T.C.
DOKUZ EYLÜL ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
EKONOMETRİ ANABİLİM DALI
EKONOMETRİ PROGRAMI
YÜKSEK LİSANS TEZİ

METİN MADENCİLİĞİ TEKNİKLERİ İLE ŞİRKETLERİN VİZYON İFADELERİNİN ANALİZİ

Cemile MELEK

Danışman
Doç. Dr. İpek DEVECİ KOCAKOÇ

İZMİR - 2012

YÜKSEK LİSANS
TEZ/ PROJE ONAY SAYFASI

2008800185

Üniversite : Dokuz Eylül Üniversitesi

Enstitü : Sosyal Bilimler Enstitüsü

Adı ve Soyadı : Cemile MELEK

Tez Başlığı : Metin Madenciliği Teknikleri ile Şirketlerin Vizyon İfadelerinin Analizi

Savunma Tarihi : 10.01.2011

Danışmanı : Doç.Dr.İpek DEVECİ KOCAKOÇ

JÜRİ ÜYELERİ

Ünvanı, Adı, Soyadı	Üniversitesi	İmza
Doç.Dr.İpek DEVECİ KOCAKOÇ	DOKUZ EYLÜL ÜNİVERSİTESİ	.....
Yrd.Doç.Dr.Murat TANIK	DOKUZ EYLÜL ÜNİVERSİTESİ	.....
Doç.Dr.Pınar SÜRAL ÖZER	DOKUZ EYLÜL ÜNİVERSİTESİ	.....

Oybirliği (X)
Oy Çokluğu ( )

Cemile MELEK tarafından hazırlanmış ve sunulmuş "**Metin Madenciliği Teknikleri ile Şirketlerin Vizyon İfadelerinin Analizi**" başlıklı Tezi(X) / Projesi( ) kabul edilmiştir.

Prof.Dr. Utku UTKULU
Enstitü Müdürü

YEMİN METNİ

Yüksek Lisans Tezi olarak sunduğum "**Metin Madenciliği Teknikleri ile Şirketlerin Vizyon İfadelerinin Analizi**" adlı çalışmanın, tarafımdan, bilimsel ahlak ve geleneklere aykırı düşecek bir yardıma başvurmaksızın yazıldığını ve yararlandığım eserlerin kaynakçada gösterilenlerden oluştuğunu, bunlara atıf yapılarak yararlanılmış olduğunu belirtir ve bunu onurumla doğrularım.

Tarih

..../..../.....

Cemile MELEK

İmza

ÖZET

Yüksek Lisans Tezi
Metin Madenciliği Teknikleri ile Şirketlerin Vizyon İfadelerinin Analizi

Cemile MELEK

Dokuz Eylül Üniversitesi
Sosyal Bilimler Enstitüsü
Ekonometri Anabilim Dalı
Ekonometri Programı

Günümüzde mevcut veritabanlarında bulunan ham verilerin her geçen gün artması, ham verilerden elde edilmek istenen bilgilerin de doğru ve güvenilir olma ihtiyacını da arttırmıştır. Bu nedenle veri madenciliği önemli bir çalışma alanı haline gelmiştir. Veri madenciliği ile elde bulunan sayısal haldeki verilerin analizi rahatlıkla yapılabilmekteyken, metin halde bulunan verilerin analiz edilmesi de önemli bir ihtiyaç halinde gelmiş ve metin madenciliği konusunda yapılan çalışmaları da artmıştır. Metinsel verilerin sayısallaştırılarak veri madenciliği algoritmalarına girdi oluşturabilecek hale dönüşmesini sağlayan metin madenciliği günümüzde büyük önem teşkil etmektedir.

Bu çalışmada, metinsel veri kaynağı olarak ele alınan şirketlerin vizyon ifadelerinin incelenip itibar kriterlerinin analiz edilmesi amacıyla, Capital dergisi "En Beğenilen Şirketler" araştırmasında yer alan şirketlerin vizyon ifadeleri Statistica programı ile sayısallaştırılmış ve metin madenciliği yöntemleri aracılığıyla analiz edilmiştir.

Anahtar Kelimeler: Veri Madenciliği, Metin Madenciliği, İtibar Yönetimi ve Vizyon İfadeleri

ABSTRACT

Master's Thesis

Analysis of Vision Statements of Firms by Using Text Mining Techniques

Cemile MELEK

Dokuz Eylul University

Graduate School of Social Sciences

Department of Econometrics

Econometrics Program

The amount of raw data in databases available today increases each passing day. The knowledge that desired to be obtained from the raw data also increased the need for information to be accurate and reliable. For this reason, data mining has become an important area of study. Mining and analysis of data that is obtained in numerical form could easily be done; however, analyzing the data in the text form has a major case of need. Thus, studies that have been made in text mining area have increased. Text mining, which makes the conversion of textual data available for being input of data mining algorithms, is crucial today.

In this study, corporate vision statements are taken as a source of textual data. The corporate investigated are taken from the Capital Magazine's "Most Admired Companies 2010" survey. These vision statements have been analyzed via Statistica software and results of the analyses are interpreted according to criteria of reputation.

Keywords: Data Mining, Text Mining, Reputation Management and Vision Statements

İÇİNDEKİLER
MUHASEBE BİLGİLERİNİN GÜVENİRLİĞİNDE
MESLEKİ YARGININ ÖNEMİ

BİRİNCİ BÖLÜM
VERİ MADENCİLİĞİNİN TEORİK YAPISI

İKİNCİ BÖLÜM

METİN MADENCİLİĞİ

ÜÇÜNCÜ BÖLÜM

İTİBAR YÖNETİMİ VE VİZYON

DÖRDÜNCÜ BÖLÜM

UYGULAMA

ŞEKİLLER LİSTESİ

TABLOLAR LİSTESİ

GİRİŞ

Veri madenciliği büyük ölçekli veriler arasından araştırma için değerli olan bilgiyi elde etmede çeşitli istatistiksel tekniklerden yararlanarak önceden bilinmeyen ve veriler içinde gizli olan bilginin çıkarılmasıdır. Günümüzde veri tabanlarında bulunan verilerin her geçen gün daha da artması verileri analiz etmede birçok çalışma yapılmasının gerekliliğini ortaya koymuştur.

Veri madenciliği veritabanlarındaki sayısal halde bulunan verilerin çeşitli istatistiksel, analitik yöntemlerle analiz edilmesi ve elde edilen sonuçların yorumlanması ile ilgilenir. Fakat sayısal halde bulunmayan verilerin analiz edilmesi ihtiyacı sonucu metin halinde bulunan verilerin analizi hususunda da çeşitli çalışmalar yapılması gerekliliği duyulmuş ve sonuçta metin madenciliği alanı oluşmuştur. Metin madenciliği günümüzde kullanılan fakat çok yeni bir alandır. Bu alanda yapılan çalışmalar kullanılacak olan veri tabanında bulunan kelimelerin bir sözlüğünün oluşturulması ve Visual Basic gibi bir programlama dili ile kelimelerin saydırıldığı bir program kurulmasını veya piyasada bulunan mevcut paket programlar vasıtası ile kelimelerin saydırılarak sayısal hale dönüşmesini içermektedir.

Çalışmada uygulama bölümünde yapılacak analizlerden bir çıkarım sağlanabilmesi adına, itibar yönetimi kavramı, itibarlı şirket olma ölçütleri, itibarın ölçülmesi, şirketlerin vizyon ifadelerinin hangi kriterlere göre oluşturulması gerektiği, iyi bir vizyon ifadesinde bulunması gereken kavramlar gibi konular teorik açıdan incelenmiştir.

Bu tezde, metinsel veri kaynağı olarak Capital dergisi “En Beğenilen Şirketler” araştırmasında 2010 yılında yer alan şirketlerin vizyon ifadeleri, itibar boyutları ile ilişkilendirilmeleri amacıyla veri olarak alınmış ve Statistica paket programının metin madenciliği modülü kullanılarak veriler sayısal hale dönüştürülmüştür. Sayısal hale dönüştürülen verilere veri madenciliği teknikleri uygulanmış, faktör analizi yapılarak ilgili şirketlerin vizyon ifadelerinde en çok önem verdikleri kriterlerin çıkarımı yapılmış ve itibar boyutlarıyla ilişkilendirilmiştir.

Çalışma kapsamında ilk bölümde veri madenciliği konuları incelenmiş, veri madenciliğinde kullanılan analiz yöntemleri incelenmiş, ikinci bölümde ise veri madenciliğinin yetersiz olduğu metinsel verilerin analiz edilebilmesinde veri madenciliği tekniklerinin uygulanabilmesine hazır hale getirilmesini sağlayan metin madenciliği konusu incelenmiştir. Üçüncü bölümde ise, uygulama kısmındaki metinsel verileri oluşturan şirket vizyon ifadelerinin incelenmesi sonucunda bir çıkarım elde edebilmek amacıyla itibar yönetimi, itibar boyutları, itibarlı şirket olma ölçütleri ve vizyon kavramı ile ilgili teorik konular incelenmiştir. Dördüncü bölümde vizyon ifadelerinden elde edilen veriler analiz edilmiş, metin madenciliği kullanılarak kelimeler sayısallaştırılmış ve sayısallaştırılan verilere veri madenciliği teknikleri uygulanılarak, itibar kriterleri de göz önüne alınarak vizyon ifadeleri analiz edilmiştir. Sonuç ve öneriler kısmı olan son kısımda da ilgili çalışmada elde edilen sonuçlar yorumlanmış, metin madenciliği ile ilgili ilerleyen zamanlarda yapılabilecek çalışmalar önerilmiştir.

BİRİNCİ BÖLÜM

VERİ MADENCİLİĞİNİN TEORİK YAPISI

1.1. VERİ MADENCİLİĞİ NEDİR?

Veri madenciliği büyük ölçekli veriler arasından "değeri olan" bir bilgiyi elde etme işidir. Veri madenciliği verilerin, belirli yöntemler kullanarak var olan ya da gelecekte ortaya çıkabilecek gizli bilgiyi su yüzüne çıkarma süreci olarak değerlendirilebilir. Bu açıdan bakıldığında, veri madenciliği işinin kurumların karar destek sistemleri için önemli bir yere sahip olabileceğini söyleyebiliriz (http://mf.dumlupinar.edu.tr).

Gartner grubuna göre veri madenciliği, yeni korelasyonlar örnekler ve trendleri stoklanmış büyük miktardaki verilerden eleyerek istatistiksel ve matematiksel tekniklerde olduğu kadar örnek tanımlama teknolojilerini kullanır (Larose, 2005: 21).

Önceden bilinmeyen ve potansiyel olarak faydalı olabilecek, veri içinde gizli bilgilerin çıkarılmasına veri madenciliği denir. Diğer bir tanım ise, veri madenciliği, büyük veri kümesi içinde saklı olan genel örüntülerin ve ilişkilerin bulunmasıdır (Adsız, 2006: 9).

Veri madenciliği büyük veri yığınlarında gizli olan örüntüleri ve ilişkileri ortaya çıkarmak için istatistik ve yapay zeka kökenli çok sayıda ileri veri çözümleme yönteminin tercihen görsel bir programlama ara yüzü üzerinden kullanıldığı bir süreçtir. Veri madenciliği algoritmaları; istatistiksel algoritmalar, matematiksel algoritmalar ve yapay zeka algoritmalarını (sinir ağları, karar ağaçları, kohonen ağlar, birliktelik kuralları vb.) bir arada içerir (Dolgun ve diğerleri, 2009: 49).

Veri madenciliği aslında klasik istatistiksel uygulamalara çok benzer. Ancak klasik istatistiksel uygulamalar yeterince düzenlenmiş ve çoğunlukla özet veriler üzerinde çalıştırılır. Veri madenciliğinde ise milyonlarca ve hatta milyarlarca veri ve çok daha fazla değişken ile ilgilenilir. Veri sayısı çok olunca, bazı özel analiz

algoritmalarının geliştirilmesi gerekmiş, ayrıca verinin saklandığı ortamların da örneğin veri ambarı biçiminde yeniden düzenlenmesini gerekli kılmıştır (http://mf.dumlupinar.edu.tr).

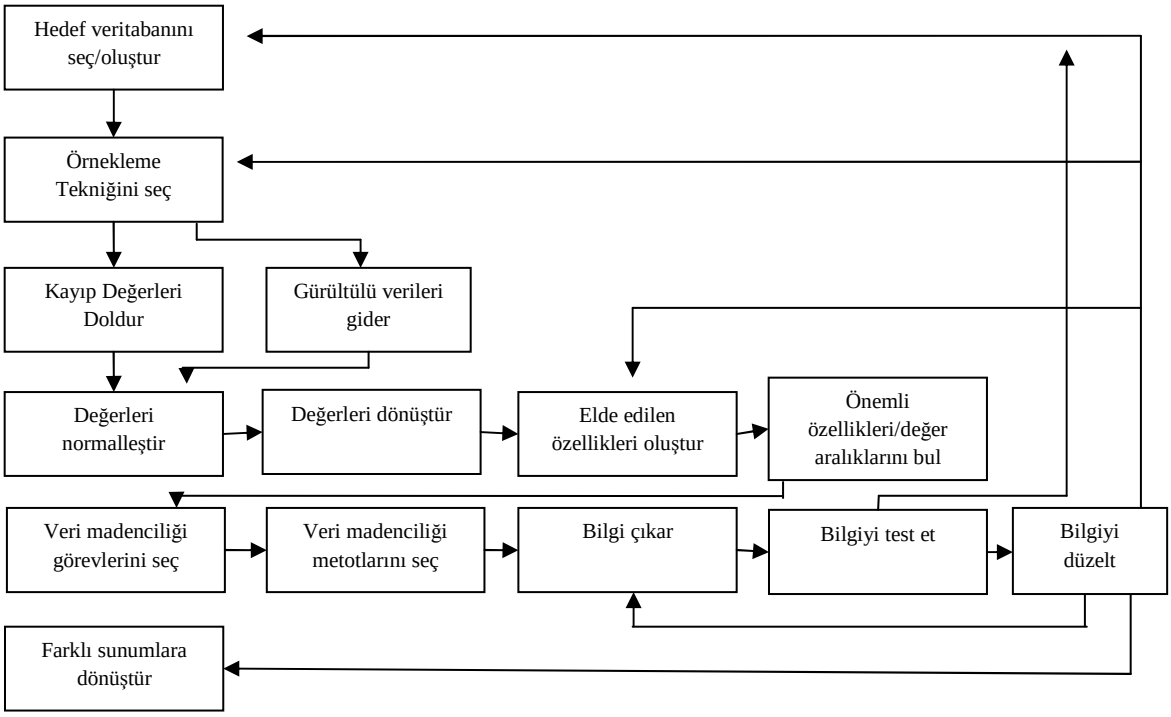
Veri Madenciliği, büyük ölçekli veriler arasından bilgiye ulaşma, veriyi madenleme işlemidir. Veri tabanlarındaki, veri ambarlarındaki veya dosyalarda bulunan veriler arasında bulunan ilişkiler, örüntüler, sapma ve eğilimler, belirli yapılar gibi bilgilerin ortaya çıkarılması ve keşfi veri madenciliğinin temelini oluşturur. "Veri Tabanlarından Bilgi Keşfi" (Knowledge Discovery in Databases) uygulamaları ile birlikte faaliyet alanına yönelik karar destek mekanizmaları için gerekli ön bilgileri temin etmek için kullanılır. Geleneksel yöntemler kullanılarak çözülmesi çok zaman olan problemlere veri madenciliği süreci kullanılarak daha hızlı bir şekilde çözüm bulunabilir (Tekerek, 2011).

Veri madenciliği açıklayıcı veri analizinin bir uzantısıdır ve verilerdeki bilinmeyen ve beklenmeyen yapının keşfedilmesi gibi temelde aynı amaçlara sahiptir. Temel ayrım veri setlerinin içerdiği büyüklük ve boyutluluğa uzanır. Veri madenciliği genelde, tam olarak uygulanabilir olmayan yüksek interaktif analizler için daha büyük kütleli veri setleri ile ilgilenir (Rao ve diğerleri, 2005: 9).

Veri madenciliği; veri ambarlarındaki çeşitli verileri kullanarak yeni bilgileri ortaya çıkarmak ve bu bilgileri karar verme ve uygulama aşamasında kullanma sürecidir. Veri Madenciliği kendi başına bir çözüm üretmemekte, ancak çözüm için gerekli bilgileri sağlamakta ve karar verme aşamasında yardımcı olmaktadır (Küçüksille, 2009: 28).

Aşağıdaki şekil veri madenciliği işlem süreçlerini göstermektedir (Rao ve diğerleri, 2005: 14);

Şekil 1: Veri madenciliği akış şeması



Kaynak: Rao ve diğerleri, 2005, s. 14.

1.2. DOKÜMAN AMBARLARI

Veri ambarlarının veri madenciliğinde kabul edilen tanımlamalarını karşılaştırdığımızda, doküman ambarı için dört tanımlama özelliği çıkarabiliriz,

1) Çoklu doküman tipleri,

2) Çoklu doküman kaynakları,

3) Doküman ambarındaki dokümanların önemli özelliklerini depolama ve otomatik olarak çizme,

4) Manasal olarak ilişkili dokümanları birleştirmek.

Doküman ambarının anahtar unsuru; doküman ambarının, sorgu ve analiz için gereksinimleri karşılamada işlem metnini yeniden yapılandırmak için kolayca erişilebilir ham verilerin gerektirdiği bilgiyi yapabilmesidir. Doküman ambarı elektronik postaları, tam metin dokümanları, HTML dosyaları gibi olan doğal dile

dayalı yapılandırılmamış ya da yarı yapılandırılmış yazılı kaynakların büyük miktarlarını depolamak için tasarlanmıştır. Bu metinsel bilginin kesin doğası tüm dokümanı, dokümanın otomatik olarak türetilmiş özetlerini, dokümanın çeşitli dillerdeki çevirilerini, dokümanlar hakkındaki metadataları, yazar adı gibi, yayınlanma tarihi ve konu anahtar kelimeleri, benzer dokümanlar hakkındaki kümeleme bilgilerini içeriksel ya da başlık içeriklerini içerebilir. Sonuç olarak, doküman ambarlama boyunca metinde uygulanan temel faaliyet alanını elde edebiliriz: özetleme, kümeleme, özellik çıkarımı, kategorizasyon ve konu izleme.

Doküman ambarları, kendilerini cevap vermeleri için tasarlanmış soru tipleri tarafından veri ambarlarından ayırırlar. Veri ambarları kim, ne, ne zaman, nerede ve ne kadar gibi sorulara cevap vermede kusursuzdurlar fakat doküman ambarlarının güçlü noktası olan neden soruları ile ilgilenirken güçlerini kaybederler. Veri ambarları ile doküman ambarları arasında en çok ayır edici özellik pratikte veri ambarlarının iç odaklı (internally focused) olmasıdır. Onları organizasyonumuzda operasyonel bilgiyi analiz etmede daha iyi kullanabiliriz (Gao ve diğerleri, 2005).

1.3. BİLGİ KEŞFİ VE VERİ MADENCİLİĞİ

Düşük seviye bilgiden yüksek seviye bilgiyi çıkarma süreçlerinin tümünü göstermek için bilgi keşfi terimi kullanılır. Bilgi keşfi için kullanılan kelimelerin anlamı veri ya da bilgi toplama, veri arkeolojisi, fonksiyonel bağımsızlık analizi, bilgi çıkarımı ve örnek analizini içerir. Tarihsel olarak, istatistikte özellikle veri madenciliğinde doğrulanacak bir ön hipotez olmadan yarım yamalak bir açılayıcı veri analizine başvurur. Basit bir tanım olarak; basit bir yüksek seviye bilgi keşfi tanımı, bilgi keşfi veritabanlarındaki önemsiz olmayan potansiyel olarak kullanışlı ve verideki nihai olarak anlaşılabilir örneklerin doğruluğunu tanımlama sürecidir. Bilgi etki (domain) bağımlı terimlerle ilgili; faydayı, doğruluğu, yeniliği ve anlaşılabilirliği ölçer. Bu tanımdaki *örnekler* ifadesi ya modelleri ya da örnekleri belirtmektedir. Genelde verilerin bir alt kümesinin bazı özet sunumunu belirler. Bilgi keşfi çoğunlukla deneme, yineleme, kullanıcı etkileşimi ve birçok tasarım kararı ve özelleştirmeyi içerir. Verilerden bilgi çıkarma kolaylıkla karmaşık ve bazen de zor

bir sürece dönüşebilir. Veri madencililiği verilerden örnekler ya da modeller çıkarır. Büyük bir veri ambarından verileri almak, çalışılacak uygun altkümeyi seçmek, uygun bir örnekleme stratejisine karar vermek, verileri temizlemek ve kayıp verilerle uğraşmak, uygun dönüşümleri boyutluluk azaltmayı ve gösterimleri uygulama gibi birçok veri madenciliği adımı bulunmaktadır. Tüm bu adımlardan sonra veri madenciliği adımı modeli oluşturur ya da ön işlenmiş verilerden örnekleri çıkarır. Bu çıkarılmış bilginin "bilgi"yi sunduğuna karar vermek için birinin bu bilgiyi değerlendirmesi ve görselleştirmesi ve sonuç olarak da onu var olan bilgi ile sağlamlaştırması gerekir. Açıkça, bu adımların hepsi veriden bilgiye giden kritik bir yoldadır. Herhangi bir adım, yeni seçenekler ve ayarlar ile sıfırdan başlamayı gerektirebilen önceki ya da sonraki adımların değişmesine sebep olabilir. Bundan dolayı veri madenciliği tüm Bilgi Keşfi sürecinin sadece bir adımıdır. (Sumathi ve Sivanandam, 2006: 187-188).

Bilgi keşfi veri hakkında "bilgi" olarak tarif edilebilecek örnekler için verilerin geniş hacimlerine otomatik olarak ulaşma sürecidir.

Veritabanlarında bilgi keşfi, bilgi keşfi sürecinin amaçlarını karşılayan örnek ya da modelleri tanımlamayı içerir. Bu yüzden bir bilgi keşfi mühendisi keşfedilmiş örneğin geçerliliğini, örneğin faydasını ölçebilmeye ihtiyaç duyar. Bu ölçümler keşfedilmiş bir örneğin "ilginçliği"nin derecesini tanımlamada yardımcıdır. Veri madenciliği bilgi keşfi sürecinde bir adım olarak tanımlanabilir. Bilgi keşfi süreci ise bilginin ne olduğunu çıkarmada veri madenciliği metotlarının kullanılması süreci olarak tanımlanabilir (Wegman ve Solka, 2005: 9).

1.4. VERİ MADENCİLİĞİ UYGULAMA ALANLARI

Veri madenciliği teknolojisi bir karar verilmesi gereken her yerde kullanılabilir, geçmişteki uygulamaların çeşitliliği aşağıdaki gibidir (Nisbet ve diğerleri 2009: 26) :

- Satış tahmini: veri madenciliği teknolojisinin ilk örnekleridir
- Raf Yönetimi: satış tahmininin mantıksal devamı

- Bilimsel keşif
- Oyun: müşterilerin yüksek harcama potansiyelini tahmin etme
- Spor: yüksek skor için en iyi potansiyele sahip olan durumu keşfetme metodu
- Müşteri ilişkileri yönetimi
- Müşteri edinme

Veri madenciliğinin geçmişteki uygulamaları ile ilgili verilebilecek ek örnekler de aşağıdaki gibidir;

- Pazarlama; müşterilerin satın alma alışkanlıklarının belirlenmesi, müşterilerin demografik özellikleri arasındaki bağlantıların bulunması, posta kampanyalarında cevap verme oranının artırılması, mevcut müşterilerin elde tutulması, yeni müşterilerin kazanılması, pazar sepeti analizi, müşteri ilişkileri yönetimi, müşteri değerlendirmesi, satış tahmini, çapraz satış analizi, mevcut müşterilerin elde tutulması için geliştirilecek pazarlama stratejilerin oluşturulması.
- Bankacılık; farklı finanssal göstergeler arasında gizli korelasyonların bulunması, kredi kartı dolandırıcılıklarının tespiti, kredi kartı harcamalarına göre müşteri gruplarının belirlenmesi, kredi taleplerinin değerlendirilmesi, müşteri dağılımı, usulsüzlük tespiti, risk analizleri.
- Sigortacılık, yeni poliçe talep edecek müşterilerin tahmin edilmesi, sigorta dolandırıcılıklarının tespiti, riskli müşteri örüntülerinin belirlenmesi.
- Perakendecilik, satış noktası veri analizleri, alış-veriş sepeti analizleri, tedarik ve mağaza yerleşim optimizasyonu, hisse senedi fiyat tahmini, genel piyasa analizleri, alım-satım stratejilerinin optimizasyonu.
- Endüstri, kalite kontrol analizleri, lojistik, üretim süreçlerinin optimizasyonu olarak belirtilebilir (Şen, 2008: 11).

1.5. VERİ MADENCİLİĞİ SÜREÇLERİNİN GÜÇLÜ YANLARI

Geleneksel istatistiksel çalışmalar bir sistemin gelecek durumunu belirlemede geçmiş bilgileri kullanır, böylece veri madenciliği çalışmaları sadece tek girdi verilerine değil aynı zamanda bu verilerin yerel mantıksal sonuçlarının örneklerini kurmada geçmiş bilgileri kullanır. Bu süreç ayrıca tahmin olarak

adlandırılır, fakat istatistiksel analizlerde bu kayıp hayati elemanlarını içerir: sırasıyla gelecekte ne olabileceğinin ifadesi, geçmişte ne olduğunun kıyaslanması (Sever ve Oğuz, 2003).

Veri madenciliği 1) önceden görülmeyen örneklerin bulunmasıyla verilerin tamamen anlaşılmasını sağlamada ve 2) tahminlenen modelleri yapmak böylece insanların daha iyi kararlar vermesini, harekete geçmesini ve gelecek olayları kalıplaştırmayı sağlar.

1.6. VERİ MADENCİLİĞİNDE KARŞILAŞILAN ZORLUKLAR

Veri madenciliği uygulanacak veri setleri büyük olduklarında analizde yavaş çalışmaya gereksiz ve hatalı sonuçlar elde edilmesine sebep olabilirler, küçük veri setlerinde doğru sonuçlar veren bir veri madenciliği sistemi büyük verilere uygulandığında hatalı sonuçlar üretilmesi ile sonuçlanabilir. Hataların sebebi genellikle büyük veri setleri söz konusu olduğunda verilerin hatalı, gürültülü olması veri setlerinde boş değerlerin bulunması gibi nedenlerdir.

Veri madenciliği girdi olarak kullanılacak ham veriyi veritabanlarından alır. Bu da veritabanlarının dinamik, eksiksiz, geniş ve net veri içermemesi durumunda sorunlar doğurur. Küçük veri kümelerinde hızlı ve doğru bir biçimde çalışan bir sistem, çok büyük veri tabanlarına uygulandığında tamamen farklı davranabilir (Şen, 2008: 12).

Veri madenciliğinde veritabanlarında karşılaşılan ve veri madenciliği sürecini olumsuz etkileyen nedenler aşağıda alt başlıklar halinde ele alınmıştır.

1.6.1. Veri Tabanı Boyutu

Büyük veri kümeleri çoğunlukla eksik, kirli ve hatalı veri noktalarını içerecektir. Bu tip hatalara sahip olmayan veri kümeleri az rastlanılan veri kümeleridir. Veri kümesinin büyüklüğü zorluklara yol açarken, standart istatistiksel uygulamalarda sık karşılaşılmayan bir takım özellikler ortaya çıkabilir. Veri madenciliğinde veriler, veri madenciliği uygulamak üzere değil diğer bazı amaçlar için toplanmaktadır. Tersi bir biçimde, pek çok istatistiksel çalışmada veriler akıldaki

belirli sorular için toplanır ve bu sorulara yanıt bulmak için analiz edilir. İstatistik, deney tasarımı ve alan araştırması gibi alt disiplinleri içermektedir. Bu disiplinler, veri toplamak için en iyi yollarla ilgili ipucu sağlarlar (Oğuzlar, 2003).

Veri tabanı boyutları inanılmaz bir hızla artmaktadır. Pek çok makine öğrenimi algoritması birkaç yüz tutanaklık oldukça küçük örneklemeleri ele alabilecek biçimde geliştirilmiştir. Örneklemenin büyük olması, örüntülerin gerçekten var olduğunu göstermesi açısından bir avantajdır ancak böyle bir örneklemeden elde edilebilecek olası örüntü sayısı da çok büyüktür. Bu yüzden veri madenciliği sistemlerinin karşı karşıya olduğu en önemli sorunlardan biri veri tabanı boyutunun çok büyük olmasıdır. Dolayısıyla veri madenciliği yöntemleri ya sezgisel bir yaklaşımla arama uzayını taramalıdır, ya da örneklemeyi yatay/dikey olarak indirgemelidir. Yatayda indirgeme veri alanının örneklenmesi, dikeyde indirgeme ise özelliklerin bulunduğu kolonların azaltılma çalışmasıdır (Şen, 2008: 13).

1.6.2. Gürültülü Veri

Verilerdeki gürültü ölçülmüş bir özelliğin rassal bir hatası ya da varyansı olan bir değer olarak tanımlanır. Verilerdeki miktarına bağlı olarak, gürültü bilgi keşfi sürecini tehlikeye atabilecek olan önemli bir problem olabilir. Verilerdeki gürültünün etkisi veriler girildiğinde anormallikleri tespit etmede özeliklere kısıtları uygulayarak önlenebilir. Gürültü oluştuysa bu özellik değerlerinin önceden belirlenmiş kısıtları kullanarak elle kontrol, ambarlama (binning) ve kümeleme metotları kullanılarak silinebilir (Cios ve diğerleri, 2007: 40).

Büyük veritabanlarında pek çok niteliğin değeri yanlış olabilir. Bu hata, veri girişi sırasında yapılan insan hataları veya girilen değerin yanlış ölçülmesinden kaynaklanır. Veri girişi ya da veri toplanması sırasında oluşan sistem dışı hatalara gürültü adı verilir. Veri kümesi gürültülü ise bozuk veri ihmal edilmelidir.

1.6.3. Boş Değerler

Birçok veri seti boş değer problemi ile karşılaşmaktadır. Bu problem tamamlanmamış veri girişinden, yanlış ölçümlerden, donanım hatalarından vb. dolayı meydana gelmiş olabilir. Her boş değer "NULL", "*" ve "?" ile gösterilir. Boş değerler silinerek ya da doldurularak giderilebilir (Cios ve diğerleri, 2007: 40).

1.6.4. Eksik ve Artık Veriler

Veri madenciliğinde kullanılacak olan veri kümesinde bir değer bilinmiyor olabilir ya da girilmemiş eksik girilmiş olabilir veya artık nitelikler içerebilir. Veri madenciliği yöntemlerinde ise her verinin bir özellik belirtiyor olmasından dolayı eksik veriler analizde sorun teşkil eder. Artık veriyi önlemek için özellik seçimi yapılmalıdır.

