T.C.
BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
YÜKSEK LİSANS PROGRAMI

VERİ MADENCİLİĞİ ALGORİTMALARI KULLANILARAK WEB GÜNLÜK ERİŞİMLERİNİN ANALİZİ

İsmail HABERAL

Yüksek Lisans Tezi
Ankara – 2007

VERİ MADENCİLİĞİ ALGORİTMALARI KULLANILARAK WEB GÜNLÜK ERİŞİMLERİNİN ANALİZİ

ANALYSIS OF WEB LOG USING DATA MINING ALGORITHMS

İsmail HABERAL

Başkent Üniversitesi
Lisansüstü Eğitim Öğretim ve Sınav Yönetmeliğinin
BİLGİSAYAR Mühendisliği Anabilim Dalı İçin Öngördüğü
YÜKSEK LİSANS TEZİ
olarak hazırlanmıştır.

2007

Fen Bilimleri Enstitüsü Müdürlüğü'ne,

Bu çalışma, jürimiz tarafından **BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI 'nda YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Başkan :….....
Prof. Dr. Berna DENGİZ

Üye :.…..... ….
Yard. Doç. Dr. Hasan OĞUL

Üye :….....
Yard. Doç. Dr. Ergün ERASLAN

ONAY

Bu tez ...../...../..... tarihinde Enstitü Yönetim Kurulunca belirlenen yukarıdaki jüri üyeleri tarafından kabul edilmiştir.

...../...../.....
Prof.Dr. Emin AKATA
FEN BİLİMLERİ ENSTİTÜSÜ MÜDÜRÜ

VERİ MADENCİLİĞİ ALGORİTMALARI KULLANILARAK WEB GÜNLÜK ERİŞİMLERİNİN ANALİZİ

İsmail HABERAL
Başkent Üniversitesi Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

ÖZ

Veri madenciliği, günümüz bilgi çağında en güncel teknolojilerden birisidir. Bilgisayar sistemlerinin her geçen ucuzlaması kullanımının yaygınlaşmasını sağlarken, güçlerinin artması da, daha büyük miktarlarda veri saklanabilmesini olanaklı kılmaktadır. Günümüzde küçülen dünyanın ortak ama büyük problemleri nedeniyle büyük miktarlarda verilerin işlenmesi gerekli olmaktadır. Bu yüzden, büyük miktardaki verileri isleyebilen teknikleri kullanabilmek, büyük önem kazanmaktadır. Veri madenciliği bu gibi durumlarda kullanılan, büyük miktardaki veri setlerinde saklı durumda bulunan desen ve eğilimleri ortaya çıkarmak ve bilgiye dönüştürmek işlemidir.

Web madenciliği ise veri madenciliği tekniklerinin world wide web verileri üzerinde uygulanmasıdır. Web madenciliği kullanıcıların web sitelerindeki davranışlarını inceler, kullanıcı isteklerini belirler, bu şekilde elde edilen veriye dayanarak web sitelerinin yeniden tasarım veya geliştirilmesi konusunda karar vermeyi sağlayan bilgiyi sunar. Hangi sayfalara daha sık girildiği, hangi sayfaların birlikte ziyaret edildiği gibi bilgiler sitenin yeniden düzenlenmesinde faydalı olacak bilgilerdir. Bu uygulamaların amacı, istatistiksel yöntemlerle kullanıcıların davranışlarını öğrenmek, böylece site içeriği ve tasarımını bu bilgiler ışığında yenilemektir.

Bu çalışmada Başkent Üniversitesi web sitesinin geliştirilmesi amacıyla kullanılabilecek gerekli bilgiler http://www.baskent.edu.tr günlük web erişimleri veri madenciliği teknikleri kullanılarak elde edilmiştir.

Anahtar Sözcükler: Veri Madenciliği,Web Madenciliği, Web Kullanım Madenciliği.

ANALYSIS OF WEB LOG USING DATA MINING ALGORITHMS

İsmail HABERAL
Başkent University Institute of Science
The Department Of Computer Engineering

ABSTRACT

Data mining is one of the mosr upto-adte Technologies of the information area. The decreasing cost of computer systems enables a wider use, besides the increasing capacity makes storing very big quantities of data possible. Today, common but significant problems of the world requires processing large quantities of data. Therfore, it is getting more and more important to be able to use techniques processing these quantities of data. Data mining is the explaration of the patterns and tendencies inherent in these huge sets of data and transforming them into information.

Web mining is the application of data mining techniques on the world wide web data. Web mining studies behaviors of users in the web sites, identifies users' demands and provides information which will enable web designer to re-design and develop the web sites. The frequency of visits to a certain web site, and the correlation of one site visit with that of another are useful data to be used in the re-development/design of a web site. The overall aim of these practices is to renovate the content and design of a web site in the light of the information obtained through the statistical techniques applied to user behaviour.

The information used in the development of the web site of Başkent University in this study has been obtained from http://www.baskent.edu.tr web log using data mining techniques.

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

İÇİNDEKİLER LİSTESİ

Sayfa

ŞEKİLLER LİSTESİ

Sayfa

ÇİZELGELER LİSTESİ

Sayfa

SİMGELER VE KISALTMALAR LİSTESİ

BK	Bilgi Keşfi
CLF	Common Log Format
DB	Database
ELF	Extended Common Log Format
OLAP	On-Line Analytical Processing
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
KDD	Knowledge Discovery in Database
MÖ	Makine Öğrenimi
URL	Uniform Resource Locator
VM	Veri Madenciliği
VT	Veri Tabanı
VTBK	Veri Tabanlarında Bilgi Keşfi
VTYS	Veri Tabanı Yönetim Sistemi
WUM	Web Utilization Miner
WWW	World Wide Web

GİRİŞ

Günümüzde Internet'in yaşantımızı adeta yeniden şekillendirmesi ile web sitelerinin etkin yönetimini konu alan araştırmaların sayısı hızla artmaktadır. Bu nedenle web sitesinin içerik ve teknik anlamdaki kalitesi, işletmenin sunduğu ürün ya da hizmet kadar işletmenin, ürünün veya hizmetin rekabetteki yerinin belirlenmesinde önemli bir parametre olabilmektedir.

Bunun yanı sıra oldukça yaygınlaşan elektronik ticaret ve online alışveriş mekanizmalarının artması, bu alanda birbirlerine rakip olan firmaların çalışmaları, veri madenciliğinin önemini ön plana çıkarmaktadır.

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 [16].

Varolan verilerden bilgiyi yani kullanılabilecek örüntüleri elde etmeye geniş çapta ihtiyaç duyulmuştur. Bu ihtiyacı gidermek için araştırma kurumları ve üniversiteler çalışmalarıyla yeni disiplinler ortaya çıkarmıştır. Veri madenciliği bu yeni disiplinlerden biridir. Veri madenciliğinin veri tabanları üzerine uygulanmasıyla Veri Tabanlarında Bilgi Keşfi (VTBK) ortaya çıkmıştır.

Veri Tabanlarında Bilgi Keşfi genelde çok büyük hacimli verileri ele almakta kullanılmaktadır. Verileri ya tam ya da yarı otomatik olarak analiz eden yeni sistemlerle, bu disiplin son zamanlardaki en güncel konulardan biri haline gelmiştir. Veri seçimi, veri temizleme, veri ön işleme, veri indirgeme, veri madenciliği algoritmasının uygulanması ve sonuçların değerlendirmesi VTBK'yi oluşturan basamaklardır. Kısaca büyük ölçekli veri tabanlarından anlamlı ve gizli örüntülerin

çıkarılması olarak anılan Veri Madenciliği (VM) VTBK içinde bir adım olarak kabul edilir [47].

Veri madenciliği, makine öğrenimi (MÖ), veri tabanı yönetim sistemi (VTYS), veri ambarlama (VA), koşut programlama (KP) ve istatistik gibi bir çok farklı disiplinin kullandığı teknikleri kullanmaktadır. Bundan dolayı VM çok disiplinli bir yaklaşımdır [48].

VM, MÖ ve istatistik birbirlerine yakın disiplinlerdir. Bu üç disiplinin ortak noktaları veri içindeki ilginç örüntüleri bulmayı amaçlamalarıdır. VM algoritmalarının çekirdeğini MÖ'de kullanılan algoritmalar oluşturur. Makine öğreniminde kullanılan sınıflama, kümeleme ve ilişkilendirme algoritmaları gibi birçok algoritma veri madenciliğinde kullanılmaktadır. MÖ ile VM arasında bu söylenen benzerliklerin bulunmasının yanı sıra aralarında çok büyük farklar da vardır [49]. Örneğin, MÖ küçük hacimli ve genelde deneysel verilerle uğraşırken, VM büyük hacimli gerçek dünya verileriyle uğraşır. MÖ'nün örneklem kümesi genelde 100-1000 arasındayken, VM uygulandığında milyonlarca veriden söz edilmektedir. VM ve MÖ arasındaki diğer bir fark da, MÖ'nün aksine VM'nin gürültülü, eksik, artık ve boş (NULL) değerleri işleyebilmesidir[48].

Büyük boyutlu yapısal veriyi saklama ve bu verilere etkin bir şekilde erişim sağlamakla yükümlü olan VTYS'lerde veri düzenlemesi, ilgili organizasyonun işletimsel veri ihtiyacı doğrultusunda gerçekleştirilir. Bu işlem her zaman bilgi keşfi (BK) perspektifi ile birebir örtüşmez. Bu açıdan VM algoritması uygulanmadan önce veri ön işleme basamakları gerçekleştirilir. VT'deki veriler üzerinde gerçekleştirilen bu basamaklar kısaca şöyledir: temizleme, boyut indirgeme, tür dönüşümleri, transfer, vb işlemlerdir[48].

Veritabanları veri organizasyonunda kullanılan araçlardan birisidir. Bu çalışmadaki kullanım hedeflerinden biri, farklı olarak tam yapısal olmayan ve hızla değişen veriyi çeşitli kullanıcı tiplerinin erişebileceği bir ortamda sunmaktır. Hedeflerden bir diğeri ise web madenciliği teknolojilerini bu veriye uygulayabilmektir.

Bu veritabanı ile sağlanabilecek faydaların bir kaç tanesi şöyle sıralanabilir;

i)kullanıcıların profilleri çıkarılabilir ve zaman içindeki değişimleri takip edilebilir,

ii)sitedeki beğenilen ya da beğenilmeyen köşeler tespit edilebilir,

iii)kullanıcıların gezinti şekli/hızı sitenin içerik, yapılandırma ve alt-yapı açısından performansı hakkında fikir verebilir.

Üniversite web sitesindeki sayfalara giriş isteklerinin bilinmesi, sitenin yeniden düzenlenmesi ve daha aktif hale getirilmesine yardımcı olacak bilgileri sağlayacaktır.
Hangi sayfalara daha sık girildiği (giriş sıklık bilgisi), hangi sayfaların birlikte (ardışık olarak) ziyaret edildiği gibi bilgiler sitenin yeniden düzenlenmesinde faydalı olacak bilgilerdir. Bu uygulamaların amacı, istatistiksel yöntemlerle kullanıcıların davranışlarını öğrenmek böylece site içeriği ve tasarımını bu bilgiler ışığında gözden geçirerek yenilemeler yapmaktır. Böylece siteden yararlanan ziyaretçilere daha kaliteli ve amaçlarını sağlayabilecek bilgi istekleri doğrultusunda sunulabilecektir.

Bu tezde giriş bilgilerinin ardından ikinci bölümde veri madenciliği tanımı, üçüncü bölümde web madenciliği, aşamaları ve teknikleri anlatılmaktadır. Dördüncü bölümde web kullanım madenciliği, web verileri üzerinde madencilik, beşinci bölümde ise üniversite web sitesi üzerinde yapılan web madenciliği uygulaması yer almaktadır.

2.VERİ MADENCİLİĞİ

Gelişen ve değişen çevre koşulları, sınırların kalkması ile küreselleşen dünya, farklı pazarlama ve ar-ge(araştırma geliştirme) yöntemleri “veri”nin değil “bilgi”nin önemini her geçen gün daha da artacak şekilde ortaya koymaktadır. İnternetin yaygınlaşması ve kolaylaşması ar-ge ekiplerinin “bilgi”ye erişmelerini zorlaştırmaktadır. İnternette arama motorları kullanılarak yapılan araştırmalar çoğu zaman istenilenden farklı bir şekilde sonuçlanmaktadır. Tıbbi bir araştırma sonucunda elde edilen verilerin yorumlanıp analiz edilmesiyle bilgiye ulaşılabilmektedir. Büyük bir perakendecinin, fatura bilgilerinden müşteri eğilimlerini belirleyip ona göre pazarlama taktikleri üretebilmesi, rakiplerinin önüne geçmesini sağlayacaktır. Verilen örneklere dikkat edilirse, “veri”nin “bilgi”ye dönüşme işleminin vurgulandığı görülecektir. Bilginin matematiksel ve istatistiksel yöntemler ile analiz edilmesi ve çıkan sonuçların bir uzman gözüyle yorumlanmasıyla geçmiş verilerden gelecek tahminleri yapma işlemi veri madenciliğidir.

Veri madenciliği, eldeki verilerden üstü kapalı, çok net olmayan, önceden bilinmeyen ancak potansiyel olarak kullanışlı bilginin çıkarılmasıdır. Bu da; kümeleme, veri özetleme, değişikliklerin analizi, sapmaların tespiti gibi belirli sayıda teknik yaklaşımları içerir[28].

Başka bir deyişle, veri madenciliği, verilerin içerisindeki desenlerin, ilişkilerin, değişimlerin, düzensizliklerin, kuralların ve istatistiksel olarak önemli olan yapıların yarı otomatik olarak keşfedilmesidir.

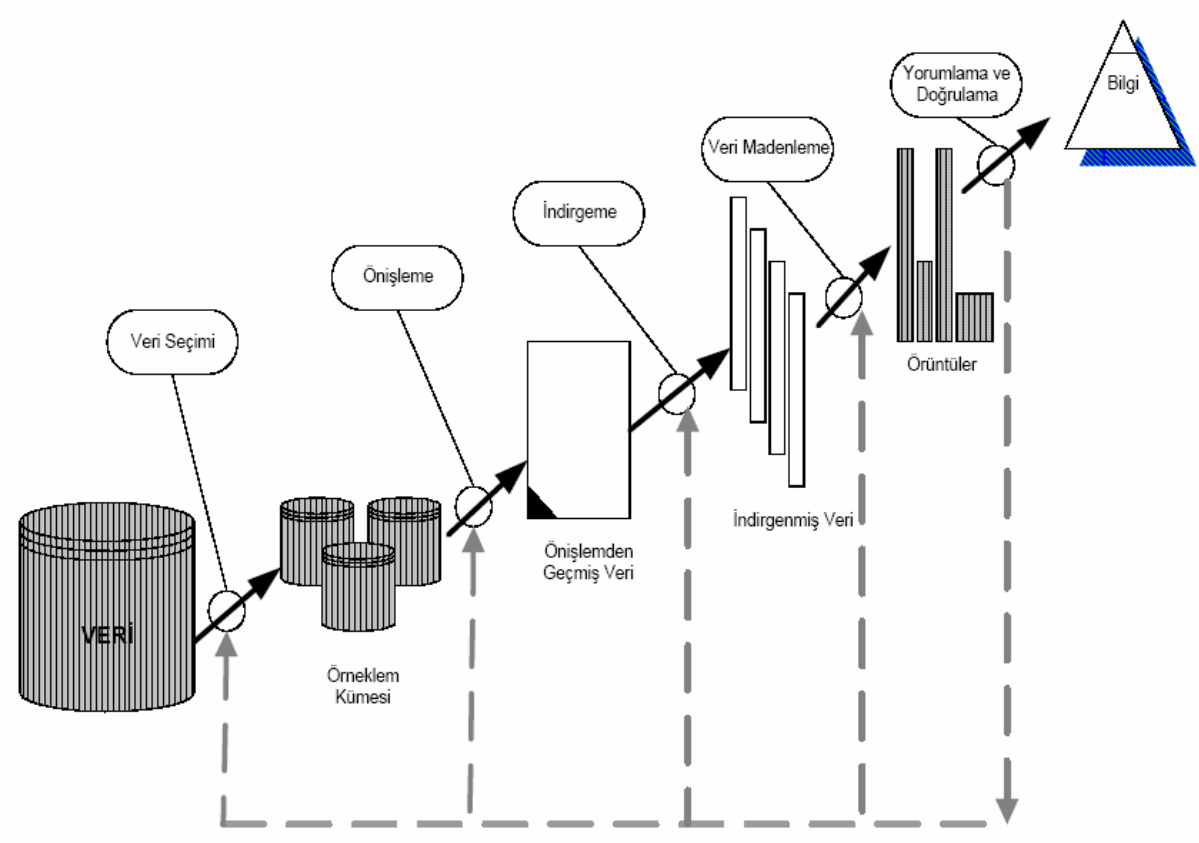
Temel olarak veri madenciliği, veri setleri arasındaki desenlerin ya da düzenin, verinin analizi ve yazılım tekniklerinin kullanılması ile ilgilidir. Veriler arasındaki ilişkiyi, kuralları ve özellikleri belirlemekten kullanılan teknikler sorumludur. Amaç, daha önceden fark edilmemiş veri desenlerini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

Veri madenciliğini istatistiksel bir yöntemler serisi olarak görmek mümkündür. Ancak veri madenciliği, geleneksel istatistikten birkaç yönde farklılık gösterir. Veri madenciliğinde amaç, kolaylıkla mantıksal kurallara ya da görsel sunumlara

çevrilebilecek nitel modellerin çıkarılmasıdır. Bu bağlamda, veri madenciliği insan merkezlidir ve bazen insan – bilgisayar arayüzü birleştirilir.

Veri madenciliği, istatistik, arama teknikleri, makine bilgisi, veri tabanları ve yüksek performanslı işlem gibi temelleri de içerir.

Veri madenciliği konusunda bahsi geçen “geniş veri”deki geniş kelimesi, tek bir iş istasyonunun belleğine sığamayacak kadar büyük veri kümelerini ifade etmektedir. Yüksek hacimli veri ise, tek bir iş istasyonundaki ya da bir grup iş istasyonundaki disklere sığamayacak kadar fazla veri anlamındadır. Dağıtık veri ise, farklı coğrafi konumlarda bulunan verileri anlatır. [15]
Veri tabanlarında bilgi keşfi sürecinde (VTBK) yer alan adımlar Şekil 1’de gösterilmiştir [12]:



Şekil 1: VTBK Sürecinde Yer Alan Adımlar

Veri Seçimi (Data Selection): Bu adım birkaç veri kümesini birleştirerek, sorguya uygun örneklem kümesini elde etmeyi gerektirir.

Veri Temizleme ve Önişleme (Data Cleaning & Preprocessing): Seçilen örneklemde yer alan hatalı tutanakların çıkarıldığı ve eksik nitelik değerlerinin değiştirildiği aşamadır ve keşfedilen bilginin kalitesini artırır.

Veri İndirgeme (Data Reduction): Seçilen örneklemden ilgisiz niteliklerin atıldığı ve tekrarlı tutanakların ayıklandığı adımdır. Bu aşama ile seçilen veri madenciliği sorgusunun çalışma zamanını iyileştirir.

Veri Madenciliği (Data Mining): Verilen bir veri madenciliği sorgusunun (sınıflama, güdümsüz öbekleme, eşleştirme, vb.) işletilmesidir.

Değerlendirme (Evaluation): Keşfedilen bilginin geçerlilik, yenilik, yararlılık ve basitlik kıstaslarına göre değerlendirilmesi aşamasıdır[12].

2.1. Veri Madenciliğine Genel Bakış (Tarihçe)

VM yaklaşımı ortaya çıkmadan önce, büyük veri tabanlarından faydalı örüntüler elde etmek için, çevrim-dışı veri üzerinde çalışan istatistiksel paketler kullanılırdı. İstatistiksel yaklaşımların kullanımında bu paketlerin dezavantajları ortaya çıkmaktaydı. Bu dezavantajlardan en önemlisi; istenen verilerin toplanmasından ve amacın belirlenerek istatistiksel yaklaşımların uygulanmasından sonra bir uzman tarafından değerlendirilmesi gerekliliğidir. Başka bir dezavantajı ise her farklı ihtiyaç için bu işlemlerin tekrarlanmasıdır. Bu sorun VTBK'de kısmen aşılmıştır. VTBK [50] çok büyük hacimli verilerden anlamlı ilişkileri otomatik keşfeder.

2.2 Veri Madenciliği Teknikleri

İstatistiksel Yöntemler: Veri madenciliği çalışması esas olarak bir istatistik uygulamasıdır. Verilen bir örnek kümesine bir kestirici oturtmayı amaçlar[15]. İstatistik literatüründe son elli yılda bu amaç için değişik teknikler önerilmiştir. Bu teknikler istatistik literatüründe çok değişkenli analiz (multivariate analysis) başlığı

altında toplanır ve genelde verinin parametrik bir modelden (çoğunlukla çok değişkenli bir Gauss dağılımından) geldiğini varsayar. Bu varsayım altında sınıflandırma (classification; discriminant analysis), regresyon, öbekleme (clustering), boyut azaltma (dimensionality reduction), hipotez testi, varyans analizi, bağıntı (association; dependency) kurma gibi teknikler uzun yıllardır kullanılmaktadır [12].

Bellek Tabanlı Yöntemler: Bellek tabanlı veya örnek tabanlı bu yöntemler (memory-based, instance-based methods; case-based reasoning) istatistikte 1950'li yıllarda önerilmiş olmasına rağmen o yıllarda gerektirdiği hesaplama büyüklüğü ve bellek yetersizliği yüzünden kullanılamamış ama günümüzde bilgisayar teknolojilerindeki gelişmeyle ve kapasitelerinin artmasıyla, özellikle de çok işlemcili sistemlerin yaygınlaşmasıyla, kullanılabilir olmuşlardır. Bu yöntemlere en iyi örnek en yakın k komşu algoritmasıdır (k-nearest neighbor). [12]

Yapay Sinir Ağları: 1980'lerden sonra yaygınlaşan yapay sinir ağlarında (artificial neural networks) amaç fonksiyon birbirine bağlı basit işlemci ünitelerinden oluşan bir ağ üzerine dağıtılmıştır. Yapay sinir ağlarında kullanılan öğrenme algoritmaları veriden üniteler arasındaki bağlantı ağırlıklarını hesaplar. YSA istatistiksel yöntemler gibi veri hakkında parametrik bir model varsaymaz yani uygulama alanı daha geniştir, ve bellek tabanlı yöntemler kadar yüksek işlem ve bellek gerektirmez.[12]

Karar Ağaçları: İstatistiksel yöntemlerde veya yapay sinir ağlarında veriden bir fonksiyon öğrenildikten sonra bu fonksiyonun insanlar tarafından anlaşılabilecek bir kural olarak yorumlanması zordur. Karar ağaçları ise veri oluşturulduktan sonra ağaç kökten yaprağa doğru inilerek kurallar (IF-THEN rules) yazılabilir. Bu şekilde kural çıkarma (rule extraction), veri madenciliği çalışmasının sonucunun geçerlenmesini sağlar. Bu kurallar uygulama konusunda uzman bir kişiye gösterilerek sonucun anlamlı olup olmadığı denetlenebilir. Sonradan başka bir teknik kullanılacak bile olsa karar ağacı ile önce bir kısa çalışma yapmak, önemli değişkenler ve yaklaşık kuralların elde edilmesine yardımcı olur. [12]

2.3 Veri Madenciliğinde Karşılaşılan Problemler

Küçük veri kümelerinde hızlı ve doğru bir biçimde çalışan bir sistem, çok büyük veri tabanlarına uygulandığında tamamen farklı davranabilir. Bir VM sistemi tutarlı veri üzerinde mükemmel çalışırken, aynı veriye gürültü eklendiğinde kayda değer bir biçimde kötüleşebilir. İzleyen kesimde günümüz VM sistemlerinin karşı karşıya olduğu problemler incelenecektir.[12]

i) Veri Tabanı Boyutu

Veri tabanı boyutları inanılmaz bir hızla artmaktadır. Pek çok makine öğrenimi algoritması birkaç yüz tutanaklık oldukça küçük örneklemleri ele alabilecek biçimde geliştirilmiştir. Aynı algoritmaların yüzbinlerce kat büyük örneklemlerde kullanılabilmesi için azami dikkat gerekmektedir. Örneklemin büyük olması, örüntülerin gerçekten var olduğunu göstermesi açısından bir avantajdır ancak böyle bir örneklemden elde edilebilecek olası örüntü sayısı da çok büyüktür. Bu yüzden VM sistemlerinin karşı karşıya olduğu en önemli sorunlardan biri veri tabanı boyutunun çok büyük olmasıdır. Dolayısıyla VM yöntemleri ya sezgisel bir yaklaşımla arama uzayını taramalıdır, ya da örneklemi yatay/dikey olarak indirgemelidir[12].

