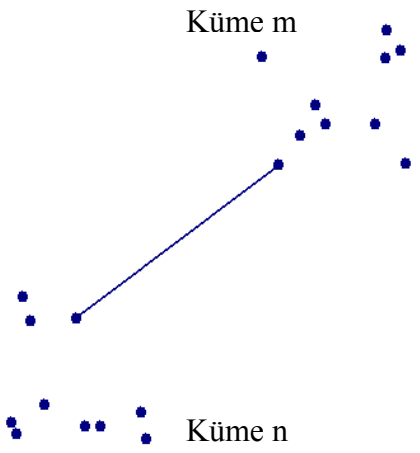
Kümeler arasındaki yakınlık yada benzerlik ilişkisini bulmak için kullanılan değişik teknikler nedeniyle birden fazla toplayıcı hiyerarşik kümeleme yöntemi bulunmaktadır[24].

Tekil bağ ile kümeleme;

Bu yöntem en basit toplayıcı hiyerarşi kümeleme yöntemidir ve en yakın komşu tekniği olarak da bilinir. Bu teknik, iki küme arası yakınlığı, kümelere ait belgeler içindeki birbirine en yakın belgeler arasındaki uzaklık ile ölçer [22]. Bunu aşağıdaki formül ile göstermek mümkündür [24]:

$$\text{Uzaklık (küme\_m, küme\_n)} = \text{minimum } \{ \text{ uzaklık (i,j)} : i \in \text{küme\_m ve } j \in \text{küme\_n } \}$$

Bu formül gereği, küme_m ve küme_n'e ait her i, j belge ikilisi için uzaklık ölçümü yapılır ve en sonunda birbirine en yakın ikili arasındaki uzaklık, küme m v n arasındaki uzaklık olarak kabul edilir. Toplayıcı hiyerarşik kümeleme işleminin her adımında birbirine en yakın iki küme birleştirilir. Kümeler asındaki tekil bağ tekniği ile hesaplanan uzaklığın nasıl belirlendiği görsel olarak Şekil 4-4'de verilmiştir



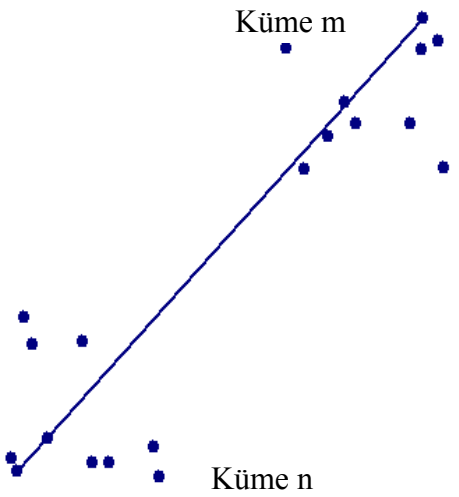
Şekil 4-4 Tekil Bağ ile Hiyerarşik Toplayıcı Kümeleme

Tam bağ ile kümeleme;

Bu teknik tekil bağ kümelemenin tam tersidir. İki küme arasındaki yakınlık ölçüsü olarak kümelere ait belgelerden birbirine en uzak olanların arasındaki uzaklık kullanılır. Bu yöntem en uzak komşular olarak da bilinir [22]. Bu hesaplama aşağıdaki formül ile yapılır :

Uzaklık (küme_m, küme_n) = maksimum { uzaklık (i,j) : i ∈ küme_m ve j ∈ küme_n }

Bu formül gereği, küme_m ve küme_n'e ait her i, j belge ikilisi için uzaklık ölçümü yapılır ve en sonunda birbirine en uzak ikili arasındaki uzaklık, küme m v n arasındaki uzaklık olarak kabul edilir. Toplayıcı hiyerarşik kümeleme işleminin her adımında birbirine en yakın iki küme birleştirilir. Kümeler asındaki tam bağ tekniği ile hesaplanan uzaklığın nasıl belirlendiği görsel olarak Şekil 4-5'da verilmiştir [24].



Şekil 4-5 Tam Bağ ile Hiyerarşik Toplayıcı Kümeleme

4.4.2.2 Kümeleme Sonuçlarının Değerlendirilmesi

Kümeleme sonuçlarının başarımının değerlendirilmesi için entropi ve f-measure ölçümleri kullanılır (Özgür vd., 2004) (Wang vd., 2006).

Her iki ölçümün de yapılabilmesi için kümelenecek belgelerin belirli sınıflara önceden atanmış olması gereklidir. Kümeleme işleminin başarısı elde edilen kümelerin, önceden belirlenmiş bu sınıflarla karşılaştırılması ile ölçülür.

Entropi

Entropi, elde edilen bir kümenin ne kadar homojen olduğunun göstergesidir. Düşük entropi değeri kümenin daha homojen olduğunu ifade eder ve bu beklenir. Toplam entropinin belirlenmesi için, önce her kümenin entropisi ve daha sonra tüm kümelerin toplam entropisi hesaplanır. Bir kümenin entropi hesabı (Özgür vd., 2004).:

$$E_j = - \sum P(i,j) * \log P(i,j) \qquad (4.1)$$

Burada P(i,j) *i* sınıfına ait belgelerin *j* kümesinde bulunma olasılığıdır. Toplam entropi hesabı:

$$E = \sum n_j / n * E_j \qquad (4.2)$$

Burada n_j, j. kümenin boyutunu, n ise belge seti içindeki toplam belge sayısını ifade eder.

F-measure

F-maesure ölçüsü kümelemenin kalitesini gösterir. F-measure hesaplanırken bilgiye erişim sahasında kullanılan iki standart ölçümden yararlanılır.

Duyarlılık (Precision) bir kümedeki belgelerin kaçının doğru kümede olduğunu gösterir.

Çağırma (Recall) bir kümede olması gereken belgelerden kaçının o kümede olduğunu gösterir.

Toplam f-measure hesaplanırken önce her kümenin f-measure'ı hesaplanır (Özgür vd., 2004). :

$$F(i,j) = \frac{2 * \text{Çağırma } (i,j) * \text{Duyarlılık } (i,j)}{\text{Çağırma } (i,j) + \text{Duyarlılık } (i,j)} \qquad (4.3)$$

Daha sonra toplam F-measure hesabını yapmak için

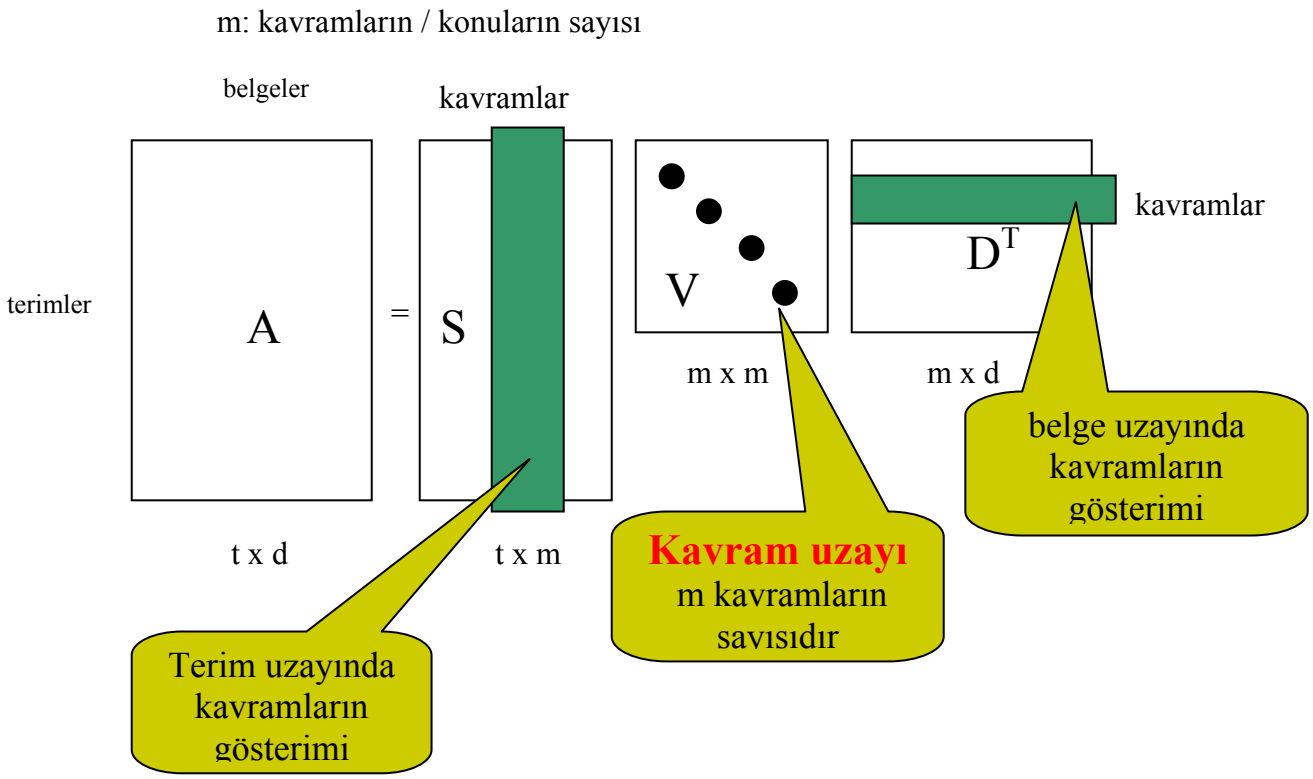
$$F = \sum n_i / n * \max F(i,j)$$

Burada n_i, i. kümenin boyutunu, n ise belge seti içinde ki toplam belge sayısını ifade eder. [20]

4.4.2.3 GAD ile Belge Kümeleme

TDA yöntemini anlatırken, bu yöntem ile elde edilen daha düşük boyutlu matris karşılığı olan üç matris sayesinde terimlerin ve belgelerin kendi aralarındaki ilişkileri de gösteren bir netice elde

edildiğinden bahsedilmişti. Bu işlemi tekrar hatırlarsak, A matrisi TDA ile üç matrise parçalanıyordu. Bu durum Şekil 4-7'de gösterilmiştir.



Şekil 4-6 Tekil Değer Ayrıştırmanın Anlamı

Görüldüğü her bir terim ve belge m boyutlu uzayda birer vektördür. İki terimin ya da iki belgenin birbiriyle ilişkisini bulmak için m boyutlu uzaydaki vektörleri arasındaki kosinüs açısına bakabileceğimizi vektör uzay modelini incelerken belirtmiştik. A ve B vektörlerinin iç çarpımı formülünü kullanarak terimlerin terimlerle, belgelerin belgelerle ilişkisini bulmak için kullanabileceğimiz formülü çıkabiliriz (Berry vd., 1994) (Tang vd., 2005).

$$A \cdot B = |A| \cdot |B| \cos \propto$$

$\propto$; A ve B vektörleri arasındaki açıdır

Terimlerin birbiri ile karşılaştırılmaları demek, her terim vektörünün, diğer terim vektörleri ile iç çarpımlarının hesaplanması demektir. Bunun matris cebirindeki karşılığı

$$A \cdot A^T$$

Benzer şekilde belgelerin birbiri ile karşılaştırılmaları demek, her belge vektörünün, diğer belge vektörleri ile iç çarpımlarının hesaplanması demektir. Bunun matris cebirindeki karşılığı

$$A^T \cdot A$$

A matrisi yerine TDA karşılığı üç matrisi (SVD^T) kullanarak (Deerwester vd., 1990) :

$$\begin{aligned} A^TA &= (SVD^T)^T SVD^T \\ &= D^{TT}V^TS^TSVD^T \\ &= DV^T(S^TS)VD^T \end{aligned}$$

$S^TS = I$ çünkü S ortonormal kolonlara sahip

$$\begin{aligned} &= DV^TVD^T \end{aligned}$$

$V = V'$ çünkü V diyagonal

$$\begin{aligned} &= DVV^TD^T \\ &= DV(DV)^T \end{aligned} \qquad (4.4)$$

elde edilen formül D matrisine dayalıdır ve D matrisi kavram uzayında belgelerin gösterimini içerir. Yani elde ettiğimiz formül mantiken de geçerlidir.

Belgelerin kümelenmesi işleminde her bir belgenin diğer bir belge ile arasındaki uzaklık/yakınlık kullanılacağını belirtmiştik. $DV(DV)^T$ çarpımı ile elde edilen matrisin elemanları (i,j) tam olarak bu değerleri içermektedir; yani i. belge ile j. belge arasındaki yakınlık değerini vermektedir (Zukas vd., 2006) (Bradford 2006).

5. N-GRAM DESTEKLİ GAD

Vektör uzay modelinin kısıtlarını belirtirken üç kısıta sahip olduğunu belirtmiştik. Birinci kısıt olan eşanlamı ve çokanlamlı kelimeleri ayırt edememe sorununun GAD yöntemi ile çözümlendiğini ortaya koymuştuk. Diğer iki kısıt olan, kelimelerin bir arada kullanılmaları nedeniyle, tekil olarak kullanımlarından daha farklı anlam taşımaları ve kelimelerin cümle içindeki yerinin anlam için önemli olduğu noktalarının GAD ile çözümlenemediğini ve buna yönelik tez çalışması kapsamında GAD yöntemini geliştiren bir çözüm önerdiğimizi söylemiştik.

GAD yöntemi kelimeleri tek başlarına, anlamlarını hesaba katmadan istatistiki yöntemlerle ele alır ve kullanım örüntülerini bulur. Bu yaklaşım doğru sonuçlar vermekle birlikte Türkçe gibi, kelimelerin yan yana kullanım şekillerine göre anlamlarının değiştiği, kuvvetlendiği veya zayıfladığı diller için iyileştirilmesiyle, daha uygun sonuçlar üretecek hale getirilebilir.

Türkçe'de kelimelerin cümle içinde fiile yakınlığına bağlı olarak, anlama kattığı değer farklılaşmaktadır. Örneğin "Ahmet bugün camı kırdı" cümlesinde vurgu cam üzerinedir. "Ahmet camı bugün kırdı" dersek konu zaman odaklı hale gelmiş olur. Aynı cümleyi "Bugün camı Ahmet kırdı" şeklinde kurduğumuzda ise anlam Ahmet üzerine yoğunlaşır. Görüldüğü gibi dört kelimeden oluşan basit bir cümlenin kelimelerinin yerlerini değiştirdiğimizde, ana konu farklılaşabilmektedir.

Diğer bir husus da kelime gruplarının, kendilerini oluşturan kelimelerden daha farklı bir anlam içermesinde ortaya çıkar. Mesela "kurt adam öldü" cümlesindeki "kurt adam" ile "kurt adam öldürdü" cümlesindeki kurt ve adam kelimeleri tamamen farklı anlamlarda kullanılmıştır.

Daha çarpıcı bir örnek aşağıdadır.

- verimli bir şirket nasıl olur ?
- bir şirket nasıl verimli olur ?
- nasıl bir şirket verimli olur ?

Bu üç cümlenin sorgu olduğunu düşünürsek, kelimeleri tek başlarına ele alan bir yöntem olan GAD, hepsi için aynı belgeleri döndürecektir. Oysa aranılan belgeler farklıdır.

Özellikle büyük belge yığınları içinden arama yapıldığı durumlar söz konusu olduğu için, bu sorgu ifadelerine uygun farklı belgeler bulunabilir ve bu belgelere ulaşmak için bir yöntem geliştirilmelidir (Hornick vd., 2004).

5.1 N-gram Yöntemi

N-gram yöntemi, bir metnin hangi dilde yazıldığını bilgisayar tarafından belirleyebilmek amacıyla kullanılan istatistiki bir yöntemdir. Bunu kelimeleri oluşturan harflerin yan yana gelme örüntülerine bakarak yapar (Ekmekçioglu vd., 1996). Örneğin bilgisayar kelimesinin n-gramları:

2-gram

b bi il lg gi is sa ay ya ar r

3-gram

b bil ilg lgi gis isa say aya yar r

her dilin kelimelerinin 2-gram, 3-gram gibi n-gram örüntüleri farklıdır. Bu yaklaşımla bir metnin hangi dille yazıldığı belirlenebilir [9] .

5.2 GAD ile N-gramın Birleştirilmesi

GAD yönteminin kelimeleri tek tek ele almasına alternatif olarak metin içinde kelimelerin yan yana kullanımları ile oluşan ikili, üçlü ve n'li kullanım şekillerini ele alması sağlanabilir. Bu durumda Türkçe gibi, kelime anlamlarının tek başlarına değil, birbirlerine göre cümle içindeki yerlerine göre değiştiği diller için daha uygun bir indeksleme ve sorgulama elde edilmiş olur.

Kelime içindeki harflere uygulanan n-gram yöntemini, cümle içindeki kelimelere uygulayarak, oluşacak ikili, üçlü, n'li kelime gruplarını GAD yöntemi ile işlemek mümkündür.

Tez çalışması kapsamında GAD yönteminin kullandığı terim uzayı 2-gram ve 3-gram terimlerle genişletilmiş ve belge madenciliği çalışmaları bu uzay üzerinde gerçekleştirilmiştir. GAD ve n-gram destekli GAD yöntemleri arasındaki performans karşılaştırmaları için hem İngilizce hem Türkçe belge setleri üzerinde test yapılmıştır. Bu testlerde gerek sorgulama gerekse de kümeleme çalışmaları yapılarak somut sonuçlar üretilmiştir. Bunlar 6. bölümde detaylı olarak anlatılacaktır.

5.3 Konuyla İlgili Çalışmalar

Türkçe için olmamakla birlikte terimler ve 2-gram terimlerle oluşturulan bir uzayda belgelerin gösterimini yapan ve bu uzayda belge madenciliği çalışmalarını gerçekleştiren girişimler bulunmaktadır. Tez çalışmamız ile karşılaştırıldığında bu yöntemler

1- Türkçe için daha önce yapılmamıştır
2- GAD yöntemini kullanmamıştır
3- 3-gram terimleri, terim uzayına dahil etmemiştir.

Bu çalışmaları iki sınıfta toplamak mümkündür.

1- terimlerle birlikte 2-gram terimleri birlikte kullananlar
2- terimleri kullanmayıp sadece 2-gram terimleri kullananlar

Sezgisel ve istatistiksel olarak ikinci yöntemin birinciye göre daha zayıf bir performans gösterdiği sonucuna varılmıştır. Sadece terimleri kullanan yöntemlerin başarısı ispatlanmıştır ve 2-gram terimler kullanarak bu performans artırılabilir.

Bu konuyla ilgili çalışmalar aşağıda paylaşılmıştır:

- Diederich (Diederich vd., 2003) çalışmasında belgelerde, belge yazarlarının etkisini ölçümlemek için, belgeleri sınıflamış ve bu sınıfların yazarlarla ilişkisine bakmıştır. Çalışmasında Destek Vektör Makinesi (DVM) - SVM (Support Vector Machine) - tekniğini kullanmıştır. Bu tekniği belgeleri oluşturan terimler üzerinde ve belgelerden elde ettiği fonksiyonel kelimeler ve bu kelimelerin 2-gram biçimlerinden oluşan terim uzayında çalıştırmıştır. Burada bahsedilen fonksiyonel kelimelerle belgelerden elde edilen özel kelimeler anlatılmaktadır. Bu özel kelimelerin içine atılabilir kelimeler (stopwords), isimler, fiiller ve sıfatlar dahil değildir. Sonuç olarak çalışmaları göstermiştir ki sadece terimlerin kullanımı ile daha başarılı sonuçlar üretilmektedir.
- Zhang and Lee (Zhang vd., 2003) çalışmalarında tekil olarak terimleri kullanarak metin sınıflaması yapmışlar ve bu işlemi gerçekleştirmek için beş farklı sınıflama yöntemini test etmişlerdir. Bunlar kNN (k en yakın komşular)-k-nearest neigbour, Naive Bayes, karar ağacı - Decision Tree, SNoW ve DVM teknikleridir. Daha sonra bu teknikleri n-gram

terimlere de uygulamışlardır. Kullandıkları veri seti ise TREC10 QA data setidir. Çalışmalarının neticesi olarak iki yaklaşım arasında fala bir değişiklik görmediklerini belirtmişlerdir

- Koster ve Seutter (Koster vd., 2003) ise çalışmalarında tekil terimler ve 2-gram terimleri kullanarak EPO1A veri seti üzerinde Rocchio and Winnow sınıflama teknikleri kullanarak test gerçekleştirmişlerdir. Elde ettikleri neticelerle 2-gram terimleri kullanmanın her iki teknik için de başarılı olmadığı ancak hem terimler hem de 2-gram terimleri kullanmanın ise her iki teknik için de başarı oranını büyük ölçüde artırdığını ortaya koymuşlardır.
- Tan (Tan vd., 2002) Naive Bayes sınıflama tekniğini kullanarak tekil terimleri 2-gram terimlerle birleştirerek belgeleri sınıflandırmıştır. Elde edilen sonuçlar başarısız olmuştur. Bunun sebebi olarak Naive Bayes tekniğinin sınıflama için zaten güçlü bir teknik olmaması gösterilebilir. Ayrıca 2-gram terimleri oluştururken kullandıkları özel bir yöntem de bu başarısızlığa katkıda bulunmuş olabilir.
- Caropreso (Caropreso M.F., Matwin S., and Sebastiani F., 2001) n-gram yaklaşımını Roochio sınıflama tekniği ile Reuters veri kümesinde test etmiştir. N-gram terimleri oluşturmak için izledikleri yöntem ise, bir cümledeki atılabilir kelimeleri attıktan sonra n adet ardışık kelimenin alfabetik olarak sıralanmasıdır. Test sonuçları ile 2-gram terimlerin genel olarak sınıflandırma sonucunu olumlu olarak etkilediği ispatlamışlardır.