Özellik seçimi, tümevarıma dayalı öğrenmede budama öncesi yapılan bir işlemdir. Başka bir deyişle, özellik seçimi, verilen bir ilişkinin içsel tanımını, dışsal tanımın taşıdığı (veya içerdiği) bilgiyi bozmadan onu eldeki niteliklerden daha az sayıdaki niteliklerle (yeterli ve gerekli) ifadeleyebilmektir. Özellik seçimi yalnızca arama uzayını küçültmekle kalmayıp, sınıflama işleminin kalitesini de arttırır (Sever ve Oğuz 2003).

1.6.5. Eksik Verilerin Doldurulması

Veri doldurma bir ya da çoklu doldurma metotları ile alt bölümlere ayrılabilen birkaç farklı algoritma kullanarak uygulanır. Tekli veri doldurma metotlarında kayıp değer tek bir değer ile doldurulur. Çoklu veri doldurma metotlarında ise kayıp değeri doldurmada olasılık hesapları ile değerler hesaplanır ve "en iyi" değer seçilir (Cios ve diğerleri, 2007: 44).

Belirli durumlarda veriden bir değişken eksikse, eğer mümkünse, bunu sezgisel verilerle doldurmak çok önemlidir. Bu değişken için uygun bir verinin

makul bir tahminini eklemek boş bırakmaktan daha iyidir. Verilerdeki bu boşlukları doldurma işlemine *veri doldurulması* denir.

Liste-boyunca (ya da durum-boyunca) silme: bu analizden tüm kayıtların silindiği anlamına gelir. Bu teknik genellikle birçok istatistik ve otomatik öğrenme algoritmaları tarafından kullanılan varsayılan metottur. Bu tekniğin birkaç avantajı vardır;

- Herhangi bir veri madenciliği analizinde kullanılabilir
- Başarmak için herhangi özel bir istatistiksel metoda ihtiyaç duymaz
- Değişkenlerin tamamen bağımsız olduğu veriler için iyidir
- Doğrusal regresyon ve hatta Lojistik ve Poisson regresyonu ile kullanmak için daha uygun olan veri setleri için uygulanabilirdir.

İkili silme: Bu bir değişkenin değerleri ile tüm durumlarda bu değişkenin kovaryansını hesaplamada kullanılacağı anlamına gelir. Bu yaklaşımın avantajı bir doğrusal regresyonun sadece örnek ortalaması ve kovaryans matrisinden tahminlenmesidir.

Uygun bir değer atfetme: kayıp olamayan durumların ortalaması ile kayıp değerlerin atfedilmesinde sık sık ortalama ikamesi anlamına gelir. Eğer kayıp değerin özel bir değerini uygulayan bir karar kuralını güvenli bir şekilde uygulayabilirseniz, o zaman ortalama ikamesinden bile doğru bir değere yaklaşmış olursunuz.

1.7. VERİ MADENCİLİĞİ MODELLERİ VE KULLANILAN ALGORİTMALAR

Veri madenciliği modelleri işlevlerine göre 3 temel grupta toplanır (Şen, 2008: 15):

1. Sınıflama (Classification) ve Regresyon,
2. Kümeleme (Clustering),
3. Birliktelik kuralları ve sıralı örüntüler

1.7.1. Sınıflama

Sınıflama, yeni bir veri elemanını daha önceden belirlenmiş sınıflara atamayı amaçlar. Sınıflama ve regresyon, önemli veri sınıflarını ortaya koyan veya gelecek veri eğilimlerini tahmin eden modelleri kurabilen iki veri analiz yöntemidir. Sınıflama kategorik değerleri tahmin ederken, regresyon süreklilik gösteren değerlerin tahmin edilmesinde kullanılır. Sınıflama, verinin önceden belirlenen çıktılara uygun olarak ayrıştırılmasını sağlayan bir tekniktir. Çıktılar, önceden bilindiği için sınıflama, veri kümesini denetimli olarak öğrenir (Şen, 2008: 16).

Sınıflandırmada, örneğin yüksek gelir, orta gelir ve düşük gelir gibi üç gruba ya da kategoriye bölümlenebilen gelir kategorisi gibi bir hedef değişken vardır. Veri madenciliği modeli girdi ya da tahminci değişken setindeki gibi hedef değişkenler üzerinde bilgi içeren her bir kaydın büyük setlerini inceler (Larose, 2005: 46).

Veri madenciliğinde sınıflama, önceden tanımlanmış sınıfların birinde görülmeyen verileri sınıflandırmada kullanılabilen önceden sınıflandırılmış veri nesnelerinden bir model çıkarmayı gerektirir. Bir veri nesnesi özellikler ya da değişkenler seti olarak tanımlanmış bir örnek olarak ifade edilir. Özelliklerden her biri örneğin ait olduğu ve böylece sınıf özelliği ya da sınıf değişkeni olarak tanımlanan bir örnek sınıfı tanımlar. Diğer özellikler çoğunlukla bağımsız ya da tahminci özellikler (değişkenler) olarak tanımlanır. Sınıflandırma modelini öğrenmede kullanılan örnek setleri "eğitim veri seti" olarak tanımlanır. Sınıflandırma ile ilgili görevler sayısal veriler tahminlemede eğitim veri setinden bir model kuran regresyonu, kategorilerden örneklerli gruplandıran kümelemeyi içerir. Sınıflandırma "denetleyici (supervised) öğrenme" kategorisine aittir. Denetleyici öğrenmede eğitim verileri, girdi veri çiftlerini ve istenilen çıktıları içerir. Sınıflandırmanın bir hasta veritabanından hastanın belirtilerine dayalı olarak hastalığı teşhis etme, kredi kartı işlemlerini analiz ederek hileli işlemleri belirleme, el yazısı örneklerine dayalı olarak harflerin otomatik olarak tanımlanması gibi çeşitli uygulamaları vardır (Wang, 2006: 175).

Çoğu uygulamada sınıf etiketlerinden ziyade bazı kayıp ya da uygun olmayan veri değerleri tahmin edilmek istenebilir. Bu durum genelde tahminlenmiş değerlerin sayısal veriler olduğu durumda gerçekleşir ve tahmin olarak adlandırılır. Tahminin hem veri değeri tahmini hem de sınıf etiketi tahmini olarak adlandırılmasına rağmen, çoğunlukla veri değeri tahmini ile sınırlıdır ve böylece sınıflamadan farklıdır. Tahminleme aynı zamanda eldeki verilere dayalı olarak dağılım trendlerinin tanımlanmasını kapsar. Sınırlama ve tahminleme, sınıflama ya da tahminleme sürecine katkıda bulunmayan özellikleri tanımlamayı amaçlayan uygunluk analizlerinden önce yapılmalıdır (Han ve Kamber, 2000: 30-31).

Sınıflama ve regresyon modellerinde kullanılan başlıca teknikler şunlardır

1 - Karar Ağaçları (Decision Trees)

2- Yapay Sinir Ağları (Artificial Neural Networks)

3- Genetik Algoritmalar (Genetic Algorithms)

4- K-En Yakın Komşu (K-Nearest Neighbor)

5- Bellek Temelli Nedenleme (Memory Based Reasoning)

6- Naive-Bayes

Karar ağaçları tahmin etmede kullanılan bir tekniktir. Karar ağaçları aynı zamanda kural çıkarma algoritmalarıdır. Bu algoritmalar bir veri kümesinden kullanıcıların çok kolay anlayabileceği “eğer doğruysa” (IF-THEN) türündeki kuralları bir ağaç yapısında türetebilirler. Ağacın her dalı bir kural ve yaprakları da bu kuralın sağlanması durumunda dahil olunacak sınıfı gösterir. Karar ağaçları kolayca anlaşılabilir kurallar çıkarması nedeniyle çok kullanılan bir tekniktir. Bu teknikte dikkat edilmesi gereken nokta; ağacın tek bir kayıt kalana kadar büyümesidir. Bu durumdan kuralları oluşturma sırasında çok fazla zaman gerektireceği için mümkün olduğunca kaçınılmalıdır (Küçüksille, 2009: 38).

Sinir ağları, tanımlayıcı ve tahminci veri madenciliği algoritmalarındandır. İnsan beyninin fizyolojisini taklit ederler. Komplike ve belirsiz veriden bilgi üretirler. Keşfettikleri örüntü ve trendler, insanlar ya da bilgisayarlarca kolay

keşfedilemez. Bu tür karmaşık problemlerde birbirleriyle etkileşimli yüzlerce değişken bulunur. Bu teknik, veritabanındaki örüntüleri, sınıflandırma ve tahminde kullanılmak üzere genelleştirir. Sinir ağları algoritmaları sayısal veriler üzerinde çalışırlar (Şen, 2008: 19).

Sinir ağlarının avantajları;

- Genel sınıflayıcıdırlar. Birçok parametre ile problemleri alabilirler ve nesnelerin dağılımı N-boyutlu parametre uzayında çok kompleks olduğunda bile nesneleri sınıflamada çok iyidirler
- Tahminci değişkenlerdeki doğrusal olmayanlığın büyük miktarlarını ele alabilirler.
- Sayısal tahminci problemleri için kullanılabilirler (regresyon gibi)
- Verilerin dağılımı hakkında hiçbir varsayım altıda bulunmamayı gerektirir
- Doğrusal olmayan ilişkileri bulmada çok iyidirler. Sinir ağı yapısının saklı tabakası yüksekçe doğrusal olmayan fonksiyonları etkili bir şekilde modelleme yeteneğini sağlar.

Sinir ağlarının dezavantajları;

- Göreli olarak yavaş olabilirler, özellikle eğitim aşamasında ve ayrıca uygulama aşamasında
- Ağın kararını nasıl yaptığını açıklamak zordur. Bu yüzden sinir ağları bir "kara kutu" olmanın şöhretine sahiptir.
- Hiçbir hipotez test edilmez ve hiçbir p değeri değişkenleri karşılaştırmak için çıktılarda mümkün değildir (Nisbet ve diğerleri, 2009: 133).

Genetik algoritmalar diğer veri madenciliği algoritmalarını geliştirmek için kullanılan optimizasyon teknikleridir. Sonuç model veriye uygulanarak gizli kalmış kalıpları ortaya çıkarılmakta ve bu sayede tahminler yapılabilmektedir. Doğrudan postalama, risk analizi ve perakende analizlerinde kullanılabilir (Küçüksille, 2009: 41).

Veri uzayında birbirine yakın olan aynı tip kayıtlar, birbirlerinin komşusu durumundadırlar. Bu anlayış doğrultusunda, çok kolay fakat güçlü olan k – en yakın komşu algoritması geliştirilmiştir. k - en yakın komşu algoritmasının temel felsefesi

komşunun yaptığını yaptır. Belirli bir bireyin (kayıtın) davranışı (özelliğini) tahmin etmek istenirse, veri uzayında o bireye yakın olan örneğin 10 bireyin davranışına bakılabilir. Bu 10 komşunun davranışının ortalaması hesaplanır ve bu hesaplanan ortalama bireylerin tahmini olur. k - en yakın komşudaki k harfi araştırdığımız komşu sayısıdır. Örneğin, 5 - en yakın komşuda 5 komşuya bakılır (Şen, 2008: 19).

1.7.2. Kümeleme

Kümeleme grup kayıtlarını, gözlemleri, ya da vakaların benzer nesnelerle sınıflandırılmasını ifade eder. Bir küme benzer olan kayıtların toplamından oluşur ve diğer kümenin kayıtlarından farklıdır. Kümeleme hedef değişkenin olup olmaması ile sınıflandırmadan ayrılır. Kümelemede hedef değişken yoktur. Kümeleme hedef değişkenin sınıflandırılmaya çalışılması, tahmin edilmesi ya da değerlendirilmesi değildir. Aksine, kümeleme algoritmaları küme içerisindeki benzer kayıtların maksimize edildiği ve küme dışındaki benzer kayıtların minimize edildiği, ilişkili homojen alt gruplar ya da kümelerin tüm veri setindeki parçalarını araştırır (Larose, 2005: 46).

Kümeleme analizi, benzer özelliklere sahip bireylerin belirlenip gruplandığı çok değişkenli bir çözümleme tekniğidir. Kümeleme analizi sayesinde dağılımdaki yoğun ve seyrek alanlar belirlenebilir ve farklı dağılım örnekleri uygulanabilir (Küçüksille, 2009: 36).

Sınıflama ve tahminlemeden farklı olarak, sınıf etiketli veri nesnelerini analiz eden, kümeleme veri nesnelerini bilinen bir sınıf etiketi olmadan analiz eder. Genelde, sınıf etiketleri başlangıçta bilinmediklerinden dolayı eğitim verilerinde kolaylıkla bulunmaz. Kümeleme her bir etiketi türetmede kullanılabilir. Nesneler, sınıflar içi benzerliği maksimize eden ve sınıflar arası benzerlikleri minimize eden temele dayalı olarak sınıflandırılır ya da gruplandırılır. Böylece, nesnelerin kümeleri bir diğeri ile karşılaştırıldığında bir kümedeki nesnelerin yüksek benzerliğe sahip olduğu fakat diğer kümelerdeki nesnelere çok benzer olmadığı bir biçimde oluşur. Oluşturulan her bir küme kuralların türetilebildiği nesnelerin bir sınıfı olarak görülebilir. Kümeleme ayrıca benzer olayların birlikte gruplandırıldığı sınıfların

gözlemlerin bir hiyerarşi içinde olduğu sınıflandırma olayını kolaylaştırabilir (Han ve Kamber, 2000: 31).

Bu konu bir sonraki bölümde daha detaylı bir şekilde ele alınacaktır.

1.7.3. Birliktelik Kuralı ve Sıralı Örüntüler

Veri madenciliği için birliktelik görevi "birlikte hareket eden" katkıları bulma işidir. İş yaşamında en başta gelen benzerlik analizi ya da Pazar sepeti analizi olarak bilinen birliktelik görevi, iki ya da daha fazla özellik arasındaki ilişkiyi ölçmek için kuralları ortaya çıkarmak için uğraşmaktadır. Birliktelik kuralları kurallarla ilgili olan destek ve güvenin birlikte ölçülmesi biçimidir (Larose, 2005: 46).

Birliktelik kurallarının amacı, büyük veri kümeleri arasından birliktelik ilişkilerini bulmaktır. Depolanan verilerin sürekli artması nedeniyle şirketler, veritabanlarındaki birliktelik kurallarını ortaya çıkarmak isterler. Büyük miktarda depolanan verilerden değişik birliktelik ilişkileri bulmak, şirketlerin karar alma süreçlerini olumlu etkilemektedir (Küçüksille, 2009: 37).

1.8. VERİ ÖNİŞLEME TEKNİKLERİ

Gerçek dünyada veriler fazla olmaları, kayıp olan veriler, yanlış işlenmiş ya da kodlanmış verilerin olması, hatalı ya da sapan değerler içeren gürültülü verilerin olması gibi nedenler dolayısıyla kaliteli ve kullanışlı veri madenciliği sonuçları elde edebilmek için veri madenciliği süreçleri uygulanmadan önce önişleme tekniklerinin uygulanmasına ihtiyaç duyulur.

Analiz sürecinde eldeki verilerle ilk yapılacak şey verilerin ön işlemden geçirilmesidir. Bir veri madenciliği sürecinde önceki çalışmalar yapılan ön işlem sürecinin gerekli çalışmanın %60 kadar kısmını kapsadığı, verilerin ön işlem sürecinden geçirilmesinin ise veri madenciliği projesinin başarısına %75 ila % 90 katkı sağladığı görülmüştür.

İlk olarak verilerin istatistiksel olarak önemli örnekler ya da ilişkiler içermesi gerekmektedir. Bazı durumlarda verilerde anlamlı örnekler olsa bile bu örnekler istenen sonuçları elde etmede diğer veri setlerine göre yetersiz olabilirler. Keşfedilmiş örnekler ayrıca mevcut uygulama için çok spesifik ya da çok genel olmamalıdır. Bu durumda araştırmacı anlamlı bilgi bile sunsa veri seti gürültü içeriyor olabilir (Rao ve diğerleri, 2005: 14).

Veri kalitesi, veri madenciliğinde anahtar bir konudur. Veri madenciliğinde güvenilirliğin artırılması için, veri ön işleme yapılmalıdır. Aksi halde hatalı girdi verileri bizi hatalı çıktıya götürecektir. Veri ön işleme, çoğu durumlarda yarı otomatik olan ve yukarıda da belirtildiği gibi zaman isteyen bir veri madenciliği aşamasıdır. Verilerin sayısındaki artış ve buna bağlı olarak çok büyük sayıda verilerin ön işlemeden geçirilmesinin gerekliliği, otomatik veri ön işleme için etkin teknikleri önemli hale getirmiştir (Oğuzlar, 2003).

Veri ön işleme teknikleri şu şekilde sıralanabilir:

1. Veri Temizleme
2. Veri Birleştirme
3. Veri Dönüştürme
4. Veri İndirgeme

Pek çok işlenmemiş (ham) veri içeren veritabanları ön işlenmemiş, tamamlanmamış ve gürültülüdür.

Örneğin veri tabanları aşağıdakileri içerebilir:

- Kullanılmayan ve gereksiz dosyalar
- Kayıp veriler
- Sapanlar
- Veri madenciliği modelleri için uygun olmayan biçimdeki veriler
- Yaygın görüşe uygun olmayan değerler.

Veri madenciliği amaçlarının daha kullanışlı olabilmesi için veri tabanlarının, veri temizleme ve veri dönüşümü biçiminde bir ön işlemden geçemeye maruz kalmak zorundadır. Veri madenciliği yıllardır ilgilenilmeyen verilerle uğraşır bundan dolayı veriler eskidir, ilişkisizdir ve kolayca kaybolur (Larose, 2005: 46).

1.8.1. Veri Temizleme

Veri tabanında yer alan gürültülü veriler söz konusu olduğunda, istenilen ve doğru analiz sonuçları elde edebilmek için veri tabanının bu verilerden temizlenmesi gerekecektir. Veri temizleme, eksik verilerin tamamlanması, aykırı değerlerin teşhis edilmesi amacıyla gürültünün düzeltilmesi ve verilerdeki tutarsızlıkların giderilmesi gibi işlemleri gerektirmektedir. Herhangi bir değişkene ilişkin eksik değerlerin doldurulması için farklı yollar vardır (Oğuzlar, 2003):

1. Eksik değer içeren kayıt veya kayıtlar atılabilir.
2. Değişkenin ortalaması eksik değerlerin yerine kullanılabilir.
3. Aynı sınıfa ait tüm örneklemler için değişkenin ortalaması kullanılabilir. Örneğin aynı kredi risk kategorisine giren müşteriler için ortalama gelir değeri eksik değerler yerine kullanılabilir.
4. Var olan verilere dayalı olarak en uygun değer kullanılabilir. Burada sözü edilen en uygun değerin belirlenmesi için regresyon veya karar ağacı gibi teknikler kullanılabilir. Örneğin yaşı x, eğitim düzeyi y olan bir kişi için ücret durumu, mevcut verilerden yukarıdaki tekniklerden birinin kullanılmasıyla tahmin edilebilir.

1.8.2. Veri Birleştirme

Çoklu veritabanlarının birleştirilmesi ile eksik veriler oluşmaktadır. Bu eksik veriler sayısal veriler için korelasyon analizi ya da kategorik veriler için ki-kare testi metodu ile tespit edilebilir.

Farklı veri tabanlarından ya da veri kaynaklarından elde edilen verilerin birlikte değerlendirmeye alınabilmesi için farklı türdeki verilerin tek türe dönüştürülmesi yani birleştirilmesi söz konusu olacaktır. Eğer veri madenciliği

uygulaması için bir veri ambarı altyapısı hazırlanmış ise söz konusu veri birleştirme işleminin yapılmış olması gerekmektedir. Ancak böyle bir yapı yoksa söz konusu veri birleştirme işleminin doğrudan veri madenciliğine esas oluşturacak veriler üzerine uygulanması gerekecektir (http://mf.dumlupinar.edu.tr).

1.8.3. Veri Dönüştürme

Veri dönüştürme ile veriler, veri madenciliği için uygun formlara dönüştürülürler. Veri dönüştürme; düzeltme, birleştirme, genelleştirme ve normalleştirme gibi değişik işlemlerden biri veya bir kaçını içerebilir. Veri normalleştirme en sık kullanılan veri dönüştürme işlemlerinden birisidir (Oğuzlar, 2003).

Veri madenciliği uygulamalarında bazı değişkenlerin ortalama ve varyans değerlerinin büyük olması, ortalama ve varyansı küçük olan değişkenlerin analizdeki önemliliklerinin azalmasına neden olabilir. Bu nedenle verilerin veri normalleştirmesi ya da standartlaştırılmadı gibi işlemlerden geçirilerek dönüşüm yapılması gerekmektedir. Veri setinin dönüştürme işleminin yapılması ile elde edilen modelin istatistiksel testlerin dayandığı varsayımlara uyacak şekilde olması olasılığı artar.

Eğer veriler az ya da çok simetrik ise, çok az sapan değeri varsa (ya da hiç yoksa) ve varyans nedensel olarak homojen (reasonably homogeneous) ise veri dönüşümü yapılarak kazanılacak bir şey yoktur. Eğer belirgin olarak çarpık veriler ya da heterojen varyans varsa, veri dönüşümünün bazı biçimleri kullanışlı olabilir. Varyans ve şekilde gerekli düzenlemeleri yapan dönüşümler gereklidir. Ayrıca veriler rapor edileceği zaman, dönüştürülmüş veriler üzerinde tek yönlü varyans analizi gibi istatistiksel bir test yapılması uygun olur. Dönüşüm yapıldıktan sonra verilerin normal dağılım ya da normal dağılıma yaklaşık bir dağılım gösterip göstermediği test edilmelidir.

1.8.4. Veri İndirgeme

Veri indirgeme teknikleri orijinal verilerin bütünlüğünü koruyan, daha küçük hacimli olan veri setlerinin indirgenmiş halini elde etmek için kullanılabilir. Böylece, indirgenmiş veri setlerini madenleme aynı analitik sonuçları üretmede daha etkili olabilir (Han ve Kamber, 2000: 30-31).

Veri indirgeme teknikleri, hacim olarak daha küçük veri kümesini temsil eden fakat orijinal verilerin de bütünlüğünü koruyan indirgenmiş verileri elde etmek için uygulanır (http://www.csun.edu). Elde edilen indirgenmiş veri kümesine veri madenciliği teknikleri uygulanarak daha etkin sonuçlar elde edilebilir.
Veri indirgeme yöntemleri aşağıdaki biçimde özetlenebilir (Oğuzlar, 2003):
1. Veri Birleştirme veya Veri Küpü (Data Aggregation or Data Cube)
2. Boyut indirgeme (Dimension Reduction)
3. Veri Sıkıştırma (Data Compression)
4. Kesikli hale getirme (Discretization)

Veriyi indirgeme aşamasında verilerin çok boyutlu **veri küpleri** biçimine dönüştürmek söz konusu olabilir. Böylece çözümlemeler sadece belirlenen boyutlara göre yapılır. Veriler arasında bir seçme işlemi yapılarak, gereksiz veriler veri tabanından çıkarılır ve boyut azaltılması sağlanabilir. Veri sıkıştırma aşamasında, büyük veri kümelerinin sıkıştırılarak daha az yer işgal etmeleri sağlanır, örnekleme aşamasında ise, büyük veri topluluğu yerine onu temsil eden daha küçük veri kümelerinin oluşturulması amaçlanır. Genelleme verilerin tek tek değil genel kavramlarla ifade edilmesini sağlar (http://mf.dumlupinar.edu.tr).

1.9. VERİ MADENCİLİĞİ SÜREÇLERİ

Bir veri ambarı ilk olarak oluşturulduğunda, veri madenciliği süreci dört temel adıma bölünür; veri seçimi, veri dönüşümü, veri madenciliği ve sonuç yorumlama. Bir veri ambarı madencilik için gerekli olmayabilecek çok çeşitli veriler içerebilir. Veri madenciliğinin ilk adımı olarak hedef veriler seçilir. Örneğin bir Pazar analizinde veri ambarı müşterilerin aldıkları ürün veya hizmetler, müşterilerin

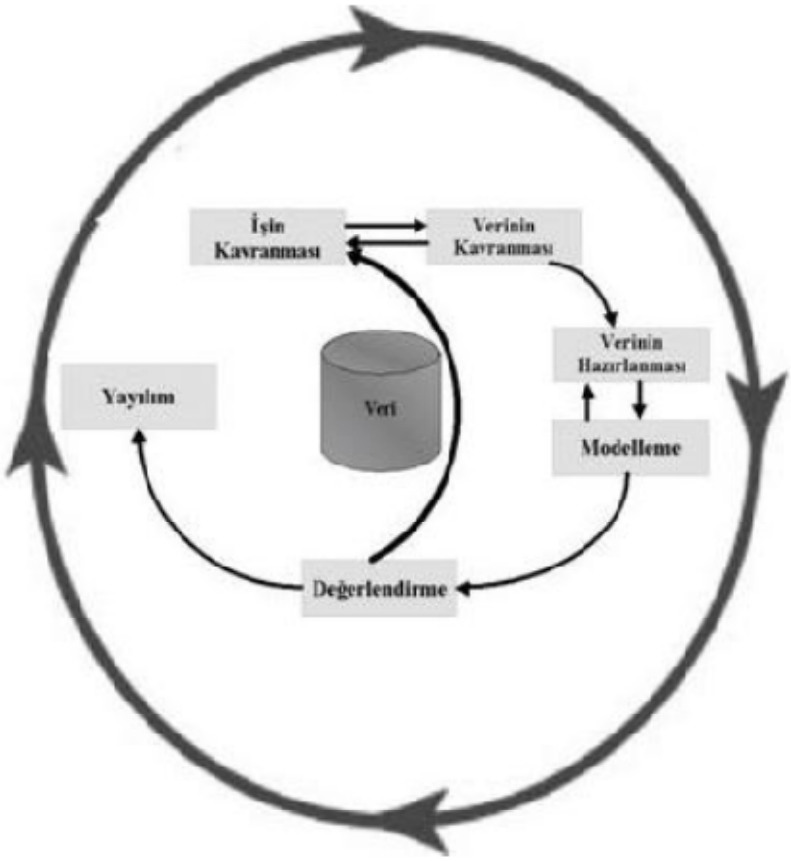
demografik veya yaşam tarzları gibi bilgileri içeriyor olabilir bunları gerekli olmayanları analize dahil edilmemelidir. Veri madenciliği için istenilen veritabanı tabloları seçildikten ve madenlenecek veriler tanımlandıktan sonra, veriler üzerinde dönüşümler yapılmalıdır. Veri dönüşümü verileri istenilen şekilde organize etme ve verileri bir türden başka bir türe dönüştürme (iki özelliğin oranını tanımlama gibi) aşamasıdır. Veri madenciliği aşamasında istenilen bilgi türünün çıkarılması için dönüştürülmüş verilerin bir ya da birden fazla teknikle madenlenmesidir. Sonuç yorumlama aşaması her sonucun en iyi bilgiyi tanımladığı, Madenlenmiş bilginin analiz edilmesidir (Sumathi ve Sivanandam, 2006: 187-188).

CRISP-DM (Çapraz Endüstri Veri Madenciliği Standart Süreci) biçimi veri madenciliği sürecini en iyi ifade edebilen biçimdir. NCR, SPSS ve Daimer-Benz şirketlerinin konsorsiyumu tarafından yaratılmıştır. Bu süreç, önemli aşamaları, genel görevler, özel görevleri ve süreç örneklerini içeren bir hiyerarşiyi tanımlar (Nisbet ve diğerleri, 2009: 35).

Bu sürecin adımları şekilden de görülebileceği üzere (Küçüksille, 2009: 31);

1. **İşin Kavranması** – işletme açısından amaçları anlama ve bu bilgiyi bir veri madenciliği problemine dönüştürme,
2. **Verinin Kavranması** – veri kalitesini belirleme, verinin ilk kez anlaşılmasının keşfi için veri toplamayla başlama,
3. **Verinin Hazırlanması** – son veri setini oluşturmak için tüm faaliyetlerin kapsama alınması,
4. **Modelleme** – değişik modelleme tekniklerinin seçilip, uygulanması ve ayarlanması,
5. **Değerleme** – modelin kalitesinin değerlendirilmesi,
6. **Yayılım** – Karar verme sürecine yardım etmek için “güncel” bir model organizasyonda uygulanmasıdır.

Şekil 2 : CRISP-DM Akış Şeması



Kaynak: http://crisp-dm.org

Veri Madenciliği sürecinin ilk aşaması olan işin kavranması; bir işletmenin bakış açısından proje amaçlarının anlaşılması ve amaçlara ulaşabilmek için bu bilginin bir başlangıç planına ve veri madenciliği problem tanımına dönüştürülmesidir (Küçüksille, 2009: 32).

İşin kavranması aşamasında iş hedeflerini anlamak, durumu değerlendirmek, veri madenciliği amaçlarının tanımı ve bir proje planının türetilmesi gerçekleştirilir (Clos, ve diğerleri 2007: 15).

Veriyi kavrama safhası veri toplanması ile başlar ve veri kalitesi sorunlarını tanımlamak, verinin ilk kavranışını keşfetmek ya da gizli bilgilere ulaşmak için ilginç alt kümeler ortaya çıkarma amaçlı faaliyetlerle devam eder (Küçüksille, 2009: 33). İç veriler toplanır, veriler tanımlanır, veriler aranır ve veri kalitesi doğrulanır.

Veri hazırlama; Bu aşama son veri setini kurmada ihtiyaç duyulacak tüm adımları içerir. Veri hazırlama aşaması veri seçme, veri temizleme, veri yapılandırma, veri dönüşümü, veri altkümelerini biçimlendirme ve özellik seçimi şeklinde gerçekleştirilir (Clos ve diğerleri 2007: 15).

Modelleme şu adımları içermektedir: a. Veriye uygun hale getirilmeye çalışılan modelin seçimi, b. Veriyle ilgili farklı modelleri değerlendiren fonksiyonların seçimi. c. Sonuç fonksiyonunu optimize etmek için algoritmaların ve hesaplama metotlarının belirtilmesi. Bu bileşenler kullanılacak veri madenciliği algoritmasını belirlemek için birleştirilir. Bu bileşenler belirli bir algoritmada önceden de derlenebilirler. Diğer bir ifade ile veri analizi açısından yüksek kaliteye sahip görünen bir ya da daha fazla model oluşturulur.

Modelin yayılma aşamasına geçmeden önce işletmenin amaçlarını tam olarak gerçekleştirdiğinden emin olmak için modelin eksiksiz bir şekilde değerlendirilmesi ve modeli gerçekleştirmek için oluşturulan adımların gözden geçirilmesi önemli bir adımdır. Temel amaç, yeteri derecede dikkate alınmayan bir işletme sorununun olup olmadığını belirlemektir. Bu evrenin sonunda veri madenciliği sonuçlarının kullanımıyla ilgili bir karara ulaşılabilir. Modelin oluşturulması çoğunlukla projenin sonu anlamına gelmemektedir.