Yatay indirgeme çeşitli biçimlerde gerçekleştirilebilir. İlkinde, belirli bir niteliğin alan değerleri önceden sıradüzensel (hierarchy) olarak sınıflandırılır ki, buna genelleştirme işlemi de denilmektedir. Sonrasında ise ilgili niteliğin değerleri önceden belirlenmiş genelleme sıradüzeninde aşağıdan yukarıya doğru seviye seviye günlenir (yani üst nitelik değeri ile değiştirilir) ve tekrarlı (mükerrer) çoklular çıkarılır.İkincisinde, oldukça sağlam (robust) olan örnekleme kuramı kullanılarak çok büyük oylumlu veri öyle bir boyuta indirgenir ki, hem kaynak veri belirli bir güven aralığında temsil edilebilir hem de indirgenen veri kümesinin oylumu makine öğrenimi teknikleri ile işlenmeye uygun/olurlu bir hale getirilebilir.

Sonuncusunda ise sürekli değerlerden oluşan bir alana sahip nitelik üzerine kesikleştirme tekniği uygulanır. Sürekli değerlerin belirli aralık değerlerine dönüştürülmesi ile ortaya çıkabilecek tekrarlı çoklular tekil hale getirilerek yatay indirgeme sağlanabilir. Aslında bu kesikleştirme tekniği, sürekli sayısal değerler için geçerli olmayan makine öğrenim algoritmaları için bir ön koşul veya ön işlemedir ki, bu konu ayrı bir alt başlık olarak verilecektir. Dikey indirgeme, artık niteliklerin indirgenmesi işlemidir ve "artık işleme" alt başlığında tartışılacaktır [12].

ii) Gürültülü Veri

Büyük veri tabanlarında pek çok niteliğin değeri yanlış olabilir. Bu hata, veri girişi sırasında yapılan insan hataları veya girilen değerin yanlış ölçülmesinden kaynaklanır. Veri girişi veya veri toplanması sırasında oluşan sistem dışı hatalara gürültü adı verilir. Günümüzde kullanılan ticari ilişkisel veri tabanları, veri girişi sırasında oluşan hataları otomatik biçimde gidermek konusunda yeterliği sağlamamaktadır. Hatalı veri gerçek dünya veri tabanlarında ciddi problem oluşturabilir. Bu durum, bir VM yönteminin kullanılan veri kümesinde bulunan gürültülü verilere karşı daha az duyarlı olmasını gerektirir. Gürültülü verinin yol açtığı problemler tümevarımsal karar ağaçlarında uygulanan metodlar bağlamında kapsamlı bir biçimde araştırılmıştır. Eğer veri kümesi gürültülü ise sistem bozuk veriyi tanımalı ve ihmal etmelidir. Quinlan (1986b), gürültünün sınıflama üzerindeki etkisini araştırmak için bir dizi deney yapmıştır. Deneysel sonuçlar, etiketli öğrenmede makine öğrenim tekniklerinin etiket niteliği üzerindeki gürültülere, diğer koşul niteliklerinde sunulan gürültülere kıyasla, daha duyarlı olduklarını göstermiştir. Buna karşın eğitim kümesindeki nesnelerin nitelikleri üzerindeki en çok %10'luk gürültü miktarı ayıklanabilmektedir. Chan ve Wong (1991), gürültünün etkisini analiz etmek için istatiksel yöntemler kullanmışlardır[12].

iii) 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 hiçbir değere eşit olmayan değerdir. Bir çokluda 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.

Lee (1992), boş değeri (1) bilinmeyen, (2) uygulanamaz, (3) bilinmeyen veya uygulanamaz olacak biçimde üçe ayıran bir yaklaşımı ilişkisel veri tabanlarını genişletmek için öne sürmüştür. Mevcut boş değer taşıyan veri için herhangi bir çözüm sunmayan bu yaklaşımın dışında bu konuda sadece bilinmeyen değer üzerinde çalışmalar yapılmıştır. Boş değerli nitelikler veri kümesinde bulunuyorsa, ya bu çoklular tamamıyla ihmal edilmeli ya da bu çoklularda niteliğe olası en yakın değer atanmalıdır [12].

iv) Eksik Veri

Evrendeki her nesnenin ayrıntılı bir biçimde tanımlandığı ve bu nesnelerin alabileceği değerler kümesinin belirli olduğu varsayılsın. Verilen bir bağlamda her bir nesnenin tanımı kesin ve yeterli olsa idi sınıflama işlemi basitçe nesnelerin alt kümelerinden faydalanılarak yapılırdı. Bununla birlikte, veriler kurum ihtiyaçları göz önünde bulundurularak düzenlenip toplandığından, mevcut veri bilgi keşfi açısından uygun olmayabilir . Örneğin hastalığın tanısını koymak için kurallar sadece çok yaşlı insanların belirtilerinin bulunduğu bir veri kümesi kullanılarak üretilseydi, bu kurallara dayanarak bir çocuğa tanı koymak pek doğru olmazdı. Bu gibi koşullarda bilgi keşfi modeli belirli bir güvenlik derecesinde tahmini kararlar alabilmelidir [12].

v) Artık Veri

Verilen veri kümesi, eldeki probleme uygun olmayan veya artık nitelikler içerebilir.Bu durum pek çok işlem sırasında karşımıza çıkabilir. Örneğin, eldeki problem ile ilgili veriyi elde etmek için iki ilişkiyi ortak nitelikler üzerinden birleştirirsek, sonuç ilişkide kullanıcının farkında olmadığı artık nitelikler bulunur.

Artık nitelikleri elemek için geliştirilmiş algoritmalar özellik seçimi olarak adlandırılır [12].

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 yalnızca arama uzayını küçültmekle kalmayıp, sınıflama işleminin kalitesini de artırır.

vi) Dinamik Veri

Kurumsal çevrim içi veri tabanları dinamiktir, yani içeriği sürekli olarak değişir. Bu durum, bilgi keşfi metodları için önemli sakıncalar doğurmaktadır. İlk olarak sadece okuma yapan ve uzun süre çalışan bilgi keşfi metodu, bir veri tabanı uygulaması olarak mevcut veri tabanı ile birlikte çalıştırıldığında mevcut uygulamanın da performansı ciddi ölçüde düşer. Diğer bir sakınca ise, veri tabanında bulunan verilerin kalıcı olduğu varsayılıp, çevrim dışı veri üzerinde bilgi keşif metodu çalıştırıldığında, değişen verinin elde edilen örüntülere yansıması gerekmektedir. Bu işlem, bilgi keşfi metodunun ürettiği örüntüleri zaman içinde değişen veriye göre sadece ilgili örüntüleri yığmalı olarak günleme yeteneğine sahip olmasını gerektirir . Aktif veri tabanları tetikleme mekanizmalarına sahiptir ve bu özellik bilgi keşif metodları ile birlikte kullanılabilir [12].

vii) Farklı Tipteki Veriler

Gerçek hayattaki uygulamalar makine öğreniminde olduğu gibi yalnızca sembolik veya kategorik veri türleri değil, fakat aynı zamanda tamsayı, kesirli sayılar, çoklu ortam verisi, coğrafi bilgi içeren veri gibi farklı tipteki veriler üzerinde işlem yapılmasını gerektirir. Kullanılan verinin saklandığı ortam, düz bir kütük veya ilişkisel veri tabanında yer alan tablolar olacağı gibi, nesneye yönelik veri tabanları, çoklu ortam veri tabanları, coğrafik veri tabanları vb. olabilir. Saklandığı ortama göre veri, basit tipte olabileceği gibi karmaşık veri tipleri (çoklu ortam verisi, zaman içeren veri, yardımlı metin, coğrafi, vb.) de olabilir. Bununla birlikte veri tipi çeşitliliğinin fazla olması bir VM algoritmasının tüm veri tiplerini ele alabilmesini

olanaksızlaştırmaktadır. Bu yüzden veri tipine özgü adanmış VM algoritmaları geliştirilmektedir [12].

2.4 Veri Madenciliği Algoritmaları

Veri madenciliği (VM) süreci sonunda elde edilen desenler kurallar biçiminde ifade edilir. Eldeedilen kurallar,
(1) koşul yan tümcesi ile sonuç arasındaki eşleştirme derecesini gösterir (if <koşul tümcesi>, then <sonuç>, derece (0..1)),
(2) veriyi önceden tanımlanmış sınıflara bölümler (partition); veya
(3) veriyi bir takım kriterlere göre sonlu sayıda kümeye ayırır. Bu kurallar veri üzerinde belirli bir tekniğin (algoritmanın) sonlu sayıda yinelenmesiyle elde edilir. Elde edilen bilginin kalitesi veri analizi için kullanılan algoritmaya büyük ölçüde bağlıdır.

VM algoritmaları, doğrulamaya dayalı algoritmalar ve keşfe dayalı algoritmalar olarak iki grupta toplanabilir . Doğrulamaya dayalı VM algoritmasında kullanıcı bir hipotez öne sürer ve sistem bu hipotezi ispatlamaya çalışır. Doğrulamaya dayalı VM algoritmalarının en yaygın olarak kullanıldığı yerler, istatistiksel ve çok boyutlu analizlerdir. Öte yandan keşfe dayalı algoritmalar otomatik olarak yeni bilgi çıkarırlar. İzleyen kesimde VM sistemlerinde kullanılan algoritmalardan önemli olanları incelenecektir.

i) 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 veri tabanındaki varlığıyla ilgili bir tahmindir . Bu tip bir analiz özellikle keşfedilmiş bilginin genişletilmesi veya damıtılması (refine) işlemleri sırasında yararlıdır.

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

hipotez seçilen veri tabanında doğruluk ve destek kıstasları baz alınarak sistem tarafından sınanır[12].

ii) Sınıflandırma ve Regresyon

Sınıflama ve regresyon, önemli veri sınıflarını ortaya koyan veya gelecek veri egilimlerini tahmin eden modelleri kurabilen iki veri analiz yöntemidir [3]. Sınıflama kategorik değerleri tahmin ederken, regresyon süreklilik gösteren degerlerin tahmin edilmesinde kullanılır. Örnegin, bir sınıflama modeli banka kredi uygulamalarının güvenli veya riskli olmalarını kategorize etmek amacıyla kurulurken, regresyon modeli geliri ve meslegi verilen potansiyel müşterilerin bilgisayar ürünleri alırken yapacakları harcamaları tahmin etmek için kurulabilir.

Sınıflama ve regresyon modellerinde kullanılan baslıca teknikler sunlardır [2]:

1 - Karar Agaçları (Decision Trees)
2- Yapay Sinir Agları (Artificial Neural Networks)
3- Genetik Algoritmalar (Genetic Algorithms)
4- K-En Yakın Komsu (K-Nearest Neighbor)
5- Bellek Temelli Nedenleme (Memory Based Reasoning)
6- Naive-Bayes

Karar ağaçları, veri madencilişinde kuruluslarının ucuz olması, yorumlanmalarının kolay olması, veri tabanı sistemleri ile kolayca entegre edilebilmeleri ve güvenilirliklerinin iyi olması nedenleri ile sınıflama modelleri içerisinde en yaygın kullanıma sahip tekniktir.

Karar ağacı, adından da anlaşsılacağı gibi bir agaç görünümünde, tahmin edici bir tekniktir [23]. Agaç yapısı ile, kolay anlasılabilen kurallar yaratabilen, bilgi teknolojileri islemleri ile kolay entegre olabilen en popüler sınıflama teknigidir.

Karar agacı karar dügümleri, dallar ve yapraklardan olusur [22]. Karar dügümü, gerçeklestirilecek testi belirtir. Bu testin sonucu agacın veri kaybetmeden dallara ayrılmasına neden olur. Her dügümde test ve dallara ayrılma işlemleri ardısık olarak gerçeklesir ve bu ayrılma islemi üst seviyedeki ayrımlara bağımlıdır. Agacın

her bir dalı sınıflama islemini tamamlamaya adaydır. Eger bir dalın ucunda sınıflama islemi gerçeklesemiyorsa, o daim sonucunda bir karar dügümü olusur. Ancak daim sonunda belirli bir sınıf olusuyorsa, o dalın sonunda yaprak vardır. Bu yaprak, veri üzerinde belirlenmek istenen sınıflardan biridir. Karar agacı islemi kök dügümünden baslar ve yukarıdan asagıya dogru yapraga ulasana dek ardışık dügümleri takip ederek gerçeklesir.

Karar ağacı tekniğini kullanarak verinin sınıflanması iki basamaklı bir islemdir [22]. İlk basamak öğrenme basamağıdır. Öğrenme basamağında önceden bilinen bir egitim verisi, model olusturmak amacıyla sınıflama algoritması tarafından analiz edilir. Ögrenilen model, sınıflama kuralları veya karar agacı olarak gösterilir. İkinci basamak ise sınıflama basamagıdır. Sınıflama basamagında test verisi, sınıflama kurallarının veya karar agacının dogrulugunu belirlemek amacıyla kullanılır. Eger dogruluk kabul edilebilir oranda ise, kurallar yeni verilerin sınıflanması amacıyla kullanılır.

Test verisine uygulanan bir modelin dogrulugu, yaptıgı dogru sınıflamanın test verisindeki tüm sınıflara oranıdır. Her test örneginde bilinen sınıf, model tarafından tahmin edilen sınıf ile karsılastırılır. Eger modelin dogrulugu kabul edilebilir bir deger ise model, sınıfı bilinmeyen yeni verileri sınıflama amacıyla kullanılabilir. Örnegin, bir egitim verisi incelenerek kredi duruma sınıfını tahmin edecek bir model olusturuluyor. Bu modeli olusturan bir sınıflama kuralı;

IF yas = "41...50" AND gelir = yüksek THEN kredidurumu = mükemmel

seklindedir. Bu kural geregince yası "41...50" kategorisinde olan (yası 41 ile 50 arasında olan) ve gelir düzeyi yüksek bir kisinin kredi durumunun mükemmel oldugu görülür.

Olusturulan bu modelin dogrulugu, bir test verisi aracılıgı ile onaylandıktan sonra model, sınıfı belli olmayan yeni bir veriye uygulanabilir ve sınıflama kuralı geregi yeni verinin sınıfı "mükemmel" olarak belirlenebilir.

Tekrarlamak gerekirse bir karar agacı, bir alandaki testi belirten karar dügümlerinden, testteki degerleri belirten dallardan ve sınıfı belirten yapraklardan olusan akıs diyagramı seklindeki agaç yapısıdır. Agaç yapısındaki en üstteki dügüm kök dügümüdür.

Belirli bir sınıfın muhtemel üyesi olacak elemanların belirlenmesi, çesitli durumların yüksek, orta, düsük risk grupları gibi çesitli kategorilere ayrılması, gelecekteki olayların tahmin edilebilmesi için kurallar olusturulması, sadece belirli alt gruplara özgü olan iliskilerin tanımlanması, kategorilerin birlestirilmesi gibi alanlarda karar agaçları kullanılmaktadır [25].

Karar agaçları, hangi demografik grupların mektupla yapılan pazarlama uygulamalarında yüksek cevaplama oranına sahip oldugunun belirlenmesi (Direct Mail), bireylerin kredi geçmislerini kullanarak kredi kararlarının verilmesi (Credit Scoring), geçmiste isletmeye en faydalı olan bireylerin özelliklerini kullanarak ise alma süreçlerinin belirlenmesi, tıbbi gözlem verilerinden yararlanarak en etkin kararların verilmesi, hangi degiskenlerin satısları etkilediginin belirlenmesi, üretim verilerini inceleyerek ürün hatalarına yol açan degiskenlerin belirlenmesi gibi uygulamalarda kullanılmaktadır [25].

iii) Kümeleme

Kümeleme, veriyi sınıflara veya kümelere ayırma islemidir[20] . Aynı kümedeki elemanlar birbirleriyle benzerlik gösterirlerken, baska kümelerin elemanlarından farklıdırlar. Kümeleme veri madenciligi, istatistik, biyoloji ve makine ögrenimi gibi pek çok alanda kullanılır. Kümeleme modelinde, sınıflama modelinde olan veri sınıfları yoktur. Verilerin herhangi bir sınıfı bulunmamaktadır. Sınıflama modelinde, verilerin sınıfları bilinmekte ve yeni bir veri geldiginde bu verinin hangi sınıftan olabilecegi tahmin edilmektedir. Oysa kümeleme modelinde, sınıfları bulunmayan veriler gruplar halinde kümelere ayrılırlar. Bazı uygulamalarda kümeleme modeli, sınıflama modelinin bir önislemi gibi görev alabilmektedir .

Marketlerde farklı müsteri gruplarının kesfedilmesi ve bu grupların alısveris örüntülerinin ortaya konması, biyolojide bitki ve hayvan sınıflandırmaları ve

islevlerine göre benzer genlerin sınıflandırılması, sehir planlanmasında evlerin tiplerine, degerlerine ve cografik konumlarına göre gruplara aynlması gibi uygulamalar tipik kümeleme uygulamalarıdır. Kümeleme aynı zamanda Web üzerinde bilgi kesfi için dokümanların sınıflanması amacıyla da kullanılabilir [21].

Veri kümeleme güçlü bir gelisme göstermektedir. Veri tabanlarında toplanan veri miktarının artmasıyla orantılı olarak, kümeleme analizi son zamanlarda veri madenciligi arastırmalarında aktif bir konu haline gelmistir.

Literatürde pek çok kümeleme algoritması bulunmaktadır. Kullanılacak olan kümeleme algoritmasının seçimi, veri tipine ve amaca baglıdır. Genel olarak baslıca kümeleme yöntemleri su sekilde sınıflandırılabilir [22]:

1 - Bölme yöntemleri (Partitioning methods)
2- Hiyerarsik yöntemler (Hierarchical methods)
3- Yogunluk tabanlı yöntemler (Density-based methods)
4- Izgara tabanlı yöntemler (Grid-based methods)
5- Model tabanlı yöntemler (Model-based methods)

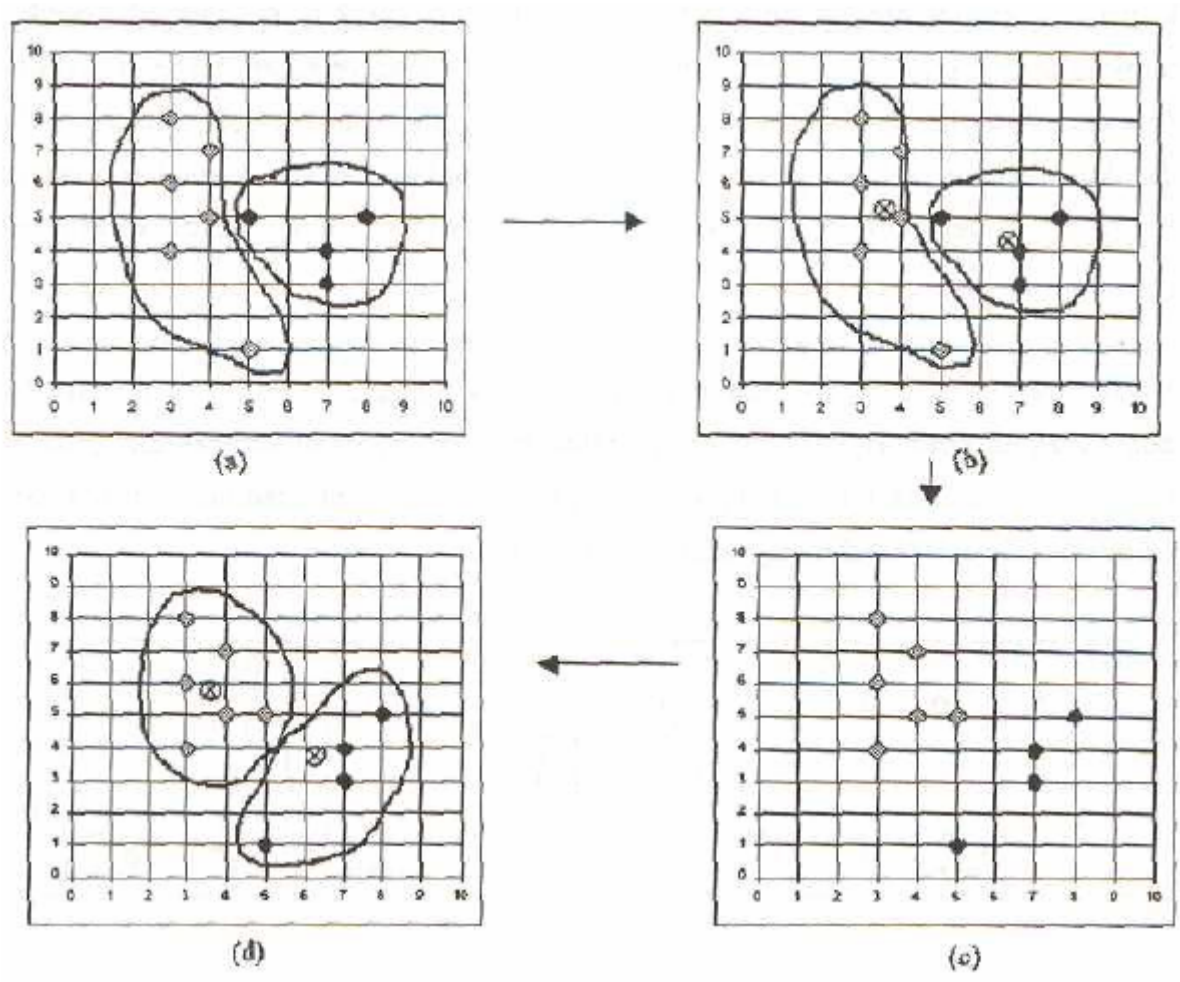
Bölme yöntemlerinde, n veri tabanındaki nesne sayısı ve k olusturulacak küme sayısı olarak kabul edilir. Bölme algoritması n adet nesneyi, k adet kümeye böler (k _ n). Kümeler tarafsız bölme kriteri olarak nitelendirilen bir kritere uygun olusturuldugu için aynı kümedeki nesneler birbirlerine benzerken, farklı kümedeki nesnelerden farklıdırlar [22].

En iyi bilinen ve en çok kullanılan bölme yöntemleri k-means yöntemi, k-medoids yöntemi ve bunların varyasyonlarıdır [23].

k-means yöntemi, ilk önce n adet nesneden rasgele k adet nesne seçer ve bu nesnelerin her biri, bir kümenin merkezini veya orta noktasını temsil eder. Geriye kalan nesnelerden her biri kendisine en yakın olan küme merkezine göre kümelere dagılırlar. Yani bir nesne hangi kümenin merkezine daha yakın ise o kümeye yerlesir.

Ardından her küme için ortalama hesaplanır ve hesaplanan bu deger o kümenin yeni merkezi olur. Bu islem tüm nesneler kümelere yerlesinceye kadar devam eder [22].

Bir nesne grubunun, Sekil 2.2'de görüldügü gibi uzayda konumlanmıs oldugu varsayılsın. Kullanıcının bu nesneleri iki kümeye ayırmak istedigi varsayılırsa, k=2 olur [22]. Sekil 2.1 (ayda ilk önce rasgele iki nesne, iki kümenin merkezi olarak seçilmis ve diger nesneler de bu merkezlere olan yakınlıklarına göre iki kümeye ayrılmıslardır. Bu ayrıma göre her iki kümenin nesnelerinin yeni ortalaması alınmıs ve bu deger kümelerin yeni merkezleri olmustur. Bu yeni merkezler Sekil 2.2(b)'de üstünde çarpı isareti bulunan noktalarla gösterilmektedir. Bu yeni çarpı isaretli merkezlere göre, her iki kümede de birer nesne diger kümenin merkezine daha yakın duruma gelmislerdir. Bu durum Sekil 2.2(c)'de görülmektedir. (5,1) koordinatındaki nesne ile (5,5) koordinatındaki nesne küme degistirmislerdir. Her iki kümedeki bu yeni katılımlar ile kümelerdeki nesnelerin ortalama degerleri ve dolayısıyla merkezleri degismistir [22]. Yeni hesaplanan merkezler Sekil 2.2(d)'de üstünde çarpı isareti bulunan noktalarla gösterilmektedir. Artık açıkta bir nesne kalmadıgı ve her nesne içinde bulundugu kümenin merkezine en yakm durumda bulundugu için k-means yöntemi ile kümelere bölünme islemi Sekil 2.2(d)'de görüldügü gibi sonlanmıstır [22].



