T.C.
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

WEB KULLANICI ERİŞİM KÜTÜKLERİNDEN BİLGİ ÇIKARIMI

Resul DAŞ

Tez Yöneticisi

Prof. Dr. Mustafa POYRAZ

Doç. Dr. İbrahim TÜRKOĞLU

DOKTORA TEZİ

ELEKTRİK – ELEKTRONİK MÜHENDİSLİĞİ ANABİLİM DALI

ELAZIĞ, 2008

T.C.

FIRAT ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

WEB KULLANICI ERİŞİM KÜTÜKLERİNDEN BİLGİ ÇIKARIMI

Resul DAŞ

Doktora Tezi

Elektrik – Elektronik Mühendisliği Anabilim Dalı

Bu tez, ... tarihinde aşağıda belirtilen jüri tarafından oybirliği / oyçokluğu ile başarılı / başarısız olarak değerlendirilmiştir.

Danışman: Prof. Dr. Mustafa POYRAZ

Üye: Prof. Dr. Muammer GÖKBULUT

Üye: Prof. Dr. Z. Hakan AKPOLAT

Üye: Yrd. Doç. Dr. Metin DEMİRTAŞ

Üye: Yrd. Doç. Dr. Mustafa TÜRK

Bu tezin kabulü, Fen Bilimleri Enstitüsü Yönetim Kurulu'nun ...../...../..... tarih ve ..... sayılı kararıyla onaylanmıştır.

TEŞEKKÜR

Doktora tez çalışmam süresince ilgi ve yardımlarını esirgemeyen danışman hocalarım Sayın Prof. Dr. Mustafa POYRAZ'a ve Sayın Doç. Dr. İbrahim TÜRKOĞLU'na, uygulamalarda Web kullanıcı erişim kütük verilerini kullanma olanağı sağlayan Fırat Üniversitesi Bilgi İşlem Daire Başkanlığı'na, çalışmalarımda fikir ve destekleriyle yardımlarını esirgemeyen çalışma arkadaşlarıma ve çalışmalarımın bu zor sürecinde beni sabır ve özveri ile destekleyen eşim Bihter DAŞ'a teşekkürlerimi ve şükranlarımı sunarım.

Doktora tez uygulama çalışmalarını gerçekleştirmek için ihtiyaç duyduğum bilgisayar yazılım ve donanım ürünlerini temin etmemde gerekli maddi desteği sağlayan Fırat Üniversitesi Bilimsel Araştırmalar Proje (FÜBAP) birimine teşekkür ederim.

Resul DAŞ

İÇİNDEKİLER

ŞEKİLLER LİSTESİ

TABLOLAR LİSTESİ

EKLER LİSTESİ

KISALTMALAR LİSTESİ

AGNES	AGlomerative NESting – Kümeleme Yöntemi
CES	Collector Engine System – Web Madenciliği Aracı
CLF	Common Log Format – Yaygın Kütük Biçimi
DIANA	DIvide ANAlysis – Bölme Analizi
ECLF	Extended Common Log Format – Genişletilmiş Yaygın Kütük Biçimi
FTP	File Transfer Protocol – Dosya Aktarım Protokolü
GA	Genetic Algorithm - Genetik Algoritma
HTML	Hyper Text Markup Language – Köprü Biçimleme Dili
HTTP	Hyper Text Transfer Protocol – Köprü Aktarım Protokolü
HTTPS	Hyper Text Transfer Protocol Secure – Köprü Aktarım Protokol Güvenliği
IE	Internet Explorer – Internet Tarayıcısı
IIS	Internet Information Server – Internet Bilgi Sunucusu
IP	Internet Protocol – Internet Protokolü
ISP	Internet Service Provider – Internet Servis Sağlayıcı
NCSA	National Center for Supercomputing Applications (Kütük Dosyaları Biçimi)
NNTP	Network News Transfer Protocol – Ağ Haber Aktarım Protokolü
OLAP	On Line Analytical Processing – Çevrimiçi Analitik İşleme
OSI	Open System Interconnection
SAS	Uluslararası Bilgisayar Yazılım Şirketinin Adı
SHTTP	Secure Hypertext Transfer Protocol – Güvenli Köprü Aktarım Protokolü
SMTP	Simple Mail Transfer Protocol – Basit Posta Aktarım Protokolü
TCP/IP	Transmission Control Protocol / Internet Protocol – İletim Kontrol Protokolü
URI	Uniform Resource Identifier – Düzenli Kaynak Belirteci
URL	Uniform Resource Language – Düzenli Kaynak Dili
W3C	World Wide Web Consortium – Internet Konsorsiyumu
WALA	Web Access Log Analyzer – Web Erişim Kütük Analizcisi
WWW	World Wide Web – Geniş Dünya Ağı
XML	Extensible Markup Language – Uzatılabilir Biçimleme Dili

SİMGELER LİSTESİ

θ	Eşik değeri
S	Oturum
t_0	Oturum başlangıç süresi
t_n	Oturum bitiş süresi
X	Değişken
Y	Değişken
q	Erişim kütüklerindeki kullanıcı URL isteği
(-)	Erişim kütüklerinde boş adres bilgisi
h-1	Oturum süresi temelli sezgisel yöntem
h-2	Sayfada kalma süresi temelli sezgisel yöntem
h-ref	Referans temelli sezgisel yöntem
$\|X\|$	X ürünü içeren alışverişlerin sayısı
$\|D\|$	Yapılan tüm alışverişlerin sayısı
$\|X.Y\|$	X ve Y ürünlerini içeren destek
$A \Rightarrow B$	Kural
$B \Rightarrow A$	Kural
P_{yx}	X bağımsız değişkeninin Y bağımlı değişkeni üzerinde yapmış olduğu doğrudan etkiyi gösteren yol katsayısı
b	Kısmi regresyon katsayısı
σ_{y^2}	Y değişkenine ait varyans
$\sigma_{X_1^2}$, $\sigma_{X_2^2}$	Birinci ve ikinci X değişkenine ait varyans
P_{YX_1}	X_1'in Y üzerindeki doğrudan etkisi
$(r_{x_1 x_2} . P_{yx_2})$	X_1'in X_2 üzerinden olan dolaylı etkisi
$(r_{x_1 x_3} . P_{yx_3})$	X_1'in X_3 üzerinden olan dolaylı etkisi
A	Değişkenler arasındaki ilgileşim katsayılarından oluşan vektör
B	Üç denklemde yer alan ilgileşim katsayılarından oluşan vektör
C	Yol katsayılarından oluşan vektör

ÖZET

DOKTORA TEZİ

WEB KULLANICI ERİŞİM KÜTÜKLERİNDEN BİLGİ ÇIKARIMI

Resul DAŞ

Fırat Üniversitesi
Fen Bilimleri Enstitüsü
Elektrik – Elektronik Mühendisliği Anabilim Dalı
2008, Sayfa: 126

Son yıllarda İnternet'in hızlıca gelişmesi ve yaygın kullanımı ile Web, dünyada erişilebilir en geniş veri kaynağı haline gelmiştir. Internet'teki bilgi yığınları aşırı şekilde artarken, Web ziyaretçi isteklerine uygun hizmetlerin sağlanabilmesi, Web site yapısının iyileştirilmesi, geliştirilmesi ve etkin olarak kullanılması gibi amaçları sağlamak için *Web Madenciliği* gittikçe daha ilgi çeken bir konu olarak görülmektedir.

Bu tez çalışmasında, metin tabanlı web kullanıcı erişim kütüklerinin temizlenmesine yönelik yeni bir süreç önerilmiştir. Önerilen sürecin uygulaması ve program kodlamaları JAVA tabanlı SAS Base yazılım ortamında geliştirilmiştir. Büyük boyutlardaki kütük dosyalarının temizlenmesinde geliştirilen veri temizleme süreci hız yönünden diğer yöntemlere göre üstünlük sağlamıştır.

Temizlenmiş kullanıcı erişim kütüklerinden anlamlı ve ilginç bilgilerin çıkarılması için üç farklı çalışma yapılmıştır;

- *Yol analizi yöntemi* ile web kullanıcı erişim kütük dosyalarından anlamlı ve ilginç örüntüleri içeren bilgiler çıkarılmıştır. Literatürde geçen ve başka alanlarda uygulaması yapılan yol analizi yönteminin, web kullanıcı erişim kütüklerine uygulaması yapılarak, anlamlı ve ilginç örüntüleri içeren bilgiler çıkarılmasın da başarılı bir şekilde kullanılabileceği gösterilmiştir.
- Aynı veri tabanı verilerine *birliktelik kuralı yöntemi* uygulanarak Web sayfaları arasındaki ilişkileri belirleyen bilgiler çıkarılmıştır.

- *İstatistiksel analiz* ile Web sitesinin üç aylık bir süre içerisindeki genel kullanımına ilişkin detaylı istatistikî bilgiler çıkarılmıştır.

Gerçekleştirilen tüm uygulama çalışmaları sonucunda, elde edilen bilgiler kullanılarak web sitesinin iyileştirilmesine, geliştirilmesine, kullanılabilirliğine ve yapısal organizasyonuna katkı sağlamak için web site tasarımcılarına ve yöneticilerine öneriler sunulmuştur. Ayrıca, HTTP durum kodları analiz edilerek web sitesi ve sunucusunun başarımını arttırmaya yönelik çözüm önerileri oluşturulmuştur. Oluşturulan öneriler ziyaretçi memnuniyetini arttırmaya yönelik kazanımları da amaçlamaktadır.

Anahtar Kelimeler: Web madenciliği, Web kullanım madenciliği, Bilgi çıkarımı, Örüntü keşfi, İstatistiksel analiz, Yol analizi, Birliktelik kuralları, SAS Yazılımı.

ABSTRACT

PhD Thesis

KNOWLEDGE EXTRACTION FROM WEB USER ACCESS LOGS

Resul DAŞ

Firat University
Graduate School of Natural and Applied Sciences
Department of Electrical - Electronics Engineering
2008, Page: 126

Recently, by rapidly developing and common usage of the Internet, Web has been largest accessible data source in the world. While extremely growing the knowledge masses up on the Internet by passing time, *Web Mining* has been seeing as more attractive subject more and more to cope with goals such as improving, growing healthy and using effectiveness of web site structure and to provide appropriate web service to the web clients requests.

In this thesis, a new process intended for purifying of text-based web user access logs are proposed. The implementation and code of the proposed process have been designed on Java-based SAS software environment. Purifying high dimensional data access logs, the improved data cleaned process is superior to the other methods in speed aspect.

Extracting meaningful and interesting knowledge from the purified user access logs, three different implementations have been realized.

- Knowledge contained the meaningful and interesting patterns from web user access log files have been extracted by using *Path Analysis Method.* The implementation of web user access logs of Path Analysis Method, present in the literature and implemented in different fields, has successfully shown that it can be used in extracting meaningful and interesting knowledge.
- Relation knowledge between web sites by applying *Association Rules Method* on the same dataset is extracted.
- The detailed statistical knowledge regarding three months usage of the Web site has been extracted by using *Statistical Analysis Method*.

In the end of all implementation, web site designers and managers are given suggestions about improvement, grow healthy and being usable of the web site and to contribute its structural organization by using obtained meaningful and pure knowledge. Furthermore, some solution suggestions with relation to increase successfulness of web site and server by analyzing HTTP state codes have been formed. The formed suggestions are intended for increase the visitor pleasure.

Keywords: Web mining, Web usage mining, Knowledge extraction, Pattern discovery, Statistical analysis, Path analysis, Association rules, SAS Software.

1. GİRİŞ

İnternet, elektronik ortamdaki verilerin uzak mesafelere aktarılması ve erişilmesi açısından dünya üzerinde var olan en büyük bilgi paylaşım ortamıdır. Bilgi paylaşımında küresel bir yapı haline gelen İnternet, birçok kişi, kurum ve kuruluşlar için farklı boyutlar kazandırmaktadır. İnternet'teki bilgi yığınlarının aşırı şekilde artması, yaygın olarak kullanılan Web sitelerinin etkin olarak kullanılabilmesi, geliştirilebilmesi ve kullanıcı taleplerine uygun hizmetlerin sağlanabilmesi ile güvenlik gereksinimleri gibi ihtiyaçların belirlenmesi amacıyla Web Madenciliği konusu aktif olarak ortaya çıkmıştır [1]. İlk kez Etzioni tarafından 1996'da ortaya atılan Web madenciliği, İnternet'teki Web bağlantı yapılarını, Web sayfa içeriklerini ve kullanıcıya ait kullanım verilerinden yararlı ve anlamlı bilgiyi keşfetmeyi amaçlamaktadır [2].

Toplumun her kesiminde İnternet kullanımının yaygınlaşması, yeni sektörleri ortaya çıkarmaktadır. Bu sektörler içerisinde elektronik haberleşme, bilgi tarama, finans, elektronik ticaret v.b. örnekler oluşturmaktadır. İnternete bağlı kişisel bilgisayarların kullanımının artması, bu sektörlerin hızla gelişmesini ve bunlara bağlı olarak alt sektörlerin oluşmasını sağlamaktadır. Bu sektörlere yapılan yatırımlar, hem alt sektörlerin gelişimini hızlandıracak hem de istihdamı arttırıcı bir etki meydana getirmektedir. Ayrıca, bu alanda meydana gelen dünyadaki gelişmeler doğrudan ülke ekonomisine yansımaktadır. Buna paralel olarak İnternet ve Web sayfaları üzerindeki verilerin hacmi ve karmaşıklığı da gün geçtikçe hızla artmaktadır. Bu bilgi karmaşıklığı ve yoğunluğun çözümünde Web madenciliği yöntemlerinin geliştirilmesi ve kullanılması her geçen gün önem kazanmaktadır.

Web sunucularının metin tabanlı olarak sakladıkları erişim kütük dosyalarında Web sitelerinin izlenilmesine ve analiz edilmesine yönelik çok önemli veriler bulunmaktadır. Bu veriler, ziyaretçilerin sayfaları gezerken bıraktıkları izlerin yanı sıra siteye kayıt yaptırırken girmiş oldukları bilgilerden oluşmaktadır. Web madenciliğinin temel görevi, metin tabanlı bu karmaşık ve anlamsız veri yığınlarından veri madenciliği yöntemlerini kullanarak site yöneticisine ve Web tasarımcısına sitenin geliştirilebilmesi ve başarımının arttırılmasının sağlanması için yararlı ve anlamlı bilgiler çıkarmaktır. Ayrıca, Web analizi çalışmaları ile ticari amaçlı hizmet veren bir Web sitesinin elde ettiği kar miktarı arttırılabilir, akademik İnternet sayfaları farklı ilgi alanlarına göre düzenlenerek ziyaretçi memnuniyeti arttırılabilir.

Gelişen teknoloji ile birlikte tasarlanan Web sayfaları sayesinde, Web kullanım madenciliğinde kullanılabilecek önemli verilerin sayısı da artmaktadır. Buna paralel olarak, ortaya konulacak yararlı bilgilerde artacaktır. Web tasarımcılarının Web sitelerini iyileştirmelerine ve geliştirmelerine, Web sunucu yöneticilerinin sistem hatalarını gidermeleri için fayda sağlayacak birçok yararlı ve gerekli bilgiler sunulabilir. Web sitesindeki kırık olan

köprüler, en yoğun girilen sayfalar, en çok indirilen dosyalar, en ilgi görülen resim, şekil ya da görüntüler, en çok meydana gelen sunucu ve istemci hatalarına ait bilgiler tespit edilerek, Web sitesinin güncellenmesi işleminde, bu bilgileri göz önüne alması sağlanır. Bu durum Web sitesine yapılan saldırı ve ataklarla ilgili sunulan bilgilerle de sistem yöneticilerine büyük destek sağlayabilir.

İnternet bankacılığını müşterilerine sunan banka yetkilileri, müşteri davranışları hakkında elde edilebilecek önemli özel bilgiler ışığında ticari kazanımlarına katkı sağlayabilirler. İnternet kullanıcı erişim kayıtları analiz edilerek, İnternet kullanıcılarına ait birçok sayısal verilerin tespit edilmesi mümkündür. Sonrasında, kullanıcılarla ilgili farklı bilgilere yönelme, kullanıcı davranışlarıyla ile ilgili tahminlerde bulunma ve benzeri araştırmalar yapma gibi konularda kolaylık sağlayarak, bu konudaki araştırmacıların ufkunu genişletecektir.

Bu doktora tez çalışmasında, Web kullanım madenciliği alanında akademik olarak yapılan araştırma ve incelemeler sonucunda, geliştirilen yeni algoritma ve önerilen yöntemler ile bazı önemli temel hususlar, gelecek yıllarda yapılacak yeni çalışmalara ışık tutarak, bu alandaki araştırmacılara farklı bakış açıları kazandıracağı beklenmektedir.

1.1. Amaç

Web madenciliği ile İnternet kullanıcılarına yönelik kurumsal anlamda araştırma ve çalışma yapılarak kurum çalışanlarına yönelik durum değerlendirilmesi yapılabilmektedir. Ayrıca kurum veya birim içerisinde ortak kullanıma açık olan yazılım, otomasyon ya da Web sitelerini kullananların davranışlarına ilişkin bilgilerin çıkarılması, kullanıcılara ait karakteristik özelliklerin tahmin edilmesi gibi birçok analiz çalışmalarının yapılabilmesi Web madenciliğine ilgi çekmektedir.

Bu tezin temel amacı, karmaşık, düzensiz ve herhangi bir anlam ifade etmeyen metin tabanlı kullanıcı erişim kütük dosyalarından Web madenciliği tekniklerine dayalı yöntemler geliştirerek anlamlı ve ilginç bilgiler çıkarmaktır. Bu bilgilerin değerlendirilmesi ile aşağıda sıralanan bulguların elde edilmesi hedeflenmektedir;

1. Web tasarımcıları ve yöneticileri için, sitesinin iyileştirilmesine, geliştirilmesine, kullanılabilirliğine ve yapısal organizasyonuna katkıda bulunmak.
2. Web sitesi kullanımı ve İnternet kullanıcılarının davranışları hakkında birçok istatistikî bilgiler elde ederek, kullanıcı davranışları hakkında tahminlerde bulunmak.

3. Web sunucu kütüklerinden çıkarılan *HTTP* durum kodları bilgisi kullanılarak Web sitesinde oluşan problemlerin giderilmesine yönelik teknik iyileştirmeler yapmak. Bunun sonucunda hem Web sitesinin hem de sunucunun çalışma başarımını arttırmak.

1.2. Literatür Taraması ve Değerlendirilmesi

Son 15 yıl içerisinde İnternet'teki büyük ve hızlı gelişmeler, Web sayfalarının vazgeçilmez derecede önemli olduğunu göstermektedir. Günümüzde 8 milyar'dan fazla çevrimiçi belge bulunmaktadır. İnternet'e her gün 20 milyon yeni Web sayfası eklenmektedir [3]. Online Computer Library Center araştırmacıları, İnternet'in büyüme hızının yavaşladığını ancak gelişmesinin devam edeceğini yayınladıkları yıllıklarında göstermektedirler [4].

İnternet kullanıcılarının Web sitelerinde gezinmesi ile oluşan kullanıcı davranış verileri Web sunucularında tutulmaktadır. Kendi başına değersiz ve anlamsız olan Web kullanıcı erişim kütükleri, ancak belli bir amaç doğrultusunda işlenirse anlamlı bilgiye dönüşmektedir. Verinin bilgiye dönüştüğü bu süreç, veri işleme ve analizi olarak adlandırılmaktadır. E-ticaret yapan Web sitelerine ait verilerin analizinde; ileriye dönük satış tahminleri çıkarılabilir, müşterilerin satın aldıkları ürünlere göre gruplandırma yapılabilir, yeni bir ürün için potansiyel müşteriler belirlenebilir, müşterilerin zaman içerisindeki hareketleri incelenerek onların davranışları ile ilgili tahminler yapılabilir. Binlerce ürün ve müşterinin olabileceği düşünüldüğünde bu analizin gözle ve elle yapılamayıp otomatik olarak yapılması gerektiği açıktır [5].

Literatürde, Web madenciliği konusunda birçok çalışmalar yapıldığı ve farklı yaklaşımlar sunulduğu görülmektedir. Özellikle son yıllarda, Web sunucularda tutulan erişim kütük dosyaları kullanılarak ilginç ve anlamlı örüntülerin bulunması, kullanıcıya ait bilgilerin ve davranışların tespit edilmesi gibi yeni çalışmalar önem kazanmaktadır. Cooley ve diğ. yaptıkları çalışmalarda, Web kullanım madenciliği için WEBMINER adlı bir sistem geliştirmişlerdir. WEBMINER, otomatik olarak kullanıcı erişim kütüklerinden birliktelik kuralları ile sıralı örüntüleri keşfetmektedir. Ayrıca, Web madenciliği teknikleri ile ilgili genel tanımlamalar ve araştırmalar teorik olarak incelemiş, genel bir durum analizi yapmışlardır [6]. Chen ve Syncara geliştirdikleri WebMate adlı sistemlerinde, Web sayfalarını inceleyerek Web içeriğinden kullanıcı ilgilerini belirlemeyi sağlamışlardır [7]. Srivasta ve diğ. çalışmalarında, Web kullanım madenciliği ile ilgili genel kavramları ve temel tanımlamaları belirtmişlerdir. Ayrıca, ISP'lerde (İnternet Service Provider) tutulan veri kayıtları ile ilgili bilgilendirmeler ve IIS'in işlevinden bahsetmişlerdir. Çalışmanın amacında, Web kullanım madenciliğinde karşılaşılan zorluklar ve bu zorlukların aşılması sonucunda umulan başarılar yatmaktadır [8].