Genellikle elde edilen bilginin müşterinin kullanabileceği şekilde düzenlenmesi ve sunulması gerekir. İhtiyaçlara bağlı olarak bir rapor oluşturma kadar basit ya da tekrar edebilen bir veri madenciliği sürecini uygulamak kadar karmaşık olabilir. Birçok durumda yayılma adımlarını gerçekleştirecek olan bir veri analisti değil, kullanıcı olacaktır (Küçüksille, 2009: 35).

1.10. VERİ MADENCİLİĞİ TEKNİKLERİ

İstatistiksel araçlar; Bayes ağları, regresyon ve kümeleme analizi ve korelasyon analizi, gibi çoğu istatistiksel araç veri madenciliği için kullanılmaktadır. Genellikle istatistiksel modeller eğitilmiş veri setinden kurulmuştur. Tanımlanmış istatistik ölçüsüne göre optimal bir model, bir hipotez uzayında aranır. Kurallar, örnekler ve devamlılıklar modellerden çizilir. Bayes ağları değişkenler arasındaki nedensel ilişkileri gösterir. Regresyon, bir çıktı değişkenine nesnelerin özelliklerinin bir kümesini haritalandıran bir fonksiyon türetmedir. Korelasyon analizi her bir değişkenin birbiri ile benzeşimini ifade eder. Kümeleme analizi uzaklık ölçülerine dayalı olarak nesneler kümesinden grupları bulur.

Otomatik öğrenme yaklaşımları; istatistiksel metotlar gibi, otomatik öğrenme metotları test verileri ile eşleşen en iyi modele ulaşır. İstatistiksel metotlardan farklı olarak; arama uzayı, n boyutlu bir vektör uzay yerine n öznitelikli (attributes) bir bilişsel uzaydır. Bununla birlikte çoğu otomatik öğrenme metodu arama (searching) sürecinde sezgileri kullanır. Veri madenciliğinde en çok kullanılan otomatik öğrenme metotları; karar ağaçları, tümevarımsal kavram öğrenme (inductive concept learning) ve kavramsal kümelemedir. Veritabanı odaklı yaklaşımlar; bu metotlar en iyi modeli aramazlar, bunun yerine veri modellemede eldeki verilerin karakteristiklerinden faydalanmada kullanılırlar (Sumathi ve Sivanandam, 2006: 217-218).

Günümüzde sayısal verilerin analiz edilmesinde kullanılan veri madenciliği tekniklerinin yanı sıra, sayısal olmayan verilerin de analiz edilmesi gerekliliği ortaya çıkmıştır. Bu nedenle metinsel verilerin analiz edilmesi için veri madenciliği ile bağlantılı olarak metin madenciliği konusu gündeme gelmiştir. Metin madenciliğinde, metin verilerinin sayısal hale dönüştürülmesinden sonraki analiz aşamaları veri madenciliği ile aynıdır. Tez konusunu oluşturan metin madenciliği kısmı bir sonraki bölümde detaylı bir şekilde ele alınacaktır.

İKİNCİ BÖLÜM
METİN MADENCİLİĞİ

2.1. METİN MADENCİLİĞİ

Metin madenciliği doğal metin dilinden anlamlı bilgi çıkarmayı amaçlayan gelişen yeni bir alandır. Özel amaçlar için gerekli olan bilginin çıkarımı metni analiz etme süreci olarak nitelendirilebilir. Veri tabanlarında depolanan veri çeşitleri ile karşılaştırıldığında, metin yapılandırılmamış, şekilsiz ve algoritmik olarak uğraşılması zordur. Fakat günümüzde metin, bilginin değişiminde resmi bir araçtır. Metin madenciliği alanı, kelime ya da bilgilerin gerçek bağlantısının fonksiyonu olan metin ile ilgilenir.

Delen ve Crossland tarafından tanımlanan, metin madenciliğinin ne yaptığına dair kısa bir özet; *Yani, metin madenciliği ne yapar? En temel seviyede, yapılandırılmamış bir metin dokümanını sayısallaştırır ve sonra, veri madenciliği araçları ve tekniklerini kullanarak, onlardan örnekler çıkarır.* Metin veri madenciliği ve metinsel veri tabanlarında bilgi keşfi olarak da bilinen metin madenciliği, analiz edilen metin kaynaklarında açıkça bulunmayan ortaklıklar, hipotezler ve trendleri çıkarım sürecidir. Metin madenciliği yapısal veri tabanları biçimleri yerine, doğal metin dilinden çıkarılan örnekler nedeniyle veri madenciline göre farklılık gösterir (Nisbet ve diğerleri, 2009: 174).

Metin madenciliğinin amacı yapılandırılmamış (metinsel) bilgiyi işlemek, metinden anlamlı sayısal içerikleri çıkarma ve böylece çeşitli veri madenciliği algoritmaları için (istatistiksel ve otomatik öğrenme) metinde içerilen bilgiye erişebilmektir. Bilgi, dokümanlarda bulunan kelimelerin özetlerinden türetilerek çıkarılabilir. Böylece dokümanlarda kullanılan kelimeleri, kelime kümeleri vs. analiz edebilir, dokümanları analiz edebilir ve aralarındaki benzerlikleri belirleyebilir ya da veri madenciliği projesinde ilgilenilen diğer değişkenlerle olan ilgisini analiz edebilirsiniz. Genel bir deyişle, metin madenciliği yapılandırılmamış öğrenme metotları (kümeleme) uygulaması vs. gibi tahminleyici veri madenciliğindeki gibi diğer analizlere birleştirilebilen “metni sayılara çevirme” (anlamlı içeriklere çevirme) işlemidir (http://www.statsoft.com)

2.2. METİN VE VERİ MADENCİLİĞİ

Veri madenciliğine benzer olarak, metin madenciliği değerli örnekleri ve eğilimleri gösteren kuralları ve belirli başlıklar hakkında önemli özellikleri kurmada, metin dosyalarındaki verileri araştırır. Veri madenciliğinin aksine metin madenciliği metin dokümanlarını yapılandırılmamış ya da yarı yapılandırılmış derlemeleri ile çalışır. Metin madenciliği ilk olarak doküman yığınları arasından anahtar kelime seçimi ile başlar. Erişim makinelerinin binlerce anahtar kelime ve ifadeleri tanımasının yanında bu makineler metnin arkasındaki içeriği analiz etmezler, araştırmacıların belirlediği içerikler ile ilgili anahtar kelimelere taban olan, bilgi kaynağı olarak kullanılacak bir sözlük kurulması gereklidir. Sözlük o zaman organize olmamış metinden anlamlı işaret ve içerikleri çevirmek için kullanılır. Metin erişim sonuçları ile, daha ileri analizler yapabilir ve başarılı bir veri tabanına dönüştürebilir (Lau ve diğerleri, 2005).

Geleneksel veri madenciliği doğal dile dayalı olan metinde, yapılandırılmamış ve yarı-yapılandırılmış yazılı malzemelerin büyük miktarları için yeterli güce sahip değildir. Metin madenciliği metin şeklindeki bilginin büyük miktarlarından kullanışlı bilgi çıkarmada gelişen bir teknolojidir. Veri madenciliği, metin madenciliğinde yeni teknoloji geliştirmek için derin temelleri ve güçlü teknikleri sağlar. Veri madenciliği ve metin madenciliği bazı yönlerden benzerdirler, fakat bazı dikkate değer farklılıkları da vardır. Veri madenciliği iş zekasına çözüm olması için lanse edilmiştir. Perakende sektörü için tüketicilerin harcama alışkanlıklarını veri madenciliği ile inceleyerek yapılan çalışmada satıcıların hangi ürünleri yan yana dizmesi gerektiği hakkında bilgi sahibi olabiliriz. Mesela bir tüketici dijital kamera alıyorsa hafıza kartı, yazıcı veya fotoğraf baskı kağıdı da almak isteyecektir. Metin madenciliği de veri madenciliğinin bir türüdür ve nispeten yeni bir disiplindir. Çoğu yeni araştırma alanı gibi, genel anlaşılabilir bir tanımı yoktur.

Metin madenciliği farklı yazılı kaynaklardan otomatik olarak bilgi çıkararak, metinde önceden bilinmeyen bilginin bir bilgisayar tarafından keşfidir. Fark edileceği üzere, metin madenciliğinin amacı metinsel verilerdeki genel eğilimleri bulma ve potansiyel hileleri tanımlamada yeni, daha önce hiç

karşılaşılmamış bilgiyi bulmaktır. Metin madenciliği bilgi yönetimi, erişimi ve analizine esnek yaklaşımlar sunabilir. Böylece metin madenciliği metinsel malzemelere değinme yeteneği ile veri madenciliğinin kapsamını genişletebilir. Metin madenciliği bir akademik boşluktan ortaya çıkmamıştır fakat ona benzer birkaç teknolojiden gelişmiştir. Bu temel teknolojiler olasılık teorisi, istatistik ve yapay zekaya dayanır (Gao ve diğerleri, 2005).

Metin yazımında standart kurallar olmadığından dolayı bilgisayar bunları anlayamamaktadır. Her bir metnin dili ve içerdiği anlam amaca bağlı olarak çeşitlilik göstermektedir. Yapısal olmayan bilgiden içerik çıkarmak için kullanılan geleneksel yöntemler; anahtar kelimeler veya mantıksal aramalar, istatistiksel veya olasılıksal algoritmalar, sinir ağları ve kalıp keşfedici sistemler gibi dilbilimsel olmayan yöntemlerdir (Dolgun ve diğerleri, 2009: 48-58).

2.3. METİN MADENCİLİĞİNİN TARİHSEL GELİŞİMİ

Manuel emek yoğun metin madenciliği yaklaşımları ilk olarak 1980'lerin ortasında görülmüştür fakat, teknolojik gelişmeler etkin olarak son on yılda hızla ilerleme göstermiştir. Metin madenciliği bilgi çıkarımı, veri madenciliği, otomatik öğrenme, istatistik ve bilişimsel dilbilimi gibi konuları da içeren disiplinler arası bir alandır. En çok bilgi (tahminler %80'den fazla olduğu yönünde) metin olarak depolanmaktadır, metin madenciliğinin ticari potansiyel değerinin yüksek olduğuna inanılmaktadır. H. P. Luhn (1958), otomatik özetleme ile ilgili çığır açan makalesinde birincil metindeki " önemli kelimelerin çözme gücü"ne değinmiştir. Lauren B. Doyle (1961) de metin madenciliğinin ruhunu ve " bilginin doğal tanımlama ve örgütlemesinin frekanslar ve kütüphanedeki kelimelerin dağılımlarının analizinden gelebileceğini" söylediği ilgili metotları yakalamıştır (burada kütüphaneden kasıt genel olarak ana kısım ya da toplanan bilgidir). Don R. Swanson (1988) bilimsel literatürün "araştırma (exploration), korelasyon ve sentez" e layık doğal bir fenomen olarak kabul edilmesi gerektiğini açıkça belirtmiştir (www.datawg.com).

Metin madenciliği ile ilgili ilk bulgular 1960'larada işlenmemiş metinlerin bulunduğu ilk bilgisayar sistemlerinin geliştirilmesi ile başlar. 1980'lerin ortalarına kadar, arama motorlarında "anahtar kelime ile arama" paradigmasına odaklanan sistemlere kadar son kullanıcı deneyimi fazla gelişmemişti. 1990'lara gelmeden yapay zeka ailesinden gelen Doğal Dil İşleme süreci başlayana kadar da ortaya çıkmamıştı. Bu süreçte geliştirilen metotlar günümüzde mevcut metin madenciliği araçlarında hala kullanılmaktadır (Bot, 2007: 3).

2.4. METİN MADENCİLİĞİ UYGULAMA ALANLARI

Metin madenciliği; ulusal güvenlik ve şirket güvenlik uygulamalarında, yasal-avukat-hukuk durumlarında, şirket finansı- iş aklı için, patent analizleri için, halkla ilişkiler- karşılaştırılabilir kurumların, işletmelerin Web sayfalarını karşılaştırma gibi pek çok çeşitli alanda uygulanabilir.

Yapılacak olan bir anket çalışmasında, belirli bir cevap formatında kısıtlamadan cevaplayıcıların görüş ve fikirlerini ifade etmeleri için sorular açık uçlu olarak hazırlanabilir. Böylece uzmanlar tarafından tasarlanmış olan yapılandırılmış sorulardan daha önce keşfedilmemiş müşteri görüş ve fikirleri elde edilebilir. Bir internet sayfası taranabilir, sitede bulunan terim ve dokümanların listesini otomatik olarak çıkarılabilir ve tanımlanmış olan en önemli özellik veya terimler belirlenebilir.

Pazar araştırması; yayınlanmış belgeler, basın bültenleri ve web sayfaları pazar etkisinin ölçülmesi için aranır ve izlenir. Metin madenciliği kantitatif yöntemler ile açık uçlu anket soruları ve mülakatların değerlendirilmesinde kullanılabilmektedir. Müşteri ilişkileri yönetimi (Customer Relationship Management, CRM); bütün müşterilerin email, işlem, çağrı merkezi ve anket gibi erişim noktalarından elde edilen metin bilgilerinden nitelikli bilgi çıkarılır. Bu nitelikli bilgi müşterinin terk etme ve çapraz satışlarını tahmin etmek üzere kullanılır (Dolgun ve diğerleri, 2009: 48-58).

Mesajları, emailleri vs. otomatik olarak işleme: metin madenciliğinin bir diğer olağan uygulaması da metinlerin otomatik olarak sınıflandırılmasına yardım etmektir. Örneğin istenmeyen bir gereksiz postayı içinde geçmesi muhtemel bazı

kelimeleri ya da ifadeleri belirterek istenmeyen bu postadan otomatik olarak filtrelemek mümkündür. Ya da çoğu otomatik sistemde olduğu gibi elektronik mesajlar sınıflandırılarak mesajların yönlendirilmesi gereken departman ya da ajanslara yönlendirilebilir.

2.5. METİN MADENCİLİĞİ İLE İLGİLİ YAZILIMLAR

Hem ticari hem de açık kaynaklı ücretsiz, mevcut bazı metin madenciliği yazılımı kaynakları aşağıdadır;

- SAS-Text Mining
- SPSS-Text Mining and Text Analysis for Surveys
- STATISTICA Text Miner
- GATE-Natural Language açık kaynaklı
- RapidMiner- Word Vector plug-in araçı ile
- R-Language Programming text mining- açık kaynaklı
- Perl İle pratik metin madenciliği- açık kaynaklı
- ODM-Oracle Data Mining
- Megaputer's "TextAnalyst"

Metin madenciliğinde kullanılan bu programlardan en yaygın olarak ilk üç program; Sas- Text Mining, Spss ve Statistica kullanılmaktadır. Uygulama bölümünde de görüleceği üzere, menü ve kullanıcı ara yüzü bakımından kullanım daha rahat kullanım kolaylığına sahip Statistica 9 programı kullanılmıştır.

2.6. METİN MADENCİLİĞİ SİSTEMLERİNİN YAPISI

Bir metin madenciliği sistemi üç ana bileşenden oluşur;

1. Bilgi besleyiciler: çeşitli metinsel yığınların ve modülleri etiketleme arasında bağlantıyı etkinleştiren bir bileşendir. Bu bileşen herhangi bir metinsel kaynak, web sitesi, içsel doküman bileşenleri ile bağlantı kurar.
2. Akıllı etiketleme: metinleri okumak ve ilgili dokümanları etiketlemede sorumlu bir bileşendir. Bu bileşen istatistiksel etiketleme (sınıflama ve terim

çıkarma), semantik etiketleme (bilgi çıkarımı) ve yapısal etiketleme (dokümanların görsel düzenlemesini çıkarma) gibi dokümanlar üzerinde etiketlemenin herhangi bir çeşidini yapabilir.

3. İş zekası devamlılığı: bütün bilgi düzeninden eş zamanlı analizlere izin veren, farklı kaynaklardan bilgiyi sağlamak için sorumlu bileşendir (Ye, 2003: 483).

2.7. METİN MADENCİLİĞİ İÇİN BAZI TEMEL TEKNOLOJİLER

Metin madenciliği, yapılandırılmış ambarda, metin madenciliği ve ilgili iş zekası işlemlerini yürütmek için dokuman ambarında depolanan toplanmış metinlerden bilgi çıkarımının sanat ve teknolojisidir. Metin madenciliğinin uygulanması bazı ilgili disiplinlerin üzerine inşa edilmiştir; Bilgi Gerikazanımı (İnformation Retrieval), Bilişimsel Dilbilim (Computational Linguistics) ve Örnek Tanımlama (Pattern Recognition).

2.7.1. Bilgi Gerikazanımı (İnformation Retrieval)

Bilgi geri kazanımı metin madenciliğinin ilk adımıdır. Bilgi geri kazanımının amacı kullanıcılarının bilgi ihtiyaçlarını karşılayan dokümanları bulmada kullanıcılara yardım etmektir. Arzu edilen bilgi bilinmemektedir ve diğer geçerli bilgi parçaları ile bir arada olur. Bir bilgi geri kazanımı sistemi, bu dokümanlar üzerinde bir ilk tur filtresi gibi hareket eder. Fakat bu bilgilerden bilgi çıkarmak için okuyan bir son kullanıcının metinle ilgili problemlerine cevap vermez. Bilgi çıkarımı, geniş bir alandır ve kullanıcıların belirli konularda dokümanlardan bulabileceği gibi metnin büyük yığınlarını temsil etmeleri için modeller geliştirmiştir. Çoğu Bilgi geri kazanımı teknolojilerinde kullanılan iki temel temsil şemaları vektör uzay modeli ve gizli anlamsal indeksleme (latent semantic indexing)dir.

Vektör uzay modeli (vector space model) temsil edilen doküman ve sorguların maliyetini minimize edebilir. Mümkün dokümanları ve belirli sorguları sırası ile temsil eden iki vektör arasındaki Öklid uzaklığı hesaplayarak belirli bir sorgunun kriterini karşılayan dokümanları bulmada etkili olabilir. Gizli anlamsal

indeksleme vektör uzay modelinin, özellikle eş anlamılık ve çok anlamlılık problemlerinin, bazı sınırlarını dengelemek için geliştirilmiştir. En iyi potansiyel değere sahip veri madenciliği tekniklerinin bazıları çoklu kümeleme süreci içerisinde yatmaktadır (Gao ve diğerleri, 2005).

Gizli anlamsal indeksleme doğal dil işlemede dokümanlar ve dokümanların içerdiği terimler arasındaki anlamsal ilişkilerin analizinde kullanılan bir tekniktir. Gizli anlamsal indeksleme doküman setlerini bir bütün olarak değerlendirir ve aranan kelimelerin geçtiği dokümanların yanı sıra yakın anlamdaki kelimelerin bulunduğu dokümanları da bularak sonuç setini genişletir (www.ce.yildiz.edu.tr).

2.7.2. Bilişimsel Dilbilim

Metin madenciliğinin doğal dile dayalı metinsel bilgi ile ilgilenmesinden bu yana, doğal dili anlamak için doğal dil ve bilgisayarın sınırlı yeteneği arasındaki kritik çatışmayı anlayabiliriz. Bilişimsel dilbilim geri kazanımı kullanıcılara okuyamayacakları kadar büyük miktarda ve gerekli olan bilgiyi gözden kaçırma gibi bir riskin olmadığı, ilgilenilen dokümanları çıkarabilir. Bilişimsel dilbilimde çalışmalar temel analiz aracının bir kümesini elde edilmesine yardım eder. Bu araçlarla, hedeflenen bilgiden daha fazla çıkarmak için metinin yapısını tarar.

2.7.3. Örnek Tanımlama

Örnek tanımlama metindeki önceden tanımlanmış olan dizilere ulaşma sürecidir. Programlama dillerindeki düzenli ifadeler ile eşleşen örneklerin aksine, örnek tanımlamanın bu çeşidi morfolojik ve sözdizimsel olan kelimelerle çalışır. Örnek tanımlamanın iki farklı aşaması kelime ve ifade eşleşmesi ve uygunluk işaretleridir (relevancy signatures). Risk analizi ve değerlendirme aşamasında olduğu gibi, metin madenciliği risk ve ödülün beklentilerine dayanarak eylemin mantıklı kararlarını tanımlayabilir. Metin madenciliği yapılması gereken makul eylemlerin ya da farklı zamanlarda verilen kararların ne olduğunu keşfetmede iyi performans sergiler. Metin madenciliği ayrıca mevcut ve gelecek riski, bir diğeri yerine bir

aksiyonu seçmenin maliyet ya da karını, bir diğeri yerine başka bir karar vermek gibi kriterleri ölçüp biçer (Gao ve diğerleri, 2005).

Geleneksel veri madenciliği teknolojisinin karar desteği sağlayabileceği gibi, metin madenciliği karar destek adımında iyi karar ve stratejileri tanımlamada yarı interaktif yazılımları kullanır. Dahası, metin madenciliği gelecek trend eğilimlerini tahmin eder. Metin madenciliği bilgiyi akıllıca kullanmaya yardımcı olabilir ve pazar değişimleri, zayıf ürün performansı gibi önemli olaylarda kullanıcıların önleyici adımlar atabileceği İş Zekası sistemlerinde kullanıcılara uyarı sağlar. Metin madenciliği kullanıcılara, satış veya pazarlama tatminini ya da personelin morali ve iş dünyasında rekabetçi avantajlar elde etmeyi sağlamada daha iyi iş kararları vermede ve analiz etmede yardımcı olması için tasarlanmıştır.

2.8. METİN MADENCİLİĞİNE YAKLAŞIMLAR

Metin madenciliği metinin "sayısallaştırılması" süreci olarak özetlenebilir. En temel seviyede, her bir dokümanda her kelimenin kaç defa bulunduğunu numaralandıran bir matrisin frekansı gibi doküman ya da kelimenin tablosunu hesaplamak için sayılacak ve işaretlenecek, girdi dokümanında bulunan tüm kelimeler bulunur. Bu temel adımdan sonra "ve", "veya" gibi anlamsız kelimeler ve kelimelerin kök hali ile çekimli halleri "tatil", "tatili" gibi yeniden düzenlenebilir. Bir dokümandan ilk defa kelimelerin tablosu türetildiğinde, kelimelerin ya da dokümanların boyutları ya da kümeleri türetmede ya da ilgilenilen değişkenin diğer çıktısının en iyi tahmincisi olabilecek önemli kelime ya da terimleri tanımlamada tüm istatistiksel ve veri madenciliği teknikleri uygulanabilirler (http://www.statsoft.com).

2.9. METİN VERİLERİNİ SAYISALLAŞTIRMA

Çok sayıda küçük doküman vs. az sayıda büyük doküman: eğer çok büyük olan az sayıda dokümandan "kavram" (concept) çıkarma niyetindeyseniz, o zaman değişken sayısı (çıkarılmış kelime) çok fazla iken durum sayınızın (doküman) çok az olmasından dolayı genel olarak istatistiksel analizler daha az güçlü olurlar.

Belirli karakter, kısa kelime, sayı vb. çıkarmak: harflerin belirli sayısından daha uzun ya da daha kısa olan kelime, karakter sırası, ya da belirli karakterleri çıkarmak, girdi dokümanlarını indekslemeye başlamadan önce yapılabilir. Ayrıca işlenen dokümanda küçük bir yüzdede görülen "seyrek (rare) kelimeleri" de çıkarabilirsiniz.

Listeye alma, listeden çıkarma (durdurma kelimeleri- stop words): sıralanacak kelimelerin belirli bir listesi tanımlanabilir, bu durum belirli kelimelere ulaşmak istediğinizde ve bu kelimelerin görülmesi ile frekansına dayalı girdi dokümanlarını sınıflandırmada kullanışlıdır. Ayrıca "durdurma kelimeleri" sıralamadan çıkarılacak olan terimleri tanımlar. İngilizcede bu kelimeler "the", "a", "of" 'dir.

Eş anlamlılar ve deyimler ve kelime kökleri: "ekmek" gibi eş anlamlılar ve belirli bir anlam ifade eden deyimler sıralama (indexing) için birleştirilebilir. Örneğin "Microsoft Office" bir bilgisayar işletim sistemi olarak deyim olarak tanımlanabilir, veya analizde kullanılan doküman veri setinde listesi çıkarılmış kelimeler tekil veya çoğul olabilirler. Analizi yapılacak kelime sayısını azaltmada kelime köklerini belirleme ve bu kelimelerin çoğul olduğu durumları kökü belirlenmiş olan kelimeye atfetme ile durum karmaşasından kurtulunabilinir. Örneğin "olmak", ile "oldu" kelimeleri metin madenciliği programı tarafından aynı kelime olarak tanımlanır. Burada amaç kelimeleri köklerine indirgemektir.

2.10. KELİME FREKANSLARINI DÖNÜŞTÜRME

Girdi dokümanı bir kere sıralandığında ve ilk kelime frekansları hesaplandığında, çıkarılmış olan bilgiyi özetlemek ve birleştirmek için birkaç ek dönüşümler yapılabilir (http://www.statsoft.com).

2.10.1 Log-Frekanslar (Log-frequencies)

İlk olarak, frekans sayısının çeşitli dönüşümleri yapılabilir. Kelime ya da terim (term) frekansları her bir dokümandaki bir kelimenin nasıl belirgin ya da önemli olduğu üzerinde durur. Özellikle bir dokümanda daha iyi frekansla görülen kelimeler bu dokümanın içeriğinin daha iyi tanımlayıcılarıdır. Kelime hesaplarının,

dokümanın tanımında olduğu gibi, kendi önemlilikleri ile orantılı olduğunu varsaymak mantıklı değildir (reasonable). Örneğin, eğer bir kelime A dokümanında 1 defa, fakat B dokümanında 3 defa görülmüşse, A dokümanı ile karşılaştırıldığında bu kelimenin B dokümanında 3 kat fazla daha önemli bir tanımlayıcı olduğu ile sonuçlandırmak akla yatkın değildir.

Böylece, ham kelime frekansını hesaplamak (wf) için genel bir dönüşüm;

$$f(wf) = 1 + \log(wf)\,, for\ wf > 0$$

Bu dönüşüm ham frekansları ve daha sonraki hesaplamaların sonuçlarının nasıl etkileyeceğini köreltecektir (dampen).

2.10.2 İkili (binary) Frekanslar:

Bir dokümanda kullanılan bir terim olup olmadığını numaralandırmada kullanılabilecek daha basit bir dönüşüm;

$$f(wf) = 1, for\ wf > 0$$

Kelimeler matrisin tarafından doküman sonuçları ilgili kelimelerin olması ya da olmamasını göstermek için sadece 1 ve 0 değerlerini içerecektir. Bu bilgi daha sonraki hesaplamalar ve analizler üzerinde ham frekans hesaplarının etkilerini köreltecektir.

2.10.3 Ters Doküman Frekansları:

Farklı kelimelerin bağıl (relative) doküman frekanslarıdır(df). İleriki analizlerde kullanılacak olan endeksleri (indices) daha dikkatli göz önünde bulundurmak ve yansıtmak isteyebilirsiniz. Örneğin; “tahmin” gibi bir terim tüm dokümanlarda sık sık, diğer bir terim “yazılım” ise sadece çok azında görülebilir.

Görünürlülüklerinin genel frekanslarının (kelime frekansları) yanı sıra kelimelerin belirliliğinin (doküman frekansları) mevcut ve çok kullanışlı dönüşümleri sözü edilen ters doküman frekansıdır (i.inci kelime ve j.inci doküman).

$$idf(i,j) = \begin{cases} 0 & if\ wf_{i\,j} = 0 \\ \left(1 + \log\left(wf_{i\,j}\right)\right) log\,\frac{N}{df_i} & if\ wf_{i\,j} \geq 1 \end{cases}$$

Bu formülde N tüm doküman sayısıdır ve dfi i.inci kelime (bu kelimeyi içeren doküman sayısı) için doküman frekansıdır. Bundan dolayı, bu formülde hem log fonksiyonu (log function) üzerinden basit kelime frekanslarının azaltmayı (dampening) hem de tüm dokümanda kelime görüldüğünde 0 değeri veren *(log(N/N=1)=0)*, ve bir kelimenin tek bir dokümanda görüldüğünde *(log(N/1)=log(N))* maksimum değeri veren bir ağırlık faktörünü içerir. Bu bilginin hem kelimenin görünürlülüğünün bağıl frekanslarını (relative frequencies) yansıtan hem de analizde mevcut olan dokümanların üzerinde anlamsal özelliklerinin indislerini nasıl yaratacağı kolaylıkla görülebilir

2.11. TEKİL DEĞER AYRIŞIMI İLE ÖRTÜK ANLAMSAL ENDEKSLEME (Latent Semantic Indexing)

Daha önce bahsedildiği gibi, girdi dokümanlarında bulunan kelimelerin ilk endekslemesinin temel sonuçları her bir dokümanda görülen farklı kelimelerin kaç defa görüldüğü gibi, basit hesaplama ile bir frekans tablosudur. Genellikle, bu ham hesaplar, kelimelerin "önemini" ya da onların girdi dokümanlarının kümesinin anlamsal özelliğini daha iyi yansıtacak olan endekslere (indices) dönüştürülür.

Çıkarılacak kelimeler ve böylece analiz edilen dokümanlar tarafından tanımlanacak "anlam" ya da "anlamsal uzay"ı yorumlamak için ortak bir analitik araç, kelime ve dokümanların ortak uzaya, hesaplanan kelime frekansları ya da dönüştürülmüş kelime frekanslarından (ters doküman frekansları gibi) bir haritasını (mapping) yaratmaktır. Bu genelde şu şekilde çalışır: yeni arabaları olan müşteri yorumlarının bir toplamını endekslediğinizi varsayalım. Her seferinde bir yorumun "gaz-kilometre" kelimesini içerdiğini ve ayrıca "ekonomi" terimini de içerdiğini

bulabilirsiniz. Daha sonra, raporlar “güvenirlilik” kelimesini içerdiğinde ayrıca “kusurlar” terimini de içerir. Ancak, bazı dokümanların en az birini ya da ikisini de içerdiği “ekonomi” ve “güvenirlilik” terimlerinin kullanımı ile ilgili tutarlı örnek yoktur. Diğer bir deyişle, bu dört kelime “gaz-kilometre” ve “ekonomi”, “güvenirlik” ve “kusurlar”, iki bağımsız boyutu ifade eder. İlki aracın tüm kullanma maliyeti ile ilgili, diğeri ise kalite ve işçilik ile ilgilidir. Gizli anlamsal endeksleme ile ilgili fikir, kelimelerin ve dokümanların haritalandırılabileceği her temel boyutu (anlamı) tanımlamaktır. Sonuç olarak, girdi dokümanındaki tanımlanan ya da tartışılan temel (underlying) (gizli) konuları tanımlayabiliriz ve ayrıca çoğunlukla ekonomi, güvenirlilik ya da her ikisi ile ilgili olan dokümanları tanımlayabiliriz. Bunun sonucu olarak, çıkarılmış kelime ya da terimleri ve girdi dokümanlarını ortak gizli anlamlar uzayına haritalandırmak isteyebiliriz (http://www.statsoft.com).