6. GAD ve N-GRAM DESTEKLİ GAD İLE MADENCİLİK

İngilizce başta olmak üzere, pek çok Avrupa ülkesi dili ve Çince ve Japonca dilleri için, bu dillerle yazılan belgelerin veri madenciliği teknikleri ile incelenmesini sağlayacak yöntemler ve araçlar geliştirilmiş ve geliştirilmektedir. Bu çalışmaların arkasındaki amaç, insan ile doğal dil yolu ile iletişim kuracak bilgisayarlar yapmaktır. Bu sayede doğal dil ile yazılmış belgeleri okuyup anlayacak ve kendi kendine öğrenecek sistemler yapmak mümkün olabilecek, bu sistemler büyük belge yığınlarını öğrenerek, doğru bilgiye hızla ulaşmayı sağlayacaktır.

Doğal dil söz konusu olduğunda farklı diller arasındaki yapısal farklılıklar göz önüne alınmak zorundadır. Örneğin Çince'de her cümle için neredeyse bir kelime kullanılırken, İngilizce'de benzer bir yapının varlığı söz konusudur. İngilizce'de bir fiilin değişik zaman çekimlerindeki karşılıkları, kurala oturtulamayacak farklı kelimelerle ifade edilir. Bu kısıt nedeniyle, İngilizce konuşan kişiler, tüm kelimeleri ve anlam karşılıklarını bilmek zorundadır. Bu durum insanlar için kısıt iken, bilgisayarlar için kolay bir işleme yöntemi sunar. Tüm kelimeler büyük bir veri tabanında, anlam karşılıkları ve diğer kelimelerle olan ilişkileri içerilecek şekilde tutulur. Princeton Üniversitesi tarafından sürekli güncellenerek geliştirilmesine devam edilen wordnet projesi [15] ile İngilizce dili için bir sözlük oluşturmuş ve ingilizce metinlerin incelenmesinde bu veritabanı herkes tarafından kullanılan bir standart sağlayarak bilgisayar ortamında İngilizce belgelerin incelenmesini kolaylaştırmıştır.

Türkçe ise tamamen farklı bir yapıya sahiptir ve kelimeler, kök halleri üzerine gelen kurallara bağlı eklerle türetilir. Bu sayede insanlar tarafından anlamları kolayca çıkartılabilir. Ancak bilgisayar ortamında bu belgelerin işlenebilmesi için kuralların bilgisayar ortamına aktarılması gerekir. Pek çok dil için uzun zamandan beri yapılan çalışmalar Türkçe için yeni yeni yapılmaktadır. Bu nedenle, Türkçe dili için, metin veri madenciliği yapanlar tarafından standart olarak kullanılabilecek bir çalışma henüz yoktur. Bazı projeler gelecek için böyle bir standardı oluşturmaya çalışmaktadır.

Tez konumuz Türkçe belgelerin madenciliğini yapmayı amaçlamaktadır. Yukarıda bahsedilen eksiklikler nedeniyle, sadece yeni bir yöntem geliştirilmemiş, aynı zamanda Türkçe için bir belge madenciliği altyapısı da hazırlanmıştır.

6.1 Türkçe'nin Zenginliği ve Zorluğu

"Victor Hugo şiirlerini 40.000 kelime ile yazmıştır. Türkçe'yi en zengin kullananlardan Yaşar Kemal'in romanları 3.500 kelimeyi geçmez" görüşü çok yaygındır. Bu görüş haklıdır zira Türkçe'nin Fransızca'ya oranla daha az sözcük içerdiği doğrudur. İngilizce'ye, Almanca'ya, İspanyolca'ya oranla da daha az sözcük içeriyor olması gerekir. Ne var ki bu Türkçe'nin daha yetersiz bir dil olduğu anlamına gelmez çünkü Türkçe az sözcük ile çok şey anlatabilen bir dildir. Başka bir dilden Türkçe'ye çeviri yapan herkes sözlüğü açtığında, aralarında minik anlam farkları olan bir çok sözcüğün Türkçe karşılığında çoğu zaman aynı kelimeyi okur. Bu, ilk bakışta bir eksiklik gibi görünebilir, oysa öyle değildir. Çünkü yukarıda adı geçen diller kelimelerin statik olan anlamlarını öğrenmeye, Türkçe ise bu anlamları bulup çıkarmaya, yani dinamik anlamlandırmaya dayanmaktadır.

Türkçe'de anlamları, sözlükteki tanımlar değil, kelimelerin cümle içindeki konumları belirler. Buna dayanarak Türkçe'nin, referans olmak üzere sadece gerektiği kadarı sözlüklere alınmış, sonsuz sayıda kelime içerdiği bile öne sürülebilir. İngilizce-Türkçe sözlükte "sick", "ill" ve "patient" ın karşısında hep "hasta" yazar. Bu bağlamda İngilizce'nin üç kat daha fazla sözcük içerdiği söylenirse bu doğrudur. Ancak, aradaki farkların Türkçe'de vurgulanamadığı söylenmeye kalkılırsa bu yanlış olur: "beyin hastası olmak","böbrek hastası olmak", "Internet hastası olmak", "filanca şarkının hastası olmak" arasındaki farkı Türkçe konuşan herkes bir çırpıda anlar. Bunun nasıl olabildiğini görmek zor değildir. Bir kalem alıp, alt alta:

3 + 5 =

12 + 5 =

38 + 5 =

yazmak, sonra da bunları toplamak yeterlidir. Hepsinde aynı "+ 5" yazdığı halde sonuçlar farklı çıkıyorsa, Türkçe'de de hepsinde aynı "hastası olmak" ifadesi geçtiği halde sonuçlar farklı olacaktır. Türkçe'nin az araç ile çok iş yapmasının sırrı matematikte yatar. 0 dan 9 a kadar 10 tane rakam, artı, eksi, çarpı, bölü dört işlem işareti ve bir ondalık ayracı

virgül, yani topu, topu 15 simge ile sonsuz sayıda işlem yapılabilir. Türkçe de benzer özellikler gösterir. Türkçe matematiğe dayalı olmaktan da öte, neredeyse matematiğin kılık değiştirmiş halidir.

Türkçe'deki herhangi bir fiilin çekiminin ve kelimelerin nasıl çoğul yapılacağının öğrenilmiş olması, henüz varlığı bile bilinmeyen, 5 yıl sonra Türkçe'ye girecek fiillerin nasıl çekileceğinin ve 300 yıl önce unutulmuş kelimelerin çoğullarının ne olduğunun biliniyor olması demektir. Bu tıpkı birinci dereceden 2 bilinmeyenli bir denklemin nasıl çözüleceği öğrenildiğinde, sadece x = 6, y = 23 olan denklemlerin değil, aynı dereceden bütün denklemlerin nasıl çözüleceğinin öğrenilmiş olması gibidir. Oysa sözgelimi İngilizce'de "go", " went" olurken "do", "did" olur. Çoğul ekleri için de durum aynıdır: "foot", "feet" olurken "boot", "beet" değil "boots" olur. Bunun tutarlı bir iç mantığı yoktur, tek çare böyle olduklarının bellenmesidir. Türkçe'de ise, statik kelimeleri ezberlemek yerine dinamik kuralları öğrenmek gerekir. Türkçe'de neredeyse istisna bile yoktur. Olanlar da, ses uyumu gereği alma olması gereken meyve isminin elma biçimine dönmesi gibi birkaç minör istisnadır.

Bu anlatılanlar perspektifinde Türkçe belgelerin madenciliği çalışmalarında, farklı diller için geliştirilmiş teknikleri birebir kullanmak zordur. Hatta bu teknikler işe yarasalar bile, Türkçe'nin kendisine özgü çalışmalarla daha başarılı sonuçlar üretmek mümkün olacaktır. Ancak bu konuda daha gidilecek yol vardır. Türkçe'ye has yöntemler geliştirmek için yapılan çalışmalar bir sonraki bölümde özetlenmiştir.

6.2 Türkçe Doğal Dil Projeleri

Türkiye´deki çalışmaların tarihçesi, Hacettepe Üniversitesi´nde, 1970´li yıllarda bir doktora tezi ile başlamıştır. Türkçe'nin morfolojisi ile ilgili bu çalışma sayesinde ilk temel atılmıştır. Benzer çalışmalar paralel olarak Bilkent Üniversitesi´nde bilgisayarlı dilbilimden çok iyi anlayan Kemal Oflazer, Orta Doğu Teknik Üniversitesi´nde Cem Bozşahin ve Boğaziçi Üniversitesinde Türkçe'nin morfolojisi ile ilgilenen Selahattin Kuru, Tunga Güngör ve Levent Akın öncülüğünde başlamıştır. Bu ekip 1990´ların hemen başında Türkçe dil işleme konusunu canlandırmıştır ve günümüzde farklı üniversitelerimizde de çalışmalar olmakla beraber esasen Sabancı Üniversitesi´nde çalışmalarına devam eden Kemal Oflazer, ODTÜ´de Cem Bozşahin´le çalışma

arkadaşları ve Boğaziçi Üniversitesinde Tunga Güngör ile Cem Say tarafından devem ettirilmektedir. Son dönemde ITÜ'de bu yönde çalışmalar yapılmaktadır.

Tüm bu çalışmalar tez çalışması için bir altyapı sunmamıştır. Bazıları akademik seviyede, çoğunluğu ise belge madenciliği için değil, doğal dil işlemenin alt konularına has kalmıştır.

6.2.1 Türkçe Sözcük Veritabanı Projesi

Bu proje Sabancı Üniversitesi ile University of California at Berkeley 'in ortak çalışmasıdır ve TÜBİTAK ile ABD-NSF tarafından desteklenmektedir. Proje, University of California - Berkeley'deki TELL projesi ile birleştirilecektir. Proje ile Türkçe için kapsamlı bir biçimbirimsel sözlük oluşturulmuştur. Örneğin gözlükçüler kelimesinin biçimbirimsel çözümlemesini aşağıdaki gibi vermektedir [9].

gözlükçü (gözlükçü+Noun+ A3pl+ Pnon+ Nom)
gözlük (gözlük+Noun+ A3sg+ Pnon+ Nom^DB+ Noun+ Agt+ A3pl+ Pnon+ Nom)
göz (göz+Noun+ A3sg+ Pnon+ Nom^DB+ Adj+ FitFor^DB+ Noun+ Agt+ A3pl+ Pnon+ Nom)

Bu çözümlemeyi bir insan olarak yorumlamak kolayken, bilgisayar ortamında bir kelimenin bir belge içindeki anlamını ortaya çıkarmak içim otomatik olarak kullanmak zordur.

6.2.2 Türkçe Kavramsal Sözlük

BalkaNet Projesi dahilindeki Türkçe Kavramsal Sözlük, Avrupa Birliği IST Programı tarafından IST-2000-29388 numaralı fon çerçevesinde desteklenmektedir [9].

Türkçe Kavramsal Sözlük Projesi, Türkçe,Yunanca, Bulgarca, Çekçe, Romence ve Sırpça için her bir dilin kendi kavramsal sözlüklerinin (-wordnet- kelimelerin biçimlerinden çok anlamlarına göre elektronik sözlükler) birleştirilmesi ile oluşturulacak çok dilli bir sözcüksel veritabanının tasarımını ve geliştirilmesini hedefleyen BalkaNet Projesi'nin, bir parçasıdır. Bu sözlük yardımı ile gözlükçü kelimesi incelendiğinde Tablo 6-1'deki gibi bir sonuç elde edilir:

Tablo 6-1 Gözlükçü Kelimesinin Türkçe Kavramsal Sözlükteki Karşılığı

Gözlükçü		
Tanım	**İlişki**	**Açıklama**
İsim--Noun (ILI: ENG20-09707270-n)	ilişki	gözlükçü /1 Türkçe Tanım--Turkish Gloss: Gözlük satan veya onaran kimse Temel Kavram Tipi--Base Type: 4
İsim--Noun (ILI: ENG20-09914482-n)	Üst kavram	vasıflı işçi /1 nitelikli işçi /1 kalifiye işçi /1 Türkçe Tanım--Turkish Gloss: İstenilen nitelikleri taşıyan, iyi yetişmiş, usta işçi. Temel Kavram Tipi--Base Type: 1
	Alt kavram	Aradığınız kelime bir yapraktır, alt kavramı yoktur
	Bölümün bütünü	Aradığınız kelime Bölümün Bütünü ilişkisine sahip değildir.
	Bütünün bölümü	Aradığınız kelime Bütünün Bölümü ilişkisine sahip değildir.
	Üyenin bütünü	Aradığınız kelime Üyenin Bütünü ilişkisine sahip değildir
	Bütünün üyesi	Aradığınız kelime Bütünün Üyesi ilişkisine sahip değildir
	Parçanın bütünü	Aradığınız kelime Parçanın Bütünü ilişkisine sahip değildir
	Bütünün parçası	Aradığınız kelime Bütünün Parçası ilişkisine sahip değildir
	Alt olay	Aradığınız kelime Alt Olay ilişkisine sahip değildir
	olayıdır	Aradığınız kelime Olayıdır ilişkisine sahip değildir
	Sebep olur	Aradığınız kelime Sebep Olur ilişkisine sahip değildir
	sonucudur	Aradığınız kelime Sonucudur ilişkisine sahip değildir
	durumundadır	Aradığınız kelime Durumundadır ilişkisine sahip değildir
	durumudur	Aradığınız kelime Durumudur ilişkisine sahip değildir
	Yaklaşık zıt anlamlı	Aradığınız kelime Yaklaşık Zıt anlamlı ilişkisine sahip değildir
	Yaklaşık eşanlamlı	Aradığınız kelime Yaklaşık Eşanlamlı ilişkisine sahip değildir

Bu sonucun da belge madenciliği çalışmalarında kullanımı zordur.

6.2.3 Sözlüksüz Köke ulaşma

"İTÜ Doğal Dil İşleme Grubu", İ.T.Ü. Bilgisayar Mühendisliği Bölümü çatısı altında 2002 yılından itibaren çalışmalarına başlamıştır. İlk olarak Doğal Dil İşleme alanında yapılan çalışmaların Türkçe'ye özgü olarak yeniden gözden geçirilmesi hedeflenmiş ve bu konuda çalışmalar yapılmıştır. Bu kapsamda "Sözlüksüz Köke Ulaşma Yöntemi", sentetik kelime üretimi gibi çalışmalar yapılmış ve yapılmaktadır [10]. Bu çalışmalar da incelenmiş ancak tez çalışmasına yardımcı olabilecek bir fonksiyonu bulunmamıştır.

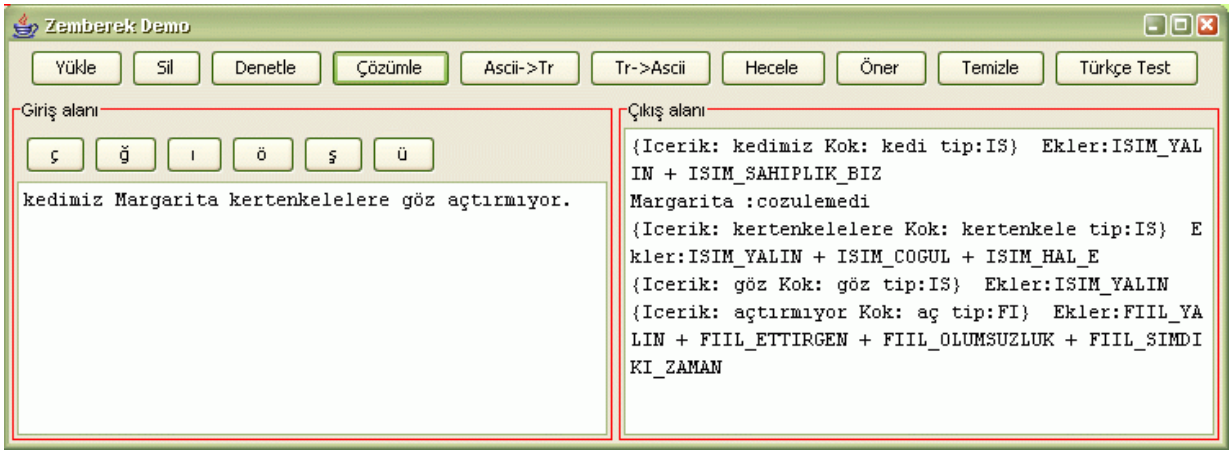
6.2.4 Zemberek

Zemberek projesi, Türkçe diline ilişkin çeşitli bilgi işlem problemlerinin çözümlenmesi için açık kodlu, platform bağımsız bir kütüphane oluşturulması amacı ile başlatılmıştır. Proje şu anda Türkçe kelime denetleme, kelime çözümleme, kelime önerme, oluşturma, Türkçe karakter kullanılmadan yazılan yazıların dönüştürülmesi, heceleme gibi işlemleri gerçekleştirmektedir [11].

Zemberek kütüphanesi ile gerçekleştirilebilecek işlemler :

Kelime denetleme:
Kelime denetleme imla denetimi (Spell Checker) işlemini gerçekleştirmek için kullanılır. Ara yüzde giriş alanındaki yazının yapısını bozmadan çıkış alanına taşır, sadece yanlış yazıldığına karar verilen kelimelerin başına "#" işareti koyulur [11].

Kelime çözümleme:
Bu özellik ile bir kelime kök ve eklerine ayrılır. Eğer birden fazla çözüm varsa tüm olası çözümler bulunur. Kök, kök tipi ve ekler şeklinde yazılır. Eğer kelime yalın ise "YALIN: tipinden bir ek varmış gibi gösterilir, Şekil 6-1 [11].



Şekil 6-1 Zemberek Kelime Çözümleme

ASCII->Tr :

Özellikle yabancı ülkede yaşayanların ya da bazı yazılımların türkçe uyumsuzluğu nedeniyle çoğumuzun düştüğü bir durum, Türkçe yazıları türkçe karakter kullanmadan yazma zorunluluğudur.. ASCII->Türkçe (Deasciifier) işlemi ile girilen kelimelerin muhtemel türkçe karşılıkları elde edilmeye çalışılır [11].

Tr->Ascii :

Yukarıdaki işlemin tam tersini yapar. Türkçe karakterleri ingilizce harflere benzer şekle dönüştürür [11].

Heceleme:

Basit heceleme algoritmasi. Girişteki tüm kelimeleri ayrı ayrı heceler. Hecelenemeyen kelimelerin başına # işareti koyulur [11].

Hatalı kelime öneri sistemi:

Türkçe kelime öneri sistemi. Hatalı yazılan kelimeler için doğru olabilecek kelimeleri önerir [11].

Türkçe metin tespiti:

Bu özellik ile kısa ya da uzun bir yazının türkçe olup olmadığı belirler. Türkçe karakter kullanılmadan yazılan yazılar için de sonuç elde edilir [11].

Kelime üretimi

Zemberek ile istenilen kök ve ek listesi verildiğinde kelime üretimi mümkündür. Kelime üretimi sırasında tüm ek ve kök özel durumlarına uyulur [11].

Kelime ayrıştırma

Bu özellik ile bir kelime görsel olarak kök ve eklerine ayrıştırılır. Örneğin giriş "kedilerim" ise çıkış bir dizi şeklinde {kedi-ler-im} şeklinde elde edilir [11].