Sunuculardaki tutulan farklı biçimdeki kütük dosyalarındaki metinsel verilerin biçimi birbirinden farklıdır. Metin tabanlı verilerden sağlıklı bilgi çıkarımı yapılabilmesi için, kütük dosyalarının gürültülü ve gereksiz verilerden ayıklanması gerekmektedir. Çünkü kütük dosyalarındaki gereksiz ve gürültülü veriler, anlamlı bilgilerin çıkarılmasında hata oranını arttırmaktadır. Cooley ve diğ. yaptıkları çalışmalarında İnternet sunucularında tutulan kütüklerin ön işlem sürecinin analizi ile ilgili kapsamlı olarak inceleme yapmışlardır [9]. Araya ve diğ. yaptıkları Web kullanım madenciliği çalışmasında yeni bir yöntem geliştirmişlerdir. Geliştirdikleri bu yöntem ile standart bilgi keşfi işlem basamaklarına amaçların tanımlanması ve mesleklerin birleştirilmesi şeklinde iki yeni işlem basamağı ilave etmişlerdir. İnternet bankacılığını kullanan banka müşterilerine ait bilgilerin tutulduğu kullanıcı erişim kütüklerine bu yöntemlerini uygulayarak, kayıtlı banka müşterileri arasında istatistiksel analiz yapmışlardır [10]. Uğur ve Kınacı yaptıkları yapay zekâ tekniği çalışmalarında, kategorilere ayrılmış dmoz.org [11] Web sitesindeki verilere yapay sinir ağları yöntemini uygulayarak Web sayfalarını sınıflandırmışlardır [12]. Uğuz ve diğ. çalışmalarında, kullanıcı erişim kütük dosyalarındaki verilere Apriori algoritması uygulayarak kullanıcı erişim örüntülerinden kullanıcı bilgilerini çıkarmışlardır [13]. Uğuz ve diğ. yaptıkları diğer çalışmalarında, Web sunucusunun İnternet erişim kütüklerine genetik algoritma ile Web sayfası ziyaretçilerinin en sık eriştiği sayfa çiftlerini, üniversite içi ve dışı kullanıcı erişim dağılımı gibi tanımsal ilişkileri tespit etmişlerdir. Böylece, yapılan çalışma ile Web üzerinden arama işlemlerinde kolaylık sağlamışlardır [14].

İşeri yaptığı tez çalışmasında, geliştirdiği yazılım ile Web günlüğünden zaman sınırlı bulanık bağıntı kuralları ve sıralı örüntülerin çıkarılmasını sağlamıştır [15]. Şakiroğlu ve diğ. yaptıkları çalışmada, Web erişim kütük dosyalarından genetik algoritma yöntemiyle sıralı erişimleri tespit etmişlerdir [16]. Tuğ ve diğ. yaptıkları makale çalışmalarında, Web erişim kayıt dosyalarından genetik algoritma yöntemiyle sıralı erişimleri (peşpeşe en çok ziyaret edilen sayfaları) tespit etmişlerdir [17]. Burada kütük dosyalarında toplanan ham verilerden yola çıkılarak kullanıcı davranışlarının tespiti yapılmıştır. Birkaç makalede, bilgi çıkarımı ve analiz işlemleri için akıllı yapı olarak makine öğrenmesi tekniklerinden genetik algoritma yöntemi kullanılmıştır [16 – 18]. Gezer ve diğ. yapmış oldukları Web kullanım madenciliği analiz çalışmasında, İstanbul Üniversitesi Uluslararası Akademik İlişkiler Kurulu AB Eğitim birimine ait Web sitesi sunucu kayıt dosyalarına WUMprep ve WUMWeb yazılımlarını kullanarak analiz yapmışlardır [3]. Carus ve Mesut geliştirdikleri Web kullanım madenciliği yazılımı ile farklı biçimlerdeki erişim kütük dosyalarının analizini yaparak, istatistiksel sonuçlar elde etmişlerdir [19]. Belen ve diğ. yaptıkları çalışmada, Web madenciliği tekniklerini kullanarak kullanıcı ara yüzü ile veri tabanı entegrasyonunu sağlayan istatistiksel analiz yapabilen bir kütük

araştırmacısı yazılımı geliştirmişlerdir. Geliştirilen WALA (Web Access Log Analyser) adlı sistem, bir Web sitesinin kullanım analizi için gerekli olan araçları sağlar ve kullanıcıların en çok ziyaret edilen sayfalar, en yoğun sayfalar, bir arada ziyaret edilen sayfalar gibi bilgilerin belirlenmesi için bir sunucu erişim kütüğü analiz programıdır. Geliştirilen yazılımın hedefi, Web tasarımcıları ve Web yöneticileri için bir çeşit karar destek sistemi oluşturmaktır [20]. Özakar ve diğ. çalışmalarında, İzmir İleri Teknoloji Enstitüsü sunucularından aldıkları Web kütüklerindeki ham verileri temizleyip, JAVA sınıflandırıcı kullanarak ilişkisel veritabanına aktarılmaya hazır hale getirmişlerdir. Veri hazırlama bölümünde geçersiz veri ayıklanıp, veri madenciliği uygulanabilecek biçime dönüştürülmüştür. Bu aşamada; erişim kütüğü, hata kütüğü, kullanıcı verisi, Web içerik verisi incelenerek işlemler yapılmıştır. Daha sonra sorgu mekanizması ile yorumlanması gereken kısımlarda veri tabanı üzerinde tanımsal sorgular yapılmaktadır [21]. Luca Iocchi'nin çalışmasında, geliştirdiği Web-OEM modeli Webden yarı yapısal bilgilerin çıkarılması için tasarlanmış bir bilgi modelidir. Bu model, Web deki dağınık bilgi yığınlarının büyük bir kısmından bilgi keşfi yapmaktadır. Klasik bilgi modellerinin yetersiz oluşuna alternatif olarak geliştirilmiştir [22]. Habegger ve Quafafou yaptıkları çalışma da, XML tabanlı WetDL dili ile ayrıştırma yaparak Webden bilgi çıkarımı uygulaması yapmak için önerilerde bulunmuşlardır. Uygulama örneği olarak amazon.com [23] İnternet sitesi belirtilmiştir [24]. Benzer uygulama örneği ise, T.C. Kimlik numarasını bulmak için Web sayfalarından girilen kişisel bilgilerin sonucunda, kişiye ait numaranın çıkartılması olayıdır. Takci ve Soğukpınar yaptıkları makalelerde, kütüphane kullanıcılarına ait erişim kütüklerini kullanarak, kullanıcıların Web üzerindeki davranışları ile ilgili analiz yapmışlardır [25, 26]. Web kullanım madenciliği için geliştirilen birçok yazılım, sunucularda tutulan kütük dosyaları (access.log, agent.log, error.log, referrer.log) için istatistiksel analizler yapılabilmektedir [27-41]. Pedro Lineu Orso tarafından C programlama dilinde yazılmış olan SARG programı, Linux tabanlı işletim sistemlerinin bulunduğu sunucularda çalışmaktadır. Bu program, sunucuda tutulan metin tabanlı erişim kütük dosyalarını otomatik olarak ön işlem aşamasından geçirip, kullanıcıların anlayabileceği anlamlı ve düzenli bilgiler şeklinde çıkarmaktadır. Bu bilgiler kolay anlaşılabilir tablolara dönüştürülerek HTML biçiminde günlük, haftalık ve aylık zaman dilimlerine göre kullanıcıya sunulmaktadır [42]. Di Guo çalışmasında, Web madenciliği sistemi ile ortak arama sistemi bütünleştirerek elektronik ticaret için yeni bir araç geliştirmiştir. Bu çalışmada, bilgi düzenleme ve Web madenciliği uygulamaları için çoklu etken yaklaşımı (Collector Engine System- CES) önerilmiştir. Sistem içerisine IBM tarafından geliştirilen IBM Aglets ve Objectspace Voyager Agents yazılımları modül şeklinde eklenerek, tek bir yazılım olarak bütünleştirilmiştir [43].

İnternet'in yaygın kullanımıyla beraber Webe dayalı uzaktan eğitim sistemleri de hızlıca gelişmektedir. Hazırlanan eğitim materyallerinin geliştirilmesi, güncellenmesi ve analiz edilebilmesi için Web kullanım madenciliği yöntemlerini kullanmak gerekmektedir. Ayrıca, Webe dayalı uzaktan eğitim sisteminin geliştirilebilmesi için materyalleri kullanan öğrenci davranışlarınında analiz edilmesi gerekmektedir. Literatürde yapılan çalışmalara bakıldığında, Web kullanım madenciliği ile bu konuda başarılı ve etkin sonuçlar alındığı görülmektedir. Guo ve diğ. yaptıkları çalışmada, çevrimiçi öğrenme aktiviteleriyle ilişkili olan modellerin bir kümesini oluşturarak, Web kullanım madenciliği teknikleriyle öğrenci davranışlarını değerlendirmek için bir yaklaşım sunmuşlardır. Bu çalışma eğitim yazılımının modelleri, öğrenme işlemi ve Web erişim kütüklerinden dokümanları kullanan öğrencilerin davranışlarının çıkarılması gibi önem arzeden detaylı bilgileri tespit etmektedir [44].

Missouri-Columbia Üniversitesi'nden Laffey ve Ai yaptıkları araştırma çalışmasında çevrimiçi öğrenmeyi anlama konusunda bir Web madenciliği yazılım aracı geliştirmişlerdir [45]. Uzaktan eğitimde Web'e dayalı eğitim materyali olarak dünya genelinde yaygın olarak kullanılan WebCT (Blackboard) e-öğrenme sisteminin Web madenciliği ile nasıl daha da faydalı ve verimli hale getirilebileceğini anlatmaktadırlar. Araştırmacılar, e-öğrenme sisteminde Web madenciliği tekniklerinin başlıca şu üç noktada ciddi fayda getireceğini önemle belirtmişlerdir.

1. Eğitim materyallerinden yararlanan öğrenci davranışlarının daha iyi anlaşılabilmesi için modellenmesi,
2. E-öğrenme sistemlerinin eğitimde ve öğrenciler üzerinde ne kadar verimli olduklarının daha iyi belirlenmesi,
3. Web'e dayalı geliştirilen eğitimsel tasarımların faydalarının ölçülebilmesi gerekmektedir.

Eğitim amaçlı yapılan çalışmalardaki en somut örneklerden birisi de, Luan tarafından 2002'de gerçekleştirilmiş olan bir Web madenciliği ve öngörü sistemidir. Bu çalışma ile iki yıllık meslek yüksek okullarında okuyan öğrencilerden hangilerinin dört yıllık üniversitelere geçiş yaptıklarını büyük başarı ile tahmin etmişlerdir. Üniversite yönetimi ve öğretim elemanları öğrencilerin geleceğe yönelik kararlarını herkesten önce tahminde bulunurlarsa hedefe yönelik akademik hizmet sunmaları da mümkün olacaktır. Bu sistem, kaynak ve zaman verimliliğini etkin kullanma konusunda çok büyük katkılar sağlayacaktır. Ayrıca, Luan çalışmasında, WebCT yazılım aracı tarafından üretilen Web kullanıcı erişim kütük bilgilerinin ne tür önişlemlerden geçirildiği ve makine öğrenme algoritmalarından biri olan ikili karar ağacı kullanılarak öğrenci başarımlarını tahmin etmeye yönelik modelin nasıl kurulduğu

anlatılmaktadır. Sonuçta geliştirilen sistem sayesinde bir öğrencinin bir haftalık WebCT materyali üzerindeki hareketlerine bakılarak dönem sonundaki başarı notunu %70 olasılıkla tahmin etmek mümkün hale gelmektedir. Yapılan uygulamalarda, daha çok veri kullanıldığında bu tahmin oranı %90′a kadar çıkabildiği belirtilmektedir. Böyle bir geleceğe yönelik tahmin aracına kavuşan üniversite öğretim elemanlarının birebir öğrencilere odaklanması ve gerçekçi tavsiyelerde bulunması, erken uyarı mekanizmaları kurmaları da kolaylaşmış olduğu açıkça görülmektedir. Yazarın çalışmada belirttiği diğer bir nokta ise, WebCT ve benzeri e-öğrenme sistemlerinden daha ayrıntılı veri elde edilmesi, örüntü keşfi yöntemleriyle analizlerin çoğaltılıp, anlamlı parametrelerin keşfedilip analiz ve tahmin modellerinin kalitesinin yükseltilmesi ile uzaktan eğitimin verimini ve kalitesinin arttırılabileceğinin mümkün olduğunu belirtmektedirler [46]. Web'e dayalı uzaktan eğitim hizmeti sunan Web materyallerini ve Web sitelerini değerlendirmek için yapılmış birçok çalışma bulunmaktadır [47-53]. Ayrıca, Web madenciliği konusu ile ilgili teorik bilgilerin verildiği birçok çalışma da incelenmiş olup, bu çalışmalardan [54-71], elde edilen kazanımların yanı sıra, önemli ve faydalı bilgiler tez içerisinde işlenmiştir.

Yukarıda sunulan literatür araştırmasının ışığında, tezin yönelim gerekçeleri aşağıdaki gibi belirlenmiştir:

1. Web kullanıcı davranışlarının analizine yeni bakış açıları kazandırmak için farklı istatistiksel yöntemlerin Web kullanım verileri üzerindeki başarımları denenmesi gerekmektedir.
2. Web kullanımının artmasına paralel olarak Web sayfalarının sayısı da hızla artmaktadır. Bu durum hem ticari hem akademik olarak Web kullanım madenciliğine olan ihtiyacı da zaruri hale getirmektedir. Bu durum, mevcut Web kullanım madenciliği uygulamalarının gelişen teknolojiye uygun olarak geliştirilmesi gerektiğini ortaya koymaktadır.
3. Web kullanıcı erişim kütüklerinin Web kullanım madenciliği ile analiz edilmesi, öğrenci davranışları ve materyaller hakkında birçok istatistiksel bilgi çıkarımları yapılabilmektedir. Bu bilgi çıkarımları sayesinde Web sitesinde bulunan eğitim materyalleri değerlendirilerek, sunucu ve belge kaynaklı oluşmuş hata kayıtları analiz edilebilmektedir. Analiz sonucunda ortaya çıkan problemlerin giderilmesi için yeni düzenlemelerin yapılması gereği ortaya çıkmaktadır.
4. Bir Web sitesinde İnternet kullanıcılarının yoğun ilgi gösterdikleri Web sayfalarını tespit edilmesi, sitenin gelişimi ve etkin kullanımı açısından önem arz etmektedir.

1.3. Tezin Organizasyonu ve Katkılar

Tez çalışmasının birinci bölümünde, teze genel bir bakış açısı kazandırılmaya yönelik temel bilgiler verilmiştir. Diğer bölümlerin organizasyonu ile birlikte tezde yapılan katkılar aşağıda sunulmaktadır:

Bölüm 2'de, Web madenciliği konusu detaylı bir şekilde ele alınmış, literatür de geçen alt bölümleri ile birlikte açıklanmıştır. Özellikle tez çalışmasında temel çalışma alanı olarak belirlenen, Web kullanım madenciliği ayrıntılı olarak incelenip, çalışma ve uygulama alanları üzerinde daha geniş bir şekilde durulmuştur.

Bölüm 3'de, Web kullanım madenciliği uygulamalarının ilk ve en önemli aşaması olan verilerin toplanması ve ön işlem süreci ayrıntılı verilmiştir. Ayrıca, Web kullanım madenciliğinde kullanılan veri tipleri ve veri yapıları incelenmiş, verilerin temizlenmesi ve veri analizi için kullanılabilir bir veri kümesinin oluşturulmasındaki adımlar açıklanmıştır.

Bölüm 4'de, Web kullanım madenciliğinin ikinci aşaması olan örüntü keşfi ve analiz edilmesi konusu ele alınmıştır. Bu kısımda, literatürde geçen ve en çok bilinen örüntü keşif ve analiz yöntemlerinin önemli özellikleri vurgulanmış ve örüntü analizi uygulamaları irdelenmiştir.

Bölüm 2, 3 ve 4'de, tez çalışmasının literatürdeki yerine açıklık getirilerek dayandığı temel konular verilip, tez çalışmasında önerilen ve geliştirilen yöntemlerin anlaşılabilmesine yönelik alt yapı bilgileri, çalışma alanının yeni olması nedeni ile genişçe sunulmuştur.

Bölüm 5'de, *yol analizi* yöntemi Web kullanıcı erişim kütük verilerine uygulanarak, kullanıcı davranışlarına yönelik anlamlı ve ilginç bilgiler çıkarılmıştır. Bu uygulama çalışması literatürde karşılaşılmayan bir çalışmadır. Kütük verilerinin temizlenmesi işlemi için SAS tabanlı yazılım ortamında bir yöntem geliştirilmiştir. Geliştirilen veri temizleme yöntemi ile Web kullanıcı erişim kütükleri temizlenerek yapılan analiz uygulamasında Web sitesinin yüksek düzeyde başarım elde edildiği gösterilmiştir. Bu uygulama, Fırat Üniversitesi Web sunucularından alınan Web kullanıcı erişim kütüklerine uygulanarak, elde edilen bilgilerin kullanılabilirliği, hem Web sitesi hem de kullanıcılar açısından ortaya konulmuştur.

Bölüm 6'da, Fırat Üniversitesi Web sunucularında tutulan genişletilmiş kullanıcı erişim kütük dosyalarından Web kullanım madenciliği ile hem kullanıcıya hem de Web sitesine ilişkin

anlamlı ve yararlı önemli istatistikî bilgiler çıkarılmıştır. Bu bilgiler Web sitesinin içeriğine, yapısına ve kullanımına ilişkindir. Elde edilen bu bilgiler kullanılarak, Web sitesinin gelişimine katkı sağlayıcı öneriler sunulmuştur. Ayrıca, HTTP durum kodları çıkarılarak, Web sitesindeki hata kayıtları analiz edilmiştir. Yapılan bu analiz çalışmasından hareketle, Web sitesinde var olan kırık bağlantılar, engellenmiş sayfalar, kısıtlı erişimler, sunucu kaynaklı problemler tespit edilerek Web sitesinin geliştirilmesi ve sistem başarımının arttırılmasına yönelik çözüm önerileri oluşturulmuştur. Bu öneriler ışığında, Web sitesi üzerinde yapılacak değişikler ve geliştirmeler Web sitesinin başarımını sağlamaya yönelik niteliktedir.

Bölüm 7'de, veri madenciliği tekniklerinden birliktelik kuralları madenciliği kullanılarak Web erişim kütüklerinden Web sayfalarının birliktelikleri ile ilgili anlamlı sonuçlar çıkarılmıştır. Bu uygulama çalışmasında, Fırat Üniversitesi Web sunucularına ait Web kullanıcı erişim kütükleri kullanılmıştır. Uygulama sonucunda, birliktelik kurallarının yanı sıra istatistiksel bilgiler, kural matrisi ve köprü bağlantı grafiği bilgisi de çıkarılmıştır. Web sitesinin yeniden tasarlanmasında ziyaretçilerin kullanım memnuniyetini ve ilgilerini arttırmak için sitenin geliştirilmesi ve iyileştirilmesine yönelik çözüm önerileri getirilmiştir.

Bölüm 8'de, tezin sonuçları irdelenmiş ve orijinal katkılar vurgulanmıştır. Ayrıca ileriye dönük uygulama alanları ve öneriler tartışılmıştır.

Bu tez çalışması, Fırat Üniversitesi Bilimsel Araştırma Projeleri Birimi tarafından *FÜBAP–1526* numaralı proje ile desteklenmiştir.

2. WEB MADENCİLİĞİ

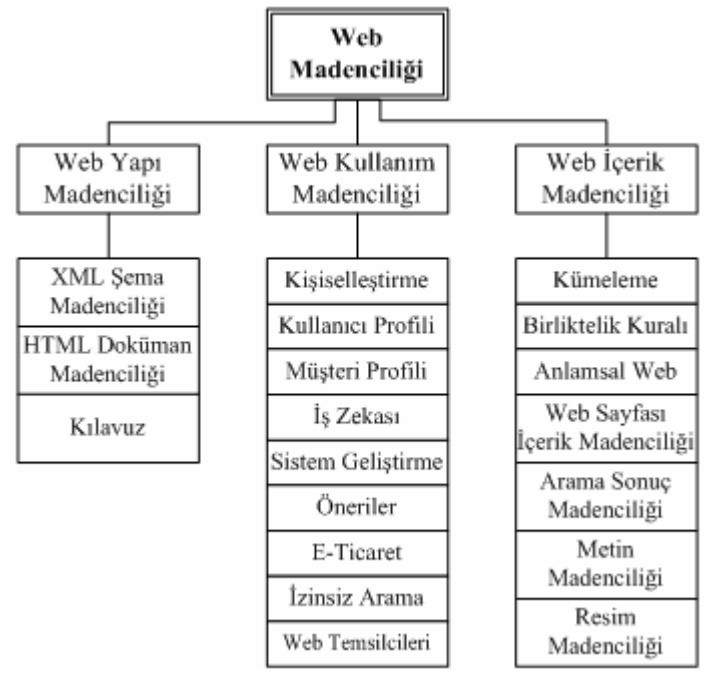
Web madenciliği, geleneksel veri madenciliği tekniklerini kullanarak World Wide Web'de bulunan dosya ve servislerden otomatik olarak örüntü bulmak ve öngörülmeyen bilgiye ulaşmaktır [2]. Web madenciliği birçok veri madenciliği tekniklerini kullanmasına rağmen, Web verilerinin heterojen, yarı yapısal ve yapısal olmayan anlamsız yapısından dolayı tamamen geleneksel veri madenciliğinin bir uygulaması değildir. Kosala ve Blockeel [62] Web madenciliğinin görevlerini dört bölüme ayırarak incelemişlerdir. Bu görevler, Web madenciliği için kullanılan anlamsız verilerden yararlı bilginin çıkarılmasının temel adımlarıdır.

1. Kaynakların bulunması: Web madenciliği işlemleri için gerekli olan Web dokümanlarının, bilgilerin ve servislerin bulunmasıdır.