Tekil değer ayrışımı; bir matrisin çarpanlarına ayrılma türlerinden biridir. Lineer cebirde tekil değer ayrışımı dikdörtgen biçiminde gerçek (rectangular real) veya karmaşık matrislerin, sinyal işleme (signal processing) ve istatistiğin pek çok uygulamalarıyla, çarpanlara ayrımıdır. Tekil değer ayrışımını kullanan pseudoinverse (kare olmayan bir matris'in bir vektör'ün boyutunu azaltmak için birleşik hızlarla birlikte kullanılıp tersine döndürülmesi) uygulamalar, verilere uygunlukta en küçük kareler, matris tahminlemeleri ve bir matrisin sırası (rank) aralığı (range) ve boş uzayını (null space) tanımlamayı içerir. Doğrusal bir modeli optimize etmede etkili bir algoritmadır. (http://en.wikipedia.org).

Değişkenleri ve durumları (gözlemleri) ortak bir uzaya çıkarmada tekil değer ayrışımı, daha çok benzerlik analizinde (correspondence anaysis), çeşitli istatistiksel tekniklerde kullanılabilir. Bu teknik ayrıca temel bileşenler analizi ve faktör analizi ile de yakından ilgilidir. Genelde, bu tekniğin amacı girdi matrisinin tüm boyutluluğunu (çıkarılmış kelimelerin sayısına göre girdi dokümanlarının sayısı) mümkün olduğunca değişkenliği (kelime ve dokümanlar arasında) her ardışık boyutu temsil eden daha düşük bir boyutluluk uzayına azaltmaktır. En iyi şekilde, kelimeler ve dokümanlar arasındaki değişkenliğin (farklılık) çoğu için hesaplanan, iki ya da üç en belirgin boyut tanımlayabilirsiniz ve böylece, analizdeki kelimeleri ve dokümanları organize eden gizli anlamsal uzayı tanımlarsınız. Bir şekilde, bir kez her

boyut tanımlanabildiğinde, dokümanlarda neyin bulunduğunun (tartışıldığı, tanımlandığı) temel "anlam"ı çıkarılmış olur (http://www.statsoft.com).

2.12. METİN MADENCİLİĞİ İÇİN ÖZELLİK SEÇİMİ

Son zamanlarda internette bulunan dokümanların sayısında muazzam bir büyüme vardır. Yapılandırılmamış veriler ile bu dokümanlar çevrimiçi olarak depolanmış baskın veriler haline gelir. "Bag-of-words" yaklaşımı şimdilerde metinsel dokümanları analiz etmede kullanılır. Bu yaklaşımda, bir doküman kelime ya da ifadelerin bir kümesi olarak kabul edilir. Terimler (kelimeler ya da ifadeler) özellikler olarak kabul edildiğinde, metinsel dokümanlara veri madenciliği tekniklerini uygularken, bir doküman bir örnek olarak kabul edilir.

Metinsel verilere etkin bir biçimde uygulanabilecek olan birçok özellik seçimi yaklaşımı mevcuttur. Bunların çoğu terimlerin bir puanlama tablosuna dayalıdır. Özelliklerin puanı, doküman veri setindeki ifadelerin kalitesini temsil eder. Yüksek puanlı bir terim onun önemli olduğu ya da veri seti ile ilgili olduğu anlamına gelir. Denetimli yaklaşımlarda, terim puanları sınıf bilgisi olan, etiketli eğitim setine dayanır. Bazı popüler denetimli özellik seçimi yaklaşımları bilgi kazanma (information gain, IG), karşılıklı bilgi (mutual information, MI) ve $\chi 2$ istatistiği (CHI)'dir. Denetimsiz özellik seçimi yaklaşımları bir veri setindeki terimlerin kalitesini tahminlemede sezgilere dayanır. Metinsel dokümanların bir veri seti için, sezgisellik genel olarak veri seti boyunca ifadelerin dağılımına odaklanır. Popüler olan denetimsiz özellik seçimi yaklaşımları, doküman frekansı ve terim gücü (term strenght, TS)'dir. Doküman frekansı basit fakat özellik seçimi için etkili bir ölçüdür. Bir terimin doküman frekansı terimin görüldüğü dokümanların sayısıdır. Özellik seçimi yaklaşımı her terim için doküman frekansını hesaplar ve önceden tanımlanmış eşik değerinin altındaki doküman frekanslarını siler. Temel varsayım sık terimlerin daha önemli olduğu ve sık olmayanlarla karşılaştırıldığında veri seti ile daha ilgili olduğudur. Terim gücü durdurma kelimelerinin kaldırılması için önerilir. Bu yaklaşım, "yakından ilgili" dokümanların nasıl muhtemel görüldüğünü bir terime dayalı gücü tahminler. Yaklaşımın iki adımı vardır, 1) benzer doküman eşleşmelerini

bulma ve terim gücünü hesaplama. Benzer dokümanların eşleşmelerini bulma adımı iki doküman vektörünün kosinüs değerini kullanarak veri setindeki dokümanların tüm çiftleri arasındaki benzerlikleri hesaplar. Eğer ön tanımlanmış eşir değerinin üzerinde bir değere sahipse veri setindeki iki doküman "benzer" olarak tanımlanır. Terim gücünü hesaplama adımında bir terimin gücü bir terimin bir dokümanda görüldüğü ardından da diğer dokümanda görüldüğündeki şartlı olasılık tahmin edildiğinde hesaplanır. Denetimsiz özellik seçimi yaklaşımları etiketlenmiş verilerin maliyetini korur ve denetimli süreçteki eğitilmiş ve test veri seti arasındaki homojenliğin yanlışlığı problemini önler. Bu özellik, çeşitli başlıklardaki dokümanların büyük miktarları ile ilgilenen metin madenciliği görevleri için özellikle önemlidir (Do ve diğerleri, 2006).

Özellik seçme süreci veri madenciliğinde veri hazırlamayı takiben çok önemli bir stratejidir. Veri madenciliğinin en önemli problemi birçok potansiyel tahminleyici ile büyük veri setlerindeki *boyutluluk lanetidir* (curse of dimensionality). Bu durum ilk olarak bir modele eklenen değişkenlerin artması problemini tanımlamak için Richard Bellman (1961) tarafından bulunmuştur. Bir modele eklenen ek değişkenler regresyon modellerini bir parça daha iyi tahminleyebilir ya da bir sınıflandırma modelinde sınıflar arasında da iyi ayrım yapabilir. Bu çözümlere *yakınsama*nın sorunu ya hata minimizasyonu sürecini ya da döngülü öğrenme süreci ek değişkenler olarak yavaş artarak analize eklenir. Özellik seçimi modeldeki değişkenleri azaltmayı amaçlar, bu sebeple alakasız veya gereksiz değişkenler veya gürültülü veriyi kaldırarak lanetin (curse) etkisini azaltır. Analiz için aşağıdaki pozitif etkileri vardır (Nisbet ve diğerleri, 2009: 77);

- Algoritma sürecini hızlandırır
- Veri kalitesini arttırır
- Algoritmanın tahminleyici gücünü arttırır
- Sonuçları daha anlaşılır yapar

2.13. BİRLİKTELİK KURALLARI

Birliktelik Kuralları büyük veri kümeleri arasındaki ilginç ilişkileri veya korelasyonları bulmak için kullanılır. Birliktelik Kuralları, verilen veri kümesi içindeki sıkça görülen özellik değer durumlarını tespit eder (Şen, 2008: 22).

Birliktelik algoritmaları basit kategorik değişkenler, ikili (dichotomous) değişkenler ve/ve ya çoklu hedef değişkenleri analiz etmede kullanılabilir. Algoritma veride ya da herhangi birliktelikte (ön tanımlı varyant hariç) mevcut farklı kategorilerin sayısını belirtmenizi gerektirmeden birliktelik kurallarını belirleyecektir. Bir çapraz çizelgeleme (corss-tabulation) biçimi değişkenler ya da kategorilerin belirlenmiş sayılarına ihtiyaç duymadan yapılandırılabilir. Böylece, bu teknik çok büyük veri setlerinin analizi için özellikle çok uygundur (Nisbet ve diğerleri, 2009: 126).

Birliktelik kuralları için verilebilecek örnek market sepeti uygulamasıdır. Bu kural, müşterilerin satın alma alışkanlıklarını analiz etmek için, müşterilerin satın aldıkları ürünler arasındaki ilişkileri bulur. Bu tür ilişkilerin analizi sonucunda, müşterilerin satın alma davranışları öğrenilebilir ve yöneticiler de öğrenilen bilgiler sonucunda daha etkili satış stratejileri geliştirebilirler. Örneğin bir müşterinin süt ile birlikte ekmek satın alma olasılığı nedir? Elde edilen müşteri bilgilerine dayanarak rafları düzenleyen market yöneticileri ürünlerindeki satış oranlarını arttırabilirler. Örneğin bir marketin müşterilerinin sütün yanında ekmek satın alma oranı yüksekse, market yöneticileri süt ile ekmek raflarını yan yana koyarak ekmek satışlarını arttırabilirler (Küçüksille, 2009: 37).

Aşağıdaki şekilde iki kelime arasındaki birliktelik kurallarına ait bir çıktı sonucu gösterilmektedir

Şekil 3: Birliktelik Kuralı Çıktı Örneği

Summary of association rules (Scene 1.sta)
Min. support = 5.0%, Min. confidence = 5.0%, Min. correlation = 5.0%
Max. size of body = 10, Max. size of head = 10

	Body	==>	Head	Support(%)	Confidence(%)	Correlation(%)
154	and, that	==>	like	6.94444	83.3333	91.28709
126	like	==>	and, that	6.94444	100.0000	91.28709
163	and, PAROLLES	==>	will	5.55556	80.0000	73.02967
148	will	==>	and, PAROLLES	5.55556	66.6667	73.02967
155	and, you	==>	your	5.55556	80.0000	67.61234
122	your	==>	and, virginity	5.55556	57.1429	67.61234
164	and, virginity	==>	your	5.55556	80.0000	67.61234
121	your	==>	and, you	5.55556	57.1429	67.61234
73	that	==>	like	6.94444	41.6667	64.54972
75	that	==>	and, like	6.94444	41.6667	64.54972
161	and, like	==>	that	6.94444	100.0000	64.54972

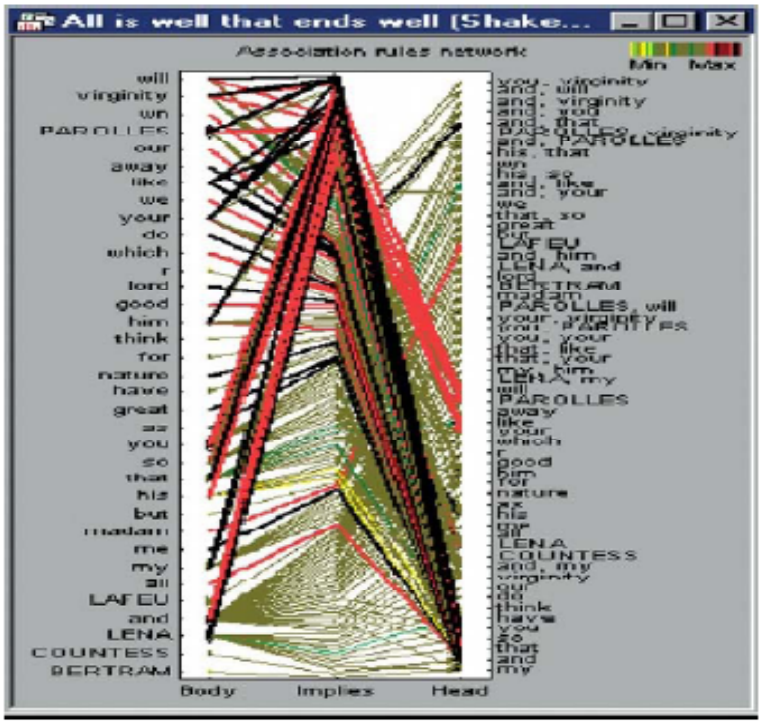
Kaynak : Nisbet ve diğerleri, 2009, s. 127.

Şekilden de görüldüğü gibi ortaklık kuralları basit kategorik değişkenler ve ikili veya çoklu hedef değişkenleri analiz etmede kullanılabilirler. Değişkenler arası destek, güvenirlilik ve korelasyon değerlerinin gösterildiği bu çapraz tablolama şekli çok büyük veri setlerinin analizinde kullanılabilirler. Destek (support) değeri iki değişkenin birlikte görülme olasılığını gösterir, şekilden de görülebileceği üzere ilk satırdaki destek değeri analizi yapılan metinde “and, that ve like” kelimelerinin birlikte görülme olasılığı %6.9 olarak ifade edilmektedir.

Güvenirlilik (Confidence) değeri ise bir durum meydana geldiğinde diğer bir durumun da birlikte görülme olasılığını ifade eder. Yukarıdaki tablodan da görülebileceği gibi ilk satırda “and, that” kelimelerinin geçtiği bir cümlede “like” kelimesinin de geçme olasılığı %83.3 olarak bulunmuştur.

Benzer şekilde birliktelik kuralları ile ilgili görsel olarak incelenebileceği seçenekler de vardır, aşağıdaki şekiller statistica programından alınmış olan birliktelik kuralları ile ilgili grafiklerdir.

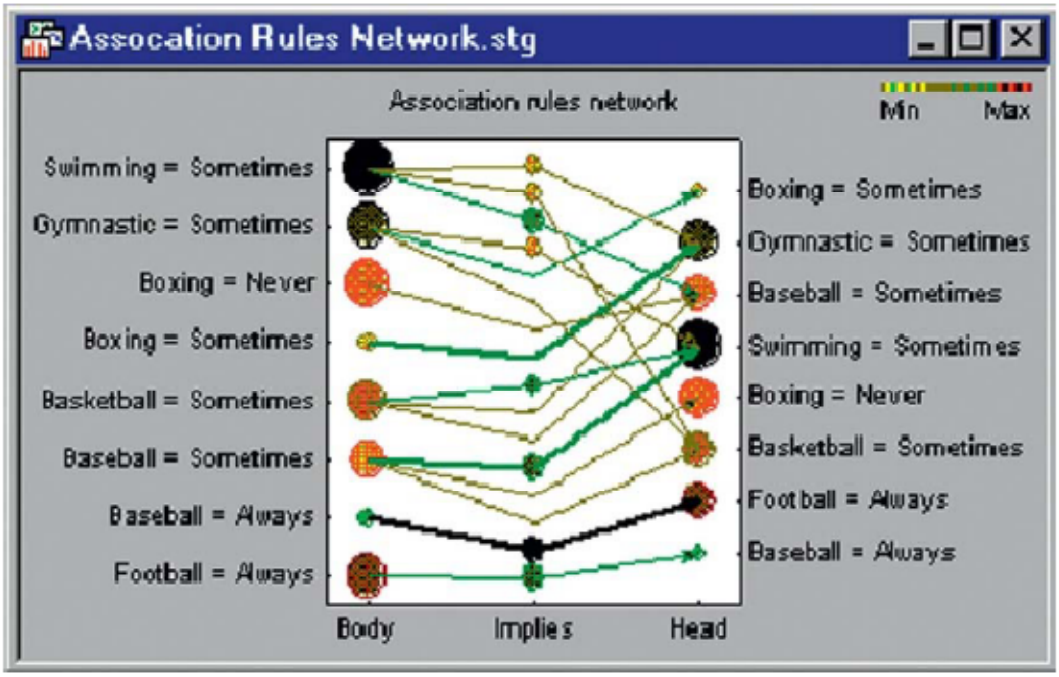
Şekil 4: Statistica Programında Birliktelik Kuralı Çıktı Örneği-1



Kaynak : Nisbet ve diğerleri, 2009, s. 128.

Yukarıdaki şekilden de görüldüğü gibi ortaklık kuralları görsel şekilde sunulmak istendiğinde şeklin her iki tarafındaki kelimeler ve bu kelimelerin birbirleri ile ilişki derecelerine bağlı olarak farklı renklerde ve kalınlıkta çizgilerle aralarındaki ortaklık kurallarını görmek mümkündür.

Şekil 5: Statistica Programında Birliktelik Kuralı Çıktı Örneği-2



Kaynak : Nisbet ve diğerleri, 2009, s. 128.

Ortaklık kurallarının görsel olarak ifade edilmesinde kullanılan bir diğer şekil de yukarıdaki gibidir. Her ortaklık kuralının "vücut" ve "başı" için destek değerleri her birinin büyüklük ve renkleri tarafından gösterilir. Şekilde her çizginin kalınlığı ilgili ortaklık kuralı ile ilgili güven değerini (verilen vücuttan kafanın şartlı olasılığı) gösterir. Ortadaki dairelerin büyüklükleri ve renkleri, İmalar etiketinin (implies label) üstünde, ilgili ortaklar kurallarının bileşenleri ilgili baş ve boyunun ortak desteğini (joint support) (ortak olaylar [co-occurances]) gösterir.

2.14. TEMEL BİLEŞENLER ANALİZİ

Çok sayılı değişkenlerin sayısını daha küçük setlere azaltma yeteneği kullanışlıdır. Bu özellikle bir metinde hesaplanacak çok sayıda dilbilimsel girdilerle uğraşılıyorsa zaman kaybını önlemek ve hata payını azaltmada oldukça faydalıdır. Bunu yapmak çoğunlukla bilgi kaybına neden olur. Burada amaç, değişken azaltma ve bilgi kaybı arasında kabul edilebilir bir takas elde etmektir.

Temel bileşenler analizi mümkün olduğunca çeşitliliği koruyarak değişken sayısını azaltır. Bu yaklaşım, çoğu durumda akla yatkın olan, verilerin kullanışlı olduğu kısımdaki değişkenliği varsayım olarak kullanır.

Temel bileşenler analizi yaklaşımı basittir, örneğin elimizde x_1, x_2, …, x_n değişkenli bir veri setimiz olduğunu varsayalım. c_1, c_2,…, c_n olarak adlandırılan n tane değişken, aşağıdaki dört şartı karşılamak için düzenlenmiştir. İlk olarak her bir bileşen c_i orijinal x_i değişkenlerinin doğrusal bir fonksiyonudur.

$$c_1 = e_{1\,1}x_1 + e_{1\,2}x_2 + \cdots + e_{1\,n}x_n$$

$$c_2 = e_{2\,1}x_1 + e_{2\,2}x_2 + \cdots + e_{2\,n}x_n$$

$$c_3 = e_{3\,1}x_1 + e_{3\,2}x_2 + \cdots + e_{3\,n}x_n$$

$$\cdot \cdot \cdot$$

$$c_n = e_{n\,1}x_1 + e_{n\,2}x_2 + \cdots + e_{n\,n}x_n$$

$$c = Ex$$

İkinci olarak, c vektörü birim uzunluğa sahiptir. Bu da;

$$c^T c = 1$$

Üçüncü olarak her bir c_i ve c_j ($i \neq j$) ilişkisizdir (uncorrelated). Böylece c'nin korelasyon matrisi birim matristir, I. Dördüncü olarak, c_1, c_2,…, c_n varyansları büyükten küçüğe sıralanmıştır. Bu dört durum tek olarak E matrisini belirtmektedir. Temel bileşenler analizi bir veri setindeki güçlü tahminleyici değişkenleri tanımlamada kullanılır. Temel bileşenler analizi, temel bileşenlerin bir grubunu tanımlayarak bir veri setindeki değişkenler arası ilişkiyi açıklar. Bu temel bileşenler verilen bir çıktı (ya da hedef değişken) ile ilgili olan girdi değişkenlerinin belirli kombinasyonlarının dönüşümünden oluşmaktadır. Her bir temel bileşen ham veri setindeki varyasyonların azalan bir miktarını açıklar. Sonuç olarak, ilk birkaç temel bileşen veri setinin alt yapısının çoğunu ifade eder. Temel bileşenler veri setindeki

ham verilerin sayısını azaltma çalışmalarında kullanılmaktadır. Bu gerçekleştirildiğinde, orijinal verilerin yerini ilk birkaç temel bileşen alır. Sonuç olarak, temel bileşenler hedef değişken ile ilgili olan sınıf farklılıklarını gizleyebilir. Klasik temel bileşenler analizi; verilerde homojenliğin bozulması durumunda (yani sapan değerlerin varlığında) sağlıklı sonuçlar vermemektedir. Bu durumda sapan değerlere karşı dayanıklı olan tahminciler kullanılarak analizin yapılması gerekmektedir (Bilisoly, 2008: 240).

Temel bileşenler analizinin ana konusu, en büyük varyans ile boyutları toplayarak boyutluluk azaltmaktır. Matematiksel olarak, tekil değer ayrışımı aracılığı ile en düşük sıraya yaklaşıma eşittir. Bu gürültü azaltma özelliği tek başına Temel Bileşenler Analizinin etkinliğini açıklamada yetersizdir. Temel bileşenler analizi daha sonra pazarlamada daha fazla uygulanabilecek sıralama ve kümeleme uygulamalarında sosyal ağ madenciliğinde temel bir yöntemdir (Santhanalakshmi ve Alagarsamy, 2011: 193-198).

Popüler çok değişkenli istatistik tekniklerinden biri olan temel bileşenler analizi, değişkenler arasındaki bağımlılık yapısının yok edilmesi ve/veya boyut indirgeme amacını taşımakta; başlı başına bir analiz tekniği olduğu gibi, başka analizler için veri üreten yardımcı bir teknik olarak da kullanılmaktadır. Temel bileşenler analizinde p sayıda başlangıç değişkenine karşılık elde edilen p sayıda temel bileşenin her biri, orijinal değişkenlerin doğrusal bir bileşimidir. Dolayısıyla, her bir temel bileşen bünyesinde tüm değişkenlerden belirli oranda bilgiyi barındırır. Bu özelliği sayesinde Temel bileşenler analizi, p boyutlu veri seti yerine, ilk m önemli temel bileşenin kullanılması yoluyla boyut indirgemesi sağlayabilmektedir. İlk m temel bileşen toplam varyansın büyük kısmını açıklıyorsa, geriye kalan p-m temel bileşen ihmal edilebilir. Temel bileşenler analizi bu yönüyle, başka analizlerle birlikte kullanıldığı hallerde, gerek diğer analizin öncesinde değişken sayısının azaltılması ya da bağımlı değişkenlerden bağımsız yeni değişkenlerin türetilmesi amacıyla, gerekse diğer analizin sonucunda elde edilen çok sayıdaki çözüm kümesini daha az boyutta ya da kavramsal anlamlılığı ortaya çıkarmak üzere kullanılabilmektedir (Yıldırım, 2010: 141-153).

2.15. FAKTÖR ANALİZİ

Faktör analizi daha sonraki kümeleme dönemi için bir içerik bağımlı önemsiz kelime filtreleyici olarak kullanılmaktadır. Bir veri tabanından çıkarılan kelimelerden bir faktör matrisi türetilir ve tüm faktörlerin içerisinden en düşük faktöre sahip olan kelimeler belirlenir ve elenir. En az bir faktör için en yüksek yüklü faktör değerleri ve bu faktörün içeriğinin tanımlanmasında etkili olan kalan kelimeler, kümeleme algoritması için girdileri oluşturur. Hem kantitatif hem de kalitatif analizler göstermektedir ki, faktör matris filtrelemesi yüksek kaliteli kümeler ve sonraki sınıflandırmalar için yol göstericidir. İçerik bağımlı önemsiz kelimeleri kümeleme için bir yol faktör analizidir. Son zamanlarda metin veritabanlarındaki korelasyonları ve sonradan bu konuların gruplarını belirlemede faktör analizi kullanılır. Faktör analizi, özellikle hiyerarşik kümeleme metotlarında, konular arasındaki yapısal ilişkilerin iyi bir tahminini sağlamaya dayanan kümeleme analizlerinden, bileşenlerin konuları arasındaki kantitatif ilişkileri iyi bir tahminini sağlamaya dayanır. Faktör ve kümeleme analizi algoritmaları yıllar önce çıkarılmış ve geçerliliği onaylanmıştır. Faktör matris filtreleme yaklaşımı ilk olarak faktör analizinin kullanıldığı ham metinden yüksek teknik içerikli kelimeleri tanımlar ve kalanları önemsiz olarak atar. Bir metin veri tabanındaki faktör analizi bir sistemdeki kelime/ifade sayılarını azaltmayı amaçlar ve kelime/ifadeler arasındaki ilişkiyi belirlemeye çalışır. Kelime/ifade korelasyonları hesaplanır ve yüksek korelasyonlu gruplar (faktörler) tanımlanır. Bu kelime/ifadelerin ilişkileri çıkan faktörler satırları kelime/faktörler olan ve sütunları faktörler olan faktör matrisinde görülebilir. Faktör matrisinde, matris elemanları M_{ij} faktör yükleridir, ya da kelime/ifade i'nin faktör içeriği olan j'ye katkısıdır. Her bir faktörün içeriği faktör yüklerinde en büyük değerlere sahip olan bu kelime/ifadeler tarafından belirlenir. Her bir faktör pozitif değer kuyruğuna ve bir negatif değer kuyruğuna sahiptir. Her bir faktör için kuyruklardan biri mutlak değer büyüklüğüne hakimdir. Bu baskın (dominant) kuyruk her bir faktörün ana temasını (central theme) belirlemede kullanılır.

Faktör analizinin anahtar zorluklarından biri seçilecek faktörlerin sayısını belirlemektir. Literatürde farklı yaklaşımlar önerilmiştir, fakat çoğunlukla kullanılan iki tanesi; Kaiser kriteri ve Scree testidir . Kaiser kriteri tüm değerlerden ziyade

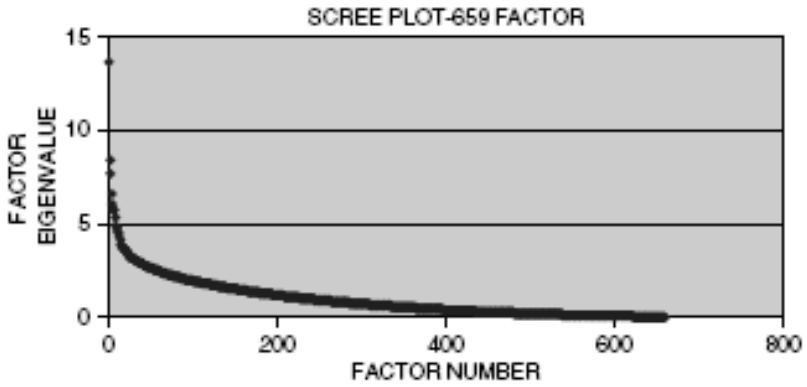
özdeğer faktörleri üzerinde durur. Esasında bir faktör bir orijinal değerin varyansına eşit olan en az bir faktöre gerek duyar. Scree testi ise faktör numarasına karşın faktör özderğerini (varyansını) çizelgeler ve koruncak asıl varyansın çıkarılmasını sağlayan faktörleri önerir. Scree plotun yorumlanması kısmen sübjektiftir ve farklı yorumlayıcılar asarındaki tutarlılık düşüktür. Daha önceki çalışmalarda Kaiser kriteri faktör matrisindeki faktörlerin sayısını seçmek için kullanılmıştır. Faktör analizi uygulanacak olan bir çalışmada kullanılacak doküman veri setinden önce tüm kelimeler çıkarılır daha sonra bu kelime grupları içinden önemsiz kelimeler veri setinden çıkarılır. Önemsiz olmayan kelimeler, her bir kelimenin görünürlülüğüne (kelime frekansı) göre dokümanlar veri tabanından manuel olarak konunun uzmanı olan kişiler tarafından çıkarılır. Bu çalışmadan sonra elde kalan kelimeler faktör analiz yapılabilmesi için hazır hale gelir. Aynı dokümanda bulunan kelime çiftlerinin eş oluşumu (co- accurance) hesaplanır ve kelime çiftlerinin korelasyon matrisi türetilir. Tüm değişkenler çarpanlarına ayrılır (factorized) ve bir faktör matrisi türetilmiş olur (Kostoff ve Block, 2005).

Faktör analizi, p tane değişkene sahip bir durumda birbiri ile ilişkili olan değişkenleri birleştirerek faktör adı verilen az sayıda ortak ilişkisiz değişken bulmayı amaçlayan çok değişkenli bir tekniktir. Temel bileşenler analizi gibi boyut indirgeme ve değişkenler arasındaki bağımlılık yapısını yok etme özelliğine sahiptir. Faktör analizinde de öncelikle değişkenlerin korelasyon matrisi belirlenir. Analizde kullanılan değişkenler arasında korelasyon mutlak değer olarak 0.40 dan daha az ise, ilgili değişkenin analize dahil edilmesi uygun olmayabilir. Bu teknikte, faktörlerin belirlenmesinde birçok yöntem olmasına rağmen en kullanışlı olanı temel bileşenler yaklaşımıdır. Bu yaklaşım, bütün değişkenlerdeki maksimum varyansı açıklayacak faktörü hesaplar. Geriye kalan maksimum miktardaki varyansı açıklamak üzere ikinci faktör hesaplanır. Bu durum, değişkenlerdeki toplam varyansı açıklayıncaya kadar devam eder (www.ekonometridernegi.org)

2.15.1. Faktör Matrisi Türetme

Faktör sayısını belirlemede şekildeki Scree plotun dirseği fikir edinmemizi ve uygun faktör matrisi türetilmesini sağlar. Bu işlemden sonra faktör matrisi, mevcut analiz için önemli teknik kelimeleri tanımlamada bir filtre olarak kullanılabilir. Faktör matrisi, bir kümeleme algoritmasının girdileri için kaynak-bağımlı (content dependent) yüksek teknik kelimeleri seçmek için basit bir önemsiz kelimeler listesini tamamlayabilir. Faktör matrisi ön filtreleme uygulama kaynağında olan önemsiz kelimeleri eleyerek kümelemenin bağımsızlığını geliştirebilir. Temel bileşenler analizi tarafından türetilen her bir temel faktör (özdeğer) için varyans hesaplanır. Aşağıdaki şekil doğrusal bir ölçekte dönüştürülmemiş 659 faktör için faktör özdeğer-faktör numarası plotunu gösterir. Eğrinin dirsek ya da kırılma noktası uyarınca hemen hemen 14 faktör var gibi görünmektedir (Kostoff ve Block, 2005).

Şekil 2.4 : Faktör Değerlerine Ait Scee Plot



Kaynak: Kostoff ve Block, 2005.

2.15.2. Faktör Matrisi Filtreleme

Faktör yükleme örneklerini çeşitlendirmek ve her bir faktörün yorumunu basitleştirmek için, varimax ortogonal rotasyon kullanılır.

Faktör matrisi filtreleme, metin veri tabanındaki ana temaları tanımlamada, konuyu tanımlamada, kritik kelimeleri tanımlamada, bu kritik kelimeleri kümeleme

durumuna seçmeye ve incelenen belirli veri tabanının durumunda hangi kelime çeşitlerinin birleştirilebileceğinin belirlenmesinde etkili bir metottur.

