T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

SAKARYA ÜNİVERSİTESİ WEB SİTESİ ERİŞİM KAYITLARININ WEB MADENCİLİĞİ İLE ANALİZİ

YÜKSEK LİSANS TEZİ

Halil ARSLAN

Enstitü Anabilim Dalı : ELEKTRONİK VE BİLGİSAYAR EĞİTİMİ

Tez Danışmanı : Yrd. Doç. Dr. Ahmet Turan ÖZCERİT

Haziran 2008

T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

SAKARYA ÜNİVERSİTESİ WEB SİTESİ ERİŞİM KAYITLARININ WEB MADENCİLİĞİ İLE ANALİZİ

YÜKSEK LİSANS TEZİ

Halil ARSLAN

Enstitü Anabilim Dalı : ELEKTRONİK VE BİLGİSAYAR EĞİTİMİ

Bu tez 02 / 06 / 2008 tarihinde aşağıdaki jüri tarafından Oybirliği ile kabul edilmiştir.

Yrd. Doç. Dr. A. Turan ÖZCERİT Jüri Başkanı	Yrd. Doç. Dr. Ahmet ZENGİN Üye	Yrd. Doç. Dr. Hayrettin EVİRGEN Üye

TEŞEKKÜR

Bu tez çalışmasının hazırlanışında bana yol gösteren tüm hocalarıma, özellikle danışman hocam Yrd. Doç. Dr. Ahmet TURAN ÖZCERİT'e, Endüstri Mühendisliği Bölümü Öğretim Üyesi Yrd. Doç. Dr. Gültekin ÇAĞIL'a, İstanbul Üniversitesi Endüstri Mühendisliği öğretim üyesi Yrd. Doç. Dr. Numan ÇELEBİ'ye, tez için gerekli teknik çalışmalara izin veren Bilgi İşlem Dairesi Başkanı Yrd. Doç. Dr. Hayrettin EVİRGEN ve Bilgi İşlem Şube Müdürlüğü personeline, yüksek lisans çalışmalarım boyunca burs imkânı sağlayan Türkiye Bilimsel ve Teknolojik Araştırma Kurumu'na (TÜBİTAK) ve desteklerini esirgemeyen aileme teşekkür ederim.

İÇİNDEKİLER

ŞEKİLLER LİSTESİ

TABLOLAR LİSTESİ

ÖZET

Anahtar kelimeler: Veri Madenciliği, Web Madenciliği, Web Kullanım Madenciliği, Web Kayıt Dosyaları, Web Log.

Veri madenciliği büyük miktarda veri içinden gelecekle ilgili tahmin yapmamızı sağlayacak bağıntı ve kuralların bilgisayar programları kullanarak aranmasıdır. Veri madenciliğinin en önemli hedeflerinden biri anlamsız görünen veri yığınlarının arasındaki gizli bağımlılıkları, desenleri tespit etmek ve elde edilen bilgiyi yararlı bir sonuç üretmek için kullanmaktır.

Web madenciliği ise web dokümanlarından ve servislerinden otomatik olarak bilgi çıkarmak ve keşfetmek için veri madenciliği tekniklerinin kullanılması olarak tanımlanır. Veri madenciliğinin hedefi olarak gösterilen anlamsız görünen veri yığınlarından gizli desenler elde edilmesi web madenciliği açısından da hayati öneme sahiptir. Çok hızlı bir büyüme gösteren web dokümanları ve kullanım verileri gizli bağlantılar çıkarılması açısından web madenciliği çalışmalarına ihtiyaç duymaktadır. Web madenciliği, web içerik madenciliği, web yapı madenciliği ve web kullanım madenciliği olmak üzere üç ana başlıkta incelenebilir.

Bu çalışmanın amacı, Web Madenciliği teknikleri kullanılarak yapılan çalışmaları incelemek ve bu çalışmalar ışığında Sakarya Üniversitesi Web Sitesini analiz etmektir. Bu tez çalışmasında, öncelikle veri madenciliğine değinilmiş, ardından veri tabanlarında bilgi keşfi sürecinden bahsedilmiş, devamında web madenciliği disiplini detaylı olarak sunulmuştur. Literatür çalışmaları ışığında Sakarya Üniversitesi web sitesinin web madenciliği ile analiz uygulaması verilmiş, uygulamanın net sonuçlar ortaya koyabilmesi açısından web sitesi erişim kayıtları web madenciliği tekniklerine uygun olarak hazırlanmıştır. Uygulama sonuçları ile ne tür kazanımlar elde edilebileceği vurgulanmıştır.

WEB MINING AND ANALYSIS OF A WEB SITE

SUMMARY

Keywords: Data Mining, Web Mining, Web Usage Mining, Web Access Log, Web Log.

Data mining is to search for relations and rules through large amount of data using computer software, in order to make predictions about future. One of the most important goals of data mining is to detect the secret relations and patterns through data which seems to be meaningless, and use this information to produce a beneficial result.

Web mining is described as using data mining techniques to explore and extract information from web documents and web services. The goal of data mining, detecting secret patterns through data which seems to be meaningless has a very big importance for web mining. Web documents are growing very fast, and usage data need web mining studies for detecting secret relations. Web mining can be examined under three topics: Web content mining, web structure mining, web usage mining. The goal of this study is to examine the studies made by using Web Mining techniques, and to analyze Sakarya University Web Site using similar techniques.

In this thesis firstly, data mining and the data exploration process on databases are studied. After that, disciplines of web mining are introduced. With the knowledge gained by the recent studies, web mining analysis of Sakarya University web site has been made, and in order to gain exact results, web access logs are prepared for web mining techniques. What kind of benefits have been gained are from the application result emphasized.

BÖLÜM 1. GİRİŞ

Son yıllarda World Wide Web (WWW) çok büyük gelişme göstermiştir. WWW'ye her gün 20 Milyon yeni web sayfası eklenmektedir [1]. Aralık 2006'da 105 Milyon'un üstünde Web sitesi ve 800 Milyon aktif internet kullanıcısı olduğu rapor edilmiştir[2]. Bu gelişmelerle birlikte veri boyutları da aynı oranda artmaktadır. Yüksek kapasiteli işlem yapabilme gücünün ucuzlamasının bir sonucu olarak, veri saklama hem daha kolaylaşmış, hem de verinin kendisi ucuzlamıştır [3]. Büyüme işlevleri cinsinden ifade edecek olursak, veri saklama kapasitesi her 9 ayda bir tahmini olarak ikiye katlanmaktadır [4]. Veri tabanlarında saklanan veri, bir arşive benzetilirse, bu veri arşivi tek başına değersizdir ve kullanıcı için çok fazla bir anlam ifade etmez. Ancak bu veri arşivi, belirli bir amaç doğrultusunda sistematik olarak işlenir ve analiz edilirse, değersiz görülen veri yığınında, amaca yönelik sorulara cevap verebilecek çok değerli bilgilere ulaşılabilir. Saklı ve işlenmemiş bilgiye olan bu büyük ihtiyaç Veritabanlarında Bilgi Keşfi (VTBK) ve Veri Madenciliği (VM) gibi yeni alanların keşfiyle anlaşılabilir ve yorumlanabilir bir hale gelmiştir.

Bazı kaynaklara göre; VTBK daha geniş bir disiplin olarak görülmekte ve veri madenciliği terimi sadece bilgi keşfi metotlarıyla uğraşan VTBK sürecinde yer alan bir adım olarak nitelendirilmektedir [5].

VTBK, veri içerisindeki geçerli, yeni, yararlı ve sonuç olarak anlaşılabilir örüntülerin çıkarılması sürecidir. Bu süreç, uygulama alanının öğrenilmesi ile başlar ve uygulamanın amaçları doğrultusunda hedef veri seti seçilir. Daha sonra, gürültülü ve tutarsız verilerin çıkarıldığı veri temizleme ve ön işleme basamağı gelir. Gerekli durumlarda veri, madenciliğe uygun bir forma dönüştürülür. Beşinci basamak olan veri madenciliği, zeki yöntemler aracılığıyla büyük miktarda veriden anlamlı bilgilerin çıkarılması sürecidir. Daha sonra, çıkarılan örüntüler, içlerinden yararlı olanların belirlenmesi için değerlendirilir. VTBK'nin son basamağı ise, elde edilen

bilginin görüntüleme ve bilgi gösterimi yöntemleri kullanılarak kullanıcıya sunulmasıdır [2].

Veri madenciliği VTBK'nın tanımından yola çıkarak büyük miktardaki veriden anlamlı bilginin çıkarılması ile ilgili bir disiplindir. Tanımı detaylandırırsak önceden bilinmeyen fakat yararlı bilginin büyük miktardaki veri arasından bulunup çıkarılmasıdır. Büyük miktardaki veri içindeki örüntünün keşfedilmesini ve geleceğe ilişkin tahminler yapılmasında kullanılabilecek ilişkilerin çıkarılmasıdır [6]. Bu çıkarımların web sitelerinde uygulanmasına ise web madenciliği denilmektedir.

Web madenciliği, WWW üzerinden kullanışlı bilgiyi keşfetme ve analiz etme işlemi, şeklinde geniş olarak tanımlanır. Bu geniş tanım bir yandan, milyonlarca siteden ve çevrimiçi (online) veritabanlarından veri ve kaynakların otomatik olarak aranması ve elde edilmesi işlemi olan Web İçerik Madenciliği'ni tarif ederken, diğer yandan, bir yada daha çok Web sunucusu veya çevrimiçi servisten kullanıcı erişim desenlerinin keşfi ve analizi işlemi olan Web Kullanım Madenciliği'ni tarif eder. Daha sonradan bu iki kategoriye, Web sitelerinin bağlantı (link) yapılarını da kapsayan yapısal özetini üreten Web Yapı Madenciliği de eklenmiştir [3]. Web madenciliği, ilk olarak Etzioni tarafından Web doküman ve servislerinden otomatik olarak bilginin elde edilmesi olarak tanımlanmıştır.

Bu çalışmanın amacı, Web Kullanım Madenciliği teknikleri kullanılarak yapılan çalışmaları incelemek ve bu çalışmalar ışığında Sakarya Üniversitesi Web Sitesini analiz etmektir. 6 bölümden oluşan çalışmanın 2. bölümünde Veri Madenciliği, 3. Bölümünde Veri Tabanından Bilgi Keşfi Süreci, 4. Bölümünde Web Madenciliği, 5. Bölümünde Sakarya Üniversitesi Web Sitesinin Analizi uygulaması ve son olarak 6. Bölümünde Değerlendirme ve Sonuçlar sunulmuştur.

BÖLÜM 2. VERİ MADENCİLİĞİ

2.1. Veri Madenciliği Nedir

Yüksek kapasiteli işlem yapabilme gücünün ucuzlaması ile birlikte veri saklama işlemi kolaylaşmıştır. Fakat son yıllarda, veriyi toplama ve saklama kapasitesindeki çok ani büyüme, yeni arayışlara yol açmıştır. Bir bilgisayarın işleyebileceği veriden daha fazlası üretilmektedir. Verilerin bu hızla büyümesi, yorumlama ve özümsemede akıllı veritabanı analizi için, yeni nesil araçlara ve tekniklere olan ihtiyacı doğurmuştur. Geleneksel sorgu veya raporlama araçları veri yığınları karşısında yetersiz kalmıştır.

Veri madenciliği büyük miktarda veri içinden gelecekle ilgili tahmin yapmamızı sağlayacak bağıntı ve kuralların bilgisayar programları kullanarak aranmasıdır [5]. Veri madenciliği ve veri tabanlarında bilgi keşfi süreci kavramları birçok kaynakta birbirinin yerine kullanılmaktadır. Veri madenciliği, veri tabanlarında bilgi keşfi sürecinde bir adım olmasına rağmen birçok çalışmada tüm süreci anlatmak için kullanılmaktadır. Veri madenciliği ile büyük veri yığınlarından veri tabanı sistemleri içerisinde gizli kalmış bilgilerin çekilmesi sağlanır. Bu işlem, istatistik, matematik disiplinleri, modelleme teknikleri, veri tabanı teknolojisi ve çeşitli bilgisayar programları kullanılarak yapılır [6].

Veri Madenciliği çok büyük veri yığınlarından kritik bilgileri elde etmeyi sağlar. Böylelikle normal şartlar altında uzun zaman süren araştırmalarla doğruluğu kesin olmayacak şekilde elde edilen bilgi veri madenciliği ile kısa sürede ve kesin olarak elde edilir. Elde edilen bu bilgi objektif değerlendirmeler yapılmasında ya da stratejik kararlar almada kullanılır. Bu bilgiler kurumsal veri kaynaklarının iyi analiz edilmesine ve iş dünyasındaki yaklaşımlara ilişkin tahminlerde bulunulmasına yardımcı olur. Kısaca veri madenciliği sayesinde şirketler stratejik adımlar atarken

çok büyük veri yığınları arasından kendilerine yol gösterecek kritik verileri ayıklayarak analiz edebilir [8].

Veri madenciliği uygulamalarından fayda sağlanmasına neden olan en önemli faktörlerden birisi çok miktarda veriyi istediğimiz şekilde işleyebilme olanağıdır. Altı çizilmesi gereken husus ise bu verileri işleme yöntemlerinin aslında senelerdir en temel işletme istatistiği derslerinde de okutulan yöntemler olmasıdır. Yeni olan uygulama, teknolojik olanaklar sayesinde yapılması gereken analizlerin çok daha düşük maliyet ve sürelerde yapabilmesidir. Dolayısıyla zaten çok önemli olduğu kimse tarafından inkâr edilmeyen istatistiksel analizlerin bu denli kolay uygulanabilir olmasından dolayıdır ki istatistik uygulamalarını iş dünyası yeniden keşfetmiştir ve bunun yararlarını her dönemden daha çok ve daha somut bir biçimde şahit olmaktadır. Günümüzde farklı olan bilgisayar ve internet teknolojilerinin sağladığı olanaklardır [10].

Temel olarak veri madenciliği, veri setleri arasındaki desenlerin ya da düzenin, verinin analizi ve yazılım tekniklerinin kullanılması ile ilgilidir. Veriler arasındaki ilişkiyi, kuralları ve özellikleri belirlemekten bilgisayar sorumludur. Amaç, daha önceden fark edilmemiş veri desenlerini tespit edebilmektir.

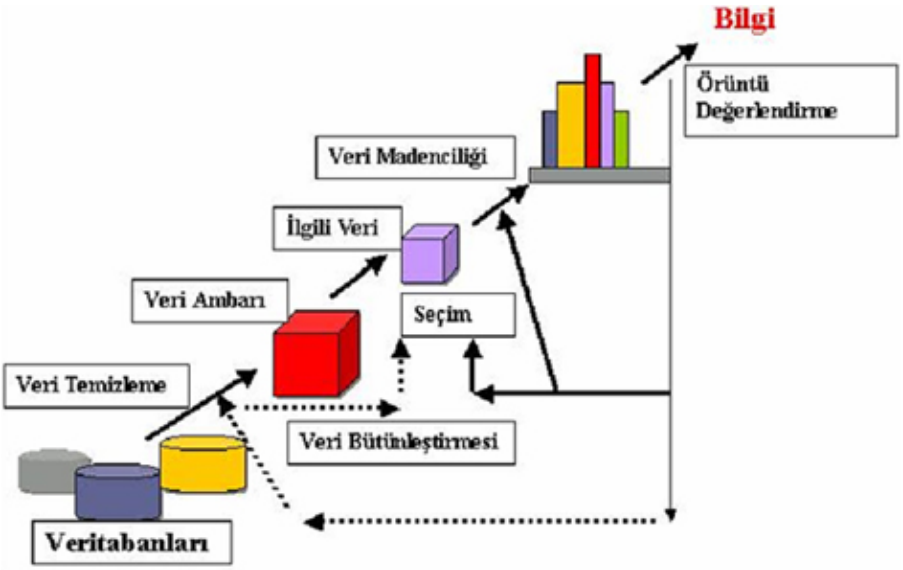
2.2. Veri Madenciliğine Gereksinim Duyulma Nedenleri

Otomatik veri toplama araçları ve veri tabanı teknolojilerindeki gelişme, veritabanlarında, veri ambarlarında ve diğer bilgi depolarında çok miktarda bilgi depolanması sonucunu doğurmuştur. Çok fazla veri var, ancak bilgi yok… Veri ambarları ve veri madenciliği büyük miktarlardaki veriler içindeki gizli örüntüler, geleneksel çözümleme araçlarıyla bulunamaz. Toplanan veri miktarı büyüdükçe ve toplanan verilerdeki karmaşıklık arttıkça, daha iyi çözümleme tekniklerine olan gereksinim de artmaktadır. Bu tür bilgiler, bilgi bulma/keşfetme (knowledge discovery) ya da veri madenciliği (data mining) olarak bilinen teknikler yardımıyla çözümlenebilir [7].

Veritabanı yönetim sistemleri (VTYS) büyük miktardaki yapısal bilgiyi saklama ve etkin bir biçimde erişim sağlamakla yükümlüdür. VTYS'lerde veri düzenlemesi, ilgili organizasyonun işletimsel veri ihtiyacı doğrultusunda gerçekleştirilir ki bu her zaman bilgi keşfi perspektifi ile bire bir çakışmaz. Bu açıdan veri tabanındaki veriler temizleme, boyut indirgeme, transfer, vb. işlemlerinden geçirilerek veri madenciliği kullanımına sunulur. Veri madenciliği teknikleri ayrı araç olarak sağlanabileceği gibi bir VTYS ile de entegre olabilirler.

Günümüzde, işletmeler rekabet ortamında varlıklarını koruyabilmek için daha hızlı hareket etmeli, daha yüksek kalitede hizmet sunmalı, bütün bunları yaparken de minimum maliyeti ve en az insan gücünü göz önünde bulundurmalıdır. Bu tip hedef ve kısıtların yer aldığı iş dünyasında veri madenciliği, temel teknolojilerden biri haline gelmiştir. Çünkü veri madenciliği sayesinde müşterilerin ve müşteri faaliyetlerinin yarattığı fırsatlar daha kolay tespit edilebilmekte ve riskler daha açık görülebilmektedir.

Veri madenciliğinde araştırma çok farklı disiplinlerin içinde uygulanmaktadır. Veri tabanını yöneten araştırmacılar veri madenciliğinin avantajını sorgu işleminden almaktadır. İlgi çeken alanlardan biri de sorgu genişletmek ve veri madenciliğini kolaylaştırmaktır [9].



Şekil 2.1. Veri madenciliği süreci

2.3. Veri Madenciliğinin Kullanım Alanları

Günümüzde veri madenciliği teknikleri başta işletmeler olmak üzere çeşitli alanlarda başarı ile kullanılmaktadır. Veri madenciliğinin asıl amacı, veri yığınlarından anlamlı bilgiler elde etmek ve bunu eyleme dönüştürecek kararlar için kullanmaktır. Örnek kullanım alanlarını aşağıdaki disiplinlere göre belirtirsek;

Web uygulamalarında:

- Kullanıcı taraflı bilgiler (tarayıcı, dil vb..) ışığında altyapı düzenlemelerine gidilebilir [13].
- Kullanıcıların profilleri çıkarılabilir ve zaman içindeki değişimleri takip edilebilir, sitedeki beğenilen ya da beğenilmeyen köşeler tespit edilebilir [11].
- Kullanıcı profillerine göre site perspektifi düzenlenebilir.
- Site haritası, linkler vs.. düzenlemeleri yapılabilir.
- Kullanıcıların gezinti şekli/hızı sitenin içerik, yapılandırma ve alt-yapı açısından performansı hakkında fikir verir [11].
- Kullanıcı profillerine uygun ürünlerin reklam kampanyaları en çok ziyaret ettikleri sayfalara koyulabilir [12].
- En sık beraber ziyaret edilen çift sayfalar belirlenebilir [12].
- Farklı web şablonları, temaları arasında kullanıcı istekleri değerlendirilebilir.
- Form verilerinin toplanmasındaki zorlukları en aza indirme yöntemleri geliştirilebilir.
- Kötü niyetli kullanıcı istekleri belirlenip bunlara karşı alınması gereken önlemler belirlenebilir [13].

İşletme alanındaki uygulamalar [14]:

- Bir işletme kendi müşterisiyken rakibine giden müşterilerle ilgili analizler yaparak rakiplerini tercih eden müşterilerinin özelliklerini elde edebilir ve bundan yola çıkarak gelecek dönemlerde kaybetme olasılığı olan müşterilerin kimler olabileceği yolunda tahminlerde bulunarak onları kaybetmemek, kaybettiklerini geri kazanmak için strateji geliştirebilir.

- Ürün veya hizmette hangi özelliklerin ne derecede müşteri memnuniyetini etkilediği, hangi özelliklerinden dolayı müşterinin bunları tercih ettiği ortaya çıkarılabilir.
- Müşterilerin kredi riskleri hesaplanarak hangi müşterilerin kredi riskinin yüksek olduğu, hangi müşterilerin geri ödemesini zamanında yapamayabileceği kestirilebilir.
- Kredi kartı ödemelerini aksatan, gecikmeli olarak yapan veya hiç yapmayanların özelliklerinden yola çıkılarak bundan sonra aynı duruma düşebilecek muhtemel kişiler saptanabilir.
- Ürün talebi bazında müşteri görünümlerini belirleyerek, müşteri segmentasyonuna gitmek ve çapraz satış olanakları yaratmakta kullanılabilir.
- Piyasada oluşabilecek değişikliklere mevcut müşteri portföyünün vereceği tepkinin firma üzerinde yaratabileceği etkinin tespitinde kullanılabilir.
- En karlı mevcut müşteriler saptanarak, potansiyel müşteriler arasından en karlı olabilecekler belirlenebilir. Karlı müşteriler tespit edilerek onlara özel kampanyalar uygulanabilir. En masraflı müşteriler daha masrafsız müşteri haline dönüştürülebilir. Örneğin en çok bankacılık işlemi yapanlar ortaya çıkarılıp bunlar şube bankacılığı yerine daha masrafsız internet bankacılığına yönlendirilebilir.
- Bir ürün veya hizmetle ilgili bir kampanya programı oluşturmak için hedef kitlenin seçiminden başlayarak bunun hedef kitleye hangi kanallardan sunulacağı kararına kadar olan süreçte veri madenciliği kullanılabilir.
- Kurum teknik kaynaklarının en uygun şekilde kullanılmasını sağlamakta kullanılabilir.
- Geçmiş ve mevcut yapı analiz edilerek geleceğe yönelik tahminlerde bulunulabilir. Özellikle ciro, karlılık, pazar payı, gibi analizlerde veri madenciliği çok rahat kullanılabilir.

Perakendecilik alanındaki uygulamalar:

- Satış noktası veri analizlerinde,
- Alış-veriş sepeti analizlerinde,
- Tedarik ve mağaza yerleşim optimizasyonunda.

Borsa alanındaki uygulamalar:

- Hisse senedi fiyat tahmininde,
- Genel piyasa analizlerinde,
- Alım-satım stratejilerinin optimizasyonunda.

Telekomünikasyon alanındaki uygulamalar:

- Kalite ve iyileştirme analizlerinde,
- Hisse tespitlerinde,
- Hatların yoğunluk tahminlerinde,
- İletişim desenlerinin belirlenmesinde,
- Kaynakların daha iyi kullanılması,
- Servis kalitesinin arttırılmasında.

Sağlık alanındaki uygulamalar:

- Test sonuçlarının tahmininde,
- Ürün geliştirmelerinde,
- Tıbbi teşhislerde,
- Tedavi sürecinin belirlenmesinde,
- Semptomlara göre hastalık tespitinde,
- Magnetik rezonans verileri ile sinir sistemi bölge ilişkilerinin belirlenmesinde.

Endüstri alanındaki uygulamalar:

- Kalite kontrol analizlerinde,
- Lojistikte,
- Üretim süreçlerinin optimizasyonunda.

Tablo 2.1.'de 2003 yılında yapılan bir araştırma sonucuna göre veri madenciliğinin sektörler bazında kullanımına ilişkin sonuçlar yer almaktadır [15].

Tablo 2.1. Veri madenciliği uygulama alanları.

131 Kişiden Toplam 279 oy	
Bankacılık (37)	13%
Bioteknoloji / Genetik (27)	10%
Pazarlama / Organizasyon (29)	10%
Web (15)	5%
Eğlence / Haber (4)	1%
Sahtekârlık Tespiti (24)	9%
Sigortacılık (23)	8%
Yatırım / Hisse Senedi (8)	3%
İmalat (5)	2%
Medikal (16)	6%
Perakende (17)	6%
Bilimsel Çalışmalar (24)	9%
Güvenlik (6)	2%
Tedarik Zinciri Analizi (3)	1%
Telekomünikasyon (21)	8%
Seyahat (5)	2%
Diğer (12)	4%
Bilinmeyen (3)	1%

2.4. Veri Madenciliği Modelleri

Veri madenciliğinde kullanılan modelleri tahmin edici (Predictive) ve tanımlayıcı (Descriptive) olmak üzere iki ana başlık altında toplayabiliriz. Tahmin edici modeller ile tanımlayıcı modeller arasındaki fark kesin sınırlarla ayrılmamıştır. Tahmin edici modeller anlaşılabilir olduğu ölçüde tanımlayıcı model olarak, tanımlayıcı modeller de tahmin edici model olarak kullanılabilirler [19].

2.4.1. Tanımlayıcı modeller

Tanımlayıcı modeller analiste daha önceden bir hipoteze sahip olmaksızın, veri kümesinin içinde ne tür ilişkiler olduğunu anlama imkânı sunar. Analizcinin çok geniş veri tabanlarındaki bilgileri incelemek, örüntüleri keşfetmek için doğru soruları sorup hipotezler geliştirmesi pratikte zor olduğundan, ilginç örüntüleri keşfetme önceliği veri madenciliği programına bırakılır. Keşfedilen bilginin kalitesi ve zenginliği, uygulamanın kullanışlılığını ve gücünü oluşturur [20].

Tanımlayıcı modellerde karar vermeyi, rehberlik etmede kullanılabilecek mevcut verilerdeki örüntülerin tanımlanması sağlamaktadır. 25 yaş altı bekar kişiler ile, 25 yaş üstü evli kişiler üzerinde yapılan ve ödeme performanslarını gösteren bir analiz tanımlayıcı modellere örnek olarak verilebilir [16]. Kümeleme, birliktelik kuralları, çok kullanılan tanımlayıcı modellerdir.

2.4.2. Tahmin edici modeller

Tahmin, geçmiş tecrübelerden elde edilen bilgiler ve mantık kullanılarak, gelecekte olması muhtemel durumlar hakkında öngörüde bulunmaktır.

Tahmin edici modellerde, sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuç değerlerin tahmin edilmesine çalışılmaktadır [16]. Örneğin bir sınıftaki öğrencilerin bir dersle ilgili almış oldukları vize ve ödev notları gibi veriler bir veritabanında toplanabilir. Bu verilere uygun olarak kurulan model öğrencilerin o dersin sonunda finalden alacağı notun tahmininde kullanılmaktadır.