2. Bilginin çıkarılması: Elde edilen Web kaynaklarından istenilen bilginin otomatik olarak çıkarılmasıdır.

3. Genelleştirme: Çoklu veya bireysel Web sitelerindeki genel örüntülerin bulunmasıdır.

4. Analiz etme: Keşfedilen genel örüntülerin yorumlanması ve doğruluğunun onaylanmasıdır.



Şekil 2.1. Web madenciliğinin sınıflandırılması

İnternet'te var olan verilerin sürekli olarak değişmesi, güncellenmesi ve yeni bilgilerin eklenmesi gibi işlemler Web'den bilgi çıkarılması uygulamalarında karşılaşılan büyük bir zorluktur. Web madenciliğinin çalışma alanlarının kapsamlı ve detaylı olması bu alanda düzenli bir sınıflandırmayı da gerektirmektedir. Literatürde yapılan birçok akademik çalışmada Web

madenciliği, *Web yapı madenciliği, Web içerik madenciliği* ve *Web kullanım madenciliği* olmak üzere üç temel alanda kategorize edilmektedir [1, 3, 15, 62-63, 65]. Şekil 2.1'de Web madenciliğinin sınıflandırılması ve alt çalışma alanları gösterilmektedir [1].

Srivasta ve diğ. [8] Web madenciliğinde kullanılabilecek verileri, içerik, yapı, kullanıcı profili ve kullanım olmak üzere dört farklı tipte tanımlamışlardır. Bu veriler Web sunucusu (Web server), istemci (client) ve vekil (proxy) sunucusu gibi farklı kaynaklardan elde edilebilir. Web madenciliğinde kullanılabilen bu veri çeşitleri kısaca aşağıda açıklanmaktadır:

Web içerik verisi: Kullanıcıların eriştiği ve kullandıkları grafik, resim, şekil, ses ve görüntü dosyaları gibi gerçek verilerden oluşan Web sayfalarıdır. Bunların dışında bir Web sitesi, tanımlayıcı kelimeler, anlamsal etiketler, doküman özellikleri gibi anlamsal ve yapısal veriler de içermektedir.

Web yapı verisi: Bir Web sitesinden diğer bir Web sitesine ya da bir Web sayfasından diğer bir Web sayfasına yapılan bağlantı yapısının kesin ve açık olarak belirtilmesidir. Yani, Web bağlantılarının organizasyonunu gösteren bilgilerdir. Bu bilgiler, Web tasarımcısının siteye bakış açısını göstermektedir. Web sitesi yapı verisi, site haritalama araçları ile otomatik olarak oluşturulan sitenin harita bilgisidir.

Web kullanım verisi: İnternet kullanıcılarının Web kaynaklarını kullanmalarından yansıyan erişim kayıt verileridir. Bu veriler vekil sunucu kayıtlarında, Web sunucu kayıtlarında ya da Web tarayıcısının geçmişinde bulunan İnternet geçici dosyalarında tutulmaktadır. Bu dosyalar içerisinde kullanıcı IP adresi, sayfa referansları, bağlantı saatleri ve tarihleri, kullanıcının İnternet tarayıcısının adı ve sürümü gibi birçok önemli bilgiler yer almaktadır.

Web kullanıcı profili verisi: Web sitesine kayıtlı olmuş kullanıcılar hakkında demografik bilgilerin sağlandığı verilerdir. Bir siteye kayıt olmak isteyen kullanıcı ya da müşterilerden alınan bilgiler, bu veriler içerisinde yer almaktadır. Bu tür verilerin elde edilebilmesi için İnternet kullanıcısının Web sitesine kayıt yaptırması gerekmektedir. Web madenciliğinde kullanılan verilerin sınıflandırılması Tablo 2.1'de gösterilmektedir [1].

Tablo 2.1. Web madenciliğindeki verilerin incelenmesi

	Web Madenciliği		
	Web İçerik Madenciliği	***Web Yapı Madenciliği***	***Web Kullanım Madenciliği***
Veri	Metin belgeleri, HTML	HTML köprü bağlantıları	Sunucu ve tarayıcı kütükleri, çerezler, kullanıcı profilleri, sorgular, meta-veri.
Verinin Şekli	Yapısız ve karışık	Link Yapısı	Kullanıcı etkileşimi ve davranışı
Gösterilimi	İlişkisel, sınıflandırmalı	Grafiksel	İlişkisel tablolar ve grafiksel

2.1 Temel Tanımlamalar

Web madenciliği uygulamalarında sıkça kullanılan temel terimler mevcuttur. Bu tez içerisinde de kullanılacak olan ve World Wide Web Konsorsiyumu'nun (W3C) önermiş olduğu bu önemli terimler Tablo 2.2'de gösterilmektedir [103, 105].

Tablo 2.2. Web madenciliğinde kullanılan temel terimler

Terim	Açıklama
Web Tarayıcı	Web sayfalarında gezinmeyi sağlayan istemci yazılımlarıdır (IE, Mozilla, vb.).
Kaynak	Yararlanılan ve özdeşliği olan her şey bu sınıfa koyulabilir.
Web Kaynağı	HTTP protokolleri kullanılarak ulaşılabilen herhangi bir İnternet kaynağıdır.
Web Sunucu	Web materyallerini İnternet üzerinde erişim hizmetine açan yüksek hızdaki bilgisayarlardır.
Web Sayfası	İnternet ortamında yayınlanabilecek farklı uzantılardaki dokümanlardır (asp, html, php, vb).
Web Sitesi	Sunucuların İnternet'e sunduğu veritabanları ve bunlarla bağlantılı belgeler, dosyalardır.
Kullanıcı	Web tarayıcısını kullanan kişi, İnternet kullanıcılarıdır.
Web İsteği	Kullanıcının bir Web kaynağına yapmış olduğu isteklerdir.
Ziyaret	Kullanıcı oturumlarında yapılmış olan Web sayfalarının görüntülenmesidir.
Kullanıcı Oturumu	Bir kullanıcının sunucu üzerinde tanımlanmış ölçüde belli bir süre içerisinde kullanıcının ardı ardına görüntülediği Web istekleridir.
URI	İnternet Web sayfalarının fiziksel kaynağını tanımlayan karakterler zinciridir.
Oturum Tanımlama	Kullanıcının bir siteye oturum açmasından kapattığı zamana kadarki yapılan işlemlerinin belirli zaman aralığına göre sınıflandırılmasıdır.
Kullanıcı Tanımlama	Siteye erişen kullanıcıların tarayıcı ve kullanım özelliklerine göre sınıflandırılmasıdır.

2.2 Web İçerik Madenciliği

Web içerik madenciliği video, ses, görüntü, bağlantılı ve bağlantısız metinler içeren ve çoğu belli bir düzene sahip olmayan çoklu Web dokümanlarından otomatik olarak bilgi çıkarmayı amaçlamaktadır. Web içeriğinin aşırı derecede büyük verilerden oluşması nedeniyle, Web içerik madenciliğinde çok farklı madencilik teknikleri kullanılabilir. Web içerik madenciliği metotlarının çoğu, yapısal olmayan metin verileri ya da yarı yapısal HTML dokümanları temeline dayanmaktadır. Bu nedenle Web içerik madenciliğinin daha detaylı çalışma alanı metin madenciliği ya da metin sınıflandırılması olarak da adlandırılmaktadır [72]. Kosala ve Blockeel çalışmalarında Web içerik madenciliğinin uygulama alanlarını *veritabanı görünümü* ve *bilginin yeniden elde edilmesi* olmak üzere iki kısımda incelemişlerdir [62].

1. *Veritabanı görünümü* ile Web'de bilgi sorgulama ve karmaşık bilgilerin yönetiminin daha iyi yapılabilmesi için Web verilerini modellemektir. Yani, Web sitesini veri tabanına dönüştürmek veya Web sitesinin yapısından sonuç çıkarmaya çalışmaktadır. Bu uygulamadaki madencilik sonuçları, Web veri tabanları ve Web veri ambarlarını oluşturmak için kullanılabilir.

2. *Bilginin yeniden elde edilmesi* ise filtrelenen bilgilerde, çıkarılan sonuçlarda ya da istenen kullanıcı profillerinde bulunan bilgileri iyileştirmek veya değerlendirmektir. Bu uygulamanın madencilik sonuçları, Web arama motorlarına, Web'in kişiselleştirilmesine, Web sitesinin uyarlanan arabirimlerine ve önerilen sistemlere uygulanabilir.

Web içerik madenciliği uygulamaları, içerik temelli uyarlamalı öneri sistemlerinde sınıflandırılmış Web dokümanlarını, kümelenmiş Web sayfalarını, Web site içeriklerinin karşılaştırılması, doküman yapısının modellenmesi ve desteklenen diğer Web madenciliği uygulamalarını kapsamaktadır.

2.3 Web Yapı Madenciliği

Web yapı madenciliği, Web sitesinin yapısını oluşturan bağlantılardan yararlı bilgileri keşfetmeye çalışmaktadır. Web içerik madenciliği Web sayfasının içeriği ile ilgilenirken, Web yapı madenciliği ise doğrudan Web sayfaları arasındaki bağlantıları incelemektedir. Web sayfaları ya da bir Web sayfasındaki bağlantılar (grafik-yazı, grafik-grafik, resim-yazı vb.) arasındaki ilişkileri inceleyerek sonucunda bilgi üretir. Örneğin, kullanıcı Google arama motorunu kullanarak arama yapılır. Arama öncesinde kullanıcı için önemli olan Web sayfaları Google arama motorunda tanımlanırsa, arama sonucunda Google o sayfaları bulduğunda önemli sayfalar olarak işaretler.

Web yapı madenciliğinin amacı, Web sitesi ve Web sayfaları içerisindeki ilgili bağlantı verisine bakarak istenilen bilgiyi keşfetmektir. Web yapı madenciliğinin kullandığı yapısal bilginin tipi iki farklı kısımda incelenmektedir.

1. **Sayfa bağlantıları:** Aynı Web sayfası içindeki bir noktaya ya da farklı bir Web sayfasına bağlantı sağlayan sayfalarının fiziksel kaynağını tanımlayan karakterler zinciridir. Aynı sayfa içindeki bir noktaya yapılan bağlantıya *iç doküman bağlantısı*, farklı sayfalara yapılan bağlantıya da *dış doküman bağlantısı* denilmektedir.
2. **Doküman yapısı:** HTML ya da XML biçimindeki bir Web sayfası içerisindeki yapı, ağaçsal yapı şeklinde organize edilebilir. Bir Web sitesinin organizasyonu ve dokümanlarının düzenlenmesi bu alana girmektedir. Özellikle, bir Web sitesindeki dokümanların yapısını otomatik olarak çıkarmak için yapı madenciliği yöntemleri uygulanmaktadır.

2.4 Web Kullanım Madenciliği

Web kullanım madenciliği, Web sitesinin kullanım analizi için geleneksel veri madenciliği yöntemlerini kullanarak Web sunucularında tutulan kullanıcı erişim kayıtlarından en yoğun ve en ilginç kullanıcı erişim örüntülerini keşfetmeyi ve anlamlı verileri çıkarmayı amaçlamaktadır. Web ve vekil sunucularında tutulan kullanıcı erişim kayıtları, tarayıcı kayıtları, kullanıcı profilleri, çerezler, fare klikleri, sayfa kaydırmaları ve kullanıcıların Web ile olan etkileşimlerinden oluşan tüm kayıtlar Web kullanım verilerini içermektedir. Web kullanım madenciliğinde kullanılan verilerinin sınıflandırılması kullanım verilerinin türüne bağlıdır.

Web sunucu verisi: Web sunucusu tarafından kaydı tutulan kullanıcı erişim kayıtlarıdır. Farklı biçimlerde tutulan Web kayıtları ile ilgili daha ayrıntılı bilgi 3. bölümde verilmektedir.

Uygulama sunucu verisi: Elektronik ticaret uygulamalarında kullanılan ticari uygulama sunucularında tutulan çok önemli verilerdir. Yani, uygulama sunucusunda bulunan müşteri özelliklerine ait kayıtların ve iş olaylarının yapıldığı izlerin tutulduğu önemli bilgilerdir.

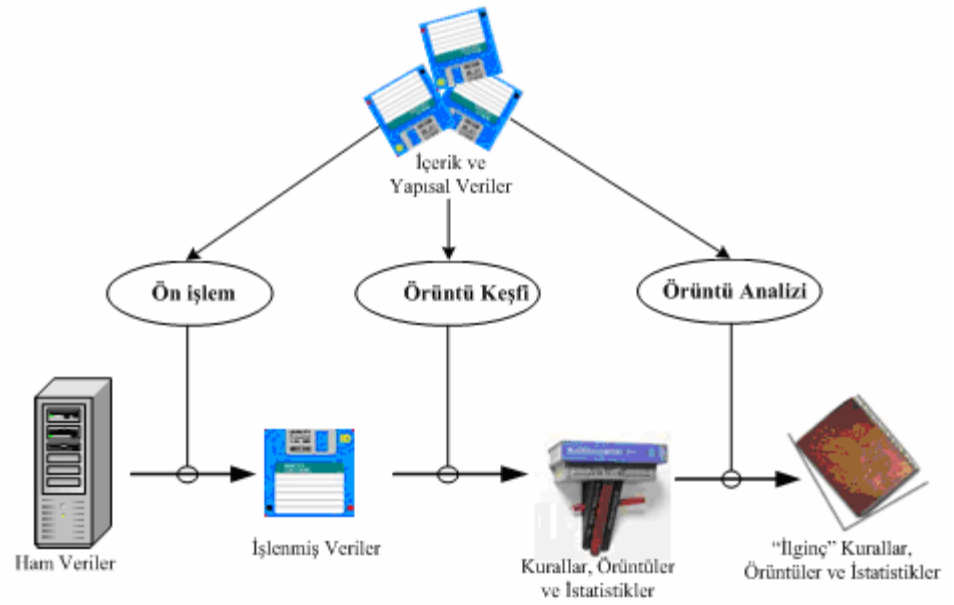
Uygulama seviye verisi: Bir Web uygulamasında yapılan olayların yeni çeşitlerinin tanımlanmasıdır.

2.4.1 Uygulama Aşamaları

Web kullanım madenciliği süreci iki yaklaşıma göre hareket etmektedir. Birinci yaklaşım, bir veri madenciliği tekniği uygulamadan, öncelikle ilişkisel tabloların içerisine Web sunucunun kullanım verisinin planını yapmaktır. İkinci yaklaşım ise, özel veri temizleme teknikleri kullanarak doğrudan günlük veri kayıtlarını kullanmaktır. Tipik veri madenciliği tekniklerini uygulamadan önce veri temizleme işlemleri, veri kalitesini arttırmak için çok önemli bir rol oynamaktadır.

Web kullanım madenciliğinin üst seviyedeki işlemlerin uygulama adımları *ön işlem, örüntü keşfi* ve *örüntü analizi* olmak üzere 3 kısımda incelenmektedir [62]. Bu aşamalar Şekil 2.2'de gösterilmektedir [1].

Ön işlem: Web kullanım madenciliğinin ilk aşamasında, Web sunucularından alınan karmaşık ve düzensiz biçimdeki erişim kayıt dosyalarında, analiz değeri olmayan ilişkisiz alanlardan arındırılarak belirli bir düzene getirilmesi sağlanmaktadır. Karmaşık ve zor olan bu işlem süreci, içeriğe göre farklılık gösterebilir. Web kullanım madenciliğinin bu süreci ile ilgili ayrıntılı bilgiler tezin 3. bölümde sunulmaktadır.



Şekil 2.2. Web kullanım madenciliğinin genel uygulama adımları

Örüntü keşfi: Örüntülerin keşfi aşamasında, düzenli ama anlamsız olan verilerden önemli ve gerekli olan bilgiyi ortaya çıkarma yapılmaktadır. Web madenciliğinde örüntü keşfi için birçok yöntem ve algoritma bulunmaktadır [8, 20, 67]. İstatistiksel analiz, ilişkilendirme kuralları, kümeleme, sınıflandırma, sıralı örüntüler ve bağımlı modelleme gibi teknikler kullanılmaktadır. Bu aşamaya ait konular tezin 4. bölümde ayrıntılı olarak verilmektedir.

Örüntü Analizi: Son aşama olan örüntü analizi sürecinde, örüntü keşfi aşamasında elde edilmiş ilginç olmayan kurallar, istatistikler ya da örüntülerin filtrelenmesidir [8, 20, 67]. Yaygın olarak kullanılan bilgi sorgulama mekanizmaları SQL, MySQL gibi veritabanı uygulamaları ve On-Line Analytical Processing (OLAP) uygulamaları bu aşamada gerçekleştirilmektedir. Web kullanım madenciliğinde örüntü analizi aşaması tezin 5. bölümde ayrıntılı olarak derlenmiştir.

2.4.2 Uygulama Alanları

İnternet kullanıcı talepleri ile ilgili hizmetlerin yeterliliği, Web sayfalarının kullanma durumları, kullanıcı oturumları ve kullanıcı davranışlarıyla üretilen erişim kayıtlarının analiz edilmesi gibi birçok konu Web kullanım madenciliğinin uygulama alanlarına girmektedir. Bu konular ise *kişiselleştirme, sistem geliştirme, Web sitesi yenileme, iş zekâsı* ve *kullanım karakteristiği* başlıkları altında toplanmaktadır. Web kullanım madenciliği, sunucu erişim kayıtlarının yanı sıra, hareket şablonu, site topolojisi, kavram hiyerarşileri gibi arka plan veya alan bilgilerini de kullanmaktadır. Web kullanım madenciliği için başlıca uygulama alanları Şekil 2.3'de gösterilmiştir [10].

Şekil 2.3. Web kullanım madenciliğinin başlıca uygulama alanları

Kişiselleştirme: Bir Web kullanıcısının kişiselleştirilmesi, ziyaret ettiği Web sitesi üzerindeki kullanım davranışları, kullanıcı profili gibi bilgilere göre sınıflandırma yaparak, sonraki davranışları için tahminler yaparak bu kullanıcıya istekleri doğrultusunda öneriler sunabilmektir. Örneğin, elektronik ticaret yapan bir Web sitesi üzerinde alışveriş yapan kayıtlı bir kullanıcının aldığı ürünlere ve site üzerindeki davranışlarına göre bir sonraki ziyarette bu kullanıcının davranışlarına yönelik tahminler yapılabilmektedir. Web kullanım madenciliği bu uygulamaları gerçekleştirmek için mükemmel bir yaklaşımdır. Tablo 2.3'de görüldüğü gibi kişiselleştirilme amacıyla WebWatcher [73], Letizia [74], Krishnapuram [75], Analog [40], WebPersonalizer [76], SiteHelper [77] gibi birçok projeler geliştirilmiştir. Bu yazılım projelerinde, kullanıcıların sahip olduğu benzer erişim örüntülerinin kümelemelerini keşfetmek için Web sunucu kayıtları kullanılmıştır.

Sistem Geliştirme: Web kullanıcı memnuniyetini ve Web kullanım aktivitelerini yüksek kaliteye çıkarmak için Web sunucu başarımını ve diğer servis özelliklerini arttırmak gerekmektedir. Benzer şekilde, bir yerleşke ağındaki kullanıcıların Web servisleri ve sistem sunucularının hizmetlerindeki kaliteyi arttırmakla mümkündür [65].

Web kullanım madenciliği bir yerleşke ağındaki sisteme yönelik ağ iletimi, Web yükleme, yük dengeleme, ağ güvenliği veya veri dağıtımı gibi Web trafik davranışlarını anlamak için temel anahtar görevini teşkil etmektedir. Bunlara ilave olarak, yerleşke ağı içinden ya da dışından sunuculara yapılan saldırı ve ataklar ile sisteme zarar veren, dolandırıcılık ve hile ile kullanıcı şifrelerini elde etmeye çalışan kötü niyetli kullanıcıların tespitinde Web kullanım madenciliği ile keşfedilen örüntülerle destek vermektedir. Özellikle ağ sistemi üzerine yerleştirilen güvenlik duvarı cihazının tutmuş olduğu kullanım kayıtlarından sistem ve kullanıcılar hakkında birçok anlamlı bilgiler çıkarılabilir. Ağın kullanımı ile ilgili genel bilgiler, ağ trafik raporları, kullanılan bant genişliği oranları, sisteme saldırı ve atak yapan kullanıcılar, virüs kaynaklı kullanıcılar gibi birçok detay bilgi ve grafikler çıkarılabilir. Elde edilen bu anlamlı bilgiler ile sistem üzerindeki problemli noktalar veya sistemin aksayan yanları

çıkarılarak çözüm sağlanabilir. Sonuçta ağ sisteminin gelişimine ve başarımının artırılmasına büyük ölçüde katkı sağlanmaktadır. Tablo 2.3'de görüldüğü gibi sistem geliştirme alanında da yapılmış bir çok çalışmalar mevcuttur.

Web Sitesi Yenileme: Birçok Web uygulamalarının işleyişi, gerekliliği ve kullanımı açısından Web sitesinin çekiciliği hem içerik hem de yapı bakımından çok önemlidir. Örneğin; şirketlerin e-ticaret için kullandıkları bir ürün katalogu, çevrimiçi satış modülleri, üniversitelerin Web sitesi üzerinde aktif olarak kullandıkları öğrenci işleri otomasyonu, personel maaş otomasyonu, akademik bilgi sistemleri veya bankaların yaygın olarak kullandıkları İnternet bankacılığı Web modüllerinin etkili, yararlı ve son derece önemli oldukları aşikârdır. Bu Web sitelerini değerlendirmek ve geliştirmek için İnternet kullanıcılarından detaylı geri dönüş bilgileri almak gerekmektedir. Web kullanım madenciliği İnternet kullanıcı davranışlarını derin bir şekilde inceleyip, Web sitelerinin güncellenmesi ya da yeniden tasarlanması konusunda Web tasarımcılara, Web yöneticilerine ayrıntılı olarak rapor sunmaktadır.

Uyarlamalı Web sitesi projesi, [78, 79] bir sitenin içeriği ve yapısının yeniden tasarlanması için Web erişim kayıtlarından SCML algoritması ile otomatik olarak bilgi çıkarmaya yönelmiştir. Bu projede, sayfaların kümelenmesi direk olarak bağlantılı olan sayfaları tanımlamak için kullanılmıştır.