2.16. KÜMELEME ANALİZİ

Kümeleme analizi; birimleri, değişkenler arası benzerlik ya da farklılıklara dayalı olarak hesaplanan bazı ölçülerden yararlanarak homojen gruplara bölmek belirli prototipler tanımlamak amacıyla kullanılır. Kümeleme analizi için başka bir tanım da şu biçimde yapılmaktadır. " Kümeleme analizi, temel amacı nesneleri (birimleri) sahip oldukları karakteristik özellikleri baz alarak gruplamak olan çok değişkenli teknikler grubudur. Kümeleme analizi, nesneleri küme içerisinde çok benzer biçimde, kümeler arasında farklı olacak biçimde kümeler. Kümeleme işlemi başarılı olursa, bir geometrik çizim yapıldığında nesneler küme içerisinde birbirine çok yakın, kümeler ise birbirinden uzak olacaktır (www.ist.yildiz.edu.tr*).*

Kümeleme analizi veri nesnelerini yalnızca nesneleri tanımlayan ve ilişkilerini ortaya koyan verilerden çıkarılacak bilgiler ışığında gruplar. Amaç aynı grup içerisindeki nesnelerin birbirine benzer veya ilişkili olması; farklı gruptakilerin ise birbirinden farklı olması ya da ilişkilerinin bulunmamasıdır. Aynı gruptakilerin birbirine benzeme oranı ya da farklı gruptakilerin ise birbirinden farklı olma oranları kümelemenin ne kadar iyi olduğunun ya da kümelerin birbirlerinden ne kadar kesinlikle ayrıldıklarının göstergesidir (bilmuh.gyte.edu.tr).

Teknik bir metin birçok özellik içerebilir, bir metin gövdesinden birçok sınıflandırma türetilebilir. Yaygın olarak iki kümeleme tipi kullanılmaktadır, içerik kümeleme ve doküman kümeleme. İçerik kümeleme, bir metin veri tabanından teknik konuları belirlemede ilişkili kelime ve ifadelerin kümelenmesidir. Web aramalarını kolaylaştırmak, literatür sınıflandırmaları türetmek, metinleri özetlemek, hipotez geliştirmek ve keşfetmek ve eş anlamlıları türetmek için kullanılır. Doküman kümeleme, konular tarafından ilişkili dokümanların gruplandırılmasıdır. Kümeler, çalışılan disiplinin (bilim dalının) bir sınıflandırma ya da sınıflandırma planını sağlamak için, hiyerarşik bir yapıda toplanabilir. Sonuç kümeler ya da sınıfın kalitesi, faktör ve kümeleme analizi için seçilen girdi kelimelerin kalitesine çok

bağlıdır. Eğer önemli yüksek teknik bileşen kelime ya da ifadeler girdilerde ihmal edilmişse bu kelimelerden çıkartılan konular (themes) sonuçlarda kaybedilecektir. Eğer girdiler için çok fazla teknik olmayan kelimeler seçilirse, teknik olmayan kelimelerle çakışmaya dayalı yapay kümeler türetilecektir ve/veya kelimeler/ifadeler teknik olmayan bağlantılar nedeniyle kümeler arasında yeniden atanmış olacaktır. Yanıltıcı bir sınıflandırma ile sonuçlanacaktır (Kostoff ve Block, 2005).

Çok değişkenli istatistik teknikler ortalama ve kovaryans yapılarını incelemeye dayanan yöntemlerin yanında sınıflama ve gruplamaya dayanan yöntemleri de içermektedir.

Diğer çok değişkenli istatistik tekniklerinde önemli olan verilerin normalliği varsayımı, kümeleme analizinde çok önemli olmayıp, uzaklık değerlerinin normalliği yeterli görülmektedir (Bülbül ve diğerleri , 2009)

Kümeleme analizi, temel olarak dört değişik amaca yönelik işlev yerine getirir.

a) n sayıda birimi, nesneyi, oluşumu p değişkene göre saptanan özelliklerine göre olabildiğince kendi içinde türdeş ve kendi aralarında farklı alt gruplara ayırmak,

b) p sayıda değişkeni, n sayıda birimde saptanan değerlere göre ortak özellikleri açıkladığı varsayılan alt kümelere ayırmak ve ortak faktör yapıları ortaya koymak,

c) Hem birimleri hem de değişkenleri birlikte ele alarak ortak n birimi p değişkene göre ortak özellikli alt kümelere ayırmak,

d) Birimleri, p değişkene göre saptanan değerlere göre, izledikleri biyolojik ve tipolojik sınıflamayı ortaya koymak

Kümeleme analizinin uygulama aşamaları aşağıdaki gibi verilebilir.

1) Birim ya da değişkenlerin doğal gruplamaları hakkında kesin bilgilerin bulunmadığı popülasyonlardan alınan n sayıda birimin p sayıda değişkenine ilişkin gözlemlerin elde edilmesi (veri matrisinin belirlenmesi)

2) Birimlerin/değişkenlerin birbirleri ile olan benzerliklerini ya da farklılıklarını gösteren uygun bir benzerlik ölçüsü ile birimlerin/değişkenlerin birbirlerine uzaklıklarının hesaplanması (Benzerlik ya da farklılık matrisinin belirlenmesi)

3) Uygun küme yöntemi yardımı ile benzerlik/farklılık matrisine göre birimlerin/değişkenlerin uygun sayıda kümelere ayrılması

4) Elde edilen kümelerin yorumlanması ve bu kümeleme yapısına dayalı olarak kurulan hipotezlerin doğrulanması için gerekli analitik yöntemlerin uygulanması

Yukarıdaki açıklamadan da anlaşılacağı gibi kümeleme analizi çok sayıda değişik işlevi yerine getiren yöntemler topluluğudur. Bu nedenle farklı amaçlar için farklı yöntemler uygulanır. Ayrıca değişkenlerin ölçü birimlerinin ve ölçümleme tekniklerinin farklı olmasından dolayı birimlerinin benzerliklerinin ortaya konmasında da değişik ölçüler kullanılır (www.ist.yildiz.edu.tr).

Kümeleme tekniği çok sayıda kayıt içeren veritabanlarında iyi bir şekilde uygulanabilir. Bu tür veritabanlarında her bir kayıt belirli bir grupta bir üye olarak sunulur. Kümeleme algoritması aynı gruplara uyan tüm üyeleri bulur. Bu üyeler içerisinde herhangi bir gruba uymayan üyeler de olabilir. Bu üyeler gürültü olarak nitelendirilir. Gürültüler kümeleme algoritmasının gücü açısından önemlidir. Örneğin veritabanında bir sigorta şirketinin müşteri bilgileri tutulduğunu varsayılsın ve benzer davranışlara göre bu müşteriler kümelenecek olsun. Bir gürültü farklı davranışlar gösteren bir müşteriyi belirtecektir. Bu gibi bir durumda örneğin şirkete karşı yapılabilecek olası bir dolandırıcılık girişimi gizlenebilir ve daha ileride araştırılmaya gerek duyulabilir. Bu aşamada kümeleme dolandırıcılık tespiti yapmak için kullanılabilir (Küçüksille, 2009: 36)

2.16.1. Farklı Kümeleme Türleri

2.16.1.1. Hiyerarşik(iç içe) kümelemeye karşın bölmesel (iç içe olmayan) kümeleme

Hiyerarşik kümeleme yönteminde başlangıçta her birey bir küme olarak kabul edilir ve birbirine en yakın iki birey ya da küme birleştirilir. Hiyerarşik

kümeleme yönteminde özellikle işleyişin daha kolay anlaşılabilmesi için dendogram (ağaç grafiği) dan yararlanılır. Dendogram birleştirici hiyerarşik kümeleme tekniği yöntemi içinde yer alan bir grafiktir. Hiyerarşik kümeleme yönteminde anlatılan işlemlere dayalı olarak kullanılan hiyerarşik metotlardan en çok kullanılanları; Tek bağlantılı, Tam bağlantılı, Ortalama bağlantı, Merkezi ve Ward metodudur (Bülbül ve diğerleri, 2009).

Üzerinde en çok tartışmanın yapıldığı kümeleme türlerini birbirinden ayrıma kriteri onların iç içe olup olmadıkları ile ilgilidir, ya da daha geleneksel bir ifade ile hiyerarşik ya da bölmesel olmaları ile ilgilidir. Bir bölmesel kümeleme basitçe veri nesnelerinin örtüşmeyen alt kümelere ayrılmasıdır öyle ki; her bir veri nesnesi yalnızca bir kümede bulunur. Kümelerin alt kümelere sahip olması durumunda ise hiyerarşik kümeleme yapmış oluruz. Hiyerarşik kümeler ağaçlar şeklinde organize edilmiş iç içe geçmiş alt kümelerden oluşur. Yaprak düğümler (leaf node) dışında ağaçtaki her bir düğüm(küme), kendi alt kümelerinin bir birliği ve ağacın kökü ise tüm nesneleri içeren bir kümedir (bilmuh.gyte.edu.tr).

2.16.1.2. Hiyerarşik Olmayan Kümeleme Yöntemi

Bu yöntem küme sayısı hakkında bir ön bilginin olması ya da araştırmacının anlamlı olacak küme sayısına karar vermiş olması durumunda tercih edilmektedir. Hiyerarşik olmayan kümeleme yönteminde en çok tercih edilen iki yöntem Mac Queen tarafından geliştirilen k-ortalama tekniği ve en çok olabilirlik tekniğidir. Bu teknikte başlangıçta ilk k kadar birimin her biri bir küme olarak alınır daha sonra her biri bir küme ortalaması olarak kabul edilerek diğer birimlerle olan uzaklıklar tespit edilir. k birim dışında kalan birimlerin her biri kendine en yakın kümeye atanarak işlem tamamlanır. Atama işlemlerinden sonra yeniden küme ortalaması hesaplanarak en yakın ortalama esasına dayalı olarak birbirine eş ya da benzer olan birimler bir araya getirilene kadar devam edilir (Bülbül ve diğerleri, 2009).

2.16.1.3. k-ortalama Kümelemesi

Klasik k-ortalamalar algoritması, Hartigan tarafından ortaya konmuştur (1975). Kümelerin verilen bir sabit sayısı (k), kümeler arasındaki (bütün değişkenler

için) ortalamaların bu kümelere gözlemleri atamak mümkün olduğunca bir diğerinden farklıdır. Gözlemler arasındaki farklılık çoğunlukla Euclidean, Squared Euclidean, City-Block ve Chebychev'i içeren birçok uzaklık ölçüleri açısından ölçülür.

Kategorik değişkenler için, tüm uzaklıklar ikilidir (0 ya da 1). Bir kümedeki en yüksek frekans ile aynı olan bir gözlemin kategorisi 0 olarak belirlenir, diğer durumda, 1 olarak belirlenir. Bu yüzden, Chebychev uzaklığının istisnası ile kategorik değişkenler için, farklı mesafe ölçüleri aynı (identical) sonuçlar verecektir (Nisbet ve diğerleri, 2009: 147).

K-ortalamalar yönteminin uygulanabilmesi için en önemli koşul, veri setindeki değişkenlerin en azından aralık ölçekte bulunmasıdır. Çünkü küme merkezleri oluşturulurken her bir iterasyonda oluşan kümeler için değişkenlerin ortalamaları alınır. İkinci önemli koşul ise, oluşturulacak olan küme sayısının başlangıçta biliniyor olmasıdır. K- ortalamalar yönteminin kullandığı algoritma aşağıdaki gibidir:

- K adet birim başlangıç küme merkezleri olarak rasgele seçilir.
- Küme merkezi olmayan birimler belirlenene uzaklık ölçütlerine başlangıç küme merkezlerinin ait oldukları kümelere atlanır
- Yeni küme merkezleri oluşturulan k adet başlangıç kümesindeki değişkenlerin ortalamaları alınarak oluşturulur.
- Birimler en yakın oldukları oluşturulan yeni küme merkezlerine birimlerin uzaklıkları hesaplanarak kümeye atlanır.
- Bir önceki küme merkezlerine olan uzaklıklar ile yeni oluşturulan küme merkezlerine olan uzaklıklar karşılaştırılır.
- Uzaklıklar makul görülebilir oranda azalmış ise 4. adıma dönülür.
- Eğer çok büyük bir değişiklik söz konusu olmamış ise, iterasyon sona erdirilir.

İterasyonun durdurulması için kullanılan ölçütlerden birisi, kareli hata ölçütleridir. Bu ölçüt p veri uzayında bir nokta, m_i ise C_i kümesine ait ortalama ya da küme merkezi olmak üzere şu biçimdedir (www.ist.yildiz.edu.tr) :

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2$$

K-Ortalama kümelemesinin dört belirgin özelliği vardır. Bunlar ;

1) Her zaman K sayıda küme olması

2) Her küme de en az bir nesne olması

3) Kümeler hiyerarşik olmalı, ayrıca her hangi bir örtüşme de olmamalıdır.

4) Kümelerin her elemanı, kendine diğer kümelerden daha yakın olmalıdır.

Çünkü yakınlık her zaman kümelerin merkezlerini kapsamaz. Veri sayısı çok fazla olan hesaplamalarda, K-ortalama hesaplaması, eğer k küçük ise hesaplamaları hiyerarşik kümelemeden daha hızlı yapar. Yine k-ortalama hesaplaması eğer kümeler özellikle küresel ise hiyerarşik kümelemeden daha sıkı bir kümeleme yapacaktır. Bunun yanında k-ortalamalar algoritmasının en büyük eksikliği k değerini tespit edememesidir. Bu nedenle başarılı bir kümeleme elde etmek için farklı k değerleri için deneme-yanılma yönteminin uygulanması gerekmektedir. K-ortalama algoritmasının küresel kümelerde, her zaman doğru kümeleri bulamadığı ancak küme sayısı doğru seçildiğinde ayrık ve sıkışık bulutlar şeklindeki kümeleri etkili bir şekilde bulabildiği söylenebilir (Çolakoğlu, 2010).

2.16.2. Farklı Küme Türleri

Kümeleme nesneleri faydalı gruplara(kümelere) ayırmayı amaçlar, burada fayda veri analizinin hedefleri tarafından tanımlanır. Doğal olarak, pratikte fayda sağlayan değişik türde kümeler vardır (bilmuh.gyte.edu.tr);

2.16.2.1. İyi Ayrılmış

Böyle bir küme bir nesneler setidir öyle ki; küme içindeki her bir nesne aynı küme içindeki bir diğer nesneye benzer ya da yakın iken küme dışındaki nesnelerden

farklı veya bu nesnelere uzaktır. Kimi zaman küme içindeki nesnelerin birbirlerine yeterince benzer olduklarını belirtmek için belirli bir eşik kullanılır. Kümenin bu ideal tanımı yalnızca verinin doğal sınıfları yani birbirlerinden yeterince uzak olan sınıfları içermesi durumunda geçerli veya doyurucu olabilir. Farklı gruplar içinde bulunan herhangi iki nokta arasındaki uzaklık aynı grup içindeki herhangi iki nokta arası uzaklıktan daha fazladır. İyi ayrılmış sınıflar küre biçiminde olmak zorunda değillerdir, fakat bir şekle sahip olabilirler.

2.16.2.2. Prototip Tabanlı

Böyle bir küme bir nesneler setidir öyle ki; küme içindeki her bir nesne kümeyi tanımlayan prototipe benzer ya da yakın iken diğer küme prototiplerinden faklı ya da bu prototiplere uzaktır. Sürekli özelliklere sahip veriler için, prototip bir ağırlık merkezidir yani kümedeki tüm noktaların ortalaması. Ağırlık merkezinin anlamlı olmadığı durumlarda, örneğin veri kategorik özeliklere sahip ise; bu durumda prototip bir **medoid**' dir yani kümeyi en iyi temsil edecek noktadır. Birçok veri türü için, prototip en merkez nokta olarak düşünülebilir ve bu gibi durumlar için prototip tabanlı sınıfları **merkez tabanlı sınıflar** olarak değerlendiririz. Doğal olarak bu kümeler küresel şekle sahip olma eğilimindedirler.

2.17. METİN MADENCİLİĞİ SONUÇLARINI VERİ MADENCİLİĞİ PROJELERİNE BİRLEŞTİRME

Girdi doküman kümesinden önemli kelimeler çıkarıldıktan sonra ve/veya belirgin anlamsal boyutları çıkarmak için tekil değer ayrışımı uygulandıktan sonra, tipik olarak sonraki ve en önemli adım bir çıkarılmış bilginin bir veri madenciliği projesinde kullanılmasıdır (www.statsoft.com).

Grafikler (Görsel veri madenciliği metotları): analizlerin amacına bağlı olarak, girdi dokümanından neyin bulunduğunun altyapısını açıklarsa, bazı örneklerde anlamsal boyutların tek başına çıkarımı, kullanışlı bir çıktı olabilir. Örneğin, Pazar araştırmaları için bu yöntem yararlı ve önemli sonuçlar çıkarabilir.

Girdi dokümanlarından çıkarılan anlamsal uzayı tanımlanması ve görselleştirilmesine yardımcı olması için grafikleri (2D scatterplots ya da 3D scatterplots) kullanılabilir.

Kümeleme Ve Çarpanları Bulma (factoring): dokümanların gruplarını tanımlamada, benzer girdi metinlerinin gruplarını tanımlamada, kümeleme analizi metotları kullanılabilir. Bu analiz tipi özellikle Pazar araştırması çalışmalarında oldukça yararlı olabilir. Ayrıca faktör analizi, temel bileşenler analizi ve sınıflandırma analizini kullanabilirsiniz.

Tahminleyici (Predictive) Veri analizi: diğer bir olasılık da; tahminleyici veri madenciliği projelerinde tahminci olarak ham ya da dönüştürülmüş kelimeleri kullanmaktır.

ÜÇÜNCÜ BÖLÜM

İTİBAR YÖNETİMİ VE VİZYON

Bu bölümde tezin uygulama kısmında yapılacak analizlerden bir çıkarım elde edebilmek amacıyla itibar yönetimi konusunun teorik yapısını ve vizyon ne demektir, vizyon nasıl oluşturulmalıdır gibi konuları inceleyeceğiz.

3.1. İTİBAR

İtibar enstitüsü "itibarı" bir şirket hakkında tüm paydaşların güven, beğeni, saygı ve iyi hislerinin derecesi olarak tanımlar. Sıkı istatistiksel analizlere göre, itibar enstitüsü bir şirketin itibarının yedi anahtar boyut etrafında gruplandığını belirtir; ürün/hizmet, performans, yenilik, çalışma ortamı, vatandaşlık (citizenship), liderlik ve hükümet (Fombrun ve diğerleri, 2007: 74).

İtibar, bir şirketin – uzun zaman boyunca koruduğu- eşsizliğinden ve kimlik şekillendirici uygulamalarından gelişir. Bunlar paydaşların şirketi inanılır, emniyetli, güvenilir ve sorumluluk sahibi olarak algılamalarını sağlar. Oluşan şirket itibarı, uygulamalarını taklit etmek için çok uğraşan rakiplerden şirketin korunmasına yardımcı olur. İtibar, rakiplerin aşmakta çok zorlanacakları bir rekabet avantajı sağlayarak şirkete stratejik değer katar (Er, 2008).

İtibar, kurumlarda ne yapıldığına ve nasıl yapıldığına odaklanan ve paydaşların deneyimlerine bağlı olarak algıya dayanan çok yönlü bir bileşendir. Kurumların finansal sorumluluklarına paralel olarak üstlendikleri kurumsal sosyal sorumluluk, kuruma duyulan güven, paydaşlarla ilişkiler ve hizmetin kalitesi gibi bileşenlerin toplam değerlendirilmesi ile ilişkilendirilen itibar, halkla ilişkilerin yeni rekabet koşulları içinde kurumsal becerilerinin kaynağını oluşturmaktadır. Bu nedenle belli bir vizyona sahip kurumlar, gelecekte olası kötü duyurumun zararlı etkilerinden proaktif olarak kaçınmaya çalışırlar. Kurum itibarı, bu amacın başarılması için iyi bir araç olarak kabul edilmektedir. Çünkü itibar, kurumlarda ne yapıldığına ve nasıl yapıldığına odaklanan ve paydaşların deneyimlerine bağlı olarak algıya dayanan çok yönlü bir bileşendir (http://if.kocaeli.edu.tr).

Amerikan Heritage Sözlügü kurumsal itibarı; "Hedef kitlelerin kurum hakkındaki toplam fikirleri" olarak tanımlamaktadır. Kurumsal itibar; müşterilerin, yatırımcıların, çalışanların ve genel kamuoyunun kurum hakkındaki iyi veya kötü, zayıf veya güçlü gibi duygusal ve etkileyici tepkilerini ifade etmektedir (Ural, 2002: 85).

Penguen İngilizce Sözlüğü itibarı " 1: diğerleri tarafından görülen ya da kıyaslanana tüm kalite ya da özellikler, 2: ün, tanınma, 3: bazı özellikler ya da yetenekler ile diğerleri tarafından tanınma" olarak tanımlamaktadır. Bu tanım basit gibi görünse de şirket itibarı konusuna gelindiğinde durum daha fazla karmaşık hale gelmektedir. Bazı uzmanlar şirket itibarını " diğer tüm lider rakiplerle karşılaştırıldığında firmanın tüm cazibesini tanımlayan, şirketin geçmiş aktiviteleri ile gelecek vaat eden beklentilerinin algısal temsili" olarak tarif etmektedir. Bu durum da "paydaşların firma hakkındaki görüşü" kavramını popüler hale getirmektedir. İtibar ile ilgili diğer uzman görüşleri ise "tüm paydaşların görüşlerini ele almak için tasarlanan yeni farklı ve tamamlayıcı itibar ölçme yaklaşımlarını paydaşların şirket itibarını anlamaları ile tamamlamak" şeklindedir (Griffin, 2008: 57).

"İtibar tüm deneyimlerin, izlenimlerin, inançların, hislerin ve şirketin performansı hakkında paydaşların sahip oldukları bilginin etkileşimidir". Böylece, bugünün itibarı yarının performansına etki eder. Günümüze kadar yapılan pek çok çalışma şirket itibarının finansal değere dönüştüğünü kanıtlamıştır. İstatistiksel olarak itibarındaki %60 artışın pazar değerinde %7 artışla ilişkilendirilmektedir. İtibarın kesin değerinin tanımlanamamasına rağmen genellikle iyi bir itibarın bir şirketin çalıştığı önemli pazarlara erişim sağlayan bir sigorta olduğu kabul edilmektedir (Kim, 2003).

İtibar ile ilgili yapılan tanımlamaların özeti olarak itibarın diğerlerinin kurum hakkındaki düşünceleri olduğu söylenebilir.

3.2. İTİBARI OLUŞTURMA VE YÖNETME

İtibar bazı yönlerden iyi bazı yönlerden de kötü olabilir ya da bir organizasyon faaliyetlerinin özel bir bölümü için bazı insanlar tarafından iyi diğer insanlar tarafından da kötü olan bir itibara sahiptir. İtibar hakkında bilgi (knowledge) kaynakları (Elearn Limited, 2005: 1-2);

1. Organizasyonla doğrudan iş ilişkisi içinde bulunmak
2. Arkadaşlardan iş arkadaşlarından ve tanıdıklardan duyulan söylentiler
3. Gazete makaleleri, televizyon belgeselleri ve yayınlanmış araştırmalar
4. Broşürler, yıllık raporlar ve reklamlar gibi organizasyonlarca oluşturulan bilgiler

Kurumsal itibarı inşa etmeye yardımcı olan faktörler (Er, 2008);

- Finansal performans
- Hizmet ya da ürünlerin kalitesi
- Marka değeri ve vaadi
- İnovasyon ve yaratıcılık
- Müşterilerin hizmet tatmini
- Sosyal sorumluluk- kurumsal vatandaşlık
- Kurumsal politikalar ve örgüt yapısı
- Başarılı rekabetçi konumlandırma
- Vizyon ve liderlik
- Tepe yöneticisinin performansı
- Kanuni düzenlemelere tam uygunluk
- Çalışan memnuniyeti ve sadakati
- Öz yeteneklerin geliştirilmesi
- İşbirliği ağlarının ve müttefiklerin oluşturulması

Schultz ve Werner'e göre itibar yönetiminin amaçları şunlardır (Er, 2008);

- Çalışma çevresinde ve piyasadaki olumlu itibarı devam ettirmek
- Şirketin olumlu ismini ve itibarını oluşturmak ve iyileştirmek

- İtibarın zarar görmesini önleyecek kabul edilebilir uygulamalar, politikalar, prosedürler, sistemler ve standartlar oluşturmak
- Şirketin itibarını zedeleyecek bir olay meydana geldiğinde durumun üstesinden gelmek için yönergeler oluşturmak
- Şirket itibarını yönetmek üzere tam sorumluluk alacak yönetim takımını hazırlamak ve donatmak

3.3. İTİBAR YÖNETİM SÜREÇLERİ

Şirketin itibar yönetimi ile ilgili süreçleri "kurumsal iletişim birimi" tarafından doğrudan üst yönetime raporlanacak şekilde yönetilmelidir. Böylece itibar yönetiminin nasıl ve ne şekilde yapılacağı, organizasyonun pazarlama, finans, üretim, insan kaynakları gibi diğer faaliyetleri tanımlanmış olacaktır. Kurumsal itibarın hammaddesi "sürdürülebilir kalkınma" ile ilgili temel öğelerdir. Buna ek olarak Fortune, Financial Times ve Reputation Institude gibi yayın ve kurumların geliştirdiği itibar kriterleri arasında yer alan ürün ve hizmetlerin kalitesi, global pazarlara entegrasyon, kaynakların kullanımı, müşteri memnuniyeti, yatırımcı değeri, araştırma-geliştirme, ileri teknoloji ve kurumsal performans gibi başlıklar farklı sosyal ortaklar nezdindeki kanaatleri etkilemeleri nedeniyle değerlendirmeye alınmaktadır (Kadıbeşegil, 2006).

Rene ve Van Dam itibar yönetimi özelliklerini söyle sıralamaktadır (Öksüz, 2008: 57);

• *Farklılık:* Güçlü itibarlar kurumun paydaşları zihninde sahip oldukları farklı konumun sonucunda oluşmaktadır.

• *Odaklanma:* Güçlü itibarların oluşturulabilmesi kurumun davranışlarının ve iletişimlerinin tek bir tema çerçevesinde gerçekleştirilmesi gerekmektedir.

• *Tutarlılık:* Güçlü itibarların oluşumu kurumun tüm paydaşlarla davranışlarında ve iletişimlerinde tutarlı olmasına bağlı olmaktadır.

• *Kimlik:* Güçlü itibarlar kurumun davranışlarının benimsenmiş kimliğe uyumlu olmasının sonucunda gerçekleşmektedir.

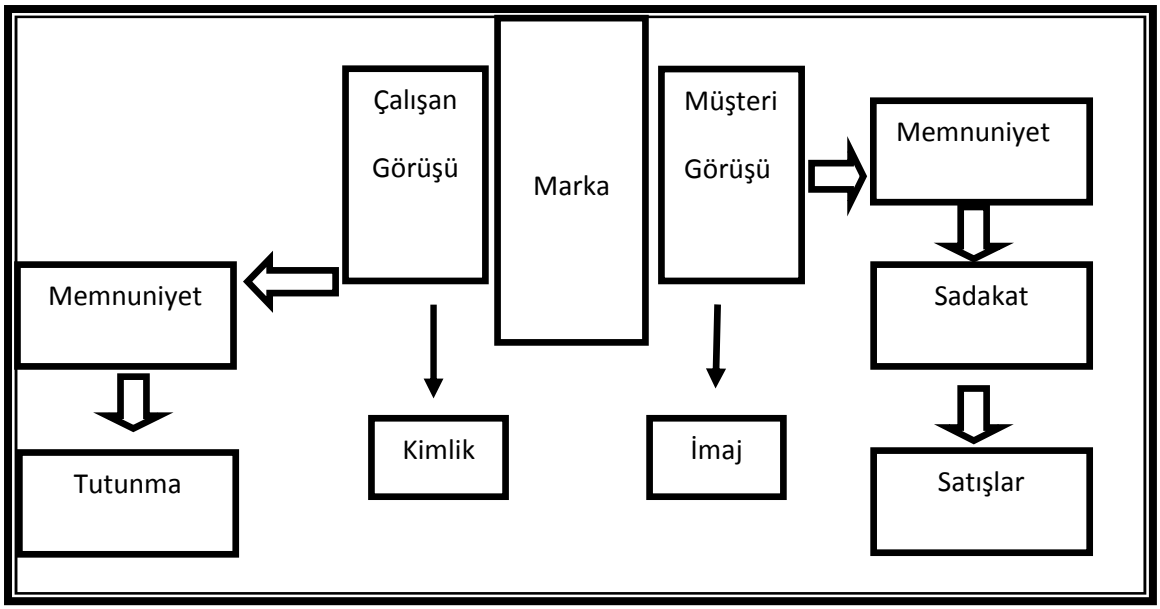
• Ş*effaflık:* Güçlü itibarlar kurumun tüm ilişkilerindeki şeffaflığının sonucunda oluşmaktadır.

3.4. İTİBAR VE KURUM

Organizasyonun itibarı tüm hissedarları için önemli bir düşüncedir. Hissedarlar organizasyonun aktivitelerinde doğrudan ilgisi olan bir grup insandır ve organizasyonun itibarına sermayesine ya da zamanına yatırım yapmış olan kişilerdir. Çünkü hissedarlar, organizasyonda gerçek bir taahhütte bulunmaya hazırdırlar ve onların fikirleri önemlidir. Aynı zamanda, bu kişiler organizasyonun müşterilerinin olduğu kadar organizasyonun itibarının da yaratıcılarıdır. Çeşitli gruplar arasında hissedarlar için iyi olan bir şeyin çalışanlar için iyi olmaması ve yönetim için iyi olan bir şeyin müşteriler için iyi olmaması gibi çatışan çıkarlar olduğunda temel zorluk ortaya çıkar. Bunun anlamı sık sık yöneticilerin çelişkili baskıları ortadan kaldırmak zorunda olmalarıdır (Elearn Limited, 2005: 7).

Şekildeki zincir ideal olan ve çoğu zaman deneyimlerimiz dışında olan bağlantıların bulunması gereken bir diziyi sunar. Zincirin temeli imaj ve kimliğin anahtar kelimeleridir. Kurumsal itibarın en önemli iki sorusu dış paydaşlara sorulması gereken“kim olduğunuz” ve iç paydaşlara sorulması gereken “biz kimiz”dir. İlk soru organizasyonun dış imajını ilgilendirirken ikinci soru kimliği, yani çalışanların çalıştıkları organizasyonu nasıl gördükleri ile ilgilidir (Davies ve diğerleri, 2003: 137).

Şekil 7: Kurumsal İtibar Zinciri



Kaynak: Davies ve diğerleri, 2003, s. 137.

Kurumsal itibar; müşteriler, rakipler, kreditörler, endüstri analistleri ve diğer insanların işletmeyi algılayış biçimidir. Daha açık bir tanımla kurumsal itibar, işletmenin yönetim kapasitesi, stratejileri, finansal durumu, sosyal ve toplumsal sorumlulukları, uzun dönem yatırımlarının değeri, rekabetteki etikliği, gelişme düzeyi, personelinin kalitesi, nitelikli iş göreni çekme becerisi gibi konularda insanlar tarafından nasıl algılandığını belirlemektedir ve bu kriterlerle ölçülmektedir. Fombrun'a göre kurumsal itibar müşteriler, yatırımcılar, iş görenler ve genel çevrenin gözünde işletmenin iyi ya da kötü, güçlü ya da zayıf olduğunu göstermektedir (Karakılıç, 2005).