Tahmin edici modeller karar alma süreçlerinde önemli bir rol oynar. Tahmin edici modellerde sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuç değerlerinin tahmin edilmesi amaçlanır [21]. Tahmin edici modellerin temel iki türü sınıflandırma ve regresyondur.

2.5. Veri Madenciliği Teknikleri

Gerek tanımlayıcı gerekse tahmin edici modellerde yoğun olarak kullanılan belli başlı teknikler; Hipotez Testi Sorgusu, Sınıflama ve Regresyon Sorgusu, Kümeleme Sorgusu, Ardışık Örüntüler, Birliktelik Kurulları olarak sıralanabilir. Sınıflama ve Hipotez Testi modelleri tahmin edici, kümeleme, birliktelik kuralları ve ardışık örüntü modelleri tanımlayıcı modellerdir [17].

2.5.1. Hipotez testi sorgusu

Hipotez testi sorgusu algoritması, doğrulamaya dayalı bir algoritmadır. Bir hipotez öne sürülür ve seçilen veri kümesinde hipotez doğruluğu test edilir. Öne sürülen hipotez genellikle belirli bir örüntünün veritabanındaki varlığıyla ilgili bir tahmindir. Bu tip bir analiz özellikle keşfedilmiş bilginin genişletilmesi veya rötuşlanması işlemleri sırasında yararlıdır.

Hipotez ya mantıksal bir kural ya da mantıksal bir ifade ile gösterilir. Her iki biçimde de seçilen veritabanındaki nitelik alanları kullanılır. X ve Y birer mantıksal ifade olmak üzere "IF X THEN Y" biçiminde bir hipotez öne sürülebilir.

Verilen hipotez, seçilen veritabanında doğruluk ve destek kıstasları temel alınarak sistem tarafından sınanır.

2.5.2. Sınıflandırma ve regresyon sorgusu

Sınıflandırma, veri nesnesini daha önceden belirlenen sınıflardan biriyle eşleştirme sürecidir [28]. Verileri ve karşı gelen sınıfları içeren eğitim kümesi ile eğitilen sistem, sonraki aşamalarda sınıf bilgisine sahip olunmayan verilerin ait olduğu sınıfların bulunması için kullanılır. Sınıflama sorgusu, yeni bir veri elemanını daha önceden belirlenmiş sınıflara atamayı amaçlar. Veritabanında yer alan ifadeler bir sınıflama fonksiyonu yardımıyla kullanıcı tarafından belirlenmiş ya da karar niteliğinin bazı değerlerine göre anlamlı alt sınıflara ayırılır. Sınıflama algoritması bir sınıfı diğerinden ayıran örüntüleri keşfeder. Müşteri segmentasyonu, kredi analizi, iş modellemesi ve benzeri birçok alanda kullanılan sınıflandırma yöntemi günümüzde en çok kullanılan veri madenciliği yöntemidir.

Regresyon, sürekli sayısal bir değişkenin, aralarında doğrusal ya da doğrusal olmayan bir ilişki bulunduğu varsayılan diğer değişkenler yardımıyla tahmin edilmesi yöntemidir [22].

Regresyon modeli, sayısal değerleri tahmin etmeye yönelik olması dışında sınıflandırma yöntemine benzetilebilir. Sınıflama gruplanacak verileri tahmin ederken, regresyon süreklilik gösteren değerlerin tahmin edilmesinde kullanılır. Çok

terimli lojistik regresyon gibi kategorik değerlerin de tahmin edilmesine olanaklı tekniklerin geliştirilmesi ile sınıflandırma ve regresyon modelleri giderek birbirine yaklaşmakta ve dolayısıyla aynı tekniklerden yararlanılması mümkün olmaktadır.

Sınıflama ve regresyon modellerinde kullanılan başlıca teknikler [17],

- K-En Yakın Komşu,
- Genetik Algoritmalar,
- Yapay Sinir Ağları,
- Naïve-Bayes,
- Doğrusal Regresyon, Lojistik Regresyon,
- Karar Ağaçları olarak verilebilir.

2.5.2.1. K-En yakın komşu algoritması

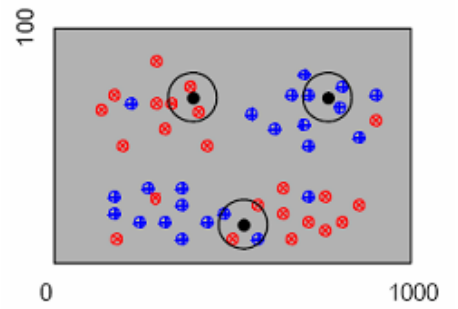
İnsanlar yeni problemleri çözmeye çalışırken genellikle daha önce çözdükleri benzer problemlerin çözümlerine bakarlar. Bu teknikte yeni bir durum daha önce sınıflandırılmış benzer, en yakın komşuluktaki k tane olaya bakılarak sınıflandırılır. K en yakın komşuluğundaki olayların ait olduğu sınıflar sayılır ve yeni durum sayısı fazla olan sınıfa dahil edilir [27]. Bu yöntemde ilk olarak nitelikler arasındaki mesafeyi ölçmek için bir ölçme yöntemi oluşturulur. Olaylar arasındaki uzaklıklar hesaplandıktan sonra, yeni olayların sınıflandırılması için hâlihazırda sınıflandırılmış olan durumlar temel olarak alınır. Uzaklık karşılaştırmasına kaç adet olayın dahil edileceği (k'nın belirlenmesi) ve komşuluk hesaplamalarının nasıl yapılacağına karar verilir. Komşuluk hesaplamaları yapılırken, daha yakın komşulara daha büyük ağırlık değerleri atanabilir [29].

Bu yöntemin tercih edilme sebebi, sayısı bilinen veri kümeleri için hızlı ve verimli olmasıdır [25]. Kayıtlar, bir veri uzayındaki noktalar olarak düşünülürse, birbirine yakın olan kayıtlar, birbirinin civarında (yakın komşusu) olur. K en yakın komşuluğunda temel düşünce "komşunun yaptığı gibi yap" tır. Eğer belirli bir kişinin davranışı tahmin edilmek isteniyorsa, veri uzayında o kişiye yakın, örneğin on kişinin davranışlarına bakılır. Bu on kişinin davranışlarının ortalaması hesaplanır ve bu ortalama belirlenen kişi için tahmin olur. K en yakın komşuluğunda, K harfi

araştırılan komşuların sayısıdır. 5-yakın komşuluğunda, 5 kişiye ve 1-yakın komşuluğunda 1 kişiye bakılır [18]. K en yakın komşuluğu bir öğrenme tekniği değildir. Daha çok bir araştırma yöntemidir. K en yakın komşuluğu, veri kümesini daha iyi anlamaya yardımcı olur.

K en yakın komşuluk yönteminde sınıflandırılmak istenen olay sayısı arttıkça hesaplamalar için gereken sürede hızlı bir şekilde artar, k en yakın komşuluk modelinin işlem hızını artırmak için genellikle bütün veri hafızada tutulur.

K en yakın komşuluğu tekniği ile n tane kayıttan oluşan bir veri kümesinde, her bir kayıt için tahmin yapılmak istendiğinde, her kayıt, diğer kayıtlarla karşılaştırılmak zorundadır. Bu da büyük veri kümelerinde karesel karmaşıklığa yol açar. Eğer, bir milyon kayıtlı veri tabanında basit bir K en yakın komşuluğu incelemesi yapılacaksa, bir milyar karşılaştırma yapılması gerekir. Bu, araştırmada sorunlara neden olur. Genelde veri madenciliği algoritmaları n kayıt sayısı kadar karmaşıklığa sahip olmalıdır. Bu nedenle K en yakın komşuluğu tekniği alt örneklemlerle ya da sınırlı sayıda veri kümesinde kullanılmalıdır. Şekil 2.2.'de K en yakın komşuluğu yapısı genel anlamda gösterilmiştir.

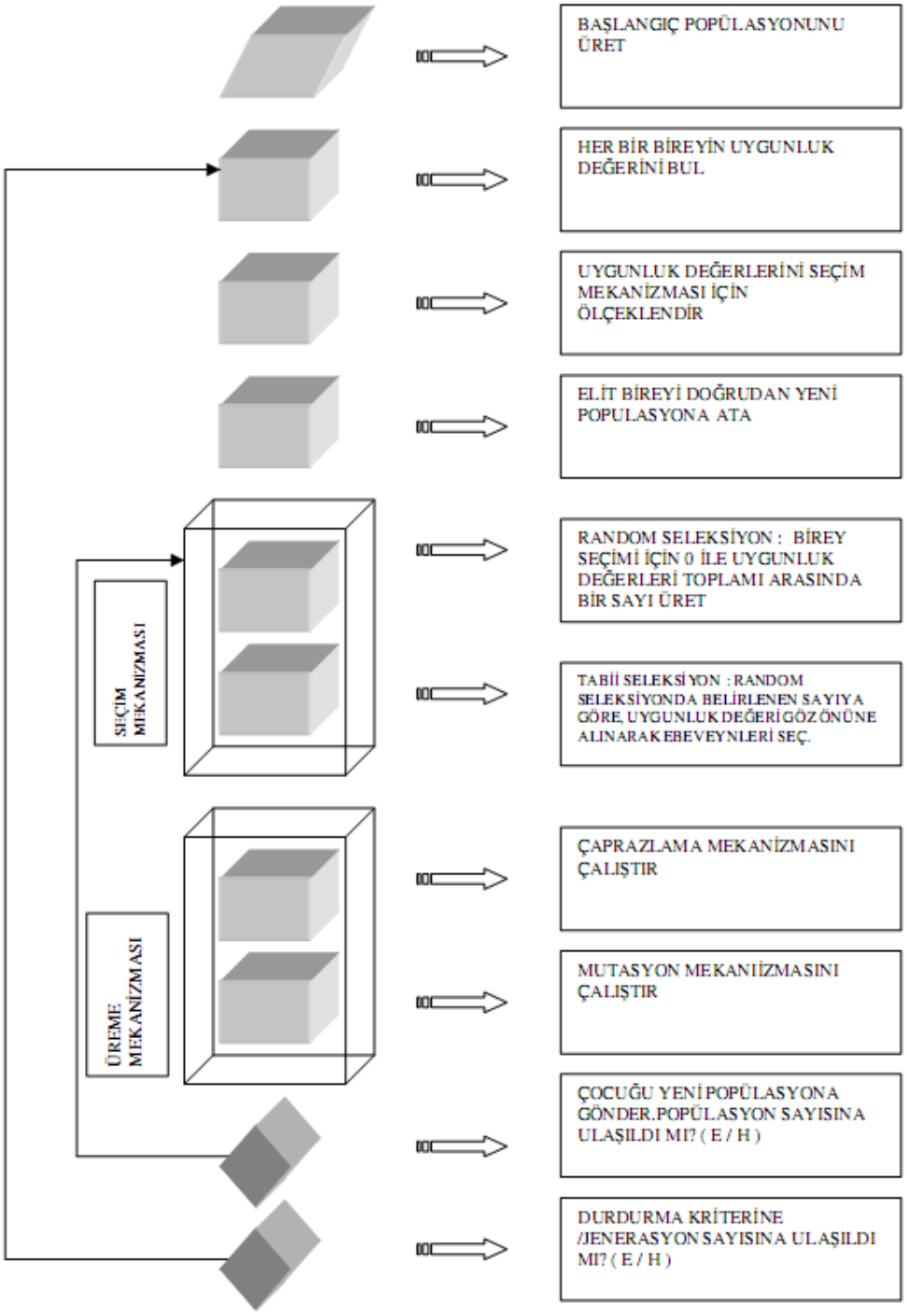


Şekil 2.2. K-En yakın komşu algoritması yapısı

2.5.2.2. Genetik algoritmalar

Genetik algoritma, Darwin tarafından geliştirilen "evrim teorisini"ne dayalıdır. Algoritma ilk olarak popülasyon adı verilen bir çözüm kümesi (öğrenme veri kümesi) ile başlatılır. Bir popülasyondan alınan sonuçlar bir öncekinden daha iyi olacağı beklenen yeni bir popülasyon oluşturmak için kullanılır. Evrim süreci (yeni popülasyonlar yaratma iterasyonu) tamamlandığında bağımlılık kuralları veya sınıf modelleri ortaya konmuş olur [23]. Genellikle genetik öğrenme şöyle olur: Rastgele

oluşturulmuş kuralları içeren ilk popülasyon(sayı kümesi) oluşturulur. Her kural, bir bit dizisi şeklinde gösterilir. Genetik algoritmalar, optimizasyon problemlerinde olduğu gibi sınıflandırma için de kullanılabilir. Basit bir örnekle açıklamak gerekirse; eğitim kümesinde A1 ve A2 boolean (evet veya hayır) niteliklerinin ve C1 –C2 sınıflarının verildiğini varsayalım. Kural "IF A1 AND NOT A2 THEN C2" bit dizisiyle 100 olarak ifade edilir. A1 ve A2 soldaki 2 bitle sınıfta sağdaki bitle gösterilir. Benzer şekilde "IF NOT A1 AND NOT A2 THEN C1" kuralıda 001 şekline kodlanır. Eğer bir nitelik k (k>2) değerlerine sahipse, niteliklerin değerleri k bitleri kullanılarak kodlanabilir[18]. Standart genetik algoritma akış diyagramı Şekil 2.3.'deki gibidir [30].



Şekil 2.3. Genetik algoritmalar akış diyagramı

2.5.2.3. Yapay sinir ağları

İlk kez 1943'te ortaya çıkan yapay sinir ağlarının bilgisayarlarda kullanımı 1980'lerde başlamıştır. Yapay sinir ağları (artificial neural networks), beynin yapısından esinlenilmiş bir bilgi işleme sistemidir. Nöronlara benzeştirilmiş işlem öğeleri arasındaki ilişkilerle yapılandırılmıştır. İnsan beyni gibi yapay sinir ağı da

birbirine bağlı birçok işlem biriminden oluşmuştur. Birçok düğüm (işlem birimi) ve arkla (iç bağlantılar) yönetilen bir grafik olarak yapılandırılır. Bu işlem birimleri birbirlerinden bağımsız işlev görürler ve yalnızca yerel veriyi (düğüme gelen girdi ve düğümden çıkan çıktı) kullanırlar. Bu özellik, yapay sinir ağlarının dağıtık ya da paralel ortamlarda kullanımını kolaylaştırır. Yapay sinir ağları, kaynak (girdi), çıktı ve iç (gizli) düğümlerle yönetilen bir grafik olarak görülebilir. Girdi düğümü girdi katmanında, çıktı düğümü ise çıktı katmanında bulunur. Gizli düğümler, bir ya da daha çok gizli katmanda bulunur. Veri madenciliğinde, çıktı düğümü tahmini belirler. Tek bir girdi düğümünün olduğu (ağacın kökü) karar ağaçlarından farklı olarak yapay sinir ağlarında, her öznitelik değeri için bir girdi düğümü vardır. Yapay sinir ağları karmaşık sorunları çözebilir, ayrıca temel uygulamalardan "öğrenebilir". Yani ağ, soruna kötü bir çözüm bulunduysa, bu soruna bir dahaki sefer daha iyi bir çözüm bulacak biçimde değiştirilir.

Yapay sinir ağları günümüzde bilgi sınıflama ve bilgi yorumlamanın içinde bulunduğu değişik problemlerin çözümünde kullanılmaktadır [29]. Karmaşık ve belirsiz veriden bilgi üretirler. Keşfettikleri örüntü ve eğilimler, insanlar ya da bilgisayarlarca kolay keşfedilemez. Bu tür karmaşık problemlerde birbirleriyle etkileşimli yüzlerce değişken bulunur [31]. Bu teknik, veritabanındaki örüntüleri, sınıflandırma ve tahminde kullanılmak üzere genelleştirir. Yapay sinir ağları algoritmaları sayısal veriler üzerinde çalışırlar [30].

Yapay sinir ağları üç bölümden oluşur [7]:

- Yapay sinir ağının veri yapısını tanımlayan yapay sinir ağı grafiği.
- Öğrenmenin nasıl gerçekleşeceğini belirten öğrenme algoritması.
- Bilginin ağdan nasıl elde edileceğini belirleyen teknikler.

Yapay Sinir Ağları, bağlantı ve öğrenme türlerine göre sınıflandırılabilir.

- İleri beslemeli bağlantıda bağlantılar yalnızca yapıdan daha sonraki katmanlaradır.
- Geri beslemeli bağlantıda ise bazı bağlantılar daha önceki katmanlaradır.

Yapay Sinir Ağları öğrenme türleri ise

- Denetimli (supervised) öğrenme,
- Denetimsiz (unsupervised) öğrenmedir

Denetimli öğrenme, temel olarak iki aşamalı bir işlemdir:

- Yapay sinir ağını, örnek dizileri göstererek verideki farklı sınıfları tanıyacak biçimde eğitmek.
- Önceden görmediği bir veri grubu sağlayarak yapay sinir ağının bu örneklerden ne kadar öğrendiğini denemektir.

Denetimsiz öğrenmede ise sinir ağına, sunulan verinin doğru olarak sınıflandırılmasına ilişkin hiçbir ön bilgi verilmez. Sinir ağı, denetimsiz öğrenmeyi, o veride doğal olarak var olan kümeleri ve altkümeleri bulmak amacıyla çok boyutlu bir veri grubunu çözümlemek için kullanır. Sinir ağları denetimsiz öğrenme tekniği, sağlanan verinin yapısını temel alarak kendi sınıflandırma şemalarını tanımlamak için kullanır [31].

2.5.2.4. Naïve-bayes

Naive Bayes, hedef değişkenle bağımsız değişkenler arasındaki ilişkiyi analiz eden tahminci ve tanımlayıcı bir sınıflama algoritmasıdır [9].

Naive Bayes, sürekli veri ile çalışmaz. Bu nedenle sürekli değerleri içeren bağımlı ya da bağımsız değişkenler kategorik hale getirilmelidir. Örneğin; bağımsız değişkenlerden biri yaş ise, sürekli değerler “<20” “21-30”, “31-40” gibi yaş aralıklarına dönüştürülmelidir.

Naive Bayes, modelin öğrenilmesi esnasında, her çıktının öğrenme kümesinde kaç kere meydana geldiğini hesaplar. Bulunan bu değer, öncelikli olasılık olarak adlandırılır. Örneğin; bir banka kredi kartı başvurularını “iyi” ve “kötü” risk sınıflarında gruplandırmak istemektedir. İyi risk çıktısı toplam 5 vaka içinde 2 kere meydana geldiyse iyi risk için öncelikli olasılık 0,4’tür. Bu durum, “Kredi kartı için başvuran biri hakkında hiçbir şey bilinmiyorsa, bu kişi 0,4 olasılıkla iyi risk grubundadır” olarak yorumlanır. Naive Bayes aynı zamanda her bağımsız değişken /

bağımlı değişken kombinasyonunun meydana gelme sıklığını bulur. Bu sıklıklar öncelikli olasılıklarla birleştirilmek suretiyle tahminde kullanılır [34].

2.5.2.5. Doğrusal regresyon, lojistik regresyon

Regresyon analizi bir bağımlı değişken ile bir veya daha fazla sayıda bağımsız değişken arasındaki ilişkiyi sayısal hale dönüştürmek için kullanılan istatistiksel analiz yöntemidir. Regresyon analizi esas olarak değişkenler arasındaki ilişkinin niteliğini saptamayı amaçlar. Bağımsız değişken olarak bir değişken kullanılırsa basit regresyon, iki veya daha fazla değişken kullanılırsa çoklu regresyon analizi olarak adlandırılır.

Regresyon analizinde amaç her bağımsız değişkenin bağımlı değişkendeki değişmeye katkısının hesaplanması, dolayısıyla tahmin değişkenlerinin değerinden hareketle bağımlı değişkenin değerinin tahmin edilmesidir [35].

Veri madenciliğinde yaygın olarak kullanılan regresyon modellerinden doğrusal regresyonda tahmin edilecek olan hedef değişken sürekli değer alırken; lojistik regresyonda hedef değişken kesikli bir değer almaktadır. Doğrusal regresyonda hedef değişkenin değeri; lojistik regresyonda ise hedef değişkenin alabileceği değerlerden birinin gerçekleşme olasılığı tahmin edilmektedir [36]. Doğrusal regresyon aşağıdaki formülle tanımlanabilir:

$Y_i = b_0 + b_1 X_i + e_i$

b_0: Doğrunun y eksenini kestiği nokta

b_1: Regresyon katsayısı

e_i: Hata değeri

Lojistik regresyonda, veriler düz bir çizgi kullanılarak modellenir. Lojistik regresyon, kestirim (prediction) çeşitlerinden en basit olanıdır. İki değişkenli (bivariate) lojistik regresyon rastgele değerler üretir; Y ve lojistik fonksiyonun diğer değişkeni olan X. Lojistik regresyon aşağıdaki formülle tanımlanabilir:

$Y=\alpha + \beta X$

Bu fonksiyonda Y'nin bir sabit olması varsayılmaktadır α ve β, sırayla Y'nin eğilimli ve durdurulabilir olmasını belirlemektedir. Bu katsayılar, asıl veride hataları en aza indirgeyen ve doğruyu değerlendiren en küçük kareler yöntemiyle çözülebilir. (x1,y1),(x2,y2),…….,(xs,ys) formunda s tane örnek ya da veri verilmiş olsun:

X; x1,x2,…..,xs 'lerin ortalamasıdır. Y; y1,y2……..,ys'lerin ortalamasıdır. A ve β diğer regresyon eşitliklerine göre daha iyi bir yaklaşım sunar [18].

2.5.2.6. Karar ağaçları

Karar ağaçları, veri madenciliğinde, yorumlanmalarının kolay olması, veri tabanı sistemleri ile kolayca bütünleştirilebilmeleri ve güvenilirliklerinin iyi olması nedenleri ile sınıflama modelleri içerisinde en yaygın kullanıma sahip tekniktir.

Karar ağacı, adından da anlaşılacağı gibi bir ağaç görünümünde, tahmin edici bir tekniktir [38]. Karar ağaçları veri oluşturulduktan sonra ağaç kökten yaprağa doğru inilerek kurallar (if-then) yazılabilir. Karar ağaçlarında kök ve her düğüm bir soruyla etiketlenir [21]. Düğümlerden ayrılan dallar ise ilgili sorunun olası yanıtlarını belirtir. Her dal düğümü de söz konusu sorunun çözümüne yönelik bir tahmini temsil eder [12]. Kök düğüm olarak da adlandırılan ilk eleman en yüksek karar düğümüdür, kullanılan algoritmaya bağlı olarak her düğüm iki veya daha fazla dala sahip olur. İki dala sahip olan karar ağaçları ikili ağaç, daha fazla dala sahip olanlar ise çok yollu ağaç olarak adlandırılır. Her dal bir başka karar düğümüyle, ya da ağacın sonuyla yani yaprak düğümle sonlanır. Karar düğümlerinde gerçekleştirilen her bölünmede oluşturulan gruplar arasındaki mesafenin maksimum olması bir başka değişle elde edilen grupların mümkün olduğu kadar saf olması istenir.

Karar ağacı temelli analizlerin yaygın olarak kullanıldığı sahalar [39],

- Belirli bir sınıfın muhtemel üyesi olacak elemanların belirlenmesi (Segmentation),
- Çeşitli vakaların yüksek, orta, düşük risk grupları gibi çeşitli kategorilere ayrılması (Stratification),

- Gelecekteki olayların tahmin edilebilmesi için kurallar oluşturulması,
- Parametrik modellerin kurulmasında kullanılmak üzere çok miktardaki değişken ve veri kümesinden faydalı olacakların seçilmesi,
- Sadece belirli alt gruplara özgü olan ilişkilerin tanımlanması**,**
- Kategorilerin birleştirilmesi ve sürekli değişkenlerin kesikliye dönüştürülmesidir.

Karar ağacı temelli tipik uygulamalar ise,

- Hangi demografik grupların mektupla yapılan pazarlama uygulamalarında yüksek cevaplama oranına sahip olduğunun belirlenmesi (Direct Mail*)*,
- Bireylerin kredi geçmişlerini kullanarak kredi kararlarının verilmesi (Credit Scoring),
- Geçmişte işletmeye en faydalı olan bireylerin özelliklerini kullanarak işe alma süreçlerinin belirlenmesi,
- Tıbbi gözlem verilerinden yararlanarak en etkin kararların verilmesi,
- Hangi değişkenlerin satışları etkilediğinin belirlenmesi,
- Üretim verilerini inceleyerek ürün hatalarına yol açan değişkenlerin belirlenmesidir.

Karar Ağacı oluşturmak için CHAID (Chi-Squared Automatic Interaction Detector), CART (Classification and Regression Trees), QUEST (Quick, Unbiased, Efficient Statistical Tree), ID3, C4.5, C5.0 gibi algoritmalar kullanılır.

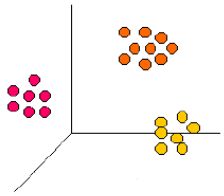
2.5.3. Kümeleme sorgusu

Kümeleme tekniğinde amaç üyelerinin birbirlerine çok benzediği, ancak özellikleri birbirlerinden çok farklı olan kümelerin bulunması ve veri tabanındaki kayıtların bu farklı kümelere bölünmesidir. Kümeleme analizinde; veri tabanındaki kayıtların hangi kümelere ayrılacağı veya kümelemenin hangi değişken özelliklerine göre yapılacağı, konunun uzmanı olan bir kişi tarafından belirtilebileceği gibi veri tabanındaki kayıtların hangi kümelere ayrılacağını geliştirilen yazılımlar da yapabilmektedir. Kümeleme; web madenciliği, istatistik, biyoloji ve makine öğrenmesi gibi pek çok alanda kullanılır. Kümeleme tekniğinde, sınıflama tekniğinde

olan veri sınıfları yoktur. Sınıflama tekniğinde, verilerin sınıfları bilinmekte ve yeni bir veri geldiğinde bu verinin hangi sınıftan olabileceği tahmin edilmektedir [18].

Kümeleme yöntemi, danışmansız sınıflama modeli olarak da bilinir [26]. Kümeleme heterojen veri kümelerini veri karakteristikleri homojen sayılabilecek gruplara bölme bir başka değişle diğerlerinden çok farklı ancak üyeleri çok benzer olan grupları bulma işidir. Kümeleme modelinde; veri tabanındaki kayıtların hangi kümelere ayrılacağı veya kümelemenin hangi değişken özelliklerine göre yapılacağı, konunun uzmanı olan bir kişi tarafından belirlenebilir [21].

Kümeleme algoritması veritabanını alt kümelere ayırır. Her bir kümede yer alan elemanlar dahil oldukları grubu diğer gruplardan ayıran ortak özelliklere sahiptir. Kümeleme modellerinde amaç, Şekil 2.4.'de görüldüğü gibi küme üyelerinin birbirlerine çok benzediği, ancak özellikleri birbirlerinden çok farklı olan kümelerin bulunması ve veri tabanındaki kayıtların bu farklı kümelere bölünmesidir.



Şekil 2.4. Kümeleme sorgusu

Tahmin edici modeller kümeleme modelini, homojen veri grupları oluşturması için veri ön işleme aşaması olarak da kullanmaktadırlar.