İş Zekâsı: Elektronik ticaretle uğraşan şirketlerin *Web sitesini hangi müşteri kitlesi nasıl kullanıyor* sorusuna cevap veren bilgilerin tespit edilmesi için araştırmalar yaptığı bir çalışma alanıdır. Buchner ve diğ. [80, 81], Web kayıtlarından akıllı alışveriş işlemlerini tespit etmek için bir bilgi keşfi uygulaması geliştirmişlerdir. Geliştirdikleri bu elektronik ticaret uygulamasında çok büyük olan alışveriş verileri ile aşırı derecede büyük olan kullanım veri kümelerini birleştirmişlerdir. Uygulamalarında geliştirdikleri bilgi keşfi teknikleri müşteri ilişkileri yaşamında müşterinin ilgisi, müşterinin devamlılığı, çapraz satış ve müşterinin Web sitesinden ayrılışı olmak üzere dört ayrı basamak tanımlamışlardır. Blue Martini çalışmasında, karar ağacı kuralları örüntülerini keşfetmek için içerik sunucusundan direk olarak elde ettiği fare klik verilerini kullanmıştır. Elektronik ticarette iş zekâsının amacına ulaşması için İnternetteki Web trafiklerini analiz eden SurfAid, Accrue , NetGenesis, Aria, Hitlist, WebTrends 14 gibi birçok ticari yazılımlar mevcuttur [65]. Ayrıca son yıllarda SAS Base [82] ve SPSS Clementine [83] yazılımları ticari şirketler tarafından iş zekası alanında aktif olarak kullanılmaktadır. Bununla birlikte Accrue, NetGenesis ve Aria yazılımları kullanım istatistiklerini kullanarak elektronik ticaret yapan site üzerindeki fare klikleri, ürün alışverişleri ve reklâmlar gibi olayları analiz etmek için geliştirilmişlerdir. Accrue yazılımı aynı zamanda Yol analizini görselleştiren bir

araçtır. IBM tarafından geliştirilen SurfAid yazılımı ise, sayfa görünüm istatistiklerinin yanı sıra kullanıcıların kümelenmesi ve veri küpü yoluyla OLAP işlemlerini desteklemektedir. Han ve diğ. [84] geliştirdikleri WebLogMiner sistemleri ile zaman serisi analizi (örneğin; eğilim analizi, değişim analizi, sıralı analiz gibi) sınıflandırma uygulamaları ve birliktelik kurallarını çıkarmak için kullanılmaktadır. İş zekâsı çözümünde SAS Enterprise BI yazılımı [83], kolay kullanımı ile raporlama, sorgulama, analiz, OLAP, görselleştirme ve ofis verilerinin birleşimi gibi uygulamalarla madencilik alanında çözümler üretmektedir. Tablo 2.3'de görüldüğü gibi iş zekası alanında da yapılmış çalışmalar mevcuttur.

Kullanım Karakteristiği: Web kullanım madenciliği ile Web karakterizasyon araştırması arasında büyük oranda bir örtüşme vardır. Pitkow ve diğ. [85, 86, 87] Georgia Teknoloji Enstitüsü'nde geliştirmiş oldukları Xmosaic adlı Web tarayıcı yazılımı ile istemci taraflı aktivitelerin kaydetme işlemlerini gerçekleştirmişlerdir. Özellikle tarayıcılar aracılığıyla bir Web sitesiyle etkileşim halinde bulunan kullanıcılardan elde edilen kayıtların sonuçları, kullanıcıların davranışları, kullanım karakteristiği hakkında detaylı bilgiler sunmaktadır. Tablo 2.3'de gösterildiği gibi kullanım karakteristiği alanında da yapılmış birçok çalışmalar bulunmaktadır [65].

Web kullanım madenciliği ile ilgili yapılmış araştırma projelerinde birbirinden farklı birçok yazılım geliştirilmiştir. Bu araştırma projelerini uygulama alanlarına, kullandıkları veri kaynaklarına ve veri tiplerine göre sınıflandırmak mümkündür. Tablo 2.3'de geliştirilen yazılımların çoğu sunucu temelli verileri kullanmaktadır. Tabloda görüldüğü üzere yazılım projelerinin tümü kullanım verilerini, birkaçı ise kullanımın yanı sıra yapı, içerik veya profil verilerini kullanarak analiz yapabilmektedir. Tek kullanıcılı projeler genellikle *kişiselleştirme* uygulama alanını içermektedir. Çoklu site analizini destekleyen projelerde ise birden fazla Web sitesinin kullanım verilerine kolayca erişebilmek için ya istemci ya da vekil sunucu seviyesinde giriş verileri kullanılmaktadır. Çoğu Web kullanım madenciliği projelerinde tek ve çok kullanıcılı siteler, Web sunucu kayıtları gibi sunucu temelli kullanım verileri kullanılmaktadır.

Tablo 2.3. Web kullanım madenciliği araştırma projeleri ve yazılımları [65]

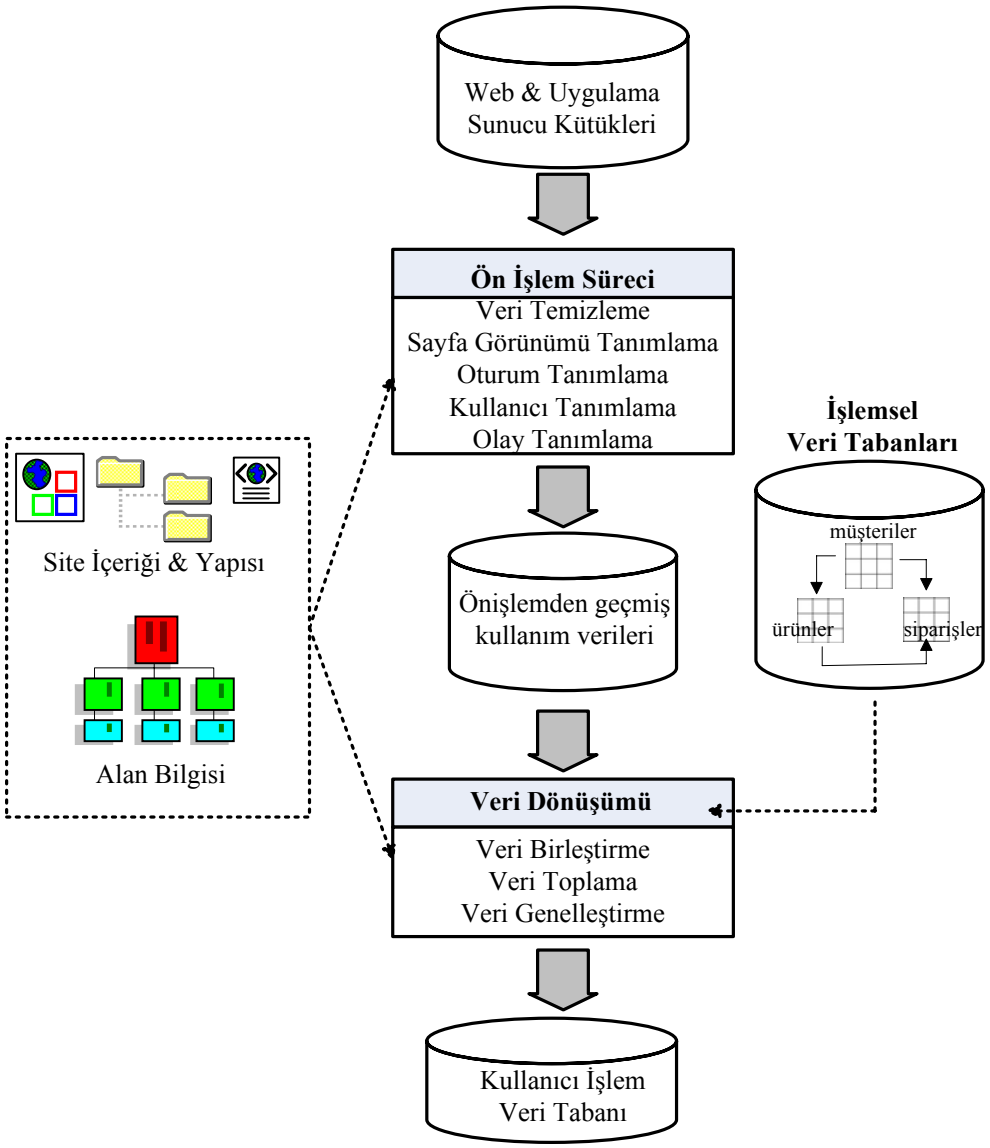
Proje Adı	Uygulama Alanları	Veri Kaynağı			Veri Tipi				Kullanıcı		Site	
		Sunucu	Vekil	İstemci	Yapı	İçerik	Kullanım	Profil	Tek	Çok	Tek	Çok
WebSIFT	Genel	X			X	X	X			X	X	
SpeedTracer	Genel	X					X			X	X	
WUM	Genel	X			X		X			X	X	
Shahabi	Genel			X	X		X			X	X	
Site Helper	Kişiselleştirme	X					X		X		X	
Letizia	Kişiselleştirme			X			X		X			X
Web Watcher	Kişiselleştirme		X				X	X		X		X
Krishnapuram	Kişiselleştirme	X					X			X	X	
Analog	Kişiselleştirme	X					X			X	X	
Web Personalizer	Kişiselleştirme	X			X		X			X	X	
Tuzhilin	İş	X					X			X	X	
SurfAid	İş	X					X			X	X	
Buchner	İş	X					X	X		X	X	
WebTrends,Hitlist,Accrue	İş	X					X			X	X	
WebLogMiner	İş	X					X			X	X	
PageGather,SCML	Site Yenileme	X			X		X			X	X	
Manley	Karakterize etme	X					X			X		X
Arlitr	Karakterize etme	X					X			X		X
Pitkow	Karakterize etme	X		X			X			X		X
Almeida	Karakterize etme	X					X			X		X
Rexford	Sistem Geliştirme	X	X				X			X	X	
Sxhechter	Sistem Geliştirme	X					X			X	X	
Aggarwal	Sistem Geliştirme		X				X			X	X	

3. VERİLERİN TOPLANMASI VE ÖN İŞLEM SÜRECİ

Web kullanım madenciliği uygulama sürecinin en önemli aşamalardan birisi de, veri madenciliği ve istatistiksel algoritmaların uygulanabileceği uygun hedef veri kümesinin oluşturulmasıdır. Web kullanım madenciliğinde ön işlem süreci, tıklanılan verinin karakteristikleri ve çoklu kaynaklardan toplanan diğer ilgili verilerin birbirleriyle olan ilişkilerinden dolayı özellikle büyük önem arz etmektedir. Web kullanım madenciliği sürecinde veri hazırlama işlemi yoğun zamanın harcandığı ve hesaplama uygulamalarının yoğunlaştığı bir aşamadır. Bu aşamada, özellikle metin tabanlı verilerden yararlı ve gerekli örüntüleri başarılı bir şekilde çıkarmak çok önemlidir. Özel veri madenciliği çalışmalarında, giriş verisinin uygun formatta bir veri kümesi haline getirilmesi için orijinal ham verilerin ön işlemi yapılmalı, çoklu kaynaklardan alınan verilerin birleştirilmesi sağlanmalı ve birleştirilen verilerin ortak bir veri tabanına dönüşümü gerçekleştirilmelidir. Web kullanım madenciliği uygulamasının bu aşaması için birçok farklı yazılımların kullanılması mümkündür [27-42].

Web kullanım madenciliğinin ön işlem sürecinde, hem pratik uygulamalarda hem de araştırma çalışmalarının çoğunda esas odaklanılan nokta, farklı analizler için web kullanım verilerinin hazırlanması ve bütünleştirilmesidir. Web kullanım verilerinin hazırlanmasında birçok yazılım uygulamaları ve gerekli bilgilerin toplanması için sezgisel (heuristic) teknikler kullanılmaktadır. Örneğin, verilerin birleştirilmesi ve temizlenmesi, kullanıcı ve oturum tanımlanması, sayfa görünümü tanımlanması ve web olaylarının tanımlanması gibi bilgilerin elde edilebilmesi için sezgisel teknikler kullanılmaktadır [9]. Web kullanım verilerine veri madenciliği yöntemlerinin başarılı bir şekilde uygulanabilmesi, ön işlem sürecindeki işlemlerin doğru uygulanmasına büyük oranda bağımlıdır. Örneğin, elektronik ticaret yapan bir web sitesinin içeriğinin veri analizinde site metrikleri, bilinçli kullanıcıların tespiti gibi önemli keşifler için erişim kayıtlarındaki anlamlı bilgiler doğru olarak ortaya konulmalıdır.

Web kullanım verilerinin ön işlem aşamasındaki ana uygulamalar ve işlemlerin genel yapısı Şekil 3.1'de gösterilmektedir. Web kullanım madenciliğinin ön işlem aşamasında yapılması gereken temel anahtar işlemler verilerin temizlenmesi, kullanıcı ve oturumların tanımlanması, sayfa görünümü tanımlanması, verilerin birleştirilmesi ve verilerin dönüştürülmesi gibi uygulamalardır.



Şekil 3.1. Web kullanım madenciliği için veri ön işlem aşamaları

Ön işlem aşamasında, sunucu üzerinden alınan karmaşık ve düzensiz biçimde bulunan erişim kayıt dosyalarındaki verilerin, analiz değeri olmayan ilişkisiz sahalarından arındırılması, ayıklanması ve belirli bir düzene getirilmesi gerekmektedir. Karmaşık ve zor olan bu işlem süreci, içeriğe göre farklılık gösterebilir. Bu süreçte önemli olan toplanan veriler içerisindeki ilgisiz ve gereksiz satırların ayıklanarak, kullanılmaya hazır veritabanlarındaki bilgilerle bütünleştirilmesi, örüntü keşfi aşamasında kullanılabilir hale getirmektir.

3.1 Veri Kaynakları ve Veri Tipleri

Web kullanım madenciliğinde kullanılan ana veri kaynakları sunucularda tutulan kayıt dosyalarıdır. Bu kayıt dosyaları *web sunucu erişim kayıtları* ve *uygulama sunucu kayıtları* olmak üzere iki kısımdan oluşmaktadır. Bunların yanı sıra, hem veri hazırlama hem de örüntü keşfi için gerekli olan ilave veri kaynakları da bulunmaktadır. Bu kaynaklar ise site dosyaları, meta veriler, uygulama şablonları ve etki alanı bilgisi gibi bilgileri içermektedir. Çeşitli kaynaklar yoluyla elde edilen verileri, içerik, yapı, kullanıcı profili ve kullanım verisi olmak üzere dört ana grupta kategorize edilebilir. Bu veriler web sunucusu, istemci ve vekil sunucusu gibi farklı kaynaklardan elde edilebilir.

3.1.1 Kullanım Verisi

Web ve uygulama sunucuları üzerinde otomatik olarak toplanan ve kaydedilen, Internet kullanıcılarının durumsal davranışlarına ait bilgilerin bulunduğu erişim kayıt dosyalarıdır. Web kullanım madenciliğinde kullanılan verinin ana kaynağı kullanım verileridir. Kullanıcının sunucuya karşı yapmış olduğu her bir HTTP isteği, sunucu erişim kayıtlarında tek bir satır olarak üretilmektedir. Kayıt biçimine bağlı olarak, genelde kayıt girişi içerisinde kullanıcı IP adresi, istenilen URL adresleri, sayfa referansları, bağlantı saatleri ve tarihleri, kullanıcı Internet tarayıcısının adı ve sürümü, kullanıcı taraflı çerezler, web uygulamalarında kullanılan muhtemel parametreler gibi birçok önemli bilgiler yer almaktadır [88].

Kullanıcı erişim kütükleri, farklı işletim sistemleri üzerinde çalışan web sunucuları ile vekil sunucuları üzerindeki yazılımların sakladığı erişim kayıtlarının biçimleri birbirinden farklı olabilir. Örneğin, Linux işletim sistemi üzerinde çalışan bir Apache web sunucusu ile Windows Server 2003 işletim sistemi üzerinde çalışan bir IIS (Internet Information Server) web sunucusunun ürettiği erişim kütüklerinin biçimi birbirinden farklıdır. Tablo 3.1'de Linux işletim sistemi tabanlı bir vekil sunucunun tutmuş olduğu kullanıcı erişim kütükleri görülmektedir [1].

Tablo 3.1. Vekil sunucusunda tutulan erişim kütüklerinden örnek kesit

```
CP_IMS_HIT/304 253 GET http://img.sabah.com.tr/i/topbar_kaydet.gif - NONE/- image/gif011623
4612 10.6.2.20 TCP_IMS_HIT/304 254 GET http://img.sabah.com.tr/i/yazar_yukari.gif - NONE/-
0 1335 GET http://anket.memurlar.net/images/common/member6.gif - NONE/- image/gif0116236177
200 6040 GET http://img245.imageshack.us/my.php? - DIRECT/38.99.76.207 text/html01162361772
10.6.2.20 TCP_IMS_HIT/304 253 GET http://img.sabah.com.tr/i/y/t/0002.gif - NONE/- image/gif
254 GET http://img.sabah.com.tr/i/tumhisseler_hdr.gif - NONE/- image/gif01162361772.626
rswebclubhouse.com/club/chat_member.php? - DIRECT/193.239.90.199 text/html01162361772.685
6.2.20 TCP_IMS_HIT/304 253 GET http://www.sabah.com.tr/i/anket_icin_tiklayiniz.gif - NONE/-
M_HIT/200 1110 GET http://www.sabah.com.tr/2006/11/01/gny/im/0647977A493AB745A63DE345e.gif
/302 615 GET http://ad.e-kolay.net/getad.a2? - DIRECT/83.66.160.10 text/html01162361772.940
_MISS/200 44604 GET http://www.internethaber.com/news_detail.php? - DIRECT/89.106.24.67 tex
http://ankt.memurlar.net/images/piechart.aspx? - DIRECT/209.85.10.99 image/gif01162361773.0
01162361773.086    650 10.6.2.95 TCP_MISS/200 381 GET http://kpss.osym.gov.tr/default.aspx
_MISS/200 312 GET http://ads.sabah.com.tr/adserver/adlog.ads? - DIRECT/213.74.5.114 image/g
.239 text/html01162361773.200     54 10.1.3.23 TCP_MISS/200 381 GET http://kpss.osym.gov.tr
10.6.2.20 TCP_IMS_HIT/304 253 GET http://www.sabah.com.tr/i/_spacer.gif - NONE/- image/gif0
p://kpss.osym.gov.tr/default.aspx - DIRECT/193.140.115.113 text/html01162361773.344     98
```

Microsoft IIS (Internet Information Server) web sunucusunda CLF (Common Log Format), ECLF (Extended Common Log Format), NCSA (National Center for Supercomputing Applications) olmak üzere 3 farklı biçimde kullanıcı erişim kayıt dosyaları tutulmaktadır.

Tablo 3.2. CLF biçimindeki erişim kütüklerinden örnek bir satır

1.2.3.4- - [31/May/2007:08:01:38 +0300] "GET /images/green.gif HTTP/1.1" 304 1346 "http://web.firat.edu.tr/index.asp" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)"

Tablo 3.2'de CLF biçiminde, Tablo 3.3'de ECLF biçiminde ve Tablo 3.4'de NCSA biçiminde kullanıcı erişim kütüklerinden örnekler gösterilmektedir.

Tablo 3.3. ECLF biçimindeki erişim kütüklerinden örnek bir satır

2008-01-14 22:45:27 W3SVC1 ICME 3.4.5.6 GET /default.asp - 80 – 5.6.7.8 HTTP/1.1 Mozilla/4.0+ (compatible;+MSIE+7.0;+Windows+NT+5.1;+.NET+CLR+2.0.50727;+.NET+CLR+3.0.04506.30; +.NET+CLR+1.1.4322) - /index.asp web.firat.edu.tr 200 0 0 13720 608 78

Tablo 3.4. NCSA biçimindeki erişim kütüklerinden örnek bir satır

3.4.5.6 - - [02/Jun/2007:00:00:12 +0300] "GET /Default.asp HTTP/1.1" 200 29191

Tablo 3.5'de ise Tablo 3.3'deki ECLF kütük örneğindeki tüm alanların bulunduğu genişletilmiş erişim kayıtlarındaki alanların açıklaması verilmektedir [104].

Tablo 3.5. Genişletilmiş erişim kütüklerindeki alanların açıklanması

Alan Adı	Örnek Veri	Açıklama
Tarih	2008-01-14	Kullanıcının istek yaptığı tarih bilgisini gösterir.
Zaman	22:45:27	Kullanıcının istek yaptığı saat bilgisini gösterir.
Servis adı ve durum numarası	W3SVC1	Internet servis adı ve istemci çalışmalarının kaydedildiği yer
Sunucu adı	ICME	Erişim kayıtlarının üretildiği ve tutulduğu sunucunun adı
Sunucu IP adresi	3.4.5.6	Web hizmeti veren sunucunun IP veya DNS adres bilgini gösterir (gizlilik nedeniyle gerçek IP değiştirilmiştir)
Metot	GET	İstenilen web isteği metodu
URI gövdesi	/enformatik/default.asp	İstenilen web isteği ya da adresi
URI kuyruğu	-	RFC931 veya kimlik tanımlamalarıdır. Özel tanımlamalar yapılmadığı sürece bilgi bulunmaz (sadece dinamik sayfalar için geçerlidir)
Sunucu port no	80	Kullanıcı tarafından web sunucusuna yapılan isteğin sunucudaki port numarası gösterir.
Kullanıcı adı	-	Web sitesini kullanan yetkili isimlerin listesidir. Sistem parola korumalıysa ve kimlik denetlemesini başarıyla geçmiş ise bu alanda kullanıcının adı gözükecektir (Genelde isimsizdir).
İstemci IP adresi	5.6.7.8	İstemciye ait IP adresi (gizlilik nedeniyle gerçek IP değiştirilmiştir)
Protokol sürümü	HTTP/1.1	İstemcinin kullandığı HTTP protokol sürümü
Kullanıcı etmenleri	Mozilla/4.0+(compatible;+MSIE+7.0;+Windows+NT+5.1;+.NET+CLR+2.0.50727;+.NET+CLR+3.0.04506.30;+.NET+ CLR+1.1.4322)	İstemcinin kullandığı tarayıcının tipi ve diğer özellikleri
Çerezler	-	Eğer varsa, gönderilen veya alınan çerezlerin içeriği
Referans	/index.asp	Kullanıcının en son ziyaret ettiği site adresi
Ana sunucu adresi	web.firat.edu.tr	Sunucunun çalıştığı sitenin başlangıç URL adresi
HTTP durumu	200	Bu kısım, sunucunun cevap verdiği durum kodunu içermektedir. (Örnekteki 404 nolu kod, ilgili adrese bağlantıda başarının sağlanamadığını belirten bir hata kodunu göstermektedir.) Bu kodlar RFC2616 teknik belgesinde belgelendirilmiştir.
Protokol alt durumu	0	Protokol alt durum hata kodu
Win32 durumu	0	Windows durum kodu
Gönderilen veri boyutu	13720	Sunucu tarafından gönderilen verinin boyutu
Alınan veri boyutu	608	Sunucu tarafından alınan verinin boyutu
Harcanan zaman	78	Kullanıcının sayfadaki hareketleri boyunca milisaniye olarak harcadığı zaman

Tablo 3.6'da görüldüğü gibi bir web sunucu erişim kütüklerinden altı bölümden oluşan tipik kütük örneklerinden bölümler gösterilmektedir. Ancak, erişim kayıtlarındaki kullanıcı IP adresleri kişisel gizlilikten dolayı değiştirilmiştir.