6.3 Zemberek ile Kelimelerin En Yalın Halini Bulma

Zemberek çalışması içindeki kelime çözümleme fonksiyonu, tez çalışmamızda, bir kelimenin anlamını kaybetmeyeceği en yalın hali bulmak konusunda yardımcı olabilecek yetenektedir. Örneğin gözlükçü keliemesinin çözümlemesi sonucu elde edilen netice aşağıdadır [11].

{Icerik: gözlükçü Kok: gözlük tip:IS} Ekler:ISIM_YALIN_BOS + ISIM_ILGI_CI
{Icerik: gözlükçü Kok: göz tip:IS} Ekler:ISIM_YALIN_BOS + ISIM_BULUNMA_LIK + ISIM_ILGI_CI

Oluşan çözümlemenin ilk hali, kelimenin anlamını koruduğu haldir- gözlük. Bir başka örnek, süzgeçteki kelimesi çözümlendiğinde

{Icerik: süzgeçteki Kok: süzgeç tip:IS} Ekler:ISIM_YALIN_BOS + ISIM_KALMA_DE + ISIM_BULUNMA_KI

Süzgeç kökü de kelimenin en yalin halidir. Daha çarpıcı bir örnek, şehirleştiremediklerimiz için çözümleme sonucu

{Icerik: şehirleştiremediklerimiz Kok: şehir tip:IS} Ekler:ISIM_YALIN_BOS + ISIM_DONUSUM_LES + FIIL_ETTIRGEN_TIR + FIIL_ISTEK_E + FIIL_OLUMSUZLUK_ME + FIIL_BELIRTME_DIK + ISIM_COGUL_LER + ISIM_SAHIPLIK_BIZ_IMIZ

{Icerik: şehirleştiremediklerimiz Kok: şehir tip:IS} Ekler:ISIM_YALIN_BOS + ISIM_DONUSUM_LES + FIIL_ETTIRGEN_TIR + FIIL_YETERSIZLIK_E + FIIL_OLUMSUZLUK_ME + FIIL_BELIRTME_DIK + ISIM_COGUL_LER + ISIM_SAHIPLIK_BIZ_IMIZ
{Icerik: şehirleştiremediklerimiz Kok: şehir tip:IS} Ekler:ISIM_YALIN_BOS + ISIM_DONUSUM_LE + FIIL_BERABERLIK_IS + FIIL_ETTIRGEN_TIR + FIIL_ISTEK_E + FIIL_OLUMSUZLUK_ME + FIIL_BELIRTME_DIK + ISIM_COGUL_LER + ISIM_SAHIPLIK_BIZ_IMIZ
{Icerik: şehirleştiremediklerimiz Kok: şehir tip:IS} Ekler:ISIM_YALIN_BOS + ISIM_DONUSUM_LE + FIIL_BERABERLIK_IS + FIIL_ETTIRGEN_TIR + FIIL_YETERSIZLIK_E + FIIL_OLUMSUZLUK_ME + FIIL_BELIRTME_DIK + ISIM_COGUL_LER + ISIM_SAHIPLIK_BIZ_IMIZ

Zemberek, sözlük tabanlı bir çözümdür ve bazı kelimeler henüz sözlüğe eklenmemiştir. Örneğin verimadenciliği kelimesinin çözümlenmesi sonucu alınan sonuç

verimadenciliği :çözülemedi

Böyle durumlarda kelime olduğu gibi alınıp, belgeleri oluşturan terimlerden biri olarak kabul edilmiştir.

6.4 Uygulamada Kullanılan Belge Setleri

Tez çalışması temelde Türkçe belgelerin madenciliğine odaklanmıştır. Belge madenciliği konusunda Türkçe dili üzerinde yeterince çalışma olmaması nedeniyle, tez çalışması kapsamında geliştirilen yöntemlerin olumlu sonuçlarını ispatlayabilmek için, bu yöntemler İngilizce belgeler üzerinde de test edilmiştir. Belge madenciliği çalışmalarında, değişik araştırmacıların geliştirdiği farklı yöntemleri karşılaştırabilmek için, tüm araştırmacıların kullanabilecekleri standart belge setleri kullanılır. İngilizce için standart olmuş pek çok belge seti olmakla beraber, Türkçe için standart olmuş bir derlem ya da belge kümesi henüz yoktur.

6.4.1 Türkçe Belge Seti

Üzerinde yeterince çalışılmış ve kabul edilmiş bir Türkçe belge seti yoktur. Tez çalışmasında kullanılmak üzere bir belge seti hazırlanmıştır. Son beş yıla ait yönetim dergileri içindeki makalelerden 615 adedi seçilmiş ve bir belge seti haline getirilmiştir. Bu makaleler 4-10 sayfa uzunlukla belgelerdir. Tüm makaleler 13 ana konuda toplanmıştır. Konu başlıkları

1- Aile şirketleri yönetimi
2- Bilgi yönetimi
3- Bütçe
4- Müşteri ilişkileri yönetimi
5- Belge yönetimi
6- Elektronik iş
7- Kurumsal kaynak planlama
8- İnsan kaynakları yönetimi
9- Liderlik
10- Pazarlama
11- Turizm
12- Verimlilik
13- Yönetim

Görüldüğü gibi hem belge sorgulamayı hem de kümelemeyi test edecek kadar farklı ve birbiri ile benzer ana kelimelere sahip bir belge seti oluşturulmuştur.

6.4.2 İngilizce Belge Seti

İngilizce belge seti olarak Reuters21578 koleksiyonu kullanılmıştır [12]. Bu koleksiyon 1987 yılına ait Reuters yayınlarında çıkan yazılardan oluşturulmuştur. Belgelerin derlemesi ve sınıflaması Reuters personeli tarafından elle yapılmıştır. 1996 Ağustosunda gerçekleştirilen ACM SIGIR '96 konferansında bir grup araştırmacı daha önce kullanılan Reuters-22173 belge setinin belge madenciliği çalışmalarında daha karşılaştırılabilir sonuçlar vermesini sağlaması için biçim ve içerik olarak daha yapısal bir koleksiyon oluşturulması gerektiğini düşünerek Reuters-21578'i ortaya çıkarmıştır. Bu çalışma Steve Finch and David D. Lewis tarafından yapılmıştır.

Reuters-21578 koleksiyonundaki belgeler kısa belgelerdir ve her belge birden fazla sınıfa aittir. 20.000 belge içeren bu setten, belirli sınıfları ifade etme gücü yüksek 1680 adet belge seçilmiş ve tez çalışmasında kullanılmıştır.

6.5 Atılabilir Kelimeler Listesi

6.5.1 Türkçe Atılabilir Kelimeler Listesi:

Standart bir "Türkçe Atılabilir Kelimeler Listesi" yoktur. Oluşturulan Türkçe belge setinin içinden elle seçilmiş bir liste hazırlanmıştır. Bu liste içinde belgelerin anlamını etkilemeyen kelimeler yer almaktadır. Bu liste Ek 1'de verilmiştir.

6.5.2 İngilize Atılabilir Kelimeler Listesi:

İngilizce için standart "İngilizce Atılabilir Kelimeler Listesi" vardır. PorterStemmer algoritması içinde kullanılan liste kullanılmıştır. Bu liste Ek 2'de verilmiştir.

6.6 Uygulama

İngilizce ve Türkçe belgeler üzerinde yapılan belge madenciliği çalışmaları için hangi adımlarla, ne gibi işlemler yapıldığını özetlersek

1. Belge havuzu oluşturma : Türkçe belge seti olarak Son beş yıla ait iş dünyası ve yöneticilikle ilgili değişik dergilerden toplanan 615 adet makale derlemiştir. Bu makaleler 13 sınıfta toplanmıştır.İngilizce belge seti olarak, belge madenciliği çalışmalarında uluslararası standart kabul edilmiş Reuters21578 belge seti kullanılmıştır [12].
2. Türkçe kelimelerin anlam kaybetmeyecekleri en yalın hallerine çevrilmeleri için Zemberek [13] projesi kapsamında geliştirilen sözlük ile Türkçe kelimelerin çözümlenmesi için modül oluşturulmuştur. İngilizce belgeler için standart Porter Stemmer [13] algoritması kullanılmıştır.
3. Türkçe için, belgelere anlam katmayan kelimelerin tespiti yapılmıştır. İngilizce için stoplist adı altında var olan standart liste kullanılmıştır.

4. Her kelimenin, içinde geçtiği belge için önem derecesini belirten matris değeri, bilgiye erişme yöntemlerinde sıkça kullanılan tfidf yöntemi ile hesaplanmış. Elde edilen matrise Tekil Değer Ayrıştırması (TDA) uygulanmıştır (Iogerner, 1998).
5. TDA işlemi sonucu elde edilen daha düşük boyutlu uzayda hiyerarşik kümeleme algoritması ile kümeleme yapılmıştır (Berry vd., 1993). Aynı uzayda belge sorgulama örnekleri gerçekleştirilmiştir.

6.7 GAD ve n-gram destekli GAD ile sınıflama

GAD yönteminin en büyük avantajı eğitmensiz bir sistem oluşturmaya katkısıdır. Eğitmensiz sistemlerde sistemin önceden bir eğitim setine göre eğitilmesi gerekmez ancak sınıflama işlemi söz konusu olduğunda, bir metin yığınının daha önceden belirlenen sınıflara dağıtımı önem kazanır ki bu durumda bu sınıfı GAD kavram uzayında en iyi temsil edecek kavramların tesbiti kritik hale gelir. Bunun yanında doğru SDD değerinin belirlenmesi de sınıflama işleminin başarımını belirleyecek diğer bir kritik parametreyi teşkil eder.

GAD yöntemi ile sınıflama yapma işlemini, GAD kavram uzayında ifade edilmiş bir sınıf ile yine bu uzaydaki metinleri benzerlikleri ölçüsünde bir araya getirme işlemi olarak yapılandırabiliriz. Bu GAD ile sorgulama yapmaya çok yakın bir işlemdir. Nasıl ki bir sorgu ifadesi GAD kavram uzayına taşınıp, bu sorgu vektörü ile diğer metin vektörleri arasındaki benzerlik değeri (Öklid uzaklığı) belirli bir eşik değerini aşan metinler sorgu neticesi olarak dönüyorsa, sınıf tanımı ile arasındaki benzerlik değeri belirli bir eşiği aşan metinler de o sınıfa ait olarak değerlendirilebilir.

Çalışmanın asıl amacı GAD ile n-gram destekli GAD yöntemlerinin sınıflama konusundaki başarımlarını karşılaştırmak olduğu için her iki yöntem için de sabit bir eşik değeri belirlemek gerekecektir. Bu eşik değeri kavram uzayını oluşturmakta kullanılan metin sayısına göre değişiklik gösterecektir çünkü daha fazla kavramın değerlendirilmesi ile elde edilen kavram uzayında kavramlar arasındaki uzaklıklar, daha az kavramın değerlendirilmesi neticesinde elde edilen kavram uzayındaki uzaklıklardan farklı olacaktır.

7. BULGULAR

7.1 Kümeleme

Belge madenciliği alanına yönelik, hızla artan belge sayısı faktörüne karşı, yeni nesil belge madenciliği teknikleri geliştirme amacımız gereği, bu alanda başarılı bir teknik olan GAD yöntemini, n-gram tekniği ile birleştirerek yeni bir yöntem önermiştik.

Bu bölümdeki amaç, mevcut kümeleme tekniklerinden daha iyi bir kümeleme tekniği geliştirmektir. Bu amaçla GAD yöntemi n-gram yaklaşımı ile geliştirilmiştir. Geliştirilen yeni yöntemin de eğitmensiz bir sistem olması, incelenecek belge sayısındaki artışa göre dışarıdan müdahale gerektirmeden ölçeklenebilmesi, bilgiye erişim alanına yönelik başarılı ve önemli bir katkı olacaktır.Önerilen yöntemin karşılaştırılabilir olması için GAD ve *n-gram kelimelerle GAD* ile elde edilen belge dizinleri hiyerarşik kümeleme tekniği kullanılarak kümelenmiştir. Tablo 7-1'de kullanılan kümeleme yaklaşımları gösterilmiştir.

Tablo 7-1 Karşılaştırma için Kullanılan Algoritmalar

Algoritma	Açıklaması
Tfidf_GAD_HK	GAD tabanlı hiyerarşik kümeleme (Tfidf tabanlı vektör uzayı, tam bağlı)
Tfidf_ngram_GAD_HK	n-gram destekli GAD tabanlı hiyerarşik kümeleme (Tfidf tabanlı vektör uzayı, tam bağlı)

7.1.1 Türkçe Belgelerin Kümelenmesi

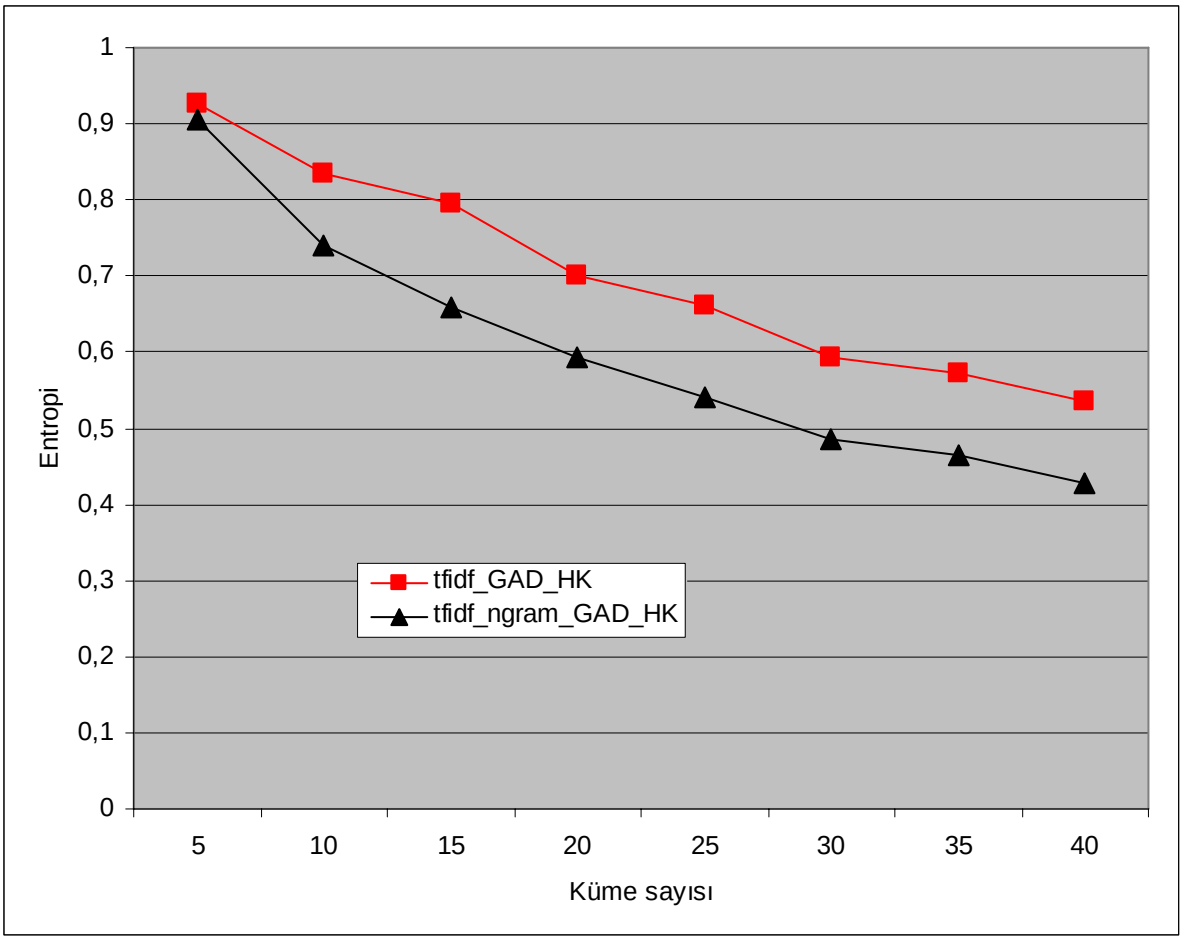
13 sınıfta kümelenmiş 615 Türkçe belgea ait entropi ve f-measure ölçümleri Tablo 7-2 ve Tablo 7-3' de gösterilmektedir. Bu tablolara ait grafikler Şekil 7-1 ve Şekil 7-2'te paylaşılmıştır.

Görüldüğü gibi *n-gram kelimelerle GAD* yöntemi hem kümelerin homojenliği (düşük entropi değeri) hem de kümeleme işlemenin kalitesi (yüksek f-measure değeri) açısından GAD yöntemine göre ciddi üstünlük sağlamıştır. Türkçe belgelerin normalde 13 sınıfta toplandığı belirtmiştir. Bu belgelerin 13 kümeye dağıtımında n-gram destekli GAD daha başarılı olduğu gibi daha fazla kümeye ayrıştırıldığında da daha kaliteli ve homojen kümeler oluşturmuştur. Bunun

anlamı şudur; bir kümeye ait belgeler da kendi içinde farklı konulara ayrıştırılabilir ve önerilen *n-gram kelimelerle GAD* yöntemi bu ayrıştırma işlemini çok daha başarılı bir şekilde yapmaktadır.

Tablo 7-2 -Türkçe Belge Seti için Entropi Ölçümleri

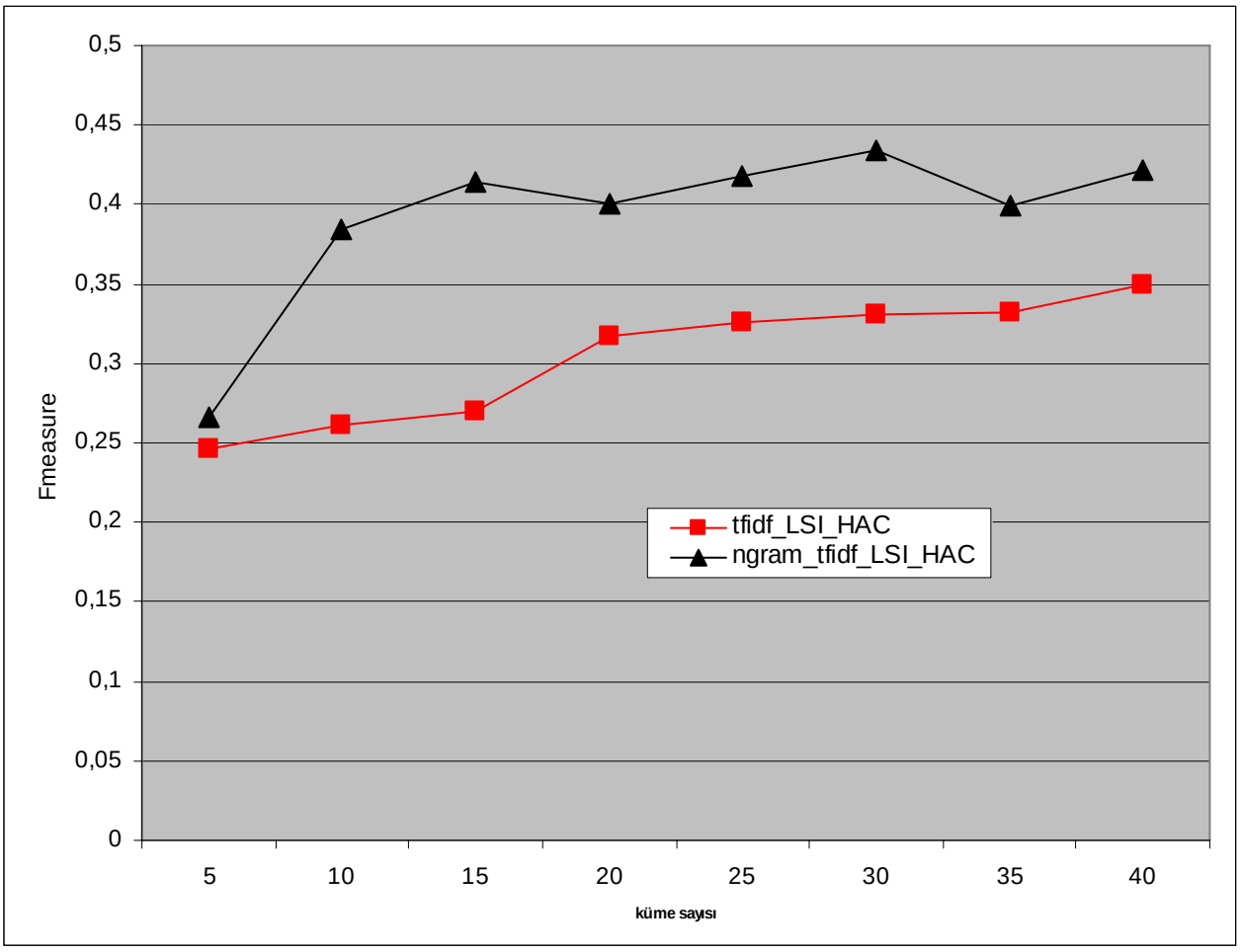
Küme sayısı	Tfidf_GAD_HK	Tfidf_ngram_GAD_HK
5	0,926381306	0,90608821
10	0,834725048	0,73937549
15	0,795165679	0,65872017
20	0,700502409	0,59382918
25	0,660731649	0,54026323
30	0,594019001	0,48636027
35	0,571697287	0,46587803
40	0,534692964	0,42687611