Kurumsal itibarın şirketin yarınlarını güvence altına alacak şekilde yönetilmesi hususunda aşağıdaki konular vurgulanmalıdır (Kadıbeşegil, 2006);

- Şirket vizyonunun içselleştirilmesi, kurum kültürü ve değerlerinin tanımlanması

- Etik ve ahlaki değerler ile birlikte hesap verilebilirlik uygulamaları
- Uluslar arası muhasebe standartlarının benimsenmesi ve şeffaflık yönetimi
- Kurumsal sosyal sorumluluk anlayışı ve yönetimi
- Çalışan memnuniyeti ve çalışanların kariyer gelişim planları politikası
- Müşteri memnuniyeti politikaları ve müşteri odaklılık
- Ar-Ge ve inovasyon yetkinliği
- Üçlü raporlama (finansal, sosyal ve ekolojik çevre uygulamaları)

İtibar algısının temelinde bütünsellik, tutarlılık, kalıcılık ve süreklilik kavramları bulunmaktadır. Müşteriler ürünle (kalitesi, ürün güvenliği, servis v.b.) ilgili, çalışanlar iş (iş güvenliği, etiği, aidiyet, yönetim, kurum içi iletişimin işleyişi v.b.) ile ilgili bilgilere ihtiyaç duyarken yatırımcılar finansal konularla (net kâr, ana para, v.b.) ilgili bilgilere gereksinim duyarlar. Müşteriler ve yatırımcılar bu bilgileri değerlendirerek kurumun onlar için ne anlama geldiğini ve o kurumla işbirliği yapmanın temelini oluştururlar. Çalışanlar ise, o kuruma duydukları aidiyet ve inanmışlık ile dışarıdaki paydaşların kurum hakkındaki güvenin temelini oluştururlar. Bir kurumun hizmetine ve ürünlerine duyulan güven, çalışanların katkısı olmadan gerçekleştirilemez (http://if.kocaeli.edu.tr).

Örgütsel itibar kavramına yönelik anlayışları *faydacı (pragmatic)* ve *yansıtmacı (reflexive)* olarak adlandırılan ve birbirlerini tamamladıkları düşünülen iki ana baslık altında toplamak mümkündür. *Faydacı* yaklaşım, örgütün nihai amacının kazancını azamileştirmek olduğu ve yöneticilerin performansının da birincil olarak karlılık düzeyine göre değerlendirileceği fikri üzerine kurgulanmıştır. Bu anlayış çerçevesinde itibar, örgütleri ve yöneticilerini nihai amaca götürecek bir araçtan ibarettir. *Yansıtmacı* bakış açısına göre ise örgüt bir para basma makinesi değildir. Örgütün etkileşim içerisinde olduğu gruplara karsı bazı sorumlulukları vardır ve itibar öncelikle kazancı azamileştirme güdüsünün değil, bu sorumlulukları yerine getirmeye yönelik *diğerkamcı* çabaların bir tezahürüdür (Eryılmaz, 2008: 155-174).

3.5. İTİBARSAL SERMAYE

İyi bir itibarın hem soyut hem de somut getirileri vardır. Müşterilerden çalışanlara şirket ile ilgili olan tüm kitlelerin organizasyonla ilgili iyi hislere sahip

olmaları önemlidir. Bir organizasyonun zor zamanlarda iyi bir itibarı sürdürmesi önemlidir. İyi itibarlı şirketler çalıştırmak için iyi ve daha çok aday çekerler, kaynak bulmada daha az öderler, reklamın yapabileceğinden daha fazla değerli olan bedava basın haberi elde ederler ve karlılıklarına katkıda bulunacak diğer katkıları arttırırlar. İtibar, pazar sermayesinin defter değeri ya da yatırımlarının likidite değeri gibi şirketin gerçek değerine değer ekler. Pazar sermayesinin itibar bileşeni, "iyi niyeti" ile yakından ilgili olan ve birçok büyük şirketin milyon dolarlarına değecek olan itibarsal sermayedir (Doorley ve Garcia, 2007: 4).

İtibarsal sermaye şirketin daima "risk altında" olan pazar değerinin bir parçasıdır. Yöneticilerin gazetecileri ve analistleri şirket hakkında övmeye ikna etmeleri ile itibarları ve hisseleri artar. Şirketin ürünlerine, yöneticilerine dair inançları azalırsa zarar görür (Fombrun ve diğerleri, 2007: 78).

Dowling'e göre de iyi itibarın isletmelere faydaları su şekildedir (Öksüz, 2008) :

- İyi itibar kurumun ürün ve hizmetlerine ekstra değer kazandırmakta.
- Müşterilerin ürün ya da hizmet satın alırken algıladıkları riskin azalmasına yardımcı olmakta.
- Müşterilerin fonksiyonel olarak benzer algıladıkları ürünler (televizyon vs.) ve hizmetler (eğitim, hukuki hizmetler) arasında seçim yapmasına yardım etmekte.
- Çalışanların iş tatminini arttırmakta (iyi kurumlar çalışanların is tatmin oranlarında hala etkisi bulunmaktadır).
- İstihdam sürecinde nitelikli işgücünün çekilmesine yardımcı olur.
- Reklam ve satış-gücü etkinliğini artırmakta.
- Yeni ürün tanıtımlarını desteklemekte.
- Kurumların rakiplerine karsı kullandığı güçlü bir araç olmakta.
- Kurumların en iyi hizmet sağlayıcılara ulaşabilmesini sağlamakta (örneğin en iyi reklam ajansları en iyi müşterilerle çalışmak istemektedir).
- Dağıtım kanallarında kurumların pazarlık gücünü artırmaktadır.

3.6. İTİBARIN ÖLÇÜLMESİ

Formbrun itibarsal sermayenin market sermayesi ve yatırımların likiditasyon değeri arasında farklı olduğunu belirtir. Fakat itibarsal sermayenin değerinin farklı olarak abartıldığına inanan çoğu finans müdürü bu formülle aynı fikirde değildir. İtibarı ölçmede en genel yaklaşım benzer organizasyonlar ile karşılaştırmalı ölçümlerini almaktır. Fortune dergisinin yıllık Amerika'nın en Hayran olunan şirketleri araştırması en çok bilinen ve hem endüstri liderleri hem de akademisyenlerin saygı duyduğu araştırmadır. Bu konuda daha kapsamlı bir araştırma çalışanları, müşterileri ve basını da içeren en önemli katılımcıların araştırılmasını da içermesi olacaktır (Doorley ve Garcia, 2007: 16).

İtibar kavramının tanımından şirketin hisse senedi fiyatının şirketin tüm deneyimlerin, izlenimlerin, inançların, hislerin ve şirketin performansı hakkında paydaşların sahip oldukları bilginin etkileşimi olduğu sonucuna varılabilir. Böylece itibarin ölçümü hisse senedi fiyatlarının çok boyutlu eşitliğidir. Açık olarak tanımlanmış "itibar-tabanlı ölçüm" metodu yoktur (Kim ve Dam, 2003).

İtibar birçok yoldan ölçülebilir. Şirket itibarını ölçmede çoğunun şirket sıralamalarına odaklandığı birçok çeşitli ölçümler vardır. Fortune dergisinin yaptığı ölçülmede ise bazı ölçüler itibar için kullanılan örnek uzayının geniş olmasından (uzmanlar ve analistler) ve çalışanlar ve müşteriler gibi önemli hissedarları hariç tutması nedeniyle aşırı derecede finansal performansa odaklandığından teorik bir temele sahip olmayan değerlendirmeler nedeniyle kriterler eleştirilmiştir. Buna tepki olarak birçok ölçüm tekniği geliştirilmiştir, bunlardan en ünlü olanları sadece bir hissedar tipine odaklananlardır. Belirli bir Pazar sektörünü belirlemeye yönelik belirli birçok ölçümler vardır (Davies ve diğerleri 2003: 159).

1995 yılında Fortune Dergisi'nin yayınladığı Amerika'nın en takdir edilen şirketleri raporunda bu konuda söyle denilmektedir; "İtibar her zaman değerli olmuştur. Fakat bilgi ekonomisinde şirketlerin itibarı ve genel kimlik anlayışı giderek artan bir değer taşımaktadır. Hatta bireyler ve şirketler arasındaki bağlar sıradan şirketlerde yıpranırken, yüksek performans gösteren şirketlerde bu bağlar daha

önemli hale gelmiştir. Bu, aynı zamanda onların neden yüksek performans gösteren şirketler olduğunu açıklamaktadır." (Ural, 2002: 87).

İtibar ile ilgili yapılan bazı araştırmalar ve kullanılan bileşenler aşağıdaki gibidir;

Tablo 1: Şirketlerin İtibar İle İlgili Sıralama Araştırmaları

Şirket Sıralama Araştırmaları
Avustralya Business Review Weekly
Pazar değeri
Finansal performans
Çevresel etki
Sosyal sorumluluk
Çalışan ilişkileri
Yönetim/etik
Finansal Times
Güçlü ve iyi planlanmış strateji
Müşteri memnuniyetinin ve sadakatinin en üst düzeye çıkartılması
İsletme liderliği
Ürün ve hizmet kalitesi
Güçlü ve istikrarlı kar performansı
Güçlü başarılı değişim yönetimi ve insancıl kurum kültürü
İsletmenin küreselleşmesi.
Management Today
Finansal sağlamlık
Yetenekli çalışanları çekme
Geliştirme ve elde tutma kabiliyeti
Ürün ve servis kalitesi
Uzun dönemli yatırım değer
Yaratım kapasitesi
Fortune AMAC, Fortune GMAC, Manager Magazin, Management Today, Asian Business, Far Eastern Ekonomic Review, Financial Times, Industry Week dergilerinin arastırmaları incelendiginde kullanılan bilesenler söyledir:
Pazarlama kalitesi
Toplumsal ve çevresel sorumluluk
Kurum değerlerinin kullanımı

Tablo 1: Devam

Asian Business
Genel yönetim
Yönetim kalitesi
Ürün ve hizmet kalitesi
Yerel ekonomiye katkısı
İyi çalışanlara sahip olması
Gelecekteki kazanç potansiyeli
Değişen ekonomik çevreye uyum yeteneği
Fortune
Yönetim kalitesi
Ürün ve hizmet kalitesi
Yenilikçilik-yaratıcılık
Uzun dönemli yatırımların değeri
Finansal güçlülük
Nitelikli çalışanları çekme
Geliştirme ve elde tutma becerisi
Sosyal sorumluluk
Kurum değerlerini/kaynaklarını kullanabilme niteliği
Capital
Pazarlama ve satış stratejileri
Hizmet ve ürün kalitesi
Çalışanların nitelikleri
Finansal sağlamlık
Toplumsal sorumluluk
Yatırımcıya değer yaratma
Uluslararası pazarlara entegrasyon
Yönetim kalitesi
Çalışana sunulan sosyal olanaklar
Ücret politikası ve seviyesi
Yönetim ve şirket şeffaflığı
Çalışanların niteliklerini geliştirme
Bilgi ve teknoloji yatırımları
İletişim ve halkla ilişkiler
Rekabette etik davranma
Çalışan memnuniyeti
Yeni ürün geliştirme
Müşteri memnuniyeti

Kaynak: Öksüz, 2008, s. 95-96.

Kurumsal itibar performansı, algılama araştırmaları ile ölçülmektedir. Kurumsal itibarın marka ve ürün performansı ile arasındaki fark ise kurumsal itibarda tüm sektörlerin rakip olmasıdır. Marka ve ürün rekabetinde ait olunan sektördeki rakipler değerlendirilir. Kurumsal itibarını etkin bir şekilde yöneten şirketler, ürün ve marka performanslarını, kurumsal itibar yönetimi politikaları ile ilişkilendirmelidirler (Kadıbeşegil, 2006: 176-177).

Lewellyn itibar ölçümü için, kurumsal kimlik ve imajın boyutlarıyla ilgili kapsamlı bir çerçevenin geliştirilmesini zorunlu görmektedir. Chun itibarın imaj ve kimlikle yapısal ortaklığından dolayı ölçümünde "müşteri ve çalışan" gibi çeşitli değişkenler arasındaki ilişkilerin test edilebilir bir duruma geldiğini ifade etmiştir. Castro, Lopez ve Saez ise kurumsal itibarın "iş itibarı" ve "sosyal itibar" olmak üzere iki bölümde ölçülmesi gerektiğinden söz etmiştir. İmaj ve kimliğin her ikisini de ölçebilmek için, sosyal itibarının ölçümünde kurumların "insan" gibi sorgulanması gerektiği düşünülmüştür. Aaker tarafından kullanılan yaklaşımdan sonra Davies ve Chun çalışan algılarını (kimliği temsil eden) ve tüketici algılarını (imajı temsil eden) ölçen "kurumsal kişilik skalasını" geliştirmişlerdir. Böylece, bir kurumun hem içteki hem de dıştaki algılanmasının değerlendirildiği kapsamlı bir ölçüm aracı oluşturulmuştur (Demir, 2010: 247-262).

Tablo 2 : Kurumsal Tabanlı Ölçümler

Kurumsal Tabanlı Ölçümlerin Özeti		
Yazar(lar)	**Ölçü**	**Değişkenler**
Javalgi vd. (1994)	Kurumsal İmaj	Ürün/Hizmet Yönetim Kar Motivasyonu Toplum Katılımı Müşteri İhtiyaçlarının Karşılanması Çalışma Ortamı
Brown ve Dacin (1997)	Kurumsal Ortaklar	**Kurumsal Sorumluluk Ortakları** Endüstri Liderliği Araştırma ve Geliştirme Yeteneği Kurumsal Liderlik **Kurumsal Sosyal Sorumluluk Ortakları** Çevre Kaygısı Yerel Toplulukların Katılımı Değerli Amaçları Olan Kurumlar
Keller (1998)	Kurumsal İmaj Ortakları	Mevcut Ürün Özellikleri, Getirileri, Davranışları İnsanlar ve İlişkiler Değerler ve Programlar Kurumsal Güvenilirlilik
Fombrun vd. (2000)	İtibar Katsayısı	Duygusal İlgi Ürün ve Hizmetler Vizyon ve Liderlik Çalışma Ortamı Sosyal ve Çevresel Sorumluluk Finansal Performans
Davies vd. (2001)	Kişileştirme Metaforu	Dürüstlük Heyecan Yeterlilik Çok Yönlülük Sağlamlık
Melewar ve Jenkins (2002)	Kurumsal Kimlik	İletişim ve Görsel Kimlik Tutum Kurum Kültürü Pazar Şartları Firma, Ürün ve Hizmetler

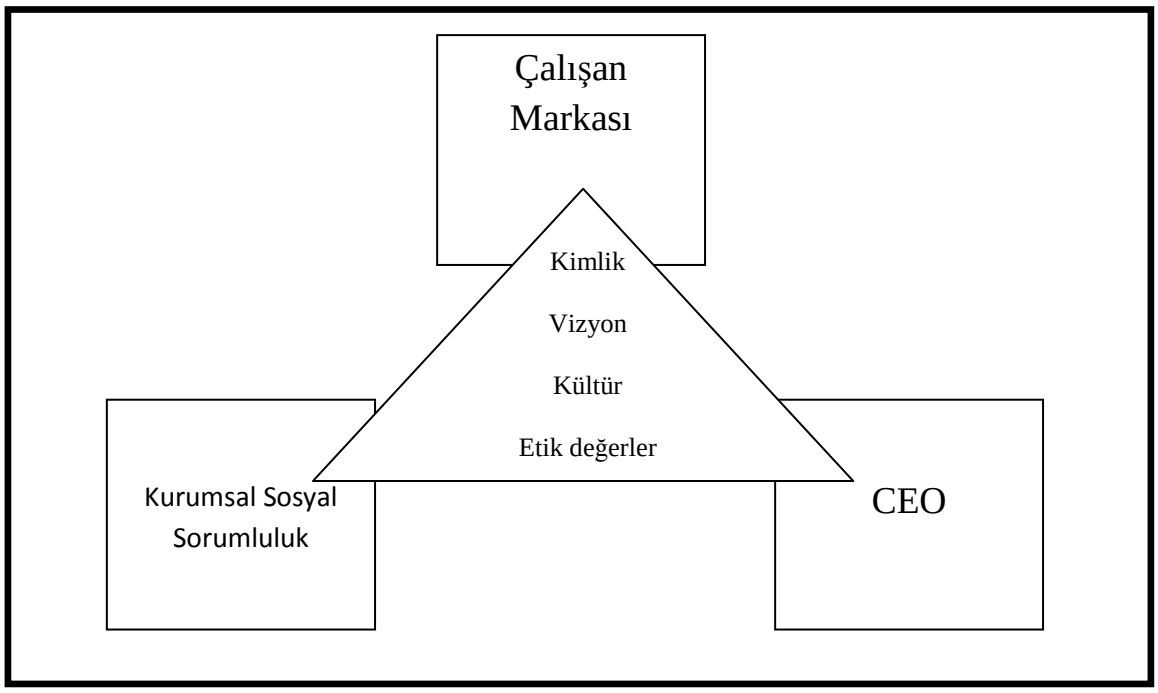
Tablo 2: Devam

Brady (2003)	Kurumsal İtibar	Bilgi ve Beceriler Duygusal Bağ Liderlik, Vizyon ve İstek Kalite
Cravens vd. (2003)	İtibar İndeksi	Ürünler Çalışanlar Dış İlişkiler İnovasyon ve Değer Yaratma Finansal Güçler ve Canlılık Strateji Kültür Maddi Olmayan Yükümlülükler
Berens ve Van Riel (2004)	Kurumsal Ortaklar	Sosyal Beklentiler Kurumsal Kişilik Özellikleri Şirkete Olan Güven
Davies vd. (2004)	Kurum Karakteri	Kabul Edilebilirlik Girişim Yeterlilik Modaya Uygunluk Acımasızlık Formaliteye Uymama Maçoluk
Helm (2005)	Kurumsal İtibar	Ürün Kalitesi Çevrenin Korunması Taahhüdü Kurumsal Başarı Çalışanlara Davranışlar Müşteri Odaklılık Yardımseverlik ve Sosyal Konularda Bağlılık Ürünlerin Para Değeri Finansal Performans Yönetim Kalitesi Reklamların Güvenirliliği

Kaynak: Shamma, 2007, s. 61.

Aşağıdaki şekilden de görülebileceği gibi itibarlı şirket olma ölçütleri; kurumsal sosyal sorumluluk projeleri, kurumun etik değerleri, vizyon misyon strateji ifadeleri ve bunları uyguladıklarını gösteren faaliyetleri, itibar yönetiminde şirket CEO'sunun rolü, çalışanlara verilen değer gibi faktörlerdir.

Şekil 8: İtibarlı Şirket Olma Ölçütleri



Kaynak: http://ww2.kalder.org/

(Argüden, 2003: 12), itibarı yönetmek için düzenli olarak ölçülmesi gerektiğini vurgulamış ve iyi bir itibar yönetimi sistemi kurulmasında hedef kitlenin belirlenmesi ve sağlanması, ölçülecek itibar boyutlarının tespit edilmesi, ölçümlerinin yapılması ve zaman içinde ulaşılması gereken hedeflerin belirlenmesinin gerekli olduğunu belirtmektedir. Özetle;

1. Net bir vizyonu olan, bunu kurum içinde ve dışında açık bir şekilde paydaşları ve kurumun tüm davranışlarında bu vizyonla uyumlu hareket etmesini sağlayan kurumların itibarı artmaktadır.
2. Etik standartlara önem veren ve bunu uygulamalarına yansıtabilen kurumlar da itibar kazanmaktadırlar.
3. Bir kurumun başta üçüncü taraflarla ilişkilerini yürüten çalışanları ve üst yönetimi olmak üzere çalışanlarının yüksek nitelikli kişilerden oluşuyor olması, kurumun itibarına olumlu etki yapmaktadır.

4. Kurumun başta müşterileri olmak üzere tüm paydaşlarının çıkarlarını gözetiyor olması, kurumun itibarını geliştirmektedir.

5. Kurumsal sosyal sorumluluğa verilen önem itibarı olumlu etkilemenin bir başka yolu olarak belirlenir.

6. İtibar oluşturmak için kurumsal iletişim araçlarının da etkin ve tutarlı olarak kullanılması, itibarın hedef kitlesine kurumsal davranışların nedenlerinin iyi anlatılması için çaba gösterilmesi gerekir.

7. İtibarı oluşturmak kadar onu korumak da önemlidir. Dolayısıyla, itibar yönetimi için risk yönetimi konusunda da hazırlıklı olmak gerekir.

İtibar Katsayısı

Merkezi ABD'de bulunan İtibar Enstitüsü (Reputation Institute) Harris araştırma şirketi ile birlikte bir araştırma gerçekleştirmiş ve kurumsal itibarın ölçümlenmesine yönelik İtibar Katsayısını (RQ-Reputation Quotient) geliştirmiştir. Bu ölçümleme aracı ile altı boyuta dayanarak kurumların itibarı değerlendirilmektedir. Bu boyutlar şöyledir (Öksüz, 2008: 94):

- *Duygusal Cazibe*, kurum hakkında olumlu duygulara sahip olunması ve kurumun takdir edilmesi.
- *Ürün ve Hizmetler*, kaliteli, geliştirilen, değerli ve güvenilir ürün ve hizmetler sunulması.
- *Finansal Performans*, rekabet edebilirlik, karlılık, büyüme olasılığı ve risk durumu.
- *Vizyon ve Liderlik*, açık ve net bir vizyon gösterilmesi, güçlü liderlik, pazar fırsatlarını görme ve faydalanabilme yeteneği.
- *Çalışma Ortamı*, iyi yönetilmesi, çalışılacak iyi bir şirket görüntüsü çizip çizmediği, nitelikli çalışanlara sahip olunması.
- *Sosyal Sorumluluk*, toplumla olan ilişkilerinde yüksek standartların oluşturulması, çevresel ve toplumsal konularda çalışmalar yapılması.

İtibarlı şirket olma yolunda en önemli başlangıç noktası olarak şirketlerin açık ve net bir vizyon tanımı oluşturmaları ve bu vizyon tanımına göre hareket

etmeleri gerekmektedir. Bu bağlamda vizyon oluşturma konusunda da dikkat edilmesi gereken hususlar vardır.

3.7. VİZYON

Bir şirketin stratejik planlama süreci vizyon ifadesidir. Bir vizyon ifadesi bir şirketin nerde olmak istediğini özetler. Bir şirket için vizyon sanatı ileriye odaklı şirketin liderinin oynadığı en önemli roldür. Bir vizyonun müşteriler, çalışanlar ve satıcılar gibi bileşenleri vizyonun anlaşılmasına ve vizyondaki rollerinin ne olduğunun anlaşılmasına katkı sağlarlar. Vizyon hissedarların gelecek hakkında nasıl düşünmeleri gerektiğini gösterir. Gelecekte olaylar genişlediğinde, vizyon ile şirkete etkileri belirlenebilir. Hissedarlar ayrıca çeşitli olaylar karşısında neler olabileceği ile ilgili kendi senaryolarını oluşturabilirler (Cuppet, 2008: 26).

Doğru Vizyonu Seçme

1. Ne ölçüde geleceğe yönelik
2. Ne ölçüde ideal
3. Ne ölçüde uygun (şirketin tarihini, kültürünü ve değerlerini belirle)
4. Ne ölçüde mükemmellik standartlarında
5. Ne ölçüde amacınızı ve yönünüzü açıklar
6. Ne ölçüde istek yaymaktadır
7. Ne ölçüde eşsizliği yansıtacak
8. Yeterince iddialı mi

İyi Bir Vizyon Nasıl Olmalı

1. Organizasyon ve zaman için uygundur
2. Mükemmellik standartlarını ve yüksek idealleri belirler
3. Amaç ve yönleri açıklar
4. İstek ve taahhüt uyandırır
5. Oldukça açık ve kolay anlaşılır
6. Eşsizliği yansıtır
7. İddialıdır

Bir şirketin vizyonu; organizasyonun nihai, basit –genellikle ulaşılamaz- durumu olmalıdır. Ulaşılmaz olmasının sebebi de organizasyonların süresiz olarak kurulmuş olmalarındandır. Eğer bir organizasyon nihai amaçlarına ulaşmışsa, vizyonunu değiştirmelidir. Bir şirketin vizyonu geniş olmalı fakat açık olmamalıdır.

Çoğu şirketin vizyon, misyon gibi temel değerleri vardır, fakat etkili olarak kullanılamamaktadırlar. Vizyon, misyon gibi kavramların kullanılması şirketlerin yönlerini net olarak belirlemesinde, çalışanların morallerini ve performanslarının artmasında etkili olmaktadır.

3.8. VİZYON NASIL OLUŞTURULMALIDIR?

Vizyon tahmin ve gelecekle ilgili öngörülerden oluşmaz. Vizyon, organizasyonun belirlenen kriterler doğrultusunda gelecekte nerelerde olabileceğinin tanımlanması ve oraya ulaşılması için amaçların konulması olarak görülmelidir. Vizyon, gelecekteki durumun tahmin edilmesi değil; bilakis gelecekte olunması istenen durumun tanımlanmasıdır (Sönal, 2007: 1-4).

1. Üst yönetimin inancı. Vizyon, kurulusun stratejik planı i!e uyum içinde olmalı ve piyasa koşulları tarafından yönlendirilmelidir. Şirket vizyonu, “Büyüdüğümüz zaman ne olmak istiyoruz?” sorusunun yanıtı olmalıdır. Vizyon, yukarıdan aşağıya dayatılanı bir kavram olmamalıdır.

2. Yönetimin tam olarak katılımı. Sizin katılmadığınız bir surece, çalışanlarınızın gönülden katılmasını asla ve asla beklemeyin.

3. İnsanların katılımı. “Biz” ve “Onlar”ın söz konusu olduğu günler geride kaldı. Bir şirkette çalışan herkes en az yöneticiler kadar önemli. Onları dinlemeniz gerekir. İnsanların katılımını sağlamak için onları dinlemek önemli bir adımdır.

4. Bireyin katılımı. Hepimiz bir gruba ait olmayı ve grup halinde çalışmayı istesek de, birey olarak da var olmak için aynı derecede güçlü bir istek duyarız. Çalışanları yetkilendirme konusunda hazırlanan her programda, bireysel tatmin unsuruna yer verilmelidir.

5. Sürekli gelişme ekipleri. Ekipleri oluşturan insanların destekleyici çabaları olmazsa, süreç yalnızca bir yönetim programı olarak kalacaktır. İnsanların şirket vizyonunun bilincine varmasını sağlamakta etkili bir yol, onları hem şirketin

hem de kendilerinin misyonunu açıklayan bildiriler oluşturma surecine katmaktır. Vizyonun yaratılmasına katkıda bulunduklarını bilmek, çalışanları motive eder.

Bir firmanın vizyonu olduktan sonra geriye yapılacak iki adım kalır. İlk adım vizyonu test etmektir. Firmanın ilerlemesini düzenli olarak izlemek fakat bu izlemeler arasında bir yıldan fazla zaman geçmemesi gerekir. İhtiyaç olduğunda gerekli ayarlamaları yapmak gerekir. İkinci adım ise eylemdir. Vizyon firmadaki herkese iletilmelidir. Vizyon bir sır değildir. Vizyon müşterilerle ve iş ortakları ile paylaşılmalıdır. Ayrıca internet sitesinde ve broşürlerle yayılmalıdır. Firma vizyonunun geliştirilmesi için zaman ve çaba harcanmalıdır. Firmanın imajını oluşturmak için vizyon kullanılmalıdır. Son olarak, firmanın kimliği yeniden sorgulanmalı. Firmalar büyür ve değişir, kimlik de bu durumda değişmeli ve vizyon oluşturma süreci yeniden başlatılmalıdır (Cuppet, 2008: 61).

DÖRDÜNCÜ BÖLÜM

UYGULAMA

Tezin uygulama aşamasında metinsel veri kaynağı olarak Capital dergisi "En Beğenilen Şirketler 2010" araştırmasında adı geçen şirketlerin vizyon ifadeleri kullanılmış olup, ilgili şirketlerin vizyon ifadelerinde itibar kavramına vurgu yapıp yapmadıkları belirlenmek istenmiştir.

4.1. YÖNTEM

Çalışmada ilk aşamada metinsel veri Statistica paket programına aktarılıp, veriler sayısal ve analiz yapılabilecek hale getirilmiştir. Uygulama aşaması için veri kaynağımızı, Capital dergisinin her yıl düzenli olarak yapmış olduğu en beğenilen şirketler araştırmasında yer alan 20 şirketin vizyon ifadeleri oluşturmaktadır. Çalışmanın bir diğer amacı da kaynak şirketlerin vizyon ifadelerinde kullandıkları kelimelerin neler olduğu ve bunların itibarlı şirket olma hususunda itibar kavramına önem veren şirketlere bir strateji oluşturabilmeleri konusunda fikir vermektir.

Capital dergisinin her yıl geleneksel olarak düzenlediği ve 2010 yılında 11.si yapılan "Türkiye'nin En Beğenilen Şirketleri" araştırmasının sonuçlarına göre listede yer alan en beğenilen şirketlerin vizyon ifadeleri metin madenciliği uygulamamızda girdi verilerini oluşturmaktadır. 2010 yılındaki araştırma kapsamında online araştırma tekniği kullanılmış olup iş dünyasını temsilen 7 bin 500 yöneticiye online anket gönderilerek yapılmıştır. (Capital dergisi Aralık 2010: 84)

4.2. ANALİZ

Araştırmanın uygulama kısmı için ilk olarak dergide yer alan şirketlerin listesinden bu şirketlerin vizyon ifadeleri alınmıştır. Vizyon ifadelerinin analizi için Statistica programı kullanılmıştır.