Tablo 3.6. Web sunucundaki erişim kütüklerinden örnek veriler

No	Web Kullanıcı Erişim Kütük Verileri
1	2008-01-14 00:29:09 1.2.3.4 - GET /ffmu/arsiv/makaleler.html - 200 8225 HTTP/1.1 web.firat.edu.tr Mozilla/4.0+(compatible;+MSIE+7.0;+Windows+NT+6.0;+SV1)+;+SLCC1; +.NET +CLR+2.0.50727) - http://www.firat.edu.tr
2	2008-01-14 00:29:09 1.2.3.4 - GET /ffmu/arsiv/19-4/sayi_4.pdf - 200 4096 HTTP/1.1 web.firat.edu.tr Mozilla/4.0+(compatible;+MSIE+7.0;+Windows+NT+6.0;+SV1)+;+SLCC1; +.NET+CLR+2.0.50727) - http://web.firat.edu.tr/ffmu/arsiv/makaleler.html
3	2008-01-14 22:29:42 3.4.5.6 - GET /ffmu/arsiv/19-4/9_Resul_Das.pdf - 200 391822 HTTP/1.1 web.firat.edu.tr Mozilla/4.0+(compatible;+MSIE+7.0;+Windows+NT+6.0) http://www.google.com.tr/search?hl=tr&q=hyperlink+analysis+in+the+web+usage+mining
4	2008-01-15 21:34:52 7.8.9.10 - GET /ffmu/arsiv/eski/ilan.html - 200 2897 HTTP/1.1 web.firat.edu.tr Mozilla/4.0+(compatible;+MSIE+7.0;+Windows+NT+6.0;+SV1) http://web.firat.edu.tr/ffmu/arsiv/eski/
5	2008-01-15 21:34:52 7.8.9.10 - GET /ffmu/arsiv/eski/styles.css - 200 1792 HTTP/1.1 web.firat.edu.tr Mozilla/4.0+(compatible;+MSIE+7.0;+Windows+NT+6.0;+SV1) http://web.firat.edu.tr/ffmu/arsiv/eski/ilan.html
6	2008-01-15 21:34:52 7.8.9.10 - GET /ffmu/arsiv/eski/banner.gif - 200 1696 HTTP/1.1 web.firat.edu.tr Mozilla/4.0+(compatible;+MSIE+7.0;+Windows+NT+6.0;+SV1) http://web.firat.edu.tr/ffmu/arsiv/eski/ilan.html

Tablo 3.6'daki 1 numaralı kütük verisinde, bir internet kullanıcısı 1.2.3.4 numaralı sunucu (*web.firat.edu.tr*) IP adresindeki */ffmu/arsiv/makaleler.html* web kaynağına eriştiği bilgisi görülmektedir. Aynı zamanda erişim kayıtlarındaki kullanıcı etmeni alanında, kullanıcıya ait Internet tarayıcısının tipi ve sürümü, kullandığı işletim sistemi ve sürümü gibi bilgilerde görülmektedir. Son kısımdaki *http://www.firat.edu.tr* URL adres bilgisi, internet kullanıcısının */ffmu/arsiv/makaleler.html* web adresine gelmeden önce girdiği referans bilgisini belirtmektedir. 2 numaralı erişim kütük verisindeki hedef dosya (*/ffmu/arsiv/19-4/sayi_4.pdf*) bilgisi görülmektedir. 3 numaralı kütük verisinde ise Google arama motorunda (*hyperlink analysis in the web usage mining)* sorgulamasını yapan bir kullanıcı, karşısına gelen arama sonuç listesinde */ffmu/arsiv/19-4/9_Resul_Das.pdf* adresini tercih etmiştir. 4-6 arasındaki 3 satırlık erişim kayıt verileri dikkat edilirse kullanıcının tek *HTTP* isteği sonucunda kaydedilmiş erişim kayıt verileridir. */ffmu/arsiv/eski/ilan.html* adresine giren kullanıcı 5 ve 6 numaralı kütük verilerini de erişim kayıt dosyasına beraber ekletmiştir. Kullanıcının girmiş olduğu sayfa içerisine gömülü olan banner resim dosyası (*banner.gif*) ve sayfa sitillerinin tanımlandığı sitil dosyası (*styles.css)* belirtmektedir.

Web kullanıcı erişim kayıtlarından istenilen anlamlı ve ilginç örüntülerin çıkarılması, farklı seviyelerdeki karmaşık erişim kayıtlarının düzenli biçimde toplanılarak gerekli dönüşümlerin sağlanılmasına ve ortak veritabanında bir araya getirilmesine bağlıdır. Web

kullanım madenciliğinde, veri soyutlamanın en temel işlemi sayfa görünümü/izleme yönteminin kullanılmasıdır. Sayfa izleme metodunda, bir Internet kullanıcısının web tarayıcısı üzerindeki tüm hareketlerini görüntülemek ve kullandığı tüm web nesnelerini bir araya getirerek kümeleyerek sunmaktır. Kavramsal olarak her bir sayfa görünümü/izleme, web sitesinde gezinen kullanıcıya ait özel işlemlerde kullanılan web kaynaklarının ya da web nesnelerinin kümelenmesi olarak tanımlanabilir. Örneğin, web sayfalarında gezinen bir kullanıcının bir makaleyi okuması, alışveriş sitesinde bulunan bir ürün sayfasını incelemesi ya da bir ürünü satın alması gibi olayların kümelenmesini sağlayarak, araştırma ve incelemeye sunmaktır. Kullanıcı davranışlarının toplanmasındaki en temel seviye ise oturum tanımlamaktır. Oturum tanımlama, tek ziyaret esnasında tek bir kullanıcı tarafından sayfa görünümlerinin bir sıralanmasıdır. Bir oturum kavramı, kayıt analiz işlemleri için anlamlı ve önemli olan bir oturumdaki sayfa izlerinin bir alt kümesi seçilerek, açıklanabilir.

Web erişim kütük dosyalarının analizi ile ilgili yapılmış birçok akademik çalışma bulunmaktadır [3,8-10,13-21,25-26, 47-50, 65, 69, 89]. Bu çalışmalarda kayıt dosyaları irdelenmiş ve kayıtların analizi ile ilgili yapılan uygulama çalışmaları ayrıntılı olarak sunulmuştur [27, 30-32, 42]. Ayrıca, web kayıtları ile ilgili çeşitli akademik çalışma ve yazılımlar mevcuttur [28-29, 33, 56, 61, 71].

3.1.2 İçerik Verisi

Kullanıcıların eriştiği ve kullandıkları grafik, resim, şekil, ses ve görüntü dosyaları gibi gerçek verilerden oluşan web içerik sayfalarıdır. Ayrıca, bir web sitesindeki tanımlayıcı kelimeler, anlamsal etiketler, doküman özellikleri, *java/asp* kodlama dili ile üretilen dinamik *HTML/XML* sayfaları anlamsal ve yapısal veriler içermektedir. Web içerik madenciliğinde, içerik verisine göre çalışma ve uygulamalar yapılmaktadır. Özellikle bir sitenin oluşturulmasında ve hiyerarşik bir yapı kazandırılmasında web içerik madenciliğinin etkin bir rolü bulunmaktadır.

3.1.3 Yapısal Veri

Bir web sitesinden diğer bir web sitesine ya da bir web sayfasından diğer bir web sayfasına yapılan bağlantı yapısının kesin ve açık olarak belirtilmesidir. Yani, site içerisindeki içerik organizasyonunun tasarım görünümünü sunan yapısal verilerdir. Bu veriler, web tasarımcısının siteye bakış açısını göstermektedir. Web sitesi yapı verisi, site haritalama araçları ile otomatik olarak oluşturulan sitenin harita bilgisidir.

3.1.4 Kullanıcı Profili Verisi

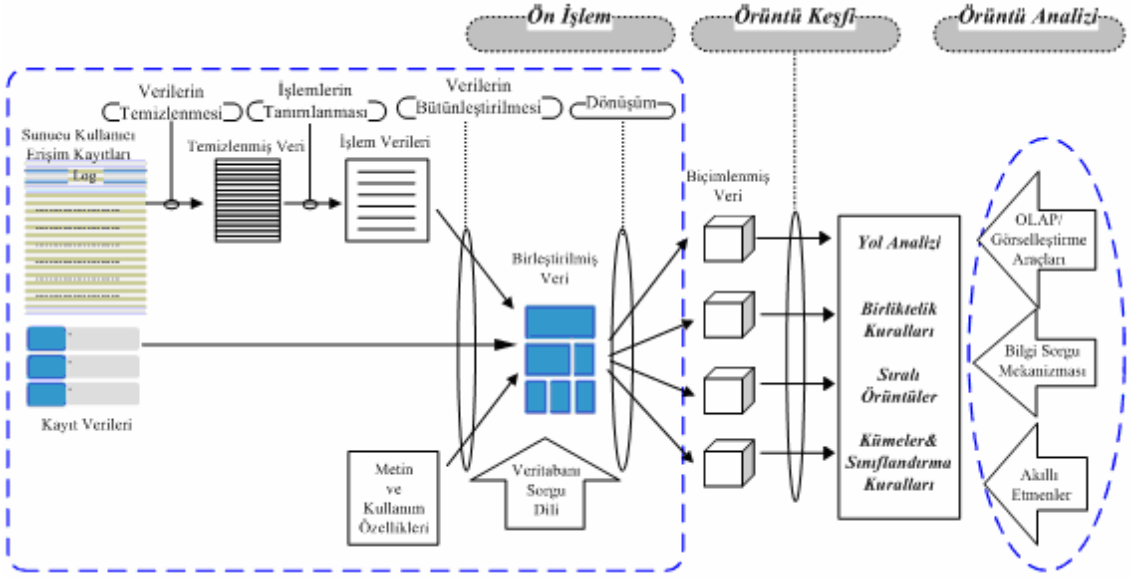
Web sitesine kayıtlı kullanıcılar hakkında demografik bilgilerin bulunduğu veritabanlarındaki ek verilerdir. Bir siteye kayıt olmak isteyen kullanıcı ya da müşterilerden alınan özel kayıt bilgileri, kullanıcının ilgi duyduğu alanlar, geçmişte yapmış olduğu alışveriş bilgileri ve ilgilendiği ürünler gibi birçok kişisel bilgiler, bu veriler içerisinde yer almaktadır. Bu tür verilerin tutulabilmesi için Internet kullanıcısının web sitesine doğru bilgilerle kayıt yaptırması gereklidir. Ancak, isimsiz olarak kayıt yaptıran kullanıcıların kaydettiği bilgiler, diğer kullanıcıların bilgilerinden kolaylıkla ayırt edilebilir.

3.2 Veri Temizleme

Verilerin temizlenmesi işlemi, web kullanım madenciliğinin ön işlem aşamasında uygulanması gereken ilk adımdır. Elde edilen ham erişim kayıtların tamamı örüntü keşfi için geçerli olan veriler değildir. Bu nedenle, kullanıcı erişim kayıtları içerisindeki verilerden geçerli ve gerekli olan kayıtlar seçilerek alınmalıdır. Bu nedenle, veri temizleme aşamasında aşağıdaki adımlar izlenilerek erişim kayıt dosyalarındaki ilgisiz girişler elimine edilmelidir.

1. ***Adım:*** Veri temizleme işleminde, kullanıcı erişim kayıtlarındaki başarısız ve hatalı giriş satırları kaldırılmalıdır.
2. ***Adım:*** Arama motorları tarafından otomatik olarak üretilen bazı erişim kayıt satırları tanımlanmalı ve kullanılacak erişim kayıt verilerinin içinden kaldırılmalıdır.
3. ***Adım:*** Kullanıcılar bir web sayfasına istekte bulunduğu zaman o sayfa içerisindeki resim, grafik, görüntü, ses gibi içerik verileri de erişim kayıtları içerisinde otomatik olarak kaydedilmektedir. Bunlara ilaveten sayfa içerisindeki Java scriptleri ve sayfa sitil dosyaları da erişim kayıtları içerisinde yerini almaktadır. Bu nedenle, analiz yapılacak erişim kayıtları içerisindeki resim, görüntü, script ve sitil (jpg, jpeg, gif, ico, avi, js, css) ve benzeri gibi uzantılı dosyaların bulunduğu kayıt satırları elimine edilmelidir.
4. ***Adım:*** Kullanıcı erişim kayıtlarındaki her satırın bir HTTP durum kodu bulunmaktadır [90]. Erişim kayıtlarının analizi işlemlerinde dikkat edilmesi gereken bir alan bilgisidir. 200 ile 299 arasındaki durum kodları başarılı girişler olduğundan, bunlar dışındaki durum kodlarının bulunduğu satırlar elimine edilebilir. Ancak, kayıtlar içerisindeki hata kayıtları veya engelli girişler gibi veri analizi yapılacaksa, durum kodlarının numaraları göz önünde bulundurulmalıdır.

Web kullanım madenciliğinin genel mimari yapısı ve ön işlem aşamasındaki işlem basamakları Şekil 3.2'de gösterilmektedir [105].



Şekil 3.2. Web kullanım madenciliğinin genel mimari yapısı

3.3 Sayfa Görüntüleme

Sayfa görüntüleme, öncelikli olarak site alan bilgisi ve sayfa içeriklerine bağlı olmasına rağmen büyük oranda da sitenin içinde bulunan sayfaların yapısıyla da ilgilenmektedir. Her bir sayfa görünümü, kullanıcı davranışlarının oluşturduğu kaynak bilgiler ya da web nesnelerinin bir toplamı olarak görüntülenebilir. Örneğin, bir bağlantı adresinin tıklanması, site üzerindeki alışveriş listesine yeni bir ürün eklenmesi, bir ürün sayfasının görüntülenmesi gibi işlemler kullanıcının web sitesi üzerindeki davranışlarını ortaya koyan olaylardır. Tek çerçeveli statik bir web sitesi için her bir HTML dosyası sayfa görünümü ile birebir uygun hale getirilmelidir. Bununla birlikte, çok çerçeveli sitelerde de verilen sayfa görünümü ile bütünlük içinde bir oluşum sağlanmalıdır. Dinamik siteler için bir sayfanın görünümü, uygulama sunucuları tarafından üretilen içerik ve statik şablonlarının bir araya getirilerek sunulmasıdır. Yani, sayfa görünümü tanımlanması işlemi, yüksek düzeyde elde edilen sayfalar içerisinde arzu edilen içeriğe ve kategoriye göre sınıflandırma işlemidir. Veri madenciliği aktivitelerinin çeşitliliğine esnek bir yapı sağlamak için her bir sayfa görünümü sayfa öznitelikleri ile kaydedilmelidir. Bu öznitelikler sayfa görünümü numarası, statik sayfa görünümünün tipi (örneğin; sayfa bilgisi, ürün görünümü, kategori görünümü ya da indeksleme) ve diğer meta veri (örneğin; içerik özellikleri, anahtar kelimeler veya ürün özellikleri) gibi bilgileri içermektedir.

3.4 Kullanıcı Tanımlama

Web kullanım analizinde bir kullanıcının kimlik bilgilerinin tespit edilmesine gereksinim yoktur. Ancak, web sitelerini ziyaret eden farklı kullanıcıları birbirlerinden ayırt etmek gereklidir. Çünkü bir kullanıcı birden fazla siteyi veya aynı siteyi birden çok oturum süresi içerisinde ziyaret edebilir. Çoklu oturumlarda her kullanıcıya ait davranış bilgileri sunucu kayıtlarında tutulmaktadır. Web kullanım madenciliğinde, *kullanıcı aktivite kaydı* aynı kullanıcıya ait olan sıralı olan kayıt aktiviteleri olarak tanımlanmaktadır.

Kullanıcı kimlik doğrulama mekanizmasının olmaması, yaygın olarak bilinen birçok kullanıcı tanımlama metodunda eşsiz kullanıcıları birbirinden ayırt etmek için istemci taraflı çerezler kullanılmaktadır. Çerezler (cookie) sayesinde istemcinin bilgisayarında bazı bilgiler saklanabilir. Çerezlerdeki bu bilgiler isim-değer çiftleri şeklinde tutulurlar ve kullanıcının bilgisayarına yollanırken de bu şekilde yollanırlar. Bu veriler, kişisel bilgilerin veya sitede o kullanıcıyla ilgili bilgilerin saklanması için kullanılabilir. Ancak, bütün siteler için çerezlerin saklanması mümkün değildir. Kullanıcı özel bilgilerinin gizliliğinden dolayı kullanıcılar çoğu zaman istemci taraflı çerezlerin tutulmasını engellemektedirler.

Çerez, web sayfası sunucusunun Internet kullanıcı bilgisayarının hard diskine yerleştirdiği ufak bir metin dosyası için kullandığı bilgiye denilmektedir. Yani, o web sitesinin bir sonraki ziyarette sizi tanıması için geçerli olan bir bilgidir. Bu metin dosyası bir komut dosyası olmadığı gibi, virüs de değildir. Kullanıcıya özel bir tanımlamadır ve sadece bu kodu tanımlayan web sunucusu tarafından okunabilir. Çerezlerin amacı; sizin o siteyi yeniden ziyaret ettiğinizi web sunucuya bildirmekten ibarettir.

Vekil sunucuların bulunduğu sistemlerde site ziyaretçileri için kullanıcı IP adresi bütün kullanıcılar için aynı gözükecektir. Kullanıcı erişim kayıtlarından aynı IP adresine sahip kullanıcıları birbirinden ayırt etmek için farklı metotlar bulunmaktadır. En yaygın kullanılan metotlardan biri de kullanıcı etmen (agent) bilgisi alanıdır. Kullanıcı erişim kayıtlarındaki IP adresi alan bilgisi ile birlikte kullanıcı etmen bilgisi alanı birlikte sorgulanmıştır. Sonraki adımda, IP adresi ile kullanıcı etmen bilgileri birebir aynı olan kullanıcılar için aynı kişidir söylenebilir. Tablo 3.7 ‘de IP adresi ve kullanıcı etmen alan bilgileri kullanılarak kullanıcı belirlemenin bir örneği gösterilmiştir.

Tablo 3.7. IP adresi ve kullanıcı etmen alanlarını kullanarak kullanıcı belirleme

Zaman	IP	URL	Ref	Agent
00:01	1.2.3.4	A	-	IE5;Win2k
00:09	1.2.3.4	B	A	IE5;Win2k
00:10	2.3.4.5	C	-	IE6;WinXP;SP1
00:12	2.3.4.5	B	C	IE6;WinXP;SP1
00:15	2.3.4.5	E	C	IE6;WinXP;SP1
00:19	1.2.3.4	C	A	IE5;Win2k
00:22	2.3.4.5	D	B	IE6;WinXP;SP1
00:22	1.2.3.4	A	-	IE6;WinXP;SP2
00:25	1.2.3.4	E	C	IE5;Win2k
00:25	1.2.3.4	C	A	IE6;WinXP;SP2
00:33	1.2.3.4	B	C	IE6;WinXP;SP2
00:58	1.2.3.4	D	B	IE6;WinXP;SP2
01:10	1.2.3.4	E	D	IE6;WinXP;SP2
01:15	1.2.3.4	A	-	IE5;Win2k
01:16	1.2.3.4	C	A	IE5;Win2k
01:17	1.2.3.4	F	C	IE6;WinXP;SP2
01:26	1.2.3.4	F	C	IE5;Win2k
01:30	1.2.3.4	B	A	IE5;Win2k
01:36	1.2.3.4	D	B	IE5;Win2k

Kullanıcı -1	00:01	1.2.3.4	A	-
	00:09	1.2.3.4	B	A
	00:19	1.2.3.4	C	A
	00:25	1.2.3.4	E	C
	01:15	1.2.3.4	A	-
	01:16	1.2.3.4	C	A
	01:26	1.2.3.4	F	C
	01:30	1.2.3.4	B	A
	01:36	1.2.3.4	D	B
Kullanıcı -2	00:10	2.3.4.5	C	-
	00:12	2.3.4.5	B	C
	00:15	2.3.4.5	E	C
	00:22	2.3.4.5	D	B
Kullanıcı -3	00:22	1.2.3.4	A	-
	00:25	1.2.3.4	C	A
	00:33	1.2.3.4	B	C
	00:58	1.2.3.4	D	B
	01:10	1.2.3.4	E	D
	01:17	1.2.3.4	F	C

Tablo 3.7'deki örneğe dikkat edilirse, kısmen temizlenmiş olan kullanıcı erişim kayıt dosyasının bir parçası gösterilmektedir. Kullanıcı tanımlama örneğindeki önişlem aşamasından geçirilen kayıtlar sol tarafta bulunan tablodaki görünümü almıştır. Erişim kayıtlarındaki verilerde IP adresi ve kullanıcı etmeni bilgilerin bulunduğu alanların birleşimi ile Tablo 3.7'nin sağ tarafında tanımlandığı gibi üç ayrı kullanıcı için erişim kayıtları sınıflandırılmıştır.