Şekil 7-1- Türkçe Belge Seti için Küme Sayısına Bağlı Entropi Değişim Grafiği

Tablo 7-3 Türkçe Belge Seti için F-measure Ölçümleri

Küme sayısı	Tfidf_GAD_HK	Tfidf_ngram_GAD_HK
5	0,246469381	0,26671552
10	0,261534557	0,38382652
15	0,269808758	0,41365287
20	0,316816249	0,39988699
25	0,325498299	0,4177481
30	0,331417705	0,43376812
35	0,331566884	0,39960739
40	0,349225421	0,42172488



Şekil 7-2 Türkçe Belge Seti için Küme Sayısına Bağlı F-Measure Değişim Grafiği

7.1.2 İngilizce Belgelerin Kümelenmesi

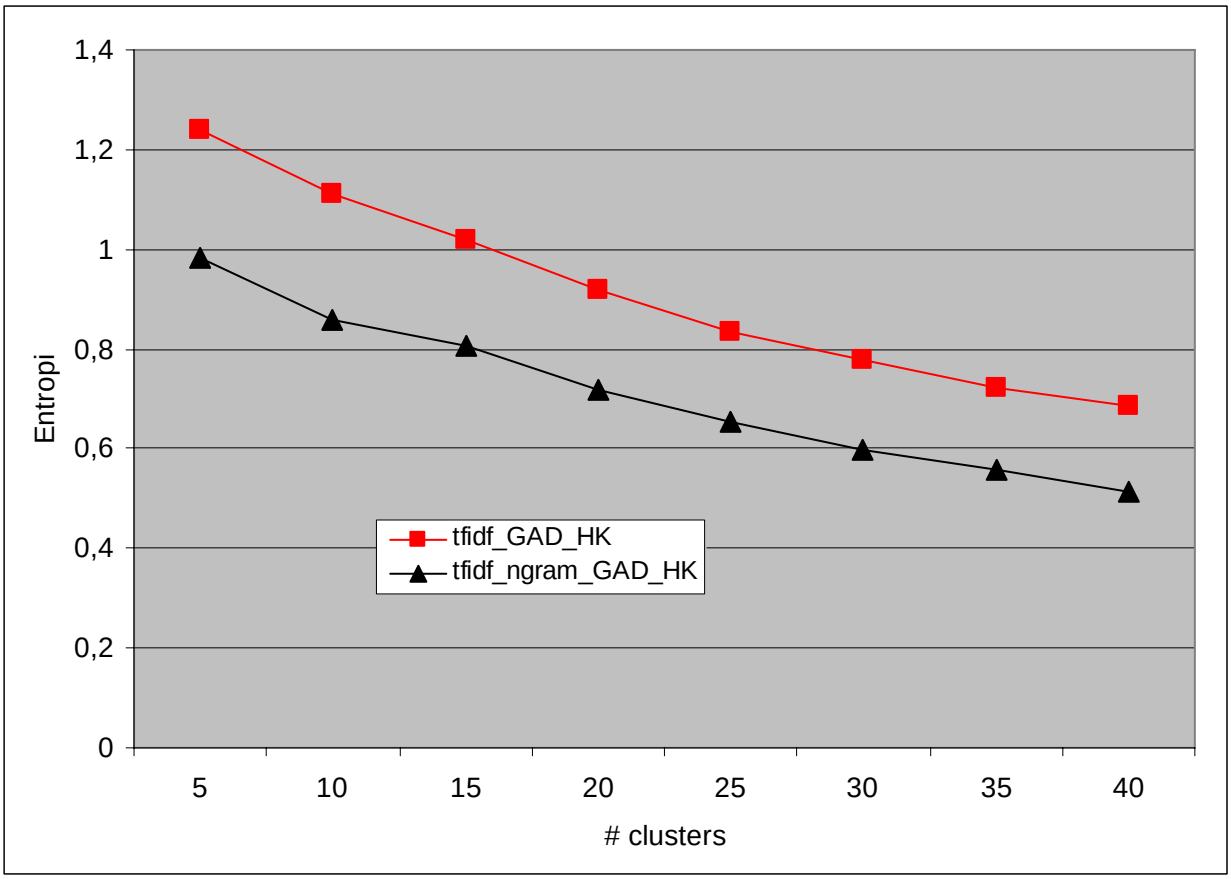
İngilizce belge seti olarak Reuters21578 setinden rasgele seçilen 1680 belge kullanılmıştır. Bu sette 20000 belge bulunur ve bu belgeler 1987 yılında Reuters yayınlarında yayınlanmış yazılardan derlenmiştir. Bu sete ait entropi ve f-measure ölçümleri Tablo 7-4 ve Tablo 7-5'de bulunmaktadır ve bu tablolara ait grafikler Şekil 7-3 ve Şekil 7-4'de gösterilmiştir. Bu set içindeki belgeler 20 sınıfta toplanmıştır [14].

Tablo 7-4 İngilizce Belge Seti için Entropi Ölçümleri

Küme sayısı	Tfidf_GAD_HK	Tfidf_ngram_GAD_HK
5	0,926381306	0,90608821
10	0,834725048	0,73937549
15	0,795165679	0,65872017
20	0,700502409	0,59382918
25	0,660731649	0,54026323
30	0,594019001	0,48636027
35	0,571697287	0,46587803
40	0,534692964	0,42687611

Tablo 7-5 İngilizce Belge Seti için F-Measure Ölçümleri

Küme sayısı	Tfıdf_GAD_HK	Tfidf_ngram_GAD_HK
5	0,246469381	0,26671552
10	0,261534557	0,38382652
15	0,269808758	0,41365287
20	0,316816249	0,39988699
25	0,325498299	0,4177481
30	0,331417705	0,43376812
35	0,331566884	0,39960739
40	0,349225421	0,42172488

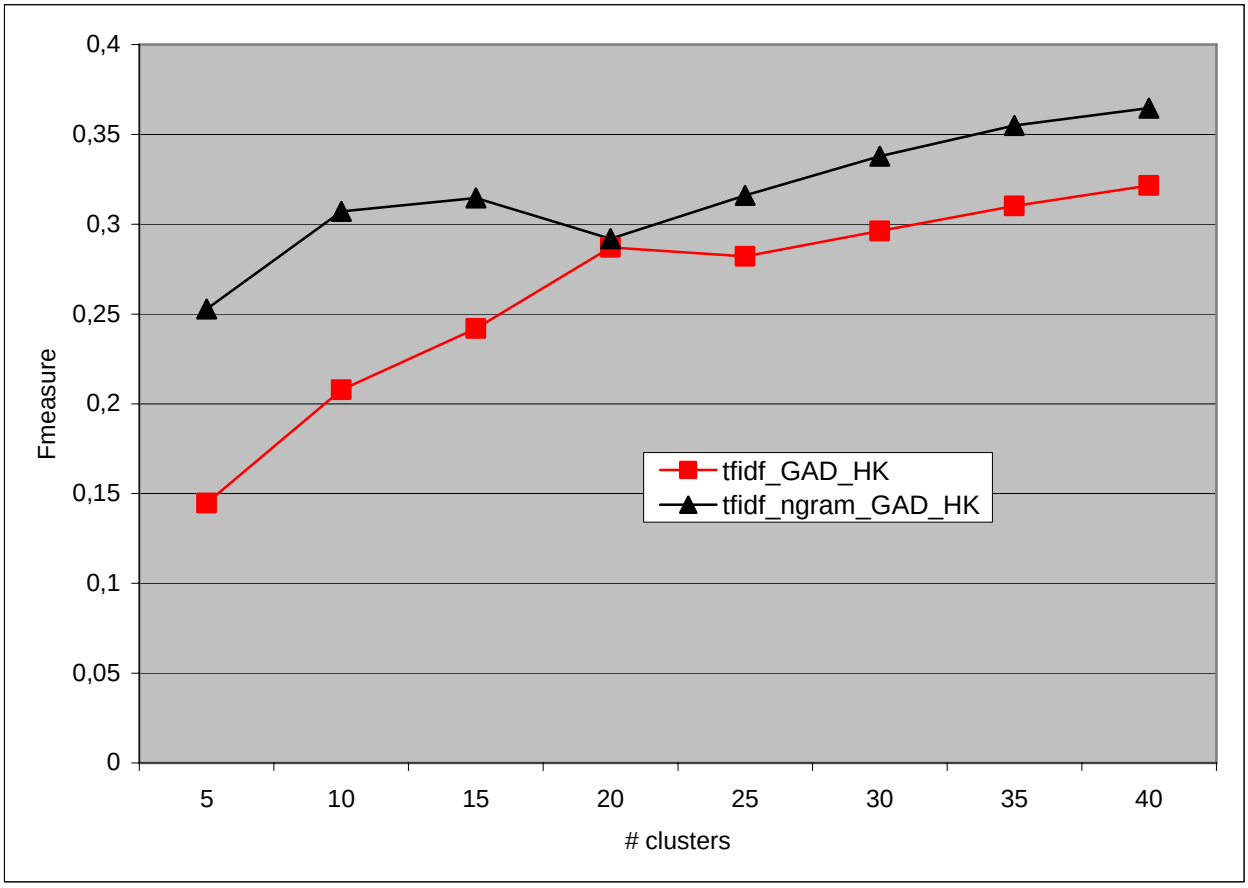


Şekil 7-3 İngilizce Belge Seti için Küme Sayısına Bağlı Entropi Değişim Grafiği

İngilizce belgelerin kümelenmesi işleminde de önerdiğimiz *n-gram kelimelerle GAD* yöntemi, normal GAD yöntemine göre çok daha başarılı sonuçlar üretmiştir. Bu setteki belgelerin uzmanlar tarafından 20 grupta sınıflandırıldığı söylemiştik. Normal GAD yöntemi sadece bu sayıda küme için *n-gram kelimelerle GAD* yöntemini yakalamıştır ancak bu küme sayısında dahi, kümelerin homojenliği açısından n-gram destekli GAD daha başarılı netice vermiştir.

7.2 Sorgular

Kümeleme testlerinde elde edilen sonuçlara paralel şekilde gerek Türkçe belgeler, gerek İngilizce belgeler üzerinde yapılan sorgulama çalışmalarında –gram destekli GAD, normal GAD'ye kıyasla daha başarılı olmuştur.



Şekil 7-4 İngilizce Belge Seti için Küme Sayısına Bağlı F-Measure Değişim Grafiği

Sorgulama söz konusu olduğunda da, kümeleme işleminin değerlendirilmesinde kullanılan entropi ve f-measure ölçümlerine benzer ölçümler bulunmaktadır. Bunları daha önce çağırma ve Duyarlılık olarak tanımlamıştık. Ancak ikinci nesil belge madenciliği diye nitelendirdiğimiz yeni belge analizi ihtiyaçlarında bu ölçütleri kullanmak pek mümkün değildir. Bu ölçütler belgelerin içerikleri ile ilgili ana konularda gerçekleştirilen sorgulamalarda, sorgulama işleminin başarısını ölçebilmektedir. Oysa bizim belge sorgulama ihtiyacımız, belgelerin ana konularını aşan, ana konular arasındaki bağlantıları ve ilişkileri sorgulayan daha detay bir ihtiyaç için yapılmaktadır. Yani ortada bir ana konu ve bu ana konu hakkında olduğunu bildiğimiz belgeler yoktur. Bu nedenle yapılan örnek sorgulamaların sonucunu bir ölçüt ile ifade etmek mümkün değildir. Bu bölümde İngilizce ve Türkçe belgeler üzerinde yapılan örnek sorgulamalardan örnekler verilerek önerilen n-gram destekli GAD yönteminin GAD yöntemi ile karşılaştırıldığında nasıl bir üstünlük ortaya koyduğu paylaşılacaktır.

7.2.1 Türkçe Belgelerin Sorgulanması

Tablo 7-6'da "bilgi yönetiminin pazarlanması" konusunda yapılan bir örnek sorgulama görülmektedir. Bu sorgulama sonucunda beklenen, bilgi yönetimi anlayışının organizasyon içinde veya dışında pazarlanması ile ilgili belgelerdir. Görüldüğü gibi normal GAD yöntemi konuyu pazarlama ağırlıklı bir içerik aranması şeklinde kabul etmiş ve belge seti içinden pazarlama konusuyla ilgili belgeleri sorgulama sonucu olarak döndürmüştür. Oysa n-gram destekli GAD konunun bilgi yönetimi ile ilgili olduğunu fark ederek bu konuyla ilgili belgeleri döndürmüştür. Pazarlama ile ilgili belgelerden neredeyse hiçbirini, ilgili belge sınıfında değerlendirmemiştir. Normal GAD kelimeleri birbirinden bağımsız olarak değerlendirdiğinden hata yaparken, n-gram destekli GAD yöntemi kelimeleri hem tekil halleri hem de 2-gram ve 3-gram halleri ile ele aldığından istenilen sonucu üretmiştir. Sorgulama sonuçlarını bir kez daha karşılaştırdığımızda, doğru sonuç vermeyen bir yöntemin, kendisinden beklenen faydanın tam tersi yönde, ne kadar zarar verdiği de ortaya çıkmaktadır.

Tablo 7-6 "Bilgi Yönetiminin Pazarlanması" Sorgusu

Bilgi Yönetiminin Pazarlanması	
Normal GAD	**Uygunluk**
Pazarlama Stratejileri	0,573
Crm - 12 Soruda Bire-Bir Pazarlama	0,565
Pazarlama - Başarinin Sirri Farklilaşmada	0,481
Pazarlama - Gerilla Usulu Pazarlamanin Onbeş	0,419
Pazarlama Performansini Ölçüyor Musunuz	0,388
Pazarlama Şirketlere Hocadan Not	0,363
Pazarlama Kotlerden Yarina Dersler	0,326
Pazarlama Yarim Einstein Yarim Picasso	0,322
Pazarlamada Yeni Trendler	0,297
n-gram destekli GAD	**Uygunluk**
Bilgi Yönetimi Nedir	0,557
Bilgi Yönetimi	0,471
Bilgi Yönetimi Ve Uygulamaları	0,421
Bilgiyi İyi Yöneten Karini Artırıyor	0,399
Bilgi Yön-Mitoslar Ve Gerçekler	0,39
Bilgi Yön-Etkin Bilgi Yönetimi	0,385
Pazarlama Stratejileri	0,322
Bilgi Yön-Şirketiniz Bilgi Yönetimine Hazır Mı	0,328
Crm - 12 Soruda Bire-Bir Pazarlama	0,327

Bir başka çarpıcı örnek Tablo 7-7'de gösterilen "kurumsal kaynak planlama" sorgusunda incelenebilir. Normal GAD kurum, kaynak ve plan terimlerini birbirinden bağımsız olarak ele almış ve bunlarla ilgili belgeler içinde en fazla bu terimleri içeren belgeleri sorgulama sonucu olarak döndürmüştür. Özellikle "kurumsal yönetim" konusunu ön plana çıkartmıştır. Oysa "kurumsal yönetim" ile "kurumsal kaynak planlama" birbirinden apayrı konulardır. Normal GAD yine yanıltmıştır. Diğer taraftan n-gram destekli GAD yöntemi ise, aynı sorgu için doğru belgeleri döndürerek başarısını göstermiştir.

Tablo 7-7 "Kurumsal Kaynak Planlama" Sorgusu

Kurumsal Kaynak Planlama	
Normal GAD	**Uygunluk**
aile - Bay Kurumsal Yönetim.txt	0,324
bilgi yön-Etkin Bilgi Yönetimi.txt	0,298
yönetim - Kurumsal Yönetim Ateşi Tüm Anadolu"yu Saracak.txt	0,281
Aile şirketinden kurumsal yapıya.txt	0,239
yönetim - Şirketler nasıl konuşuyor.txt	0,223
crm - Ben müşteriyim - CRM İnstitut.txt	0,22
Bilgi Yönetimi Ve Teknoloji İlişkisi.txt	0,22
insan kay - Kurumsal Wellness Kârı Da Artırıyor.txt	0,212
erp – İşletme Kaynakları Planlamasının Dünü Bugünü Yarını.Txt	0,204
n-gram destekli GAD	**Uygunluk**
erp – İşletme Kaynakları Planlamasının Dünü Bugünü Yarını.txt	0,395
erp - Kurumsal Kaynak Planlaması Yatırımları.txt	0,336
erp - ERP KOBI'lere doğru yelken açıyor.txt	0,332
ERP KOBI'lere doğru yelken açıyor.txt	0,332
eiş - Çözümler Peş Peşe.txt	0,298
erp - E-şirketin sinir sistemi.txt	0,263
ERP uygulamalarında başarı faktörü.txt	0,234
erp - ERP uygulamalarında başarı faktörü.txt	0,234
Yönetim - Kurumsal Yönetim Ateşi Tüm Anadolu"Yu Saracak.txt	0,227

n-gram destekli GAD yönteminin, normal GAD yöntemine karşı sunduğu üstünlüğü tam anlamıyla gösteren güzel bir örnek Tablo 7-8'de "turizm sektöründe insan kaynakları yönetimi" sorgusuyla sunulmaktadır. Bu sorgu "insan kaynakları yönetimi" ana konusu içinde turizm sektörüne yönelik belgelerin döndürülmesini amaçlamaktadır. Normal GAD yöntemi tamamen turizm konusuyla ilgili ama arama amacına ile ilgisiz belgeleri döndürmüştür. N-gram destekli GAD yöntemi ise tam isabet göstererek doğru belgeleri döndürmüştür. Belge isimlerinin başında

belgelerin ilgili olduğu konuları gösteren ifadeler yer almaktadır. Bazı belgeler birden fazla konu ile ilgili olduklarından, bu belgeler belge seti içinde birden fazla isimle bulunmaktadır. Bu örnekteki sorgunun sonucunda aynı belge hem turizm alanından hem de insan kaynakları alanından, sorgu sonucu olarak döndürülmüştür.

Tablo 7-8 "Turizm Sektöründe İnsan Kaynakları Yönetimi" Sorgusu

Turizm Sektöründe İnsan Kaynakları Yönetimi	
Normal GAD	**Uygunluk**
turizm - Akdeniz Çanağının yeni yıldızı.txt	0,767
turizm - yatak kralları.txt	0,708
turizm - Turizmde Ürün Çeşitliliği Olmalı.txt	0,69
turizm - Türkiye de Turizm ve bilgi teknolojileri.txt	0,681
turizm - İç Turizmi Canlandırmaya Yönelik Bir Pazarlama Önerisi.txt	0,65
turizm - Dünyaya Otel Konsepti.txt	0,554
eiş - turizm bakanlığında e-iş.txt	0,548
turizm - 2006 büyümesi yüzde 15 olur.txt	0,534
crm - Turizm sektörü ve akıllı pazarlama.txt	0,449
n-gram destekli GAD	**Uygunluk**
insan kay - Otel İşletmelerinde İnsan Kaynakları Yönetiminin Yeri Ve Önemi.txt	0,591
turizm - Otel İşletmelerinde İnsan Kaynakları Yönetiminin Yeri Ve Önemi.txt	0,591
turizm - Turizmde Ürün Çeşitliliği Olmalı.txt	0,536
turizm - yatak kralları.txt	0,481
insan kay - İnsan Kaynakları Yönetiminin Geleceği.txt	0,459
pazarlama - Turizm sektörü ve akıllı pazarlama.txt	0,428
crm - Turizm sektörü ve akıllı pazarlama.txt	0,428
turizm - Akdeniz Çanağının yeni yıldızı.txt	0,416
turizm - İç Turizmi Canlandırmaya Yönelik Bir Pazarlama Önerisi.txt	0,375

7.2.2 İngilizce Belgelerin Sorgulanması

Türkçe belge setinin hazırlanması aşamasında, her belgenin hangi konu ile ilgili olduğu, belge başlığının önüne getirilen bir ifade ile belirlenmiştir. Benzer şekilde İngilizce belgelerin isimlerinde de hangi konu ile ilgili olduğunu gösteren ifadeler yer almaktadır. Ancak belge isimleri yerine belge numaraları kullanılmıştır. Sadece belge isimlerine bakarak n-gram destekli GAD yönteminin, normal GAD yöntemine karşı sunduğu avantajı yakalamak kolay değildir.