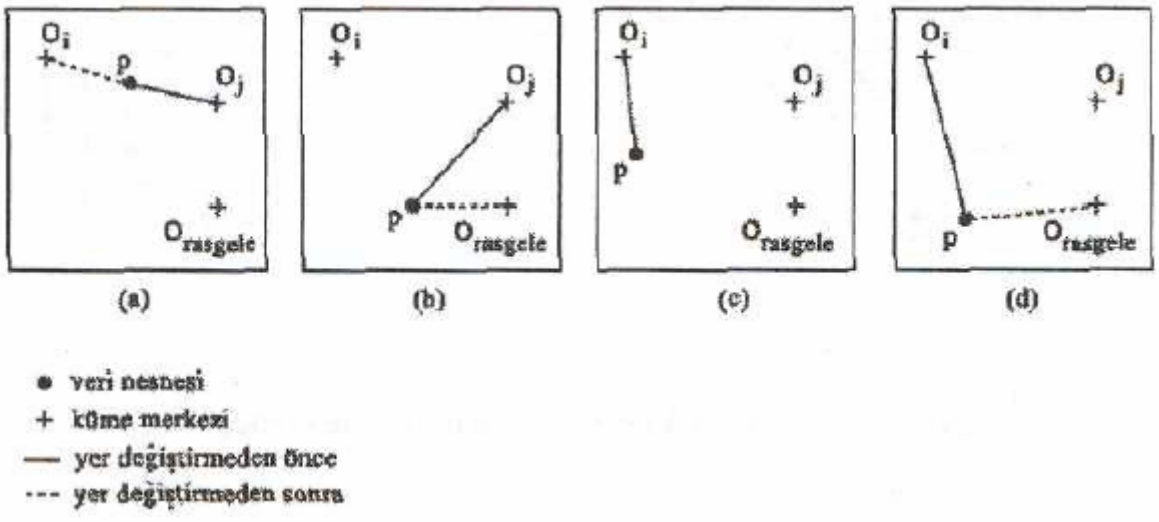
Şekil 2.2 k-means Yöntemiyle Kümeleme Örnegi-1

k-means yöntemi, sadece kümenin ortalaması tanımlanabildigi durumlarda kullanılabilir[9], Kullanıcıların k degerini, yani olusacak küme sayısını belirtme gerekliligi bir dezavantaj olarak görülebilir. Esas önemli olan dezavantaj ise dısarıda kalanlar (outliers) olarak adlandırılan nesnelere karsı olan duyarlılıktır [3]. Degeri çok büyük olan bir nesne, dahil olacagı kümenin ortalamasını ve merkez noktasını büyük bir derecede degistirebilir. Bu degisiklik kümenin hassasiyetini bozabilir.

Bu sorunu gidermek için kümedeki nesnelerin ortalamasını almak yerine, kümede ortaya en yakın noktada konumlanmıs olan nesne anlamındaki medoid kullanılabilir.
Bu islem k-medoids yöntemi ile gerçeklestirilir.

K-medoids kümeleme yönteminin temel stratejisi ilk olarak n adet nesnede, merkezi temsili bir medoid olan k adet küme bulmaktır [22]. Geriye kalan nesneler, kendilerine en yakın olan medoide göre k adet kümeye yerlesirler. Bu bölünmelerin ardından kümenin ortasına en yakın olan nesneyi bulmak için medoid, medoid olmayan her nesne ile yer degistirir. Bu islem en verimli medoid bulunana kadar devam eder
[22].

Sekil 2.3'de Oi ve Oj iki ayrı kümenin medoidlerini, Orasgele rasgele seçilen ve medoid adayı olan bir nesneyi, p ise medoid olmayan bir nesneyi temsil etmektedir. Sekil 2.3 Orasgele'nin, su anda medoid olan Oj'nin yerine geçip, yeni medoid olup olamayacagını belirleyen dört durumu göstermektedir [22].



Şekil 2.3 k-medoids Yöntemiyle Kümeleme Örnegi-2

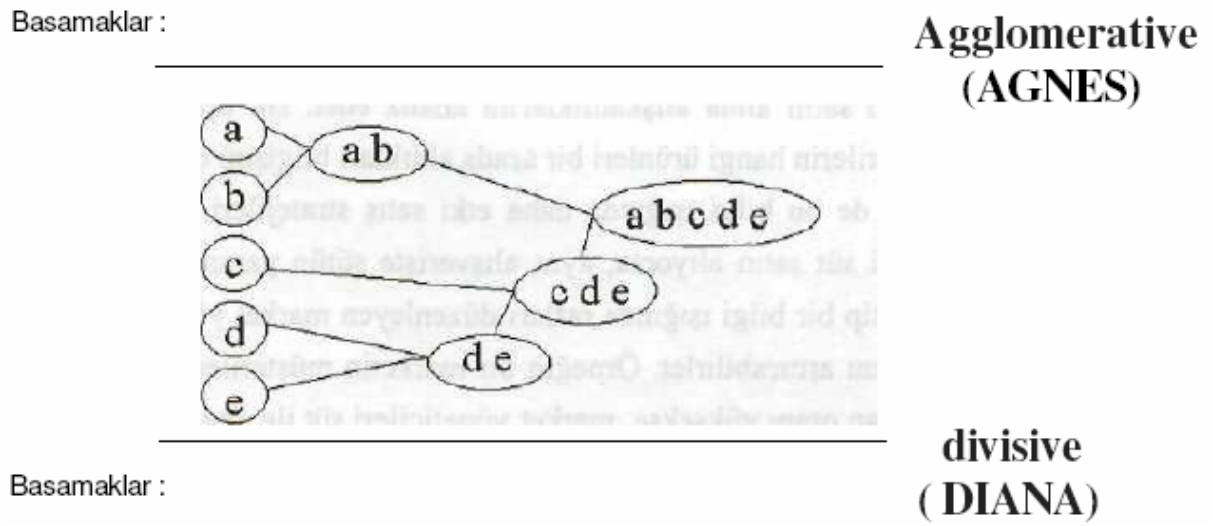
(a): p nesnesi su anda O_j medoidine baglıdır (O_j medoidinin bulundugu kümededir).
Eger O_j , Orasgele ile yer degistirir ve p O_i 'ye en yakınsa, p nesnesi O_i 'ye geçer.
(b): p nesnesi su anda O_j medoidine baglıdır. Eger O_j , Orasgele ile yer degistirir ve p
Orasgele'ye en yakınsa, p nesnesi Orasge!e'ye geçer.
(c): p nesnesi su anda 0_i medoidine baglıdır. Eger O_j , Orasgele ile yer degistirir ve p halaOrasgele 'ye en yakınsa, p nesnesi yine Oi 'ye baglı kalır.
(d): p nesnesi su anda O_i medoidine baglıdır. Eger O_i, Orasgele ile yer degistirir ve P

Orasgele ‘ye en yakınsa, p nesnesi Orasgele 'ye geçer.

Kümeleme yöntemlerinden biri olan hiyerarsık yöntemler, veri nesnelerini kümeler agacı seklinde gruplara ayırma esasına dayanır [24]. Hiyerarsık kümeleme yöntemleri, hiyerarsık ayrısmanın asagıdan yukarıya veya yukarıdan asagıya dogru olmasına göre agglomerative ve divisive hiyerarsik kümeleme olarak sınıflandırılabilir [24].

Agglomerative hiyerarsik kümelemede, Sekil 2.3'de görüldügü üzere hiyerarsik ayrısma asagıdan yukarıya dogru olur [24]. İlk olarak her nesne kendi kümesini olusturur ve ardından bu atomik kümeler birleserek, tüm nesneler bir kümede toplanıncaya dek daha büyük kümeler olustururlar.

Divise hiyerarsik kümelemede, Sekil 2.4'de görüldügü üzere hiyerarsik ayrısma yukarıdan asagıya dogru olur [24]. İlk olarak tüm nesneler bir kümededir ve her nesne tek basına bir küme olusturana dek, kümeler daha küçük parçalara bölünürler.



Şekil 2.4. Hiyerarsik Kümeleme Örnegi

Sekil 2.3, bir agglomerative hiyerarsik kümeleme yöntemi olan AGNES (AGlomerative NESting) ve bir divise hiyerarsik kümeleme yöntemi olan DIANA (DIvise ANAlysis) uygulaması göstermektedir [24]. Bu yöntemler bes nesneli

(a,b,c,d,e) bir veri setine uygulanmaktadır. Baslangıçta AGNES her nesneyi bir kümeyeyerlestirir. Kümeler, bazı kriterlere göre basamak-basamak birlesirler. Örnegin C_1 ve C_2 kümeleri, eger C_1 kümesindeki bir nesne ve C_2 kümesindeki bir nesne ile, diger kümelerdeki herhangi iki nesne arasında belirlenen uzaklık mesafesini karsılayacak bir mesafe varsa birlesebilirler. Bu birlesme islemi tüm nesneler bir kümede toplanıncaya kadar devam eder [22]. DIANA'da ise tüm nesnelerin içinde toplandıgı küme, her küme bir nesne içerecek duruma gelene kadar bölünür [24].

.

iv) Birliktelik Kuralları

Birliktelik kuralları büyük veri kümeleri arasında birliktelik ilişkileri bulurlar. Toplanan ve depolanan verinin her geçen gün gittikçe büyümesi yüzünden, sirketler veritabanlarındaki birliktelik kurallarını ortaya çıkarmak istemektedirler. Büyük miktardaki mesleki islem kayıtlarından ilginç birliktelik iliskilerini kesfetmek, şirketlerin karar alma islemlerini daha verimli hale getirmektedir[19].

Birliktelik kurallarının kullanıldıgı en tipik örnek market sepeti uygulamasıdır. Bu islem, müsterilerin yaptıkları alısverislerdeki ürünler arasındaki birliktelikleri bularak müsterilerin satın alma alıskanlıklarını analiz eder. Bu tip birlikteliklerin kesfedilmesi, müsterilerin hangi ürünleri bir arada aldıkları bilgisini ortaya çıkarır ve market yöneticileri de bu bilgi ısıgında daha etki satıs stratejileri gelistirebilirler. Örnegin bir müsteri süt satın alıyorsa, aynı alısveriste sütün yanında ekmek alma olasılıgı nedir? Bu tip bir bilgi ısıgında rafları düzenleyen market yöneticileri ürünlerindeki satıs oranını arttırabilirler. Örnegin bir marketin müsterilerinin süt ile birlikte ekmek satın alan oranı yüksekse, market yöneticileri süt ile ekmek raflarını yan yana koyarak ekmek satıslarını arttırabilirler.

Örnegin bir A ürününü satın alan müsteriler aynı zamanda B ürününü da satın alıyorlarsa, bu durum
A => B [destek = %2, güven = %60] Birliktelik Kuralı ile gösterilir.

Buradaki destek ve güven ifadeleri, kuralın ilginçlik ölçüleridir. Sırasıyla, kesfedilen kuralın kullanıslıgını ve dogrulugunu gösterirler. Yukardaki Birliktelik Kuralı için 2%

oranındaki bir destek degeri, analiz edilen tüm alısverislerden %2'sinde A ile B ürünlerinin birlikte satıldıgını belirtir. %60 oranındaki güven degeri ise A ürününü satın alan müsterilerinin %60'ının aynı alısveriste B ürününü de satın aldıgını ortaya koyar [20]. Kullanıcı tarafından minimum destek esik degeri ve minimum güven esik
degeri belirlenir ve bu degerleri asan birliktelik kuralları dikkate alınır[19].

v) Ardışık Örüntüler

Ardışık örüntü keşfi, bir zaman aralığında sıklıkla gerçekleşen olaylar kümelerini bulmayı amaçlar . Bir ardışık örüntü örneği şöyle olabilir: Bir yıl içinde Turgut Özakman'ın "Çılgın Türkler" romanını satın alan insanların %70'i Buket Uzuner'in "Güneş Yiyen Çingene" adlı kitabını da satın almıştır. Bu tip örüntüler perakende satış, telekomünikasyon ve tıp alanlarında yararlıdır.

vi) Apriori Algoritması

Büyük veri tabanlarında birliktelik kuralları bulunurken, su iki islem basamagı takip edilir:

1- Sık tekrarlanan ögeler bulunur: Bu ögelerin her biri en az, önceden belirlenen minimum destek sayısı kadar sık tekrarlanırlar.
2- Sık tekrarlanan Ögelerden güçlü birliktelik kuralları olusturulur: Bu kurallar minimum destek ve minimum güven degerlerini karsılamalıdır.

Sık tekrarlanan ögeleri bulmak için kullanılan en temel yöntem Apriori Algoritmasıdır. Asagıda Apriori algoritması bir örnekle anlatılmaktadır.

ANO	Ürün NO
A100	I1, 12,15
A200	I2,14
A300	I2,13
A400	I1,12,14
A500	I1,13
A600	I2,13
A700	I1,13
A800	I1,12,13,15
A900	I1,12,13

Tablo 2.1 Yapılan Alısveriş Bilgilerini İçeren D Veritabanı

Tablo 2.1'de bir marketten yapılan alısverislerin bilgilerini içeren D veritabanı görülmektedir. Bu veritabanında yapılan alısverislerin numaraları ANO sütununda görülmektedir. Her alısveriste satın alınan ürünler de Ürün No sütununda görülmektedir.

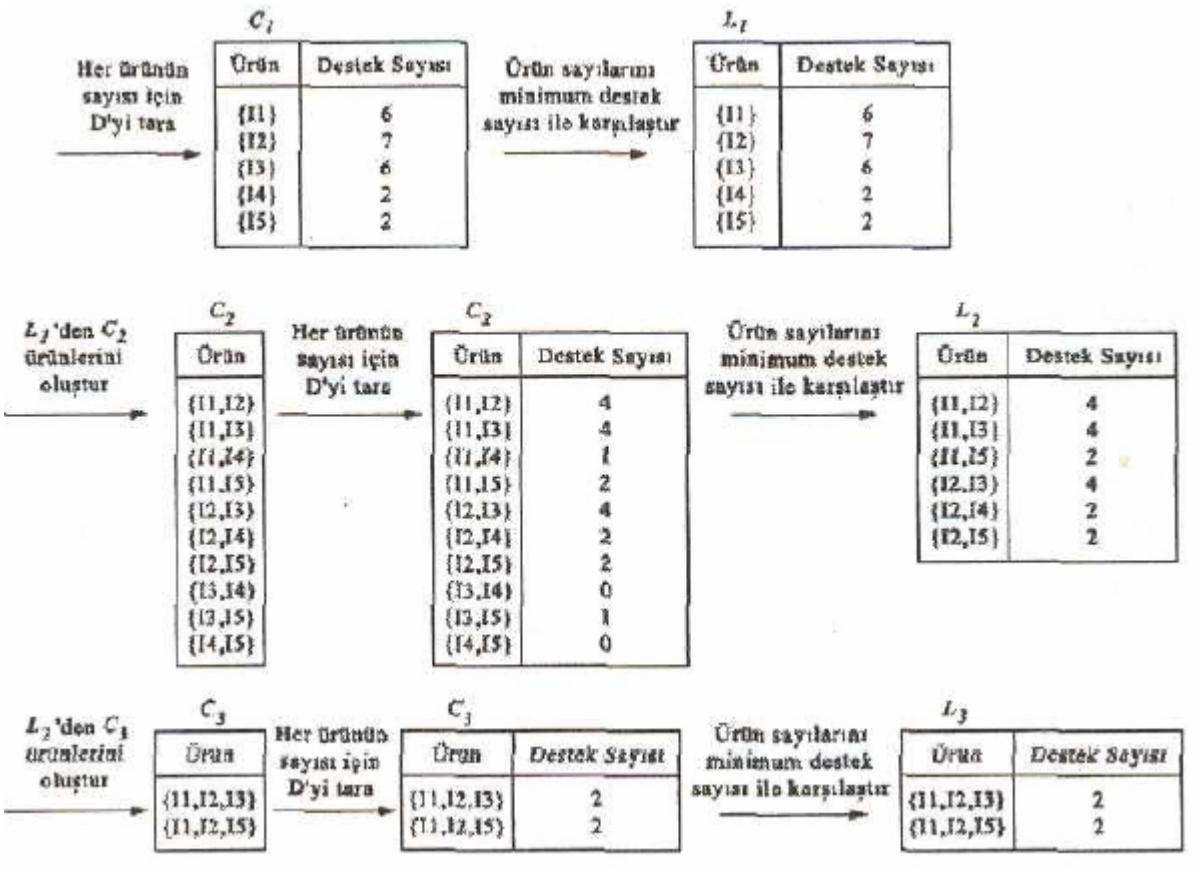
Apriori algoritmasında takip edilen basamaklar Sekil 2.4'de gösterilmektedir.

1- Algoritmanın ilk adımında, her ürün tek basına bulundugu C_1 kümesinin ele manıdır. Algoritma, her ürünün sayısını bulmak için tüm alısverisleri tarar ve elde edilen sonuçlar Sekil 2.5'de Destek Sayısı sütununda görülmektedir. Tablo 2.1'de görülebilecegi gibi D'de I1 ürününden 6 adet, 12 ürününden 7 adet, 13 ürününden 6 adet, 14 ürününden 2 adet ve 15 ürününden de 2 adet satıldıgı görülmektedir.

2- Minimum alısveris destek sayısının 2 oldugu varsayılırsa, tek baslarına sık tekrarlanan ürünler L_1 kümesinde görülmektedir. C_1 kümesindeki tüm ürünlerin destek sayısı, minimum destek esik degeri olan 2'den fazla oldugu için C_1 tüm ürünler sık tekrarlanan ürün olarak degerlendirilir ve L_1 kümesine aktarılır.

3- Hangi ürünlerin ikili olarak sık tekrarlandıgını belirlemek için L_1 kümesindeki ürünlerin ikili kombinasyonları bulunarak C_2 kümesi olusturulur.

4- C_2 kümesindeki ürünlerin destek sayılarını bulmak amacıyla D taranır ve bulunan degerler destek sayısı sütununda belirtilir.



Şekil 2.5. Apriori Algoritmasının Gösterimi

5- C_2 kümesindeki ürünlerden minimum destek esik degerini asan ürünler L_2 küme sine aktarılır.

6- Hangi ürünlerin üçlü olarak sık tekrarlandıgını belirlemek için L_2 kümesindeki ürünlerin üçlü kombinasyonları bulunarak C_3 kümesi olusturulur. Bu durumda C_3 = {{I1,I2,I3}, {I1,I2,I5}, {I1,I3,I5}, {I2,I3,I4}, {I2,I3,I5}} olması beklenir.

Ancak Apriori algoritmasına göre, sık tekrarlanan ögelerin alt kümeleri de sık tekrarlanan öge olması gerekmektedir. Buna göre yukarıdaki C3 kümesindeki elemanlar sık tekrarlanan olmadıgı için, yeni C_3 kümesi C_3 = {{I1,I2,I3}}, {I1,I2,I5}} olur.

7- C_3 kümesindeki ürünlerin destek sayılarını bulmak amacıyla D taranır ve bulunan degerler destek sayısı sütununda belirtilir.

8- C_3 kümesindeki ürünlerden minimum destek esik degerini asan ürünler L_3 küme sine aktarılır.

9- Hangi ürünlerin dörtlü olarak sık tekrarlandıgını belirlemek için L_3 kümesindeki ürünlerin dörtlü tek kombinasyonu {II, I2, I3, I5} olarak belirlenir. Ancak bu kümenin alt kümelerinin tamamı sık tekrarlanan öge olmadıgı için C_4 kümesi bos küme olur ve Apriori tüm sık tekrarlanan ögeleri bularak sonlanmıs olur. Sık tekrarlanan ögeleri bulduktan sonra , sıra birliktelik kurallarını olusturmaya gelir. Örnegin sık tekrarlanan bir öge için, bos olmayan tüm alt kümeler sunlardır :
[11]: {I1, I2}, {I2, I5}, {I2, I5}, {I1}, {I2},{I5}.
Bu durumda Tablo 2.1'deki veritabanına bakarak su birliktelik kuralları çıkartılabilir .

1- I1 Λ I2 => I5, güven = 2 / 4 = % 50
2- I1 Λ I5 => I2, güven = 2 / 2 = % 100
3- I2 Λ I5 => I1, güven = 2 / 2 = % 100
4- I1 Λ I2 => I5, güven = 2 / 6 = % 33
5- I2 Λ I1 => I5, güven = 2 / 7 = % 29
6- I5 Λ I1 => I2, güven = 2 / 2 = % 100

Eger minimum güven esik degeri %70 olarak belirlenmisse, ikinci, üçüncü ve altıncı kurallar dikkate alınır çünkü diger kurallar esik degerini asamamıs olurlar.

k-ögeküme	K adet öge içeren küme
L_k	Sık geçen k-öge küme kümesi (Bu kümeler mininum destek kıstasını sağlarlar). Bu kümenin her üyesi iki alandan oluşur. i) öge kümesi ii) destek sayacı
C_k	Aday k-ögeküme kümesi (Bu kümeler potansiyel olarak sık geçen öge_kümeleridir). Bu kümenin her üyesi iki alandan oluşur. i) öge kümesi ii) destek sayacı

Şekil 2.6 Apriori Algoritmasında Kullanılan Değişkenler

```
L1 ={sık geçen 1-ögeküme kümeleri}
SAYARAK YINELE ( k:=2, Lk-1 ≠ ∅, k:=k+1 ) [
  /* k adet ögeye sahip aday kümelerin bulunması*/
   Ck = apriori-gen(Lk-1);
   TUM t ∈ D hareketler için [
     /* t hareketinde yer alan aday kümelerin bulunması*/
     Ct = subset(Ck , t);
     TUM c ∈ Ct aday kümeler için
           c.sayac := c.sayac+1;
   ]
   Lk ={c ∈ Ck | c.sayac ≥ min-destek
]
∪k Lk DONDUR
```

Şekil 2.7 Apriori Algoritması Kesiti

```
INSERT INTO Ck
   SELECT p.öge1, p.öge2,… p.ögek-1, q.ögek-1
       FROM Lk-1 p, Lk-2 q
       WHERE p.öge1 = q.öge1 and… p.ögek-2= q.ögek-2 and p.ögek-1 <
q.ögek-1;

TUM c ∈ Ck aday kümeler için
  TUM c kümesinin (k-1) ögeye sahip tüm alt kümeleri için
     EĞER ( s ∉ Lk-1 ) ISE
         DELETE c FROM Ck;
```

Şekil 2.8: Apriori-gen Aday Küme Üretme Algoritma Kesiti

2.5 Veri Madenciliğinin Kullanım Alanları

Veri Madenciliği kullanım alanı olarak çok geniş bir yelpazeye sahiptir. Örnek uygulama alanları aşağıda belirtilmiştir [51]:

i) Finans Sektörü

Finans ve sigorta sektörü günümüzde sundukları hizmet, ürün ve servislerle bilgiye dayalı yönetime en fazla ihtiyaç duyan kuruluşlardır. Bu sektörde bilgiye dayalı yönetim özellikle ekonomik krizin yaşattığı sonuçlar göz önüne alındığında tartışmasız önemli ve zorunludur. Finans sektöründe en temel uygulamalar çapraz satış, risk derecelendirme, mevcut müşteriyi elde tutma, yeni müşteriler kazanma, maliyetleri azaltma, kayıp ve kaçakları engelleme, alternatif kanallar oluşturma, müşteri memnuniyetini sağlama olarak özetlenebilir. Hangi müşteri profilinin neyi, ne zaman ve neden tercih ettiğini anlayabilen bir kuruluş hem talep yaratma, hem de doğru zamanda doğru talebi karşılama ve sunma avantajına sahip olacaktır. Kuruluşun karlılığı artarken, müşterinin memnuniyeti de artacağından, aynı zamanda müşteri sadakati de sağlanmış olacaktır ki, ağ ekonomisinin en büyük kaosu budur. Mevcut müşteri kaybı, finans ve sigorta sektörlerinde en önemli problemi teşkil etmektedir. Yeni bir müşteri kazanmanın maliyetinin müşteriyi elde tutma maliyetinden daha yüksek olduğu, kaybedilen bir müşteriyi yeniden kazanma maliyetinin yeni müşteriler edinme maliyetinden daha fazla olduğu göz önüne alındığında şirketler müşteri odaklı gitmek ve mevcut müşteriyi ellerinde tutmak zorundadır. Bankalar, mevcut müşterilerden rakip bankaya geçme ihtimali olan müşterileri, profillerini ve kaybettikleri müşterilerin hangi sebepler yüzünden sistemden ayrıldıklarını tespit etmek istemektedir.

ii) Haberleşme Sektörü

Telekom sektöründe en önemli sorun müşteri kaybıdır. Kuruluşlar hangi müşterilerini kaybedebileceklerini önceden belirleyebildikleri taktirde bu müşterilerini elde tutma amaçlı stratejiler geliştirebilir, düşük maliyetli ve etkili

kampanyalar düzenleyebilirler. Kaybetme olasılığı olmayan bir müşteriye kalıcılığını sağlama amaçlı bir mesaj göndermek hem müşterinin kendisine verilmek istenen mesajın ne olduğunu algılamasını zorlaştıracak hem de maliyetleri artıracaktır. Örneğin Amerika'nın en büyük kablosuz iletişim sağlayıcısı olan Verizon kaybetme olasılığı yüksek olan müşterilerini ve müşteri kaybına neden olan faktörleri belirleme amaçlı bir Veri Madenciliği çalışması yapmıştır.

iii) Sağlık Sektörü

Doğru ve zamanında karar almanın hasta sağlığı üzerindeki etkisi tartışmasız çok önemlidir. Hastane bünyesinde toplanan operasyonel veriler, hasta verileri, uygulanan tedavi yöntemi ve tedavi sürecine dair veriler yöneticiler açısından bakıldığında; hastanedeki servislerin ve programların başarısının görüntülenmesi, kaynakların maliyetlerle göreceli olarak kullanımı, kaynak kullanımı ve hasta sayıları ile ilgili trendlerin tahmini, harcamalarla ilgili normal olmayan durumların anlık tespiti ve yolsuzlukların engellenmesi, hastanede uygulanan tedavi yöntemlerinin başarısının irdelenmesi açısından önemli bilgileri içermektedir. Bu veriler başarılı tedavi sonuçları almada etken faktörlerin belirlenmesi, ameliyatlarda yüksek risk faktörlerinin sınanması, hasta verilerinin yaş, cinsiyet, ırk ve tedavi yöntemi gibi faktörlere göre sınıflanması, hasta sağlığı açısından geriye dönük faktörlerin sınanması, tedavi yöntemi geliştirme vb. amaçlarla kullanılmaktadır. Dünya çapında çok sayıda başarılı uygulama örneği mevcuttur. Örneğin, San Francisco Hearth Institute; hasta sonuçlarının iyileştirilmesi, hastanın hastanede kalma süresinin azaltılması, vb amaçlarla bir çalışma başlatmış ve kurum bünyesinde toplanan verilerden hastanın geçmişine ait veriler, laboratuar verileri, kollestrol verileri, diğer medikal verileri bilgiye dönüştürmüştür.

iv) Devlet Uygulamaları

Kamu yöneticileri günümüzde verinin ve bilginin önemini kavramışlardır. Müşteriye özel hizmet sunan ticari kuruluşlarda olduğu gibi devlet kurumları da vatandaşlarının ihtiyaçlarına özel hizmet sunabilmenin önemini kavramışlardır.