3.5 Oturum Tanımlama

Internet kullanıcısının bir web sitesine ya da web sunucusuna bağlandığı zamandan ayrılıncaya kadarki geçen süre içerisinde site üzerinde gerçekleştirdiği aktivitelerin bir kümesi, bir kullanıcı oturumu olarak tanımlanabilir. Oturum tanımlama işlemi, web oturumları içindeki her bir kullanıcının davranış aktivite kayıtlarının kümelenmesidir. Kullanıcı doğrulama bilgisi olmayan web siteleri için oturum tanımlama işlemleri sezgisel yöntemleri bağlıdır. Oturum tanımlamanın amacı, bireysel erişim oturumları içerisindeki her bir kullanıcının sayfa erişimlerinin gruplandırılmasıdır. Bu amaç doğrultusunda, kullanıcıların siteye yaptığı ziyaret esnasında oluşturduğu fare klikleme verilerinin düzenlenmekte ve kullanıcıların gerçekleştirdiği web olaylarının gerçek sırasına konulmakta ve yeniden tertip edilmektedir.

Oturum tanımlarken kullanıcı erişim kütük dosyaları içerisindeki alan bilgilerine göre hedeflenen amaca uygun veri alanları seçilmelidir. Örneğin, Cooley ve diğ. [9] yaptıkları

oturum tanımlama işleminde erişim verileri içerisinde *kullanıcı IP adresi, kullanıcı numarası, erişilen URL adresi, erişilen zaman* olmak üzere dört alan bilgisi seçmişlerdir. Oturum tanımlama işlemi için *oturum süresi temelli (session-duration-based), sayfada kalma süresi temelli (page-stay-time-based method)* ve *referans temelli (referrer-basic heuristic method)* olmak üzere çeşitli sezgisel yöntemler kullanılabilir. Bu metotlar aşağıda açıklanmaktadır.

***h*-1**: Oturum süresi temelli sezgisel metotta toplam oturum süresi θ eşik değerini aşmamalıdır. Yapılan birçok çalışmada, standart olarak eşik değeri genelde minimum θ=*30sn* olarak seçilmektedir. Oturumun başlangıç süresi t_0, bitiş süresi t olsun. Oluşturulan oturum S olursa, $t - t_0 \leq \theta$ şartı sağlandıkça bu süre içerisindeki aktiviteler aynı oturum içerisinde kabul edilir [88]. Zaman temelli sezgisel yöntem ile oturum oluşturma örneği Tablo 3.8'de gösterilmektedir. Oturum zamanı eşik değeri θ=*30sn* olarak alınmıştır.

Tablo 3.8. Zaman temelli sezgisel yöntem ile oturum oluşturma

	Zaman	IP	URL	Ref.
Kullanıcı-1	00:01	1.2.3.4	A	-
	00:09	1.2.3.4	B	A
	00:19	1.2.3.4	C	A
	00:25	1.2.3.4	E	C
	01:15	1.2.3.4	A	-
	01:26	1.2.3.4	F	C
	01:30	1.2.3.4	B	A
	01:36	1.2.3.4	D	B

Oturum-1	00:01	1.2.3.4	A	-
	00:09	1.2.3.4	B	A
	00:19	1.2.3.4	C	A
	00:25	1.2.3.4	E	C
Oturum-2	01:15	1.2.3.4	A	-
	01:26	1.2.3.4	F	C
	01:30	1.2.3.4	B	A
	01:36	1.2.3.4	D	B

	Zaman	IP	URL	Ref.
Kullanıcı-1	00:01	1.2.3.4	A	-
	00:09	1.2.3.4	B	A
	00:19	1.2.3.4	C	A
	00:25	1.2.3.4	E	C
	01:15	1.2.3.4	A	-
	01:26	1.2.3.4	F	C
	01:30	1.2.3.4	B	A
	01:36	1.2.3.4	D	B

Oturum-1	00:01	1.2.3.4	A	-
	00:09	1.2.3.4	B	A
	00:19	1.2.3.4	C	A
	00:25	1.2.3.4	E	C
	01:26	1.2.3.4	F	C
Oturum-2	01:15	1.2.3.4	A	-
	01:30	1.2.3.4	B	A
	01:36	1.2.3.4	D	B

***h*-2**: Sayfada kalma süresi temelli sezgisel metotta sayfada harcanan toplam süre θ eşik değerini aşmamalıdır. Standart olarak eşik değeri genelde minimum θ=*10sn* olarak seçilmektedir. Oturumun başlangıç süresi t_1, bitiş süresi t_2 olsun. Oluşturulan oturum S olursa, $t_2 - t_1 \leq \theta$ şartı sağlandıkça bu süre içerisindeki aktiviteler aynı oturum içerisinde kabul edilir [88]. Yani, kullanıcının bağlandığı sunucu üzerindeki davranışları 10'ar saniyeler şeklinde bölütlenerek, tanımlanan oturumlara göre analiz edilir.

***h*-ref**: Referans temelli sezgisel metotta ise bir S oturumunu oluşturmak için kullanıcı erişim kayıtlarındaki referans ve kullanıcı istek alan bilgileri kullanılmalıdır. Erişim kayıtlarındaki kullanıcı URL isteği q olarak alınırsa, q'dan önceki URL adresi de referans alan bilgisidir. Oturum oluşturma işleminde erişim kayıtları içerisindeki referans alan bilgisi ile istek yapılan adres alan bilgisi arasındaki ilişki dikkate alınır. Genelde ilk istek yapılan sayfanın referans bilgisi boş (-) olarak gösterilmektedir. Bu nedenle oturum oluşturmada başlangıç olarak istek alan bilgisi alınmaktadır [88].

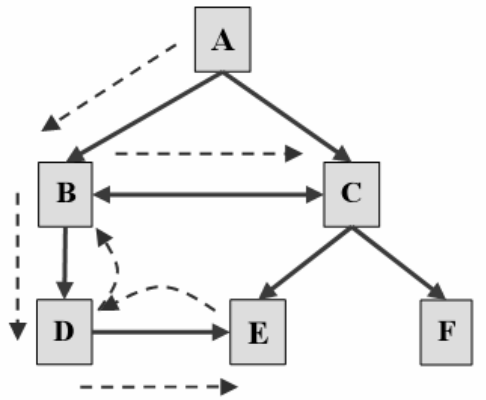
Aynı kullanıcıya ait aktivite kayıtlarına uygulanan (*h*-ref) referans temelli sezgisel yöntem ile oturum oluşturma örneği Tablo 3.9'da gösterilmektedir. Eşik değeri minimum θ=*30sn* olarak alındığından iki farklı oturum oluşturulmuştur.

Tablo 3.9. *h*-ref sezgisel yöntem ile oturum oluşturma

	Zaman	IP	URL	Ref						
Kullanıcı–1	00:01	1.2.3.4	A	-	Oturum–1	00:01	1.2.3.4	A	-
	00:09	1.2.3.4	B	A		00:09	1.2.3.4	B	A
	00:19	1.2.3.4	C	A		00:19	1.2.3.4	C	A
	00:25	1.2.3.4	E	C		00:25	1.2.3.4	E	C
	01:15	1.2.3.4	A	-		01:26	1.2.3.4	F	C
	01:26	1.2.3.4	F	C	Oturum–2	01:15	1.2.3.4	A	-
	01:30	1.2.3.4	B	A		01:30	1.2.3.4	B	A
	01:36	1.2.3.4	D	B		01:36	1.2.3.4	D	B

3.6 Yol Tamamlama

Yol tamamlama işlemi, oturumların belirlenmesinden sonra uygulanan önemli bir aşamadır. Web tarayıcının ön belleği veya kullanıcının kullandığı vekil sunucudan dolayı kayıt dışı kalan bağlantıları oluşturmak amacıyla yol tamamlama işlemi uygulanır. Örneğin, web sitesinde gezinen bir kullanıcı en son girdiği web sayfasıyla doğrudan bağlantı kurmamışsa bu sayfadan ayrıldığı zaman kullanıcı erişim kayıtlarında bir eksiklik oluşacaktır. İşte bu problemin çözümüne yol tamamlama yöntemi ile çözüm aranmaktadır. Günlük erişim kayıtlarına eklenmeyen satırlarının eklenebilmesi ve bu satırların tanımlanabilmesi için çözüm yol tamamlama metodunun kullanılmasıdır. Bir sayfanın hangi istek sayfasından geldiğini görmek için kullanıcı referans kayıtlarını kullanmak gerekir. Eğer bir kullanıcının geçmişindeki sayfa bilinmek isteniyorsa, tarayıcı sayfasından geri butonu ile geçmişe doğru gidilerek ön bellekte tutulan sayfalar görüntülenebilir. Eğer kullanıcı geçmiş kayıtlarında birden fazla istek sayfası bulunuyorsa, yol tamamlamada bu bağlantılardan en yakın olanı başlangıç olarak kabul edilir.



(a)

URL Adresi	Referans
A	-
B	A
D	B
E	D
C	B

(b)

Şekil 3.3. Yol tamamlama

Eksik referansların gösterildiği basit bir örnek Şekil 3.3'de verilmektedir. Şekil 3.3 (a)'da görüldüğü üzere A → B → D → E yollarını izleyen bir kullanıcının durum davranışına bakıldığında E sayfasından sonra tarayıcı üzerindeki geri tuşu ile önceki sayfa olan D sayfasına geri dönmüştür. Bu sayfa üzerinde tekrar geri tuşu ile B sayfasına ulaşmıştır. Daha sonra B sayfası üzerinde farklı bir bağlantıyı kullanarak C sayfasına erişmiştir. E sayfasından B sayfasına geri dönüş erişim kayıtlarındaki referans alan bilgisinde tutulmamıştır. Şekil 3.3 (b)'deki web sunucu erişim kayıtları örneğinde görüldüğü üzere tarayıcı geri tuşu ile yapılan işlemler referans alanına kaydedilmemiştir. Bu nedenle referans alanı eksik ve belirsiz bilgilerle kayıtlara geçmiştir. Bu problemin çözülebilmesi için de muhtemel sayfalardan gidilebilecek muhtemel bağlantı yolları dikkate alınmalı ve geri referansların gerekli olanlarından en az biri seçilmelidir [88].

3.7 Veri Bütünleştirme

Veri bütünleştirme aşamasında amaç, veri ön işleme aşamasındaki uygulamalar sonunda farklı alanlar ve farklı kaynaklardan elde edilen verileri bir araya getirerek bir bütünlük sağlamaktır. Örneğin, bir elektronik ticaret uygulamasında web verilerinin analizinde hem kullanıcıya ait verileri hem de ürüne ait verilerin elde edilmesi gerekmektedir. Kullanıcıyla ilgili olarak kullanıcı demografik özellikleri, oranlar, ürün alışveriş zamanlamaları gibi bilgilere ihtiyaç duyulurken ürün için ise, ürün özellikleri ve kategorileri gibi bilgiler madencilik işlemleri için gerekmektedir. Birbirleriyle ilişkili olan ve bütünlük halinde olan verilerin birleşimi sonucundaki oluşan veriler örüntü keşfi aşamasına sunulmaktadır. Akıllı iş değerleri elde etmek için bu bütünlüğü kurmak ve sağlamak zorunludur.

4. ÖRÜNTÜ KEŞFİ VE ANALİZİ

Örüntü, veri ya da veri yığınları içerisindeki herhangi bir biçimdeki yapıdır. Örüntü keşfi ise ön işlem evresinden geçirilmiş olan verilere veri madenciliği tekniklerinin uygulandığı aşamadır. Bir örüntü keşfi uygulamasında, keşfedilen örüntünün önemli olabilmesi şu temel esaslara bağlıdır;

- Çıkarılan örüntünün insanlar tarafından kolayca anlaşılabilir olmalıdır.
- Çıkarılan yeni örüntünün test edilmiş veri veya yeni veriler üzerinde belirli bir oranda geçerliliği sağlanmalıdır.
- Çıkarılan örüntülerin yararlı, ihtiyaçları karşılayan ve kullanılabilir olması gerekir.
- Çıkarılan örüntüler yeni olmalıdır.

Örüntü keşfinde "bütünlük" ve "en iyileme" kavramları sıkça kullanılmaktadır. *Bütünlük kavramı*, bütün örüntülerin bulunması demektir. *En iyileme* kavramı ise sadece en önemli örüntülerin tespit edilip çıkarılmasıdır. Bütün örüntüler bulunduktan sonra önemsiz olanlar ayıklanarak sadece önemli örüntüler tespit edilir [91].

Örüntü analizi ise, örüntü keşfi aşamasında ortaya çıkarılan kural veya örüntülerin analiz edilmesidir. Bilgi sorgulama ve OLAP (OnLine Analytical Processing-Çevrim İçi Analitik İşlem) uygulamaları ile veriler üzerinde derinlemesine analizler yapılabilmektedir [84].Bu analiz aşamasında elde edilen sonuçların (özetler gibi) anlaşılabilmesi için görselleştirme tekniklerinden faydalanılır. Görselleştirmede daireler, düğümler ve kenarlar kullanılmaktadır [25].

4.1 Örüntü Keşfi ve Analizinde Kullanılan Yöntemler

Veri madenciliği için örüntü keşfi ve analizinde kullanılan birçok yöntem bulunmaktadır. Literatürde, birçok kaynakta bu yöntemlerin detaylı ve kapsamlı olarak anlatımı mevcuttur [72, 88, 92-93]. Bu kısımda, özellikle bu tezin çalışma konusunu temel alan veri madenciliği algoritmaları anlatılmaktadır. Ön işlem aşamasından geçirilmiş Web kullanım örüntülerinden ilginç ve bilinmeyen örüntüleri keşfetmek için kullanılan yöntemler esas alınarak analiz edilmektedir. Web kullanım örüntülerinin keşfi ve analizinde kullanılan yöntemler; istatistiksel analiz, birliktelik kuralları, kümeleme, sınıflandırma ve sıralı örüntüler, OLAP ve bağımlı modellemedir [6, 72].

4.1.1 İstatistiksel Analiz

İstatistiksel yöntemler bir Web sitesinin ziyaretçileri hakkında bilgi açığa çıkarmaya yarayan en güçlü araçlardır. İstatistiksel analizin amacı, Web sitesi ve Web sayfaları içerisindeki birçok temel bilgileri elde etmektir. Kullanıcıların Web site içerisinde gezindiği sayfaların görüntülenmesi, site içerisindeki hatalı sayfaların ve kırık köprü bağlantılarının tespit edilmesi, kullanıcı sistemine ait yazılımların bulunması gibi birçok örnek verilebilir. Analizciler oturum dosyasını analiz ederken farklı değişkenler üzerinde farklı açıklamalı istatistiksel analiz tiplerini yerine getirirler. Periyodik Web sistem raporlarında bulunan istatistiksel bilgiler analiz edilerek sistem performansını arttırıcı, sistem güvenliğini genişletici, düzeltme işlemlerini kolaylaştırıcı ve pazarlama kararlarını destekleyici raporlar çıkartılabilir [6].

İstatistiksel yöntemler, Web kütük madenciliği sistemlerinde yaygın olarak kullanılan temel yaklaşımlardır. Bu yöntemler, daha çok ticari Web sitelerinin kullanıcı erişim kütüklerinden birçok istatistikî bilgilerin özet şeklinde elde edilebilmesi için Web madenciliği yazılımlarında kullanılmaktadır. Eğer Unix tabanlı sistemler kullanılıyorsa kabuk programlaması yolu ile çok hızlı şekilde Web kullanım dosyalarından aşağıdaki soruların cevabını bulan istatistik bilgilere erişilebilir.

- Hangi kullanıcılar dışarıdan Web sitemizi ve sayfalarımızı kullanmaktadır?
- Hangi Web tarayıcıları ve işletim sistemleri Web sayfalarımıza ulaşmaktadır?
- Hangi ziyaretçiler Web sayfalarında gezinmektedirler?
- Site içerisinde en çok ziyaret edilen Web sayfaları hangileridir?

Bu şekilde birçok istatistikî bilgiler elde edilebilmektedir. Web kullanıcı erişim kütük verileri kullanılarak istatistiksel analiz yapmak için internette birçok serbest yazılım bulunabilir. Bu yazılımlar ile ilgili ayrıntılı bilgiler 6. bölümde sunulmaktadır.

Web madenciliği yazılımları kullanılarak elde edilen istatistikî bilgiler ile Web sitesinin daha iyi organize edilebilmesi, Web trafiklerinin anlaşılması ve Web site performansının iyileştirilmesi için site tasarımcılarına ve yöneticilerine yardım edilmektedir. Bölüm 6'de istatistiksel yöntemler içeren yazılımlar kullanılarak, Fırat Üniversitesi Web sunucularına ait kullanıcı erişim kütükleri üzerinde yapılmış uygulama detaylı olarak verilmeltedir.

Koutsoupias [94] Web kütük istatistiklerinde kullanıcı davranışları ve tercihlerini incelemek içi *uygunluk analizini* kullanmaktadır. Megaputer Web Analyticst [29], WebTrends [30, 33], Nihuo [31], WMS ve WLE [36], Awstat [39] ve Analog [40] gibi birçok yazılım istatistiksel bilgilerin çıkarılmasını sağlayan yazılım paketlerin örnekleri olarak verilebilir.

4.1.2 Birliktelik Kuralı

Birliktelik kuralı veri tabanındaki geçmiş tarihli hareketleri analiz etmek için karar verme aşamasında örüntüleri ve ilişkileri bulmada, verilen kararların kalitesini arttırmada izlenen bir yaklaşımdır. Bu yöntemdeki amaç bir küme içerisindeki nesnelerin birbirleri ile olan bağlarının tespit edilmesidir. Bu veri madenciliği yöntemi birçok kaynakta alışveriş sistemlerinde kullanıldığını göstermesine rağmen başka uygulamalarda da kullanılabilmektedir.

Toplanan ve depolanan verinin her geçen gün artması, şirketler tarafından kendi veritabanlarındaki öğelerin birliktelik kurallarını ortaya çıkarmaya itmektedir. Birliktelik kurallarının çıkarımı katalog tasarımı, müşterilerin satın alma alışkanlıklarına göre sınıflandırılması, mağaza ürün yerleşim planı gibi pek çok uygulama alanında kullanılabilir. Birliktelik kuralına örnek verilecek olursa, A ürününün alınması ile B ürününün veya C ürünün alınması arasındaki işlemlerde bir bağlantı olup olmadığının tespit edilmesi ve eğer bağlantı var ise bu bağlantılar arasındaki kuvvet veya önem derecesinin ortaya çıkarılması sağlandığı görülmektedir. Bu analizin amacı A ürününü alan kişilerin B veya C ürünleri alımları ile ilgili olarak kuvvetli bir bağın bulunup bulunmadığını kontrol etmek eğer var ise bununla ilgili olarak örneğin müşterilere promosyonlar veya ürünlerin raflarının daha yakın yerlere yerleştirilmesini sağlamak olabilir. Bu işlem bir Web sitesi içerisinde sayfaların yapılandırılması amacı ile de kullanabilir.

Ticari bir Web sitesinde alışveriş esnasında veya birbirini izleyen alışverişlerde müşterinin hangi mal veya hizmetleri satın almaya eğilimli olduğunun belirlenmesi, müşteriye daha fazla ürünün satılmasını sağlayan yollardan biridir. Satın alma eğilimlerinin tanımlanmasını sağlayan birliktelik kuralları ve ardışık zamanlı örüntüler, pazarlama amaçlı olarak da pazar sepeti analizi yöntemleri veri madenciliğinde yaygın olarak kullanılmaktadır. Bununla birlikte bu teknikler, tıp, finans ve farklı olayların birbirleri ile ilişkili olduğunun belirlenmesi sonucunda değerli bilgi kazanımının söz konusu olduğu ortamlarda da önem taşımaktadır.

Birliktelik kuralları eş zamanlı olarak gerçekleşen ilişkilerin tanımlanmasında kullanılır. Keşfedilen örüntüler uygulamada sıklıkla birlikte geçen nitelik değerleri arasındaki ilişkiyi gösterir. Şampuan ve saç kremi satın alan müşterilerin %20 ihtimalle saç jölesi de almaları, kola satın alan müşterilerin, %75 ihtimalle patates cipsi de almaları ya da düşük yağlı peynir ve yağsız yoğurt satın alan müşterilerin, %85 ihtimalle diyet süt de almaları birliktelik kurallarına örnek olarak verilebilir. Bu tür birliktelik örüntüleri ancak, örüntüde yer alan öğelerin işlemleri birden fazla tekrarlandığında potansiyel olarak bu kuralın geçerliliği sağlanabilir.

Birliktelik kurallarının kullanıldığı en tipik örnek market sepeti uygulamasıdır. Bu işlem, müşterilerin yaptıkları alışverişlerdeki ürünler arasındaki birliktelikleri bularak

müşterilerin satın alma alışkanlıklarını analiz eder. Bu tip birlikteliklerin keşfedilmesi, müşterilerin hangi ürünleri bir arada aldıkları bilgisini ortaya çıkarır ve market yöneticileri de bu bilgi ışığında daha etkili satış stratejileri geliştirebilirler. Örneğin bir marketin müşterilerinin süt ile birlikte ekmek satın alan oranı yüksekse, market yöneticileri süt ile ekmek raflarını yan yana koyarak ekmek satışlarını arttırabilirler.

Sepet analizinde mallar arasındaki bağıntı, destek ve güven değerleri aracılığıyla hesaplanır. *Destek* veri içerisinde bu bağıntının ne kadar sık olduğunu, *güven* de *X* ürününü almış bir kişinin hangi olasılıkla *Y* ürününü alacağını ifade eder. Bağıntının önemli olabilmesi için her iki değerin de olabildiğince büyük olması gerekir.

X ve *Y* farklı ürünler olmak üzere, *X* ürünü için destek tüm alışverişler içinde X ürünün oranıdır. $|\mathrm{X}|$, *X* ürünü içeren alışverişlerin sayısını, $|\mathrm{D}|$ yapılan tüm alışverişlerin sayısını göstermek üzere;

$$Destek\ (X) = \frac{|X|}{|D|}$$ olarak ifade edilir.

X ve *Y* ürünleri için destek, $|X.Y|$ *X* ve *Y* ürünlerini birlikte içeren alışveriş sayısı olmak üzere;

$$\text{Destek}\ (X \Rightarrow Y) = \frac{|X.Y|}{|D|}$$ olarak ifade edilir.

X ve *Y* ürünleri için güven ise;

$$Güven\ (X \Rightarrow Y) = \frac{destek(X.Y)}{destek(X)}$$ olarak ifade edilir.