Tablo 7-9'de gösterilen birinci örneğimizde "petrol şirketi satın alınması" konusu ile ilgili sorgulama yapılmaktadır. Normal GAD yöntemi yine konuyu satın alma ile ilişkilendirip satın alma ile ilgili belgeleri döndürmüştür. N-gram destekli GAD ise petrol şirketleri ve satın alma konusunu ortak ele alıp, buna uygun belgeleri sorgulama neticesi olarak getirmiştir.

Tablo 7-9 "Oil Company Acquisition" Sorgusunun Sonuçları

Oil Company Acquisition	
Normal GAD	**Uygunluk**
acq,-19826	0,365
acq,crude,-4315	0,363
grain,oilseed,veg-oil,sorghum,sun-	0,293
meal-15500	0,27
acq,crude,-9913	0,253
acq,-19837	0,252
acq,crude,nat-gas,-16007	0,246
gold,-16589	0,244
acq,crude,nat-gas,-8100	0,237
heat,-5706	
n-gram destekli GAD	**Uygunluk**
gas,fuel,-17441	0,479
gas,-19083	0,367
acq,crude,-4315	0,303
heat,-19223	0,292
gas,-17173	0,29
heat,gas,-11880	0,269
acq,crude,-5116	0,229
acq,-19984	0,216
gas,-18367	0,211

İngilizce belgelerle ilgili ikinci örnek "petrol endüstrisindeki tüketici fiyatları" ile ilgili belgeleri döndürmeyi amaçlamaktadır – Tablo 7-10. Burada da n-gram destekli GAD ile normal GAD sonuçları arasındaki farklılık göze çarpmaktadır. Bu farkın sebebi, yukarıdaki örneklerdeki gibi, n-gram destekli GAD yönteminin kelimeleri ve kelimelerin birlikte kullanım örüntülerini dikkate alarak sonuç üretmesidir.

Bu örnekler dışındaki testlerde ise karşılaşılan durum, normal GAD ile n-gram destekli GAD'nin benzer sonuçlar üretmesi ile sonuçlanmıştır.

Tablo 7-10 "Oil İndustry Consumer Price " Sorgusunun Sonuçları

Oil İndustry Consumer Price	
Normal GAD	**Uygunluk**
acq,crude,-9913	0,347
grain,oilseed,veg-oil,sorghum 15500	0,312
acq,crude,-9947	0,269
grain,rice,-19059	0,258
cpi,-10260	0,256
acq,sugar,crude,-11213	0,236
heat,-5706	0,232
heat,-12670	0,23
heat,gas,-11880	0,227
n-gram destekli GAD	**Uygunluk**
cpi,-10260	0,481
acq,crude,-9913	0,333
acq,crude,-9947	0,253
cpi,-6982	0,242
cpi,-6920	0,231
cpi,-10553	0,224
cpi,-20247	0,221
cpi,-8726	0,217
cpi,-16950	0,211

7.3 Sınıflama

Bu bölümde GAD ve n-gram destekli GAD yöntemlerinin sınıflama işlemi noktasındaki başarımları karşılaştırılmıştır.

Bu karşılaştırma işleminin testi için 7 sınıfa daha önceden uzmanlar tarafından ayrıştırılmış metin yığını oluşturulmuştur. Toplam 308 adet metin içeren bu yığının sınıfları ve her sınıfa düşen metin sayısı Tablo 7-11'de verilmiştir.

Tablo 7-11 Sınıflama İşlemi için Önceden Hazırlanan Sınıflar

Sınıf	**Adet**
insan kaynakları	82
Kurumsal kaynak planlama	52
Müşteri İlişkileri Yönetimi	62
Bilgi Yönetimi	37
Aile şirketi konuları	12
turizm	37
liderlik	26

Sınıflama işleminin yapılması için sınıf tanımlarının belirlenmesi ve GAD kavram uzayına taşınması aşamasında iki farklı sınıf tanımlaması denenmiştir.

1. sınıf tanımı için çok basit anahtar kelimeler kullanılmıştır, Tablo 7-12

Tablo 7-12 Sınıf Tanımlaması 1

Sınıf	**Sınıf tanımı**
insan kaynakları	insan kaynakları Yönetimi
Kurumsal kaynak planlama	Kurumsal kaynak planlama
Müşteri İlişkileri Yönetimi	Müşteri İlişkileri Yönetimi
Bilgi Yönetimi	Bilgi Yönetimi
Aile şirketi konuları	Aile şirketi
turizm	Turizm
liderlik	liderlik

2. sınıf tanımı için örnek metinler toplanmıştır. Bunlara Ek 1'den ulaşılabilir.

Yapılan testlerde kavram uzayı çok geniş olmadığı için SDD olarak 0,1 değerinin kullanılması uygun görülmüş ve sınıf tanımı ile bu eşik değerinin üstünde benzerlik değerine sahip metinler bu sınıfa ait olarak belirlenmiştir. Yapılan sınıflama çoklu sınıflamadır yani bir metin birden fazla sınıfa ait olabilir.

Sınıf tanımları GAD kavram uzayına taşınmadan önce, 615 adet daha önceden toplanmış metin ile kavram uzayı oluşturulmuştur. Sınıflanacak metinler de bu uzaya taşınmış ve en son sınıf tanımı da bu uzaya aktarılarak, sınıf tanımı ile metinler arasında karşılaştırma yapılmıştır. Elde edilen neticeler çağırma ve duyarlılık kıstaslarına göre değerlendirilmiştir. İki farklı sınıf tanımı için elde edilen neticeler Tablo 7-13, Tablo 7-14, Tablo 7-15, Tablo 7-16'da, aşağıdadır.

Tablo 7-13 1. Sınıf Tanımına Göre Sınıflama

	Sınıflama Sonucu					**Çağırma**		**Duyarlılık**	
	Metin sayısı	GAD		Ngram GAD		Normal GAD	Ngram GAD	Normal GAD	Ngram GAD
		çağrılan	toplam	çağrılan	toplam				
İnsan kaynakları	82	16	30	22	29	0,20	0,27	0,53	0,76
Kurumsal kaynak planlama	52	9	51	21	45	0,17	0,40	0,18	0,47
Müşteri ilişkileri yönetimi	62	40	99	44	60	0,65	0,71	0,40	0,73
Bilgi yönetimi	37	26	45	26	31	0,70	0,70	0,58	0,84
Aile şirketi yönetimi	12	12	25	12	19	1,00	1,00	0,48	0,63
Turizm	37	19	27	20	27	0,51	0,54	0,70	0,74
Liderlik	26	22	49	22	28	0,85	0,85	0,45	0,79

Tablo 7-14 2. Sınıf Tanımına Göre Sınıflama

	Sınıflama Sonucu					**Çağırma**		**Duyarlılık**	
	Metin sayısı	GAD		Ngram GAD		Normal GAD	Ngram GAD	Normal GAD	Ngram GAD
		çağrılan	Toplam	çağrılan	toplam				
İnsan Kaynakları	82	7	20	32	68	0,085	0,40	0,35	0,47
Kurumsal kaynak planlama	52	47	96	43	52	0,90	0,83	0,49	0,83
Müşteri ilişkileri yönetimi	62	39	58	37	51	0,63	0,60	0,67	0,73
Bilgi yönetimi	37	24	44	22	25	0,65	0,60	0,55	0,88
Aile şirketi yönetimi	12	12	32	12	20	1	1	0,37	0,6
Turizm	37	20	28	21	26	0,54	0,58	0,71	0,80
liderlik	26	20	107	22	28	0,77	0,85	0,19	0,79

Tablo 7-15 1. Sınıf Tanımına Göre Sınıflamanın Etkinliği

Çağırma		**Duyarlılık**	
Normal GAD	Ngram GAD	Normal GAD	Ngram GAD
0,46	0,54	0,44	0,70

Tablo 7-16 1. Sınıf Tanımına Göre Sınıflamanın Etkinliği

Çağırma		**Duyarlılık**	
Normal GAD	Ngram GAD	Normal GAD	Ngram GAD
0,55	0,61	0,44	0,7

Görüldüğü gibi sınıf tanımlamaları sınıflama işleminin başarısı için önemlidir. Farklı sınıf tanımlamaları farklı neticeler verebilmektedir. GAD ile ngam destekli GAD karşılaştırması söz konusu olduğunda ise ngram destekli GAD'nin sınıflaması, çağırma kriteri için %6-8, duyarlılık kriteri için ise % 25 daha etkin sonuç oluşturmaktadır.

8. SONUÇ

Belge veri madenciliği çalışmaları, insanların öğrenme şekillerini inceleyerek, karar verme süreçlerini modellemeye çalışır. Bu öğrenme sürecinde nesneler, nesneler arasındaki ilişkiler, nesnelerin özellikleri, nesnelerin durum ve özelliklerini değiştiren fiiller öğrenilir. Farklı kültür ve sosyal ortamlarda insanlar aynı nesneleri, özelliklerini, ilişkilerini, fiilleri tanımlamak için farklı kelimeler kullanırlar (Landauer vd., 1998) (Wolfe vd., 1998).

Büyük belge yığınlarını analiz etmek (öğrenmek) için temel iki yaklaşım kullanılır.

1- Nesne, ilişki, durum, özellik ve fiiller için, daha önceden belirlenmiş ve standart olarak kabul edilmiş kavram haritaları kullanmak
2- Nesne, ilişki, durum, özellik ve fiilleri mevcut belge yığını içinden çıkarmak

Birinci grupta ontolojiler kullanılır. Ontolojiler uzmanlar tarafından elle hazırlanır ve sürekli güncellenir. Bu nedenle oldukça pahalı bir yöntemdir. İkinci grupta ise Gizli Anlambilimsel Dizinleme (GAD) gibi istatistiki yöntemler yer alır. Ontoloji kullanımı kadar net öğrenme gerçekleşmese de daha hızlı bir yöntemdir ve yeteri kadar iyi neticeleri ile kendisini ispatlamıştır.

GAD yöntemi nesneleri, ilişkilerini, özelliklerini ve fiilleri öğrenmek için belgelerdeki kelimeleri kullanır. GAD yöntemi belge yığınlarını işlemek için altyapısında vektör uzay modelinden yararlanır. Bu modelde her bir belge, kendisini oluşturan kelimelerin boyut kabul edildiği uzayda, bir vektör olarak tanımlanır. Belgelerin birbirleriyle ilişkisi, vektörel karşılıklarının birbirine vektör uzayındaki yakınlıkları ile ölçülür. Vektör uzay modelinin en büyük üç eksikliği, GAD'nin istatistiki yapısı sayesinde giderilmektedir (Kontostathis vd., 2006):

- eşanlamlı ve çok anlamlı kelimelerin oluşturduğu karışıklık
- sadece belgelerin karşılaştırılması yanında, kelimelerin kendi aralarında da karşılaştırılabilmeleri
- büyük belge yığını içindeki çok sayıda kelime nedeniyle büyüyen uzayın daha az boyut ile ifade edilmesi

GAD yönteminin eşanlamlılığı ve çok anlamlılığı çözmek ve birbiriyle ilgili kelimeleri belge yığını içindeki kullanım istatistiklerine bakarak çıkarmak için sunduğu çözüm insan öğrenmesine benzeyen bir yaklaşım sunar. Ancak belgelerin anlamını ortaya çıkarmak için daha çözülmesi

gereken başka konular da vardır. Bunlardan en önemlisi, özellikle Türkçe'deki gibi, kelimelerin yan yana geliş düzenlerine göre kazandıkları anlamı ortaya çıkarmaktır. GAD yöntemi bu noktda yetersiz kalmaktadır. Aynı kelimelerden oluşan ancak kelimelerin birbirine göre göreceli konumları farklı olan ifadeler sadece kelimeleri kullanan GAD gibi sistemler tarafından aynı anlamı ifade ediyor şeklinde algılanır. Örneğin "verimli bir şirket nasıl olur", "bir şirket nasıl verimli olur", "nasıl bir şirket verimli olur" soruları kelime bazında ele alındığında aynı anlamı ifade ediyor olsa da gerçekte kelimelerin göreceli konumları gereği farklı anlamlar taşır. GAD yönteminin bu eksikliğini gidererek, daha güçlü bir GAD elde etmek ve daha iyi öğrenen bir sistem elde etmek mümkün olacaktır.

GAD yöntemi kelimeleri baz alır ancak kelimelerin anlamını dikkate almaz. Bu nedenle anlamlı birliktelik oluşturan kelimelerin tesbiti sorundur. Ancak eğer bir belge içinde geçen kelime gruplarını sistematik bir şekilde, sanki birer kelimeymiş gibi GAD yöntemine aktarabilirsek, kelimelerin oluşturacağı anlamlı grupları da indekslemiş oluruz. Kelime gruplarının sistematik olarak, yani anlamlarını değerlendirmeden sisteme almanın yolu olarak n-gram yaklaşımı kullanılabilir. N-gram yaklaşımı, bir dili oluşturan kelimelerdeki harflerin yan yana kullanım örüntülerini tespit etmek ve buradan yola çıkarak, bir belgedaki kelimelerin uyduğu n-gram haritasına bakarak, belgenin yazıldığı dili bulmayı sağlar. Aynı yaklaşım, cümledeki kelimelere uygulanabilir.

GAD yöntemine 2-gram ve 3-gram kelimeler, terimler olarak verilmiş ve bu yöntemin tekil kelimeleri terim olarak alan GAD yöntemi ile karşılaştırması yapılmıştır.

Yaptığımız pratik çalışma neticesinde, "n-gram destekli GAD" yönteminin beklediğimiz gibi daha seçici bir belge madenciliği yöntemi olduğunu gördük. Hem Türkçe hem de İngilizce belgelerin sorgulanması ve kümelenmesi çalışmalarında önerilen "n-gram destekli GAD yöntemi" belirgin bir şekilde GAD yöntemini geride bırakmıştır. Bu yöntem ile geleneksel sorgulama ve kümelere çalışmarından elde edilecek sonuçlardan daha yararlı ve detaylı bilgiye erişmek mümkün olmuştur.

Önerilen yöntem, sunduğu avantajın yanında, kullanım kolaylığı açısından ele alındığında bir eksiye sahiptir. N-gram kelimelerden oluşan bir kelime uzayının, sadece kelimelerden oluşan

uzaydan daha büyük olduğu ortaya çıktı. Örneklerimizdeki 615 Türkçe belge için geleneksel GAD "terim x belge" matrisinin boyutları 5678 x 615 iken, 2-gram ve 3-gram kelimelerle genişletilmiş uzayda "terim x belge" matrisinin boyutları 19075 x 615 olmuştur. Matrisin büyümesiyle, bu matris üzerinde işlem yapmak için gerekli işlemci gücü ve hafıza gereksinimleri de oldukça artmıştır.

Bir başka önemli husus, n-gram kelimeler baz alınarak oluşturulan vektör uzayında, sorgu vektörünü doğru kelimelerle oluşturmanın yanında doğru sıra ile oluşturma da önem kazanmıştır. Ancak daha büyük belge yığınlarının indekslenmesi sürecinde pek çok farklı kelime gruplaşmaları ve bunlar arasındaki örüntüler indeksleneceğinden, bu ihtiyacın azalacağını söyleyebiliriz.

"n-gram destekli GAD" yöntemi, beklenen iyileştirmeyi işlemci gücü ve hafıza bedeline karşılık sunmuştur. Yapılan örnek çalışmalar esnasında, geleneksel GAD ve "n-gram destekli GAD" yöntemlerini ayrı kullanmanın yanında birlikte kullanmanın da avantajları olabileceği saptanmıştır (Kontostathis vd., 2005). Bunun için alternatif denenebilir.

- kelimeler ve n-gram kelimelerden oluşan bir kelime uzayında işlem yapma
- önce geleneksel GAD ile sorgulama yapıp sonra gelen belgeleri 2-gram kelimelerle GAD yöntemi ile tekrar indeksleme

Bir başka gelişmeye açık husus, terim ağırlıklarının belirlenmesi aşamasındadır. Gerek geleneksel GAD, gerek ise "n-gram destekli GAD" yöntemleri, terimlere ağırlık verirken tfidf yöntemini kullanır. Bu yöntem kelimelerin istatiski kullanımlarını dikkate alır. Oysa belgelerde anlamı ortaya çıkaran kelimelerin tümü, belgeye kattığı değer bakımından sadece, ne kadar sık kullanıldığına bakılarak tespit edilemez (Wu vd., 2006). Bunun için

- fiiller ve fiillere yakın nesneler için değer hesaplama işlemine, önem derecesini artıracak katkı koyma

yöntemi denenebilir.

KAYNAKLAR

Baschab J., Piot J. (2003), Executive's Guide To Information Technology, John Wiley&Sons

Beckerman R., Allan J. (2004)., Using Bigrams in Text Categorization. Technical Report IR-408

Berkan R.C., Trubatch S.L. (2002), Fuzzy Logic and Hybrid Approaches to Web Intelligence Gathering and Information Management.

Bernstein A., Provost F., Hill S. (2002), Intelligent Assistance for the Data Mining Process:An Ontology-based Approach, MIT

Berry M. W., Tumais S. T., O'brien G. W. (1994) Using Linear Algebra For Intelligent Information Retrieval

Berry M.W., Drmac Z., Jessup E.R., (1999), Matrices, Vector Spaces and Information Retrieval

Bery M.W., Fierro R.D., (1995), Low-rank Orthogonal Decompositions for Information Retrieval Applications

Bradford R.B., (2006), Relationship Discovery in Large Text Collections Using Latent Semantic Indexing

Breivik, Patricia Senn., (1999) "Take II-Information Literacy: Revolution in Education". Reference Service Review, 27(3), 1999: 271-275.

Cannataro M., Talia D., Trunfio P., (2002), Distributed Data Mining On Grid

Caropreso M.F., Matwin S., and Sebastiani F., (2001), A learner-independent evalualiton of the usefulness of statistical phrases for automated text categorization. In Amita G. Chin, editor, Text Databases and Document Management; Theory and Practice, pages 78-102. Idea Group Publishing, Hershey, US

Chakrabarti S., (2003) – Mining The Web Discovering Knowledge from Hypertext Data Morgen Kaufmann Publisher

Chouchoulas A., (2002), A Literature Review of Variable Reduction Methodologies

Cohen, W. W.,Hirsh, H., (1998), Joins that generalize text classification using WHIRL, In Proceedings of KDD-98, 4th International Conference on Knowledge Discovery and Data Mining (169–173).

Dagan, I., Karov, Y., Roth, D., (1997), Mistakedriven learning in text categorization, In Proceedings of EMNLP-97, 2nd Conference on Empirical Methods in Natural Language Processing (Providence, RI, 1997), 55–63.

Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K., Harsman R., (1990), Indexing by Latent Semantic Analysis

Diederich J., Kindermann J. L., Leopold E., and Paass G., (2003), Authorship Attribution with Support Vector machines. Applied Intelligigence, 19(1/2):109-123

Dumais S.T., Landauer T.K, (1996), A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisiton, Introduction and Representation of Knowledge

Dunlavy D.M., (2002), An Information Retrieval System for Improving Efficieny in Scientific Literature Searches: Progress Report #1

Ekmekçioglu F. Ç., Lynch M. F., Willett P., (1996) Stemming And N-Gram Matching For Term Conflation In Turkish Texts Department of Information Studies University of Sheffield, Sheffield, UK

Faloutsos C., Gibson G., Michel T.l, Moore A., Thrun S., (1997) Data mining at CALD-CMU:Tools, Experiences and Research Directions, Carnegie Melon Universty

Fayyad U. M, Piatetsky-Shapiro G., Smyth P., (1996c), From Data Mining to Knowledge Diccovery in Databases

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

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

Fayyad U. M., Uthurusamy R., (2002), Evolving Data Mining Into Solutions For Insights, Communications of ACM

Fayyad, U. M. and Irani, K. B., (1993), Multi interval discretization of continuous attributes for classification learning, In Proceedings of I3th International Joint Conference on Artificiad Intelligence (R. Bajcsy, ed.), pp. 1022-1027, Morgan Kauffmann.