Kamu yöneticileri için en önemli uygulamalar kaynakların doğru olarak kullanımını sağlama ve planlama; kamu güvenliğini sağlama amacı ile güvenlik problemlerini önceden tahmin etmek, rastlantısal olaylardaki sorunların çözümüne dair izleri keşfetme ve olası güvenlik sorunlarını eş zamanlı olarak tespit edebilme ve çözüm üretebilme; vergi ile ilgili yolsuzlukları ve izlerini belirleme, yolsuzlukları eş zamanlı olarak belirleme, sağlık ödemeleri, programların uygulanması vb. konularda şüpheli durumların tespiti, suiistimal ve israfları belirleme ve milyonlarca dolarlık zararı engelleme, örnekleri artırmak mümkündür. Kamuda enformasyon ve bilgi ihtiyacı sonsuzdur. Emniyet birimleri için suç istatistiklerine dair online raporlama, hangi profildeki insanların ne tür suçlara meyilli olduklarını belirleme, eş zamanlı suç engelleme politikaları oluşturmak ancak ileri analitik uygulamalar ile mümkündür. Günümüzde e-devlet kavramı oldukça kritiktir. E-devlet uzmanlarının en önemli hedefi bilgiye eş zamanlı olarak ulaşmak ve daha iyi hizmet vermektir. E-devlet uygulaması gerçekleştirilen ülkelerde kamu kuruluşları ziyaretçilerin sayfalarını nasıl kullandığı, ihtiyaç duyulan formlara kolayca ulaşılıp ulaşılamadığı, web sayfa tasarımın nasıl en iyi kullanılabilir hale getirilebileceği, hangi sayfaların hangi sıra ile ziyaret edildiğinin anlaşılması, geçmişteki ziyaretçi davranışlarına göre kurumun web sayfasını vatandaşın ihtiyacına daha iyi yanıt verecek şekilde yeniden düzenlemek mümkündür.

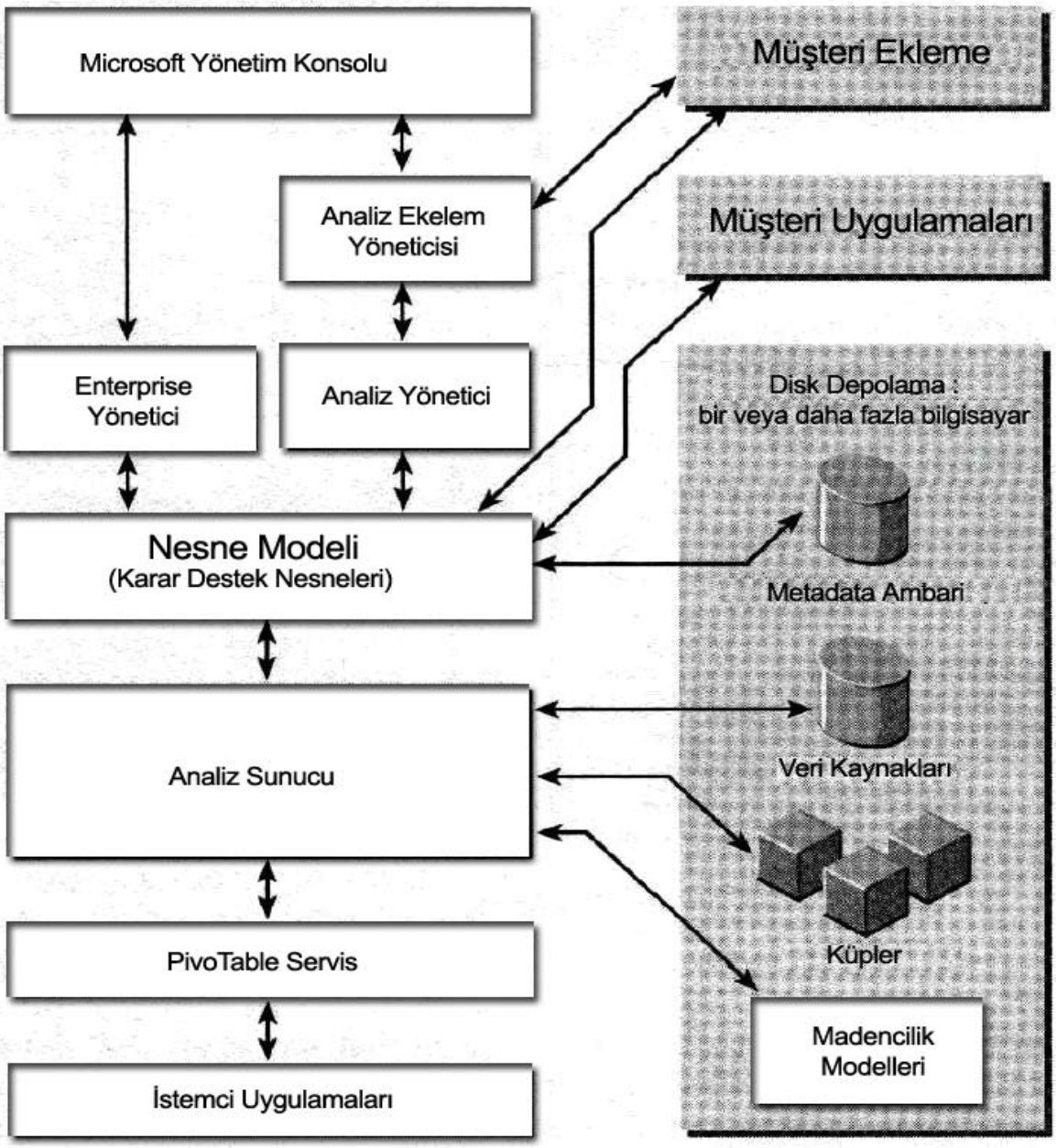
2.6 Veri Madenciliği Sistemleri Üzerine Yapılan Çalışmalar

VM teknikleriyle hem genel hem de özel amaçlı bir çok uygulamanın geliştirilmesi, VM tekniklerinin bir çok alanda gerekli olan bilgiye erişmek için uygulanabilir olmasıyla sağlanmaktadir [55].

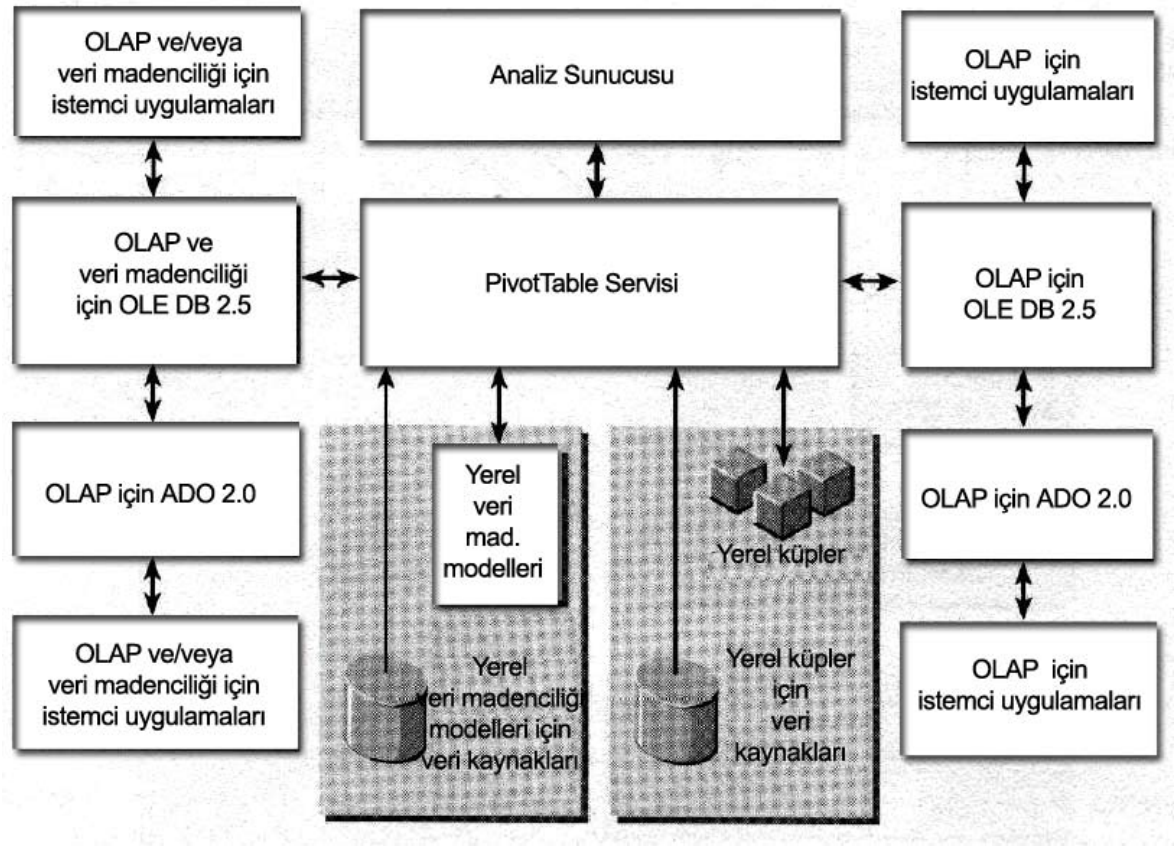
i) Analysis Manager

Analysis Manager, Microsoft firmasının VM için üretmiş olduğu ürünüdür [55]. Kümeleme analizi ve karar ağaçları için hazırlanmıştır. Analysis Manager OLAP (çevrim içi analitik işlem) küp desteği sunmaktadır. Analysis Manager'ın güçlü olduğu taraf kullanıcı-dostu (user friendly) bir ara yüze sahip olması ve uygulama kolaylığıdır. Aracın SQL SERVER 2000'le bütünleşik çalışabilmesi bu aracı etkin

hale getirmektedir. Analysis Manager'ın bir VM sorgusu için farklı algoritmaları desteklememesi en büyük eksikliğidir. Kaynak kodun açık olmaması uygulama geliştiriciler için büyük zorluklar oluşturmaktadır. Analysis Manager üretilen sonuçları farklı bir çok gösterim şekliyle kullanıcıya sunabilmektedir. Mesela karar ağaçları için karar ağacını gösterebildiği gibi sonuçları kural tabloları şeklinde yorumlama imkanı vermektedir. Analysis Manager içinde bulunduğu Analysis Services çatısı (framework) için Şekil 2.9'de sunucu mimarisi, Şekil 2.10'de ise istemci mimarisi verilmiştir [48].



Şekil 2.9 Analysis Services sunucu mimarisi



Şekil 2.10 Analysis Services istemci mimarisi

ii) Darwin

Darwin Oracle firmasının VM aracıdır [56]. Darwin regresyon ağaçları, karar ağaçları, kümeleme, yapay sinir ağları, Bayesian öğrenme, k-yakınlığında komşuluk gibi birçok algoritmayı destekleyen bir VM aracıdır. Paralel sunucular için geliştirilmiş bir VM sistemidir. Darwin kullanımı kolay bir ara yüze sahiptir. Darwin VM algoritmalarından CART, StarTree, StarNet ve StarMatch'i kullanır [48].

iii) Clementine

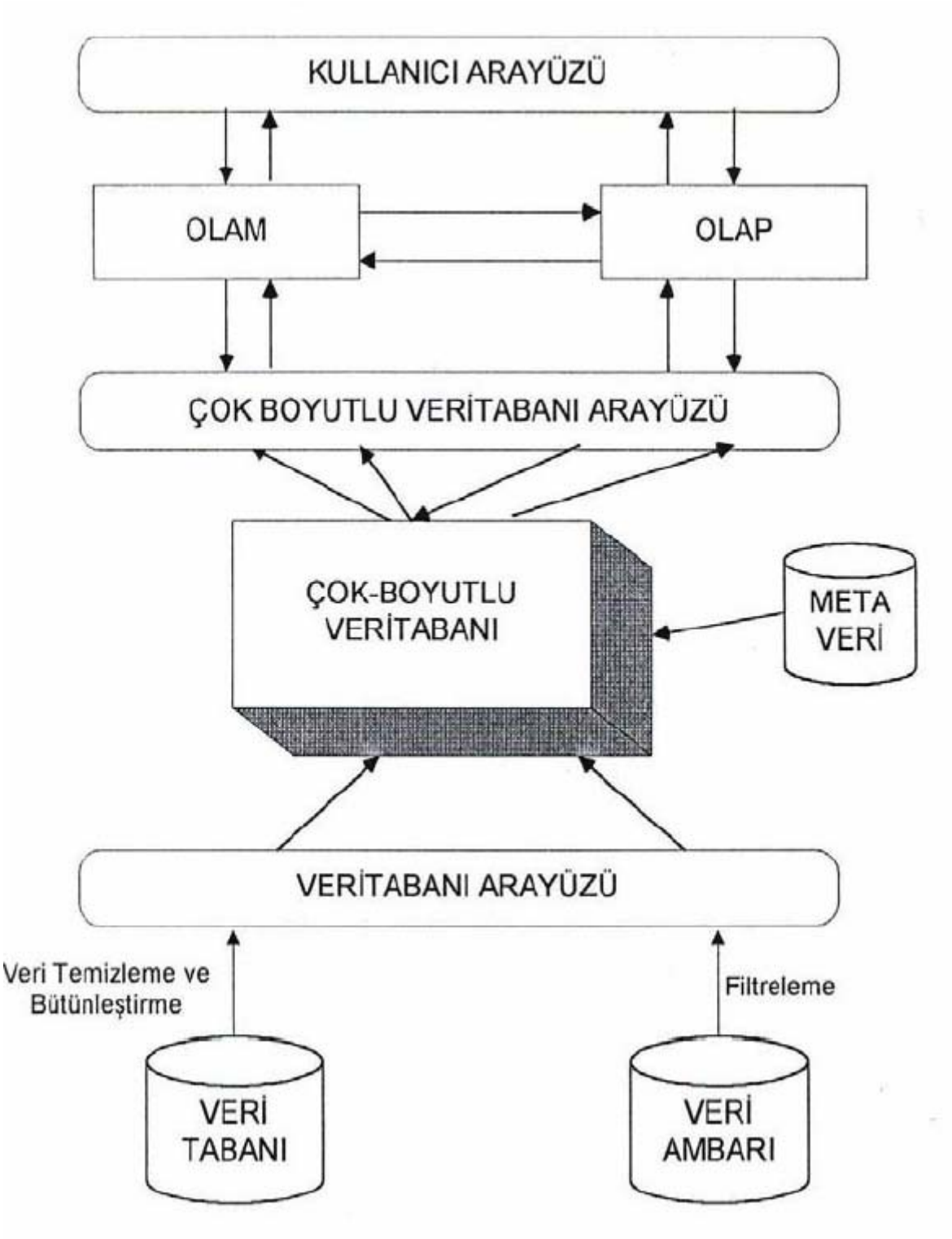
Clementine SPSS firmasının VM için geliştirmiş olduğu bir modüldür [57]. SPSS istatistiksel bir araçtır. Clementine'nin SPSS içinde bir modül olarak kullanılması kullanıcıların SPSS'in istatistiksel fonksiyonlarından faydalanmasına imkan verir. Yapay sinir ağları ve kural tümevarım yöntemlerini kullanır. Clementine müşteri

hizmetleri yönetimi, kimya sektöründe maddelerin aşındırıcılık tahmininde ve bankacılık alanında kredi kartı dolandırıcılıkları gibi konularda kendine uygulama alanı bulmuştur [48].

iv) DBMiner

Kanada Simon Fraser Üniversitesi tarafından geliştirilen bir sistemdir [58]. DBMiner sınıflama, kümeleme, eşleştirme ve sıra örüntüleri sorgularını yapabilecek VM algoritmalarını kullanır.

DBMiner çevrimiçi analitik işleme özelliğiyle VM algoritmalarının bütünleşik çalışabilme özelliği sayesinde ön plana çıkmaktadır. Bu özellik OLAM (Online Analytical Mining) olarak anılır. DBMiner OLAP ve VM yöntemlerini dinamik bir şekilde seçebilme imkanına sahiptir. Kullanıcının kolay kullanabileceği bir ara yüze sahiptir. Bu ara yüz sayesinde elde edilen sonuçlar çok yönlü bir soyutlama kullanılarak gösterilebilmektedir. DBMiner sisteminin mimarisi Şekil 2.11'de verilmiştir [48].



Şekil 2.11 DBMiner sisteminin yazılım mimarisi

Şekil 2.11'de de görüldüğü üzere DBMiner verilerini ilişkisel VT'dan ve/veya VA'dan alarak veri küpleriyle bütünleştirerek çok boyutlu VT'ye aktarır. Bu aktarım

kaynaktan ya verilerin bir bütün olarak çekilmesiyle ya da belli bir bölümünün çekilmesiyle gerçekleşir.
DBMiner'ın diğer sistemlere göre en büyük avantajı geliştirilen DMQL'i (data mining query language) kullanmasıdır. DMQL SQL benzeri bir VM sorgu dilidir. DMQL sayesinde çevrimiçi sorgular OLAM veya OLAP modülüne yönlendirilerek işlenir.
DBMiner'ın VT ara yüzü çok boyutlu veri tabanına temizlenmiş, filtrelenmiş ve bütünleştirilmiş verileri aktarmaya yarar. Veri aktarımı için ODBC ve OLE DB(Object Linking and Embedding Database) gibi bağlantılar da kullanılabilmektedir [48].

v) Data Logic/R

DataLogic/R kümeleme ve sınıflama analizi için kullanılan ticari bir VM aracıdır [59]. DataLogic/R artık nitelik ve verilerin temizlenmesi işlemlerini yapabilmektedir. Sistemin en güçlü olduğu taraf, üretilen kuralların öğrenme-test geçerliliği ve güvenlik gibi kriterlerde değerler üretmesidir. Bu değerler üretilen kuralların kalitesini belirlemek için kullanılabilmektedir. Bu araç, kimya ve ticaret sektöründeki çeşitli uygulamalarda kullanılmaktadır [48].

vi) INLEN

İlişkisel VT'den aldığı verileri makine öğrenimi teknikleriyle işledikten sonra ortaya çıkan sonuçları VT'ye yazmaktadır. Üretilen bilgi kesimi basit ya da bileşik olabilmektedir.
INLEN aracında dört işleç vardır [48].:
1. VT yönetim işleci: VT sorgularını yazmak için geliştirilen bir işleçtir.
2. Bilgi yönetim işleci: Üretilen bilgiyi yönetmek için kullanılır.
3. Bilgi üretim işleci: VT'den bilgi almak ve makine öğrenimi algoritmalarını çağırmak için kullanılır.
4. Makrolar: INLEN işleçlerini bir sırada tanımlamayı ve tek bir işleç gibi kullanabilmeyi sağlar.

vii) KDW (Knowledge Discowery Workbench)

KDW kümeleme, sınıflama, bağımlılık analizi algoritmalarını kullanan bir araçtır [60]. Etkileşimli veri analizine imkan vermektedir. INLEN sistemiyle bir çok ortak özelliği bulunmaktadır. [48].

viii) SKICAT (Sky Image Classification & Archiving Tool)

Özel amaçlı bir VM sistemidir. Özelleştiği konu astronomidir [47]. Bu araç astronomik verileri indirgemek ve karar ağacı analizi için ID3, GID3, O-Btree algoritmalarını kullanmaktadır. Görüntü işleme, veri sınıflama ve VTYS metotlarını kullanır. SKICAT adından da anlaşılabileceği gibi gökyüzü fotoğraflarındaki gök cisimlerini tanımlamak, bunları sınıflandırmak, kataloglamak için kullanılan bir araçtır [48].

ix) R-MINI

R-MINI [52], SKICAT gibi özel amaçlı bir VM sistemidir. Finansal konularda özelleşen R-MINI sınıflama ve sapma tespiti yapmak için kullanılır. R-MINI VT'den çektiği gürültü içerikli verileri kullanarak tamlık ve tutarlılık kriterlerini sağlayan en küçük kural kümesini bulur [48].

x) TASA (Telecommunication Network Alarm Sequence Analyzer)

TASA, telekomünikasyonda kullanılan özel amaçlı bir VM sistemidir. Telekomünikasyon hatlarında oluşabilecek bir hatanın önceden tahmini için kullanılır. Zaman serileri arası bağımlılıklarda kullanılan VM algoritmaları, hata tahmini için kullanılmaktadır. Hatlarda olağandışı bir olay meydana geldiğinde bu sistem tetiklenir. Tetikleme sayısının, kontrol edilebilecek sayının çok üzerinde olması böyle bir sisteme ihtiyaç doğurur [48].

xi) GCLUTO (Graphical CLUstering TOolkit)

GCLUTO Minnesota Üniversitesi tarafından gerçekleştirilmiş bir araçtır [http://www.cs.umn.edu/~mrasmus/gcluto]. Bu araç kümeleme algoritmaları için geliştirilmiştir. Girdi kütüğünden aldığı verileri istenen kümeleme algoritmasına göre işleyip sonuçları çıktı kütüğüne yazmaktadır. Kolay kullanılabilir arayüze sahip olması ve görüntüleme problemlerinin iyi çözülmüş olması, üretilen sonuçların farklı gösterimleri ile GCLUTO kümeleme analizi için güçlü bir araçtır [48].

xii) Enterprise Miner

SAS firmasının VM aracıdır. SAS'ın VA ve ÇAİ (çevrimiçi analitik işleme) araçlarıyla bütünleşik çalışabilmektedir. Enterprise Miner karar ağaçları, yapay sinir ağları, regresyon analizi, 2-aşama modelleri (two-stage models), kümeleme, zaman serileri, ilişkilendirme, vb. VM sorgularını ele alabilmektedir. Grafiksel arayüzü sayesinde kullanım kolaylığı sağlar ve kullanıcılar uygulamanın karmaşıklığından habersiz bir şekilde sadece girdi ve çıktılara yoğunlaşabilirler. 2 katmanlı mimariyi kullanır. İstemci bilgisayardaki yazılım gereksinimi Windows 98, 2000 ve NT'dir. Sunucu bilgisayardaki yazılım gereksinimi Windows 98, 2000 ve NT ile Linux'dür [48].

xiii) VM araçlarının karşılaştırmaları

Bu bölümdeki VM araçlarının karşılaştırmaları, Elder ve Abbott'un "A Comparison of Leading Data Mining Tools" [52] isimli sunum sonuçlarından faydalanılarak Tablo 2.2'de gösterilmiştir.

Özellik Belirtileri	Güçlü olduğu taraf	Zayıflıkları
Clementine	Görsel arayüz ; Algoritma genişliği	Scability
Darwin	Etkili istemci-sunucu; sezgisel arabirim seçenekleri	*No unsupervised* ; sınırlı görsellik
DataCruncher	Kolay kullanım	Basit algoritma
Enterprise Miner	Algoritma derinliği; görsel arabirim	Zor kullanım
GainSmarts	Veri dönüşümleri SAS üzerinde yapılabilmekte; algoritma derinlik seçeneği	*No unsupervised* ; sınırlı görsellik
Intelligent Miner	Algoritma genişliği; grafiksel ağaç/küme çıktısı	Az algoritma seçeneği; otomasyon yok
MineSet	Veri görüntüleme	Az algoritma; model dışarı aktarımı yapılamamaktadır
Model 1	Kolay kullanım; otomatik model keşfi	*Really a vertical tool*
ModelQuest	Algoritma genişliği	Bazı sezgisel olmayan arabirim seçenekleri
PRW	Geniş algoritma;otomatik model seçimi	Sınırlı görsellik
CART	Ağaç derinlik seçenekleri	Zor G/Ç; Sınırlı görsellik
Scenario	Kolay kullanım	sınırlı analiz
NeuroShell	Çoklu YSA mimarileri	*Unorthodox* arabirim; sadece YSA
OLPARS	Çoklu istatistiksel algoritmalar; sınıf-tabanlı görsellik	*Dated interface* ; zorlu kütük G/Ç
See5	Ağaç derinlik seçenekleri	Sınırlı görsellik; az veri seçenekleri
S-Plus	Ağaç derinlik seçenekleri; görsellik ; programlanabilir / genişletilebilir	*Limited inductive methods ; step learning curve*
WizWhy	Kolay kullanım; modeller kolay anlaşılabilir	Sınırlı görsellik

Tablo 2.2 VM araçlarının güçlü ve zayıf olduğu alanlar

3.WEB MADENCİLİĞİ

World wide web (www) günümüzde inanılmaz hızda bir büyüme göstererek kullanıcıların dünyasındaki en önemli ortamlardan birisi olmuştur. Basit bir tahminde bulunacak olursak insanoğlunun asırlar boyu edindiği bilgiler ve günlük veriler çok kısa bir sürede, 10–15 yıl, internetten erişilebilir hale gelmiştir; bu büyüklük de verinin bir yerde ve erişilebilir olması gelecek için insanı umutlandırmaktadır.