Capital dergisi Aralık 2010 sayısında yayınlanan "Türkiye'nin En Beğenilen Şirketleri" araştırmasında yer alan şirketler;

Tablo 3: Capital Dergisi "En beğenilen Şirketler 2010" Araştırması

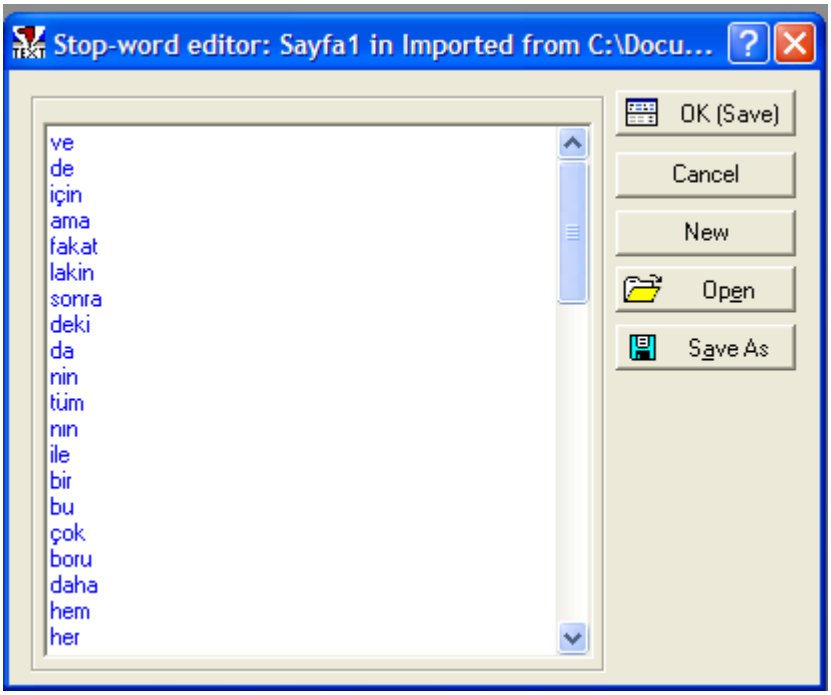
Türkiye'nin En Beğenilen İlk 20 Şirketi			
1	TURKCELL	**11**	MICROSOFT
2	GARANTİ BANKASI	**12**	SABANCI HOLDİNG
3	ARÇELİK	**13**	VODAFONE
4	KOÇ HOLDİNG	**14**	BSH/EFES PİLSEN
5	ECZACIBAŞI TOPLULUĞU	**15**	BORUSAN HOLDİNG
6	COCA-COLA	**16**	VESTEL/SİEMENS
7	UNILEVER	**17**	THY
8	PROCTER&GAMBLE	**18**	FORD
9	ÜLKER	**19**	SHELL/ENKA
10	İŞ BANKASI/DOĞUŞ HOLDİNG	**20**	TÜPRAŞ/ANADOLU GRUBU

Kaynak: Capital, Aralık 2010: 84.

En beğenilen şirketlerin vizyon ifadeleri alınıp, Statistica programı ile metin madenciliği çalışması uygulandığında; şirketlerin vizyon ifadelerinde geçen kelimeler program tarafından sayılarak, metinsel ifadelerin sayısal hale dönüşümü olan metin madenciliğinin ilk aşaması gerçekleştirilmiş olacaktır.

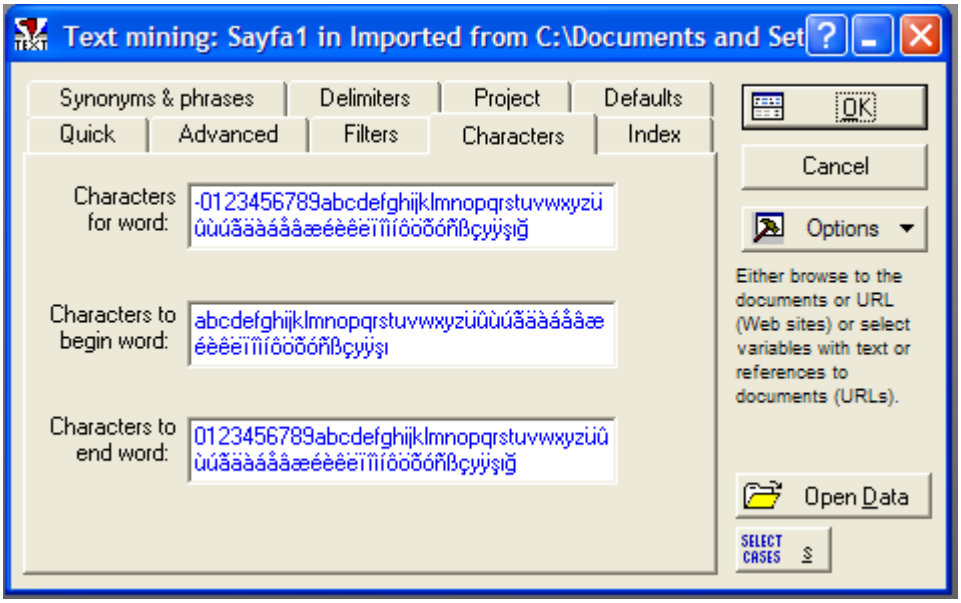
Statistica Programına vizyon ifadelerinin bulunduğu veriler girildikten sonra, programa hazırlık amaçlı olarak programın parametrelerini oluşturabilmek için metin madenciliği modülü uygulanıp, vizyon ifadelerinde geçen kelimeler saydırılır ve anlamsız, analizde bulunması gereksiz kelimelerin neler olduğuna bakılır, "ve" "de" gibi bağlaçlar, ya da fiillerin çekimli halleri tespit edilir. Anlamsız kelimler stop-words (durdurma kelimeleri) dosyasına, fillerin kök ve çekimli halleri synonyms (eş anlamlılar) dosyasına, isim tamlamaları (yan sanayi gibi tek kelime olarak program tarafından sayılacak tamlamalar) da phrases (sözcük grubu) dosyasına kaydedilir. Statistica programı pek çok dilde analiz yapabilmektedir, fakat bu dillerin arasında Türkçe bulunmadığından, Türkçe karakterleri ayrıca programa tanıtmamız gerekmektedir.

Şekil 9 : Durdurma Kelimeleri Listesi (Stop-word)



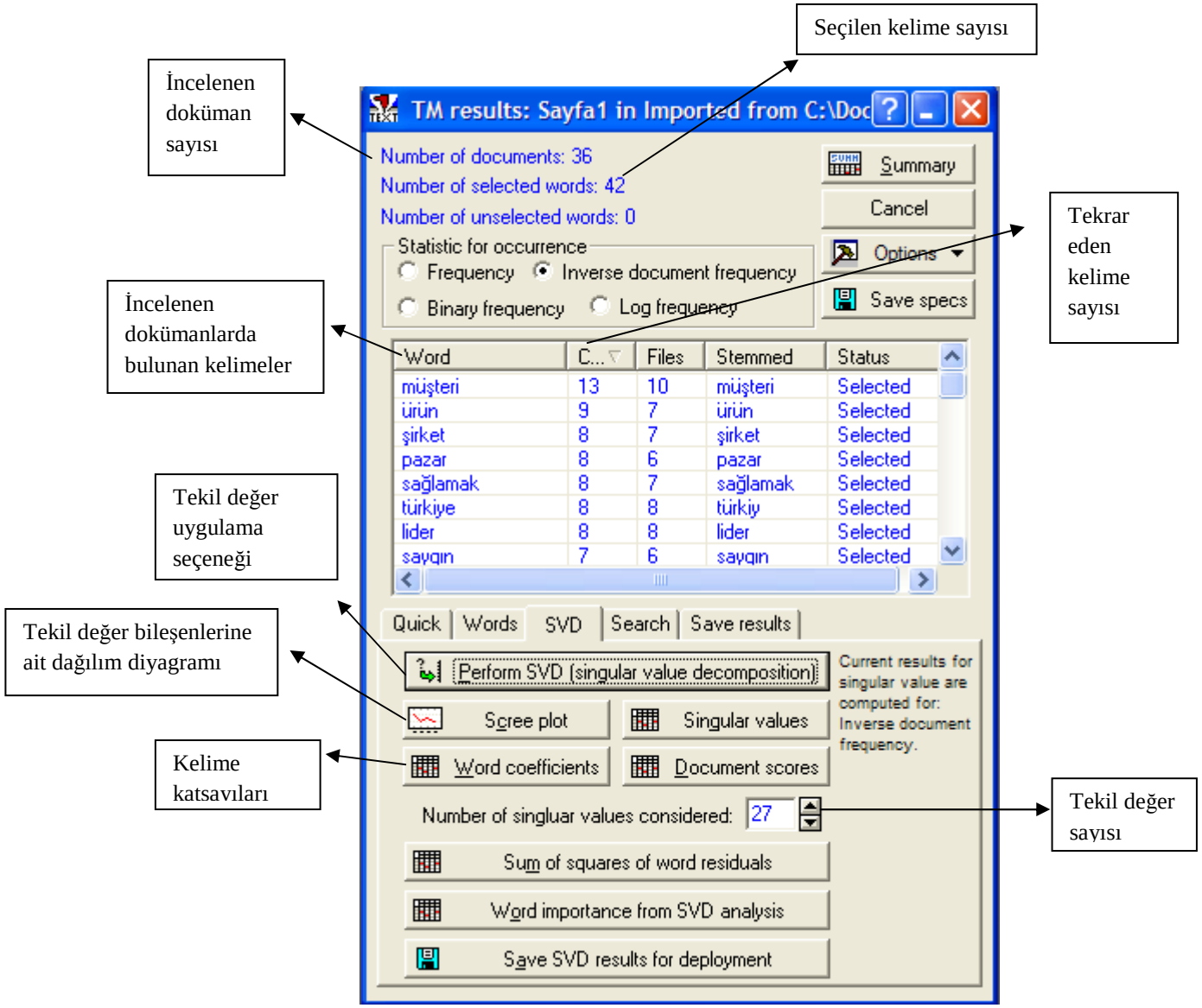
Şekil 10: Eşanlamlı Kelimeler

Şekil 11: Kelimelerde Geçen Harfler



İlk aşama tamamlandıktan sonra aşağıdaki şekilde de görüldüğü üzere program metin madenciliği sonuç ekranında incelenen doküman sayısı 36, analize dahil edilmesini istemediğimiz durdurma kelimeleri, kelimelerin çekimli halleri, anlamsız kelimeler vb, gibi kelimeler çıkarıldıktan sonra seçilen kelime sayısı 42'dir. Metin madenciliği sonuç ekranında yine şekilde görülüğü gibi saydırılan kelimelere tüm belgelerde görülen kelimeleri "0" ve sadece bir belgede görülen kelimelere "1" değerini vererek ağırlıklandırma yapan ters doküman frekansı dönüşümü uygulanmıştır. En beğenilen şirketlerin vizyon ifadelerinde en çok tekrar eden kelimelerin; müşteri, ürün, şirket, pazar gibi kelimelerin olduğu görülmektedir. Analiz için bir sonraki adım ise verilere tekil değer ayrışımının uygulanması aşamasıdır.

Şekil 12 : Statistica Text Miner Sonuç Erkanı



Aşağıdaki tabloda ise Statistica Text Miner sonuç ekranında tamamı görülemeyen kelimeler, bu kelimelerin bulunduğu doküman sayısı ve kaç defa tekrarlandığı görülmektedir. Örneğin "Müşteri" kelimesi incelenen 36 dokümandan 10 tanesinde bulunmaktadır ve toplamda 13 kez tekrarlanmıştır.

Tablo 4 : Statistica Programında Vizyon İfadelerde Geçen Kelime Sayıları

Kelime	Sayı	Geçtiği Dosya Sayısı	Kelime	Sayı	Geçtiği Dosya Sayısı
müşteri	13	10	değer	4	4
ürün	9	7	sektör	3	3
sağlamak	8	8	mümkün	2	2
lider	8	8	sürdürülebilir	2	2
Pazar	8	6	yan sanayi	2	2
türkiye	8	8	karlı	2	2
saygın	7	6	yönetim	2	2
iş	5	3	çevre	2	2
hizmet	5	5	iyi	2	2
büyüme	5	3	konumunu	2	2
marka	5	5	hedefler	2	2
gelişim	4	4	odaklı	2	2
şirket	8	7	ortakları	2	2
kalite	4	4	standartlarda	2	2
sürekli	4	4	devamlılık	2	2
topluma	6	6	beğenilen	2	2
tahminini	4	4	tercih	2	2
tüketici	3	2	avrupa	2	2
çalışanların	4	4	amaçlar	2	2
güvenilir	4	4			

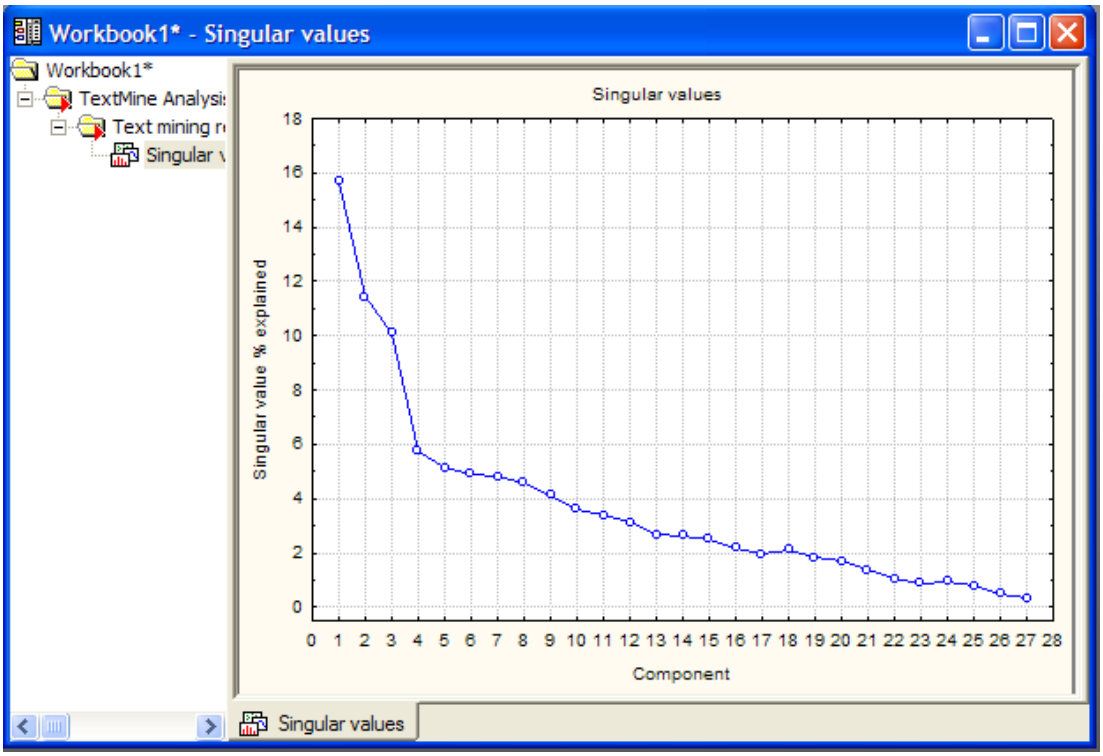
Statistica Programında Tekil Değer Ayrışımı

Tekil değer ayrışımı özellik çıkarımı ve gizli anlamsal endeksleme için metin madenciliğinde tanımlanmıştır. Statistica programı tekil değer ayrışımı için büyük matrislerle bile başa çıkabilecek olan bir algoritma kullanır. A'nın m'nin girdi dokümanlarının (dosyalarının) sayısı ve n'in ise analiz için seçilen kelimelerin sayısı olduğu mxn kelime görünürlülük matrisini gösterdiğini varsayalım. Tekil değer ayrışımı, r'nin ***A'A'***nın öz değerlerinin sayısı olduğu, mxr ortogonal matrisi U'yu, nxr ortogonal matris V ve rxr matris D ve böylece ***A = UDV'*** hesaplar. Statistica Metin Madenciliği ve Doküman Belge Alma (Statistica Text Mining and Document Retrieval) genellikle çok geniş ve seyrek olan A matrisini kullanmak için tekil değer ayrışımını hesaplamada etkili bir tekrarlamalı metot kullanır. Bu metot nispeten büyük tekil değerler için kesin değerleri üretir, fakat analizler için tipik olarak çok az

ilgilenilen küçük tekil değerler üzerideki düşük doğruluğa neden olabilir. Özellikle, bu problem (çok küçük öz değerler için bozulmuş doğruluk) tekil değer ayrışımının boyutluluk azaltma ve özellik seçimi için kullanıldığında metin madenciliğinde önemsizdir ve böylece nispeten küçük tekil değerler ilgi alanında değildir. Ayrıca Statistica programında en büyük tekil değer ayrışımı öz değeri sayısı 82 ile limitlidir.

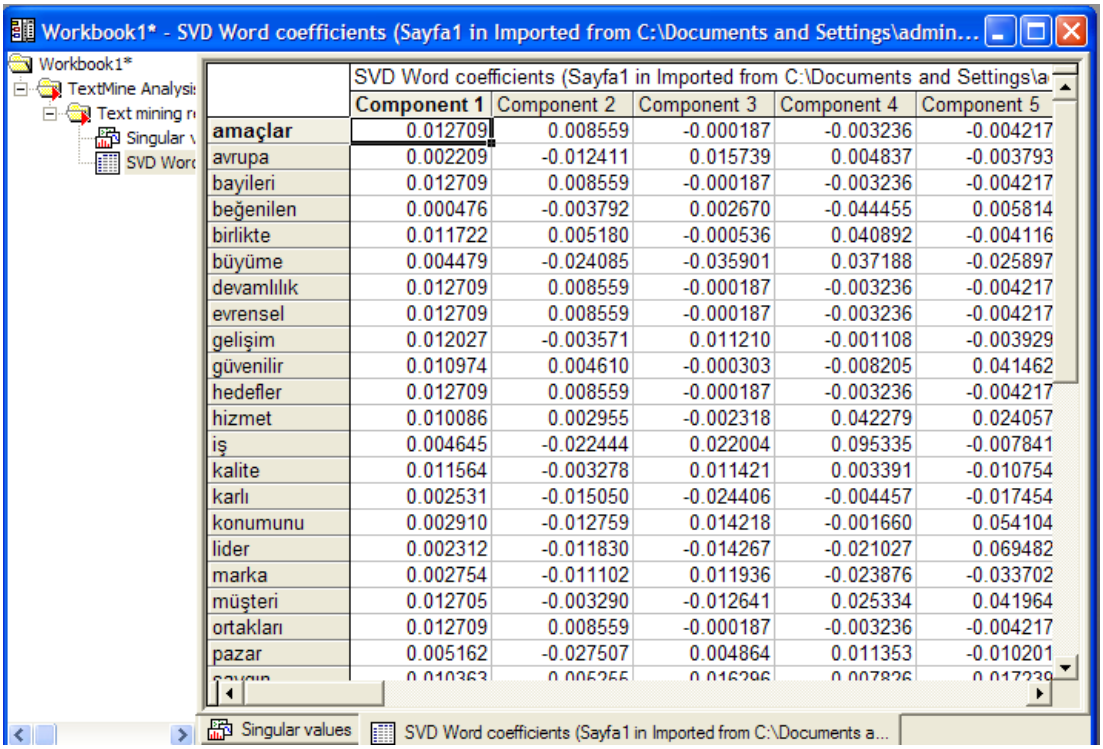
Verilere tekil değer ayrışımı uygulandıktan sonra, analizdeki tekil değer sayısı program tarafından 27 olarak bulunmuştur. Aşağıdaki şekilde 27 tekil değerin analizi açıklama gücü grafiksel olarak görülmektedir. Girdiler arasındaki varyansı açıklayan bileşenleri görsel olarak belirlememize yarayan şekilde, ilk bileşen 42 tane girdinin toplam varyansının yaklaşık olarak %16'sını, bir sonraki bileşen yaklaşık olarak %11.5'ini açıklamaktadır. İlk dört bileşen toplam varyansın yaklaşık olarak %42'sini açıklamaktadır. Bu, faktör analizi, temel bileşenler analizi gibi daha sonra kullanacağımız analizlerde faktör sayısını belirlememizde fikir sahibi olmamızı sağlamaktadır.

Şekil 13: Tekil Değerlere Ait Scree Plot



Aşağıdaki şekilde ise bileşenlerin kelime katsayıları görülmektedir. Bu kelime katsayıları program tarafından analizde belirlenen 27 tane tekil değere sahip bileşenlerden oluşmaktadır. Kelime katsayıları her bir kelimenin ilgili bileşendeki bulunma oranını göstermektedir. Örneğin “amaçlar” kelimesinin birinci bileşene (Component 1) katkısı 0.012709 olarak görülmektedir.

Şekil 14: Kelime Katsayıları



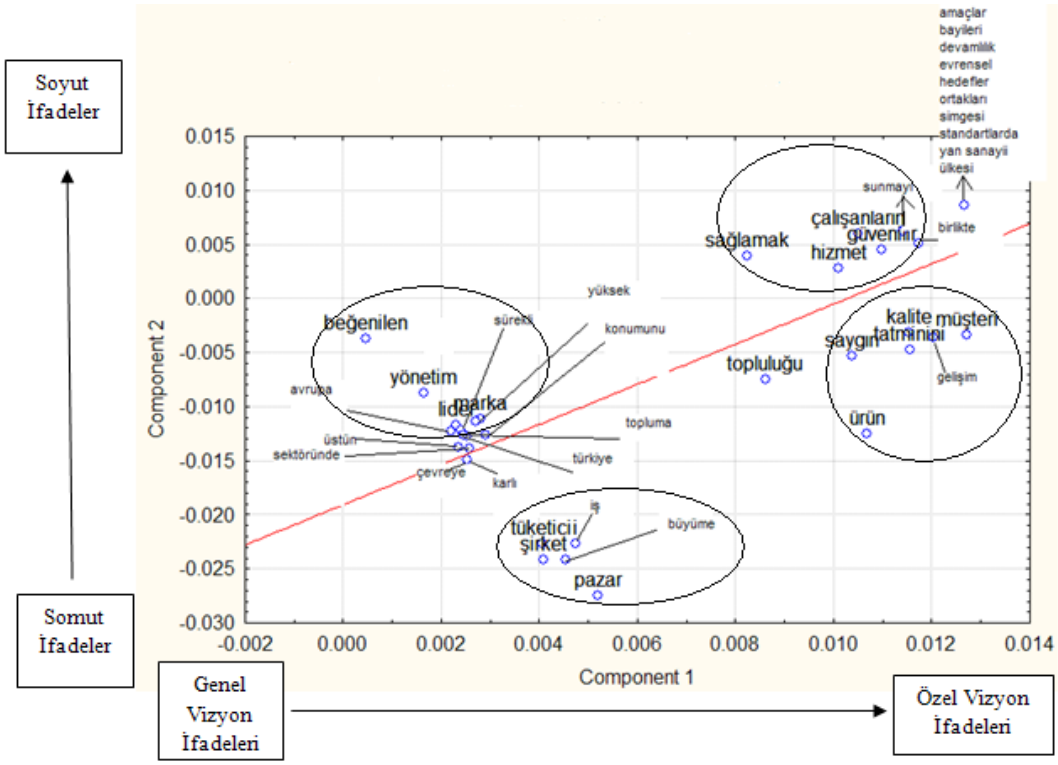
	Component 1	Component 2	Component 3	Component 4	Component 5
amaçlar	0.012709	0.008559	-0.000187	-0.003236	-0.004217
avrupa	0.002209	-0.012411	0.015739	0.004837	-0.003793
bayileri	0.012709	0.008559	-0.000187	-0.003236	-0.004217
beğenilen	0.000476	-0.003792	0.002670	-0.044455	0.005814
birlikte	0.011722	0.005180	-0.000536	0.040892	-0.004116
büyüme	0.004479	-0.024085	-0.035901	0.037188	-0.025897
devamlılık	0.012709	0.008559	-0.000187	-0.003236	-0.004217
evrensel	0.012709	0.008559	-0.000187	-0.003236	-0.004217
gelişim	0.012027	-0.003571	0.011210	-0.001108	-0.003929
güvenilir	0.010974	0.004610	-0.000303	-0.008205	0.041462
hedefler	0.012709	0.008559	-0.000187	-0.003236	-0.004217
hizmet	0.010086	0.002955	-0.002318	0.042279	0.024057
iş	0.004645	-0.022444	0.022004	0.095335	-0.007841
kalite	0.011564	-0.003278	0.011421	0.003391	-0.010754
karlı	0.002531	-0.015050	-0.024406	-0.004457	-0.017454
konumunu	0.002910	-0.012759	0.014218	-0.001660	0.054104
lider	0.002312	-0.011830	-0.014267	-0.021027	0.069482
marka	0.002754	-0.011102	0.011936	-0.023876	-0.033702
müşteri	0.012705	-0.003290	-0.012641	0.025334	0.041964
ortakları	0.012709	0.008559	-0.000187	-0.003236	-0.004217
pazar	0.005162	-0.027507	0.004864	0.011353	-0.010201

Kelime katsayıları uyarınca, tüm girdilerin toplam varyansını en büyük oranda açıklayan ilk iki bileşenden elde edilen dağılım grafiğinde (şekil 15) ise kalite, müşteri, ürün, tatmin, gelişim gibi kelimelerin bir grup, tüketici, şirket, iş, pazar, büyüme gibi kelimelerin bir grup lider, marka, yönetim, avrupa, üstün karlı gibi kelimelerin ise bir grup oluşturduğu ve bu kelimelerin birbirleri ile anlamlı ilişki içerisinde oldukları görülmektedir.

Yatay eksendeki birinci bileşen vizyon ifadelerinin genelden özele doğru bir sıralamasını içermektedir. Sol tarafta “marka”, “şirket”, “pazar” gibi kelimeler bulunurken sağ tarafta “ürün”, “müşteri” gibi özel ifadeler yer almaktadır.

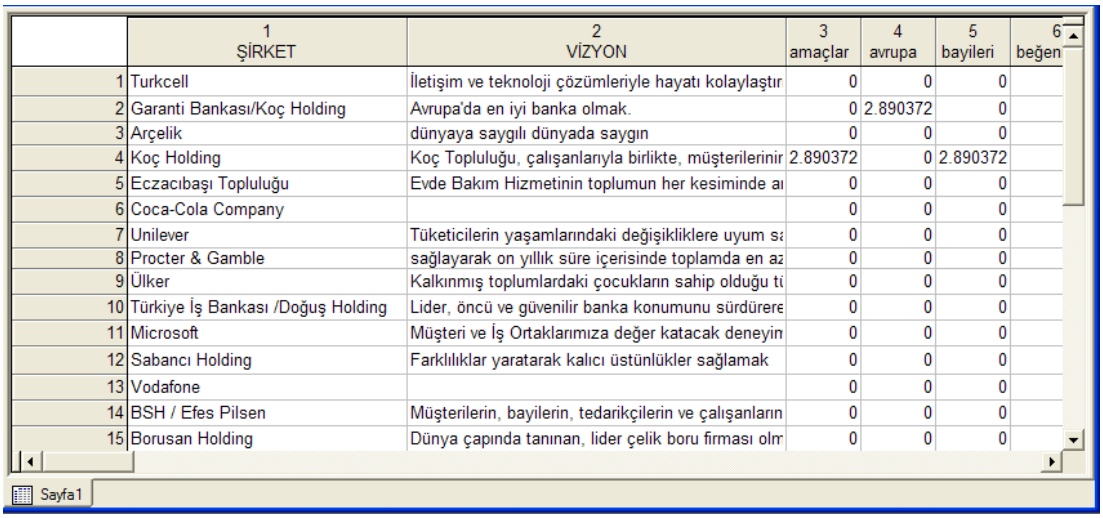
Dikey eksende bulunan ikinci bileşende ise somut değerli ifadelerden soyut değerli ifadelere doğru yayıldığını göstermektedir. “İş”, “pazar”, “çevre” gibi somut ifadelerden “sunmak”, “sağlamak”, “güvenilirlilik” gibi daha soyut ifadelere geçiş olduğu görülmektedir.

Şekil 15: En Yüksek Varyansa Sahip İlk İki Bileşene Ait Scatter Plot



Tekil değerler ve toplam varyansı açıklama yüzdeleri incelendikten sonra, metinsel verileri sayısal haldeki verilerin bulunduğu dosyaya aktarma işlemini gerçekleştirmemiz gerekmektedir. Aşağıdaki şekilde de görüldüğü üzere “amaçlar”, “avrupa”, “bayileri”, “beğenilen” gibi analizde programın çıkarmış olduğu 42 kelime ağırlık değerleri ile birlikte verilerin bulunduğu tabloya eklenir.