2.5.4. Ardışık örüntüler

Ardışık örüntü keşfi, bir zaman aralığında sıklıkla gerçekleşen olay kümelerini bulmayı amaçlar.

- Bir yıl içinde Orhan Pamuk'un "Benim Adım Kırmızı" romanını satın alan insanların %70'i Buket Uzuner' in "Güneş Yiyen Çingene" adlı kitabını satın almıştır.

- X ameliyatı yapıldığında, 15 gün içinde % 45 ihtimalle Y enfeksiyonu oluşacaktır.
- İMKB endeksi düşerken A hisse senedinin değeri % 15'den daha fazla artacak olursa, üç iş günü içerisinde B hisse senedinin değeri % 60 ihtimalle artacaktır,
- Çekiç satın alan bir müşteri, ilk üç ay içerisinde % 15, bu dönemi izleyen üç ay içerisinde % 10 ihtimalle çivi satın alacaktır.

Bu tip örüntüler perakende satış, telekomünikasyon ve tıp alanlarında yararlıdır.

2.5.5. Birliktelik kuralları

Birliktelik kuralları, bir arada olan olayların ya da özelliklerin keşfedilmesi sürecidir. Birliktelik kuralları genellikle "eğer şu olursa daha sonra bu olur" şeklindedir. Genellikle açıklayıcı veri analizinde, ayrık değerleri tespit etmede, veri ön işlemede, eğilim ve ilişkilerin bulunmasında kullanılır [24]. Bir alışveriş sırasında veya birbirini izleyen alışverişlerde müşterinin hangi mal veya hizmetleri satın almaya eğilimli olduğunun belirlenmesi, müşteriye daha fazla ürünün satılmasını sağlama yollarından biridir. Örneğin, düşük yağlı peynir ve yağsız yoğurt alan müşteriler, % 85 ihtimalle diyet süt de satın alırlar. Bununla birlikte bu teknikler, tıp, finans ve farklı olayların birbirleri ile ilişkili olduğunun belirlenmesi sonucunda değerli bilgi kazanımının söz konusu olduğu ortamlarda da önem taşımaktadır.

Bir birliktelik algoritması oluşturmadan önce kurallar belirlenmelidir. Büyük veri tabanında ilişkileri bulacak algoritmalar geliştirmek çok zor değildir. Fakat geliştirilen algoritmalar önemli ilişkileri ortaya çıkaracağı gibi önemsiz birçok ilişkiyi de ortaya çıkarır. Bu yüzden, büyük veri tabanlarında küçük alt kümeler bulunmalıdır.

Büyük veri tabanlarında birliktelik kuralları bulunurken, şu iki işlem basamağı takip edilir [42]:

1- Sık tekrarlanan öğeler bulunur. Bu öğelerin her biri en az, önceden belirlenen minimum destek sayısı kadar sık tekrarlanırlar.

2- Sık tekrarlanan öğelerden güçlü birliktelik kuralları oluşturulur. Bu kurallar minimum destek ve minimum güven değerlerini karşılamalıdır.

Ayrıca, büyük veri tabanlarında çok sayıda ilişki bulunabileceğinden, birliktelik kuralları sayısı da sınırsız olabilir. Dolayısıyla ilginç ilişkilerle önemsiz ilişkilerin ayrılması gerekir [40].

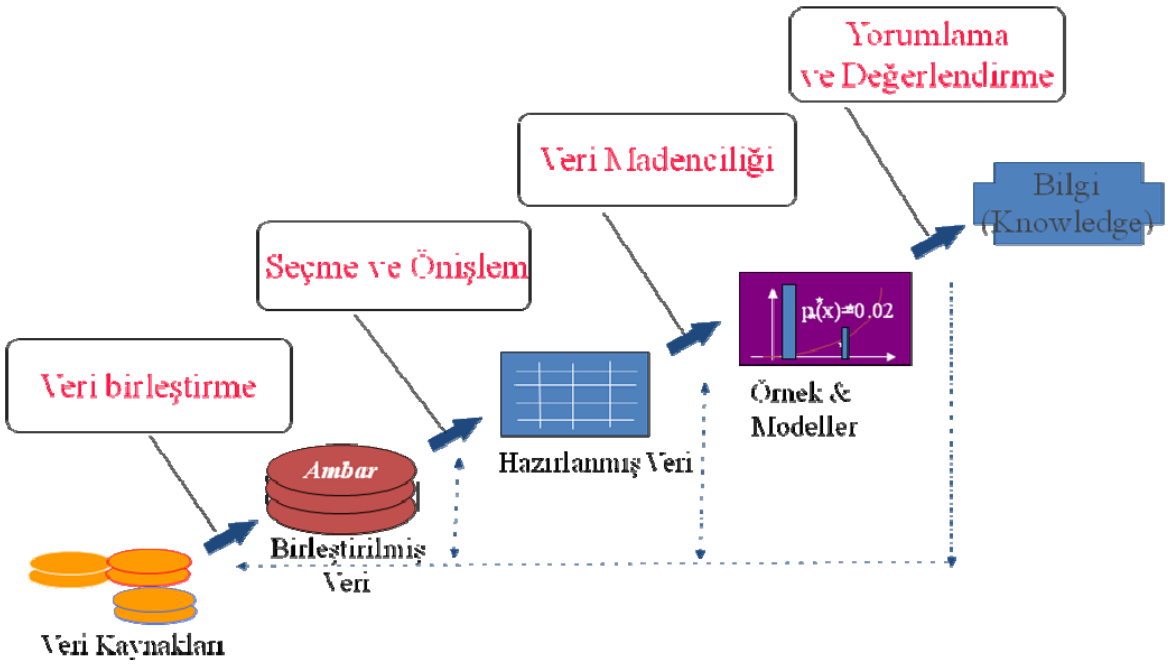
Birliktelik kuralları oluşturmada en çok kullanılan algoritmalar Apriori, GRI, AIS ve SETM'dir.

2.5.5.1. Apriori algoritması

Sık geçen öğe kümelerini bulmak için birçok kez veri tabanını taramak gerekir. İlk taramada bir elemanlı minimum destek metriğini sağlayan sık geçen öğe kümeleri bulunur. İzleyen taramalarda bir önceki taramada bulunan sık geçen öğe kümeleri, aday kümeler adı verilen, yeni potansiyel sık geçen öğe kümelerini üretmek için kullanılır. Aday kümelerin destek değerleri tarama sırasında hesaplanır ve aday kümelerinden minimum destek metriğini sağlayan kümeler o geçişte üretilen sık geçen öğe kümeleri olur. Sık geçen öğe kümeleri bir sonraki geçiş için aday küme olurlar. Bu süreç yeni bir sık geçen öğe kümesi bulunmayana kadar devam eder [41].

BÖLÜM 3. VERİ TABANLARINDA BİLGİ KEŞFİ SÜRECİ

Aktif araştırma alanlarından biri olan veri tabanlarında bilgi keşfi (VTBK), çok büyük oylumlu verileri tam veya yarı otomatik bir biçimde analiz eden yeni kuşak araç ve tekniklerin üretilmesi ile ilgilenen son yılların gözde araştırma konularından biridir [41]. VTBK Şekil 3.1.'de gösterildiği gibi veri birleştirme, veri seçimi ve ön işleme, veri madenciliği ve değerlendirme aşamalarından oluşan bir süreçtir [44]. Veri madenciliği, önceden bilinmeyen, veri içinde gizli, anlamlı ve yararlı örüntülerin büyük ölçekli veri tabanlarından otomatik biçimde elde edilmesini sağlayan VTBK süreci içinde bir adımdır [24].



Şekil 3.1. VTBK süreci

Veri tabanı yönetim sistemleri (VTYS) büyük miktardaki yapısal bilgiyi saklamak ve etkin bir biçimde erişim sağlamakla yükümlüdür. VTYS'lerde veri düzenlemesi, ilgili organizasyonun işletimsel veri ihtiyacı doğrultusunda gerçekleştirilir ki, bu her zaman bilgi keşfi perspektifi ile birebir çakışmaz. Bu açıdan veri tabanındaki veriler

temizleme, boyut indirgeme, transfer, vb. işlemlerden geçirilerek veri madenciliğinin kullanımına sunulur [41].

3.1. Veri Tabanlarında Bilgi Keşfi Aşamaları

Veri madenciliği algoritmalarının üzerinde inceleme yapılan işin ve verilerin özelliklerinin bilinmemesi durumunda fayda sağlaması mümkün değildir. Bu nedenle aşağıda tanımlanan tüm aşamalardan önce, iş ve veri özelliklerinin öğrenilmesi / anlaşılması başarının ilk şartı olacaktır [39]. Bu aşamalar:

1- Problemin Tanımlanması,
2- Verilerin Hazırlanması,
3- Modelin Kurulması ve Değerlendirilmesi,
4- Modelin Kullanılması,
5- Modelin İzlenmesi şeklinde belirtilmiştir.

3.1.1. Problemin tanımlanması

Veri madenciliği çalışmalarında başarılı olmanın ilk şartı, uygulamanın hangi işletme amacı için yapılacağının açık bir şekilde tanımlanmasıdır. İlgili işletme amacı işletme problemi üzerine odaklanmış ve açık bir dille ifade edilmiş olmalı, elde edilecek sonuçların başarı düzeylerinin nasıl ölçüleceği tanımlanmalıdır. Ayrıca yanlış tahminlerde katlanılacak olan maliyetlere ve doğru tahminlerde kazanılacak faydalara ilişkin tahminlere de bu aşamada yer verilmelidir.

3.1.2. Verilerin hazırlanması

Modelin kurulması aşamasında ortaya çıkacak sorunlar, bu aşamaya sık sık geri dönülmesine ve verilerin yeniden düzenlenmesine neden olacaktır. Bu durum verilerin hazırlanması ve modelin kurulması aşamaları için, bir analizcinin veri keşfi sürecinin toplamı içerisinde enerji ve zamanının % 50 - % 85'ini harcamasına neden olmaktadır [39].

Verilerin hazırlanması aşaması kendi içerisinde toplama, değer biçme, birleştirme ve temizleme, seçme ve dönüştürme adımlarından meydana gelmektedir.

3.2.2.1. Toplama (Collection)

Tanımlanan problem için gerekli olduğu düşünülen verilerin ve bu verilerin toplanacağı veri kaynaklarının belirlenmesi adımıdır. Verilerin toplanmasında kuruluşun kendi veri kaynaklarının dışında, nüfus sayımı, hava durumu, merkez bankası kara listesi gibi veri tabanlarından veya veri pazarlayan kuruluşların veri tabanlarından faydalanılabilir.

3.2.2.2. Değer biçme (Assessment)

Veri madenciliğinde kullanılacak verilerin farklı kaynaklardan toplanması, doğal olarak veri uyumsuzluklarına neden olacaktır. Bu uyumsuzlukların başlıcaları farklı zamanlara ait olmaları, kodlama farklılıkları (örneğin bir veri tabanında cinsiyet özelliğinin e/k, diğer bir veri tabanında 0/1 olarak kodlanması), farklı ölçü birimleridir. Ayrıca verilerin nasıl, nerede ve hangi koşullar altında toplandığı da önem taşımaktadır.

Bu nedenlerle, iyi sonuç alınacak modeller ancak iyi verilerin üzerine kurulabileceği için, toplanan verilerin ne ölçüde uyumlu oldukları bu adımda incelenerek değerlendirilmelidir.

3.2.2.3. Birleştirme ve temizleme (Consolidation and Cleaning)

Bu adımda farklı kaynaklardan toplanan verilerde bulunan ve bir önceki adımda belirlenen sorunlar mümkün olduğu ölçüde giderilerek veriler tek bir veri tabanında toplanır. Ancak basit yöntemlerle ve baştan savma olarak yapılacak sorun giderme işlemlerinin, ileriki aşamalarda daha büyük sorunların kaynağı olacağı unutulmamalıdır.

3.2.2.4. Seçim (Selection)

Bu adımda kurulacak modele bağlı olarak veri seçimi yapılır. Örneğin tahmin edici bir model için, bu adım bağımlı ve bağımsız değişkenlerin ve modelin eğitiminde kullanılacak veri kümesinin seçilmesi anlamını taşımaktadır.

Sıra numarası, kimlik numarası gibi anlamlı olmayan ve diğer değişkenlerin modeldeki ağırlığının azalmasına da neden olabilecek değişkenlerin modele girmemesi gerekmektedir. Bazı veri madenciliği algoritmaları konu ile ilgisi olmayan bu tip değişkenleri otomatik olarak elese de, pratikte bu işlemin kullanılan yazılıma bırakılmaması daha akılcı olacaktır.

Genellikle yanlış veri girişinden veya bir kereye özgü bir olayın gerçekleşmesinden kaynaklanan verilerin (Outlier), önemli bir uyarıcı enformasyon içerip içermediği kontrol edildikten sonra veri kümesinden atılması tercih edilir.

Modelde kullanılan veri tabanının çok büyük olması durumunda tesadüfîliği bozmayacak şekilde örnekleme yapılması uygun olabilir. Günümüzde hesaplama olanakları ne kadar gelişmiş olursa olsun, çok büyük veri tabanları üzerinde çok sayıda modelin denenmesi zaman kısıtı nedeni ile mümkün olamamaktadır. Bu nedenle tüm veri tabanını kullanarak bir kaç model denemek yerine, tesadüfî olarak örneklenmiş bir veri tabanı parçası üzerinde birçok modelin denenmesi ve bunlar arasından en güvenilir ve güçlü modelin seçilmesi daha uygun olacaktır.

3.2.2.5. Dönüştürme (Transformation)

Kredi riskinin tahmini için geliştirilen bir modelde, borç/gelir gibi önceden hesaplanmış bir oran yerine, ayrı ayrı borç ve gelir verilerinin kullanılması tercih edilebilir. Ayrıca modelde kullanılan algoritma, verilerin gösteriminde önemli rol oynayacaktır. Örneğin bir uygulamada bir yapay sinir ağı algoritmasının kullanılması durumunda kategorik değişken değerlerinin evet/hayır olması; bir karar ağacı algoritmasının kullanılması durumunda ise örneğin gelir değişken değerlerinin yüksek/orta/düşük olarak gruplanmış olması modelin etkinliğini artıracaktır.

3.1.3. Modelin kurulması ve değerlendirilmesi

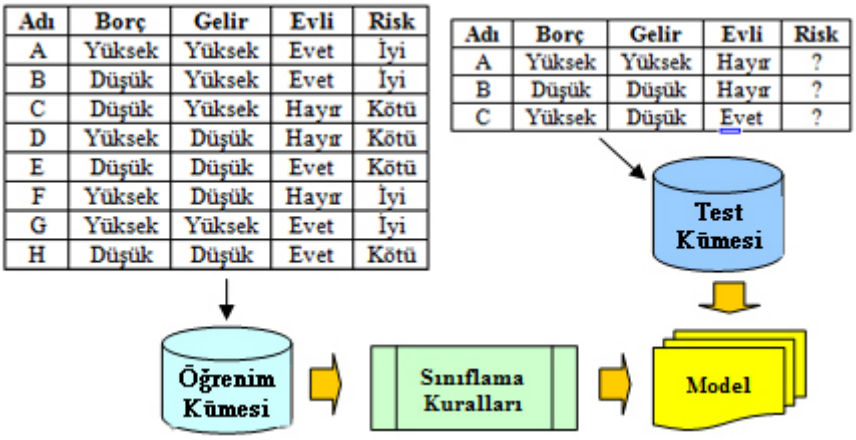
Tanımlanan problem için en uygun modelin bulunabilmesi, olabildiğince çok sayıda modelin kurularak denenmesi ile mümkündür. Bu nedenle veri hazırlama ve model kurma aşamaları, en iyi olduğu düşünülen modele varılıncaya kadar yinelenen bir süreçtir.

Model kuruluş süreci denetimli (Supervised) ve denetimsiz (Unsupervised) öğrenimin kullanıldığı modellere göre farklılık göstermektedir.

Örnekten öğrenme olarak da isimlendirilen denetimli öğrenimde, bir denetçi tarafından ilgili sınıflar önceden belirlenen bir kritere göre ayrılarak, her sınıf için çeşitli örnekler verilir. Sistemin amacı verilen örneklerden hareket ederek her bir sınıfa ilişkin özelliklerin bulunması ve bu özelliklerin kural cümleleri ile ifade edilmesidir.

Öğrenme süreci tamamlandığında, tanımlanan kural cümleleri verilen yeni örneklere uygulanır ve yeni örneklerin hangi sınıfa ait olduğu kurulan model tarafından belirlenir.

Denetimsiz öğrenmede, kümeleme analizinde olduğu gibi ilgili örneklerin gözlenmesi ve bu örneklerin özellikleri arasındaki benzerliklerden hareket ederek sınıfların tanımlanması amaçlanmaktadır.



Şekil 3.2. Denetimli öğrenme

Denetimli öğrenimde seçilen algoritmaya uygun olarak ilgili veriler hazırlandıktan sonra, ilk aşamada verinin bir kısmı modelin öğrenimi, diğer kısmı ise modelin geçerliliğinin test edilmesi için ayrılır. Modelin öğrenimi öğrenim kümesi kullanılarak gerçekleştirildikten sonra, test kümesi ile modelin doğruluk derecesi (Accuracy) belirlenir.

Bir modelin doğruluğunun test edilmesinde kullanılan en basit yöntem basit geçerlilik (Simple Validation) testidir. Bu yöntemde tipik olarak verilerin % 5 ile % 33 arasındaki bir kısmı test verileri olarak ayrılır ve kalan kısım üzerinde modelin öğrenimi gerçekleştirildikten sonra, bu veriler üzerinde test işlemi yapılır. Bir sınıflama modelinde yanlış olarak sınıflanan olay sayısının, tüm olay sayısına bölünmesi ile hata oranı, doğru olarak sınıflanan olay sayısının tüm olay sayısına bölünmesi ile ise doğruluk oranı hesaplanır. (Doğruluk Oranı = 1 - Hata Oranı).

Sınırlı miktarda veriye sahip olunması durumunda, kullanılabilecek diğer bir yöntem çapraz geçerlilik (Cross Validation) testidir. Bu yöntemde veri kümesi tesadüfî olarak iki eşit parçaya ayrılır. İlk aşamada a parçası üzerinde model eğitimi ve b parçası üzerinde test işlemi; ikinci aşamada ise b parçası üzerinde model eğitimi ve a parçası üzerinde test işlemi yapılarak elde edilen hata oranlarının ortalaması kullanılır.

Bir kaç bin veya daha az satırdan meydana gelen küçük veri tabanlarında, verilerin n gruba ayrıldığı n katlı çapraz geçerlilik (N-Fold Cross Validation) testi tercih edilebilir. Verilerin örneğin 10 gruba ayrıldığı bu yöntemde, ilk aşamada birinci grup test, diğer gruplar öğrenim için kullanılır. Bu süreç her defasında bir grubun test, diğer grupların öğrenim amaçlı kullanılması ile sürdürülür. Sonuçta elde edilen on hata oranının ortalaması, kurulan modelin tahmini hata oranı olacaktır.

Bootstrapping küçük veri kümeleri için modelin hata düzeyinin tahmininde kullanılan bir başka tekniktir. Çapraz geçerlilikte olduğu gibi model bütün veri kümesi üzerine kurulur. Daha sonra en az 200, bazen binin üzerinde olmak üzere çok fazla sayıda öğrenim kümesi tekrarlı örneklemelerle veri kümesinden oluşturularak hata oranı hesaplanır.

Model kuruluşu çalışmalarının sonucuna bağlı olarak, aynı teknikle farklı parametrelerin kullanıldığı veya başka algoritma ve araçların denendiği değişik modeller kurulabilir. Model kuruluş çalışmalarına başlamazdan önce, imkânsız olmasa da hangi tekniğin en uygun olduğuna karar verebilmek güçtür. Bu nedenle farklı modeller kurarak, doğruluk derecelerine göre en uygun modeli bulmak üzere sayısız deneme yapılmasında yarar bulunmaktadır.

Özellikle sınıflama problemleri için kurulan modellerin doğruluk derecelerinin değerlendirilmesinde basit ancak faydalı bir araç olan risk matrisi kullanılmaktadır. Tablo 3.1.'de görülen matriste sütunlarda fiili, satırlarda ise tahmini sınıflama değerleri yer almaktadır. Örneğin fiilen B sınıfına ait olması gereken 46 elemanın, kurulan model tarafından 2'sinin A, 38'inin B, 6'sının ise C olarak sınıflandırıldığı matriste kolayca görülebilmektedir.

Tablo 3.1. Fiili ve tahmini sınıflama değerleri

	Fiili		
Tahmini	**A Sınıfı**	**B Sınıfı**	**C Sınıfı**
A Sınıfı	45	2	3
B Sınıfı	10	38	2
C Sınıfı	4	6	40

Önemli diğer bir değerlendirme kriteri modelin anlaşılabilirliğidir. Bazı uygulamalarda doğruluk oranlarındaki küçük artışlar çok önemli olsa da, birçok işletme uygulamasında ilgili kararın niçin verildiğinin yorumlanabilmesi çok daha büyük önem taşıyabilir. Çok ender olarak yorumlanamayacak kadar karmaşıklaşsalar da, genel olarak karar ağacı ve kural temelli sistemler model tahmininin altında yatan nedenleri çok iyi ortaya koyabilmektedir.

Kaldıraç (Lift) oranı ve grafiği, bir modelin sağladığı faydanın değerlendirilmesinde kullanılan önemli bir yardımcıdır. Örneğin kredi kartını muhtemelen iade edecek müşterilerin belirlenmesi amacını taşıyan bir uygulamada, kullanılan modelin belirlediği 100 kişinin 35'i gerçekten bir süre sonra kredi kartını iade ediyorsa ve tesadüfî olarak seçilen 100 müşterinin aynı zaman diliminde sadece 5'i kredi kartını iade ediyorsa kaldıraç oranı 7 olarak bulunacaktır.

Kurulan modelin değerinin belirlenmesinde kullanılan diğer bir ölçü, model tarafından önerilen uygulamadan elde edilecek kazancın bu uygulamanın gerçekleştirilmesi için katlanılacak maliyete bölünmesi ile elde edilecek olan yatırımın geri dönüş (Return On Investment) oranıdır.

Kurulan modelin doğruluk derecesi ne denli yüksek olursa olsun, gerçek dünyayı tam anlamı ile modellediğini garanti edebilmek mümkün değildir. Yapılan testler sonucunda geçerli bir modelin doğru olmamasındaki başlıca nedenler, model kuruluşunda kabul edilen varsayımlar ve modelde kullanılan verilerin doğru olmamasıdır. Örneğin modelin kurulması sırasında varsayılan enflasyon oranının zaman içerisinde değişmesi, bireyin satın alma davranışını belirgin olarak etkileyecektir.

3.1.4. Modelin kullanılması

Kurulan ve geçerliliği kabul edilen model doğrudan bir uygulama olabileceği gibi, bir başka uygulamanın alt parçası olarak kullanılabilir. Kurulan modeller risk analizi, kredi değerlendirme, dolandırıcılık tespiti gibi işletme uygulamalarında doğrudan kullanılabileceği gibi, promosyon planlaması simülasyonuna entegre edilebilir veya tahmin edilen envanter düzeyleri yeniden sipariş noktasının altına düştüğünde, otomatik olarak sipariş verilmesini sağlayacak bir uygulamanın içine gömülebilir.

3.1.5. Modelin izlenmesi

Zaman içerisinde bütün sistemlerin özelliklerinde ve dolayısıyla ürettikleri verilerde ortaya çıkan değişiklikler, kurulan modellerin sürekli olarak izlenmesini ve gerekiyorsa yeniden düzenlenmesini gerektirecektir. Tahmin edilen ve gözlenen değişkenler arasındaki farklılığı gösteren grafikler model sonuçlarının izlenmesinde kullanılan yararlı bir yöntemdir.

3.2. Veri Madenciliğinde Karşılaşılan Problemler

Veri madenciliği sistemlerinin temel unsuru her şeyden önce ham veridir. Ham veriler veri ambarları ya da veri tabanlarından sağlanır. Veritabanlarının dinamik,

eksiksiz, geniş ve net veri içermemesi durumunda sorunlar ortaya çıkar. Bunun yanında verinin konu ile uyumsuzluğu, sınıflandırma gerekliliği gibi durumlar diğer sorunlar arasında sayılabilir [45].

3.2.1. Veri tabanı boyutu

Veritabanlarında tutulan verilerin boyu iki boyutu ifade etmektedir [24].

1- Yatay boyut: VT'lerde tutulan bilgilerin özelliklerini ifade eden satırların sütunsal detaylarıdır.
2- Dikey boyut: VT'lerde tutulan kayıt sayısını ifade etmektedir.

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

Örneklem kümesinin yatay ve dikey boyutta indirgenmesi,

1- Yatay indirgeme: Nitelik değerlerinin önceden belirlenmiş genelleme sıradüzenine göre, bir üst nitelik değeri ile değiştirilme işlemi yapıldıktan sonra aynı olan çokluların çıkarılma işlemidir.
2- Dikey indirgeme: Artık niteliklerin indirgenmesi işlemidir.

3.2.2. Gürültülü veri

Veri girişi veya veri toplanması sırasında oluşan sistem dışı hatalara gürültü adı verilir. Hatalı veri veritabanlarında ciddi problem oluşturabilir. Bu durum, bir veri madenciliği yönteminin, kullanılan veri kümesinde bulunan gürültülü verilere karşı daha az duyarlı olmasını gerektirir. Eğer veri gürültülü ise sistem bozuk veriyi tanımalı ve ihmal etmelidir. Herhangi bir veri toplama tekniğinin, gürültüden tümüyle arınmış olması çok zordur. Bu nedenle, veri madenciliğinde, gelecekte toplanacak verideki gürültü miktarının yaklaşık olarak o anki veridekiyle aynı olmasına dikkat gösterilmelidir [46].

3.2.3. Boş değerler

Bir veri tabanında boş değer, birincil anahtarda yer almayan herhangi bir niteliğin değeri olabilir. Boş değer, tanımı gereği kendisi de dahil olmak üzere hiç bir değere eşit olmayan değerdir. Birçokluda eğer bir nitelik değeri boş ise o nitelik, bilinmeyen ve uygulanamaz bir değere sahiptir. Bu durum ilişkisel veri tabanlarında sıkça karşımıza çıkmaktadır. Bir ilişkide yer alan tüm çoklular aynı sayıda niteliğe, niteliğin değeri boş olsa bile sahip olmalıdır. Örneğin, kişisel bilgisayarların özelliklerini tutan bir ilişkide bazı model bilgisayarlar için ses kartı modeli niteliğinin değeri boş olabilir [41]. Boş değerli nitelikler veri kümesinde bulunuyorsa ya bu çoklular tamamıyla ihmal edilmeli ya da bu çoklularda niteliğe olası en yakın değer atanmalıdır [47].