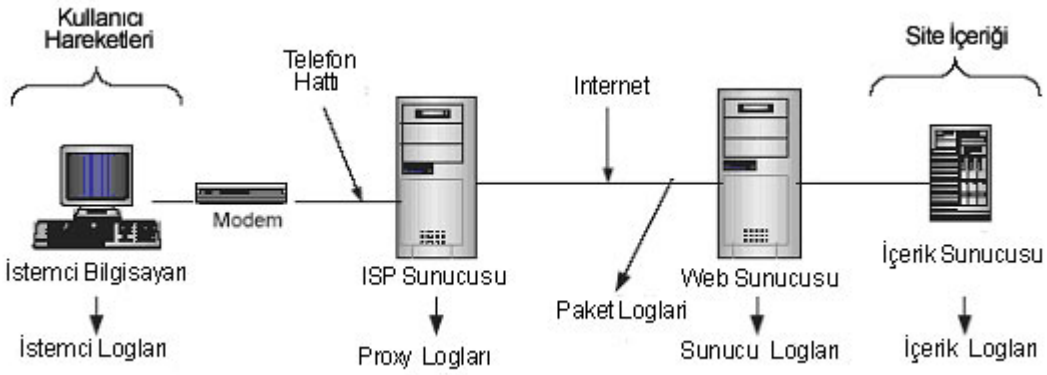
İnternet siteleri artık kullanıcılarını tanımak istemekte, onları yönlendirmekte ve onlar hakkında bilgi toplamaktadır. Bu sayede verimi ve müşteri kitlesini arttırmayı amaçlamaktadırlar. Web Madenciliği için kabaca webde bulunan bilgilerin keşfedilmesi ve yorumlanması olarak tanımlanabilir. Bu iş çevrimiçi olarak bulunan kaynakların otomatik olarak aranmasıdır [13].

Web madenciliği ilk kez 1996 yılında Oren Etzoni tarafından dile getirilmiştir[1]. Web madenciliği, veri madenciliği tekniklerinin kullanılarak web belgelerinden ve servislerinden otomatik olarak bilginin ayıklanması, ortaya çıkarılması ve analiz edilmesidir[1,2]. İşlenecek olan ham veri, ziyaretçilerin sayfaları gezerken bıraktıkları izlerin (bilgilerin) yanı sıra üye olurken verdikleri bilgilerden oluşmaktadır. Web madenciliği ile bu bilgiler farklı veri madenciliği teknikleri kullanılarak site sahibine yararlı bilgilere dönüştürülür. Bu sayede ticari amaçlı bir siteden elde edilen kar miktarı arttırılabileceği gibi, internet sayfaları farklı ilgi alanlarına göre düzenlenerek ziyaretçi memnuniyeti arttırılabilir.

Web madenciliğinin başarıyla kullanıldığı ve müşteri/ziyaretçi memnuniyetiyle site karının arttırıldığı örneklerin başında www.amazon.com alışveriş sitesi gelmektedir. Sitede yeni üye olanlardan ilgi alanlarını algılamak amacıyla farklı ürünler listelenerek en çok beğendiklerini seçmeleri istenmektedir. Üyelik işlemi sırasında kullanıcı hakkındaki ilk yararlı bilgilerin toplanmasının yanı sıra kullanıcıların sitede alışveriş yaptıkları süre boyunca ilgi duyduğu farklı alanlar da veri ambarına kaydedilmektedir. Bu veriler daha sonra işlenerek kullanıcının profiline uygun sayfalara ulaşabilmesi kolaylaştırılmaktadır.

İnternet hali hazırda büyük bir bilgi hazinesi olarak hayatımızda çok önemli bir yer tutmaktadır. Internet ortamında bilgilerin herhangi bir site içerisinde kullanıcılar için düzensiz hali bilgiye ulaşımı zorlaştırmaktadır. Bu nedenle web madenciliği, mevcut verinin anlamlı ve bir düzen içinde sunulmasının sağlanması konusunda önemli bir işlemdir.

3.1 Web Veri Kaynakları



Şekil 3.1 Web erişim diyagramı

Web Madenciliği yapılırken veri değişik kaynaklardan toplanmaktadır. Şekil 3.1'de web erişim verileri ayrıntılı olarak gösterilmiştir. Kaynaklara örnek olarak sunucu tarafı (server-side), istemci tarafı (client-side), vekil sunucu (Proxy server), kurum/kuruluş veri tabanı ve benzerleri verilebilir. Web Kullanım Madenciliğinde öncelikli veri kaynakları web sunucu kaynak dosyaları ve uygulama sunucusu kaynak dosyalarıdır. Site dosyaları, meta veriler, operasyonel veritabanı, uygulama şablonları ve tanım kümesi bilgileri verinin hazırlanması ve desen bulmada kullanılan ek kaynaklardır.

Veriyi geldiği kaynağa göre de sınıflandırabiliriz [26]:

i) İçerik (Content): Sitedeki verinin içeriği kullanıcıya iletilen objelerin ve ilişkilerin toplamıdır ve web sayfaları içerisindeki gerçek veridir. Genellikle yazı ve grafikten ibarettir ama son zamanlarda internetin gelişmesi ile başka formatlarına da rastlanmaktadır. Bunlara örnek olarak statik HTML/XML sayfalar, resimler, video klipler, ses dosyaları, betikler tarafından dinamik olarak oluşturulan sayfa kısımları, diğer uygulamalar ve operasyonel veri tabanından veri kayıtlarının birleşimi

verilebilir. Site veri içeriği bunlardan başka tanımlayıcı kelimeler, doküman özellikleri, semantik taglar yada http değişkenleri gibi semantik ve yapısal meta-verileri de içerir. Son olarak, site için tanımlı küme ontolojisi veri içeriğinin bir parçası olarak düşünülür[27].

ii) 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 ve 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[27].

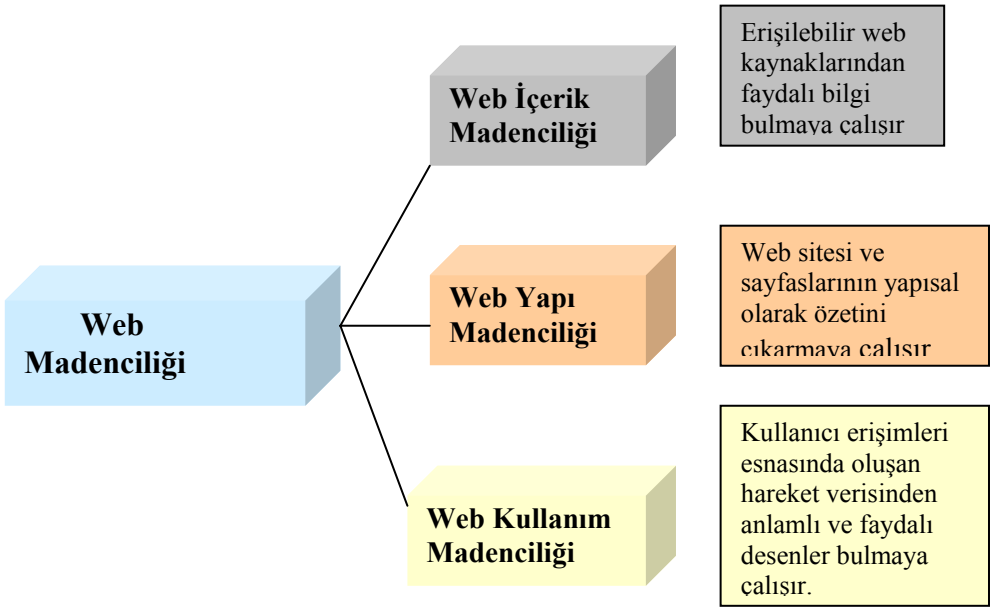
iii) Kullanım (Usage): Web sayfalarının kullanım bilgilerini gösteren veridir. Bu bilgiler içerisinde IP adresleri, sayfa referansları, bağlantı tarih ve saati verilmektedir. Web ve uygulama sunucularından otomatik olarak toplanan kayıt dosya (log) verileri kullanıcıların yönelim (navigational) davranışlarını 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[27].

iv) 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 yada kayıtlı kullanıcıların ayırıcı bilgileri, sayfalar, ürünler yada filmler, geçmiş alışverişler gibi çeşitli objelerdeki kullanıcı oranları, yada 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[27].

3.2 Web Madenciliği Sınıflandırması

Web Madenciliği ortaya atıldığı ilk zamanlarda iki sınıfa ayrılmaktaydı. Web İçerik Madenciliği (Web Content Mining) ve Web Kullanım Madenciliği (Web Usage

Mining). Web Madenciliğinin yaygınlaşması ile Web Yapı Madenciliği de (Web Structure Mining) üçüncü bir sınıf olarak literatüre girmiştir[1]. Şekil 3.2'de bu sınıflandırma görülmektedir.



Şekil 3.2 Web Madenciliği Sınıflandırması

Web İçerik Madenciliği, web dokümanları içerisinde saklı olan bilgileri çıkarmak amacıyla kullanılmaktadır. Web Yapı Madenciliği, web sayfaları ve web siteleri arasındaki bağlantıları inceleyerek bir takım bilgiler üretir. Elde ettiği bilgileri sitenin yapısal tasarımını iyileştirmek için kullanır. Web Kullanım Madenciliği ise temel olarak web sitelerinin kullanımı, site ziyaretçilerinin hareketlerinin incelenmesi ile ilgilenmektedir.

	Web Madenciliği		
	İçerik Madenciliği	**Yapı Madenciliği**	**Kullanım Madenciliği**
Veri	Metin belgeleri ve HTML sayfaları	HTML Linkleri	Server log, browser log dosyaları...
Verinin Şekli	Yapısız, karışık	Link Yapısında	Kullanıcı Etkileşimi
Gösterimi	İlişkisel ve sınıflandırmalı	Grafik	İlişkisel Tablolar ve Grafik

Tablo 3. 1 Web Madenciliği Sınıfları arasındaki temel farklılıklar

i) Web İçerik Madenciliği

Web içerik madenciliği, web kaynaklarından içeriklerine göre otomatik bilgi arama tekniklerini tanımlar. Web kaynakları içerisinde metin, resim, ses, görüntü, metadata ve hiper linkler bulunmaktadır. Web içerik madenciliğin amacı, bu kaynaklar arasından bilginin bulunması veya filtrelenmesidir.

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.

Web içerik madenciliğinde kullanılan iki yaklaşım vardır:

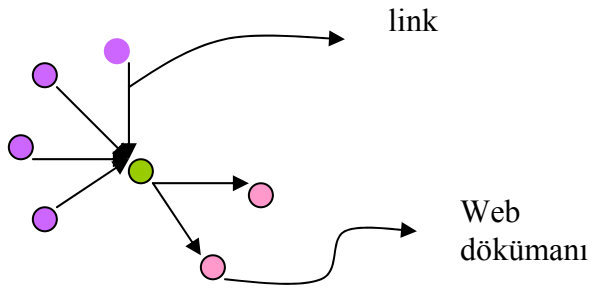
i) Bilgi Erişim Yaklaşımı: Kullanıcı profili baz alınarak kullanıcılara gösterilen bilgileri filtrelemek ve bilgiye erişimi geliştirmek için kullanılan yöntemdir.

ii) Veritabanı Yaklaşımı: Web'deki veriyi modellemek ve veriyi bütünleştirerek daha karmaşık bir yapıya sokmak için kullanılan yöntemdir. Bu yöntem sayesinde anahtar kelime temelli arama yerine daha gelişmiş sorgular çalıştırmak mümkün olur.

ii) Web Yapı Madenciliği

Web yapı madenciliği, web siteleri ve web sayfaları arasındaki bağlantı (link) verisine bakarak bilgi üretmektir. Teknik olarak, Web içerik madenciliği dokümanın içeriğine, yapı madenciliği ise dokümanlar arası bağlantılara odaklanır. Web yapı madenciliği, linklerin topolojisine dayanarak farklı siteler arasındaki benzerlik ve ilişki gibi bilgileri üretir, sayfaların link tasarımlarını ortaya çıkarmamıza yardımcı olur. İlgili araştırmalar "hyperlink" düzeyinde yapılıyorsa "Hyperlink Analysis" adını alır.

Şekil 3.3 ‘de web grafik yapısı görülmektedir [13]. Web dokümanları arasındaki oklar iki sayfa arasındaki ilişkiyi temsil etmektedir.

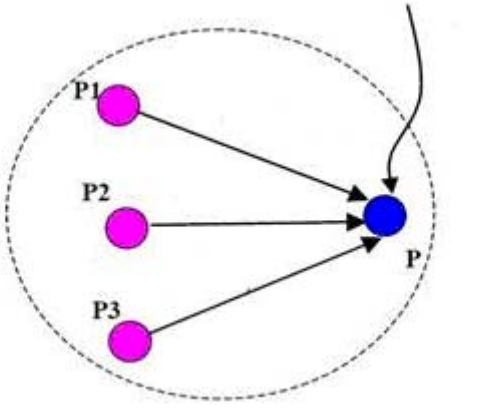


Şekil 3.3 Web sayfaları arasındaki link bağlantısı

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 yapan özelliği “Hyperlink Analyse” yöntemini başarılıyla uygulamasıdır. Google’ın PageRank (Şekil3.4) teknolojisi, link yapılarını kullanarak her bir sayfa için bir derece hesaplar[13]. Bu sayede Google istenen konu ile ilgili bir sayfayı getirirken, bu sayfa ile ilgili diğer sayfaları da getirir.



Şekil 3.4 Sayfaların derecelendirilmesi

iii) Web Kullanım Madenciliği

Web kullanım madenciliği kullanıcın siteyi kullanırken gerisinde bıraktığı erişim verilerinden bilgi üretmeyi amaçlar. Bu veriler ikinci sınıf verilerdir yani bir yere girilmiş; bir yerde yazılan ya da kullanıcın isteğiyle oluşan, linke tıklamak gibi, veri değildir. Tamamen kullanıcıdan bağımsız oluşur ve çok ciddi boyutlardadır. Bu veriler istemcilerde, sunucularda ve Proxy sunucularında depolanır. Veri kaynakları olarak sunucu erişim kayıtları, referrer kayıtları, agent kayıtları, istemci tarafında bulunan çerezler (cookies), kullanıcı profilleri (kayır bilgileri), metadata (sayfa özellikleri, içerik özellikleri, kullanılan veri) sayılabilir.

Bu konudaki çalışmalar Genel Web Kullanım Madenciliği, Site Güncelleme Sistemleri, Sistem İyileştirme ve Kişiselleştirme başlıkları altında toplanabilir. Genel Web Kullanım Madenciliği Sistemleri kullanıcıların genel davranış biçimlerini bilinen ya da önerilen veri madenciliği algoritmalarını sunucu erişim dosyalarındaki veriye uygulayarak bulmaya çalışır. Site Güncelleştirme Sistemlerinin hedefi ise site içerik ve yapısında yapılması gereken tadilatları bulmaktır. Sistem İyileştirme üzerine yapılan araştırmalar web kullanım verisini kullanarak trafiği etkinleştirmeyi hedefler. Son olarak, kişiselleştirme çalışmaları bireysel taleplere göre değişen siteler oluşturmaya çalışır.

Bu bilgilerin çoğu web sunucuların otomatik olarak tuttuğu günlük dosyalarından elde edilir. Günlük dosyaları (Şekil 3.5), istemciden sunucuya gönderilen her bir isteğin bir kayıt olarak eklenmesi ile meydana gelir.[17]

10.0.3.124	05/Dec/2005:14:03:27	GET	/~bdengiz/End-308.htm	HTTP/1.1	404	301	
65.55.246.42	05/Dec/2005:14:03:27	GET	/english/hastalan.php	HTTP/1.0	200	2988	
81.214.188.253	05/Dec/2005:14:03:28	GET	/~egfak/bolumler/ilko/snfo/index.html	HTTP/1.0	200	11292	
10.0.1.33	05/Dec/2005:14:03:28	POST	/~htinmaz/oto205d.php	HTTP/1.1	200	26180	
85.96.67.176	05/Dec/2005:14:03:29	GET	/~gkose/personal.php	HTTP/1.1	200	3851	
10.0.3.124	05/Dec/2005:14:03:29	GET	/~bdengiz/courses.html	HTTP/1.1	304	0	
10.0.3.124	05/Dec/2005:14:03:30	GET	/~bdengiz/End-421.htm	HTTP/1.1	304	0	

Şekil 3.5: Web Kayıt Dosyası

- Web kullanım madenciliği; Önişlem (Preprocessing), patern keşfi (Pattern Discovery) ve patern analizi (Pattern Analysis) aşamalarından oluşur

3.3 Web Madenciliği Teknikleri

Ön işlemden geçirilen veriler üzerinde Web Madenciliği teknikleri uygulanarak bir takım çıkarımlarda bulunulur. Yaygın olarak kullanılan Web Madenciliği teknikleri:

i) İstatistik: İstatistiksel teknikler bir web sitesi ve ziyaretçileri hakkında bilgi açığa çıkarmaya yarayan en güçlü araçlardır. Analizciler oturum dosyasını analiz ederken farklı değişkenler üzerinde farklı açıklamalı istatistiksel analiz tiplerini yerine getirirler. Bu sayede web sayfasındaki güvenlik sorunları, sistem performansı ve benzeri konularda bilgiler elde edebilirler.

- Hangi kullanıcılar tarafından hangi sayfalar kullanılıyor?

- Hangi web tarayıcıları ile sayfalara erişiliyor?

- Resim ve diğer bağlı dosyalar olmadan kaç ziyaretçi var?

İstatistik analiz yapmak için internette bir çok serbest yazılım bulunmaktadır. Bunlar arasında en çok bilinenleri AWSTAT (http://awstats.sourceforge.net/), ANALOG (http://www.analog.cx/) ve WEBALİZER (http://www.mrunix.net/webalizer/) yazılımlarıdır.

ii) İlişkilendirme Kuralları (Association Rules): Genellikle alışveriş uygulamalarında kullanıldığı için İlişkilendirme Kuralları aynı zamanda Alış Veriş Sepeti (Market Basket) analizi olarak da tanınmaktadır. Bu yöntemdeki amaç bir küme içerisindeki nesnelerin birbirleri ile olan bağlarının tespit edilmesidir. Bu Veri Madenciliği yöntemi yaygın olarak alışveriş sistemlerinde kullanıldığı görülse de başka uygulamalarda da kullanılmaktadır.

İlişkilendirme Kuralı yöntemine örnek verecek olursak A ürününün alınması ile B ürününün veya C ürünün alınması arasında bir bağlantı olup olmadığının tespit edilmesi ve eğer bağlantı var ise bu bağlantılar arasındaki kuvvet veya önem derecesinin ortaya çıkarılması sağlanır. Bu analizin amacı A ürününü alan kişilerin B veya C ürünleri alımlarıyla ilgili olarak kuvvetli bir bağlantı bulup sistemde bir takım değişiklikler gerçekleştirmektir. Örneğin, süpermarket sisteminde çeşitli promosyonların gerçekleşmesi, ürün raflarının elde edilen sonuçlar doğrultusunda

yerleştirilmesi olabilir. Bu işlem bir web sitesi içerisinde benzer olarak sayfaların yapılandırılmasında kullanılır.

iii) Sıralı Desen (Sequential Patern) : Sıralı desen yöntemiyle kullanıcı oturumları arasında desen bulunmaya çalışılır. Sıralı desen bulma işleminde, belirli zaman aralıklarında oturumlar incelenir ve karşılaştırma yapılır. Sıralı desen yönteminde, eğilim analizi, değişen nokta bulma veya benzerlik analizleri gibi bazı geçici analiz tipleri kullanır. Sıralı desenlerin bulunması, örneğin gelecekteki eğilimi tahmin edecek web pazarlamacıları için oldukça anlamlıdır. Böylece ilanlar belirli kullanıcı gruplarına yönlendirilebilinir.

iv) Kümeleme (Clustering): Kümeleme yöntemi aynı özelliğe sahip olan nesnelerin bir araya getirilmesi işlemidir. Web Madenciliğinde genel olarak iki kümeleme yaklaşımı vardır: Kullanıcı Grupları (User Clusters), Sayfa Grupları (Page Clusters).

Birinci yaklaşımda (Kullanıcı Grupları) amaç, benzer sayfa görüntülemesi yapan kullanıcıları tespit edip bir grup içerisine almaktır. Bu yöntem özellikle web kişileştirme işleminde oldukça yararlıdır. Örneğin, bir portal içerisinde oyun ve spor sayfalarına girenleri bir grup içerisinde toplayıp kişilerin bir sonraki bağlantısında oyun ve spor konulu reklamların ekrana gelmesi gibi.

İkinci yaklaşımda (Sayfa Grupları), benzer içerikli sayfaların bir arada gruplandırılması özellikle arama motorları için çok yararlı olmaktadır. Böylelikle bir kullanıcının aramış olduğu bilgilere daha hızlı bir şekilde ulaşlabilmesi sağlanır.

v) Sınıflandırma (Classification): Sınıflandırma bir veriyi daha önceden tanımlanmış sınıflara dağıtma tekniğidir. Sınıflandırma işleminde, verilen bir sınıf veya kategorinin özelliklerini en iyi biçimde açıklamak için seçim ve açığa çıkarma uygulamalarına ihtiyaç duyulur. Sınıflandırma; karar ağaçları, bayezian sınıflayıcıları, en yakın komşu ve destek vektör makineleri gibi denetlenen tümevarımsal öğrenim algoritmaları kullanılarak yapılabilir.

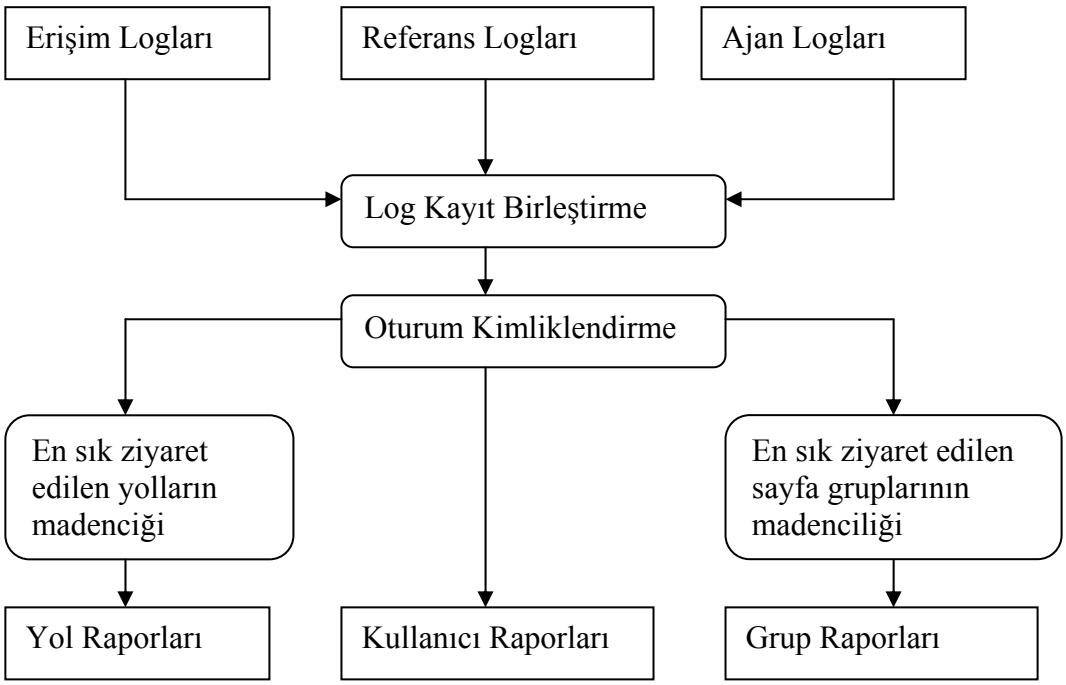
3.4 Web Madenciliğinde Kullanılan Araçlar

Web madenciliğine yardımcı olacak bir çok ticari araç bulunmaktadır. Bunlardan SpeedTracer, Clementine,Net Analysis aşağıda açıklanmaktadır.