Fayyad, U. M., Piatetsky-Shapiro G., Weir N., Djorgovski S.G., (2000), Automated analysis of a large-scale sky survey: The SKI CAT System, http://techreports.ipl.nasa.gov/1993/93-0597.pdf

Fayyad, U. M., Uthurusamy, R., (1995), Discovery and data mining, Montreal, Quebec,Canada, pp. 51-76.

Fensel D., (2001), Ontologies: A silver Bullet fo Knowledge Management and Electronic Commerce, Springer

Foltz P.W., Landuer T.K., Laham D., (1998), An Introduction to Latent Semantic Analysis

Frawley, W. J., Piatetsky-Shapiro, G., Matheus, C. J., 1991, Knowledge discovery databases: An overview, In Knowledge Discovery In Databases (G. Piatetsky-Shapiro and W. J. Frawley, eds.),

Fuhr N., Buckley C., (1991), A probabilistic learning approach for document indexing. ACM

Trans. Inform. Syst. 9, 3, 223–248.

Cambridge, MA: AAAI/M1T pp. 1-27.

Grossman R.L., Hornick, M.F., Meyer, G. (2002), Data Mining Standards Initiatives

Hand D., Manila H., Smyth P., (2001), Principles of Data Mining, MIT

Holzman L. E., Fisher T. A., Galitsky L. M., Kontostathis A., Pottenger W.M., (1994), A Software Infrastructure For Research IN Textual Data Mining

Honavar V., Andorf C., Caragea D., Silvescu A., Reinoso-Castillo J., and Dobbs D., (2001). Ontology-Driven Information Extraction and Knowledge Acquisition from Heterogeneous, Distributed Biological Data Sources. In: Proceedings of the IJCAI-2001 Workshop on Knowledge Discovery from Heterogeneous, Distributed, Autonomous, Dynamic Data and Knowledge Sources.

Hornck M. F., Yoon H., Venkayala S., (2004), Java™ Data Mining (JSR-73): Status and Overview

Ioerger T. R. (1998), What is Datamining

Iwayama, M.,Tokunaga, T., (1995), Cluster-based text categorization: a comparison of category search strategies. In Proceedings of SIGIR-95,18th ACM International Conference on Research and Development in Information Retrieval (273–281).

Joachims T., (1999), Transductive inference for text classification using support vector machines. In Proceedings of ICML-99, 16th International Conference on Machine Learning (200–209).

Johnston R., (2001), An Overview of Latent Semantic Indexing

Kiryakov a., Popov B., Terziev I, Manov D., Ognyanoff D., (2004), Semantic Annotation, Indexing, and Retrieval

Kontostathis A., Pottenger W. M., (2002), Detecting Patterns in the Latent Semantic Indexing Term-Term Matrix

Kontostathis A., Pottenger W. M., (2005), Identification of Critical Values in Latent Semantic Indexing

Kontostathis A., Pottenger W. M., (2006), A Framework for Understanding Latent Semantic Indexing Performance

Koster C. H. and Seutter M., (2003), Taming Wild Phrases. In F. Sebastiani, editor, Proceeding of ECIR'03, 25th European Conference on Information Retrieval, pages 161-176, Pisa, IT, 2003. Springer Verlag

Kreyszig E., (1999), Advanced Engineering Mathematics, John Wiley & Sons

Landauer T.K., Laham D., Foltz P., (1998), Learning Human Like Knowledge By Singular Value Decomposition

Landauer T.K., Laham D., Rehder B., Schreiner M. E., (1997), How well a Passage Meaning be Derived without Using Word Order? A comparison of Latent Semantic Analysis and Humans

Larkey, L. S., Croft, W. B., (1996), Combining classifiers in text categorization. In Proceedings of SIGIR-96, 19thACMInternational Conference on Research and Development in Information Retrieval (289–297).

Lerman K., (1999), Document Clustering in Reduced Dimension Vector Space

Lewis, D. D., (1995a). Evaluating and optmizing autonomous text classification systems. In Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval (246–254).

Lewis, D. D., (1998), Naive (Bayes) at forty: The independence assumption in information retrieval, In Proceedings of ECML-98, 10th European Conference on Machine Learning (4–15).

Li, Y. H., Jain, A. K., (1998), Classification of text documents. Comput. J. 41, 8, 537–546.

Lutu P.E.N. (2002), An Integrated Approach for Scaling up Classification and Prediction Algorithms for Data Mining

Manning C. D., Schütze H. (1999), Foundations Of Statistical Natural Language Processing, MIT Pres, London

Mitchell, T.M., (1996), Machine Learning, McGrawHill, New York, NY.

Özgür A., Alpaydın E., (2004), Unsupervised Machine Learning Techniques For Text Document Clustering,

Papadimitriou C.H., Raghavan P., Tamaki H., (1997), Latent Semanticc Inexing: A Probabilistic Analysis

Pena J.M., Menasalvas E., (2001), Towards Flexibility in a Distributed data Mining Framework

Quesada J., Kintsch W., Gomez E. (2003a), A computational Theory of Complex Problem Solving Using Lantent Semantic Analysis

Quesada J., Kintsch W., Gomez E. (2003b), A computational Theory of Complex Problem Solving Using Vector Space Model : Latent Semantic Analysis, Through the Path of Thousands of Ants.

Rehder B., Schreiner M.E., Wolfe M.B.W., Laham D., Landauer T.K., Kintsch W., (1998), Using Latent Semantic Anaysis to Assess Knowledge: Some Technical Considerations

Rosario B, (2000), Latent Semantic Analysis : An Overview

Ruiz M., (2001), Combining Machine Learning and Hierarchical Structures for Text Categorization

Schütze, H., Hull, D. A., Pederson, J. O., (1995), A comparison of classifiers and document representations for the routing problem, In Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval, 229–237.

Scott, S., Matwin, S., (1999), Feature engineering for text classification. In Proceedings of

ICML-99, 16th International Conference on Machine

Suh C. Rajagopalan a., Li X., Rajan K., (2002), The Application of Principal Component Analysis to Materials Science Data

Szymkowiak A., Larsen J., Hansen L.K., (2001), Hierarchical Clustering for Datamining

Tang B., Shepherd M., Milios E., Heywood M.I., (2005), Comparing and Combaining Dimension Reduction Techniques for Efficient Tezt Clustering

Wall M.E., Rechtsteiner A., Rocha L.M., (2002), Singular Value Decomposition and Principal Component analysis

Wang Y., Hodges J., (2006), Document Clustering with Semantic Analysis

Weimer-Hasting P., (1999), How Latent Is Latent Semantic Analysis

Wolfe M.B.W., Schreiner M.E., Rehder B., Laham D., Foltz P.W., (1998), Learning from Text; Matching Readers and Texts by Latent Semantic analysis

Wu S., Tsai T., Hsu W., (2003), Text Categorization Using Automatically Acquired Domain Ontology

Yang, Y., Chute, C. G., (1994), An example-based mapping method for text categorization and retrieval, ACMTrans. Inform. Syst. 12, 3, 252–277.

Yang, Y., (1999), An evaluation of statistical approaches to text categorization, Information Retrieval 1, 1–2, 69–90.

Yans Y., Slattery S., Ghani R., (2002), A Study of Approaches to Hypertext Categorization

Yu C., Cuadrado J., Ceglowski M, Payne J. S. (2002), Patterns in Unstructured Data - Discovery, Aggregation, and Visualization

Zaiane, Antonie, (2001) Classifiying Text Documents byAssociating Terms with Text Categories

Zaki M. J., Pan Y., (2003), Introduction: Recent Developments in Paralel and Distributed Data Mining

Zukas A., Price R.J., (2006), Document Categorization Using Latent Semantic Indexing

İNTERNET KAYNAKLARI

[1] http://www.lib.ku.edu/research/-searchengines.shtml, KU libraries

[2] http://www.lib.berkeley.edu/Collections/-crisis.html

[3] http://www.crisp-dm.org, CRISP-DM

[4] http://www.ncdm.uic.edu, DSTP

[5] http://crl.nmsu.edu/Research/proj.html

[6] http://www.captain.at/review-internet.php

[7] http://www.jarg.com

[8] http://www.ak-kurt.com/dy.html

[9] http://www.hlst.sabanciuniv.edu/

[10] http://ddi.cs.itu.edu.tr

[11] https://zemberek.dev.java.net

[12] http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html

[13] http://www.tartarus.org/martin/PorterStemmer/

[14] http://faculty.cs.tamu.edu/ioerger/datamining.html

[15] http://www.omg.org/cwm/ : Common Warehouse Metamodel

[16] http://wordnet.princeton.edu

[17] http://www.rbdenter.com/Products/CasaXPS/FactorAnalysis.pdf

[18] http://www.xmla.org

[19] http://maya.cs.depaul.edu/~classes/ds575/porter.html

[20] http://www.nag.co.uk/numeric/cl/manual/pdf/F02/f02_intro.pdf

[21] http://www-4.ibm.com/software/data/iminer/fortext/

[22] http://www.alias-i.com/lingpipe/demos/tutorial/cluster/read-me.html

[23] http://www.teknoturk.org/docking/yazilar/tt000044-yazi.htm

[24] http://www.resample.com/xlminer/help/HClst/HClst_intro.htm

EK KAYNAKLAR

Abramowich W., Zurada J., (2001), Knowledge Discovery for Business Information Systems, Kluwer Academic Publishers

Baeza-Yates R., Ribeiro-Neto Berthier, (), Modern Information Retrieval, ACM Pres

Berry M.W., (2004), Survey of Text Mining Clustering, Classification and Terrieval, Springer

Berson A., Smith S., Thearling K., (2000), Building Data Mining Applications for CRM, Mc Graw Hill

Dunham M.H., (2003), Data Mining Introductory and Advanced Topics, Prentice Hall

Fabrizio S., (2002), Machine Learning in Automated Text Categorization, ACM Computing Surveys, Vol. 34, No. 1, pp. 1–47.

Fensel D., Hendler J., Lieberman H., (2003), Spinning the Semantic Web, MIT

Frakes W.B., Baeza-Yates R., (1992), Information Retrieval data Structures & Algorithms, Prentice Hall

Freitas A.A. (2002),,Data Mining and Knowledge Discovery with Evolutionary Algorithms, Springer

Jain M.C., (2001), Vector Spaces and Matrices in Physics, Alpha Science

Kantardzic M., (2003), Data Mining Concepts, Models, Methods and Algorithms, John Wiley&Sons

Kargupta H., Chan P., (2000), Advances in Distibuted and Paralel Knowledge Discovery, MIT

Kudyba S., (2004), Managing Data Mining Advice from Experts, Cybertech Publishing

Kudyba S., Hoptroff R., (2001), Data Mining and Business Intelligence: A Guide to Productivity, Idea Group Publishing

Mller T.W., (2005), Data and Text Mning A Business Applications Approach, Pearson Prentice Hall

Mohammadian M, (2004), Intelligent Agents for Data Mining and Information Retrieval, Idea Group Publishing

Roiger R.J., Geatz M.W., (2003), Data Mining A Tutorial-Based Primer, Addison Wesley

Sullivan D., (2001), Document Warehousing and Text Mining, john Wiley&Sons

Wang J., (2003), Data Mining Opprtunities and Challenges, Idea Group Publishing

Weiss S.M., Indurkhya N., Zhang T., Damerau F.J., (2005), Text Mining Predictive Methods for Analyzing Unstructured Information, Springer

Witten I. H., Frank E., (2000), Data Mining Practical Machine Learning Tools and Techniques with JAVA Implementations, Morgan Kaufmann

Witten I.H., Frank E., (2005), Data Mining Practical Machine Learning Tools and Techniques, Elsevier

EKLER

Ek 1 Türkçe Atılabilir Kelimeler Listesi

ve	non	ı	lere	unda
va	ın	i	ları	ünde
veya	lü	j	larına	sa
vah	vs	k	leri	se
veyahut	yi	ka	lerine	si
hiç	yı	m	lardan	sı
her	yu	o	lerden	so
çok	yü	ö	li	sö
herşey	te	p	lı	sü
heryer	ta	r	la lık	sn
hepsi	tı	s	lik	um
hiçbiri	ti	ş	likle	üm
tek	tü	t	lakla	ım
tük	tu	u	luk	im
derhal	n	ü	lük	om
kesin	na	v	lak	öm
da	nda	y	lek	lu
de	ne	z	yu	lü
mi	nu	zira	yi	lo
mı	nü	zaten	yı	lö
mü	ni	w	yü	muk
mu	nı	q	vb	ki
me	nö	l	ra	nın
ma	niz	le	ra	nin
mo	a	lar	rı	nun
mö	b	ler	rın	
mesi	ç	lır	ri	
ması	çünkü	lara	rin	
mak	c	larla	ru	
mek	d	lerle	run	
mik	e	ların	rü	
mık	f	lerını	rün	
mok	fakat	larına	ında	
mük	g	lerini	ını	
mök	h	lerin	inde	

Ek 2 İngilizce Atılabilir Kelimeler Listesi

A	because	Each	how	much
abaft	been	early	howbeit	must
aboard	before	either	however	mustn't
about	behind	em	how's	my
above	being	english	i	myself
across	below	enough	id	n
afore	beneath	ere	if	near
aforesaid	beside	even	ill	'neath
after	besides	ever	i'm	need
again	best	every	immediately	needed
against	better	everybody	important	needing
agin	between	everyone	in	needn't
ago	betwixt	everything	inside	needs
aint	beyond	except	instantly	neither
albeit	both	excepting	into	never
all	but	f	is	nevertheless
almost	by	failing	isn't	new
alone	c	far	it	next
along	can	few	it'll	nigh
alongside	cannot	first	it's	nigher
already	can't	five	its	nighest
also	certain	following	itself	nisi
although	circa	for	i've	no
always	close	four	j	no-one
am	concerning	from	just	nobody
american	considering	g	k	none
amid	cos	gonna	l	nor
amidst	could	gotta	large	not
among	couldn't	h	last	nothing
amongst	couldst	had	later	notwithstanding
an	d	hadn't	least	now
and	dare	hard	left	o
anent	dared	has	less	o'er
another	daren't	hasn't	lest	of
any	dares	hast	let's	off
anybody	daring	hath	like	often
anyone	despite	have	likewise	on
anything	did	haven't	little	once
are	didn't	having	living	one
aren't	different	he	long	oneself
around	directly	he'd	m	only
as	do	he'll	many	onto
aslant	does	her	may	open
astride	doesn't	here	mayn't	or
at	doing	here's	me	other
athwart	done	hers	mid	otherwise
away	don't	herself	midst	ought
b	dost	he's	might	oughtn't
back	doth	high	mightn't	our
bar	down	him	mine	ours
barring	during	himself	minus	ourselves
be	durst	his	more	out
	e	home	most	outside

devamı

over	sometimes	true	which	
own	soon	'twas	whichever	
p	special	'tween	whichsoever	
past	still	'twere	while	
pending	such	'twill	whilst	
per	summat	'twixt	who	
perhaps	supposing	two	who'd	
plus	sure	'twould	whoever	
possible	t	u	whole	
present	than	under	who'll	
probably	that	underneath	whom	
provided	that'd	unless	whore	
providing	that'll	unlike	who's	
public	that's	until	whose	
q	the	unto	whoso	
qua	thee	up	whosoever	
quite	their	upon	will	
r	theirs	us	with	
rather	their's	used	within	
re	them	usually	without	
real	themselves	v	wont	
really	then	versus	would	
respecting	there	very	wouldn't	
right	there's	via	wouldst	
round	these	vice	x	
s	they	vis-a-vis	y	
same	they'd	w	ye	
sans	they'll	wanna	yet	
save	they're	wanting	you	
saving	they've	was	you'd	
second	thine	wasn't	you'll	
several	this	way	your	
shall	tho	we	you're	
shalt	those	we'd	yours	
shan't	thou	well	yourself	
she	though	were	yourselves	
shed	three	weren't	you've	
shell	thro'	wert	z	
she's	through	we've		
short	throughout	what		
should	thru	whatever		
shouldn't	thyself	what'll		
since	till	what's		
six	to	when		
small	today	whencesoever		
so	together	whenever		
some	too	when's		
somebody	touching	whereas		
someone	toward	where's		
something	towards	whether		

Ek 3 Sınıf Tanımları 2

İnsan Kaynakları

İnsan Kaynakları Yönetimi Nedir? Kavramsal İçeriği İnsan Kaynakları Yönetimi (İKY), işletmelerin hedeflerine ulaşabilmeleri için gerekli olan işlevleri gerçekleştirecek yeterli sayıda vasıflı elemanın işe alınması, eğitilmesi, geliştirilmesi, motive edilmesi ve değerlendirilmesi işlemidir. İnsana odaklanmış, çalışanların ilişkilerini idarî bir yapı içinde ele alan, kurum kültürüne uygun ayarlanmış politikalarını geliştiren ve bu yönüyle kurum yönetiminde kilit işlev görevi gören bir yönetim anlayışıdır.İKY, bir işletmede işgücünü etkin bir şekilde oluşturmak, geliştirmek ve devam ettirmek için, kanunî çerçevede ve çevre şartlarına da uygun bir biçimde ortaya konan faaliyetlerin bütünüdür. İKY, bir organizasyonun, vizyon ve misyon doğrultusunda, ihtiyaç duyduğu işgücünü en optimal bir biçimde meydana getirmek, motive etmek, geliştirmek, ödüllendirmek ve devamlılığını sağlamak için ortaya konulan plân, program ve stratejilerin uygulanmasıdır.İşletme içindeki çalışanlarla ilgili program, yöntem, yönetmelik ve süreçleri geliştirme, uygulama ve değerlendirmeyle ilgili bir alan olan İKY, malî ve maddî kaynaklara ek olarak, insan kaynağının da doğru yönetilmesi ile uğraşan bir disiplindir. İKY'nin İşlevleri İnsan Kaynakları Politikasının Tespiti. Personel Organizasyonu. İnsan Kaynakları Plânlanması. Performans Yönetimi. Ücret Yönetimi. Eğitim Yönetimi. Motivasyon Yönetimi. Kalite Yönetimi. Bilgi Yönetimi. Vizyon Yönetimi. İKY'nin Hedefleri Ahlâkî ve sosyal sorumluluk anlayışıyla, çalışanların kapasitelerinden gereği gibi yararlanmak ve onların örgüte-işyerine-organizasyona olan etkili katkılarını artırmak. Emek verimliliğini (insan kaynağının verimli kullanımını) ve işgücü performansını, motivasyon ve teşvik programları sâyesinde artırmak. Çalışma hayatının kalitesini artırmak ve çalışmayı, cazip hâle getirmek (işgören tatminini ve işçi sağlığını yükseltmek). Örgüt verimliliğini, örgüt ve personel bütünleştirme gayretlerinin yanında personel-örgüt ve sanayi psikolojisi yöntemleriyle arttırmak. Ekip çalışmasını ve toplam kaliteyi özendirecek çalışmalar yapmak. Uygun örgüt kültürünün temellerini atmak. İKY'nin İşletmeye Getireceği Faydaları İşçilik maliyeti azalır. İşgücü devri oranı azalır. İşe devamsızlık ve temarüz azalır. İş kazaları azalır. Hatalı üretim azalır. Ürün kalitesi artar. Moral ve motivasyon artar. İşyeri atmosferi iyileşir. Çalışma barışı sağlanır. İşyeri tatmini artar. İşletmenin rekabet edebilirliği artar. İnsan Kaynakları Yönetimi (İKY) nin İlgi Alanları İKY'nin Kavramsal İçeriği ve Tarihî Gelişimi İKY ve Personel Yönetiminin Karşılaştırılması İKY'nin Sistemi, Politikası ve Örgütlenmesi İKY'nin Görev ve Hedefleri İş Analizi ve İş Tanımlamaları ve Gerekleri İnsan Kaynakları Plânlaması İnsan Kaynakları Temini ve Seçimi Psikoteknik: Psikolojik ve Beceri Testleri Meslekî Kariyer İKY'de Eğitim ve Geliştirme Teknikleri Performans Değerlemesi İş Değerlemesi Örgütsel Davranış Liderlik Motivasyon Etkin İletişim Sendika-Yönetim Münasebetleri İşyerinde Katılımcı Yönetim Ücretlendirme ve Sosyal Sorumluluk