Örneğin bir X ürününü satın alan müşteriler aynı zamanda Y ürününü da satın alıyorlarsa, bu durumun birliktelik kuralı ile gösterimi;

X => Y [destek = %2, güven = %60]

Buradaki destek ve güven ifadeleri, kuralın ilginçlik ölçüleridir. Sırasıyla, keşfedilen kuralın kullanışlığını ve doğruluğunu gösterirler. Yukarıdaki bağıntı için %2 oranındaki bir destek değeri, analiz edilen tüm alışverişlerden %2'sinde X ile Y ürünlerinin birlikte satıldığını belirtir. %60 oranındaki güven değeri ise X ürününü satın alan müşterilerinin %60'ının aynı alışverişte Y ürününü de satın aldığını ortaya koyar. Kullanıcı tarafından minimum destek eşik

değeri ve minimum güven eşik değeri belirlenir ve bu değerleri aşan birliktelik kuralları dikkate alınır.

Büyük veri tabanlarında birliktelik kuralları bulunurken, şu iki işlem basamağı takip edilir [95–96]:

1- Sık tekrarlanan öğeler bulunur: Bu öğelerin her biri en az, önceden belirlenen minimum destek sayısı kadar sık tekrarlanırlar.
2- Sık tekrarlanan öğelerden güçlü birliktelik kuralları oluşturulur: Bu kurallar minimum destek ve minimum güven değerlerini karşılamalıdır.

Web madenciliğinde birliktelik kurallarının çıkarılma uygulaması bölüm 7.2.1'de detaylı olarak anlatılmaktadır.

Literatürde, Web kullanım madenciliği için birliktelik kurallarının çıkarılması için birçok çalışmalar yapılmıştır. Shen ve diğ. [97] yaptıkları çalışmalarında, çok ilginç Web erişim birlikteliklerini modellemek için üç adım yaklaşımı önermektedirler. Tan ve Kumar [98] daha karmaşık yapıda çeşitli birliktelik örüntülerinin birleşimini sağlayan doğrudan olmayan birliktelik olarak adlandırılan bir madencilik yöntemi geliştirmişlerdir. Bu yeni teknik aynı zamanda var olan aynı veritabanında hem negatif hem de pozitif ilişkileri tespit etmektedir. Mobasher ve diğ. [99] etkili Web kişiselleştirmeyi gerçekleştirmek için birliktelik kuralları madenciliğini önermektedirler. Yao ve diğ. [100] akıllı Web olaylarını inşa etmek için birliktelik kuralları madenciliğini kullanmışlardır.

4.1.3 Sınıflandırma

Veri madenciliğinde sınıflandırma işlemi önemli bir problemdir. Bilginin alınması problemine çözüm üreteme kadar makine öğrenmesi topluluğu yoğun bir şekilde çalışmalar yapmaktadırlar [72]. Sınıflandırma, bir veriyi daha önceden tanımlanmış sınıflara dağıtma tekniğidir. Örüntü keşfi uygulamalarında en çok kullanılan yöntemlerden biridir. Sınıflandırma, daha önceden belirlenmiş ölçütlere göre, örneğin yaşa, cinsiyete, gelir durumuna, eğitim düzeyine ve müşterinin kredi borcunu zamanında ödeyip ödememesine, bir kampanyaya olumlu cevap verip vermemesine, hedeflenen değerlerin üzerinde bulunup bulunmamasına yani ilgilenilen herhangi bir özelliğe veya birkaç ölçüte göre yapılır. Web etki alanında, sınıflandırma tekniği kullanarak müşterilerinin hangi sınıf veya kategoride bir profile sahip olduğu belirlenebilir. Bu yöntemin en önemli fonksiyonu, sınıflandırma sonrasında her kategoride yer alan kayıtların, alanların, kişilerin, nesnelerin, kurumların özelliklerini ortaya çıkarmaktır. Örneğin; internet bankacılığında yaptıkları elektronik fon transferi sıklıklarına göre sınıflandırmada internet müşterileri, "seyrek" kullanıcı, "orta sıklıkta" kullanıcı ve "sık"

kullanıcı olarak sınıflandırılabilir. Müşteriler bu şekilde gruplandıktan sonra amaç, her bir grubun özelliklerini analiz etmek, profilini ortaya çıkarmak ve bu grupların özelliklerini, tutum ve davranışlarını içeren bir davranış geliştirebilmektir. Sınıflandırma işleminde, verilen bir sınıf veya kategorinin özelliklerini en iyi biçimde açıklamak için seçim ve açığa çıkarma uygulamalarına ihtiyaç duyulur. Sınıflandırma işlemi; karar ağaçları, Bayes sınıflayıcıları, en yakın komşu ve destek vektör makineleri gibi denetlenen tümevarımsal öğrenim algoritmaları kullanılarak yapılabilir [6].

Sınıflama algoritması, yeni bir veri elemanını daha önceden belirlenmiş sınıflara atamayı amaçlamaktadır. Veri tabanında yer alan çoklular bir sınıflama fonksiyonu yardımıyla kullanıcı tarafından belirlenir veya karar niteliğinin bazı değerlerine göre anlamlı ayrık alt sınıflara ayırır. Bu nedenle sınıflama, eğiticili öğrenmeye girmektedir [95]. Sınıflama algoritması bir sınıfı diğerinden ayıran örüntüleri keşfeder. Sınıflama algoritmaları iki şekilde kullanılır:

1. *Karar Değişkeni ile Sınıflandırma:* Seçilen bir niteliğin aldığı değerlere göre sınıflandırma işlemi yapılır. Seçilen nitelik karar değişkeni adını alır ve veri tabanındaki çoklular karar değişkeninin değerlerine göre sınıflara ayrılır. Bir sınıfta yer alan çoklular karar değişkeninin değeri açısından özdeştir.
2. *Örnek ile Sınıflandırma:* Bu biçimdeki sınıflandırmada veri tabanındaki çoklular iki kümeye ayrılır. Kümelerden biri pozitif, diğeri negatif çokluları içerir. Yaygın kullanım alanları, banka kredisi onaylama işlemi, kredi kartı sahteciliği tespiti ve sigorta risk analizidir.

4.1.4 Kümeleme

Kümeleme, veri madenciliğinde öncelikli verilerde ilginç veri dağılımlarını ve örüntüleri keşfetmek için kullanılan yararlı bir yöntemdir. Kısacası, veriyi sınıflara veya kümelere ayırma işlemidir [101]. Bu yöntem, doküman içinde geçen terimlere bakarak aynı konudaki dokümanları gruplamaktadır. Her doküman içinde sık geçen terimleri bulur. Bu terimlerden ve ağırlıklarından yararlanarak bir benzerlik ölçütü geliştirir ve bu ölçüte göre kümeleme yapar. Aynı kümedeki elemanlar birbirleriyle benzerlik gösterirlerken, başka kümelerin elemanlarından farklıdırlar. Kümeleme veri madenciliği, istatistik, biyoloji ve makine öğrenimi 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 geldiğinde bu verinin hangi sınıftan olabileceği tahmin edilmektedir. Oysa kümeleme modelinde, sınıfları bulunmayan veriler gruplar halinde kümelere

ayrılırlar. Bazı uygulamalarda kümeleme modeli, sınıflama modelinin bir önişlemi gibi görev alabilmektedir [95, 102].

Web kullanım madenciliğinde genelde iki tarz kümeleme yaklaşımı vardır.

1. *Kullanıcı Kümeleri:* Buradaki amaç benzer sayfa görüntülemesi yapan kullanıcıları tespit edip onları bir grup içerisine almaktır. Özellikle Web kişiselleştirilmesi işleminde çok yararlı olmaktadır. Örneğin bir portal içerisinde oyun ve spor sayfasına girenleri bir grup içerisine alıp bir sonraki bağlantıyı yaptıklarında bu konuda reklâmların sayfalarda gelmesini sağlamak gibi.
2. *Sayfa Kümeleri:* Benzer içerikli sayfaların bir araya gruplandırılması özellikle arama motorları için çok yararlı olabilecektir. Böylelikle bir kullanıcının aramış olduğu bilgilere daha hızlı şekilde ulaşılabilmesi sağlanabilecektir.

Marketlerde farklı müşteri gruplarının keşfedilmesi ve bu grupların alışveriş örüntülerinin ortaya konması, biyolojide bitki ve hayvan sınıflandırmaları ve işlevlerine göre benzer genlerin sınıflandırılması, şehir planlanmasında evlerin tiplerine, değerlerine ve coğrafik konumlarına göre gruplara ayrılması gibi uygulamalar tipik kümeleme uygulamalarıdır. Kümeleme aynı zamanda Web üzerinde bilgi keşfi için dokümanların sınıflanması amacıyla da kullanılabilir [95, 102]. Veri kümeleme güçlü bir gelişme göstermektedir. Veri tabanlarında toplanan veri miktarının artmasıyla orantılı olarak, kümeleme analizi son zamanlarda veri madenciliği araştırmalarında aktif bir konu haline gelmiştir.

Literatürde bilinen pek çok kümeleme algoritması bulunmaktadır. Kullanılacak olan kümeleme algoritmasının seçimi, veri tipine ve amaca bağlıdır. Genel olarak başlıca kümeleme yöntemleri şu şekilde sınıflandırılabilir [95–96]:

1- Bölme yöntemleri
2- Sıradüzensel yöntemler
3- Yoğunluk tabanlı yöntemler
4- Izgara tabanlı yöntemler
5- Model tabanlı yöntemler

Bölme yöntemlerinde, n veri tabanındaki nesne sayısı ve k oluşturulacak küme sayısı olarak kabul edilir. Bölme algoritması n adet nesneyi, k adet kümeye böler. Kümeler tarafsız bölme ölçütü olarak nitelendirilen bir kritere uygun oluşturulduğu için aynı kümedeki nesneler birbirlerine benzerken, farklı kümedeki nesnelerden farklıdırlar [95–96].

En iyi bilinen ve en çok kullanılan bölme yöntemleri k-means yöntemi, k-medoids yöntemi ve bunların varyasyonlarıdır [14]. 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 dağılırlar. Yani bir nesne hangi kümenin merkezine daha yakın ise o kümeye yerleşir. Ardından her küme için ortalama hesaplanır ve hesaplanan bu değer o kümenin yeni merkezi olur. Bu işlem tüm nesneler kümelere yerleşinceye kadar devam eder [95–96].

K-means yöntemi, sadece kümenin ortalaması tanımlanabildiği durumlarda kullanılabilir. Kullanıcıların k değerini, yani oluşacak küme sayısını belirtme gerekliliği bir dezavantaj olarak görülebilir. Esas önemli olan dezavantaj ise dışarıda kalanlar olarak adlandırılan nesnelere karsı olan duyarlılıktır. Değeri çok büyük olan bir nesne, dâhil olacağı kümenin ortalamasını ve merkez noktasını büyük bir derecede değiştirebilir. Bu değişiklik kümenin hassasiyetini bozabilir. Bu sorunu gidermek için kümedeki nesnelerin ortalamasını almak yerine, kümede ortaya en yakın noktada konumlanmış olan nesne anlamındaki *medoid* kullanılabilir. Bu işlem *k-medoids* yöntemi ile gerçekleştirilir.

K-medoids kümeleme yönteminin temel stratejisi ilk olarak *n* adet nesnede, merkezi temsili bir *medoid* olan *k* adet küme bulmaktır. Geriye kalan nesneler, kendilerine en yakın olan *medoide* göre *k* adet kümeye yerleşirler. Bu bölünmelerin ardından kümenin ortasına en yakın olan nesneyi bulmak için *medoid*, *medoid* olmayan her nesne ile yer değiştirir. Bu işlem en verimli *medoid* bulunana kadar devam eder [95].

Toplayıcı sıradüzensel kümeleme yöntemi olan AGNES (AGlomerative NESting) ve bir bölücü sıradüzensel kümeleme yöntemi olan DIANA (DIvide ANAlysis) uygulaması gösterilmektedir. Bu yöntemler beş nesneli (a,b,c,d,e) bir veri kümesine uygulanmaktadır. Başlangıçta AGNES her nesneyi bir kümeye yerleştirir. Kümeler, bazı kıstaslara göre basamak-basamak birleşirler. Örneğin C1 ve C2 kümeleri, eğer C1 kümesindeki bir nesne ve C2 kümesindeki bir nesne ile diğer kümelerdeki herhangi iki nesne arasında belirlenen uzaklık mesafesini karşılayacak bir mesafe varsa birleşebilirler. Bu birleşme işlemi tüm nesneler bir kümede toplanıncaya kadar devam eder. DIANA'da ise tüm nesnelerin içinde toplandığı küme, her küme bir nesne içerecek duruma gelene kadar bölünür [95].

Kümeleme algoritması veri tabanını alt kümelere ayırmaktadır. Her bir kümede yer alan elemanlar dâhil oldukları grubu diğer gruplardan ayıran ortak özelliklere sahiptir [106–108]. Bu yüzden kümeleme, güdümsüz öğrenmeye girmektedir. Güdümsüz (etiketsiz) kümeleme, güdümlü (etiketli) sınıflama için ön işlem olarak çok sıkça kullanılmaktadır. Bilgi erişim disiplini kümeleme konusundaki çalışmalar açısından oldukça zengin bir geçmişe sahiptir ve bu çalışmalar gömü adı altında toplanabilir. Tipik bir bilgi geri erişim sistemi için gömü, terimlerin belli bir ilişkiye göre düzenlenmesidir. Gömü, dizinleme ve erişim hizmetlerinde terimlerin kullanımına rehberlik eder. Bu özelliği ile gömünün bir yetkili kütüğü olduğu söylenebilir. Gömü ile amaçlanan; kullanıcı sorgusunu, sorguda kullanmadığı ama bilgi ihtiyacı ile ilişkili

terimler ile genişletmektir. Sorgu genişletmede kullanılacak terimler gömü ile belirlenir. Böylece sorgular kullanıcının ifade seklinden kısmen bağımsızlaştırılır ve sorguya eklenen terimler ile daha fazla ilgili belgeye erişme imkânı ortaya çıkar. Bir gömünün performansı da dizinleme ve/veya erişim aşamasında kullanıldığı ve kullanılmadığı durumlarda anımsama ve duyarlılık parametrelerinin karşılaştırılması ile ölçülür. Bu alanda yapılan çalışmalar gömünün üretildiği derlemenin, benzer derlemelerde kullanılması şartıyla anımsama değerinde %20'lere yaklaşan artışlar elde edilebildiğini göstermiştir [109].

Literatürde benzer şekilde, hem sınıflandırma hem de kümeleme yöntemleri kullanılarak yapılmış birçok çalışma bulunmaktadır [72]. Fu ve diğ. [110] yaptıkları çalışmalarında, sıradüzensel sayfalar göre oturumları üretmek için özellikli-yönlü atama özelliğini kullanmışlardır. Üretilen oturumlara sıradüzensel kümeleme yöntemi uygulamışlardır. Nasraouni ve diğ. [111] çalışmalarında, kullanıcıların tipik oturum profillerinin madenciliği için ilişkisel rekabete dayalı bulanık kümeleme yöntemi kullanmaktadırlar. Cadez ve diğ. [112] çalışmalarında, bir Web sitesindeki durumsal örüntüleri görselleştirmek için model tabanlı kümeleme kullanmışlardır.

4.1.5 Sıralı Örüntüler

Sıralı (ardışık) örüntülerin keşfi, belirli bir zaman içerisinde olaylar ya da oturumlar kümesindeki bir öğeden diğer bir öğeyi takip eden örüntüleri bulmaya çalışmaktadır. 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, bir Web sitesinde yapılan ilanlar ya da ürün satışları belirli kullanıcı gruplarına yönlendirilebilecektir [6].

Sıralı örüntüler, birbiri ile ilişkisi olan ancak birbirini izleyen dönemlerde gerçekleşen ilişkilerin tanımlanmasında kullanılmaktadır. X ameliyatı yapıldığında, 15 gün içinde %45 ihtimalle Y enfeksiyonun oluşması, IMKB borsa endeksi düşerken A hisse senedinin değeri %15'den daha fazla artacak olursa, üç iş günü içerisinde B hisse senedinin değeri %60 ihtimalle artacak olması, çekiç satın alan bir müşterinin ilk üç ay içerisinde %15, bu dönemi izleyen diğer üç ay içerisinde de %10 ihtimalle çivi satın alacak olması ardışık zamanlı örüntülere örnek olarak verilebilir.

Chen ve diğ. [113] yaptıkları çalışmalarında, bir dağıtık bilgi sistemlerinde yol aykırı örüntülerde madencilik fikrini ortaya koymuşlardır. Bu çalışmalarında temel olarak iki yaklaşım sunmuşlardır. Birincisi, kütük verilerinin orijinal sırası, maksimum ileriye dönük referansların bir kümesine dönüştürülmektedir. İkincisi, sıklık aykırı örüntüleri bulmak için iki algoritma

kullanmaktadırlar. Nanopoulos ve Manolopoulos [114] çalışmalarında, aykırı örüntülerin tanımlanması için Web sitesinin öncelikli yapısında ve verilen akıllı seviye algoritmalarda örüntünün genel bir tipini önermişlerdir. Buncher ve diğ. [115] çalışmalarında, Web'in önemli özelliklerinin geniş bir oranı ile keşif için MiDAS isminde yeni bir algoritma geliştirmişlerdir.

4.1.6 Bağımlı Modelleme

Bağımlı modellemenin amacı, Web alanlarında çeşitli Web değişkenleri arasındaki önemli bağımlılıkları ortaya çıkaran modeller oluşturmaktır. Bağımlı modellemede kullanılabilen Web kullanıcı davranışlarını modellemek için birkaç olasılık öğrenme teknikleri vardır. Bu teknikler, Morkov modelleri ve Bayesian ağları teoremlerini içermektedir [72]. Chen ve diğ. [116] çalışmalarında, dağıtık Web sunucu kütük dosyalarından Bayesian ağları öğrenmesi ile kullanıcıların Web deki gezinme örüntülerini modellemişlerdir. Borges ve Levene [117–118] kullanıcı durum oturumlarını modellemişlerdir. Deshpande ve Karypis [119] çalışmalarında, seçici bir Morkov modeli ile öğrenme işlemini yapmış, daha sonra Web sayfa erişimlerini tahmin etmişlerdir.

4.1.7 OLAP ve Veri Ambarlama

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

1- Kapasitesi çok büyük boyutta olan verileri analiz etme
2- Verilerdeki geçici boyutlar için açık destek sağlama
3- Farklı tipte bulunan bilgiler için destek sağlama
4- Uzun sıra analizi

Bu özellikler gerçekleştirilirken toplam eğilimler bireysel veri elamanlarından daha önemlidir. OLAP doğrudan ilişkisel veri tabanları üzerinde çalışabilir. OLAP kullanılırken analiz için veri küplerinden faydalanılmaktadır [25].

Veri ambarlama, OLAP araçları kullanarak plansız analiz uygulama ve veri depolamak için popüler bir depolama şeklidir [72]. OLAP araçları, farklı seviyeden alınan Web erişim kütük verilerini bölme, istenilen alan bilgisine göre kısımlara ayırma, esnek verileri yuvarlama, uzun verileri kesme, istenilen şekilde görüntüleme ve analiz etmek için kullanıcılara izin vermektedir.

4.1.8 Bilgi Sorgulama Mekanizması

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.

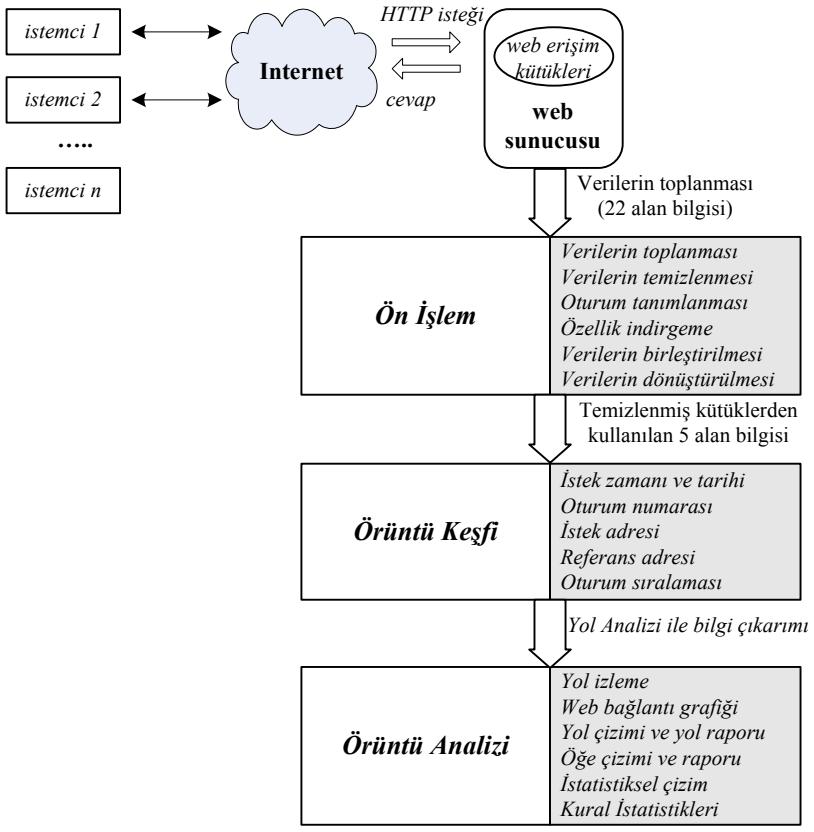
5. YOL ANALİZİ YÖNTEMİNİ KULLANARAK WEB KULLANICI ERİŞİM KÜTÜKLERİNDEN BİLGİ ÇIKARIMI

Bu bölümde, *yol analizi* (path analysis) yöntemini web kullanıcı erişim kütük verilerine uygulayarak, kullanıcı davranışlarına yönelik anlamlı ve ilginç bilgiler elde edilmiştir. Literatürde, web kullanıcı erişim kütük verilerinden, kullanıcı davranışlarına yönelik anlamlı ve ilginç bilgilerin keşfi için istatistiksel analiz [3], birliktelik kuralları [75, 96–98, 124–126], sınıflandırma ve kümelendirme [95–96, 102, 106–112], bağımlı modelleme [116–119], genetik algoritma [14, 16, 18, 127–128] gibi yöntemlerinin kullanıldığı görülmektedir. Yol analizi yönteminin, literatürde tarım ve zirai alanlarında yaygın olarak kullanıldığı görülmüş [120–121], ancak web kullanıcı erişim kütükleri ile ilgili yapılmış bir çalışmaya rastlanmamıştır. Yapılan bu çalışmalar da [120], bitki boyu, yaprak eni, yaprak boyu, yaprak alanı gibi dört farklı birbirleriyle ilişkili değişken verileri alınarak, yol analizi katsayıları hesaplanmış birbirleriyle etkileşimlerinin toplamının ilgileşim (korelasyon) katsayısına eşitliği gösterilmiştir. Sonuçta değişkenler arasındaki etkileşim şekillerinin tespiti üzerinde durulmuştur. Diğer bir çalışma da [121], arıların birbirleriyle ilişkili özellikleri kullanılarak, yol analizi ile arıların bal verimini arttırıcı özelliği tespit edilmiştir. Bu alanlardaki başarımların ümit verici olduğu görülmüş, web kullanım madenciliği alanında da kullanılarak başarımının değerlendirilmesi tezin bu bölümde gerçekleştirilmiştir.