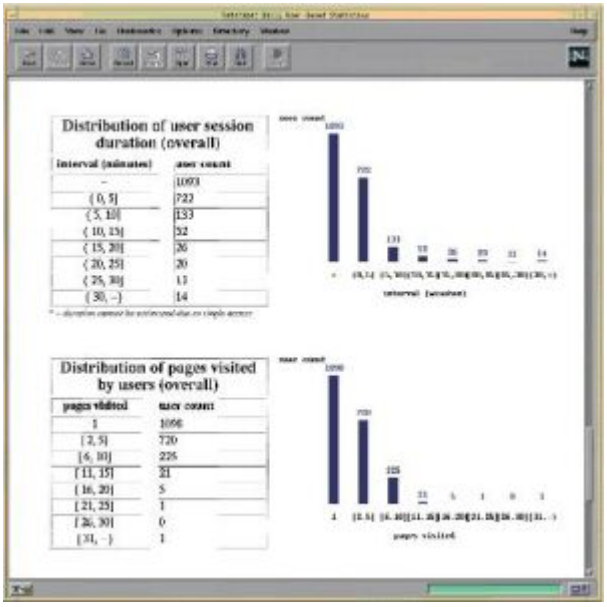
1. Ticari Araçlar

i) SpeedTracer

IBM'in geliştirdiği web madenciliği aracıdır. "SpeedTracer" kullanıcının gezindiği sayfalar hakkında şablonlar çıkaran, raporlayan ve bu sayede site yöneticilerine web sayfalarının yapısını düzenleme imkânı veren web kullanım madenciliği ve analizi aracıdır. Uygulama, yenilikçi çıkarım algoritmaları kullanarak kullanıcının izlediği yolları yeniden yaratır ve kullanıcı oturumlarını tanımlar. Gelişmiş madencilik araçları kullanıcının web sitesindeki hareketlerini açığa çıkarır. Program çıktısı olarak web sitesi yöneticilerine kullanıcı davranışlarını daha iyi anlamalarına yardımcı olacak sayfalardaki gezinti bilgilerinden oluşan bir koleksiyon sunulmaktadır.



Şekil 3.6. SpeedTracer'in gerçekleştirimi



Şekil 3.7 SpeedTracer analiz raporu örneği

SpeedTracer 3 çeşit istatistik oluşturmaktadır: Kullanıcı temelli (user-based), gezilen yol temelli (path-based), grup temelli (group-based):

- Kullanıcı temelli istatistikler, kullanıcının giriş zamanlarını ve sistemde kaldığı süreyi tam olarak göstermektedir.
- Gezilen yol temelli istatistikler sıklıkla gezilen yolları tanımlar.

Grup temelli istatistikler, sıklıkla ziyaret edilen web sayfası grupları hakkında bilgi verir.

ii) Clementine

Bölüm 2.6.3 ‘de de açıklandığı gibi, SPSS firması tarafından geliştirilmiştir.[18] Clementine, yenilikçi çıkarım algoritmaları kullanarak kullanıcının izlediği yolları yeniden yaratır ve kullanıcı oturumlarını tanımlar. Gelişmiş madencilik araçları kullanıcının web sitesindeki hareketlerini açığa çıkarır. Program çıktısı olarak web sitesi yöneticilerine kullanıcı davranışlarını daha iyi anlamalarına yardımcı olacak sayfalardaki gezinti bilgilerinden oluşan bir koleksiyon sunulmaktadır.

iii) Net Analysis

Net Genesis firmasının ödüllü çevrim içi davranış analizi çözümü olan Net Analysis, çevrim içi rekabetçi ortamlarda e-iş yatırımları için gerekli yüksek düzeyde esneklik ve genişletilebilirlik sağlamaktadır. Yüksek düzeydeki esnekliği ve işlevselliği ile Net Analysis her bir şirketin belirli e-müşteri ihtiyaçlarını karşılarken şirketin mevcut mimarisine de kolayca uyum gösterir.

2. Ücretsiz Araçlar

Web madenciliğinde kullanılan ücretsiz araçlardan Weblog_parse, Weblog, Analog kısaca aşağıda tanıtılmaktadır.

i) Weblog_parse

“ACME Labs” yazılımı olan weblog_parse, log dosyaları işleme aracıdır. Weblod_parse, web log dosyasından belirli alanları çıkarır. Web sunucunun log dosyasını okur ve sadece kullanıcı tarafından istenen alanları listeler.

ii) Weblog

Darryl C. Burgdorf tarafından geliştirilen “weblog” giriş loglarını ayrıntılı şekilde analiz eden bir araçtır. Ay, hafta, gün ve saat bazında site üzerinde aktivitelerin (toplam sayfa ziyareti, ne kadar veri transferi olduğu, popüler sayfalar,vs..) izlenmesine imkan tanır.

iii) Analog

“Cambridge Üniversitesi İstatistiksel laboratuarı” tarafından geliştirilen Analog, web sunucusu log dosyalarını analiz eden bir programdır. Program sayesinde bir web sitesindeki kullanıcı aktiviteleri hakkında bilgi sahibi olunabilir. Hangi sayfalar daha popüler, insanlar hangi ülkelerden web sitesine giriş yapıyorlar gibi bilgiler elde edilebilmektedir.

4. WEB KULLANIM MADENCİLİĞİ

Web kullanım madenciliği, bir veya birçok web sunucusundan kullanıcı erişim desenlerinin otomatik keşfinin ve analizinin yapıldığı bir tip veri madenciliğidir. Web kullanım madenciliğinin amacı, kullanıcının siteyi ziyaretinden sonra gerisinde bıraktığı erişim bilgilerinden veri üretmektir. Bu veriler ikinci sınıf verilerdir, yani kullanıcının isteği dışında oluşan verilerdir. Kuruluşlar bu yolla her gün yüzlerce megabayt veri toplamaktadır. Bu bilgilerin çoğu web sunucuların otomatik olarak tuttuğu günlük dosyalarından elde edilir [26]. Günlük erişim dosyaları (Şekil 4.1), istemciden sunucuya gönderilen her bir isteğin bir kayıt olarak eklenmesi ile meydana gelir.

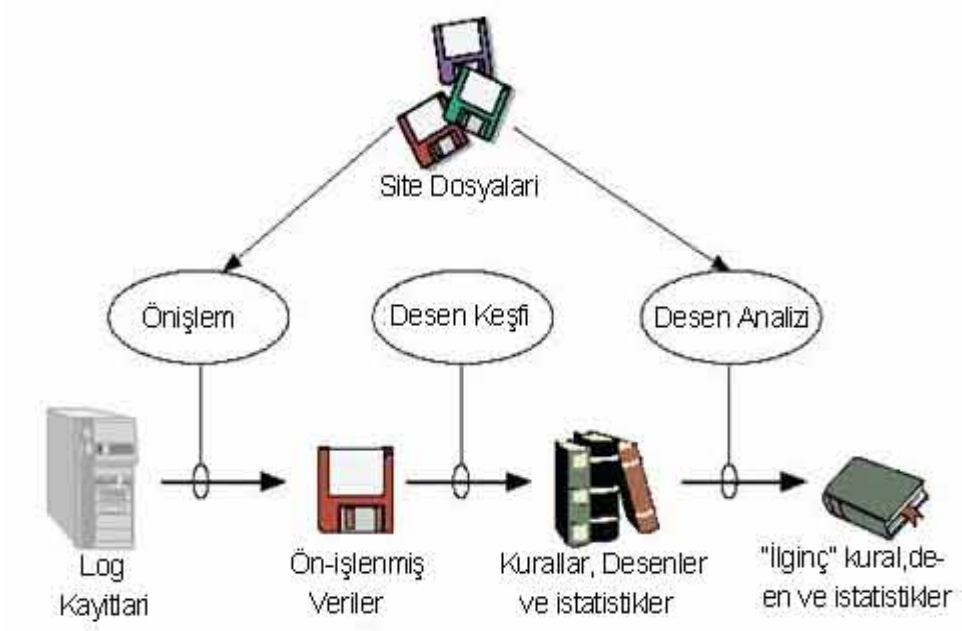
Günlük dosyalarının analizi, müşterilerin ilgi alanları, ürünler üzerinden pazar stratejileri oluşturma, promosyon kampanyalarının etkisi gibi hususlarda, kurumlara karar süreçlerinde yardımcı olur. Sunucu erişim kayıtlarının ve kullanıcı kaydı verilerinin analizi, aynı zamanda kurumun daha etkili bir sunumunun yapılabilmesi için Web sitesini nasıl daha iyi hale getirebileceği hakkında değerli bilgiler sağlar.

10.0.3.124	05/Dec/2005:14:03:27	GET	/~bdengiz/End-308.htm	HTTP/1.1	404	301	
65.55.246.42	05/Dec/2005:14:03:27	GET	/english/hastalan.php	HTTP/1.0	200	2988	
81.214.188.253	05/Dec/2005:14:03:28	GET	/~egfak/bolumler/ilko/snfo/index.html	HTTP/1.0	200	11292	
10.0.1.33	05/Dec/2005:14:03:28	POST	/~htinmaz/oto205d.php	HTTP/1.1	200	26180	
85.96.67.176	05/Dec/2005:14:03:29	GET	/~gkose/personal.php	HTTP/1.1	200	3851	
10.0.3.124	05/Dec/2005:14:03:29	GET	/~bdengiz/courses.html	HTTP/1.1	304	0	
10.0.3.124	05/Dec/2005:14:03:30	GET	/~bdengiz/End-421.htm	HTTP/1.1	304	0	

Şekil 4.1: Web günlük erişim dosyası

- Web kullanım madenciliği; Önişlem (Preprocessing), desen keşfi (Pattern Discovery) ve desen analizi (Pattern Analysis) aşamalarından oluşur (Şekil 4.2). Web kullanım madenciliği esnasında harmanlanacak veriler aşağıdaki tiplerde olabilir:
    - İçerik verisi: Web dokümanlarında, genellikle metin şeklinde yer alan verilerdir. Herhangi bir web sayfası üzerinde yer alan veriler bu tip için bir örnektir.

- Yapı verisi: Web sitesinin bağlantı yapısı hakkındaki verilerdir. Web sitesinde yer alan sayfaların hangi alt dizinler içerisinde bulunduğunu gösteren verilerden oluşur.
- Kullanım verisi: Web sitesini ziyaret eden kullanıcıların oluşturdukları veri tipidir. Kullanım verisi genellikle hangi kullanıcı, ne zaman, hangi sayfaları ziyaret etti, ne kadar süre sitede kaldı gibi soruların cevaplarını içerir.
- Kullanıcı profili: Web sitesini ziyaret eden kullanıcı hakkındaki; kullanıcı kimlik verileri gibi bilgilerden oluşur.



Şekil 4.2 Web Kullanım Madenciliği Süreci [13]

4.1 Web Günlük Erişimleri

Web üç tip veri barındırır: web içerik verileri, web kullanım verileri ve web yapı verileri. Cooley bu veri tiplerini içerik,yapı,kullanım ve kullanıcı profil verileri olarak sınıflandırmıştır[14]. Html dökümanları, imajlar,ses dosyaları vs. içerik verileridir. Linkler ve iç bağlantılar yapı verileri, Ip adresi, sayfa referansı, tarih ve zaman da kullanıcı profil verilerini gösterir.

4.2 Http Log Analizi

Web Kullanım Madenciliğinin temeli Web Sunucusunun kayıt dosyalarıdır (*web server logfiles*). Bizim çalışmamızda ihtiyacımız olan bilgiler Erişim kayıt dosyalarında (*access log files*) saklı olacaktır. Bu dosyaya ait bir satır ;

85.100.73.177- ahmet [21/May/2007:00:01:24 -0500] GET /index.php HTTP/1.1 200 3153 http://www.donanimhaber.com/index.asp Mozilla/4.0 (compatible; MSIE 6.0; Windows 98)
şeklindedir.

Görüldüğü gibi satır 9 bölümden oluşmaktadır. Web sunucuları üzerinde bulunan bu tür dosyalara Common Log Format (Şekil 4.3)veya Extended Logfile Format (*ELF*) (Şekil 4.4) adı verilmektedir.

```
123.456.78.9   [20/Jun/2000:15:13:05 +0300]''GET /courses.html HTTP/1.1 '' 304
123.456.78.9   [20/Jun/2000:15:13:05 +0300]''GET / will/courses/CS101/ HTTP/1.1'' 304
123.456.78.9 - - [20/Jun/2000:15:13:05 +0300]"GET / gif/geney.jpg HTTP/1.0 " 304 -
123.456.78.9 - - [20/Jun/2000: 15:13:05 +0300]" GET / gif/acad.gif HTTP/1.0 " 304 -
123.456.78.9 - - [20/Jun/2000:15:13:05| +0300]" GET / gif/ciz7.gif HTTP/1.0 "304 -
```

Şekil 4.3 Ortak Erişim Kütüğü Formatı (Common log format)

IP Address	Userid	Time	Method/ URL/ Protocol	Status	Size	Referrer	Agent
123.456.78.9	- -	[25/Apr/1998:03:04:41 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla/3.04 (Win95, I)
123.456.78.9	- -	[25/Apr/1998:03:05:34 -0500]	"GET B.html HTTP/1.0"	200	2050	A.html	Mozilla/3.04 (Win95, I)
123.456.78.9	- -	[25/Apr/1998:03:06:02 -0500]	"POST /cgi-bin/p1 HTTP/1.0"	200	5096	B.html	Mozilla/3.04 (Win95, I)
123.456.78.9	- -	[25/Apr/1998:03:06:58 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla (IE4.2, WinNT)
123.456.78.9	- -	[25/Apr/1998:03:07:42 -0500]	"GET B.html HTTP/1.0"	200	2050	A.html	Mozilla (IE4.2, WinNT)

Şekil 4.4 Genişletilmiş Erişim Kütüğü Formatı (Extended logfile format)

Çalışmamızın konusu olan kayıt dosyaları combined log format türünde olduğu için bundan böyle bu tarz kayıt dosyalarından bahsedeceğiz. Yukardaki örneğin her bölümünü inceleyecek olursak:

IP:

85.100.73.177- ahmet [21/May/2007:00:01:24 -0500] GET /index.php HTTP/1.1 200 3153 http://www.donanimhaber.com/index.asp Mozilla/4.0 (compatible; MSIE 6.0; Windows 98)

Bu alan istekte bulunan tarayıcının IP adresidir. Bazı durumlarda bu alanda gelen tarayıcının alan adı da bulunabilmektedir.

Identity:

85.100.73.177**-** ahmet [21/May/2007:00:01:24 -0500] GET /index.php HTTP/1.1 200 3153 http://www.donanimhaber.com/index.asp Mozilla/4.0 (compatible; MSIE 6.0; Windows 98)

Bu alanda çok özel tanımlamalar yapılmadığı sürece bilgi bulunmaz.

User:

85.100.73.177- **ahmet** [21/May/2007:00:01:24 -0500] GET /index.php HTTP/1.1 200 3153 http://www.donanimhaber.com/index.asp Mozilla/4.0 (compatible; MSIE 6.0; Windows 98)

Bu alan eğer sisteminiz parola korumalı ise ve HTTP kimlik denetlemesi sonucu onaylanmış ise ve parola doğru ise kullanıcı adı çıkacaktır. Bizim örneğimizde bu alan ahmet, ancak çoğu durumlarda burası bir önceki bölümdeki gibi – olacaktır.

Time:

85.100.73.177- ahmet **[21/May/2007:00:01:24 -0500]** GET /index.php HTTP/1.1 200 3153 http://www.donanimhaber.com/index.asp Mozilla/4.0 (compatible; MSIE 6.0; Windows 98)

Bu kısım isteğin yapıldığı saat ve tarihtir. Bizim bu örneğimizde tarih 01 ocak 2003 saat de GMT ye göre -5 zaman diliminde olan bir bölgede bulunan sunucu da gece 00:01 yi göstermektedir.

Request:

85.100.73.177- ahmet [21/May/2007:00:01:24 -0500] **GET /index.php HTTP/1.1** 200 3153 http://www.donanimhaber.com/index.asp Mozilla/4.0 (compatible; MSIE 6.0; Windows 98)

Bu kısım yapılmış olan isteğe ait ayrıntılı bilgi sunmaktadır. Bu bölümü kendi içerisinde gruplandırırsak ilk kısım istek türünü (bu örnekte GET (*Get ve Post*)) ikinci kısım istenilen dosyayı (burada index.php) ve son kısım protokolü vermektedir (burada HTTP/1.1).

Status:

85.100.73.177- ahmet [21/May/2007:00:01:24 -0500] GET /index.php HTTP/1.1 **200** 3153 http://www.donanimhaber.com/index.asp Mozilla/4.0 (compatible; MSIE 6.0; Windows 98)

Bu kısım sunucunun verdiği yanıt kodunu içermekte. Bizim örneğimizde olan '200' yanıtı hiç bir sorun oluşmadığını göstermektedir. Bu kodlar RFC2616 teknik belgesinin 10. kısmında belgelendirilmiştir. Ayrıca Tablo 2'de bu kodların bir kısmını görebiliriz.

100	Continue
101	Switching Protocols
200	OK
201	Created
202	Accepted
203	Non-Authoritative Information
204	No Content
205	Reset Content
206	Partial Content
300	Multiple Choices
301	Moved Permanently
302	Found
303	See Other

304	Not Modified
305	Use Proxy
306	(Unused)
307	Temporary Redirect
400	Bad Request
401	Unauthorized
402	Payment Required
403	Forbidden
404	Not Found
405	Method Not Allowed
406	Not Acceptable
407	Proxy Authentication Required
408	Request Timeout
409	Conflict
410	Gone
411	Length Required
412	Precondition Failed
413	Request Entity Too Large
414	Request-URI Too Long
415	Unsupported Media Type
416	Requested Range Not Satisfiable
417	Expectation Failed
500	Internal Server Error
501	Not Implemented
502	Bad Gateway
503	Service Unavailable
504	Gateway Timeout
505	HTTP Version Not Supported

Tablo 4.1**:** Sunucunun Verdiği Yanıt Kodlarından Örnek.

Burada genelikle 200-299 arasındaki kodlar başarı erişim durumunu, 300-399 arası bir uyarı durumunu,400-499 arasındaki kodlar ise istemci hatalarını, 500 ‘den sonrası sunucu hatalarını temsil etmektedir.

Size:

85.100.73.177- ahmet [21/May/2007:00:01:24 -0500] GET /index.php HTTP/1.1 200 **3153** http://www.donanimhaber.com/index.asp Mozilla/4.0 (compatible; MSIE 6.0; Windows 98)

Gönderilmiş olan dosyanın Byte cinsinden boyutu.

Referer:

85.100.73.177- ahmet [21/May/2007:00:01:24 -0500] GET /index.php HTTP/1.1 200 3153 **http://www.donanimhaber.com/index.asp** Mozilla/4.0 (compatible; MSIE 6.0; Windows 98)

İsteğe nerden ulaşıldığı konusunda bilgi verir. Böylelikle bu istek yapılmadan önce nereden gelindiği görülebilir.

User-agent:

85.100.73.177- ahmet [21/May/2007:00:01:24 -0500] GET /index.php HTTP/1.1 200 3153 http://www.donanimhaber.com/index.asp **Mozilla/4.0 (compatible; MSIE 6.0; Windows 98)**

İstekte bulunan tarayıcı bilgileri.

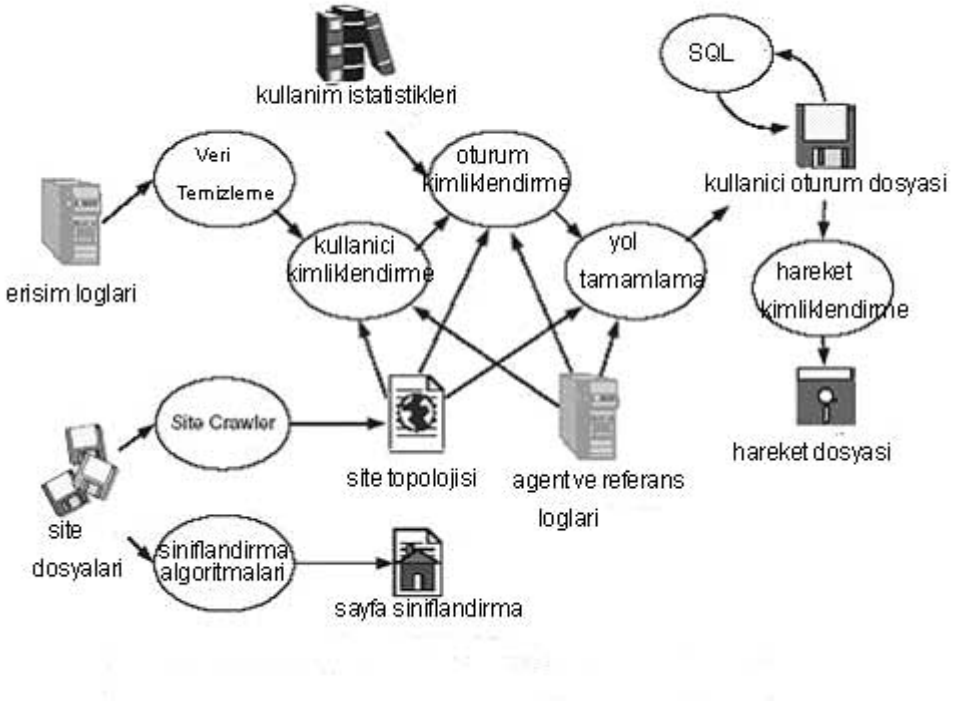
4.3 Web Kullanım Madenciliği Aşamaları

1. Ön İşlem

- Ön İşlem (Preprocessing): Ön işlem web kullanım madenciliğinin ilk aşamasıdır. Ham veri bir takım işlemlerden geçirilerek soyutlaştırılır ve desen keşfi (Pattern Discovery) için hazır hale getirilir. Soyutlaştırma bir çeşit istatistiksel özet çıkarmadır ve kullanıcı, sayfa görünümü, tıklama akışı,

kullanıcı oturumu, sunucu oturumu gibi çeşitleri olabilmektedir. Genel olarak yapılan ön işlemler aşağıdaki gibi olup Şekil 6'da görümektedir.

- Veri Ayrıştırma (Data Cleaning): Kayıt dosyalarından, gereksiz ve ilişkisiz veriler çıkarılır.
- Kullanıcı Kimliği (User Identification): Birçok kişi internete çıkışını tek bir internet adresi üzerinden gerçekleştirir. Bu nedenler çeşitli yöntemler kullanılarak (çerezler, kullanıcı girişi vb.) kişiler web kayıt dosyaları üzerinde tespit edilir.
- Oturum Kimliği (Session Identification): Kullanıcının web üzerinde yaptığı sayfa görüntülemeleri oturumlara bölünür.
- Yol Tamamlama (Path Completion): Web tarayıcının ön belleği veya kullanıcının kullandığı Proxy server'dan dolayı kayıt dışı kalan bağlantılar tamamlanır.



Şekil 4.5 Web kullanım Madenciliğinde Ön İşlem Aşaması [13]

2. Desen Keşfi

Desen keşfi, ön işlemden geçirilen verilere veri madenciliği tekniklerinin uygulandığı

aşamadır. En sık kullanılan bazı veri madenciliği yöntemleri; istatistiksel yöntemler,

eşleştirme kuralları, kümeleme, sınıflandırma ve sıralı örüntülerdir [6]. Bu tekniklere kısaca göz atmak gerekirse:

İstatistiksel Yöntemler:

İstatistiksel teknikler bir web sitesinin ziyaretçileri hakkında bilgi açığa çıkarmaya yarayan en güçlü araçlardır. Analizciler oturum dosyasını analiz ederken farklı değişkenler üzerinde farklı açıklamalı istatistiksel analiz tiplerini yerine getirirler. Periyodik web sistem raporlarında bulunan istatistiksel bilgi analiz edilerek sistem performansını artırıcı, sistem güvenliğini genişletici, düzeltme işlemlerini kolaylaştırıcı ve pazarlama kararlarını destekleyici raporlar.

Eşleştirme Kuralları:

Web etki alanında sıklıkla birbirini referans gösteren sayfalar eşleştirme kuralı üretimi uygulanarak tek bir sunucu oturumu şeklinde düzenlenebilir. Eşleştirme teknikleri bir işletimsel veri tabanında bulunan değerler arasındaki sıralı olmayan ilinti keşfinde kullanılır.

Kümeleme Analizi:

Kümeleme analizi, kullanıcıları veya sayfaları benzer özelliklerine göre birlikte gruplara ayırır. Kullanıcının veya sayfaların kümelenmesi geliştirme ve gelecek pazarlama stratejilerinin çalıştırılmasını kolaylaştırabilir (Cooley ve diğerleri, [8]). Kullanıcıların kümelenmesi benzer navigasyon örüntüsüne sahip kullanıcı gruplarını keşfetmede yardımcı olacaktır. Elektronik ticaret uygulamalarında müşterilere özel hizmet sunabilmek için gerekli olan pazar bölümlemesi kümeleme sayesinde yerine getirilebilmektedir. İlgili içeriğe sahip sayfa gruplarının keşfinde kullanılabilen sayfaların kümelenmesi, arama motorları ve web servis sağlayıcıları için de yararlı olmaktadır.

Sınıflandırma Tekniği:

Sınıflandırma bir veriyi daha önceden tanımlanmış sınıflara dağıtma tekniğidir. Web etki alanında, webmaster veya pazarlamacı sınıflandırma tekniğini kullanarak müşterilerinin hangi sınıf veya kategoride bir profile sahip olduğunu belirleyebilir. Sınıflandırma işleminde, verilen bir sınıf veya kategorinin özelliklerini en iyi biçimde

açıklamak için seçim ve açığa çıkarma uygulamalarına ihtiyaç duyulur. Sınıflandırma; karar ağaçları, bayezian sınıflayıcıları, en yakın komşu ve destek vektör makineleri gibi denetlenen tümevarımsal öğrenim algoritmaları kullanılarak yapılabilir (Cooley ve diğerleri, [8]).