Şekil 16: Vizyon İfadelerinde Geçen Kelimelerin Ana Tabloya Aktarılmış Hali

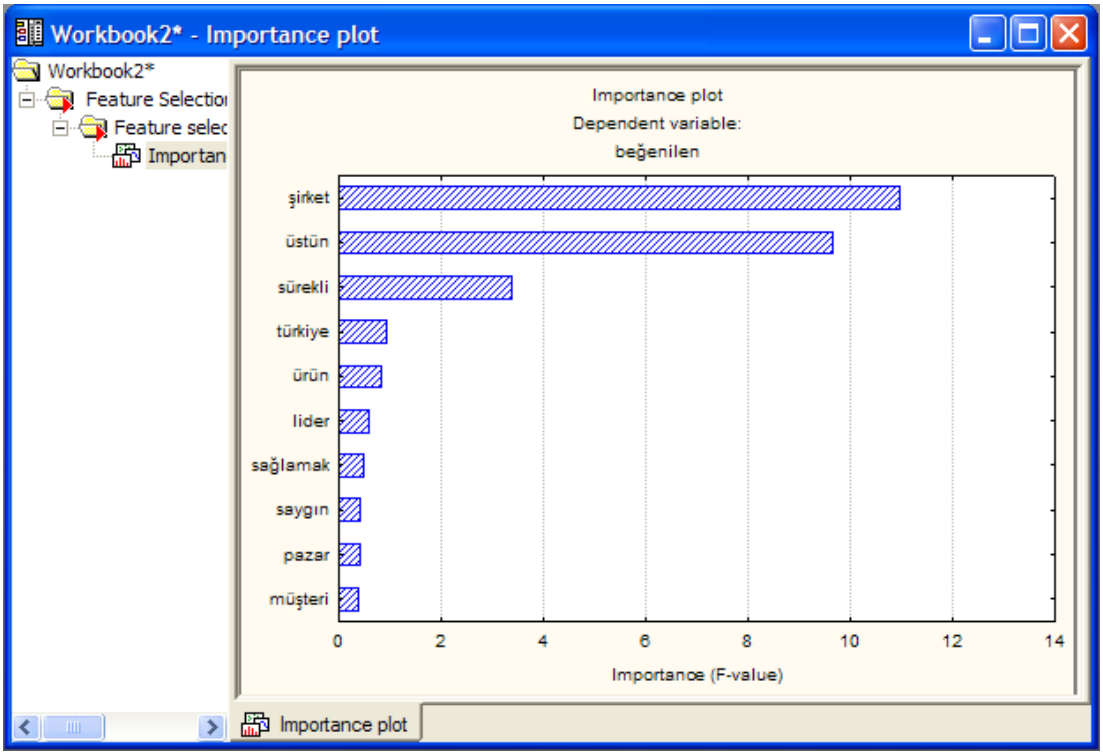


	1 ŞİRKET	2 VİZYON	3 amaçlar	4 avrupa	5 bayileri	6 beğen
1	Turkcell	İletişim ve teknoloji çözümleriyle hayatı kolaylaştır	0	0	0	
2	Garanti Bankası/Koç Holding	Avrupa'da en iyi banka olmak.	0	2.890372	0	
3	Arçelik	dünyaya saygılı dünyada saygın	0	0	0	
4	Koç Holding	Koç Topluluğu, çalışanlarıyla birlikte, müşterilerinir	2.890372	0	2.890372	
5	Eczacıbaşı Topluluğu	Evde Bakım Hizmetinin toplumun her kesiminde ar	0	0	0	
6	Coca-Cola Company		0	0	0	
7	Unilever	Tüketicilerin yaşamlarındaki değişikliklere uyum sa	0	0	0	
8	Procter & Gamble	sağlayarak on yıllık süre içerisinde toplamda en az	0	0	0	
9	Ülker	Kalkınmış toplumlardaki çocukların sahip olduğu ti	0	0	0	
10	Türkiye İş Bankası /Doğuş Holding	Lider, öncü ve güvenilir banka konumunu sürdüre	0	0	0	
11	Microsoft	Müşteri ve İş Ortaklarımıza değer katacak deneyin	0	0	0	
12	Sabancı Holding	Farklılıklar yaratarak kalıcı üstünlükler sağlamak	0	0	0	
13	Vodafone		0	0	0	
14	BSH / Efes Pilsen	Müşterilerin, bayilerin, tedarikçilerin ve çalışanların	0	0	0	
15	Borusan Holding	Dünya çapında tanınan, lider çelik boru firması olm	0	0	0	

Sayfa1

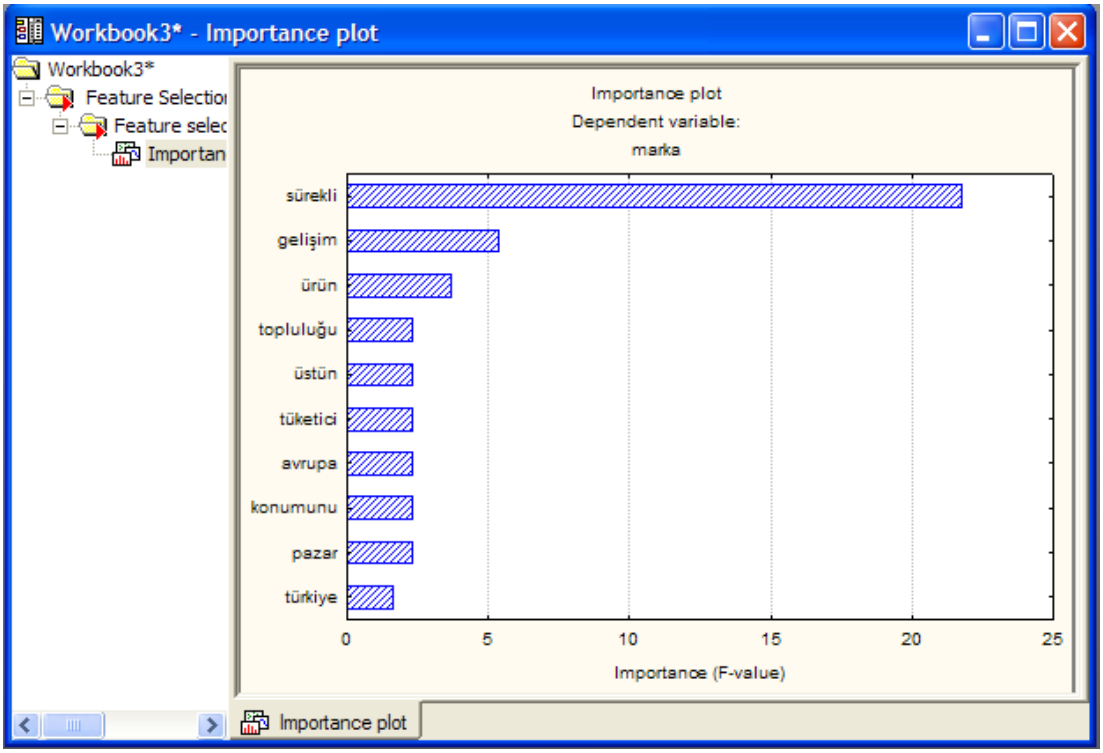
Metinsel ifadeler sayısal hale dönüştürüldükten sonra, veriler üzerinde veri madenciliği analiz yöntemleri uygulanmaya başlanabilir. Bu bağlamda, ilk olarak özellik seçimi aracı ile bağımlı değişken olarak belirlenen bir kelime ve bu kelime ile birlikte tekrar eden kelimelerin histogram grafiği elde edilebilir.

Şekil 17: Bağımlı Değişkenin "Beğenilen" Kelimesi Olması Durumda Özellik Seçimi



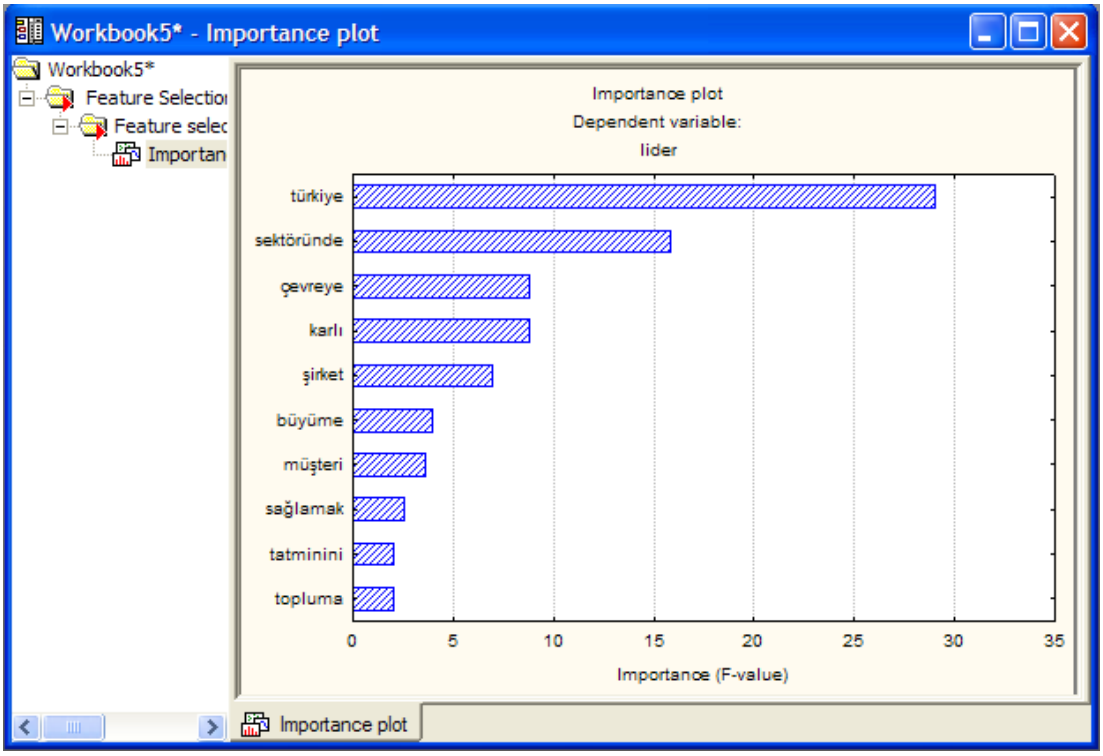
Yukarıdaki şekilde bağımlı değişken olarak belirlenen "beğenilen" kelimesi ve bu kelime ile birlikte en sık tekrarlanan kelimeler görülmektedir. Beğenilen kelimesi ile en sık tekrar eden kelime "şirket", daha sonra "üstün" kelimesi olarak görülmektedir. Şirketlerin "üstün" olma isteklerini rekabette öne çıkmada beğenilme ile ilişkilendirilebilir. Aynı şekilde beğenilen şirket olmak isteyen şirketlerin "sürekli" olmayı, "Türkiye"de beğenilen şirket olmayı hedefledikleri, vizyon ifadelerinde dünyada beğenilen şirket olmayı vurgulamadıkları görülmektedir. Ayrıca şekilden de görülebileceği gibi "ürün" kelimesi "şirket", "üstün" kelimelerinden sonra yer almaktadır bu da göstermektedir ki şirketlerin beğenilen şirket olmada ürün vurgusuna yeterince önem vermedikleri anlaşılmaktadır. "Pazar", "müşteri" kavramları ise listenin en alt sıralarında yer almışlardır, şirketlerin vizyon ifadelerinden yola çıkılarak beğenilen şirket olmada Pazar ve müşteri odaklılık yeterince vurgulanmamaktadır.

Şekil 18: Bağımlı Değişkenin "Marka" Olması Durumunda Özellik Seçimi



Bağımlı değişken "marka" olarak belirlendiğinde ise en çok geçen kelimeler sırası ile "sürekli", "gelişim", "ürün" gibi kelimelerdir. Şirketler açısından marka olmanın önemliliği markanın sürekliliğinin olması, markanın kendini geliştirmesi, şirketlerin ürettiği ürünlerin marka olması, pazarda üstün olmak istemeleri gibi kriterleri göz önüne alacak olursak özellik seçimi aracının çıkardığı sonuçların anlamlı olduğunu söyleyebiliriz. Analize veri oluşturan şirket vizyonlarından yola çıkılarak yukarıdaki şekilden de anlaşılacağı gibi söz konusu şirketlerin "Avrupa"da marka olmayı istemeleri "Pazar"da ve "Türkiye"de marka olmaktan daha öncelikli görmektedirler. Bu durum söz konusu şirketlerin ülke çapında markalaştıklarına inandıkları ve Avrupa'da markalaşmayı ön planda istedikleri düşüncesini akla getirmektedir.

Şekil 19: Bağımlı Değişkenin "Lider" Olması Durumunda Özellik Seçimi



Özellik seçim aracını kullanarak "lider" kelimesinin bağımlı değişken yapıldığı (Şekil 19) analizde ise şirketlerin vizyon ifadelerinde Türkiye'de lider olma, sektörde lider olma gibi kavramları önemsedikleri gibi bir sonuca varılabilir. Şirketlerin lider olmada aynı zamanda "karlı" olmak istedikleri de görülmektedir. "büyüme" ise daha düşük bir paya sahiptir.

Elde edilen verilere yapılan analizler sonucu tekil değer bileşenlerinin dağılım grafiklerinde toplam varyansın büyük bir oranını açıklayan bileşenlerin ilk dört bileşen olduğu gözlemlenmişti. Bu verilere faktör analizi yapılmadan önce faktör sayısını belirlemede bir önsezi oluşturması açısından temel bileşenler analizi yapıldığında program otomatik olarak beş model belirlemiştir. Bu ön analizler sonucunda beş faktör belirlenerek yapılan faktör analizinde faktörler ve aldıkları yükler karşılaştırıldığında dördüncü ve beşinci faktörlerin anlamlı olmadığına karar verilip, üç faktörlü, varimax yönteminin kullanıldığı ve en düşük faktör yükü sınırının 0.40 olarak belirlendiği faktör analizi ve kelimelerin faktör yükleri aşağıdaki gibidir.

Tablo 5 : Vizyon İfadelerine Ait Faktörler Ve Yükleri

Değişkenler	Faktör 1	Faktör 2	Faktör 3
amaçlar	0.99232		
avrupa			0.688464
bayileri	0.99232		
beğenilen			
birlikte	0.82774		
büyüme		-0.904207	
devamlılık	0.99232		
evrensel	0.99232		
gelişim	0.70990		
güvenilir	0.73093		
hedefler	0.99232		
hizmet	0.64787		
iş			0.628694
kalite	0.68337		
karlı		-0.967783	
konumunu			0.675144
lider		-0.578758	
marka			0.488750
müşteri	0.65565		
ortakları	0.99232		
pazar			0.702349
saygın	0.49204		
sağlamak	0.44191		
sektöründe		-0.867660	
simgesi	0.99232		
standartlarda	0.99232		
sunmayı	0.80750		
sürekli			0.602825
tatminini	0.69060		
topluluğu			0.710235
topluma		-0.671259	
tüketici			0.868373
türkiye		-0.599697	

Tablo 5: Devam

yan sanayii	0.99232		
yönetim		-0.487170	
yüksek			0.558797
çalışanların	0.71649		
çevreye		-0.967783	
ülkesi	0.99232		
ürün			0.709151
üstün			0.774386
şirket			0.662513
Expl.Var	15.45962	6.164422	6.107625
Prp.Totl	0.36809	0.146772	0.145420

Şekil 20: Üç Faktörün Özdeğerleri, Varyansları, Toplam Özdeğerleri ve Toplam Varyansları

Value	Eigenvalues (25.08.11) Extraction: Principal components			
	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %
1	15.50920	36.92668	15.50920	36.92668
2	6.20945	14.78440	21.71865	51.71107
3	6.01301	14.31670	27.73167	66.02777

Yukarıdaki tablodan da görülebileceği gibi birinci faktör toplam varyansın %36.93, ikinci faktör %14.78, üçüncü faktör de %14.32'lik kısmını açıklamaktadır. Üç faktör toplamda tüm varyansın %66.02'lik kısmını açıklamaktadır.

Faktör analizi sonuçlarına göre; birinci faktöre yer alan ifadelerin ortak özellikleri dikkate alındığında itibar kriterlerine göre, ilk faktör "Beklenen İmaj" olarak adlandırılabilirken, faktör iki "Yönetsel performans göstergeleri", faktör üç ise "Konumlandırma" olarak adlandırılmıştır.

Faktör analizi sonucunda tablodan da görülebileceği gibi "beğenilen" kelimesinin 0.40 lık faktör yükü limitinde her üç faktör grubuna da ait olmadığının gözlemlenmesidir.

Her bir faktöre ait olan kelimeler ve faktör isimleri aşağıdaki tabloda gösterilmiştir;

Tablo 6 : İtibar Boyutlarına Göre Faktörlere Verilen İsimler Ve Her Bir Faktöre Ait Kelimeler

Faktör 1: Beklenen İmaj	**Faktör 2: Yönetsel performans göstergeleri**	**Faktör 3: Konumlandırma**
Amaçlar	Büyüme	Avrupa
Bayileri	Karlı	İş
Birlikte	Lider	Konumunu
Devamlılık	Sektöründe	Marka
Evrensel	Topluma	Pazar
Gelişim	Türkiye	Sürekli
Güvenilir	Yönetim	Topluluğu
Hedefler	Çevreye	Tüketici
Hizmet		Yüksek
Kalite		Ürün
Müşteri		Üstün
Ortakları		Şirket
Sağlamak		
Saygın		
Simgesi		
Standartlarda		
Sunmayı		
Tatminini		
Yan sanayi		
Çalışanların		
Ülkesi		

Faktör analizi mevcut verilerden elde edilen değişkenlerin sayısını azaltmamıza yardımcı olmuştur, ayrıca her bir faktöre ait olan kelimelerin ortak özellikleri göz önüne alınarak faktörlerin her birine itibar boyutları uyarınca isim vermek, analiz için girdi oluşturan şirketlerin vizyon değerlerinin, itibarın hangi boyutları ile örtüştüğünü görebilmek adına faktör analizinin yapılması uygun görülmüştür.

4.3 SONUÇLAR VE YORUMLAR

Analizde şirketlerin vizyon ifadelerinde geçen kelimelere yapılan faktör analizi sonucunda, en dikkat çeken unsur “beğenilen” kelimesinin hiçbir faktörde yük almaması olmuştur. Bu durum analize girdi oluşturan şirketlerin vizyon ifadelerinde “beğenilme” konusuna yeterince vurgu yapmamış olduklarını göstermektedir. Bununla birlikte “devamlılık, gelişim, güvenilir, kalite, hizmet, müşteri, saygın, çalışanlar” gibi kelimelerin birinci faktörde yer alarak itibarın “Beklenen İmaj” boyutu ile örtüşmesi, “büyüme, karlı, sektör, toplum, yönetim, çevre” gibi anlam olarak yakın ve itibarın “Yönetsel Performans Göstergeleri” boyutu ile ilişkilendirilebilecek şekilde ikinci faktörde yer alması, “Avrupa, iş, marka, pazar, tüketici, ürün” gibi kelimelerin de aynı şekilde itibarın “Konumlandırma” boyutu ile ilişkilendirilebilmesi analizin anlamlı bir sonuç verdiğinin bir göstergesi olmuştur.

SONUÇ

Veri madenciliği, yöneylem araştırması alanında sıkça kullanılan bir konu olması nedeniyle günümüzde sürekli gelişen, araştırma yapılan, teorik ve uygulama alanlarında çalışmalar yapılan bir alandır. Verilerin analiz edilebilmesinde kullanılacak istatistiksel yöntemlerden önce verilerin analize hazır hale getirilmesinde, mevcut ve çok sayıda bulunan veritabanlarından kullanışlı, yararlı ve doğru bilgi çıkarılabilmesinde veri madenciliği önemli bir yer teşkil etmektedir. Veri madenciliği sayısal halde bulunan verilerden bilgi elde etmede çeşitli istatistiksel yöntemleri kullanabiliyorken, metinsel halde bulunan verilerden de çıkarım yapılabilmesi ihtiyacı doğrultusunda metin madenciliği gibi bir alan da günümüzde oldukça popüler bir alan haline gelmiştir. Tezde teorik yapılarıyla incelenen veri madenciliği ve metin madenciliği konularından sonra uygulama kısmında şirket vizyon ifadelerinin oluşturduğu verilerin Statistica programında analiz edilmesi ile ilk aşamada vizyon ifadelerinde geçen kelimler program tarafından sayılmış, gereksiz olan kelimeler (edatlar, bağlaçlar ve anlamsız kelimeler) analizden çıkarılmış, programda Türkçe dili bulunmadığından Türkçe'de bulunan "ş,ç,ğ,ı" gibi harflerin program tarafından algılanması sağlanmıştır. Şirketlerin vizyon ifadelerinde geçen kelimeler saydırılmış, hangi kelimenin kaç defa tekrarlandığı ile ilgili tablolar çıkalımıştır, vizyon ifadelerinde en çok geçen kelimelerin "müşteri, ürün, lider, Pazar" gibi kelimeler olduğu gözlemlenmiştir. Bu doğrultuda şirketlerin vizyon ifadelerinde bu kelimelerin en çok tekrarlanan kelimeler olması, şirketlerin müşteri kazanma, sahip olunan müşterileri memnun etme, ürünlerini satabilme, Pazar paylarını büyütmek istemeleri, pazarda lider konumunda olmak istemeleri nedeniyle anlamlı olduğu sonucu çıkarılmıştır. Kullanılan verilere veri madenciliği tekniklerinden temel bileşenler analizi, kümeleme analiz, faktör analizi gibi yöntemler uygulanmış, temel bileşenler ve kümeleme analizi sonuçlarının çıkarım yapabilmede faydalı sonuçlar çıkarmadığından sadece faktör analizi sonuçları analize dahil edilmiştir. Faktör analizi sonucu program otomatik olarak beş faktör belirlemiş ancak her bir faktöre ait olan kelimeler incelenip, faktörlere itibar boyutları uyarınca isim verme sırasında analizdeki faktör sayısının üç olarak belirlenmesine karar verilmiştir. Üç faktörlü varmiax yöntemi kullanılarak ve 0.40 olarak belirlenen faktör

yükü ile yapılan faktör analizi sonucunda, her bir faktöre ait olan vizyon ifadelerinde geçen kelimelerden, itibar boyutları uyarınca yapılan çıkarım sonucunda birinci faktör “Beklenen İmaj”, ikinci faktör “ Yönetsel Performans Göstergeleri”, üçüncü faktör ise “Konumlandırma” olarak adlandırılmıştır. Çalışmanın en dikkat çeken noktası ise “beğenilen” kelimesinin söz konusu şirketlerin vizyon ifadelerinde yeterince vurgulanmamasından dolayı hiçbir faktör grubuna dahil olmaması olmuştur. Bu konuda söz konusu şirketlerin, itibarlı şirket olma ve itibarlarını sürdürebilmeleri açısından beğenilen şirket olma hususunda vizyon ifadelerinde değişikliğe gitmeleri tavsiye edilebilir.

Metin madenciliği hususunda bir diğer önemli nokta ise, mevcut verilerin fazla olması analizden çıkan sonucun güvenirliliği ve doğruluğu açısından önemli olmasının yanı sıra eldeki metinsel verilerin düzgün ifadelerle yazılmış olması gerekliliğidir. Konu ile ilgili daha önce yapılan ve öğrenciler tarafından doldurulması istenen bir anket çalışmasında, ifadelerin düzgün olarak yapılmadığı ve elde edilen anket geribildirimlerinin sayıca fazla olmamasından dolayı analiz sonuçlarından anlamlı bir çıkarım yapılamamış olması nedeniyle vazgeçilmiştir.

Metin madenciliği gelişmekte olan bir alandı ve konu ile ilgili bir çok uygulama yapılarak kullanılan bilgisayar programlarının geliştirilmesi gerekmektedir. Metinsel olarak bulunan çeşitli veri kaynakları bu programlar aracılığı ile kolaylıkla analiz edilebilmekte, araştırmacıların metinsel veri analizi yapabilmesiyle yeni ufuklar belirleyebilmesini sağlayacaktır.

KAYNAKÇA

Adsız, A. (2006). *Metin Madenciliği*. Dönem projesi. Ahmet Yesevi Ünversitesi Bilişim Sistemleri ve Mühendislik Fakültesi Bilgisayar Mühendisliği Bölümü

Argüden, Y. Kuyucu, B. (2003). www.**arguden**.net/**arguden**/UserFiles/File/kitaplar/itibaryonetimi.pdf. (01.08.2011).

Arı, İ. (2011). http://ismailari.com/blog/tekil-deger-ayrisimi-1/. (21.04.2011)

Bilisoly, R. (2008). *Practical Text Mining With Perl*. Amerika: A John Wiley & Sons, INC Publication.

Bot, J. (2007). *Text-mining in the Life-Sciences, an Exploration*. Netherlands: Bioinformatics Track Delft University of Technology Delft.

Bülbül, Ş. Güler, F. Kandemir, A. (2009). *Propensity Skor Uygulamalarında Kümeleme Analizinin Test Amaçlı Kullanımı*. 10.Ekonometri ve İstatistik Sempozyumu, Düzenleyen Erzurum Üniversitesi. Erzurum. 27-29 Mayıs 2009.

California State University. http://www.csun.edu/~twang/595DM/Slides/Week2.pdf. (16.07.2011).

Cios, K. Perdycz, W. Kurgan, L. (2007). *Data Mining A Knowledge Discovery*. Springer Science and Business Media, LLC.

Clos, J. Pedrycz, W. Swiniarski, R. Kurgan, L. (2007). *Data Mining a Knowledge Discovery Approach*. USA, New York: Springer Science and Business Media, LLC.

Cuppet, G. (2008). *Evaluating Methods For Developing A Vision As They Apply To Small Landscape Architecture Firms*. Yayınlanmış Yüksek Lisans Tezi. Texas: Faculty of the Graduate School of The University of Texas at Arlington.

Çolakoğlu, E. (2010).
http://akademik.maltepe.edu.tr/~ttbilgin/BIL518/presentations/EsraCOLAKOGLU/Sunum2/Esra%20%C7olako%F0lu%20%20KMeans.pdf (05.04.2011).

Davies, G. Chun, R. Sılva, R. Roper, S. (2003). *Corporate Reputation and Competitiveness*. Taylor & Francis e-Library.

Demir, F. (2010). Kurumsal İtibar Ölçümünde Kişiselleştirme Metaforu. fbe.emu.edu.tr/journal/doc/9-10/13.pdf (05.09.2011).

Do, T. Hui,S. Fong, A. (2006). "Associative Feature Selection for Text Mining". *International Journal of Information Technology*. 12 (4), 59-68.

Dolgun, Ö. Özdemir, T. Doruk, O. (2009). "Veri madenciliği'nde yapısal olmayan verinin analizi: Metin ve web madenciliği". *İstatistikçiler Dergisi*. 2(2009) , 48-58.
.
Doorley, J. Garcia, H. (2007). *Reputation Management The Key to Successful Public Relations and Corporate Communication*. Taylor & Francis Group, LLC.

Edward J. Wegman and Jeffrey L. Solka. (2005). *Handbook of Statistics, Vol. 24*. Published by Elsevier B.V.

Elearn Limited, (2005). *Management Extra Reputation Management*. Wordwide Learning Limited adapted by Elearn Limited.

Er,G. (2008). "*Sanal Ortamda Kurumsal İtibar Yönetimi*". Yüksek Lisans Tezi. İzmir: Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü.

Eryılmaz, M. (2008). Örgüt İtibarı Kavramı Ve Yönetimi İle İlgili Bazı Sorunlar. *Anadolu Üniversitesi Sosyal Bilimler Dergisi*. 8(1), 155–174.

Fombrun, C. Neilsen, K. Trad, N. (2007). *The Two Faces Of Reputation Risk: Anticipating Downside Losses While Exploiting Upside Gains*. www.reputationinstitute.com (01.09.2011).

Gao, L. Chang, E. Han, S. (2005). World Academy of Science, Engineering and Technology. http://www.waset.org/journals/waset/v8/v8-21.pdf (15.01.2011).

Griffin, A. (2008). *New Strategies For Reputation Management*. Kogan Page Limited London and Philedalphia.

Han, J. Kamber, M. (2000). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers.

History of Text Data Mining. www.datawg.com/data-mining/49-history-of-text-data-mining.html. (02.05.2011).

Gaizauskas, R. (2002). An Information Extraction Perspective on Text Mining: Tasks, Technologies and Prototype Applications. www.itri.brighton.ac.uk/.../**Text**%20**Mining**%20Event/Rob_Gaizauskas.pdf. (03.05.2011)

Gebze Yüksek Teknoloji Üniversitesi Bilgisayar Mühendisliği Bölümü. bilmuh.gyte.edu.tr/~htakci/vm/kumeleme_analizi.doc. (12.06.2011).

Gülbandılar, E. http://mf.dumlupinar.edu.tr/~eyup/DM/dm2.pdf. (10.05.2011)

Gürgen, H. (2008). www.iso.org.tr/kongre/kongre_2008/.../2a-3-**haluk-gurgen**.ppt (20.08.2011)

Kadıbeşegil, S. (2006). *İtibar Yönetimi-İtibarınızı Yönetmekten Daha Önemli İşiniz Var Mı?*. İstanbul: Kapital Medya Hizmetleri A.Ş.

Karakılıç, N. (2005). Kurumsal İtibarın Müşteri Tercihleri Üzerine Etkileri : Afyon'da Perakende Sektöründe Faaliyet Gösteren İşletmeler Üzerine Bir Araştırma. *Afyon Kocatepe Üniversitesi, İ.İ.B.F. Dergisi.* 7(2), 181-196 .

Kenar, S. *www.ce.yildiz.edu.tr/mygetfile.php?id=1393*. (20.04.2011)

Kim, R. Dam, E. (2003). "*The Added Value Of Corporate Social Responsibilty*". http://www.triple-value.com/upload/docs/addedvalueofcsr.pdf (13.07.2011)

Kostoff, R. Block, J. (2005). "Factor Matrix Text Filtering and Clustering". *Journal Of The Amerıcan Socıety For Informatıon Scıence And Technology*, 56(9), 946–968.

Küçüksille, E. (2009). *Veri Medenciliği Süreci Kullanılarak Portföy Performansının Değerlendirilmesi ve İMKB Hisse Seneleri Piyasasında Bir Uygulama*. Doktora Tezi. Süleyman Demirel Üniversitesi Sosyal Bilimler Enstitüsü İşletme Anabilim Dalı

Larose, D. (2005). *Dıscovering Knowledge In Data An Introduction To Data Mining*. New Jersey: Published by John Wiley & Sons, Inc.

Lau, K. Lee, K. Ho, Y. (2005). Text Mining for the Hotel Industry. *Cornell Unıversıty DOI:0.1177/0010880405275966* 46 (3,) 344-362.

Nisbet, R. Elder, J. Miner, G. (2009). *Handbook of Statistical Analysis and Data Mining Applications*. Canada: Elsevier.

Oğuzlar, A. (2003). Veri Önişleme. *Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 21 (Temmuz-Aralık), 67-76.

Öksüz, B. (2008). "*Kurumsal İtibar Ve İnsan Kaynakları Yönetimi İlişkisinin İncelenmesi*". Yayınlanmış Yüksek Lisans Tezi. İzmir: Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü.

Rao, C. Wegman, E. Solka, J. (2005). *Handbook Of Statistics Vol.24 Data Mining and Data Visualization*. Elsevier B.V.

Santhanalakshmi, R. Alagarsamy, K. (2011). An Innovatice Approach In Text Mining. *Int. J. Comp. Tech. Appl.*. 2 (1), 193-198.

Sever, H. Oğuz, B. (2003). *Veritabanlarında Bilgı Keşfine Formal Bir Yaklaşım Kısım I: Eşleştırme Sorguları Ve Algorıtmalar*. eprints.rclis.org/bitstream/10760/7348/1/173-204.pdf (12.08.2011.)

Shamma, H. (2007). *A Stakeholder Perspective For Examining Corporate Reputation: An Empirical Study Of The U.S. Wireless Telecommunications Industry*. Yayınlanmış Doktora Tezi. Washington: Department of Marketing School of Business The George Washington University.

Sharp, M. (2001). *TextMining*. http://comminfo.rutgers.edu/~msharp/text_mining.htm. (20 Nisan 2011).

Sönal, M. (2007). *Firmaların Vizyon ve Misyon Bildirgelerinin Analizi*. Yüksek Lisans Tezi. Kahramanmaraş Sütçü İmam Üniversitesi Sosyal Bilimler Enstitüsü İşletme Anabilim Dalı.

StatSoft. http://www.statsoft.com/textbook/text-mining/?button=3. (21.04.2011)

Sumathi, S. Sivanandam, S.N. (2006). *Introduction to Data Mining and Its Applications* . Berlin: Springer-Verlag.

Şatır, Ç. Sümer, F. (.2006). http://if.kocaeli.edu.tr/hitsempozyum2006/kitap/06cigdem_satir_Fulya_Erendag.pdf . (27.08.2011).

Şen, F. (2008). *Veri Madenciliği ile Birliktelik Kurallarının Bulunması*. Yüksek Lisans Tezi. Sakarya Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar ve Bilişim Mühendisliği.

Tekerek, A. (2011). *Akademik Bilişim Konferansı*. ab.org.tr/ab11/bildiri/29.pdf. (01.05.2011).

Ural, E. (2002). *İstanbul Ticaret Üniversitesi*. www.iticu.edu.tr/yayin/dergi2.htm (01.09.2011).

Wang, J. (2006). *Encyclopedia of Data Warehousing and Mining*. Idea Group Inc.

Ünsal, A. Duman, S. *Türkiye' Deki Bankaların Performanslarının Temel Bileşenler Yaklaşımı İle Karşılaştırmalı Analizi* www.ekonometridernegi.org/bildiriler/o1s1.pdf (01.05.2011).

Wikipedia. http://en.wikipedia.org/wiki/Singular_value_decomposition. (21.04.2011)

Ye, N. (2003). *The Handbook of Data Mining*. Lawrence Erlbaum Associates, Inc.

Yıldırım, E. (2010). "Veri Zarflama Analizinde Girdi Ve Çıktıların Belirlenmesindeki Kararsızlık Problemi İçin Temel Bileşenler Analizine Dayalı Bir Çözüm Önerisi". *İstanbul Üniversitesi İşletme Fakültesi Dergisi*. 39 (1), 141-153.

Yıldız Teknik Üniversitesi İstatistik Bölümü. www.ist.yildiz.edu.tr/dersler/dersnotu/Kum-Analiz.doc. (12.03.2011)