3.2.4. Eksik veri

İstenen problemin çözümüne ulaşabilmek için gereken örneklem kümesindeki 2 boyutun (yatay ve dikey boyutun) eksik olmaması gerekir. Bu boyuttaki eksiklikler şu şekilde olabilir [48]:

- Yatay boyutta: Yatay boyuttaki eksiklik, örneklem kümesinde olması gereken nitelik veya niteliklerin olmamasıdır. Örneğin: eğer insanların göz rengiyle alakalı bir hastalığın neye bağlı olduğu bulunmaya çalışılıyorsa, niteliklerden göz renginin örneklem kümesinde bulunması gerekir.
- Dikey boyutta: Dikey boyuttaki eksiklik örneklem kümesindeki kayıtların eksik olmasıdır. Örneğin bir süper markette yaşı 10 ve 25 yaşındaki kişiler her yaptıkları alışverişte bir ürünü sürekli alıyorlarsa, bu örüntünün keşfedilmesi için örneklem kümesinde yeterli sayıda 10-25 yaş aralığına giren kayıtların bulunması gerekir. Eğer örneklem kümesinde bu kayıtlar bulunmazsa gerçek hayatta var olan bir örüntü kaçırılmış olur.

3.2.5. Artık veri

Veri kümesi, eldeki probleme uygun olmayan veya artık nitelikler içerebilir. Artık veri, problemde istenilen sonucu elde etmek için kullanılan örneklem kümesindeki

gereksiz niteliklerdir. Artık nitelikleri elemek için geliştirilmiş algoritmalar, özellik seçimi olarak adlandırılır. Özellik seçimi, tümevarıma dayalı öğrenmede bir ön işlem olarak algılanır. 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) ifade edebilmektir. Özellik seçimi arama uzayını küçültür ve sınıflama işleminin kalitesini de artırır [41].

3.2.6. Dinamik veri

On-line içeriği değişen veri tabanlarında karşılaşılan başlıca problemdir. Bir veri tabanındaki içeriğin sürekli değişmesi veri madenciliği uygulamalarının uygulanabilmesini önemli ölçüde zorlaştırıcı sorunlar doğurmaktadır. Bu sorunlardan bazıları şunlardır [48]:

- Ortaya çıkan veri madenciliği örüntülerinin sürekli değişim halinde olan verilerden hangisini ifade ettiğinin tespitinin zorluğu ve bu üretilen sonuçların zaman içinde eski üretilen sonuçlardan farkının tespiti ve gereken yerlerin güncellenme zorluğu,
- Veri madenciliği algoritmalarının çalışabilmesi için verilerin üzerine okuma kilidi konulması gerektiğinde, bu verilerin başka uygulamalar tarafından değişime açık olmaması,
- Veri madenciliği algoritmalarının ve çevrimiçi VT uygulamalarının aynı anda uygulanmasından kaynaklanan ciddi performans düşüşlerinin olması, vb.

3.2.7. Farklı tipteki verileri ele almak

Kullanılan verinin saklandığı ortam düz bir kütük veya ilişkisel veritabanlarında yer alan tablolar olabileceği gibi nesneye yönelik veritabanları, çoklu ortam veritabanları, coğrafik veritabanları vs. olabilir. Bununla birlikte veri çeşitliliğinin fazla olması bir veri madenciliği algoritmasının tüm veri tiplerini ele alabilmesini olanaksızlaştırmaktadır. Bu yüzden veri tipine özgü, veri madenciliği algoritmaları geliştirilmektedir [41].

BÖLÜM 4. WEB MADENCİLİĞİ

4.1. Web Terimleri

Web madenciliği disiplini içerisinde sıkça rastlanan ve W3C konsorsiyumunca tanımlanan önemli terimler aşağıda belirtilmiştir [57,58].

Kaynak (Resource): W3C'nin Değişmez kaynak tanımlayıcısı tarifine göre (Uniform Resource Identifier - URI) özdeşliği olan her şey olabilir.

URI: Kaynağın fiziksel adresini tanımlayan karakterler katarı olarak ifade edilir. Örnek olarak http://www.sakarya.edu.tr/dosyalar/hedef_dosya.doc, adresi verilebilir.

Web Kaynağı (Web Resource): HTTP protokollerinden (örneğin HTTP 1.1 vb.) herhangi bir sürümüne ulaşabilen kaynaktır.

Web Sunucusu (Web Server): Bir veritabanı içeren ve internet üzerinde belgelere erişim hizmetlerini sunan bilgisayardır.

Web Sayfası (Web Page): URI tarafından tanımlanan bir veya birden fazla web kaynağının veri kümesidir.

Web Sitesi (Web Site): Bir web sunucusu tarafından web'te sunulan veritabanları, ilgili belgeler ve dosyalar. Bir web sitesindeki belgeler birbirleriyle ilgili birkaç konuyu kapsayıp, aralarında üst metin linkleri ile bağlantılar kurulur.

Sayfa Görüntüleme (Page View yada Kısaca Hit): Bir web tarayıcısının (web browser) belli bir zamanda bir web sayfasında bulunması.

Web Tarayıcı (Web Browser): Internet üzerinde bilgi kaynaklarını aramaya elverişli, bağlantılı, metin ve ortamların olanaklarını kullanan istemci yazılımı. Başka bir ifade ile istenen URI'yi görüntüleyen yazılım (IE, Mozilla, Opera, vb.).

Kullanıcı (User): Web tarayıcısını kullanan kişi.

Web İsteği (Web Request): İstemcinin bir web kaynağına yapmış olduğu istek. Bunlar açık (kullanıcı taraflı, explicit) ya da dolaylı (web istemci taraflı, implicit) olarak ikiye ayrılır. Açık web istekleri (aynı zamanda tıklama – click olarak da adlandırılır) kendi içersinde iki sınıfa ayrılır; gömülü (embedded) ve kullanıcı girişli şeklinde adlandırılır. Gömülü web isteğine örnek olarak bir web sayfası içerisinde bulunan bağlantılardan yapılan istekler verilebilir. Kullanıcı girişli web istekleri ise kullanıcının web tarayıcısı üzerinden yazarak ya da seçerek yapacağı isteklerdir. Dolaylı web istekleri çağrılmış olan web sayfası içerisinde gömülmüş olan öğelerin (örneğin sayfa içerisindeki resim, betik (script) dosyaları vb.) getirilmesini sağlayan isteklerdir.

Kullanıcı Oturumu (User Session): Bir kullanıcının bir veya daha fazla web sunucusu üzerinde yapmış olduğu sınırlandırılmış sayıda kullanıcı taraflı web istekleridir.

Ziyaret (Visit): Belli bir zaman süresince kullanıcı oturumu esnasında yapılmış olan sayfa görüntülemesi eylemidir.

İçerik (Content): Sitedeki verinin içeriği kullanıcıya iletilen objelerin ve ilişkilerin toplamıdır ve web sayfaları içerisindeki gerçek veridir. Site veri içeriği bunlardan başka tanımlayıcı kelimeler, doküman özellikleri, semantik taglar ya da http değişkenleri gibi semantik ve yapısal meta verileri de içerir. Son olarak, site için tanımlı küme ontolojisi veri içeriğinin bir parçası olarak düşünülür. Tanımlı küme ontolojisi açıkça sitenin içinde yakalanabilir, ya da bazı formlarda bulunabilir. Tanımlı küme ontolojisinin açık gösterimi ürün kategorileri, gösterilen yapısal hiyerarşiler ve dosya yapısı gibi site içeriğinde depolanmış kavramsal hiyerarşileri, semantik içerik ve ilişkilerini rdf (resource description framework) ya da veritabanı şeması ontoloji dili ile açıkça gösterimlerini içerebilir [54].

Yapı (Structure): İçeriğin organizasyonunu gösteren veridir. Yapı verisi sitedeki içerik organizasyonunun tasarımcı bakış açısı ile nasıl göründüğünü gösterir. Bu organizasyon sayfalar arasındaki linkler ile belirlenir. Örneğin, sayfa içerisinde HTML veya XML dokümanları ağaç yapısı gibi gösterilebilir. Site için yapı verisi normalde otomatik olarak oluşturulan site haritasıdır. Site haritalama aracı sayfalar

arası ve sayfa içindeki ilişkileri yakalama ve gösterme yetisine sahip olmalıdır. Aynı fiziksel sayfada gösterilen pageview'lerden oluşan frame tabanlı sitelerde daha da önem kazanır. Dinamik olarak oluşturulan sayfalarda site haritalama aracı uygulama ve scriptlerle etkileşim bilgisine sahip olmalı ya da uygulama veya scripte örnek parametreler geçirerek segment içerik oluşturma kabiliyetine sahip olmalıdır [54].

Kullanım (Usage): Web sayfalarının kullanım bilgilerini gösteren veridir. Bu bilgiler içerisinde IP adresleri, sayfa referansları, bağlantı tarih ve saati verilmektedir. Web ve uygulama sunucularından otomatik olarak toplanan kayıt dosya (log) verileri kullanıcıların yönelim (navigational) davranışlarını gösterirler. Analizin amacına göre bu veri değişik şekillere dönüştürülmeli ya da bir araya getirilmelidir. Web Kullanım Madenciliğinde en temel seviye verinin ayrıştırılması olan sayfa görüntülenmesidir. Fiziksel olarak, sayfa görüntülenmesi kullanıcının web tarayıcısı ile yapmış olduğu istekten kaynaklanan web objelerinin birleşiminin gösterimidir. Bu web objeleri birden fazla sayfa (frame tabanlı site), resimler, gömülen bileşen ya da betik ve veri tabanı sorgularından oluşur. Kavramsal olarak sayfa görüntüleme, kullanıcının sitedeki makaleyi okuması, arama sorgusunun sonuçlarını görüntüleme, ürün sayfasını görme, alışveriş sepetine bir ürün eklenmesi gibi belirli bir tipteki eylemini gösterir. Diğer taraftan, kullanıcı seviyesinde en temel ayırım kullanıcı oturumudur. Oturum (ziyaret olarak da anılır) bir kullanıcının bir ziyareti sırasında belli süre içerisinde art arda görüntülediği sayfa görüntülemesidir [54].

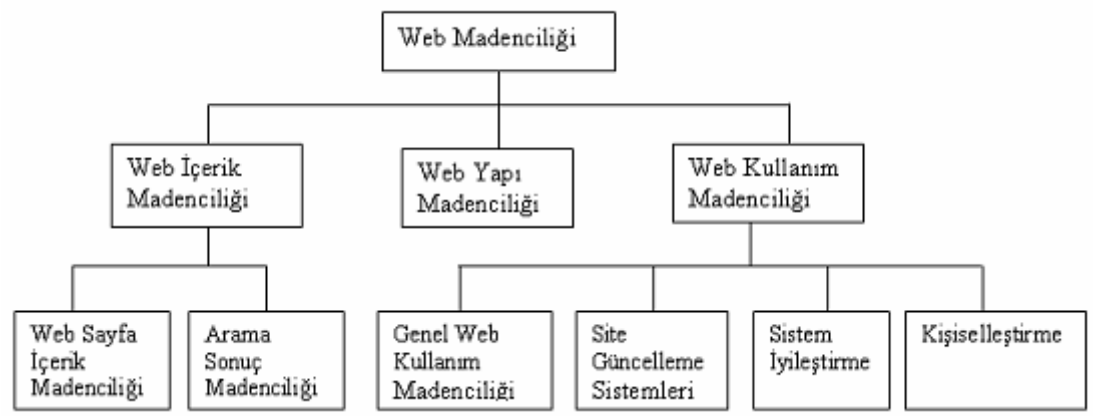
Kullanıcı Profilleri (User Profiles): Web site kullanıcısının demografik bilgisini gösteren veridir. Kayıt olunduğunda alınan bilgiler buna dahildir. Operasyonel veri tabanları ek olarak kullanıcı profil bilgilerini içerebilirler. Bu veri demografik ya da kayıtlı kullanıcıların ayırıcı bilgileri, sayfalar, ürünler, filmler, geçmiş alışverişler gibi çeşitli objelerdeki kullanıcı oranları, kullanıcıların ziyaret geçmişlerinden oluşabilir. Böyle bir verinin elde edilebilmesi için kullanıcının site ile açıkça etkileşime girmesi gerekir. Bu verinin bir kısmı anonim olarak bir kullanıcının tanımlayıcı bilgileri olmadan elde edilebilir. Örneğin istemci tarafındaki çerezlerde (cookies) bulunan anonim bilgiler kullanıcı profil bilgisi olarak düşünülebilir ve siteye tekrar gelen ziyaretçileri ayırt etmek için kullanılabilir [54].

4.2. Web Madenciliği Nedir

Web madenciliği, veri madenciliği teknikleri kullanarak, World Wide Web dokümanları ve servislerinden, otomatik olarak, anlamlı bilgi çıkarmaktır [49]. Web madenciliğinin işi bu bilgilerin farklı veri madenciliği teknikleri kullanılarak site sahibine yararlı bilgiler çıkarmasıdır. Bu sayede ticari amaçlı bir siteden elde edilen kar miktarı arttırılabileceği gibi, internet sayfaları farklı ilgi alanlarına göre düzenlenerek ziyaretçi memnuniyeti arttırılabilir. İşlenecek olan ham veri, ziyaretçilerin sayfaları gezerken bıraktıkları bilgilerin yanı sıra üye olurken verdikleri bilgilerden oluşmaktadır. Bu verilerden sağlanabilecek faydaların bir kaçı şöyle sıralanabilir; kullanıcıların profilleri çıkarılabilir ve zaman içindeki değişimleri takip edilebilir, sitedeki beğenilen ya da beğenilmeyen köşeler tespit edilebilir, kullanıcıların gezinti şekli/hızı sitenin içerik, yapılandırma ve altyapısı açısından performansı hakkında fikir verebilir.

Web sitelerinin bulunduğu sunucular üzerindeki erişim ve hata kayıt dosyalarında kullanıcının site içinde gezinirken yaptığı her bir tıka karşılık bir ya da birden çok hareket kaydı birikir. Kullanıcı adeta gezindiği her noktada parmak izlerini bırakmaktadır. Bu hızla büyüyen dosyalar yer kazanmak için periyodik olarak temizlenmektedir. Oysa bu veriler, site içerik verisi ve kayıtlı kullanıcılara ait veri ile birleştirildiğinde fayda sağlanabilecek bir veritabanı oluşturmak mümkün olabilecektir [50].

Web Madenciliği ortaya atıldığı ilk zamanlarda iki kategoriye ayrılmaktaydı. Web İçerik Madenciliği (Web Content Mining) ve Web Kullanım Madenciliği (Web Usage Mining). Web Madenciliğinin yaygınlaşması ile beraber Web Yapı Madenciliği de (Web Structure Mining) üçüncü bir kategori olarak literatüre eklenmiştir [51].



Şekil 4.1. Web madenciliğinin sınıflandırılması

4.2.1. Web içerik madenciliği

Web içerik madenciliği web kaynaklarından otomatik bilgi arama tekniklerini tanımlar[50]. Web kaynakları içerisinde metin, resim, ses, görüntü, metadata ve hiper linkler bulunmaktadır. Web içerik madenciliğin amacı, bu kaynaklar arasından bilginin bulunması veya filtrelenmesidir. Verinin farklı tiplerde oluşu ve yapısal olmayışı bu konudaki tekniklere daha karışık yaklaşımlar kazandırır. Web içerik madenciliği, text madenciliği ve veri madenciliği ile ilgili olmasına rağmen aralarında bir takım farklılıklar vardır. Web içerik madenciliği, veri madenciliği ile ilgilidir çünkü web dokümanları içerisindeki verileri çıkarmak için veri madenciliği tekniklerini kullanır. Veri madenciliğinde, tam olarak yapısal veriler kullanılırken; web verileri kısmı yapılı ve yapısız verilerdir. Aynı şekilde, web içerik madenciliği text madenciliğiyle ilgilidir çünkü web üzerindeki bilgilerin çoğu text tabanlıdır. Web içerik madenciliği ile text madenciliği arasındaki fark ise text madenciliğinin tamamen yapısal olmayan veriler üzerinde odaklanmış olmasıdır [52].

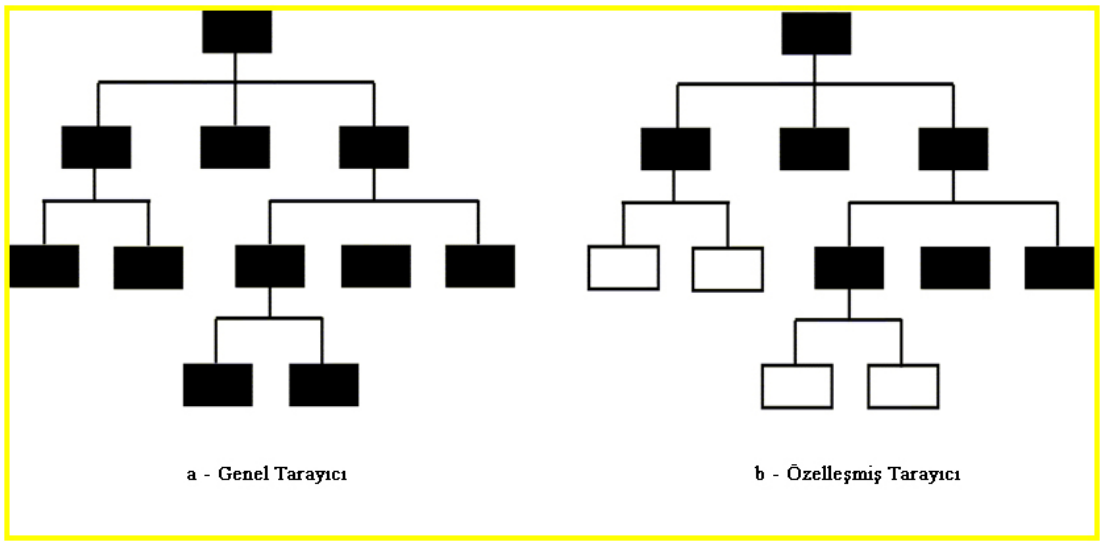
Web içerik madenciliğinde kullanılan iki yaklaşık vardır [51]:

– Information Retrieval Approach (IR) (Bilgiye Erişim Yaklaşımı): IR kısaca belirli yapısı olmayan ya da yarı-yapısal verilerin araştırılma yönteminin adıdır. Bunlara herhangi bir sitedeki metni örnek olarak verebiliriz (bir haber sitesindeki haber metni). Bunlar neredeyse tamamen yapısal olmayan belirsiz kaynaklardır ya da metin dosyalarıdır.

Burada veriyi yorumlarken yaygın kullanılan tekniklerden faydalanılır: Sınıflama, Kümeleme ve Birliktelik vb.. Yaygın bir kullanım alanı vardır. Özellikle arama motorları bu tekniğe başvururlar, ayrıca web belgelerinin sınıflandırılmasında, farklı sunuculardaki aynı içeriğe sahip web sayfalarının bulunmasında ve web belgelerinin çeşitli konularda temsil edilmesinde kullanılırlar.

– Database Approach (Veritabanı Yaklaşımı): Genelde içerik madenciliğinde veri olan sayfaların içerikleri düzgün bir yapıda bulunmamaktadır. Kaynağın DB yapısında olması işleri çok rahatlatır. Web'deki veriyi modellemek ve veriyi bütünleştirerek daha karmaşık bir yapıya sokmak için kullanılan yöntemdir. Bu yöntem sayesinde keyword temelli arama yerine daha gelişmiş sorgular çalıştırmak mümkün olur.

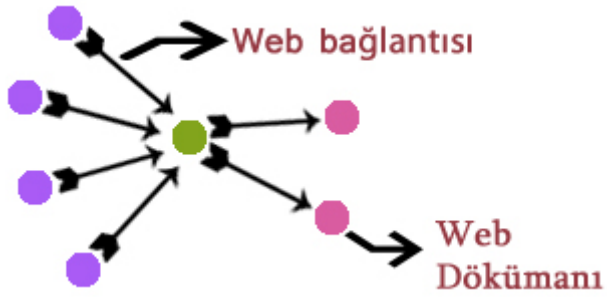
Şekil 4.2.'de genel tarayıcı (Crawling) ve özelleşmiş tarayıcı arama mantığı görülmektedir. Şekildeki siyah gölgeli kısımlar tarayıcının değerlendirmeye aldığı sayfaları temsil etmektedir. Buna göre özelleşmiş tarayıcı bir sayfayı ilgili bulduysa sayfanın linklerini değerlendirmeye almakta, aksi halde diğer sayfaları değerlendirmeye geçmekte bir alt seviyeye inmemektedir [59].



Şekil 4.2. Tarayıcı çeşitleri

4.2.2. Web yapı madenciliği

Web yapı madenciliği sitenin yapısal tasarımını iyileştirmek için kullanılır. Konusu siteler arası bağlantılardır[60]. Web sayfaları arasındaki bağlantılarının (hyperlink) ilişkilerini keşfetmekle ilgilenir. Yani HTML kodlarındaki <a href> etiketleri arasında yer alan veriyi yorumlar. İçerik madenciliği dokümanın içeriğine, yapı madenciliği ise dokümanlar arası bağlantılara yoğunlaşır. Şekil 4.3. 'de web sayfaları arası bağlantı görülmektedir. Web dokümanları arasındaki oklar iki sayfa arasındaki ilişkiyi temsil etmektedir.

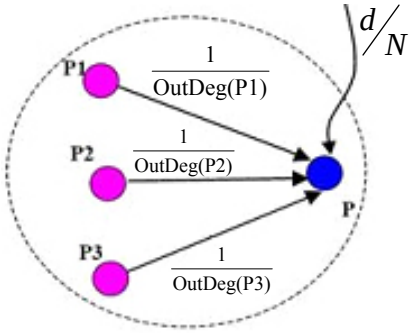


Şekil 4.3. Web sayfaları arasındaki link bağlantısı

Web dokümanları arasındaki linkler bir araya getirildiğinde "Web Graph Structure" elde edilir. Bu yapı sayesinde iki nokta arasındaki en kısa yola ulaşabiliriz. Bu bilgi web sayfaları arasındaki ilişkiyi belirlemek açısından son derece önemlidir. İki sayfa arasında doğrudan bir link yoksa o link arasındaki bağlantıya ve komşuluk ilişkisine kolay bir şekilde erişebiliriz.

Sonuç olarak; web yapı madenciliği sayesinde, araştırılan konu ile ilgili bir sayfayı sisteme vererek onunla ilgili tüm sayfalara erişebilir, web sayfaları arasındaki benzerlik ilişkilerini çıkarabiliriz.

Google'ı dünyanın en önemli arama motoru yapan özelliği "Hyperlink Analyse" yöntemini başarılıyla uygulamasıdır. Google'ın PageRank teknolojisi Şekil 4.4.'de ifade edildiği gibi link yapılarını kullanarak her bir sayfa için bir derece hesaplar. Bu sayede Google istenen konu ile ilgili bir sayfayı getirirken, bu sayfa ile ilgili diğer sayfaları da getirir [51].



$$PR(P) = d/N + (1-d)\left(\frac{PR(P1)}{OutDeg(P1)} + \frac{PR(P2)}{OutDeg(P2)} + \frac{PR(P3)}{OutDeg(P3)}\right)$$

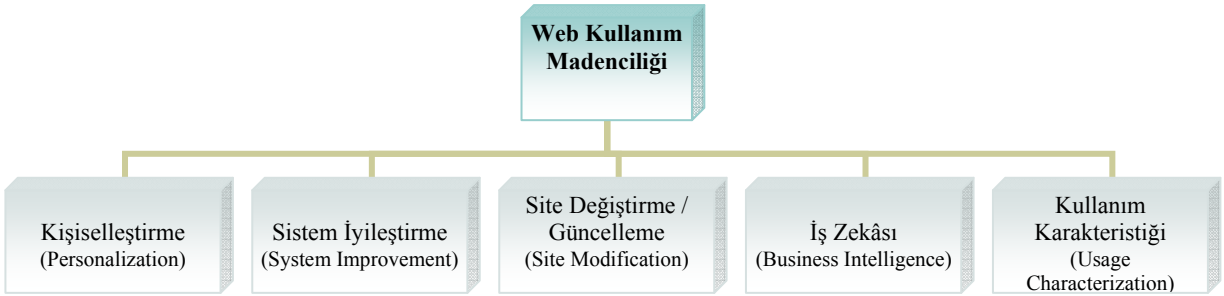
Şekil 4.4. Page rank örneği

4.2.3. Web kullanım madenciliği

Web kullanım madenciliği, kullanıcıların web'de dolaşırken yaptıkları erişim hareketlerince oluşturulan veriden (günlük dosyaları) bilgi üretmeyi hedefler. Web kullanım madenciliği, kullanıcıların genel davranış biçimlerini bilinen ya da önerilen veri madenciliği algoritmalarını, günlük dosyalarındaki veriye uygulayarak bulmaya çalışır [50].

Veri kaynakları olarak günlük dosyalarının yanı sıra, yönlendirme kayıtları, uzak bilgisayar kayıtları, istemci tarafında bulunan çerezler (cookies), kullanıcı profilleri, veri nesneleri (sayfa özellikleri, içerik özellikleri, kullanılan veri) sayılabilir [56].

Web kullanım madenciliği uygulama alanları Şekil 4.5.'de belirtildiği gibi Kişiselleştirme, Sistem İyileştirme, Site Değiştirme / Güncelleme, İş Zekâsı ve Kullanım Karakteristiği başlıkları altında toplanmıştır [61].



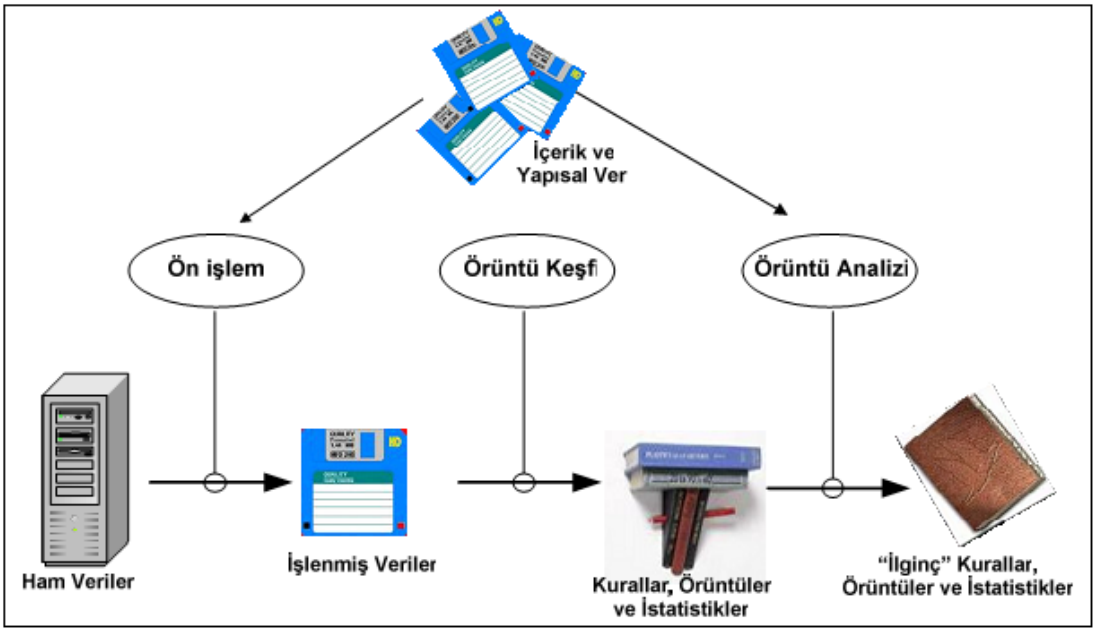
Şekil 4.5. Web kullanım madenciliği uygulama alanları

Günlük dosyalarının ve kullanıcı kaydı verilerinin analizi, aynı zamanda kurumun daha etkili bir sunumunun yapılabilmesi için web sitesini nasıl daha iyi hale getirebileceği hakkında değerli bilgiler sağlar. Bunun yanında, müşterilerin ilgi alanları, ürünler üzerinden pazar stratejileri oluşturma, promosyon kampanyalarının etkisi gibi hususlarda, kurumlara karar süreçlerinde yardımcı olur. Örnek bir günlük dosyası Tablo 4.1.'de gösterilmiştir [61].