Sıralı Desenler:

Sıralı Desenler; oturumlar arasında örüntü bulmaya çalışır. Sıralı örüntü bulma işleminde, belirli zaman aralıklarında oturumlar incelenir ve karşılaştırmalar yapılır. Sıralı örüntülerin bulunması gelecekteki eğilimi tahmin edecek web pazarlamacıları için oldukça anlamlıdır. Böylece ilanlar belirli kullanıcı gruplarına yönlendirilebilecektir. Sıralı örüntüler için, eğilim analizi, değişen nokta bulma veya benzerlik analizleri gibi bazı geçici analiz tipleri kullanır (Cooley ve diğerleri, [8]).

3. Desen Analizi

Desen keşfi aşamasında ortaya çıkarılan kural veya örüntülerin analiz edilmesi işlemidir. Bilgi sorgulama ve OLAP (OnLine Analytical Processing-çevrim içi analitik işlem) uygulamaları ile derinlemesine analizler yapılabilmektedir (Zaiane ve diğerleri, [7]). Aşağıda bazı örüntü analiz seçenekleri bulunmaktadır:

Görselleştirme teknikleri:

Desen keşif aşamasında elde edilen sonuçların (özetler gibi) anlaşılabilmesi için görselleştirme tekniklerinden faydalanılır. Görselleştirmede daireler, düğümler ve kenarlar kullanılır.[17]

OLAP teknikleri:

OLAP, iş ortamında veri tabanlarının stratejik analizi için çok güçlü bir uygulama alanı olarak ortaya çıkmıştır. Stratejik analizin bazı önemli özellikleri şunlardır:

1) Çok büyük boyutlu veri.
2) Geçici boyutlar için açık destek.
3) Çeşitli bilgi tipleri için destek sağlama.
4) Uzun-sıra analizi, ki orada toplam trendler bireysel veri elamanlarından daha önemlidir. OLAP doğrudan ilişkisel veri tabanları üzerinde çalışabilir. OLAP kullanılırken analiz için veri küplerinden faydalanılır. [17]

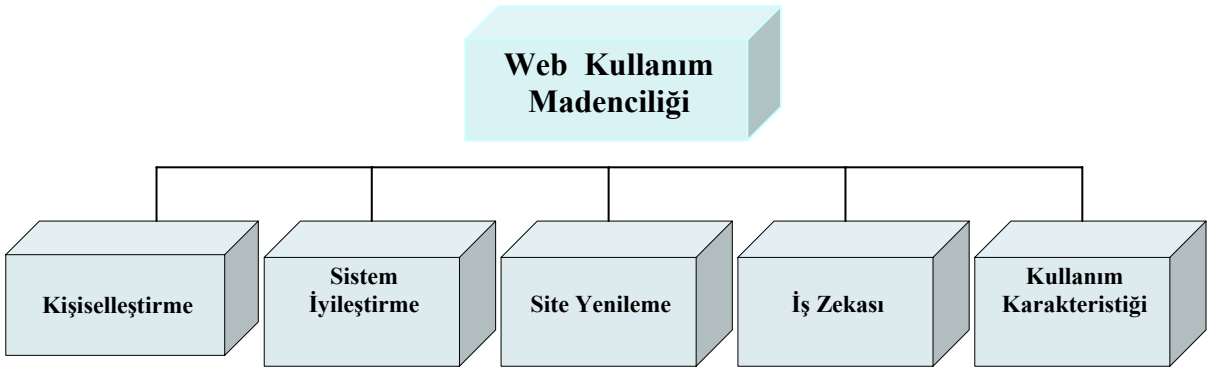
Veri ve Bilgi Sorgulama:

Veri ve bilgi sorgulama için iki yol bulunmaktadır. Birincisinde bildiriler şeklinde bir dil kullanılarak veri elde edilirken, ikincisinde SQL'e (Structured Query Language-Yapısal Sorgu Dili) benzeyen diller kullanılarak bilgi sorgulanabilir.

Kullanılabilirlik Analizi:

Bulunan veya ortaya konulan çözümlerin başarılı sonuçlar verebilmesi için kullanılabilir olmaları gerekmektedir. Veri analizlerinde de takip edilen yöntemin başarısı kullanılabilirlik analizleri ile yerine getirilir. Bu konunun şu an hedefi, kullanılabilirlik için sistematik bir yaklaşım geliştirme çabasıdır. İlk adımda, yazılım kullanılabilirliği için geliştirme metotları bir araya toplanır. Veri, hesaplanmış modeller oluşturmada kullanılır. Son olarak, çeşitli veri sunum ve görselleştirme teknikleri ile verinin anlaşılması sağlanır. Bu şekilde web kullanıcılarının davranışları bir model ile anlaşılabilir halde gösterilebilir.

4.4 Web Kullanım Madenciliği Uygulama Alanları



Şekil 4.6: Web Kullanım Madenciliği Uygulama Alanları

Kişiselleştirme:

Web sayfalarının kullanıcılara hitap etmesini sağlayacak şekilde, web sitesinin kişiye özel olarak tasarlanması.

Sistem İyileştirme :

Web madenciliği ile ortaya çıkan eksikler göz önünde bulundurularak yazılım ve donanımsal olarak sitemin ve diğer bileşenlerinin güçlendirilmesi.

Site Değiştirme/Güncelleme:

Siteyi ortaya çıkan analizler doğrultusunda yeniden tasarlamak, güncelleştirmeler yapmak ve site içeriğinin değiştirilmesi ile geliştirilmesini sağlamak.

İş Zekası:

Veri madenciliği teknikleri ile ortaya çıkan keşfedilmemiş bilgileri kullanarak optimizasyon ve pazarlama için yeni işler geliştirmek.

Kullanım Karakteristiği:

Web arayüzü içerisinde kullanıcı etkileşimlerine ve browserın nasıl kullanılacağına yardımcı olur.

5. WWW.BASKENT.EDU.TR LOG ANALİZİ

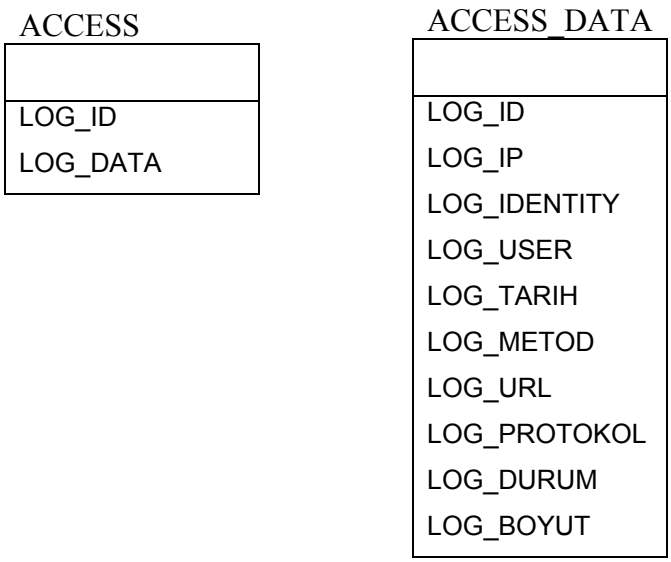
Bu Çalışmamızda Başkent Üniversitesi web sayfaları Web Kullanım Madenciliği ile analiz edilmiştir. Bunun için web madenciliğinin Apriori algoritması ile en sık ziyaret edilen url gruplarını(ikili, üçlü, dörtlü), Microsoft Clustering algoritması ile verideki kümelenmeyi ve Microsoft Decision Tree algoritması ile veri içindeki sınıflandırma bilgisini (url, ip, gün, durum kodu, boyut) veren WEKULA adını verdiğimiz, java ortamında geliştirilmiş bir uygulama ve serbest kullanıma sunulmuş olan WUM (*Web Utilization Miner*) , WUMprep [30] ile Weka [65] yazılımları kullanılmıştır.

Başkent Üniversitesi web sunucularında bulunan ve 01/02/2007 ile 29/05/2007 aralığında sistemin üretmiş olduğu web kayıt dosyaları alınmıştır. Web günlük erişim dosyası ayrıntılı bilgisi Tablo 5.1'de verilmiştir.

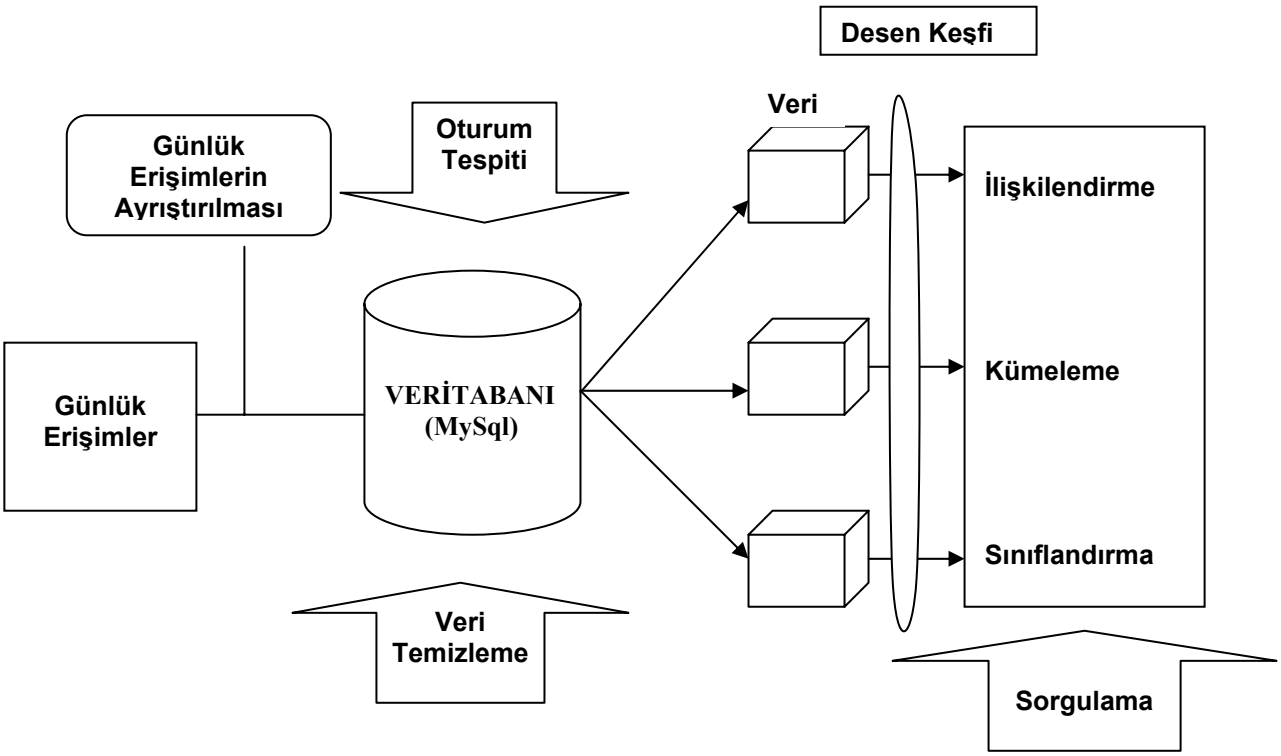
Günlük erişim dosyası bilgilieri	
Dosya adı	www.baskent.edu.tr-access_log-02_05
Boyut(MB)	2 462,61
Periyot	01 Şubat 2007- 29 Mayıs 2007
Erişim sayısı	27 381 257

Tablo 5.1 Günlük erişim dosyası bilgileri

Web günlük dosyasındaki veriler Şekil 5.2'de ver gösterilen WEKULA uygulamasıyla ACCESS adlı veritabanına aktarılır. Bundan sonra ise WUMprep adlı uygulamayıda kullanarak, veri ayrıştırması yapılarak web kayıt dosyasında bulunan gereksiz olan veriler uzaklaştırılır. Böylelikle web kayıt dosyası içerisinde bulunan tüm çoklu ortamlar ve çıkarılması istenilen diğer uzantılı (.ico,#,.css,.js,JPG,GIF) satırlar temizlenmiş olur. Bu aynı zamanda Internet'teki yavaşlık nedeni ile oluşabilecek fazla tıklarıda temizlemektedir. Elde edilen verilerden ziyaretçiler oturumlara ayrıştırılır. Tüm bu işlemlerin sonunda istenmeyen bilgilerden arındırılmış olan yeni bir web kayıt dosyası oluşmuş olur. Oluşan bu dosyadaki veriler de ACCESS_DATA adlı veritabanına Şekil 5.1 'de gösterildiği gibi bileşenlere ayrıştırılarak kaydedilir.

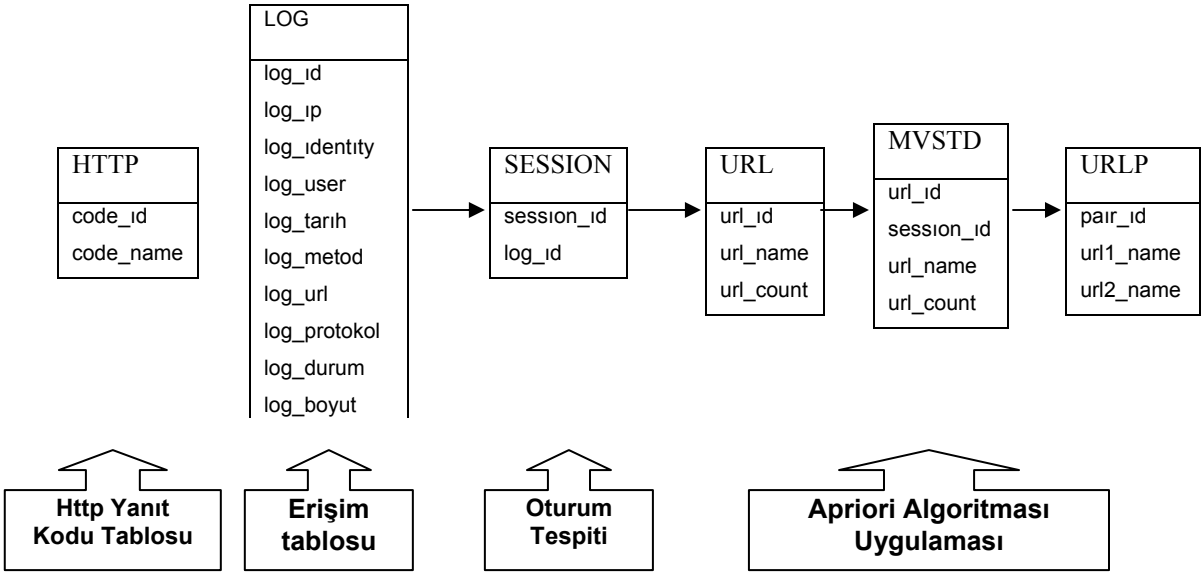


Şekil 5.1 WEKULA Veritabanı



Şekil 5.2 WEKULA mimarisi

Bu yeni dosya WEKULA ve WUM uygulaması ile analiz yapılmaya hazır hale getirilmiş olur. Ayrıca bu bilgiler kullanılarak Şekil 5.3 'te verilen veritabanı sistemi oluşturulur ve Apriori algoritmasının uygulanma aşamasında kullanılır.



Şekil 5.3 Apriori algoritmasının uygulanış adımları

Sitemizi ziyaret eden kullanıcının site içersinde durabileceği süreyi Spiliopoulou ve ekibinin [28] önerileri doğrultusunda 30 dakika olarak ayarladıktan sonra analizi başlattık.

Genel İstatistikler:

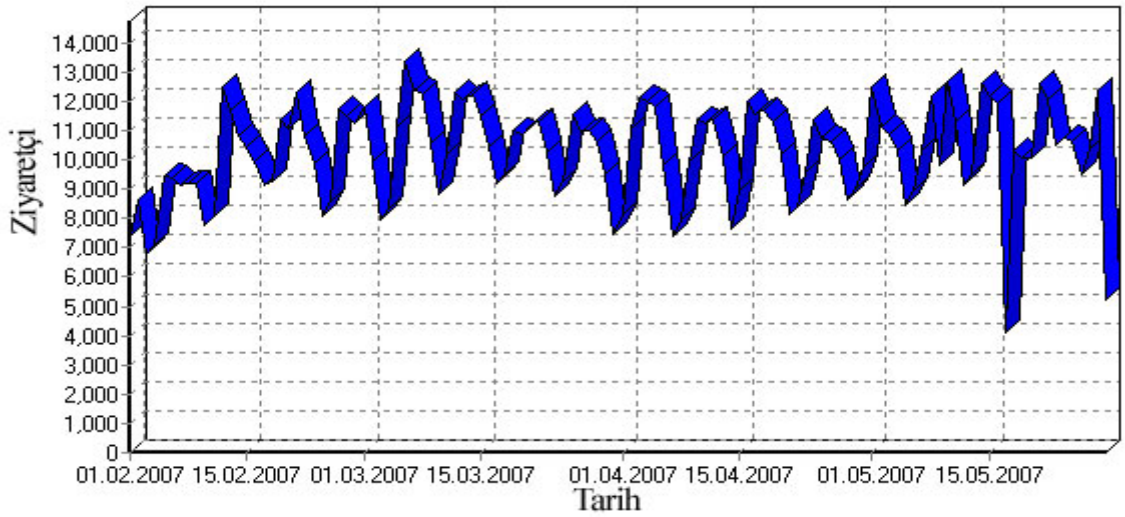
Tablo 5.2 ‘de site üzerinde verilen periyottaki erişim istatistikleri, ziyaretçi istatistikleri ve sayfa görüntüleme istatistikleri verilmişltir.

Erişimler	
Toplam Erişim	27,381,257
Günlük Ortalama Erişim	232,044
Ziyaretçi başına Ortalama Erişim	22.69
İstek sayısı	10,524,440
Başarısız İstek sayısı	954,457
Sayfa Görüntüleme	
Toplam Sayfa Görüntüleme	5,512,074
Günlük Ortalama Sayfa Görüntüleme	46,712
Ziyaretçi Başına Ortalama Sayfa Görüntüleme	4.57
Ziyaretçiler	
Toplam Ziyaretçi	1,206,710
Ortalama Günlük Ziyaretçi	10,226

Tablo 5.2 Genel site istatistikleri

Günlük Ziyaretçi

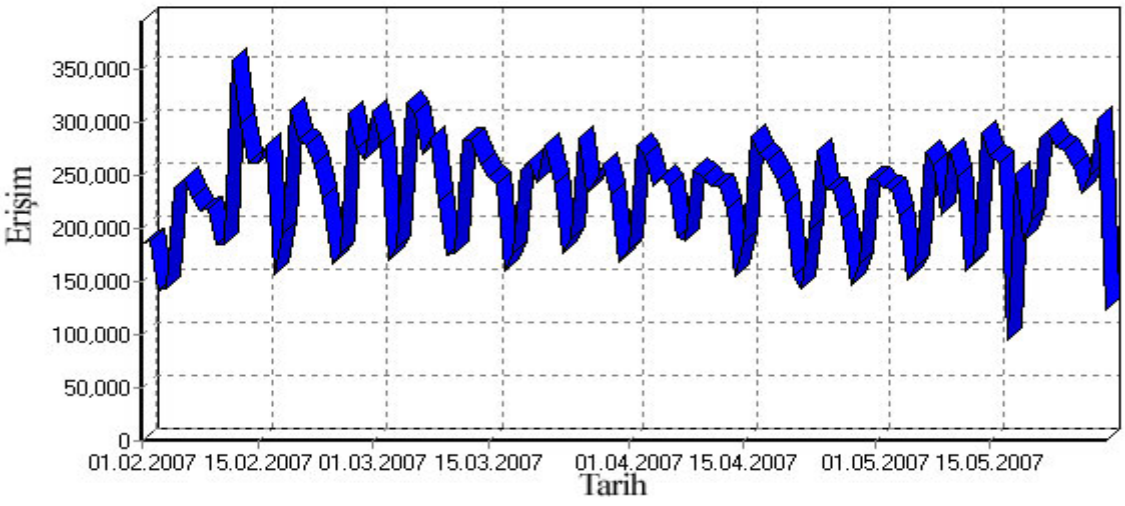
Kullanıcıların ele alınan periyotta günlük olarak siteyi ziyaret sayıları Şekil 5.4 de grafiksel olarak gösterilmiştir. Burada site ziyaretinin en çok ve en az olduğu aralığı görmemiz mümkündür.



Şekil 5.4 Günlük ziyaretçi çizelgesi

Günlük Erişim(Hit)

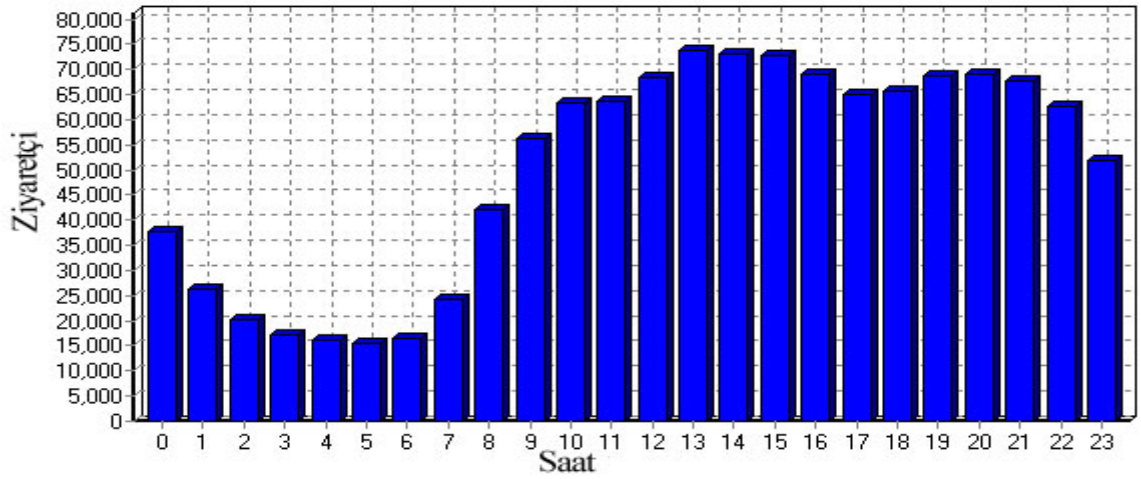
Siteyi ziyaret eden kullanıcıların günlük erişimleri Şekil 5.5'de grafiksel olarak gösterilmiştir. Hangi aralıklarda erişimlerin maximum ve minimum olduğu bu grafikten elde edilebilir.



Şekil 5.5 Günlük erişim çizelgesi

Günün Saatlerine Göre Ziyaretçi

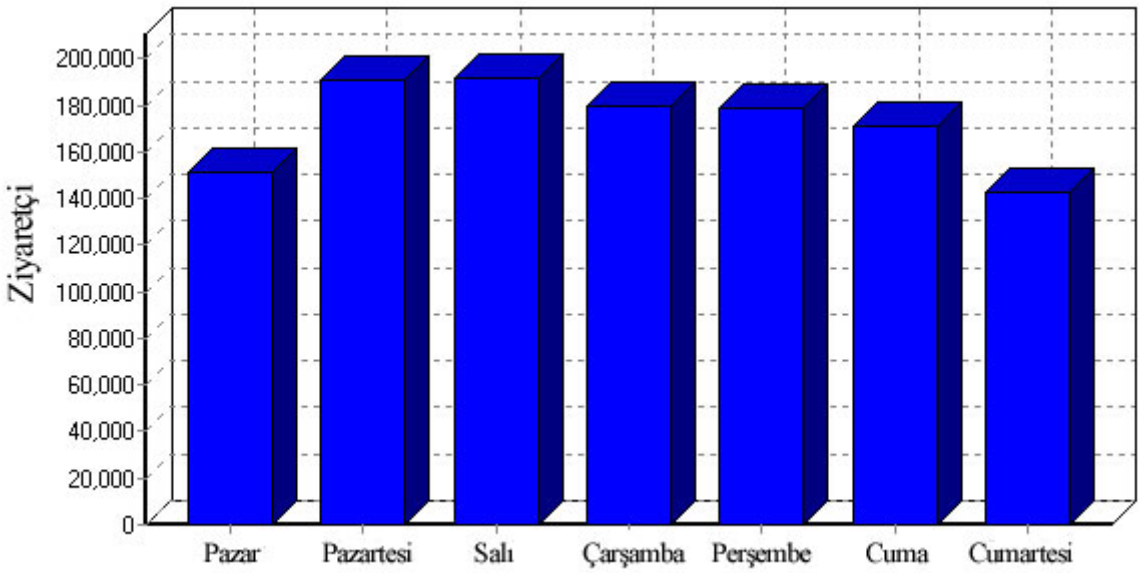
Ziyaretçilerin günün en çok ve en az hangi saatlerinde siteyi ziyaret ettikleri Şekil 5.6 de gösterilmiştir.