Kurumsal Kaynak Planlama

Enterprise Resource Planining (ERP-Kurumsal Kaynak Planlama), şirketlere yönelik bir yazılım programı. Kullanımı hızla artan bu yazılım, şirket içindeki tüm üretim ve yönetim sürecini planlıyor. Üstelik planın işleyişini kontrol ediyor, çalışmasını da sağlıyor. Türkiye'de SAP, Oracle ve Baan tarafından sağlanan bu yazılımı kullananların sayısı hızlı artıyor.global

ekonominin kuralları bir şirketin veya bir ülkenin rekabetçi pazar koşullarında ayakta kalabilmesi için, ı teknolojik yenilikleri takip etmesini zorunlu kılıyor. Gelişen ve değişen ekonomik ortama ayak uydurmak isteyen firmalar, değişim mühendisliği ve toplam kalite yönetimi gibi pek çok yönetim stratejisini uygulamaya koyuyor. Ancak bu stratejilerin, rekabetçi pazar ortamında, etkili bir teknolojik alt yapı olmadan tek başına başarıya ulaşması mümkün değil. Başarıya ulaşabilmek için daha az insanla, daha çok veriyi işleyebilmek ve yönetebilmek gerekiyor. Bunun tek çözümü ise etkin bir teknolojiyi şirkete adapte etmek.Günümüzde "anahtar" konuma gele bu yeni teknolojik sistemlere Enterprise Resource Planning (ERP-Kurumsal Kaynak Planlama) sistemleri adı veriliyor. Capital, şirketlerde neredeyse bir zorunluluk haline gelen ERP sistemlerinin nasıl uygulandığını ve neler sağladığını araştırdı.Nasıl bir sistem? ERP sistemleri, bir şirket içindeki tüm üretim ve yönetim fonksiyonlarını planlayabiliyor, kontrol edebiliyor ve çalıştırabiliyor. Sistem üretim, satış ve pazarlama, finans, insan kaynakları, stok ve depo yönetimi ve lojistik gibi, birçok fonksiyonu aynı çatı altında toplayabilen, bütünleşik ve geniş kapsamlı bir yapıya sahip. Bu sistemlerin öncesinde pazara sürülen ilk jenerasyon sistemler ise MRP- (Ma-nufacturing Resources Planning) ve MRP II olarak adlandırılıyordu.ERP yazılımlarını üreten ve satan firmalar, sektörlere özel EKİ çözümleri de sunuyor. ERP pazarında, otomotiv, gıda, tüketim mallan, ilaç ve kimya, ulaşım, elektrik-elektronik, petrol, tekstil mağazacılık, sağlık, savunma, sigorta, bankacılık ve kamu sektörlerine yönelik çözümler bulunuyor.Sistemin getirdikleri Peki ERP sistemini kullanan bir finansman neler kazanıyor? Bu sorunun yanıtını Oracle Uygulama Yazılımları Müdürü İbrahim Göğüs şöyle veriyor: "İşletim verimliliği ve ürün kalitesi artıyor. Hizmet ve ürün kalitesinde yeniliklere olanak sağlı yor. Satış gücünüzü güçlendiriyor, internet modülü sayesinde yeni pazarlara acıtabiliyorsunuz. Özellikle çokuluslu şirketlerde ve dış ticaret yapanlarda giderek karmaşıklaşan finansal operasyonlar, birden daha fazla döviz kuruyla çalışabilen muhasebe sistemleri ile yönetilebiliyor. Bütçeleme ve çok boyutlu finansal analizlere olanak sağlıyor. Finansal verilerinizi konsolide edebiliyor ve daha pek çok finansal yönetim işlevini yerine getirebiliyor. Tedarik zinciri modülleri ile müşterilerden depolara ya ela dağıtım merkezlerine, fabrikadan tedarikçilere kaçlar olan tüm süreçlerde etkin bir iletişim sağlanarak verimlilik artışı sağlıyor." SAP Pazarlama ve İş Ortaklan Grup Yöneticisi Arzu Gençoğlu ise bu konuda şöyle bir örnek veriyor: "Havacılık ve savunma sanayii konusunda çalışan bir müşterimiz şu sonuçları elde etti; teklif ile tahsilat arasında geçen süre 270 günden 65-80 gün arasına indi. Sipariş giriş zamanı 6 günden l güne çekildi. Hatalı faturaların oranı önemli düzeyde azalarak, yüzde 18'den yüzde 10'a düştü." ERP rekabeti kızışıyor Dünyada ERP pazarının ulaştığı büyüklük 1.5 milyar dolar ve pazar her geçen yıldızla büyüyor. Şu an pazar lideri olarak SAP gösteriliyor. Bu alandaki diğer iki şirket ise Haan ve Oracle. Pazardaki hızlı gelişme nedeniyle, daha önce Digital ile birlikte çalışmış olan QAD firması ise kendi ofisini açma çalışmalarını yürütüyor. Interpro' nun verilerine göre Türkiye'de yazılım pazarının toplam büyüklüğü, 1997 yıl sonu itibariyle 140 milyon dolar düzeyinde. Bunun içinde ERP türü yazılımlarının payı ise 15 ile 20 milyon dolar arasında değişiyor.Hedef KOBİ'ler... Donanım fiyatlarının düşmesiyle birlikte KRP pazarı bütün dünyada hızlı bir büyüme sürecine gireli. Böylece ERF yazılımlarının fiyatları da orta ölçekli pek çok firmanın erişebileceği düzeye indi. Baan Pazarlama Direktörü Hanım Doğan bu konuya dikkat çekerek şöyle diyor: "Türkiye'de son 3-4 yıldır büyük firmalar yanında yurt dışına açılan, sıkı rekabet ortamına ayak uydurmak isteyen küçük ve orta ölçekli firmalar da ERP çözümlerine yönelmeye başladılar. Orta ölçekli firmaların da çok ciddi birtakım raporlara, maliyet analizlerine, kârlılık analizlerine ve ürün kalitesinde belli standartları tutturmaya ihtiyaçları var. Bu ihtiyaçlarını karşılamak için birbirinden ayrı, adacıklar halinde sistemler yerine bütün sistemliyoruz. Örneğin bir Kayseri iline çok önem veriyoruz." Anadolu'dan büyük talep SAP Pazarlama ve İş Ortaklan Grup

Yöneticisi Arzu Gençoğlu Anadolu Kaplanı olan illere daha rahat ulaşmak için o bölgelerdeyerleşik"işortağı"olarakadlandırdıklarıfirmalarileçalıştıklarını leşik tercih ediyorlar. Biz KOBİ'lerin yoğun olduğu Anadolu kaplanı denilen illere ulaşmayı hedefliyoruz. belirtiyor ve şöyle devam ediyor: "Denizli'de bir iş ortağımız bulunuyor. Adana'da ise bir firma ile görüşmelerimizi sürdürüyoruz. Kayseri, Konya ve Eskişehir illerine ise Ankara kökenli bir firma aracılığıyla hizmet veriyoruz" eliyor. Oracle Uygulama Yazılımları Müdürü İbrahim Göğüs ise KO-Bİ 'lere önem verdiklerini vurguluyor ve şöyle devam ediyor: "KOBİ' ler rekabet avantajı yaratabilmek için bilgi üretebilmeli ve tüketebilmeli... Bu anarak ERP yazılımlarını kullanarak mümkün." ERP YAZILIMLARI NEREDE KULLANILIYOR?Bir erp yazılımının ne gibi yeteneklere sahip olduğunu tek tek saymak ve bu sayfalara sığdırmak pek mümkün değil .. çünkü her konuya özel, farklı bir uygulama tipi ortaya çıkıyor. Biz burada örnrk olarak insan kaynakların da kullanılan erp yazılımının ne gibi olanaklar sağladığını özetlemeye çalıştık. Erp insan kaynakları grubu, insan kaynaklarınızı yönetmeniz için gereken personel yönetim görevlerinin tümünü kapsıyor. Sistemi kullanarak, şunları yapabiliyorsunuz: insan kaynakları ana verileri toplama, Personel yönetimi, İşe alma ve yerleştirme,Seyahat yönetimi ve planlaması Ücret ve performans yönetimi, Personel maliyet planlaması, Bordro hesaplamaları, " Planlama senaryoları, Eğitim ve olay yönetimi, Zaman yönetimi, Qut sourcing ve tazminat yönetimi .Arzu Gençoğlu, Anadolu kaplanı olarak nitelendirilen illere daha iyi hizmet vermeyi hedeflediklerini belirtiyor. ÇÖZÜM ORTAĞI OLMAK İÇİN NELER GEREKİYOR?Baan, SAP ve Oracle gibi dünyanın büyük yazılım firmaları özellikle sanayinin geliştiği Anadolu illerinde ERP konusunda "çözüm ortağı" olarak çalışabilecek firmalar arıyorlar. Sistek firmasının Genel Müdürü Cem Kobaner, çözüm ortağı olabilmek için gereken özelliklen şöyle anlatıyor:"Biz 1992 yılında bu işe bağımsız bir yazılım firması olarak başladığımızda bütün şirket ortakları olarak ortaya sermaye değil bilgi ve tecrübemizi koyduk. Önceleri Türkiye'nin İlk 500 büyük şirketine özel, butik tarzı diyebileceğimiz nitelikte yazılımlar geliştirerek büyüdük.Müşterilerle ilişkilerimiz uzun vadeli... Kaliteden ödün vermiyoruz. Referanslarımız çok güçlüydü. Beko, Otosan, Arçelik, Otoyol, Paşa bahçe... Almanya'da Schott diye bir cam firması,.. SAP da bizi araştırmış, iş anlayışımızı, ilkelerimizi kendi yapısına uygun bulmuş ve geçen ay iş ortaklığı teklif ettiler, anlaşmamızı imzaladık. Oracle ile tanışmamız şöyle oldu: Biz Tofaş için pek çok proje gerçekleştirmiştik. Oracle, Tofaş ile çalışmaya başlayınca bizim ismimizi duymuşlar, Veri tabanı uygulamaları konusunda birlikte çalışmak için teklif getirdiler, kabul ettik, 5 yıldır birlikte çalışıyoruz...

Müşteri İlişkileri Yönetimi

Müşteri İlişkileri Yönetimi (CRM)Günümüzde CRM'i olmayan, müşteri sadakatini garantilemeyen bir markanın pazar payını uzun vadede koruması ve artırması çok zor. Geleneksel pazarlamacı; ürünleri için mümkün olan daha fazla müşteriyi bulmayı amaçlarken, CRM odaklı bir şirket mevcut müşterileri için daha fazla ürün ve hizmet bulmayı amaçlar. CRM nedir? Yeni bir müşteri kazanmak, mevcut müşterilerinize satış yapmaktan 6 kat daha maliyetlidir. Müşteri ilişkileri Yönetimi (Customer Relationship Management) veya kısa adıyla CRM, mevcut ve görece sadık müşterinizle birebir ilgilenerek onların cebinden alacağınız payı yükseltmeyi hedefler.CRM'e, müşteri ilişkilerini teknolojik bir altyapıyla ölçülebilir biçimde yönetme şekli de denebilir. CRM, kapsamlı bir müşteri perspektifiyle üretim birimlerinin, çalışanların ve iş süreçlerinin entegrasyonunu içerir. İlişkilerin yönetimini gerektirir ama işlemlerin yönetimini gerektirmez. ERP, SCM gibi işlemlerin yönetiminde kullanılan iş

modelleri ile entegre edilebilir. CRM teknolojik bir uygulamadır ama net bir CRM stratejisi belirlemeden, süreçleri önceden planlamadan işe başlanırsa sonuç başarısız olabilir.Siemens Business Services Türkiye, bu noktada işe danışmanlık hizmetiyle başlar. Mevcut durumu inceler, CRM stratejinizi en etkin şekilde kurmanız için size yol gösterir. Dünyaca kabul gören teknoloji uygulamalarıyla, esnek ve zaman içinde geliştirilebilir bir altyapı kurar. Siemens Business Services, önce şirketinizdeki mevcut durumu inceler. Şirketinizin CRM'e hazırlık düzeyi belirlenir. Pazarlama, satış ve servis bölümlerindeki geliştirmeye açık alanlarının belirlenmesi amacıyla ekiplerinizle görüşmeler yapılır. İş süreçleri incelenir, problem yaratan alanlar belirlenir. Müşteri bilgileriniz incelenir, problemler tespit edilir. Müşterilerinizin düzenli eğilimleri incelenir. …nedenleri belirler. Satış, SWOT, ABC ürün ve müşteri analizleri gerçekleştirilir. Müşteri segmentasyon kriterleri belirlenir. Müşteri profili ve veri sözlüğü çıkarılır. Strateji belirlenir. Hedefler ve strateji bütünleştirilir. …çözümler geliştirir.Anahtar Performans Göstergeleri (AGP) için Best Practice değerleri saptanır. Çözümler için gerekli zaman riske göre önceliklendirilir. Kaba maliyetler çıkarılır. Öncekli çözümlerin ROI ve EVA hesaplamaları yapılır.…altyapıyı kurar.

Bilgi Yönetimi

Bilgi Yönetimi Nedir Kurumun sahip olduğu bilgi kullanıma sunulduğunda, performansta o güne değin beklenmedik sonuçlar yaratır. BİLGİNİN YÖNETİLMESİ İÇİN cep telefonları, bilgisayarlar, internet ya da diğer gelişmiş teknolojilere gerek yoktur. Ne var ki, bu teknolojilerin varlığı, bilgi yönetiminde günümüze kadar hayal bile edilememiş olasılıkları birer gerçeğe dönüştürmektedir. Bugün bilgi yönetimi üzerine düşündüğümüzde bir çoğumuzun aklına gelişmiş intranet ve internet uygulamaları gelmektedir. Bunun bir sonucu olarak da bilgi yönetimi tanımlan kesin bir dille değil de imâ yolu ile olsa bile teknolojiden söz etmektedir. Bilgi yönetimi pek Çok şekilde tanımlanmıştır. Tanımlar arasında her ne kadar ortak yönler varsa da her biri farklı bir unsuru vurgulamaktadır. Aşağıda bu tanımlar için bazı örnekler verilmiştir. Bilgi yönetimi; "ister somut biçimlerde ifade edilen açık bilgi, ister bireyler ya da toplulukların sahip olduğu saklı bilgi biçiminde olsun, entelektüel varlıkların belirlenmesi, optimizasyonu ve aktif bir biçimde yönetilmesidir" Bilgi yönetimi; "duyarlılık düzeyinin ve yenilikçiliğin artırılması amacı ile kolektif aklın geliştirilmesidir." Bilgi yönetimi "bir işletmenin enformasyon varlıklarının belirlenmesi, yakalanması, erişilebilir durumda tutulması, paylaşılması ve değerlendirilmesinde bütünleşik bir yaklaşım getiren disiplindir. Bu enformasyon varlıktan arasında veri tabanları, belgeler, politikalar ve prosedürlerin yum sıra yakalanmamış, bireylerin zihninde kalmış saklı uzmanlıklar ve deneyimler de bulunur." Bilgi yönetimi; "organizasyonel performansın yükseltilmesi ve değer yaratılması amacıyla bilginin üretilmesi, sürdürülmesi, kullanılması, paylaşılması, yenilenmesine ilişkin süreçlerin uygulanması anlamına gelir." Bilgi yönetimi; "verimlilik, yenilik ve daha hızlı, daha etkili kararlar aracılığı ile rekabet avantajı ve müşteri sadakati kazanmak amacı ile entelektüel sermayenin ortaya çıkarılması ve kullanılması uygulamasıdır. "(Barth. 2000). Bu tanımların tümünde başlıca iki temanın varlığı görülmektedir. Bunlardan bir tanesi yararlı bilginin elde edilmesi ve yayılmasıyla ilgilidir. Kuruluşlar gerek şirket içinde, gerekse şirket dışında işe yarayabilecek türden bilgiyi çeşitli kaynaklarda aramak ve bulduklarında da almak; sonra da öğrenilenleri o enformasyonu kullanabilecek çalışanlarına (ve diğerlerine) yaymak zorundadır. İkinci bir tema, bilginin rekabet avantajı yaratılması yolunda uygulamaya geçirilmesidir. Bu ikinci tema bilginin, yeni ürün ve hizmetler yaratmak, müşteri hizmetini daha verimli hale getirmek ve yeni değer yaratma kaynakları ortaya çıkaracak türden beceriler geliştirmek amacıyla kullanılmasına ve paylaşılmasına odaklanır.

Aile Şirketleri

Aile Şirketleri ve Yönetim Aile şirketi sahipleri, kendileri işin başında olduğu, çok çalıştıkları, hızlı karar verdikleri ve işe fazla asıldıkları için kısa zamanda önemli karlar elde edebilmektedirler. Ancak iş belli bir büyüklüğe geldiği ve işin kendisinin yanında yönetim ve organizasyon, insan kaynakları, verimlilik, kalite, maliyet muhasebesi, insan ilişkileri gibi - aile şirketinin sahibinin alışık olmadığı hatta lüks gibi gördüğü - kavramlar işin içine girdiğinde sıkıntılar da başlar. İşi bilmek, işin çekirdeğinden gelmek önemlidir.Ama kalıcılık için yeterli değildir. Aile şirketi yöneticilerini kapsayan bir araştırmaya göre aile şirketi sahiplerinin % 85'inin dile getirdikleri sorunlar şunlardır Kurumsallaşamama En eski şirketin Hacı Bekir olduğu Türkiye'de aile şirketlerinin çoğu kurum kimliği kazanamamıştır. Gerek ülkemizde ve gerekse dünyada aile şirketlerinin üçüncü kuşağa ulaşma oranı % 15-20, ömürleri ise 25-30 yıldır.Aile şirketlerinin başarısızlık nedenleri arasında yönetimde yetersizlik ve kurumsallaşamama ilk sırada yer almaktadır. Bir aile şirketinin en zayıf noktası, aile ve şirket kavramlarının birbirine karıştırılmasıdır. Şirkette yetenek ve performansın yerine kan bağının ön plana çıkmasına nipotizm adı verilmektedir. Yetenek ve deneyimlerine bakılmaksızın çocuklar işe alınıp hızla yükseltilmekte ve performans değerlendirilmeden ömür boyu iş olanağı veril- mektedir. Hatta bazen şirket içinde sadece onlara özgü konumlar yaratılmaktadır. Bu örnekler aşırılaştığında şirket ailenin oyun bahçesi görünümüne bürünebilmektedir.İkinci nesil - sermaye sahibinin kan bağı yakını olması sebebi ile - kendisini yılların profesyonel yöneticilerinin tecrübe ve birikimlerine rağmen üstün görmeye kalktığı anda felaket başlar.Sadece oğul olmak hiçbir değerli birikimin, tecrübenin önünde yer almaya yetmez. Aile şirketlerinin - genellikle rahat yetişmeleri ve işin çekirdeğinden gelmemeleri nedeni ile - ikinci kuşağın elinde geriye doğru gittiği bilinmektedir.İkinci kuşak işe tepeden gelmemelidir. Bu tür bir tepeden inme ham ağaca gerekli işlemler yapılmadan vernik sürmeye benzer. Vernik tırnakla kazındığında ağacın işlenmemiş olduğu hemen belli olacaktır. Kurucu patronun kendisine güven duyması ve çalışanlarını motive edebilmesinin yanı sıra her şeyi denetim altında tutmak isteyen otokratik bir yapısı vardır. Kimi zaman, kırtasiye alımından önemli yatırımlara kadar her türlü kararı elinde tutan "one man show" tipi bir yönetim sergilemektedir. Oysa şirketin büyümesi, nitelikli personel istihdamını ve sistemli kontrol mekanizmalarını gerektirir. Patronun "tek adam" rolünü rafa kaldırıp "orkestra şefliği"ne yönelmesi şart. Patron şirket karlılığının gözeticisidir. Patronluk ayrı bir meslektir, öğrenilmesi gerekir ve yürütücü yönetimle çakışmamalı, onun bir tamamlayıcısı olmalıdır. Kurumsallaşma patronların işi bırakması değildir. Patronun görevi tepe yöneticisini atama, kurumu direk ve dolaylı olarak denetlemektir. Bazen de şirket kurumsallaşır ama aile ilişkileri kurumsallaşamaz. Önemli olan aile ilişkilerinde de sistematik hale gelmektir. Kurumsallaşma bir felsefe ve inanç meselesidir. Başlangıçta aile olarak özveride bulunmak gerekecektir. Çünkü kurumsallaşma tek adam yönetimine zıt, tamamen yönetim bilgisine dayalı olarak yürütülmesi gereken profesyonelce bir sistemdir. Saltanat. Sermaye sahibi aynı zamanda yönetici olduğu zaman sistemi denetleyen organ noktasında zafiyet doğuyor. Şirket asla saltanat zihniyetine teslim edilmemelidir. En güçlü saltanat rejimleri nasıl ki bir gün gelip yok olabiliyorsa; iç yönetim rekabetine açılmayan saltanatlaşmış şirketlerin de aynı olumsuz akıbetten kurtulmaları mümkün olmayacaktır. Profesyonelleşememek ve Yüksek İşgücü Devir Oranı. Aile şirketlerinde genel olarak insana yatırım yapılmamaktadır. Organizasyon şeması, görev tanımları ve yetki ve sorumluluk dengesi genel olarak yoktur.İş gücü devir oranı yüksektir.Çalışanlar ve sahipler arasında "biz" ve "onlar" ayrımı vardır. Kurum sahipleri genellikle kendilerini daha deneyimli, bilgili, zeki ve işi daha fazla biliyor gördüklerinden kendilerini aşan yöneticilerle çalışmaktan zorlanırlar. Raporlara,