Tablo 4.1. Günlük dosyası kayıt örneği.

#	IP Adress	Time	Method/URL/Protocol	Status	Size	Referrer	Agent
1	123.456.78.9	[25/Apr/1998:03:04:41]	GET A.html http/1.0	200	3290	-	Mozilla/3.04(Win95, I)
2	123.456.78.9	[25/Apr/1998:03:05:34]	GET B.html http/1.0	200	2050	A.html	Mozilla/3.04(Win95, I)
3	123.456.78.9	[25/Apr/1998:03:05:39]	GET L.html http/1.0	200	4130	-	Mozilla/3.04(Win95, I)
4	123.456.78.9	[25/Apr/1998:03:06:02]	GET F.html http/1.0	200	5096	B.html	Mozilla/3.04(Win95, I)
5	123.456.78.9	[25/Apr/1998:03:06:58]	GET A.html http/1.0	200	3290	-	Mozilla/3.01(X11,I,IRIX6.2,IP22)
6	123.456.78.9	[25/Apr/1998:03:07:42]	GET B.html http/1.0	200	2050	A.html	Mozilla/3.01(X11,I,IRIX6.2,IP22)
7	123.456.78.9	[25/Apr/1998:03:07:55]	GET R.html http/1.0	200	8140	L.html	Mozilla/3.04(Win95, I)
8	123.456.78.9	[25/Apr/1998:03:09:50]	GET C.html http/1.0	200	1820	A.html	Mozilla/3.01(X11,I,IRIX6.2,IP22)
9	123.456.78.9	[25/Apr/1998:03:10:02]	GET O.html http/1.0	200	2270	F.html	Mozilla/3.04(Win95, I)
10	123.456.78.9	[25/Apr/1998:03:10:45]	GET J.html http/1.0	200	9430	C.html	Mozilla/3.01(X11,I,IRIX6.2,IP22)
11	123.456.78.9	[25/Apr/1998:03:12:23]	GET G.html http/1.0	200	7220	B.html	Mozilla/3.04(Win95, I)
12	123.456.78.9	[25/Apr/1998:05:05:22]	GET A.html http/1.0	200	3290	-	Mozilla/3.04(Win95, I)
13	123.456.78.9	[25/Apr/1998:05:06:03]	GET D.html http/1.0	200	1680	A.html	Mozilla/3.04(Win95, I)

Web kullanım madenciliği Şekil 4.6.'da belirtildiği gibi, Önişlem (Preprocessing), Örüntü keşfi (Pattern Discovery) ve Örüntü analizi (Pattern Analysis) aşamalarından oluşur [61,63].

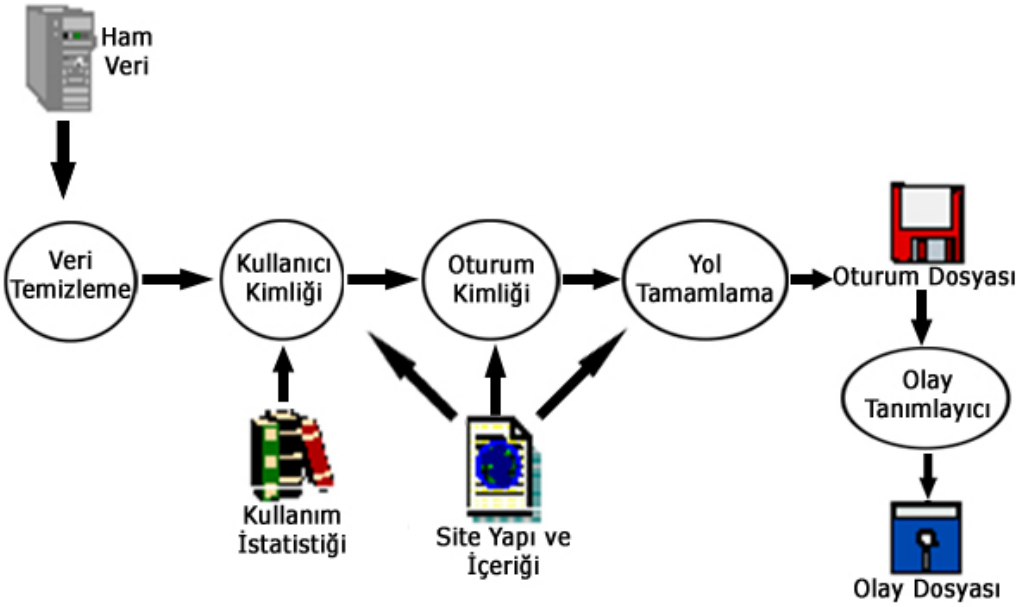


Şekil 4.6. Web kullanım madenciliği süreci

4.2.3.1. Web kullanım madenciliği aşamaları

1- Ön İşlem (Preprocessing): Ön işlem web kullanım madenciliğinin ilk aşamasıdır. Ham veri bir takım işlemlerden geçirilerek soyutlaştırılır ve örüntü keşfi (*Pattern Discovery*) için hazır hale getirilir. Soyutlaştırma bir çeşit istatistiksel özet çıkarmadır ve kullanıcı (users), sayfa görünümü (pageviews), tıklama akışı (click stream), kullanıcı oturumu (sessions), sunucu oturumu gibi çeşitleri olabilmektedir [51]. Web kullanım madenciliğinin en önemli aşamasıdır çünkü etkili bir şekilde yapıldığında zaman ve kaynak tasarrufu sağlayacaktır. Bu adımda esas olarak veri gürültüden temizlenir [60].

Ön işlem aşamasında ham verilerdeki problemleri gidermek için uygulanan aşamalar aşağıdaki gibidir [54,62].



Şekil 4.7. Ön işlem akış şeması

Veri Temizleme(Data Cleaning): Kayıt dosyası üzerinde web kullanım madenciliği açısından birçok gereksiz satır olacaktır. Bu işlem vasıtası ile kayıt dosyası üzerinde bulunan resim, çoklu ortam ve betik dosyaları silinecektir. Ayrıca bu esnada robot (web robot, spider veya bot) adını vermiş olduğumuz web üzerinden otomatik tarama yapan yazılımların bırakmış olduğu satırları da web kayıt dosyasından çıkarmamız gerekmektedir. Örnek bir web log üzerinde veri temizle işlemi [52];

Tablo 4.2. Web loglarının ilk 20 satırı

	IP Address	Date / Time	HTTP Request	Status Code	Transfer Volume
1	141.243.1.172	[29:23:53:25]	GET /Software.html HTTP/1.0	200	1497
2	query2.lycos.cs.cmu.edu	[29:23:53:36]	GET /Consumer.html HTTP/1.0	200	1325
3	tanuki.twics.com	[29:23:53:53]	GET /News.html HTTP/1.0	200	1014
4	wpbfl2-45.gate.net	[29:23:54:15]	GET / HTTP/1.0	200	4889
5	wpbfl2-45.gate.net	[29:23:54:16]	GET /icons/circle_logo_small.gif HTTP/1.0	200	2624
6	wpbfl2-45.gate.net	[29:23:54:18]	GET /logos/small_gopher.gif HTTP/1.0	200	935
7	140.112.68.165	[29:23:54:19]	GET /logos/us-flag.gif HTTP/1.0	200	2788
8	wpbfl2-45.gate.net	[29:23:54:19]	GET /logos/small_ftp.gif HTTP/1.0	200	124
9	wpbfl2-45.gate.net	[29:23:54:19]	GET /icons/book.gif HTTP/1.0	200	156
10	wpbfl2-45.gate.net	[29:23:54:19]	GET /logos/us-flag.gif HTTP/1.0	200	2788
11	tanuki.twics.com	[29:23:54:19]	GET /docs/OSWRCRA/general/hotline HT...	302	-
12	wpbfl2-45.gate.net	[29:23:54:20]	GET /icons/ok2-0.gif HTTP/1.0	200	231
13	tanuki.twics.com	[29:23:54:25]	GET /OSWRCRA/general/hotline/ HTTP/1.0	200	991
14	tanuki.twics.com	[29:23:54:37]	GET /docs/OSWRCRA/general/hotline/95...	302	-
15	wpbfl2-45.gate.net	[29:23:54:37]	GET /docs/browner/adminbio.html HTTP/...	200	4217
16	tanuki.twics.com	[29:23:54:40]	GET /OSWRCRA/general/hotline/95report...	200	1250
17	wpbfl2-45.gate.net	[29:23:55:01]	GET /docs/browner/cbpress.gif HTTP/1.0	200	51661
18	dd15-032.compuserve....	[29:23:55:21]	GET /Access/chapter1/s2-4.html HTTP/1.0	200	4602
19	tanuki.twics.com	[29:23:55:23]	GET /docs/OSWRCRA/general/hotline/95...	200	56431
20	wpbfl2-45.gate.net	[29:23:55:29]	GET /docs/Access HTTP/1.0	302	-

Web log kayıtlarındaki ilk satır ele alınacak olursa, 141.243.1.172 ip numarası ile sistemden istekte bulunan kullanıcı ayın 29. günü saat 23:53:25'te Web sunucudan

GET metodu ile "/Software.html" dosyasına HTTP/1.0 http versiyonu ile 200 kod numaralı istekte bulunmuş ve 1497 byte veri transfer etmiştir. Web log kayıtlarında bir sonraki kayıt 11 saniye sonra oluşmuştur. Bu isteğin ardından bir başka kullanıcı da 17 saniye sonra kayıt dosyasına işlenmiştir. Web log kayıtlarının yukarıda ifade edildiği gibi veri temizleme işlemine tabi tutulması gerekmektedir. Veri temizleme adımları şu şekilde ifade edilebilir.

Adım 1: Verileri çıkarmak

1- Date / Time alanından date verisinin belirlenmesi
2- Date / Time alanından time verisinin belirlenmesi
3- HTTP request alanından istek tipinin belirlenmesi (POST, GET vb.)
4- HTTP request alanından istemde bulunulan URL nin belirlenmesi
5- HTTP request alanından HTTP sürümünün belirlenmesi

Adım 2: Zaman damgasının oluşturulması

1- Web loglarının tam olarak hangi tarihler arasında tutulduğu belirlenir
2- Yıl, ay, gün, saat, dakika, saniye verileri belirlenir
3- Zaman damgası üretilir. Web log kayıtlarının ilk satırı için örnek PHP komut yapısı $zaman_damgasi = mktime(23, 53, 25, 29, 10, 2006)[64]. Bu uygulamada zaman damgası başlangıç değeri 1 Ocak 1995 olarak işleme alınmıştır.

Bu işlemlerin ardından elde edilen web log kayıtları Tablo 4.3.'de gösterilmiştir.

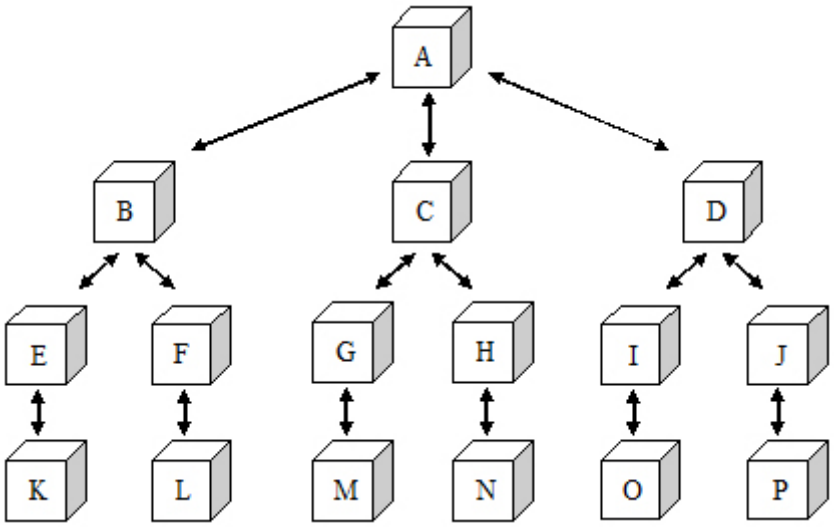
Tablo 4.3. Veri temizle işlemi sonrası web log kayıtları

	IP Address	Date	Time	Method	Page	HTTP_version	TimeStamp
1	141.243.1.172	29	23:53:25	GET	/Software.html	HTTP/1.0	20822005
2	query2.lycos.cs.cmu.edu	29	23:53:36	GET	/Consumer.html	HTTP/1.0	20822016
3	tanuki.twics.com	29	23:53:53	GET	/News.html	HTTP/1.0	20822033
4	wpbfl2-45.gate.net	29	23:54:15	GET	/	HTTP/1.0	20822055
5	wpbfl2-45.gate.net	29	23:54:16	GET	/icons/circle_logo_small.gif	HTTP/1.0	20822056
6	wpbfl2-45.gate.net	29	23:54:18	GET	/logos/small_gopher.gif	HTTP/1.0	20822058
7	140.112.68.165	29	23:54:19	GET	/logos/us-flag.gif	HTTP/1.0	20822059
8	wpbfl2-45.gate.net	29	23:54:19	GET	/logos/small_ftp.gif	HTTP/1.0	20822059
9	wpbfl2-45.gate.net	29	23:54:19	GET	/icons/book.gif	HTTP/1.0	20822059
10	wpbfl2-45.gate.net	29	23:54:19	GET	/logos/us-flag.gif	HTTP/1.0	20822059
11	tanuki.twics.com	29	23:54:19	GET	/docs/OSWRCRA/general/...	HTTP/1.0	20822059
12	wpbfl2-45.gate.net	29	23:54:20	GET	/icons/ok2-0.gif	HTTP/1.0	20822060
13	tanuki.twics.com	29	23:54:25	GET	/OSWRCRA/general/hotline/	HTTP/1.0	20822065
14	tanuki.twics.com	29	23:54:37	GET	/docs/OSWRCRA/general/...	HTTP/1.0	20822077
15	wpbfl2-45.gate.net	29	23:54:37	GET	/docs/browner/adminbio.ht...	HTTP/1.0	20822077
16	tanuki.twics.com	29	23:54:40	GET	/OSWRCRA/general/hotlin...	HTTP/1.0	20822080
17	wpbfl2-45.gate.net	29	23:55:01	GET	/docs/browner/cbpress.gif	HTTP/1.0	20822101
18	dd15-032.compuserve....	29	23:55:21	GET	/Access/chapter1/s2-4.html	HTTP/1.0	20822121
19	tanuki.twics.com	29	23:55:23	GET	/docs/OSWRCRA/general/...	HTTP/1.0	20822123
20	wpbfl2-45.gate.net	29	23:55:29	GET	/docs/Access	HTTP/1.0	20822129

Kullanıcı Tanımı (Kimliği) (User Identification): Burada amaç kayıt dosyalarında oluşan satırların hangi kullanıcılar tarafından oluşturulduğunun belirlenmesidir. Eğer kullanıcı web sitesine kullanıcı adı ve parola ile erişmiş ise bu bilgilerin belirlenmesi gayet kolay olacaktır. Bu yapının olmadığı durumlarda birçok kişi internet çıkışını tek bir internet adresi (IP Adress) üzerinden yaptığı için kullanıcı tanımlama işleminde farklı yöntemler kullanılmalıdır. Bunlar arsında çerezler, oturum kimliği gömme, agent bilgisi, referrer sayılabilir. Bu kişilerin web kayıt dosyası üzerindeki adımlarından kullanıcı kimliklerini tanımlayan örnek:

Tablo 4.4. Kullanıcı tanımı için örnek web log dosyası

IP Adres	Zaman	Metod	Referrer	Agent
87.65.43.21	00:00:02	"GET A.html HTTP/1.1"	-	Mozilla/4.0(Windows NT 5.1, MSIE6.0)
87.65.43.22	00:00:05	"GET B.html HTTP/1.1"	A.html	Mozilla/4.0(Windows NT 5.1, MSIE6.0)
87.65.43.23	00:00:06	"GET A.html HTTP/1.1"	-	Mozilla/5.0(Linux 1.0, Firefox/0.9.3)
87.65.43.24	00:00:10	"GET E.html HTTP/1.1"	B.html	Mozilla/4.0(Windows NT 5.1, MSIE6.0)
87.65.43.25	00:00:17	"GET K.html HTTP/1.1"	E.html	Mozilla/4.0(Windows NT 5.1, MSIE6.0)
87.65.43.26	00:00:20	"GET C.html HTTP/1.1"	A.html	Mozilla/5.0(Linux 1.0, Firefox/0.9.3)
87.65.43.27	00:00:27	"GET L.html HTTP/1.1"	-	Mozilla/4.0(Windows NT 5.1, MSIE6.0)
87.65.43.28	00:00:36	"GET G.html HTTP/1.1"	C.html	Mozilla/5.0(Linux 1.0, Firefox/0.9.3)
87.65.43.29	00:00:49	"GET O.html HTTP/1.1"	I.html	Mozilla/4.0(Windows NT 5.1, MSIE6.0)
87.65.43.30	00:00:57	"GET M.html HTTP/1.1"	G.html	Mozilla/5.0(Linux 1.0, Firefox/0.9.3)
87.65.43.31	00:03:15	"GET H.html HTTP/1.1"	-	Mozilla/5.0(Linux 1.0, Firefox/0.9.3)
87.65.43.32	00:03:20	"GET N.html HTTP/1.1"	H.html	Mozilla/5.0(Linux 1.0, Firefox/0.9.3)
87.65.43.33	00:31:27	"GET E.html HTTP/1.1"	K.html	Mozilla/4.0(Windows NT 5.1, MSIE6.0)
87.65.43.34	00:31:34	"GET L.html HTTP/1.1"	E.html	Mozilla/4.0(Windows NT 5.1, MSIE6.0)



Şekil 4.8. Web log kayıtlarının tutulduğu örnek site ağacı.

Tablo 4.4.'de belirtilen web log kayıtları incelendiğinde aynı ip adresinden giriş yapan tek bir kullanıcıdan oluşmuş gibi görünen kayıtlar agent bilgisinde yer alan tanımlardan ilk aşamada ayrıştırılabilir.

Birinci kullanıcı için Windows 5.1 işletim sistemi MSIE6.0 tarayıcı

İkinci kullanıcı için Linux 1.0 işletim sistemi Firefox 0.9.3 tarayıcı

Yukarıdaki tanımlar incelenerek kullanıcı hareketi,

- Kullanıcı 1: A → B → E → K → I → O → E → L
- Kullanıcı 2: A → C → G → M → H → N

Ancak oluşturulan kullanıcı tanımlarının gelinen sayfaları (referrer) incelendiğinde sayfa geçişlerinin farklı kullanıcılar tarafından yapıldığı görünmektedir. A → B → E → K geçişlerinden gelinen sayfalar (referrer) doğru bir şekilde oluşmuştur. Yani her geçiş bir önceki sayfayı gelinen sayfa (referrer) olarak göstermektedir. Ancak Şekil 4.8.'deki site ağacı ve I.html sayfasına yapılan isteğin farklı bir kullanıcıdan oluşturulduğu anlaşılmaktadır. I.html'i yeni kullanıcının ilk isteği olarak kabul edersek kayıt dosyasında bir sonraki adımının O.html olduğu belirlenir. Bu durumlar birlikte, kullanıcı tanımları

- Kullanıcı 1: A → B → E → K → E → L

- Kullanıcı 2: A → C → G → M → H → N
- Kullanıcı 3: I → O şeklinde oluşur.

Kullanıcı Tanımı adımları

1- Web log kayıtları IP Adresi ve Time'e göre sıralanır
2- Her bir IP adresi için, farklı istemci yapısı (agent) belirlenir. Farklı istemci yapıları için geçici kullanıcı tanımları oluşturulur.
3- Adım 2 de belirlenen kullanıcılar için, site ağacı ve gelinen sayfalar (referrer) uyumluluğuna göre yeni kullanıcılar belirlenir.
4- Adım 2 ve 3 her bir IP Adresi için tekrarlanır.

Oturum Tanımı (Kimliği) (Session Identification): Kullanıcının web üzerinde yapmış olduğu sayfa görüntülemeleri oturumlara bölünmelidir (sessionize). Bu konuda birçok çalışma 30 dakikayı temel almıştır [65]. Temel alınan zaman aşımı eşik değerine göre sayfa geçişlerinin bu aralığa uygunluğu belirlenir.

Örnek olarak tekrar Tablo 4.4. dikkate alınırsa, Kullanıcı 1'in zaman aşımı eşik değeri olan 30 dakika'ya uygun olarak son istek yaptığı sayfanın K.html (geçen süre 00:00:17) olduğu belirlenir. Kullanıcı 1'in K.html sayfasından sonra gelen isteği E.html(geçen süre 00:31:27)'dir. Zaman aşımı eşik değerinden (30 dakika) daha fazla bir beklemenin olduğu tespit edilen bu isteğin yeni bir oturumda yapıldığı anlaşılmaktadır. Bu durama göre örnek sitenin oturum tanımı:

- Oturum 1 (Kullanıcı 1): A → B → E → K
- Oturum 2 (Kullanıcı 2): A → C → G → M → H → N
- Oturum 3 (Kullanıcı 3): I → O
- Oturum 4 (Kullanıcı 1): E → L şeklinde oluşur.

Oturum Tanımı adımları

1- Kullanıcı tanımı adımında belirlenen her bir farklı kullanıcı için yeni bir oturum numarası belirlenir,
2- Zaman aşımı eşik değeri belirlenir (örneğin t = 30 dakika),
3- Her bir kullanıcı için

a. O kullanıcıya ait ardıl iki log satırı için zaman aralığı bulunur.
b. Hesaplanan zaman aralığı t değerinden büyük ise bu log satırı yeni oturum numarası ile ifade edilir.

4- Adım 3 her kullanıcının son sayfasına kadar tekrarlanır.

Tablo 4.5.'te yer alan 184 ve 185. log kayıtları incelendiğinde zaman damgası (timeStamp) farkları 20878108 – 20872908 = 5200 saniye = 86.67 dakika olduğu bulunur. Bu fark zaman aşımı eşik değeri olan 30 dakikadan daha büyük bir değer olduğu için bu aşamadan sonraki adımlar yeni bir oturum numarası ile ifade edilir.

Tablo 4.5. Oturum tanımı için web log kayıtları

	IP Address	method	TimeStamp	Page	Session ID
173	128.165.180.61	GET	20872485	/	Session_20
174	128.165.180.61	GET	20872511	/docs/WhatsNew.html	Session_20
175	128.165.180.61	GET	20872554	/Offices.html	Session_20
176	128.165.180.61	GET	20872737	/cgi-bin/imagemap/eparegio...	Session_20
177	128.165.180.61	GET	20872738	/docs/eparegions/region4.ht...	Session_20
178	128.165.180.61	GET	20872777	/Standards.html	Session_20
179	128.165.180.61	GET	20872787	/docs/OPPTS_Harmonized	Session_20
180	128.165.180.61	GET	20872789	/OPPTS_Harmonized/	Session_20
181	128.165.180.61	GET	20872840	/Consumer.html	Session_20
182	128.165.180.61	GET	20872866	/cgi-bin/imagemap/eparegio...	Session_20
183	128.165.180.61	GET	20872867	/docs/eparegions/region6.ht...	Session_20
184	128.165.180.61	GET	20872908	/docs/Environment.html	Session_20
185	128.165.180.61	GET	20878108	/	Session_21
186	128.165.180.61	GET	20878205	/	Session_21
187	128.165.180.61	GET	20878228	/Info.html	Session_21
188	128.165.180.61	GET	20878236	/docs/Procurement.html	Session_21
189	128.165.180.61	GET	20878266	/docs/conlist	Session_21
190	128.165.180.61	GET	20878267	/conlist/	Session_21
191	128.165.180.61	GET	20878287	/docs/conlist/conlist.html	Session_21
192	128.165.180.61	GET	20878792	/docs/OPP_TECHNICAL_S...	Session_21

Yol Tamamlama (Path Completion): Web kayıtlarında bazı bağlantıların kayıt altına alınmamış olduğu görülmektedir. Büyük çoğunlukla bunun sebebi web tarayıcısının önbelleği (cache) olabilir yada kullanıcın internet bağlantısının vekil sunucu (Proxy server) tarafından dağıtılıyor olması sonucunda bazı sayfaların önbelleğe alınmış olmasıdır. Örneğin, birçok insan önceden incelediği bir sayfaya geri dönmek için tarayıcılarının "Geri" düğmesini kullanır. Bu durumda tarayıcı yerel ön bellekte (local cache) saklanan sayfaya geri döner. Bu işlem web log kayıtlarında yol tanımlamalarının eksik oluşmasına neden olur. Bu durumu tespit ederek yol tanımlamalarını eksiksiz yapabilmek için Şekil 4.8'deki gibi site ağaçlarından yararlanılır.

2 numaralı oturum dikkate alınırsa,

Oturum 2 (Kullanıcı 2): A → C → G → M → H → N

Şekil 4.8.'de gösterilen site ağacına göre M.html ile H.html sayfaları arasında direkt bağlantının olmadığı görülmektedir. Bu durum kullanıcının M.html sayfasında iken önce G.html sayfasına oradan da C.html sayfasına geri döndüğü anlaşılır. Tarayıcının yerel bellekten aldığı bu sayfalar da oturum adımlarına eklenirse 2 numaralı oturum için yeni tanım Oturum 2 (Kullanıcı 2): A → C → G → M → G → C → H → N şeklinde oluşur. Yeni duruma göre örnek sitenin oturum tanımı,

- Oturum 1 (Kullanıcı 1): A → B → E → K
- Oturum 2 (Kullanıcı 2): A → C → G → M → G → C → H → N
- Oturum 3 (Kullanıcı 3): I → O
- Oturum 4 (Kullanıcı 1): E → L şeklinde belirlenir.

2- Örüntü keşfi (Pattern Discovery): Ön işlemden geçirilen verilere veri madenciliği tekniklerinin uygulandığı aşamadır. En sık kullanılan veri madenciliği yöntemleri; istatistiksel, birliktelik kuralları (Affinıty Analysis), kümeleme (Clustering) ve sınıflandırma (Classification) sayılabilir. Bu alanda kullanılan yöntemler Bölüm 2'de yer alan "2.5. Veri Madenciliği Teknikleri" başlığı altında anlatılmıştır.