Şekil 5.6 Günün saatlerine göre ziyaretçi çizelgesi

Haftanın Günlerine Göre Ziyaretçi

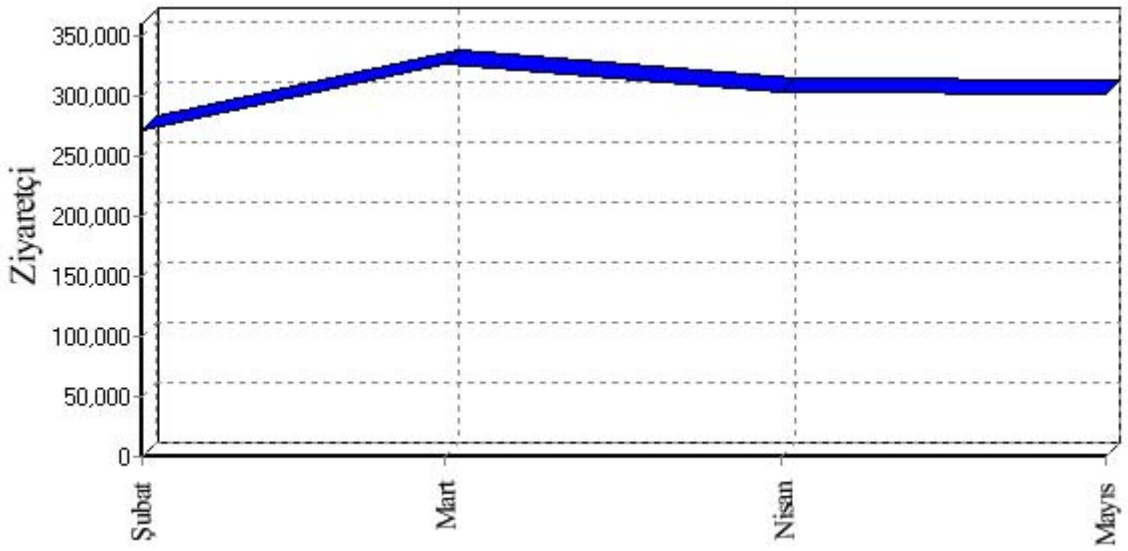
Ziyaretçilerin haftanın en çok hangi günlerinde siteyi ziyaret ettikleri Şekil 5.7 de gösterilmiştir.



Şekil 5.7 Haftanın günlerine göre ziyaretçi çizelgesi

Aylara Göre Ziyaretçi

Ziyaretçilerin hangi aylarda siteyi ne kadar ziyaret ettikleri Şekil 5.8 de gösterilmiştir.



Şekil 5.8Aylara göre ziyaretçi çizelgesi

En Çok Ziyaret Edilen Sayfalar

Kulanıcıların web sitesinde en çok ziyaret ettikleri sayfalar ve bu sayfalara erşim miktarları verilmiştir. Burada sayfalardan yapılan isteklerin ne kadarının tamamladığı bilgisine de ulaşılabilmektedir. Tablo 5.3'de en çok ziyaret edilen 20 sayfa gösterilmektedir

	Sayfa	Erişim	Tamam-lanmamiş İstek	Ziyaretçi
1	http://www.baskent.edu.tr/	808,473	0	413,074
2	**http://www.baskent.edu.tr/~ihaberal/uyebilgileri.php**	14,698	0	12,991
3	**http://www.baskent.edu.tr/goruntuler/**	15,343	0	12,574

4	http://www.baskent.edu.tr/~gurayk/	14,537	0	12,203
5	http://www.baskent.edu.tr/tip/	38,279	0	11,917
6	http://www.baskent.edu.tr/~scanan/sesweb/ses.htm	15,108	121	11,855
7	http://www.baskent.edu.tr/aday/	11,684	0	9,659
8	http://www.baskent.edu.tr/tanitim/	9,125	4	7,844
9	**http://www.baskent.edu.tr/~ihaberal/ziyaretci-genis.php**	7,777	0	7,161
10	http://www.baskent.edu.tr/genel.php	6,753	0	5,796
11	http://www.baskent.edu.tr/~fpakdil/dersler.php	13,599	0	5,693
12	http://www.baskent.edu.tr/~htinmaz/	12,006	1	5,435
13	http://www.baskent.edu.tr/~eminec/moodle/help.php	5,418	0	5,331
14	http://www.baskent.edu.tr/iletisim.php	5,969	0	5,301
15	http://www.baskent.edu.tr/akademik_birimler.php	6,381	0	5,285
16	http://www.baskent.edu.tr/~htinmaz/duyuruliste.html	6,963	64	5,252
17	http://www.baskent.edu.tr/~htinmaz/index-tr.html	6,790	2	5,167
18	http://www.baskent.edu.tr/~kilter/	13,017	2	5,033
19	http://www.baskent.edu.tr/ulasim/	5,754	0	5,012
20	http://www.baskent.edu.tr/~msert/	11,360	0	4,934

Tablo 5.3 Ençok ziyaret edilen sayfalar

Encok İndirilen Dosyalar

Ziyaret edilen sayfalardan kullanıcıların kendi bilgisayarlarına indirdikleri veya açtıkları dosyalar, erişim miktarları ile birlikte elde edilmiştir. Tablo 5.4'de en çok incelenen 20 dosya erişim miktarlarıyla birlikte gösterilmektedir.

	Dosya	Erişim	Ziyaretçi
1	http://www.baskent.edu.tr/tip/3a.pdf	134,631	5,107
2	http://www.baskent.edu.tr/aday/Etkili_Ogrenme_Yontemleri.pdf	27,612	4,112
3	http://www.baskent.edu.tr/~omadran/eskiweb/donem0304/ dersnotu0304/hafta1.pdf	4,562	2,530
4	http://www.baskent.edu.tr/~mustafak/BME-201/dokumanlar/ HÜCRE VE YAPISIdoc.pdf	7,885	1,788
5	http://www.baskent.edu.tr/~scanan/dersler2005/kurul2/01_kan2005.pdf	14,186	1,630
6	http://www.baskent.edu.tr/~scanan/dersler2005/01_cizgili_kas_sgulen.pdf	14,587	1,326
7	http://www.baskent.edu.tr/aday/SinavKaygisi.pdf	6,349	1,315
8	http://www.baskent.edu.tr/~scanan/dersler2005/kurul2/05_koagulasyon2005.pdf	11,113	1,314
9	http://www.baskent.edu.tr/~mustafak/BME-201/dokumanlar/ BAĞIŞIKLIK SİSTEMİdoc.pdf	2,297	1,305
10	http://www.baskent.edu.tr/~mustafak/BME-201/dokumanlar/ kanser doc.pdf	2,276	1,215
11	http://www.baskent.edu.tr/duyurular/301/wgfes.pdf	10,512	1,106
12	http://www.baskent.edu.tr/~scanan/dersler2005/kurul2/02_eritrositler2005.pdf	12,703	1,075
13	http://www.baskent.edu.tr/~scanan/dersler2005/SB18_dolasim1_2005.pdf	2,550	1,033
14	http://www.baskent.edu.tr/duyurular/301/gbiss.pdf	3,011	943
15	http://www.baskent.edu.tr/~scanan/dersler2005/kurul2/	6,598	940

	07_kangrup2005.pdf		
16	http://www.baskent.edu.tr/~scanan/dersler2005/SB21_dolasim4_2005.pdf	2,084	849
17	http://www.baskent.edu.tr/~mustafak/BME-201/dokumanlar/ KLONLAMAdoc.pdf	4,028	837
18	http://www.baskent.edu.tr/~scanan/dersler2005/SB26_endokrin2_2005.pdf	17,971	812
19	http://www.baskent.edu.tr/~scanan/dersler2005/SB22_bobrek_2005.pdf	10,690	772
20	http://www.baskent.edu.tr/~scanan/dersler2005/kurul2/06_antikoagulasyon2005.pdf	4,749	767

Tablo 5.4 En çok indirilen dosyalar

En Çok Ziyaret Eden Ülkeler

www.baskent.edu.tr web sitesini ziyaret eden ülkeler erişim miktarlarıyla birlikte elde belirlenmiştir. Burada siteye Türkiye sınırları dışından en fazla United States'dan ziyaret edildiği görülmektedir. Tablo 5.5' te web sitesini en çok ziyeret 20 ülke verilmiştir.

	Ülke	Erişim	Ziyaretçi	%
1	Turkey	21,786,824	754,133	62.49%
2	United States	636,910	307,370	25.47%
3	Unknown	4,628,180	108,357	8.98%
4	Germany	68,149	5,414	0.45%
5	China	12,231	3,858	0.32%
6	France	17,005	3,056	0.25%
7	United Kingdom	29,246	1,933	0.16%
8	Sweden	6,415	1,897	0.16%
9	Brazil	3,017	1,529	0.13%
10	Netherlands	18,918	1,501	0.12%
11	India	9,259	1,294	0.11%
12	Malaysia	4,136	1,158	0.10%
13	Korea, Republic of	3,160	935	0.08%
14	Japan	10,339	926	0.08%
15	Canada	9,905	841	0.07%
16	Russian Federation	7,564	758	0.06%
17	Azerbaijan	9,031	561	0.05%
18	Austria	6,084	547	0.05%
19	Switzerland	6,300	472	0.04%
20	Taiwan	2,287	461	0.04%

Tablo 5.5 En çok ziyaret eden ülkeler.

6 . SONUÇ VE ÖNERİLER

Veri madenciliğinin, bilgi sistemlerinin zeki davranışlar göstermesinde büyük bir etkisi vardır. Bu etki elektronik cihazların zeki davranışlar göstermesinde sensörlerin etkisi gibi düşünülebilir. Veri madenciliği içine girdiği bilgi sistemlerini zeki hale getirip, zekaya ihtiyacı olan ve her geçen gün büyüyen webe de uygulanmış ve ortaya web madenciliği kavramını çıkarmıştır.

Web kullanıcılarını gezindikleri sayfalardan tanıyan bu yeni teknik başta elektronik ticaret olmak üzere bir çok konuda kullanılmaktadır. Bu teknik sayesinde kullanıcılar tanınabilmekte ve onların memnun olacağı hizmetin verilmesi mümkün hale gelmektedir. Tekniğin üniversite web sitelerinde kullanılması da mümkün olup bu sayede üniversite hizmetlerinin kalitesi artırılabilmekte ve üniversite web sitelerinde yeniden tasarımlar ile en iyi üniversite web siteleri ortaya çıkabilmektedir.

Web sunucu günlük erişim dosyalarını ham veri kaynağı olarak kullanarak, verilecek tarih aralıklarındaki veriler işlenip, site hakkında yeni kararlar alabilmeye yardımcı olacak birçok grafiksel çıktı üretilmiştir.

Web günlük erişim dosyalarının ayrıştırılması ve analiz edilmesi değerli bilgi sağlamaktadır. Günlük dosyası analizleri sayesinde hedef kitleye ve özel kullanıcı grublarına (kümeler) hizmet verilebilmektedir.

Web Sayfaları bir kurumun dışarıya açılan penceresidir. Bilinçsiz şekilde ve W3C standartlarına uyulmadan hazırlanan web sayfaları, siteye olan ziyaretçi sayısını olumsuz şekilde etkileyeceği gibi, gelen ziyaretçilerin aramış oldukları bilgilere ulaşmalarında da zorluklar çıkaracaktır. Bu çalışmada başlıca hedefimiz Başkent Üniversitesi web sayfasının kullanım durumunun ortaya konması, gelen ziyaretçilerin kullanımları doğrultusunda kısa yolların oluşturulmasını sağlamaktı.

www.baskent.edu.tr web sitesinde 5 farklı kullanıcı tipi tespit edilmiştir:

- Akademisyenler, akademik çalışmaları takip ve yararlanmak için bilimsel makale inceleme ve indirme,
- Ziyaretçiler, site ziyaretlerinde akademik birimlerin ve öğretim elemanlarının kişisel sayfalarını görüntüleme,
- Öğrenciler, derslerin web sitelerini ziyaret ve ders notu indirme.
- Ajanlar, websitesi hakkında bilgi toplamak,arama makineleri için bilgi almak.
- Yayıncılar, dökümanlarını yükledikten sonra sitelerini yeniden kontrol etmek

Bu kullanıcı erişimlerinden yola çıkarak Başkent Üniversitesi'nin 01/02/2007 ile 29/05/2007 arasındaki 118 günlük veri incelenmiştir. Bu süre içerisinde 1206710 IP'den toplam olarak 5512074 sayfa izlenmiş, bu da günlük ortalama olarak 46712.49 sayfaya karşılık gelmektedir. Ziyaretçi başına ortalama olarak 16.80 MB veri transferi gerçekleşmiştir. Günlük olarak 10226.35 IP ortalama olarak 5.34 dakika sitede kalmaktadir. Bu süreçte sitemde bulunan 21396 değişik web sayfası izlenmiştir. Sisteme en çok pazartesi ve salı günleri, en çok 13 ve 14 saatlerinde erişim olmuştur.
WMV dosya tipi 74442 erişim ile trafiğin %66'sını, JPG dosya tipi ise 10891000 erişim ile trafiğin %0,42'sini oluşturmuştur , bu da bize WMV dosya tipine ulaşım az olmasına rağmen, dosya boyutlarının büyük olduğunu göstermektedir. İlginç olan ise html/htm dosya tipinin %0,04 gibi çok düşük bir trafik oluşturmasıdır.

Sayfalar içerisinde fazla kullanılmayan sayfalar, hangi ülkelerden ziyaretçilerin hangi sayfaları ziyaret ettiği, hangi dökümanların incelendiği tespit edilmiştir. Buna göre verilen tariharalığında örnek olarak 3a.pdf, Etkili_Ogrenme_Yontemleri.pdf ve SinavKaygisi.pdf dosyaları ziyaretçilerin en çok ilgisini çeken dosyalar olduğu tespit edilmiştir.

Dosya istem kodları (durum kodları) incelendiğinde en büyük oranın 200 numaralı kod olduğu, 304 numaralı kodun ise 373024 kez tekrarlandığı, yani aynı dosyanın

değişiklik olmadığı halde tekrar tekrar istenmesi olduğu görülmekte, bu tekrar istemlerin kötü amaç taşıma olasılığını ortaya çıkarmaktadır.

Bundan sonraki aşamada bu verilerden faydalanılarak web sitesinin yeniden tasarlanması veya mevcut sitenin geliştirilmesi sağlanılabilir. Ziyaretçilerin site üzerinde daha uzun süre gezinmeleri etkin kılınabilir. Başkent Üniverstisi web anasayfasından öğrenci veya personelin kişisel hesaplarına direk erişimi sağlayacak bir modulun eklenmesi ile kullanım hem kolaylaşacak hemde erişimde artışın olduğu gözlenecektir. Video ağırlıklı dosyalar trafiği çok fazla kullandığı için bu dosyaların daha az kullanılması önerilebilir.

Bunun yanında veri madenciliği işlemine tabi tutulan verilerden yola çıkarak kullanıcılar kümelendirilebilir (gruplandırılabilir), bu sayede kullanıcılara özel kişiselleştirilmiş sayfalar yapılabilmektedir. Ayrıca kullanıcının hareketlerinden yola çıkarak, bir sonraki hareketinin (hangi sayfaya girebileceğinin) ne olabileceğini tahmin etmeyi sağlayacak bir yapay sınır ağı modeli tasarlanabilir.

KAYNAKLAR LİSTESİ

[1] O. Etzioni, The World Wide Web: Quagmire or gold mine. Communications of the ACM, 39(11):65-68, (1996)

[2] S.K.Madria, S.S.Bhowmick, W.K.Ng, and E.P.Lim, Research issues in Web data mining. In Proceedings of Data Warehousing and Knowledge Discovery, First International Conference, DaWaK '99, sayfa 303-312 , (1999)

[3] R. Cooley, Web Usage Mining: Discovery and Application of Interesting Patterns from Web data. Ph.D. thesis, Dept. of Computer Science, University of Minnesota, (2000)

[4] R. Agrawal and A. Srikant, Fast algorithms for mining association rules. Proc. VLDB'94, sayfa 487-499, (1994)

[5] B. Özakar, Finding and evaluating patterns in Web Repositories using data mining algorithms and database technologies, Master Tezi, 2002, İzmir Yüksek Teknoloji Enstitüsü Bilgisayar Mühendisliği Bölümü

[6] R. Cooley, B.Mobasher, and J.Seivasta. Data preparation for mining world wide web browsing patterns. Knowledge and Information Systems,1(1), 1999.

[7] O. R. Zaiane, M. Xin, J. Han, "Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs", Proc. Advances in Digital Libraries Conf. (ADL'98), Santa Barbara, CA, April 1998.

[8] Cooley, R. Mobasher, B. and Srivastave, J. Web Mining: Information and Pattern Discovery on the World Wide Web In Proceedings of the 9th IEEE International Conference on Tool with Artificial Intelligence, 1997.

[9] S.Gunduz, M.T.Ozsu, Recommendation model for user accesses to web pages, ICANN 2003

[10] Y.F.Kanwalpreet, S.M. Shih, Clustering of Web Users Based on Access Patterns.

[12] H.Sever,B.Oguz, Veri Tabanlarında Bilgi Keşfine Formel Bir Yaklaşım

[13] R.Ayaz, Web Madenciliğine Bir Bakış

[14] R. Cooley,P. Tan, J. Srivastava. Discovery of Interesting usage patterns from web data. In Myra Spiliopoulou,editor,LNCS/LNAI Series.Springer-Verlag,(2000)

[15] A. Vahaplar, M.M. İnceoğlu: Veri Madenciliği ve Elektronik Ticaret, VII. Türkiye'de İnternet Konferansı, 1–3 Kasım 2001

[16] B. Özakar, H. Püskülcü, 'Tiklarin Dili', Internet Konferansi, ISTANBUL 2002

[17] H.Takcı, İ.Soğukpınar,Kütüphane kullanıcılarının erişim desenlerinin keşfi

[18] http://www.spss.com/clementine/

[19] S. Özekes, Veri Madenciliği Modelleri ve Uygulama Alanları, İstanbul Ticaret Üniversitesi Dergisi

[20] G. Karypis, E. Han, V. Kumar, Chameleon: Hierarchical Clustering Using Dynamic Modeling, , *IEEE Computer,* 1999 : 68-75

[21] G.D.Ramkumar, A. Swami., Clustering Data Without Distance Functions, *IEEE Bulletin ofthe Technical Committee on Data Engineering,* Vol.21 No.l, March 1998 : 9-14.81

[22] J.Han, M. Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, 1st Ed., San Francisco, USA, 2000.

[23] U.Fayyad, Mining Databases: Towards Algorithms for Knowledge Discovery, *IEEE Bulletin of the Technical Committee on Data Engineering,* Vol.21 No1, March 1998:41-48.

[24] J.Han, Cluster Analysis, *http://www-sal.es.uiuc.edu/~hanj/bk/8clst.ppt.*

[25] H.Akpınar, Veri Tabanlarında Bilgi Kesfi ve Veri Madenciligi, İ*stanbul Üniv.* İs*letme Fakültesi Dergisi,* C:29 S: 1 Nisan 2000

[26] J.Sristava, R. Cooley, M. Deshpande, Pang-Ning. Tan, "Web usage mining: Discovery and applications of usage patterns from web data." *SIGKDD Explorations*,1(2), 2000, 12-23.

[27] M.Gezer, Ç.Erol, S.Gülüseçen,bir web sayfasının web madenciliği ile analizi,Akademik bilişim 2007,Dumlupınar Üniversitesi,Kütahya 31 ocak-2 şubat 2007.

[28] M.Spiliopoulou,C. Pohle., "Data Mining for Measuring and Improving the Success of Web Sites." *Special issue on applications of data mining to electronic commerce, Journal of Data Mining and Knowledge Discovery*, 5(1-2), 2001, 85-114.

[29] W.J. Frawley, G.P. Shapiro, C.J. Matheus, Knowledge Discovery in Databases: An Overview. AI Magazine 13(3): 57-70 (1992)

[30] WUM,WUMprep, http://www.hypknowsys.org

[31] A.Carus, A. Mesut, Web Kullanım Madenciliği Uygulaması,*II. Mühendislik Bilimleri Genç Araştırmacılar Kongresi* MBGAK 2005 İstanbul 17– 19 Kasım 2005

[32] A.S. Lalani, Data Mining of Web Access Logs, Master of Applied Science Thesis, Royal Melbourne Institute of Technology, 2003

[33] J. Luo, Towards a flexible Interactive Webu Usage Mining System, Master of Science Thesis,University of Alberta,2001

[34] Peter I. Hofgesang, Web Usage Mining Structuring semantically enriched clickstream data, Master of Computer Science Thesis, Vrije Universtiy, 2004

[35] R.W.Cooley, Web usage mining: Discovery and Application of Interesting Patterns from Web Data, Doctoral Thesis, University of Minnesota,2000

[36] D.Xing, J. Shen, Efficient data mining for web navigation patterns. *Information and Software Technology* vol.46, (2004)

[37] Mixture model implementation within the DIANA Project http://www.cs.vu.nl/ci/DataMine/DIANA/

[38] Z.Chen, A. Fu, F. Tong, Optimal Algorithms for Finding User Access Sessions from Very Large Web Logs. *Proceedings of the Sixth Pacific-Asia Conference on*
Knowledge Discovery and Data Mining, (PAKDD), Taipei, 2002

[39] S.Gunduz, Recommendation Models for Web Users: User Interest Model and Click-Stream Tree,Ph.D. Thesis, Department of Computer Engineering, Institute of Science and Technology, Istanbul Technical University, 2003.

[40] F.M Facca, P.L.Lanzi, Mining Interesting Knowledge from Weblogs:a Survey, *Data&Knowledge Engineering,* 53(3), 2005, 225-241.

[41] M. Uluer, Web Günlük Analizi, 2004

[42] E. Tuğ, M. Şakiroğlu, A. Arslan, Automatic discovery of the sequential accesses from web log data files via a genetic algorithm,2003

[43] S.Gunduz, M. T. Ozsu: A Poisson Model for User Accesses to Web Pages. ISCIS 2003

[44] World Wide Web Consortium, http://www.w3.org/.

[45] D.Florescu, A.Y. Levy, A.O.Mendelzon, Database technicques fort he worl-wide web: A survey, 1998

[46] G.Paliouras,C.Papatheodrou,V.Karkaletsis, P.Tzitziras, C.D.Spyropoulos, Large-scale mining of usage data on web sites, in AAAI 2000 Spring Symposium on Adaptive User Interfaces,2000

[47] P. S. U. M. Fayyad, G. Piatetsky-Shapiro, R. Uthurusamy, Advances in knowledge discovery and data mining, Cambridge, 1996

[48] F.Aydoğan, E-ticarette veri madenciliği yaklaşımlarıyla müşteriye hizmet sunan akıllı modüllerin tasarımı ve gerçekleştirmi, Yüksek lisans tezi, Bilgisayar Mühendisliği Bölümü, Fen Bilimleri Enstitüsü, Ankara Üniversitesi,2003

[49] Data Mining, http://www.pcc.qub.ac.uk/tec/courses/datamining/ohp/dm-OHP-final_2.html

[50] C.J.Matheus, P.K. Chan, G. Piatetsky-Shapiro, Systems for knowledge discovery in databases, IEEE Trans. On Knowledge And Data Engineering, 1993

[51] Mennan's Weblog, http://mennan.kagitkalem.com/VeriMadenciligiDataMiningNedirVeNerelerdeKullanilir.aspx

[52] J.F. Elder IV, D.W. Abbott, A comparison of leading data mining tools, Fourth International Conference on Knowledge Discovery & Data Mining Friday, New York, 1998

[53] G. Piatetsky-Shapiro, , An overview of knowledge discovery databases: Recent progress and challenges. In W. P . Ziarko, editör, Rough Sets, Fuzzy Sets and Knowledge. Discovery. Proceedings of the International Workshop on Rough Sets and Knowledge Discovery (RSKD'93), Berlin, Germany , 1994

[55] C.Seidman, Data Mining with Microsoft SQL Server 2000, Microsoft Pres, 2000

[56] P.Tamayo, J. Berlin, N. Dayanand, G. Drescher, D.R. Mani, C.Wang, Darwin; A scalable integrated system for data mining, http://www.oracle.conVdatawarelıouse/products/datamining/downloads/ darwin, 2000

[57] Clementine, Data Mining-An Introduction Tutorial/Practical, QUB, http://www.pcc.qub.ac.uk /tec /courses/datamining/ohp/dm-OHP-final_2.html, 2003

[58] J.Han, J. Chiang, S. Chee, J. Chen, S. Cheng, W. Gong, M. Kamber, K.Koperski, H. Zhu, DBMiner: A system for data mining in relational datahases and data warehouses, Proc. CASCON'97: Meeting of Minds, Toronto, Canada, 1997

[59] A.J. Szladow, DataLogic/R: for database mining and decision support, In Proceedings Of The International Workshop on Rough Sets And Knowledge Discovery, (Banff, Alberta, Canada), 1993

[60] G.P. Shapiro, C.J.Matheus, Knowledge discovery workbench for exploring business databases, International Journal of Inteldigent Systems, 1992

[61] http://www.galeas.de/webmining.html

[62] http://www.kdnuggets.com/

[63] http://en.wikipedia.org/wiki/Web_mining

[64] http://infolab.stanford.edu/~ullman/mining/2006/index.html

[65] http://www.cs.waikato.ac.nz/ml/weka/