eğitime kısaca işin tüm bileşenlerine yeterince duyarlı olmayıp sadece üretim ve satıştan zevk aldıklarından sistemin tamamını görmekte zorlanırlar. Bu zorluğu aşarak rapor üzerinden hakimiyet kuramadıklarından da kendilerince kilit birimlere aileden birilerini koyma gereği duyarlar. İşin Çekirdeğinden Gelme Eski Alışkanlıkların Devamı. Aile şirketlerinde firma kültürünü, aileyi oluşturan bireylerin duygu ve düşünceleri ile kültürlerinin bir yansıması oluşturur. Genel olarak firmanın vizyon ve misyonunun iş sahibi ailenin kültürünün gölgesinde kaldığı görülmektedir.Güç Kavgası. Özellikle büyük şirketlerde patron adaylarının etrafında çıkar çevreleri oluşabilir. Özellikle aile içi geçimsizliklerin işe taşınması bu kümelenmeleri artırır. Bu çevreler çoğu kez belki bilinç altında ileride maddi ve manevi çıkarlar sağlamak üzere patron adaylarının olmayan yeteneklerinin bulunduğuna bu kişileri inandırmaya çalışırlar.Şirket içinde mevcut profesyonel yönetim, gelecekte güç sahibi olmak için sinsi bir şekilde böl ve yönet politikasını gerçekleştirmek üzere bu yönde bir güç kavgasını körükleyebilir. Bu tür entrikaların olmayacağını düşünme saflığını gösterme yerine, çok geç kalmadan önlem alma sürecine girilmelidir.Türk toplumunun lider bağımlılığı aile şirketlerinde de bir lider ihtiyacını adeta mecbur bırakmaktadır. Aile şirketlerinde yönetim diğer herhangi bir kurumdaki yönetimden ciddi farklılıklar göstermektedir. Her şeyden önce aile şirketleri sahipleri birer yönetici olarak yeterliliklerini, eksikliklerini tanımalıdırlar. Aile şirketlerinde patronların her birinin ayrıca bir patronluk yapması, birbirleri ile yarışması, görüş ayrılıklarının kişisel çatışmaya dönüşmesi, gereksiz tartışmalar, işler iyi olsa da kötüye gidişin göstergeleridir. Ne olursa olsun lider belli olmalıdır. Liderin öncelikle aile içi dengeyi sağlaması, kararların ortak biçimde alınmasını sağlaması, aile içi adaleti sağlaması önemlidir. Çoğu zaman yeri doldurulamayan aile liderinin yokluğu şirket hezimetlerinin en önemli nedeni olmaktadır. Bu nedenle gerektiğinde patron, liderliği profesyonellere devredebilmeli ve kurumsallaşmasını sağlayabilmelidir. Aile şirketlerinin gelecekte başarısı için yeter ki Aile bireyleri, şirket ve ailenin birbirinden farklı olduğunun ayrımına varsın ve aradaki dengeyi kurabilsin.Servetin yönetimi ile şirketin yönetiminin birbirinden ayrı kavramlar olduğuna inanılsın.Şirket, disiplin ve organizasyonun var olduğu profesyonel bir kimlik kazansın. Yönetim üstün nitelikli profesyonelleri işe alıp motive edebilsin. Onlara görevlerini başarı ile yerine getirebilmeleri için özgürlük alanları tanısın. Aile bireyleri de onlar "dışarıdan gelenler" olarak değil, takımın bir üyesi olarak benimsesin. Liderlik kurumsallaştırılmaya çalışılsın.Aile ve şirket içinde iletişim açık, iki taraflı ve yoğun olarak gerçekleştirilsin.Ailenin şirket içindeki rolü ve yönetim biçimi açıkça belirlensin. Aile içi geçimsizlikler işe taşınmasın.Bağımsız olarak çalışan ve üst yönetime destek veren bir yönetim kurulu oluşturulsun.Kısa vadeli günlük plan ve kazanç mantığından uzun vadeli planlama ve kazanç mantığına dönülsün.Yönetimin devredilmesi çalışmalarına bir an önce ağırlık verilsin.Patronluk kendi asli rolüne dönsün.Şirket sahibi patron, aile bireyleri yönetimde yetersiz kaldığında anında teşhis edip, gerekli önlemleri alsın. Gerektiğinde liderliği profesyonellere devredebilsin.Bütün bu koşulları yerine getirebilen bir aile şirketi kurumsal bir nitelik kazanacak ve böylece sürekliliğini sağlayabilecektir.

Turizm

Zengin yayıncılık birikimi ve deneyimi olan seçkin bir kadroyla İnternet ve Dergi yayıncılığıyla şirketimiz sizlerden aldığı güçle iletişim ve bilişim teknolojisinin hızla geliştiği, Dünya'nın yoğun bir değişim ve dönüşüm yaşadığı çağımızda Türk Turizmi'nin iç ve dış tanıtımını en düzeyli ve yaygın bir biçimde yapma uğraşı veriyor. Şirketimiz, Kobilere yönelik çalışmaları ve diğer faaliyetlerimizle de hizmetinizde olmaktan mutluluk duyar, sizlerin desteği, eleştiri ve önerisiyle daha iyiyi sunmak dileğiyle… Alışveriş Merkezleri, Araç Kiralama (Rent A Car), Büyük Elçilik ve Konsolosluklar, Cafe - Bar ve Eğlence Merkezleri, Firma ve Sektör Haberleri,

Eczane ve Hastaneler, Hava Durumu, İnşaat Araç-Gereçleri, Konaklama Merkezleri (Hotel-Motel), Mobilya Araç-Gereçleri, Mutfak Araç-Gereçleri, Odalar ve Dernekler, Restaurantlar, Lokantalar, Seyahat Acenteleri, Tarihi Yerler, Tatil Merkezleri, Turizm Araç-Gereçleri, Turizm Sektörüne Ücretsiz Hizmet, Turizm Bakanlığı Danışma Ofisleri, Yararlı Linkler, Yurtiçi ve Yurtdışı Fuarlar, Ulaşım Yolları (Hava-Kara-Deniz) tourism turkey tourism tourism in turkey tourism turkey world tourism organization tourism statistics tourism agency sustainable tourism turkish tourism tourism management world tourism organisation eco tourism tourist visit touristic turism visiting travel orbits travel agency travel agencies travel agent travel agents vacation trips cruises vacations geography locations impacts ask jeeves sightseeing visitors bahamas ecotourism central asia agriculture black sea turkey outbound tourists tourismus historical inbound manali religion kazakhstan pakistan algeria greenland regions currency bulgaria population info club med reservations climate syria agent geographical yemen detailed capital sudan cyprus history albania armenia belarus cia factbook russia uzbekistan atlas hyatt coat of arms iran time zones hungary malta nicaragua mauritius senegal guatemala angola macedonia time zone national geographic maps saudi arabia tibet denmark norway economy geology yugoslavia uganda ukraine zimbabwe latvia travel guide greece indonesia maldives japan middle east slovakia poland topography country bosnia lonely planet romania agency newspapers asia scotland europe slovenia geographic mongolia cia sweden map taiwan afghanistan lithuania south america east timor culture siberia switzerland tunisia finland portugal israel political bahrain kosovo estonia cia world factbook iceland usa austria zip codes nepal showing vegetation hawaii physical country code china iraq kuwait hotels cairns costa rica guam oman flags mexico outline flag bhutan seychelles kenya lebanon world fact book comfort best western mapquest world factbook croatia cities rivers dominican republic radisson cameroon beaches ethiopia fairfield laos mountains guyana topographic factbook great britain cuba thailand cambodia honeymoon tanzania weather tahiti ivory coast south africa inn populations cia world fact book new zealand turizm turizm bakanlığı metro turizm ets turizm boss turizm duru turizm nilüfer turizm turizm şirketleri pamukkale turizm varan turizm trek turizm turizm acentaları turizm haftası ulusoy turizm turizm bakanligi turizm gazetesi nilufer turizm turizm seyahat uludağ turizm turizm otelcilik turizm rehberi turizm acenteleri turizm bakanlıgı türkiye turizm hidayet turizm didim turizm turizm acentası kent turizm has turizm asya turizm türkiyede turizm turizm firmaları turizm acenta turizm bakanliği vip turizm turizm ve otelcilik koç turizm antalya turizm istanbul turizm turizm istatistikleri kamil koç turizm kültür ve turizm viking turizm turizm sektörü turizm meslek lisesi turizm sirketleri anı turizm fest turizm turizm acentalari turizm iş turizm il müdürlüğü accord turizm turizm gelirleri ipek turizm truva turizm turizm eğitimi turizm şirketi tura turizm alternatif turizm ersoy turizm turizm gov eko turizm vista turizm denge turizm turizm pazarlaması turizm nedir termal turizm ankara turizm izmir turizm turizm teşvik turizm fuarı ticaret turizm turizm işletmeciliği yöre turizm türkiye de turizm uludag turizm il turizm müdürlüğü radar turizm turizm bakanl turizm gov tr turizm türkiye turizm haftasi öger turizm turizm otel kontur turizm turizm rehberliği sürdürülebilir turizm bosfor turizm kültür ve turizm bakanlığı ayder turizm turizm acentesi figür turizm oyak turizm show turizm bando turizm turizm tur buzlu turizm achill turizm turizm çeşitleri turizm ekonomisi dünya turizm örgütü akdeniz turizm seyahat turizm seç turizm turizm acentalar tempo turizm dadaş turizm net turizm bursa turizm özkaymak turizm turizm haritası turizm siteleri akdeniz üniversitesi turizm jet turizm turizm fuarları turizm politikası turizm yatırımları turizm bakanı turizm işletmeleri turizm tanıtım doğa turizm mersin turizm turizm yatırımcıları derneği tc turizm bakanlığı turizm acente turizm eleman turizm rketleri turkiye turizm tatil turizm www turizm gov tr turizm acentasi uluslararası turizm vasco turizm irem turizm kırsal turizm turizm mevzuatı cafe turizm pekin turizm turizm haberleri turizm tatil

tatlıses turizm turizm rehberlik turizm teşvik kanunu turizm ve çevre zafer turizm turizm iş ilanları turizm bakanlık turizm firmalari adana turizm bolu turizm murat turizm oger turizm iç turizm jolly turizm bos turizm tekser turizm valör turizm çağlar turizm plaza turizm ani turizm diamond turizm hisar turizm turizm bakan www turizm com turizm seyahat acentaları turizm sözlüğü setur turizm turizm şirket ben turizm nil er turizm trabzon turizm turizm makale karadeniz turizm turizm dergileri turizm ekonomi turizm politikaları ykm turizm çanakkale turizm alarko turizm turizm coğrafyası turizm şurası vals turizm gürsel turizm ispanya turizm iş arayanlar turizm turizm okulları turizm oteller ak turizm diana turizm kongre turizm babil turizm beydağı turizm konya turizm turizm dergisi turizm planlaması atoll turizm erguvan turizm turizm belgesi otel hatlar şirketleri koç firmaları uçak gezi haftası denizli asya buyukada kamil vapur bilet havayollari thy kamil koç hava yollari otobüsü ulusoy seferleri eminönü otogar milliyet şehir hatları deniz otobüsü evden eve otobus seyehat burgaz bakanlığı ulaşım hafta sonu işletmesi bostancı işletmeleri acentesi feribot turları nasıl gidilir hava yolu aşti adana otobusleri turizim ulaştırma turlar uçak bileti doğa acentasi isletmeleri hızlı truzim tarifesi vapuru sirkeci kabataş nilüfer sehir hatları seyahat acentaları deniz otobüsleri varan adalar deniz otobüs hürriyet bakanlıklar ankara sakalari iett sefer burgazada ucuz bandırma otogarı acentaları yolculuğu denizyolları turızm saatleri taşımacılığı tcddy ulasım miili eğitim mersin durağı ismail ayaz havayolları esenler devlet demir yolları kalkış kamilkoç izmir etstur demiryollari seyhat şehirlerarası yurtdışı topçular sakaları isletmesi gemi kamil koc taşıma şehirhatları denizcilik deniz otobusleri uçuş hava yolları şirketi ulaşımı arabalı taşımacılık demir yolları limanları gezileri hakiki asyatur otobüsleri ulasim acentası heybeliada iskelesi büyükada acenteleri anıtur eskihisar havaalanı milli eğtim denizotobüsü heybeli yenikapı baknlığı otob durakları bakalığı iç işleri sakalar mili eğitim günübirlik denizcilik işletmesi esadaş firması yolları geziciyak kamil ko dadaş demiryolları hidayet otobüsle bakanlık turlari nakliye haftasonu balayı deniz otob anı uçak biletleri ulastırma ets havaş bosfor ulaşim tatili nakliyat otobüsler çınarcık lüks millieğitim hatları bakanlıgı denizyollari türk hava yolları rezervasyon çevre seç seferi havayolu istanbul deniz otobüsleri yollari işletmeciliği kurban bayramı seyahati kapadokya üniversitesi buzlu gazetesi denizcilik işletmeleri kültür bakanlığı vapurları lojistik köksallar acentalar tatil oteller otobüs seyahat uludağ pamukkale

Liderlik

Liderlik Nedir? Liderlik sizinle aynı amacı paylasan bir gruba o amaca ulaşmak icin onculuk etmektir. Liderden ne beklenir?- İyi planlama ve organizasyon- Teknik ve fizik yeterlilik (En az ekibin ortalama kondusyonuna sahip olmak, ilkyardım, ip teknikleri gibi konularda genel hakimiyet)- Ekiple ilgili olmakZor durumlar icin saglam sinirlere sahip olmak (Paniklemis ekip uyesi grubun moralini asagı ceker bunu engellemek)MotivasyonSabırLiderden ne beklenmez? : Herşeyi planlayıp organize etmesi Hazırlık Asaması:Lider olarak ilk siz hazır olun :Teknik ve fizik eksikliklerinizi giderin.Limitlerinizi zorlayacak faliyetlerden kacının ya da limitlerinizi yükseltin.Malzemenizi hazır ve eksiksiz tutun ve tabi kullanmayı iyi bilin.Yerleri belli olsun ki hızlı ve kolay hazırlanabilesiniz.Faliyet planınızı iyi yapın . hava yol durumu gibi bilgileri onceden edinin.Grubunuzu belirleyin ve planınızda muhakkak esneklik olsun.Misal bivak yapmak durumunda kalabilirsiniz diye ptesi gunku onemli işlerinizi salıya kaydırın ekip elemanlarınıda durumdan haberdar edin.Gerekirse yardım isteyin.kalabalık bir grupla tek bir egitmenin ilgilenmesi zor olabilir .İlk kez konusuyorsanız insanları faliyetle ilgili aradığınızda sırf bilgi alışverisinde bulunmak icin degil belli bir ilişki geliştirmek konusun.Klüp malzemesi kullanacaksanız malzemenin durumunu onceden kontrol edin.Liderin stili :Kendiniz olun ki insanlar celişkiye dusmesin size olan guvenlerini yitirmesinler veya ilk izlenimleri kotu

olmasın.Kadınların Liderligi:Kadınlar da erkekler kadar iyi lider olabilirler.Bazı erkekler kadınların dagcılık alanında basarılı olmasını cekemeyebilirler.(Halt etmişler!) Kadınlara verebilecek tavsiye liderlikte erkek gibi davranmamları aksine kendi tarzlarını yaratmalarıdır. Unutulmamalı ki uyumlu bir grupta farklılık avantajdır. Kadınlar erkeklerden erkeklerde kadınlardan cok sey ogrenebilirler.Liderlikte Karar AlmaKararı dogru zamanda verin er veya gec degil.Ani kararlar almayın.Dusunun.Derin bir nefes alın olaya odaklanın.Eger secim yapmanız gerekirse sartlara gore karar verin . Aynı zorlukta 3 yol seceneginiz varsa artısı daha cok olanı secin.(Misal ben manzaralı olanı secerdim !)Olayları otamatige baglamayın.herhangi bir mekanizmanın sizin yerinize karar almasını engelleyin. Kontrolü kaybetmeyin.Liderin Ekiptekilere Karşı Tutumu:Duyarlı olun.Gerektiğinde kendinizi o kisinin yerine koyun ve çözümleri onların kosullarını gözönüne alarak yaratın.Dinleyin, önyargılı davranmayın, güven kazanın.Yeni baslayanlara özen gösterin .Mümkün oldugunca yardımcı olun .Bu kisiler bir sonraki sene liderlikte size yardım edecek konuma gelebilirler.Kendinizi de gözardı etmeyin. (Eğitim vereceğim diye kendinizi bitirmeyin ki grup yorgunluktan tükenmiş bir lidere mecbur kalmasın.)Zorunlu kalmadıkça insanları grup önünde eleştirmeyin.hatalarını uygun bir dille anlatın. Eğer hatayı cok kisi yapıyorsa grubu etrafınızda toplayıp dogrusunu anlatın.Eleştri yaparken kişilerin ve grubun motivasyonunu kaybetmemesine özen gosterin.Liderlik pozisyonunuzdan kötü anlamda istifade etmeye çalışmayın. (yorumsuz) Etkili İletişim:Yaptıgınız konusmanın,verilerinizin dogru olmasına ozen gösterin.Bilgiyi sadece bilmesi gereken kişiye iletin. (Yanlış listeye mail atmayın sakın)Mesajınızın alındıgına ve dogru anlasıldıgına emin olun.Eğer ekiple ilk faliyetinizse gruba kendinizi tanıtın onları da mumkun oldugu kadar tanımaya calışın. Ekip Kurma:Yapılacak faliyete fizik ve teknik uygunlukta partner secin. Hedefiniz ortak ve belli olsun.Ekipte kar yagsada bivak yapar zirve yaparız diyenlerle . yok abi ben bivak yapmam donerim diyenler olursa problem yasayabilirsiniz.

ÖZGEÇMİŞ

Doğum Tarihi	18-08-1973	
Doğum Yeri	Altıntaş- KÜTAHYA	
Lise	1984-1991	Ali Güral Lisesi
Üniversite	1991-1995	Ege Üniv. Müh. Fak. Bilgisayar Mühendisliği
Yüksek Lisans	1996-1999	Ege Üniv. Fen Bilimleri Enstitüsü Bilgisayar Müh. Anabilim Dalı, Bilgisayar Mühendisliği Programı
Yüksek Lisans	2001-2001	Sabanci Üniv. İşletme Yüksek Lisans - MBA

Çalıştığı Kurumlar

1995-	Gürok Turizm Madencilik ve San A.Ş. Bilgi İşlem Yöneticisi

Anahtar Kelimeler : Veri Madenciliği, Belge İşleme, Bilgi Çıkarımı, Bilgi Erişim, Bilgi Arama

JÜRİ :

1. Prof Dr. Oya KALIPSIZ
2. Prof Dr. Eşref ADALI
3. Prof. Dr. A. Çoşkun SÖNMEZ
4. Prof. Dr. Ümit KOCABIÇAK
5. Doç. Dr. Selim AKYOKUŞ

Kabul Tarihi : 8 Şubat 2007
Sayfa Sayısı : 162

Keywords : Data Mining, Tezt Categorization, Text Retrieval, Information Retrieval, Querying

JURY :

6. Prof Dr. Oya KALIPSIZ
7. Prof Dr. Eşref ADALI
8. Prof. Dr. A. Çoşkun SÖNMEZ
9. Prof. Dr. Ümit KOCABIÇAK
10. Doç. Dr. Selim AKYOKUŞ

Date : 8 Şubat 2007
Pagesı : 162