3- Örüntü analizi (Pattern Analysis): Örüntü analizi web kullanım madenciliğinin son adımıdır. Örüntü analizinin amacı bulunan örüntülerden ilginç olmayan örüntüleri elemektir. Örüntü analizinin en çok karşılaşılan şekli SQL gibi bilgi sorgulama dilleri ile yapılan uygulamalardır. Bir başka yöntem ise verilerin veri küplerine yüklenerek OLAP işlemlerinin yapılmasıdır.

BÖLÜM 5. UYGULAMA

Uygulama Sakarya Üniversitesi CAWIS platformu üzerinde geliştirilmiştir. CAWIS, Sakarya Üniversitesi Bilgi İşlem Dairesi Başkanlığınca geliştirilen Kampus Otomasyonu Web Bilgi Sistemi (Campus Automation Web Information System) projesinin kısa adıdır. CAWIS projesinde ulaşılmak istenen hedefler, kullanıcı doğrulama işlemlerini tek ve güvenli bir noktadan yapmak, kullanıcı altyapı sistemini oluşturmak ve web üzerinden yönetmek, geniş içerikli web servisleri ile ihtiyaç duyulan hizmetleri en iyi şekilde sunmaktır.

CAWIS sistemi, Sakarya Üniversitesinde e-posta hesabı bulunan tüm kullanıcıları kapsayacak bir veritabanı içerir. Bu veritabanı zamanla başından -beri hedeflendiği gibi- kullanıcı bilgilerini otomasyon programlarından alan tam entegre bir yapıya bürünecektir. 2001 yılında temelleri atılan 2004 yılında hizmete açılmasına rağmen geliştirilme ve üzerine servisler eklenme süreci halen devam CAWIS, Sakarya Üniversitesinde geliştirilmiş en büyük yazılım tabanlı otomasyon sistemidir [66].

Bu sistemin web kullanım madenciliği açısında sağladığı en büyük avantajlardan birisi web kullanım madenciliğinin birinci aşamasını olan ve en önemli aşama olarak belirtilen Ön İşlem adımındaki problemleri minimum seviye indirebilecek yapıda olmasıdır. Sistem üzerine entegre edilen veri toplayıcı modül ile platformun sunduğu kullanıcı ve oturum bilgileri üzerinden veri ambarları oluşturulmuştur.

5.1. Uygulama Hedefleri

Geliştirilen bu uygulama ile Sakarya Üniversitesi web sayfasının Tablo 5.1.'de belirtilen analizleri yapılacak ve bu verilerden anlamlı bağlantılar kurulacaktır.

Tablo 5.1. Uygulama hedefleri

1- İZLENME ANALİZİ
Kullanıcı Sayıları
Oturum Sayıları
Sayfa Gösterimleri
Kullanıcı Başına Ortalama Sayfa Gösterimi
Kullanıcı Frekansı (Bir kullanıcının ortalama kaç oturum açtığı)
Kullanıcıların Ortalama Sayfada Kalış Süresi
Ortalama Bir Oturum Süresi
2- TEKNİK ANALİZ
İşletim Sistemleri
Browser
Proxy Bilgisi
Dil ve Bölgesel Ayarlar
3- ARAMA MOTORU ANALİZİ
Anahtar Kelime
Arama motorları
Arama Motoru Bazında Kelimeler
Kelime Bazında Arama Motorları
Arama Motorundan Giriş Yapılan Sayfalar
4- STRATEJİK ANALİZ
Gelinen Domainler
Siteye Nasıl Giriş Yapıldı
Siteye Giriş Noktaları
Ülke/Bölge
Siteden Çıkış noktaları
Ziyaret Edilen Servis
Ziyaret Edilen Sayfa
En Çok Sayfa Görüntüleyen Kullanıcılar
Sayfa Gösterimi Kullanıcı ilişkisi
Saldırı yapan kullanıcı analizi

5.2. Kullanılan Araçlar

Uygulama geliştirme adımlarında kullanılan araçlar;

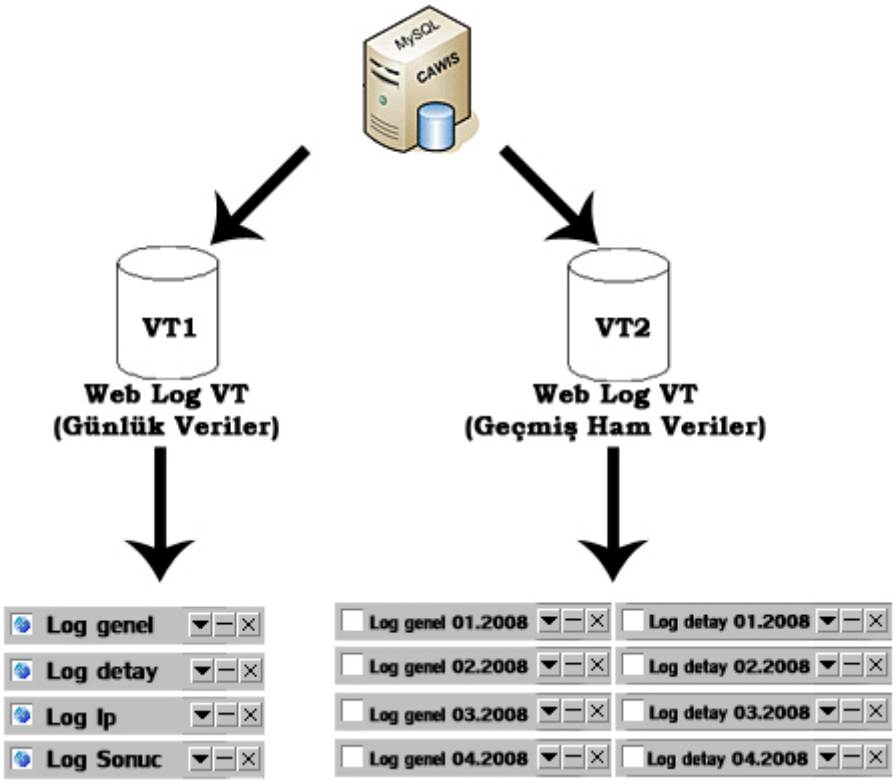
1- PHP("PHP: Hypertext Preprocessor" için kısaltılmış haldir): HTML içine gömülebilir, sunucu taraflı, açık kod lisansıyla dağıtılan bir programlama dilidir[67]. Ön İşlem ve analiz adımında kullanılmıştır.

2- MySQL: Özellikle Linux-Apache-PHP üçlüsü ile birlikte kullanılmak üzere tasarlanmış, basit ama güçlü yapısıyla özellikle web işlemlerinde yaygın olarak kullanılan bir Veritabanı Yönetim Sistemi'dir. GPL lisansı ile ücretsiz

olarak dağıtımı yapılan MySQL sunucusunun Windows ve diğer işletim sistemleri üzerinde çalışan sürümleri de bulunmaktadır. Standart SQL komutlarına direk olarak destek veren MySQL ile diğer tüm VTYS ürünlerinde olduğu gibi, yeni veritabanları ve tablolar oluşturabilir, bu tablolara yeni kayıtlar ekleyebilir, eklediğiniz kayıtları düzenleyebilir ya da silebilirsiniz. Yine diğer VTYS ürünleri üzerindeki verilerinizi MySQL'e, MySQL'den de diğerlerine rahatça aktarabilirsiniz. Uygulamada işlenecek verilerin saklandığı araçtır.

5.3. Veritabanı Mimarisi

Verilerin toplanması ve analiz aşamasına sunulması için hazırlanan MySql veri tabanı mimarisi iki ayrı veri tabanından oluşmaktadır. Birincisi günlük web loglarının tutulduğu ve bu logların analiz sonuçlarının saklandığı, ikincisi de gün sonunda o gün tutulan logların taşındığı veri tabanıdır. Sistemin iki veri tabanı üzerine tasarlanma gerekçesi, analiz aşamasında çok fazla veri karşısında sistem kaynaklarının yetersiz kalabilmesi ve yedeklemenin sadece bir veri tabanı üzerinde yapılacak olmasıdır. Web logları gün sonunda ön işlemden geçirilir ve anlamlı veriler bir tabloda saklanırken ham veri diğer veri tabanına taşınır. Veri tabanı yedeklemesi de günlük olarak sadece birinci veri tabanı için uygulanır. İkinci veri tabanı ayda bir kere yedeklenerek disk alanından tasarruf sağlanmış olur.



Şekil 5.1. Uygulamanın veritabanı mimarisi

Veritabanlarından VT1 de yer alan tablolar ve amaçları;

Log genel Tablosu: Sistemde açılan her oturumun bir satırda ifade edildiği ve o oturumu tanımlayan temel bilgilerin temizlenerek saklandığı tablodur. Tabloda tutulan veriler özetle; Oturum No, Kullanıcı No, internet adresi (IP Adress), varsa vekil sunucu adresi (proxy), işletim sistemi ve sürümü, tarayıcı adı, sürümü, dili ve tipi, ülkesi, siteye giriş yapılan ilk url, yönlendiren site (referrer), arama motoru anahtar kelimeleri ve işlem zamanıdır.

Log detay Tablosu: Açılan oturumun sistem üzerindeki her adımının bir satırla tutulduğu tablodur. Tabloda tutulan veriler özetle; Oturum No (“Log genel” tablosu ile “Log detay” tablosunun bağlantı alanıdır), istemde bulunulan url, CAWIS servisi, servis sayfası, sayfa işleme süresi, cookies, session, post, get gibi globaller ve işlem zamanıdır.

Log Ip Tablosu: Kullanıcının ip adresinin, hangi ülkeye ait olduğunun bulunmasında kullanılan verileri içeren tanım tablosudur. Tabloda tanımlanan bilgiler özetle;

start_number: IP adresinin başlangıç değerinin ondalık karşılığı,

end_number: IP adresinin bitiş değerinin ondalık karşılığı,

ulke: start_number – end_number aralığına denk gelen IP adresinin ait olduğu ülkenin ismidir.

Log Sonuc Tablosu: Web loglarının gün sonunda analiz edildikten sonra elde edilen sonuçların tutulduğu tablodur. Tabloda her satır, bir günlük analiz verisini içerir. Tabloda tanımlanan bilgiler;

Gun: Satırdaki verinin hangi güne ait olduğu,

Page View User, Page View Guest: Kayıtlı ve misafir kullanıcıların sayfa görüntüleme sayıları,

Session User, Session Guest: Kayıtlı ve misafir kullanıcıların oturum sayıları,

Count User, Count Guest: Kayıtlı ve misafir kullanıcı sayıları,

Average Time per Page User, Guest: Kayıtlı ve misafir kullanıcıların bir sayfada ortalama ne kadar kaldığı,

Average Time per Session User, Guest: Kayıtlı ve misafir kullanıcıların sitede ortalama ne kadar kaldığı,

Servis: CAWIS platformunda bulunan servislerin kullanım oranları,

Sayfa: Sayfaların görüntülenme değerleri,

IP: Sisteme giriş yapılan IP adresleri ve dağılımları,

Proxy: Sisteme giriş yapılırken kullanılan vekil sunucular,

User: Sistemde oturum başına sayfa görüntüleme değerine göre sıralanmış kullanıcılar,

Tehlikeli User: Sistemden en kısa zaman diliminde en çok istemde bulunan kullanıcılar,

OS: Uzak bilgisayarların işletim sistemi dağılımı,

Browser: Uzak bilgisayarların tarayıcı dağılımı,

Referrer: Sitenin yönlendirme aldığı web siteleri,

İlk Sayfa: Kullanıcıların ilk kez görüntüledikleri sayfalar,

Login Sayfa: Kullanıcıların hangi sayfada iken kayıtlı kullanıcı olarak sisteme giriş yaptıklarının dağılımı,

Son Sayfa: Oturum sonunda görüntülenen sayfaların dağılımı,

Arama Motoru: Siteye giriş yapılan arama motorlarının dağılımı,

Anahtar Kelime: Arama motorlarında kullanılan ifadelerin dağılımı,

Sayfa Görüntüleme Saatleri: Saatlere göre sayfa görüntüleme dağılımı,

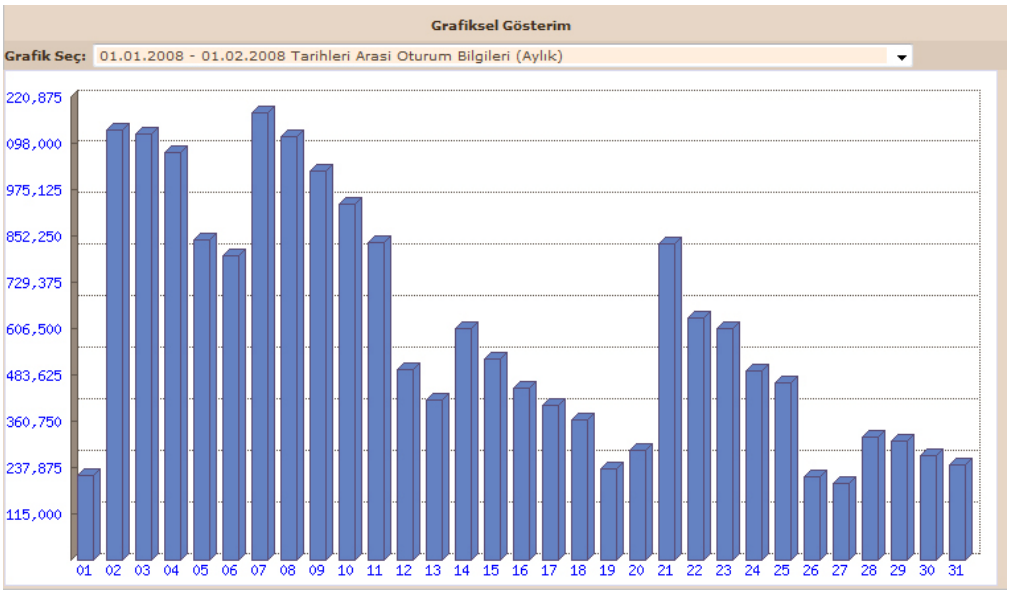
En iyi session saat: En fazla hangi saatte oturum açıldığını ifade eder,

5.4. Uygulama Arayüzü

01.02.2008 Tarihli Genel Log Bilgileri		Tarih Seç : 01.02.2008 Git	
İşlem Zamanı	: 02.02.2008 00:06:15	İşlem Süresi	: 381,935 (sn)
Sayfa Göst. Sayısı [Kullanıcı]	: 72.331	Sayfa Göst. Sayısı [Guest]	: 79.648
Oturum Sayısı [Kullanıcı]	: 10.678	Oturum Sayısı [Guest]	: 15.361
Giren Kullanıcı Sayısı	: 7.048	Giren Guest Sayısı	: 15.357
En Çok Kullanılan Server	: Srv#3	En Çok Kullanılan Servis	:
En Çok Girilen Sayfa	: page-rss	En Çok Giriş Yapan Ip	: 10.9.5.160
En Çok Giriş Yapan Proxy	: 193.243.207.122	En Çok Giriş Yapan Kullanıcı	: Guest_8375_4407
En Az Giriş Yapan Kullanıcı	: Guest_8375_4407	En Çok Kullanılan İşl. Sist.	: win-5.1
En Çok Kullanılan Browser	: msie-6.0	En Çok Yönlendiren	: www.google.com.tr
En Çok Kullanılan Arama	: saü	En Çok Sayfa Göster. Saat	: 13
En Çok Oturum Açılan Saat	: 14	En Çok Login Olunan Saat	: 14
En Çok Kullanılan Giriş Adresi	: http://www.sakarya.edu.tr/		
En Çok Login Olunan Adres	: http://www.obis.sakarya.edu.tr/		
En Çok Kullanılan Çıkış Adresi	: http://www.sakarya.edu.tr/		
En Çok Kull. Arama Motoru	: www.google.com.tr		
Ortalama Sayfa Gösterim Süresi(sn) [Kullanıcı / Guest / Ortalama] : 23,683 / 14,149 / 20,026			
Ortalama Oturumda Kalma Süresi(sn) [Kullanıcı / Guest / Ortalama] : 207,767 / 53,703 / 116,881			

Şekil 5.2. Uygulama arayüzü genel bilgiler

Şekil 5.2.‘de geliştirilen uygulamanın özet sunum ekranını görünmektedir. Bu ekran aracılığıyla seçilen tarihteki bilgilerin özet bir dökümü sunulmaktadır.



Şekil 5.3. Uygulama arayüzü grafik ekranı

Şekil 5.3.’de uygulama sonucu elde edilen çıkarımların grafiksel gösterimlerinin elde edildiği arayüz görünmektedir.

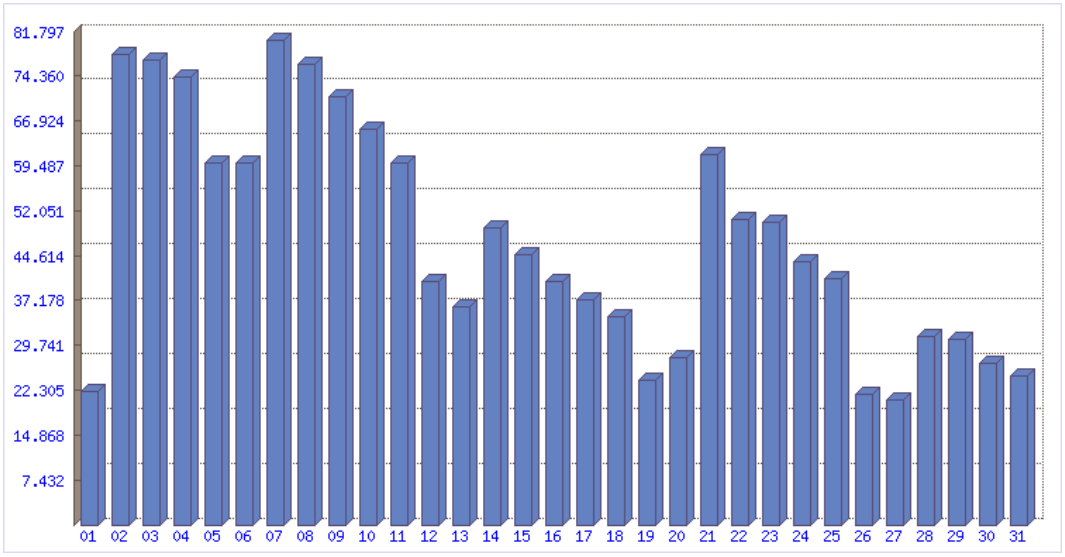
5.5. Uygulama Sonucu Elde Edilen Çıkarımlar

5.5.1. İzlenme analizi

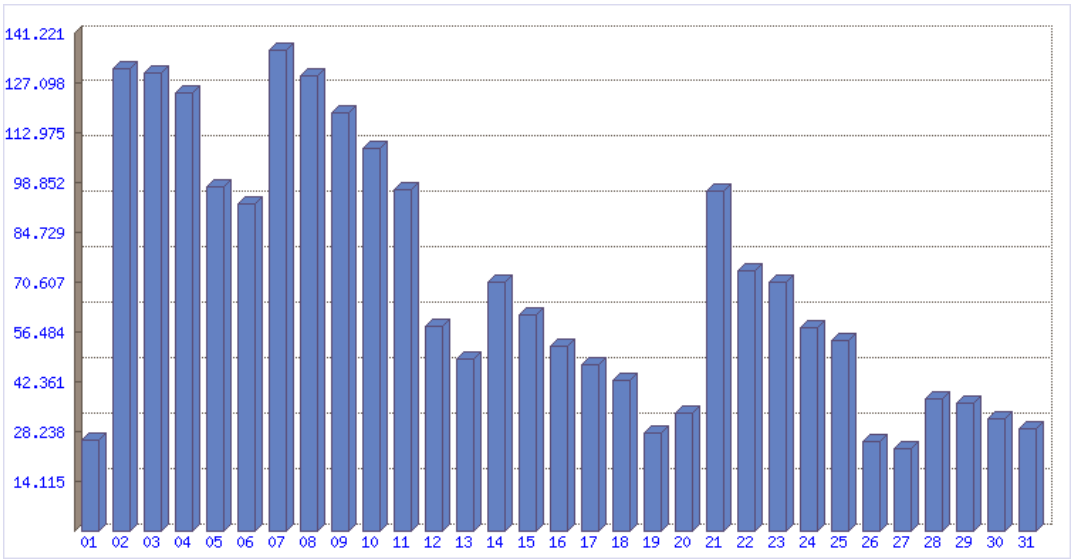
Bu bölümde Sakarya Üniversitesi web sitesi trafik ölçümleri yapılacak ve sitenin performansı ortaya konulacaktır. İzlenme analizi başlığı altında farklı kategorilerde örneklemler sunularak bu örneklemlerin teknik analizlerine değinilecektir.

5.5.1.1. Kullanıcı – oturum sayıları ve frekansları

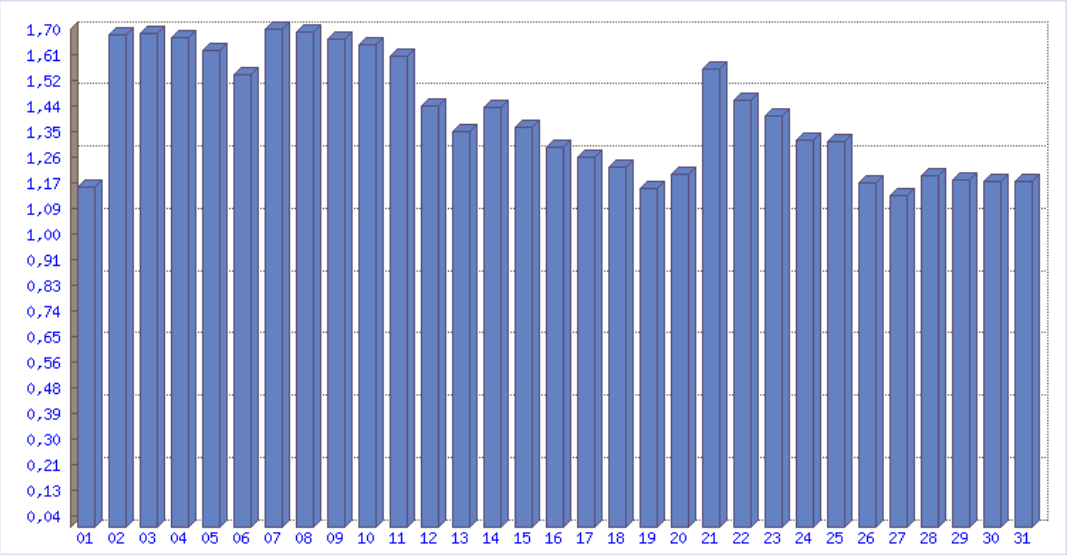
Bu bölümde sistemin kullanıcı sayıları, oturum sayıları ve frekansları belirlenip bu bilgiler ışığında kullanıcı verimliği değerlendirilmiştir.



Şekil 5.4. Ocak 2008 Günlük kullanıcı sayıları



Şekil 5.5. Ocak 2008 Günlük oturum sayıları



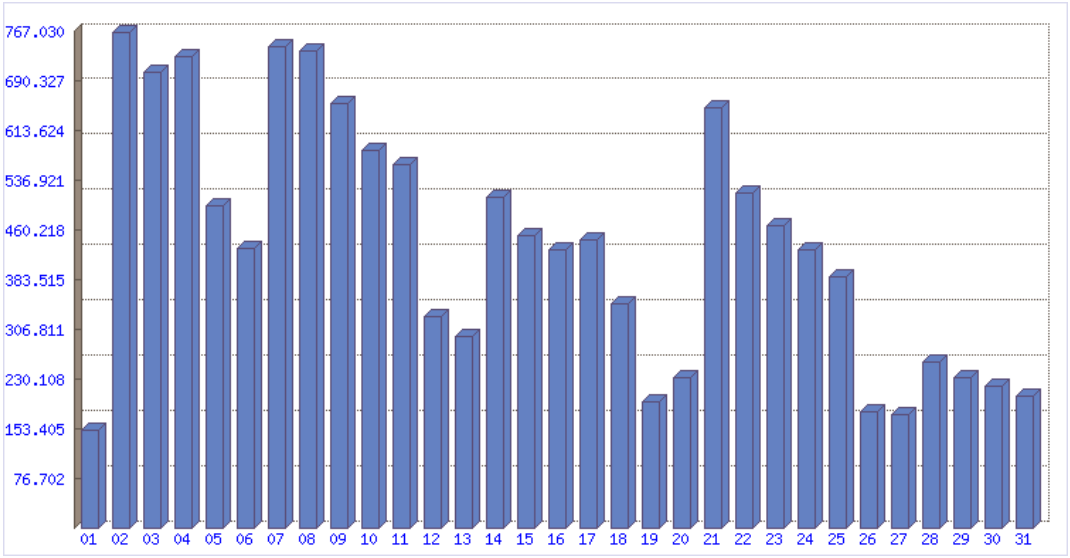
Şekil 5.6. Ocak 2008 Günlük kullanıcı frekansları

Şekil 5.4. ve 5.5. incelendiğinde hafta sonuna denk gelen günlerde kullanıcı ve oturum sayılarında ani azalmaların olduğu görülmektedir. Bu durum web sitesinin eğitim – öğretimin devam ettiği günlerde yoğun olarak kullanıldığını göstermektedir.

Kullanıcı frekansı oturum sayısının kullanıcı sayısına bölünmesi ile elde edilmektedir. Yani kullanıcının sitede günlük kaç adet oturum açtığını ifade etmektedir. Şekil 5.6. incelendiğinde günlük kullanıcı frekansının 1,30 civarında olduğu anlaşılmaktadır. Siteye giriş yapan kullanıcının aynı gün içinde tekrar siteye girme oranı %30 civarındadır.

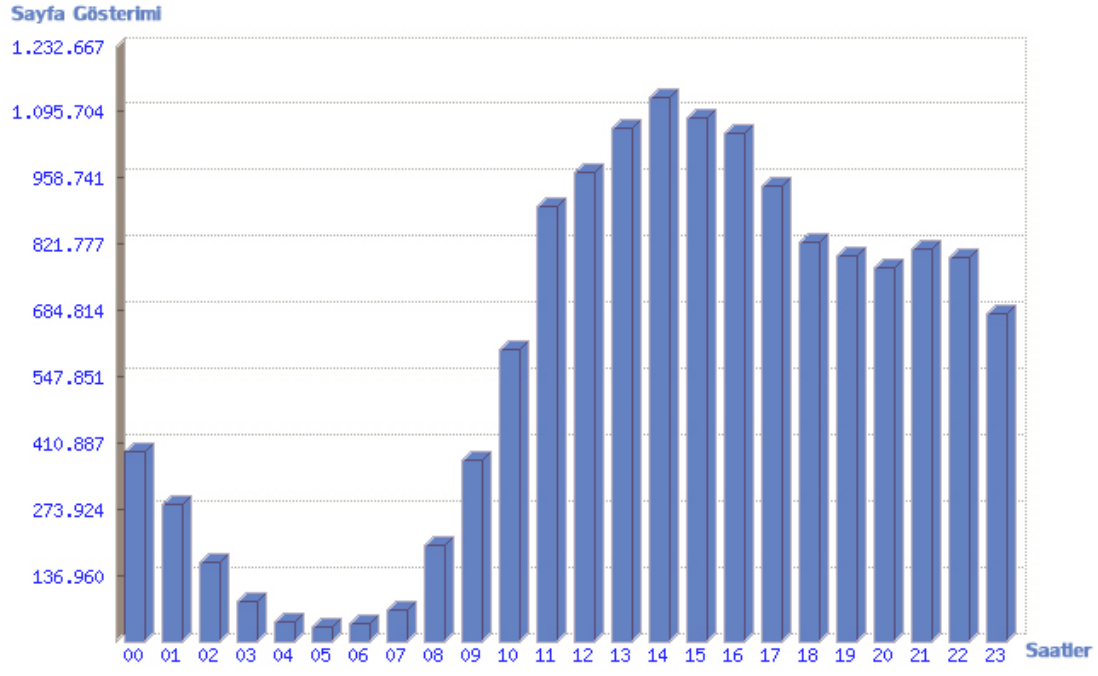
5.5.1.2. Sayfa gösterimi (page view)

Bu bölümde kullanıcı başına sayfa gösterimi ve oturum başına sayfa gösterimi çalışmaları verilmiştir.



Şekil 5.7. Ocak 2008 Toplam sayfa gösterimi

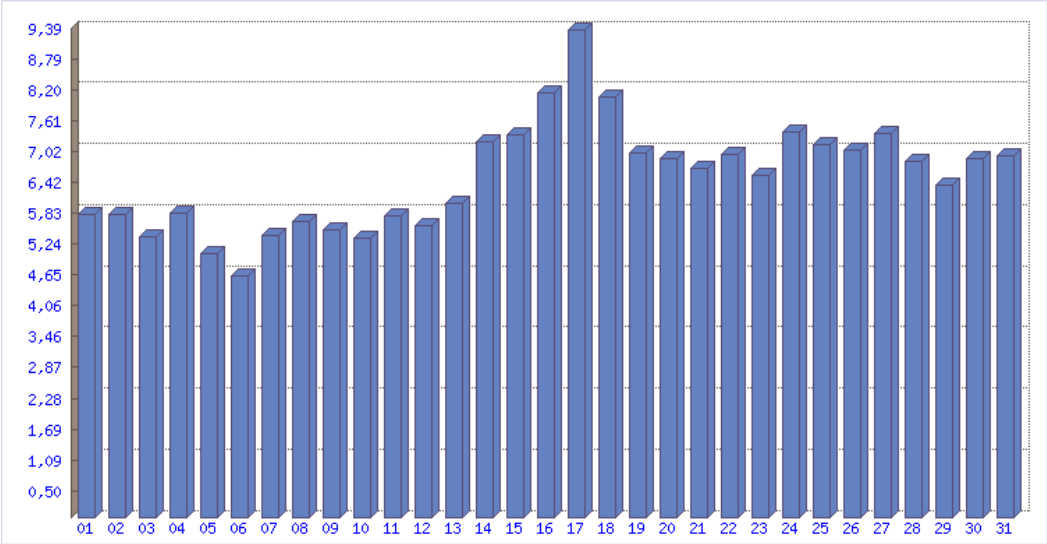
Sayfa görüntüleme değerlerinin günlük ortalama 450.635 civarlarında olması sistemin yoğun bir şekilde kullanıldığını ifade eder. Bu yoğunlukta bir sistemin kesintisiz hizmet, felaket senaryoları, sistem bakımı vb.. altyapı çalışmaları azami ölçüde iyi planlanmalıdır. Günler arasındaki sayfa gösterim değerlerindeki yüksek farklılık hafta sonları ve dönemsel derse yazılma, sınav sonuçlarının ilanı gibi nedenlerden kaynaklanmaktadır.



Şekil 5.8. Ocak 2008 Saatlik sayfa gösterimleri

Şekil 5.8.'de 2008 yılı Ocak ayına ait toplam 13.969.685 adet sayfa gösteriminin saatlere göre dağılımı verilmiştir. Bu grafik yorumlandığında sistemin en yoğun saat 14 civarların, en az da saat 05 civarlarında kullanıldığı tespit edilmiştir. Bu veriler ışığında sistem bakımları, yedekleme, sunucu güncellemeleri gibi alt yapı çalışmalarının saat 05 civarlarında yapılmasının uygun olacağı sonucu çıkarılmıştır.

Bundan sonraki adımda sayfa gösterimlerinin oturum başına ağırlıklarına bakabiliriz. Bu analizdeki amaç bir oturumda ortalama kaç sayfanın görüntülendiğini belirlemek ve elde edilen verilere göre oturum verimliliğini tespit etmektir.



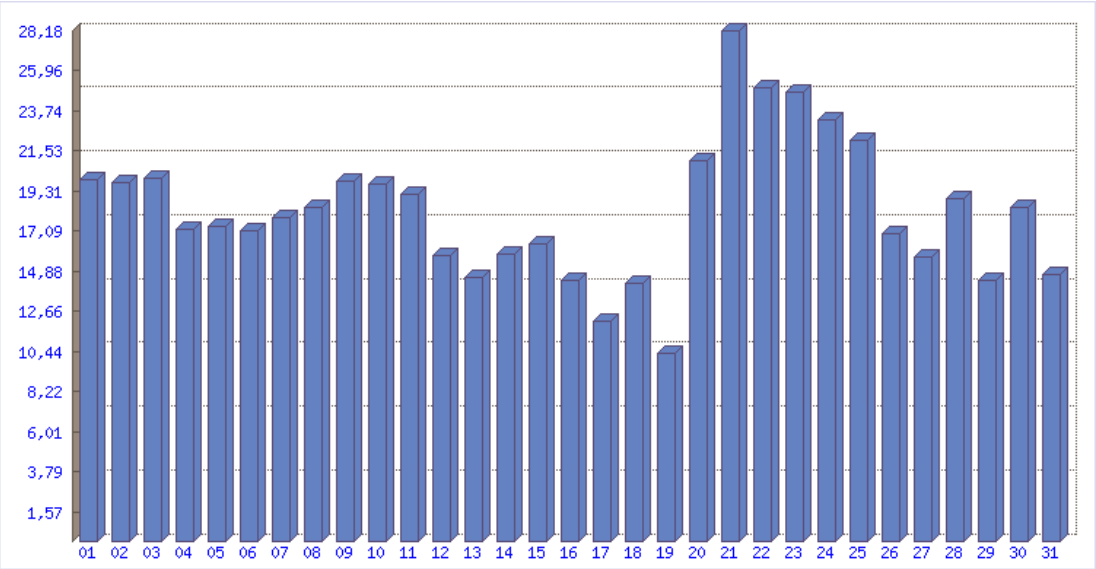
Şekil 5.9. Ocak 2008 Oturum başına sayfa gösterim değerleri

Şekil 5.9. incelendiğinde bir oturumda ortalama 6,7 sayfa gösterim yapıldığı belirlenmektedir. Bu durum eğitim kurumu web sitesi açısından normal olarak bakılabilir. Ayın 16, 17 ve 18. günlerinde (Çarşamba, Perşembe, Cuma) frekansın yüksek çıkmasının sebebi olarak 2007-2008 öğretim yılı güz dönemi yılsonu başarı notlarını öğrenmek için öğrencilerin fazla sayfa görüntüleme yapmalarından kaynaklanmaktadır.

5.5.1.3. Süreler

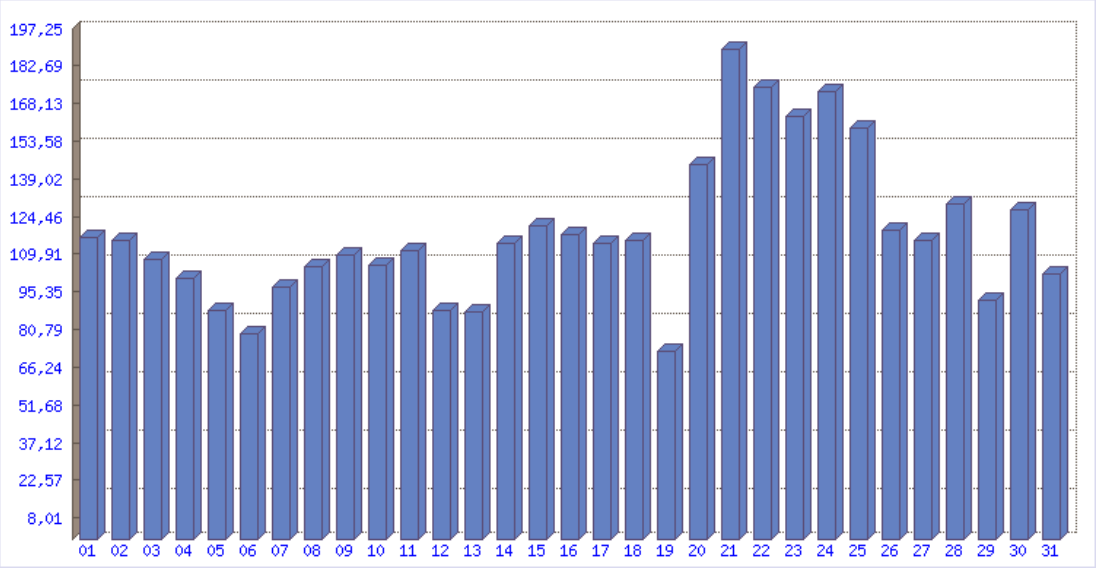
Bu bölümde kullanıcıların bir sayfada ortalama kaç saniye kaldığına ve bir oturumun ortalama kaç saniyede sonlandığına değinilecektir. Bu verilerin önemi, genelde web sitesinin, özelde ise bir sayfanın kullanıcıyı ne kadar süre tutabildiğidir. CAWIS

bünyesinde hizmet veren WebMail servisi bu sürelerin artmasına, WebMenu ve WebRehber gibi kısa sürelerde işlemlerin tamamlanacağı statik sayfalar da sürelerin azalmasına neden olabileceği öngörülmektedir.



Şekil 5.10. Ocak 2008 Ortalama sayfa görüntüleme süresi (sn.)

Ocak ayı ortalama sayfa görüntüleme süresi 18,8 sn. çıkmaktadır.



Şekil 5.11. Ocak 2008 Ortalama oturum süresi (sn.)

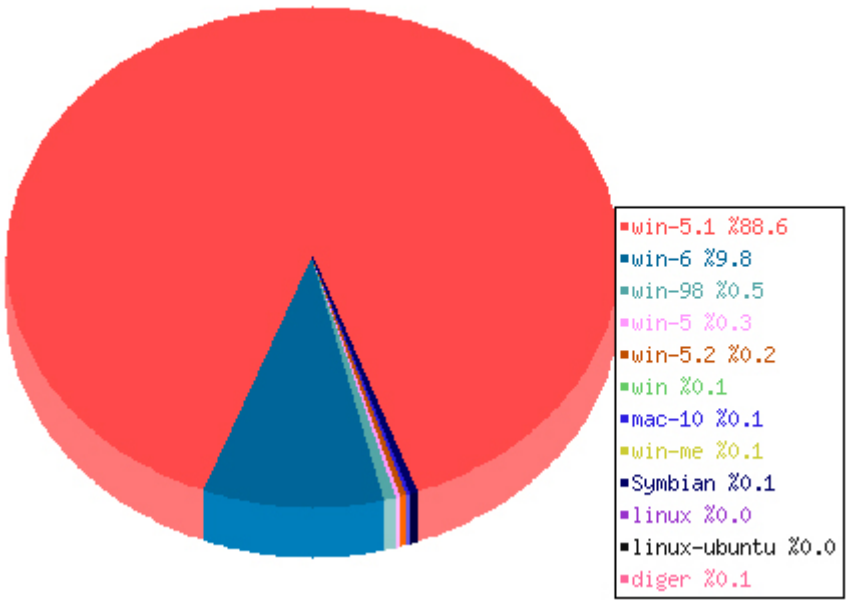
Kullanıcının bir oturum süresi hesaplanırken birçok çalışma bu süreyi 30 dakika olarak temel almıştır [65]. Bu çalışmada oturum kontrolü gerçek zamanlı yapıldığı için bu veri dikkate alınmaksızın gerçek oturumlar ve bunlara ait gerçek işlem

süreleri hesaplanmıştır. Ortalama 121,33 saniye süren oturumlar siteye giriş çıkışların fazla olduğunu göstermektedir. Bu durum yükseköğretim kurumu sitesi için normal bir durum olarak karşılanabilir. Bu sürenin artırılması ders dokümanları paylaşımı, tartışma forumlarının hazırlanması gibi kullanıcıyı web sitelerinde tutmaya yönelik çalışmalarla artırılabilir. Bu değerler online eğitim veren bir kurum sitesinden elde edilmiş olsa idi ortada büyük bir sorun olduğundan bahsedilebilinirdi. Sakarya Üniversitesi web sitesi için örneğin bir kullanıcı siteye girer yemek menüsüne bakar ve sistemi terk eder ya da öğrenci sınav sonucu açıklanmış ise sınav sonucuna bakar ve sistemi terk eder.

5.5.2. Teknik analizi

Bu bölümde Sakarya Üniversitesi web sitesi kullanıcılarının (uzak bilgisayarların) teknik analizlerinin belirlenmesi çalışmaları yapılmıştır. Teknik analiz, sistemin hitap ettiği uzak bilgisayar açısında erişilebilirliği, desteklenebilirliği, kullanabilirliliği gibi gereklilikler açısından önemli veriler sunmaktadır.

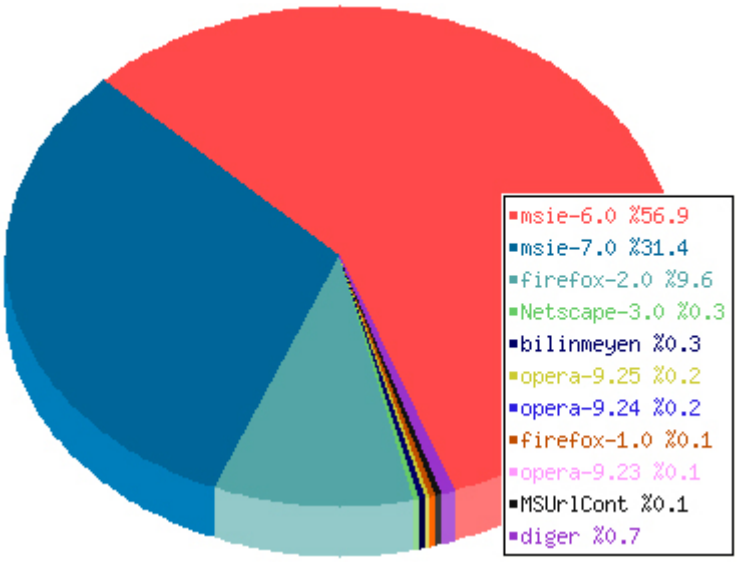
5.5.2.1. İşletim sistemleri



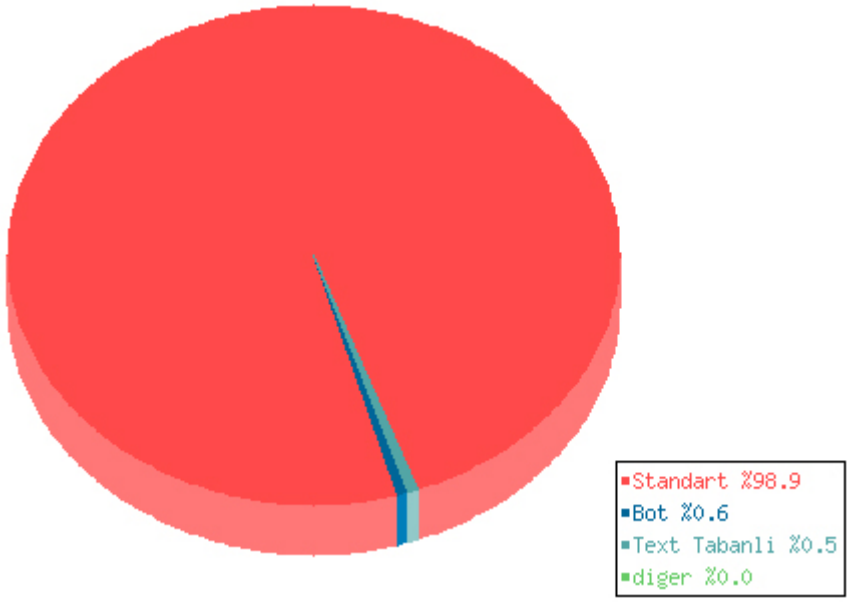
Şekil 5.12. Ocak 2008 Uzak bilgisayarların işletim sistemi dağılımı

Toplam 2.300.217 oturuma ait işletim sistemi verisinden %88'ini Windows XP oluşturmaktadır.

5.5.2.2. Tarayıcı bilgileri



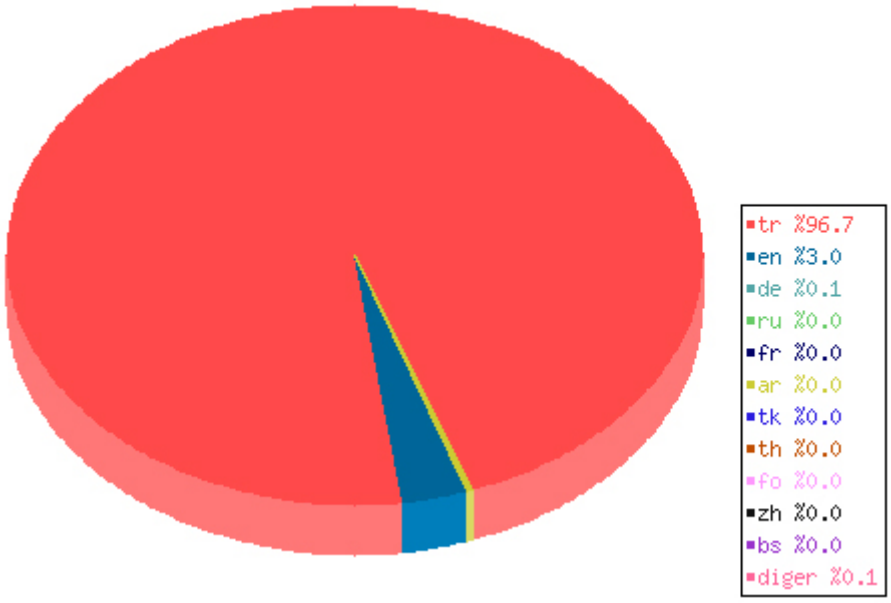
Şekil 5.13. Ocak 2008 Uzak bilgisayarların tarayıcı dağılımı



Şekil 5.14. Tarayıcı tiplerinin dağılımı

Toplam 2.322.429 oturumun tarayıcı dağılımı dikkate alındığında web sitesinin kodlanması sonrası farklı tarayıcılar üzerinde tasarım ve script denemeleri yapılması gerekliliği ortaya çıkmaktadır. Ayrıca web sitesinden istemde bulunan 2.322.429 oturumun %0,6'sını arama motorları oluşturmaktadır. Bu veri gerçek oturumların belirlenmesi aşamasında analiz dışı bırakılması gereken oturumların belirlenmesinde kullanılmıştır.

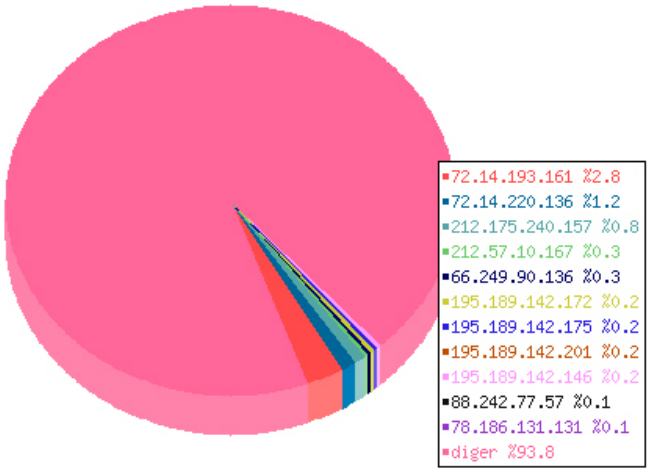
5.5.2.3. Dil bilgileri



Şekil 5.15. Ocak 2008 Tarayıcı dili dağılım grafiği

Toplam 2.289.838 oturuma ait tarayıcı dili dağılım grafiği Şekil 5.15.'de görülmektedir.

5.5.2.4. Proxy bilgileri

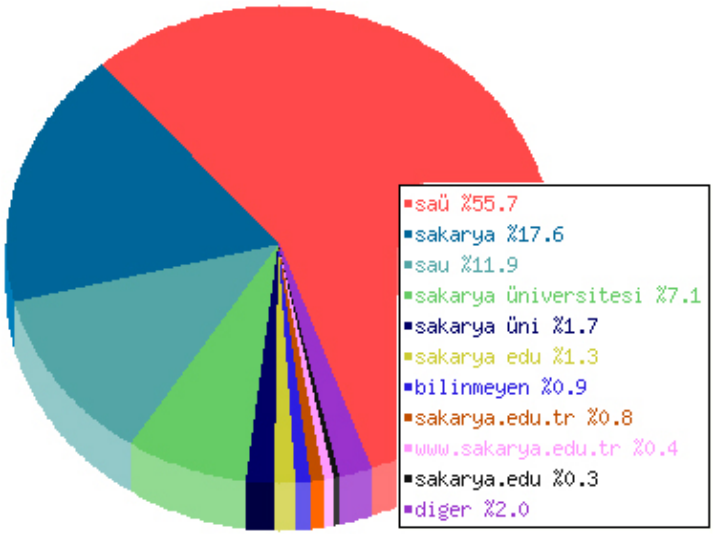


Şekil 5.16. Ocak 2008 Proxy dağılım grafiği

Toplam 2.300.000 civarındaki oturumdan sadece 107.853 tanesi tarafından vekil sunucu kullanılmaktadır. Bu durum web sitesinde oturum açan kullanıcıların %4'ünün vekil sunucu kullandığını ortaya koymaktadır. 210 farklı vekil sunucunun, en çok kullanılan 11'inin dağılımı Şekil 5.16.'daki gibi oluşmuştur.

5.5.3. Arama motoru analizi

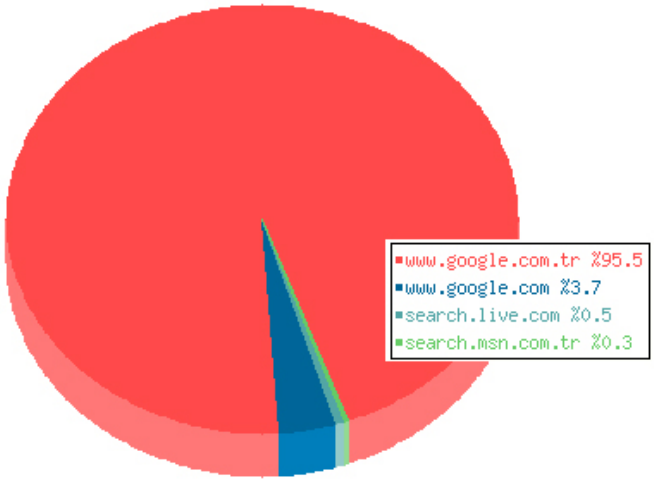
5.5.3.1. Anahtar kelimeler



Şekil 5.17. Ocak 2008 Anahtar kelimelerin dağılımı

Şekil 5.17.'de Sakarya Üniversitesi web sitesine arama motorlarını kullanarak giriş yapan 275.845 kullanıcının arama yaptığı ifadelerin dağılımı verilmiştir. Arama motorlarında yer alan sponsor linkler için ticari sitelerin bu verileri dikkate almaları gerekmektedir. Hangi ifadeye sponsor link ücreti ödemenin daha verimli olacağının değerlendirilmesi bu veriler ışığında ortaya konulabilir.

5.5.3.2. Arama motorları

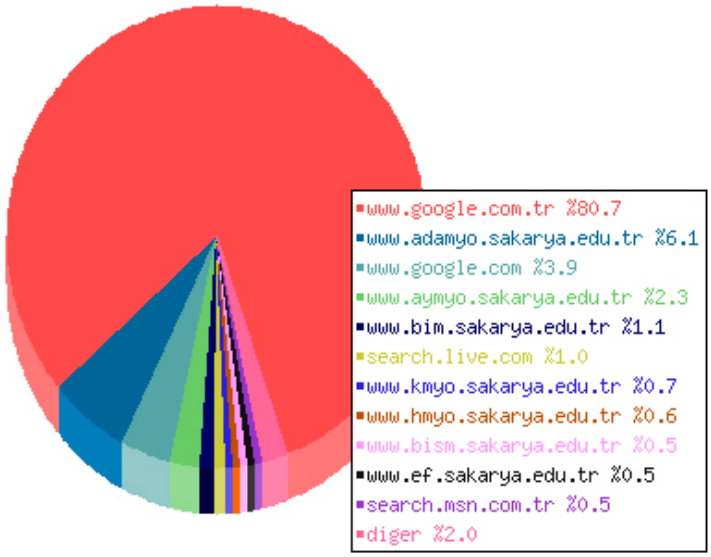


Şekil 5.18. Ocak 2008 Arama motoru dağılımı

Şekil 5.18.'de Ocak ayında Sakarya Üniversitesi web sitesine arama motorlarını kullanarak giriş yapan 275.845 kullanıcının kullandığı arama motoru dağılımı verilmiştir.

5.5.4. Stratejik analizi

5.5.4.1. Gelinen domainler (Referrer)

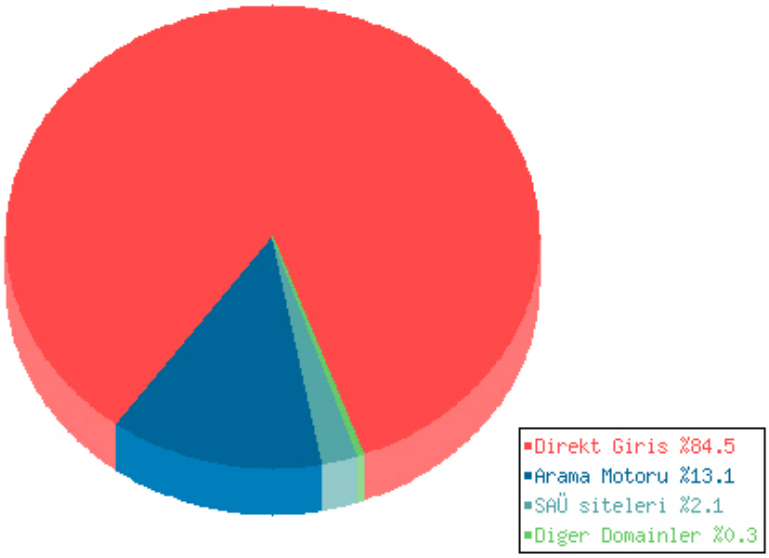


Şekil 5.19. Ocak 2008 Gelinen domain dağılım grafiği

Gelinen domain dağılımı, bir web sitesinin hangi siteler üzerinden kullanıcı topladığını ifade etmesi açısından önem taşımaktadır. Diğer sitelere reklam verme durumunu bu bilgiler ışığında değerlendirmek verimliliği artıracaktır.

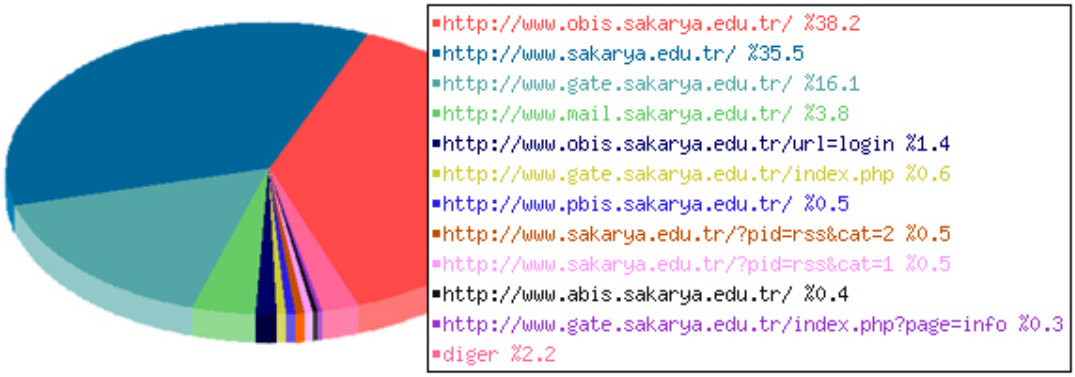
Örneğin 15.07.2007 tarihinde Hürriyet gazetesinde çıkan Sakarya Üniversitesi E-MBA reklamın ardından aynı gün "www.hurriyet.com.tr" sitesinden 127 adet farklı kullanıcı sisteme giriş yapmıştır. Bu verilere ulaşılması reklamın ne derece verimli olduğunu ortaya koymaktadır.

5.5.4.2. Siteye nasıl giriş yapıldığı



Şekil 5.20. Ocak 2008 Siteye nasıl giriş yapıldığı dağılımı

5.5.4.3. Siteye giriş noktaları



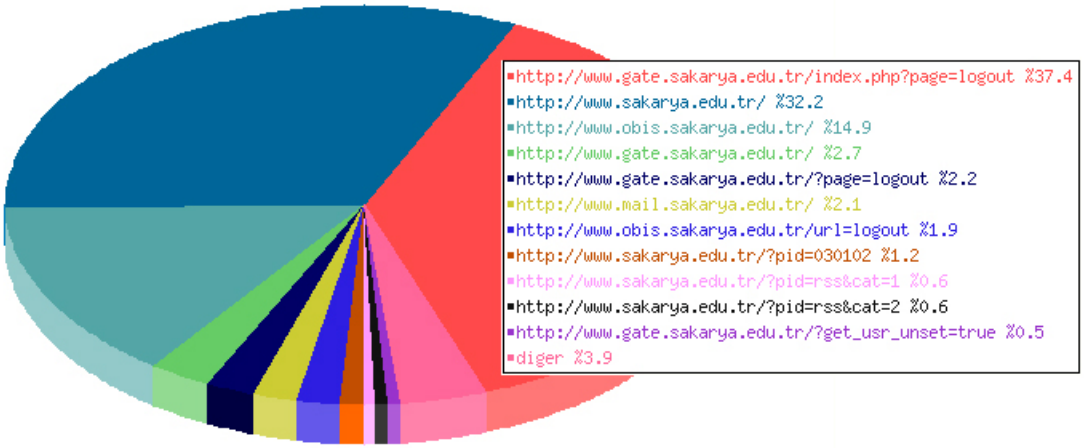
Şekil 5.21. Ocak 2008 Siteye giriş noktaları dağılımı

Toplam 2.273.737 adet kullanıcı girişi incelendiğinde yukarıdaki dağılımda belirtilen giriş noktaları yoğun olarak kullanılmıştır.

Giriş noktalarının önemi web sitesi menü tasarımı aşamasında hangi sayfaların ön planda sunulacağı bilgisini vermesidir. En çok giriş noktası olarak WebObis (Öğrenci Bilgi Sistemi) servisinin kullanılması sisteme kayıtlı öğrenciler tarafında web sitesinin yoğun olarak kullanıldığını ortaya koymaktadır. Bu veriler kurum dışı kullanımın düşük olduğu bilgisini çıkarabiliriz. Yani kurum ana sayfasının giriş

noktası sıralamasında ikinci sırada yer alması web sitesinin diğer kullanıcılar tarafından az kullanıldığını göstermektedir.

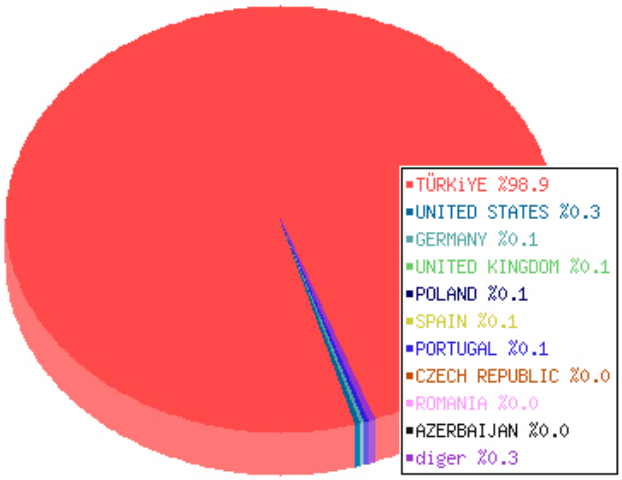
5.5.4.4. Siteden çıkış noktaları



Şekil 5.22. Ocak 2008 Siteden çıkış noktaları dağılımı

Web sitesinden çıkış noktaları siteyi ziyaret eden kullanıcının en son görüntülediği sayfayı ifade etmektedir. Bu veriler incelendiğinde kullanıcıların oturumlarını kapatarak sistemden ayrılma durumlarının %41 düzeyinde olduğunu ortaya koymaktadır (page=logout sayfaları toplamı).

5.5.4.5. Ülke bilgileri

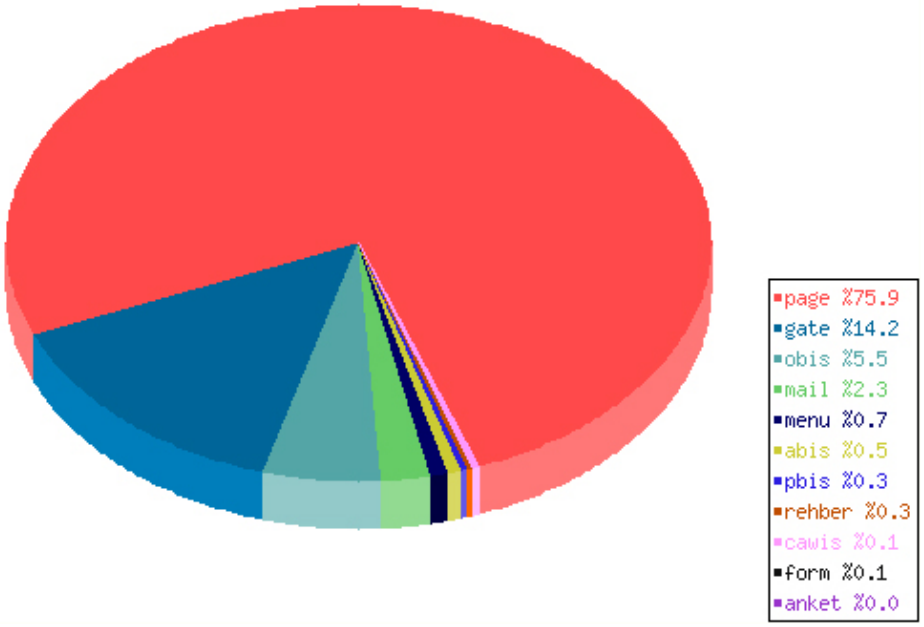


Şekil 5.23. Ocak 2008 Ülke dağılım grafiği

Toplam 2.321.580 kullanıcıya ait ülke dağılım grafiği Şekil 5.23.'de gösterilmiştir. Web sitelerinin, kurumların dünyaya açılan kapıları olduğu düşünüldüğünde bu

veriler daha da anlam kazanmaktadır. Erasmus öğrenci değişim programı yaygınlaştıkça Avrupa ülkelerinden Sakarya Üniversitesi web sitesine girişlerin, önceki yıllara ait veriler incelendiğinde, arttığı anlaşılabilmektedir.

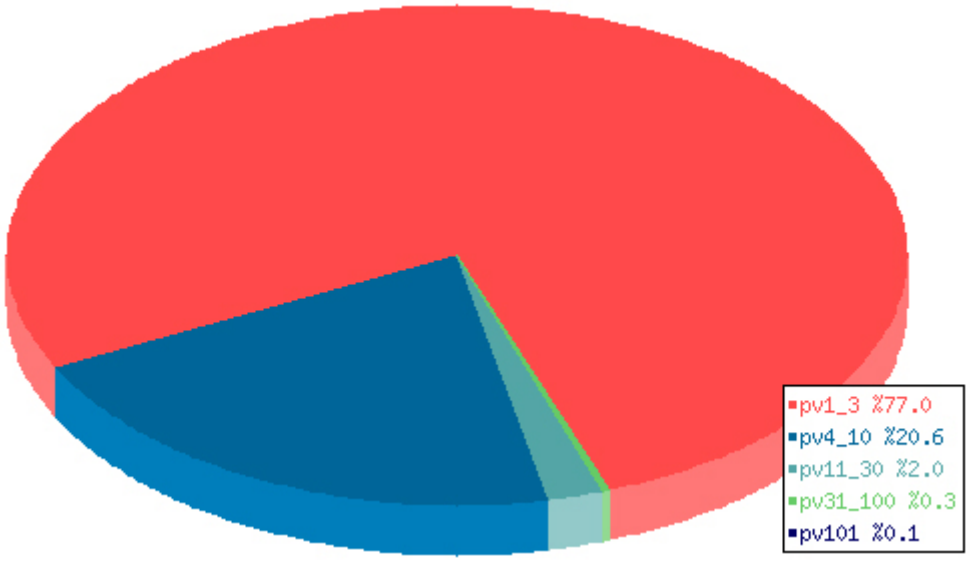
5.5.4.6. Servis kullanım bilgileri



Şekil 5.24. Ocak 2008 CAWIS servisleri kullanım grafiği

Yukarıdaki grafikte "page" olarak isimlendirilen servis Sakarya Üniversitesi ana sayfasıdır. Şekil 5.21.'de siteye giriş noktaları grafiğine bakıldığında en çok giriş noktası olarak Obis servisi bulunmuşken bu grafikte ana sayfanın en yoğun kullanıma sahip olduğu ortaya çıkmaktadır.

5.5.4.7. Sayfa gösterimlerinin grupsal dağılımı



Şekil 5.25. Ocak 2008 Sayfa gösterimlerinin grupsal dağılımı

Bu grafikte belirtilen pv1_3 ifadesi 1-3 adet sayfa görüntüleme yapan oturumları ifade etmektedir. Toplam 73.448.525 sayfa gösteriminin %77'sini, bir oturumda ortalama 1 ile 3 arasında sayfa görüntüleyenlerin oluşturduğu belirlenmiştir.

Bir oturumda 100 sayfadan fazla görüntüleyenlerin oranı % 0,1 oranında olsa da bu çok önemli bir değerdir. Bu veri, web sitesine saldırı yapan kullanıcıların belirlenmesi açısından kullanılabilir.

BÖLÜM 6. SONUÇLAR VE ÖNERİLER

Bu tez çalışmasında Sakarya Üniversitesi Web Sitesi Web Madenciliği disiplini kapsamında analiz edilmiştir. Her şeyden önce bir web sitesi ait olduğu kurumun dünyaya açılan kapısıdır. Bu kapıyı verimli şekilde kullanabilmenin en önemli adımı o sitenin iyi bir şekilde analiz edilmesinden geçer. Günümüz bilim dünyasının yeni eğilimi olan veri madenciliği ile bu işlem gerçekleştirilmiştir. Çalışmada öncelikle Veri madenciliğini tanıtılmış, ardından Veri Tabanından Bilgi Keşfine değinilmiş, devamında çalışmada kullanılan web madenciliği üzerinde durulmuş ve uygulama sonuçları sunulmuştur. Çalışmanın başlıca hedefi olan Sakarya Üniversitesi web sitesinin analizi kapsamında

- Site trafiğinin ortaya konması,
- Kullanım durumunun ortaya konması,
- Kullanıma sunulan servislerin verimliliğinin belirlenmesi,
- Site erişim noktalarının belirlenmesi,
- Arama motorlarının ve anahtar kelimelerinin tespit edilmesi,
- Ziyaretçilerin kullanımları doğrultusunda kısa yolların oluşturulması,
- Uzak bilgisayarların teknik analizlerinin yapılarak site yapısının en geniş perspektifte sunumunun oluşturulması,
- Aşırı isteklerin tespit edilerek sistem kaynaklarının gereksiz kullanımının engellenmesi
- Site içeriğinin kullanıcıları sayfada tutma süreleri,
- Kullanıcı / oturum ilişkisi ve verimliliği
- Kullanıcı bağımlılığı gibi analizler ortaya konmuştur.

Uygulama bölümünde kullanılan veriler web kayıt dosyasının ayıklanması işleminde karşılaşılan zorlukları ve veri güvenilirliği problemlerini ortadan kaldırmak için sitenin web kayıtları özel bir betik hazırlanarak oluşturulmuştur. Bu durum hem

analiz aşamasını kolaylaştırmış hem de gerçek zamanlı web kayıtlarının istenilen formatta elde edilmesi sağlanmıştır.

Yapılan tez çalışmasında web madenciliği disiplininin çok geniş çalışma alanına sahip olduğu görülmektedir. Bu durum web madenciliğinin, istatistik, matematik, işletme, bilgisayar gibi birçok bilim dalı ile ortak zemini paylaştığını ortaya koymaktadır.

Bundan sonra yapılacak çalışmalarda elektronik ticaret yapan firmalar açısından web madenciliği sonucu elde edilen verilerin ticari getirisinin öneminin ortaya konması gerekmektedir. E-ticaret sitelerinin veri paylaşımı konusunda daha hassas davranması bu çalışmaların gerçekleştirilebilmesi için hayati önem taşımaktadır.

KAYNAKLAR

[1] TANASA, D., TROUSSE, B., "Advanced Data Preprocessing for Intersites Web Usage Mining", IEEE Intelligent Systems, P:59-64, March/April 2004.

[2] OLSON, D.L., DELEN, D., "Advanced Data Mining Techniques", ISBN: 978-3-540-76916-3, Springer, 2008.

[3] CARUS, A., MESUT, A., "Web Kullanım Madenciliği Uygulaması", II. Mühendislik Bilimleri Genç Araştırmacılar Kongresi, İstanbul, S:120-127, 2005.

[4] PORTER, J., President, "Disk/Trend Disk Drives Evolution", "Magnetic Recording and Information Storage Santa Clara University, December 14, 1998", "http://www.disktrend.com/pdf/portrpkg.pdf", Mart 2008.

[5] IAN, W., EİBE, F., "Data Mining Practical Machine Learning Tools and Techniques, Second Edition", Morgan Kaufmann Publishers ISBN: 0-12-088407-0, 2005.

[6] EKER, H., "Veri Madenciliği Veya Bilgi Keşfi", "http://www.bilgiyonetimi.org/cm/pages/mkl_gos.php?nt=538", Nisan 2008.

[7] BAYKAL N., "Veri Tabanı ve Veri Madenciliği", Antalya, 2003.

[8] ALPAYDIN, E., "Zeki veri madenciliği: Ham Veriden Altın Bilgiye Ulaşma Yöntemleri" Bilişim 2000 Eğitim Semineri, 2000.

[9] THUARISINGHAM, B., M., "Web Data Mining and Applications in Business İntelligence and counter Terrorism", Auerbach Publishers, incorporated, 2003.

[10] ÖZMEN, Ş., "İş Hayatı Veri Madenciliği İle İstatistik Uygulamalarını Yeniden Keşfediyor", http://idari.cu.edu.tr/sempozyum/bil38.htm, Marmara Üniversitesi Mart 2008.

[11] BING, L., "Web Data Mining : Exploring Hyperlinks, Contents, and Usage Data", ISBN-10 3-540-37881-2, Springer-Verlag Berlin Heidelberg, 2007.

[12] GÜVENÇ, E., "Student Performance Assesment in Higher Education Using Data Mining", Yüksek Lisans Tezi, Boğaziçi Üniversitesi, İstanbul, 5-14, 2001.

[13] MANGANARIS, S., CHRISTENSEN , M., ZERKLE, D., HERMIZ, K., "A data mining analysis of RTID alarms", Computer Networks 34 571–577, 2000.

[14] HUDAIRY, H., "Data mining and decision making support in the governmental sector", Faculty of Graduate School of The University of Louisville, Kentucky, 2004.

[15] "Current data mining applications/industries" "http://www.kdnuggets.com/polls/2003/data_mining_applications_industries.htm", Nisan 2008.

[16] ZHANG, N., ZHOU, L., "Mehodologies for knowledge Discovery and Data Mining", Third Pacific-Asia Conference, Pak-dd99, China, 1999.

[17] BRAMER, M., "Principles of Data Mining", Undergraduate Topics in Computer Science ISSN 1863-7310, Springer-Verlag London Limited, 2007.

[18] HAN, J., KAMBER, S.F., "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, 2001.

[19] VELICKOV, S., SOLOMATINE, D.P., "Predictive Data Mining: Practical Examples." In: AI methods in Civil Engineering Applications, Cottbus, 2000.

[20] ÖZÇINAR, H., "KPSS Sonuçlarının Veri Madenciliği Yöntemleriyle Tahmin Edilmesi", Pamukkale Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi, Denizli, 2006.

[21] http://www.bilgiyonetimi.org/cm/pages/mkl_gos.php?nt=538, "Veri Madenciliği Veya Bilgi Keşfi", Nisan 2008.

[22] BIDGOLI, B. M., "Data Mining For a Web Based Educational System", Ph. D. Thesis, Department of Computer Science and Engineering, Michigan State University, 2004.

[23] SHAH, S., KURSAK, A., "Data Mining And Genetic Algorithms Based Gene / SNP Selection", Artificial Intelligence in Medicine, 2004.

[24] FAYYAD, U., PIATETSKY-SHAPIRO, G., SMYTH, P., "From Data Mining to Knowledge Discovery in Databases", American Association for Artificial Intelligence, 1996.

[25] BELGİN, Ö., "Finding and Evaluating Patterns in Web Repository Using Database Technology and Data Mining Algorithms.", Yüksek Lisans Tezi, İzmir Teknoloji Enstitüsü, İzmir, 2002.

[26] PRYKE, A. N., "Data Mining Using Genetic Algorithms and Interactive Visualization", PhD Thesis Faculty of Science, University of Birmingham, Birmingham, 187s, 1998.

[27] TAN, P.N., STEINBACH, M., KUMAR, V., "Introduction to Data Mining", ISBN-10: 0321321367, Addison-Wesley, 2006.

[28] WANG, W., "Classification and Pattern Matching Methods", M.S. Thesis, Com Beijing Polytechnic University, 1999.

[29] GÜVENÇ, E., "Yüksek Öğrenimde Öğrenci Performansının Veri Madenciliği Teknikleri ile Belirlenmesi", Yüksek Lisans Tezi, Endüstri Mühendisliği ABD, Fen Bilimleri Enstitüsü, Boğaziçi Üniversitesi, İstanbul, 2001.

[30] GOLDBERG, D. E., "Genetic Algorithms in Search, Optimization and Machine Learning", Addison-Wesley, USA, 1989.

[31] ELMAS, Ç., "Yapay Sinir Ağları (Kuram,Mimari,Uygulama)", Seçkin Yayınları, 2003.

[32] TEMURTAS, F., "Yapay Sinir Ağları Ders Notları" Sakarya Üniversitesi, 2005.

[33] ARYEETEY, K., "Data Analysis and Predictive Modelling Using The Variable Precision Rough Set Approach", Master Thesis, Faculty of Graduate Study and Research of University of Regina, Canada, 2003.

[34] CHEESEMAN, P., "A Bayesian Classification System.", On Machine Learning., Morgan Kaufman, 1988.

[35] XU, Y., "Using Data Mining In Educational Research: A Comparison of Bayesian Network With Multiple Regression in Prediction", Department of Educational Psychology, The University of Arizona, Arizona, 2003.

[36] HUI, S., JHA, G., "Application Data Mining for Customer Service Support", Information and Management, 2000.

[37] SORENSEN, K., JANSSENS, G.K., "Data Mining With Genetic Algorithms on Binary Trees", European Journal of Operational Research, 2003.

[38] BERRY, M.J.A., LINOFF, G.S., "Mastering Data Mining: The Art and Science of Customer Relationship Management,", John Wiley & Sons, 1st Ed., 1999.

[39] AKPINAR, H., "Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği", İ.Ü. İşletme Fakültesi Dergisi, C:29, s: 1-22, S: 1/Nisan 2000.

[40] ZHONG, N., ZHOU, L., "Methodologies for Knowledge Discovery and Data Mining", Springer, Germany, 1999.

[41] SEVER, H., OĞUZ, B., "Veri Tabanlarında Bilgi Keşfine Formel Bir Yaklaşım", Bilgi Dünyası 3(2):pp. 173-204, 2002.

[42] ZAKI, M.J., "Parallel and Dİstributed Association Mining : A Survey", IEEE, S:14-25, October–December 1999.

[43] ÖZEKES, S., "Veri Madenciliği Modelleri ve Uygulama Alanları", İstanbul Ticaret Üniversitesi Dergisi, No:3 ,pp. 65-82, 2003.

[44] MATHEUS, C.J., CHAN, P.K., PIATETSKY-SHAPIRO, G., "Systems for Knowledge Discovery in Databases", IEEE TKDE, 1993.

[45] AKBULUT , S., "Veri Madenciliği Teknikleri ile Bir Kozmetik Markanın Ayrılan Müşteri Analizi ve Müşteri Segmentasyonu", Yüksek Lisans Tezi, Gazi Üniversitesi, 2006.

[46] AMIRI, N., MATTHEWS, D., GAO, Q., "Designing a Framework of Intelligent Information Process on Dentistry Administration Data", International Journal of Computerized Dentistry, Vol. 8, No. 2, 2005.

[47] QUINLAN, J. R., "Induction of Decision Trees", Kluwer Academic Publishers, USA, p: 81-106, 1986.

[48] AYDOĞAN, F., "E-Ticarette Veri Madenciliği Yaklaşımlarıyla Müşteriye Hizmet Sunan Akıllı Modüllerin Tasarımı ve Gerçekleştirmi", Yüksek Lisans Tezi, Hacettepe Üniversitesi, 2003.

[49] ÖZAKAR, B., "Tıkların Dili", "http://inet-tr.org.tr/inetconf8/sunum/119.ppt", Nisan 2008.

[50] ÖZAKAR, B., PÜSKÜLCÜ, H., "Tıkların Dili Web içerik ve Web Kullanım Madenciliği Tekniklerinin Entegrasyonu ile Oluşmuş Bir Veri Tabanından Nasıl Yararlanılabilir?", "http://www.teknoturk.org/docking/yazilar/tt000119-yazi.htm", Nisan 2008.

[51] ELGÜN, C.Ç., "Web Madenciliği", "http://www.bilyaz.com/bMakaleGetir.php?id=56", Mart 2008.

[52] MARKOV, Z., LAROSE, D.T., "Data Mining the Web", Wiley-Interscience a John WILEY & SONS, inc., Publication, 2007.

[53] SUMATHI, S., SIVANANDAM, S.N., "Introduction to Data Mining and its Applications", Springer, 2006.

[54] SCIME, A., "Web Mining: Applications and Techniques", Idea Group Publishing, 2005.

[55] WANG, X., ABRAHAM, A., SMITH, K. A., "Intelligent web traffic mining and analysis", Journal of Network and Computer Applications 28 147–165, 2005.

[56] AYAZ, R., "Web Madenciliğine Bir Bakış", YTÜ Bilgisayar Mühendisliği, 2003.

[57] "Web Characterization Terminology & Defininiation Sheet", "http://www.w3.org/1999/05/WCA-terms/", W3C Working Draft, 1999, Mart 2008.

[58] GEZER, M., EROL, Ç., GÜLSEÇEN, S., "Bir Web Sayfasının Web Madenciliği İle Analizi", Akademik Bilişim 2007, Dumlupınar Üniversitesi, Kütahya 31 Ocak-2 Şubat 2007.

[59] DUNHAM, M.H., "Data Mining Introductory and Advanced Topics", Prentice Hall, New Jersey, 5-19 P, 195-220 P, 2003.

[60] BELEN, E., ÖZGÜR, Ç., ÖZAKAR, B., "WALA : Web Erişim Kütük Araştırmacısı", "http://inet-tr.org.tr/inetconf9/bildiri/60.doc", Mart 2008.

[61] SRISTAVA, J., COOLEY, R., DESHPANDE, M., TAN, P.-N., "Web Usage Mining: Discovery and Applications of Usage Patterns From Web Data." SIGKDD Explorations, V:1, I:2, p:12-23, 2000.

[62] MADRIA, S. K., "Web Mining : A Bird's Eye View", Department of Computer Science, University of Missouri-Rolla, USA, 2008.

[63] DAŞ, R., TÜRKOĞLU, İ., POYRAZ, M., "Web Kayıt Dosyalarından İlginç Örüntülerin Keşfedilmesi", Fırat Üniv. Fen ve Müh. Bil. Dergisi, 19 (4), 493-503, 2007.

[64] PHP Manual, mktime function, "http://www.php.net", Nisan 2008.

[65] CATLEDGE, L.D., PITKOW, J.E., "Characterizing browsing strategies in the world wide web", Computer Networks and ISDN Systems, 27: 1065–1073, 1995.

[66] Sakarya Üniversitesi Kampüs Otomasyonu Web Bilgi Sistemi, "http://www.cawis.sakarya.edu.tr/" Mayıs 2008.

[67] CANAY, Ö., "Sakarya Üniversitesi AdaMYO İnternet Programcılığı II Ders Notları", 2002.

ÖZGEÇMİŞ

Halil ARSLAN, 1982 yılında Sivas'ta doğdu. İlk ve orta öğrenimini Sivas'ta tamamladıktan sonra 2006 yılında Sakarya Üniversitesi Teknik Eğitim Fakültesi Bilgisayar Sistemleri Öğretmenliği'nden mezun oldu. Halen Sakarya Üniversitesi Bilgi İşlem Dairesi Başkanlığı'nda web yazılım uzmanı olarak görev yapmaktadır.