Yol analizi kullanılarak web kullanım madenciliği alanında yapılan bu ilk çalışma sonucunda elde edilen bulgular ve sonuçlar, mevcut literatürde kullanılan ve bu tez çalışmasında da aynı veriler kullanılarak uygulanan birliktelik kuralları yöntemi ile karşılaştırıldı. Bu karşılaştırma sonucunda her iki yönteminde birbirlerine göre farklı anlamlı bilgiler tespit ettiği görüldü. Yol analizi yönteminde yol katsayıları hesaplanmasında ölçümler niceleyici değişkenlerden elde edilmiş olmalıdır. Aksi halde yol katsayısı hatalı hesaplanır ve bu durum tahmin kümesinin güvenilirliğini azaltır. Fakat birliktelik kuralları yönteminde güven ve destek değerleri hesaplanarak geçmişteki öğeleri kullanıp, gelecekle ilgili tahmin yapılmaktadır. Yani, geçmiş veri kümesinde eksiklik veya bozukluk tespit edilerek, gelecek veri kümesi için eksiklik tamamlanır. Bu sonuçlardan hareketle web kütük madenciliği alanında, ziyaretçilerin gezindiği yolları tespit etmek için yol analizi yöntemi, ziyaretçilerin sıklıkla birlikte ziyaret ettikleri web sayfalarının keşfi için de birliktelik kuralları yönteminin kullanılabilir olduğu tespit edildi.

Bu bölümde, yol analizi işleminin öncesinde, web kullanıcı erişim kayıtlarının temizlenmesi için bir yöntem geliştirilmiştir. Söz konusu yöntemin, literatürdeki çalışmalardan [9–10, 15–18] farklı yanı tüm işlemlerin tek bir uygulama çatısı altında bütünleştirilmiş olması ile daha hızlı ve daha doğru veri elde etmek için avantaj sağlamasıdır.

Web kullanıcı erişim kütüklerinden yol analizi ile bilgi keşfi için geliştirilen yöntem, JAVA tabanlı SAS Base yazılım (SAS Yazılımı Lisans Numarası: 291468, Sürümü: 9.1.3) ortamında gerçekleştirilmiştir [82]. Yaygın olarak kullanılan tüm işletim sistemlerine uygun SAS Base yazılımının sürümleri bulunmaktadır [82]. Güçlü bir donanım yapılandırması isteyen SAS yazılımında büyük verilerle çalışırken en az 2 GB bellek gerekmektedir. Birçok farklı biçimdeki (.html, .txt, .log, .mdb v.b.) dosyalar ile uyumlu çalışabilen SAS yazılımının kendisine özgü veri tabanıyla hızlı bağlantı kurabilen yapısı nedeniyle dünyada veri madenciliği alanında yaygın olarak kullanılmaktadır. Ayrıca, elde edilen verilerin analizinde birçok yöntemlerin kullanımına izin veren, geniş ve özellikli grafiksel yapısıyla ticari ve akademik yaşamda tercih edilmektedir. Kullanıcıya özgü programlanabilir özelliği, görsel ortamı, değiştirilebilir parametre uyarlamaları ve modüler bir yapı içermesinin yanı sıra daha hızlı ve başarımı yüksek ilave sonuçlar çıkarabilmesi SAS yazılımının, diğer bilgisayar yazılım ortamlarına (MATLAB, MATCAD, v.s.) göre üstünlüklerini ortaya koymaktadır [121–122].



Şekil 5.1. Yol analizi ile bilgi çıkarımı için geliştirilen yöntemin prensip şeması

Web kütük madenciliği için geliştirilen yöntemin prensip şeması Şekil 5.1'de gösterilmektedir. Yöntem; *ön işlem, örüntü keşfi* ve *örüntü analizi* olmak üzere üç ana

bileşenden oluşmaktadır. Bu bileşenler üzerinden geliştirilen yöntemin ayrıntıları aşağıdaki alt bölümlerde sunulmuştur.

5.1 Ön İşlem Süreci

Web kullanım madenciliğinin ön işlem aşamasındaki en önemli adımlardan biri veri madenciliği yöntemlerinde kullanılmak üzere uygun bir veri kümesi oluşturmaktır. Web kütük madenciliği için sunucu kütük dosyalarında saklanan veriler ön işlem sürecinde işlenerek gerekli dönüşümler yapılır. Bu süreçte, sunucudan alınan kullanıcı erişim kütük dosyaları *veri temizleme, oturum tanımlama, özellik indirgeme, veri birleştirme* ve *dönüştürme* aşamalarından geçirilerek bir sonraki evre olan *örüntü keşfi* aşamasına uygun hale getirilir.

Uygulamada kullanılan, kullanıcı erişim kütük dosyaları, Fırat Üniversitesi Bilgi İşlem Daire Başkanlığı bünyesinde hizmet veren web sunucularına aittir. Önerilen yöntem ve geliştirilen yazılım kodlamaları SAS Base yazılımında gerçekleştirildi, program kodları bölüm içerisinde ve tez sonunda Ek-1'de sunuldu.

Bu bölümde, önerilen yöntemdeki her bir aşama detaylı olarak açıklanıp, uygulamada karşılaşılan zorluklar ifade edilmektedir. Önerilen yöntem, Bölüm 3'de verilen ve Şekil 3.1.ile gösterilen web kullanım madenciliğinin genel mimari yapısı esas alınarak oluşturuldu.

5.1.1 Verilerin Toplanması

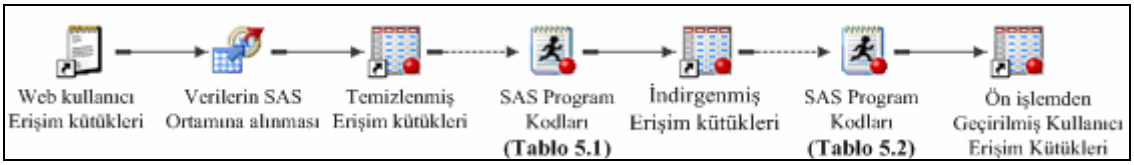
Önerilen yöntemin uygulaması için kullanılan kullanıcı erişim kütük dosyaları, Fırat Üniversitesi Bilgi İşlem Daire Başkanlığı bünyesinde hizmet veren web sunucularından alındı. Kullanıcı erişim kütüklerinin analizi sonucunda web sitesinin kullanımına ve ziyaretçilere ait detaylı bilgiler çıkarabilmek için ECLF biçiminde erişim kütükleri kullanıldı. Bunun için web sunucunun ECLF biçiminde kütükleri kaydedilebilmesi için IIS üzerinde yapılandırma ve tanımlamalar yapıldı.

5.1.2 Verilerin Temizlenmesi

Web sunucu üzerinde saklanan erişim kütük dosyaları, karmaşık, düzensiz ve anlaşılmaz biçimdedirler. Verilerin temizlenmesi, bu erişim kütük dosyalarındaki veriler içerisinden, analiz değeri olmayan ilişkisiz alanların arındırılması, ayıklanması ve belirli bir düzene getirilmesi işlemidir. Karmaşık ve çok zor olan bu işlem süreci, kütüklerin içeriğine göre farklılık göstermektedir. Bu süreçte önemli olan veri kaynağından alınan verilerin işlenirken değerlerin ve bilgilerin orijinalliğini koruması, anlamlı ve doğru sonuçların üretilebilmesi için

başarılı veri tablolarının oluşturulmasıdır. Ayrıca, kütük verileri içerisinde elde edilen alan bilgilerinden, analiz için hangi uygun alanların kullanılması gerektiğinin seçilmesidir. Bilgilerin keşfi ve analiz işlemleri bu alan verileriyle doğrudan ilişkilidir.

Kütük verilerinin temizlenmesi işlemi için SAS tabanlı yazılım ortamında bir yöntem geliştirildi. SAS yazılım ortamında geliştirilen ön işlem süreci aşamalarını gösteren akış şeması Şekil 5.2'de gösterilmektedir.



Şekil 5.2. Ön işlem süreci

Önerilen yöntemin bu aşamasında SAS Base yazılımı kullanılmış, kullanıcı erişim kütük dosyalarındaki analiz değeri olmayan ilişkisiz veriler temizlendi. Uygulama için geliştirilen program kodları Tablo 5.1'de gösterildi. Tablodan görüldüğü üzere, kütük dosyaları içerisindeki resim, grafik, ses, görüntü, stil ve diğer çoklu ortam web sunucudan alınan erişim kütükleri, SAS yazılımının kendi veritabanıyla bütünleştirildi. Veri tabanındaki bu veriler, JAVA tabanlı SAS programlama dili ile gerekli sorgulamalar yaptırılarak istenilen alan bilgileri çıkarıldı.

Tablo 5.1. Veri temizleme için geliştirilen program kodları

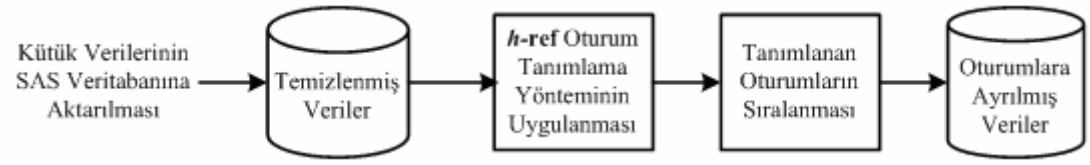
```
libname FIRAT "C:\SAS\resul\Belgelerim\MySASFiles";
data web_log_0006;
set SASUSER.IMPW_0006;

time_hex=put(F2, hex16.);
date_time=catx(" ",F1, substr(F2,1));
where F5 not contains ".png" and F5 not contains ".GIF" and F5 not contains ".gif";
where F5 not contains ".swf" and F5 not contains ".css" and F5 not contains ".CSS";
where F5 not contains "JPG" and F5 not contains "jpg" and F5 not contains ".doc";
where F5 not contains ".ico" and F5 not contains ".ICO" and F5 not contains ".js";
requested_file=lowcase(F5);
referer=lowcase(F8);
IP_adres=F3;
keep IP_adres requested_file referer;
run;
```

Temizleme işleminden sonra, binlerce satırdan oluşan metin tabanlı erişim kütük dosyaları SAS Base yazılımıyla kendi veri tabanı tablosuna aktarıldı. Kütük dosyalarındaki bilgi alanlarının her bir sütunun değişken tipleri tanımlanarak, etiketlendirildi. Veriler, program kodlarında belirtilen klasörde dosya olarak saklanarak, bu dosyalar üzerinde işlem yapıldı.

Tablo 5.1'de görüldüğü üzere *web_log_0006* isimli kütük dosyasından alınan veriler içerisinden, istenilmeyen satırlar ayıklandı. SAS ortamında geliştirilen ön işlem süreci, benzer veri temizleme çalışmalarına göre [13-18, 20-21, 89], büyük boyutlu kütük dosyaları üzerinde çalışabilmesi ve hız açısından üstünlük sağladığı görüldü.

5.1.3 Oturum Tanımlanması

Oturum tanımlama işleminde yaygın olarak kullanılan yöntemler bölüm 3.5'de detaylı olarak anlatılmaktadır. Gerçekleştirilen uygulamada, oturum tanımlama işleminde ise *h*-ref yöntemi kullanılmıştır.



Şekil 5.3. Oturum tanımlama akış şeması

Oturum tanımlamanın bu kısmındaki amaç, web sunucularından alınan web erişim kütük dosyalarındaki gerçekleşmiş olayların zamana göre oturumlarını tespit etmektir. Oturum tanımlama işlemi için geliştirilen SAS kodları, veri temizleme kodları üzerine kurulmuştur. SAS Base yazılımı ile geliştirilen SAS kodları Tablo 5.2'de gösterilmektedir.

Tablo 5.2. Oturum tanımlama işlemi için geliştirilen program kodları

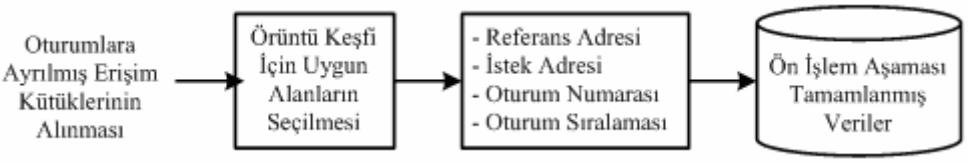
```
libname FIRAT "C:\SAS\resul\Belgelerim\MySASFiles";
session_id=lowcase(catx(" ",time_hex,substr(date_time,1)));

data SASUSER.session_sequence;
set SASUSER.web_log_0007;
    by user_id;
    format sequence 2.;
    retain sequence;
     if first.user_id then sequence = 0;
    sequence = sequence+1;
run;
```

Kullanıcı erişim kütüklerindeki verilerin düzenlenmesinde kodların yanı sıra yazılım üzerindeki temel parametre ayarlamaları ve tanımlamaları yapıldı. Tablo 5.3'de görüldüğü gibi, oturumlar içerisindeki erişim satırları tekrar oturum tanımlama numarasına göre belirlendi.

5.1.4 Özellik İndirgeme

ECLF biçiminde elde edilen erişim kütüklerinden 22 farklı alan bilgisi bulunmaktadır. Bu alan bilgilerinin açıklamaları tezin 3. bölümünde detaylı olarak tablo halinde sunuldu. Yapılan uygulamada gerekli olan alan bilgileri için 22 alan bilgisi içerisinden birbirleriyle ilişkili 4 alan bilgisi seçilerek toplamda 5 alan bilgisi ile veriler örüntü keşfi aşamasına hazırlandı. Bu durum Şekil 5.1'deki blok diyagramında gösterildi.

Örüntü keşfi ve analizi için birbirleriyle ilişkili uygun alanların seçilmesi en önemli aşamadır. Çünkü sonuçta elde edilecek anlamlı ve istatistikî bilgiler, seçilecek bu alanlar ile doğrudan ilişkilidir. Bu nedenle, yapılacak örüntü keşfi ve analizi uygulamalarında uygun özelliklerin ya da alanların seçilmesi gerekmektedir.



Şekil 5.4. Özellik indirgeme akış şeması

Önerilen *Yol analizi* yöntemi ile Web erişim kütüklerinden bilgi çıkarımı için *referans adresi, oturum numarası, istek adresi ve* tespiti yapılan *oturum sıralaması* olmak üzere beş alanlar bilgisi kullanıldı. Giriş tablosu Bölüm 3'de verilen Tablo 3.3'deki biçiminde olan erişim kütükleri ön işlem sürecinden sonra Tablo 5.3'deki çıkış tablosu halini almıştır.

Tablo 5.3. Ön işlemden geçirilmiş Web erişim kütüklerinden örnek kesit

Referans adresi	Oturum numarası	İstek adresi	Oturum sıralaması
-	67e2a4da826149c5 2008-01-14 08:38:17	/default.asp?id=7	1
/default.asp	67e2a4da826149c5 2008-01-14 08:38:17	/?git=fakulteler	2
/?git=fakulteler	67e2a4da826149c5 2008-01-14 08:38:17	/eemuh/default.asp	3
/eemuh/default.asp	67e2a4da826149c5 2008-01-14 08:38:17	/eemuh/dersler_detay.asp	4
/eemuh/dersler_detay.asp	67e2a4da826149c5 2008-01-14 08:38:17	/eemuh/default.asp	5
/eemuh/default.asp	67e2a4da826149c5 2008-01-14 08:38:17	/eemuh/foto/igallery.asp	6
-	67eb39ebce6c73c3 2008-01-14 23:43:23	/default.asp	1
/default.asp	67eb39ebce6c73c3 2008-01-14 23:43:23	/?git=enstituler	2
/?git=enstituler	67eb39ebce6c73c3 2008-01-14 23:43:23	/fenbilimleri/index.htm	3
/fenbilimleri/index.htm	67eb39ebce6c73c3 2008-01-14 23:43:23	/fenbilimleri/dersler.htm	4
/fenbilimleri/dersler.htm	67eb39ebce6c73c3 2008-01-14 23:43:23	/fenbilimleri/computer.htm	5
-	67cb8704bb370ee2 2008-01-14 13:44:04	/default.asp	1

Yol analizi yönteminde arzu edilen başarılı sonuçların elde edilebilmesi için, web kullanıcı erişim kütüklerinin Tablo 5.3'deki biçimde hazırlanması gerekmektedir.

5.1.5 Verilerin Birleştirilmesi ve Dönüştürülmesi

SAS Base yazılımı ile temizlenmiş erişim kütükleri, oturum tanımlama ve özellik indirgeme aşamalarından geçirildikten sonra Tablo 5.3'deki hale getirilmiştir. Oturum numarası çıkarılırken, oturum tanımlama ile erişim zamanı ilişkilendirilmiş, tek bir alan olarak birleştirildi. Elde edilen bu veriler SAS Base yazılımının kendi veritabanında saklandı. Bu verilerin örüntü keşfi ve analizinde kullanılabilmesi için alan bilgileri tek bir dosyada birleştirilerek, örüntü keşfi ve analizinde kullanılabilecek uygun hale dönüştürüldü.

5.2 Örüntü Keşfi

Örüntü keşfi, ön işlem aşamasından geçirilen kütük verilerine veri madenciliği tekniklerinin uygulandığı aşamadır. Madencilik uygulamalarında en sık kullanılan veri madenciliği yöntemleri; istatistiksel yöntemler, eşleştirme kuralları, kümeleme, sınıflandırma ve sıralı örüntülerdir [149].

Uygulamamızın bu kısmında elde edilen temizlenmiş verilerden örüntü keşfi işlemi yapıldı. 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 başarımı artırıcı, sistem güvenliğini genişletici, düzeltme işlemlerini kolaylaştırıcı ve pazarlama kararlarını destekleyici raporlar çıkartılabilir [6].

Yapılan uygulamanın, ön işlem aşamasından geçirilmiş web erişim kütüklerini analiz etmek için istatistiksel bir yöntem olan *Yol Analizi* yöntemi kullanıldı. Yol analizi, bir web sitesi yoluyla sayfalar arasında gezinen ziyaretçilerin gezindiği yollarını tanımlamak için kullanıldı. Aynı zamanda, web köprü bağlantıları arasındaki eşleştirme analizini ve sıralı birliktelik kurallarını çıkarmak içinde kullanıldı.

5.2.1 Yol Analizi Yöntemi

Yol analizi ilk kez Amerikalı genetik uzmanı Dr. Sewall Wright tarafından 1921 yılında tanımlanmıştır. Yol analizi yöntemi birçok bilim alanında verim ve verimli öğeler arasındaki etkileşimleri tespit etmek amacıyla kullanılmaktadır [120]. Yol analizi yönteminin en zor ve en önemli kısmı yol diyagramının oluşturulmasıdır. Yol diyagramı sayısal analizler için gerekli olmamasına rağmen değişkenler arasındaki doğrudan ve dolaylı ilişkilerin ortaya konulması açısından oldukça kullanışlıdır [120]. Değişkenler arasındaki ilişkilerin yanlış gösterilmesi

sonucunda analiz yapıldığında etkilerin toplamı, toplam korelâsyonu vermeyip toplam korelâsyondan önemli bir sapma göstermektedir.

Yol analizi tekniği, hiyerarşik çoklu bir regresyon analizidir. Her istatistik analiz tekniğinde olduğu gibi yol analiz tekniğinin de bazı önemli varsayımları bulunmaktadır. Bu varsayımlar; modelde yer alan değişkenler arasındaki ilişkiler, doğrusal, eklenebilir ve sebep sonuç ilişkisine dayanmalıdır. Model içerisindeki hatalar kendi aralarında ve modeldeki diğer değişkenlerle ilişkili olmamalıdır. Tek yönlü bir sebep akısı olmalıdır. Ölçümler niceleyici değişkenlerden elde edilmiş olmalıdır. Ölçümler hatasız olarak yapılmalıdır.

5.2.1.1 Yol Katsayılarının Hesaplanması

X bağımsız değişkenindeki bir birim sapmanın *Y* bağımlı değişkeni üzerinde yapmış olduğu etki;

$$P_{yx} = b\frac{S_x}{S_y} \tag{5.1}$$

şeklinde gösterilir [120]. Burada;

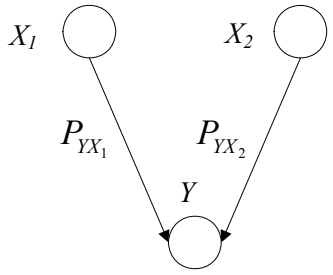
P_{xy} : *X* bağımsız değişkeninin *Y* bağımlı değişkeni üzerinde yapmış olduğu doğrudan etkiyi gösteren Yol katsayısıdır.

b: Kısmi regresyon katsayını göstermektedir.

$$S_x = \sqrt{\left(\sum (x-\bar{x})^2\right).\frac{1}{n}} \tag{5.2}$$

$$S_y = \sqrt{\left(\sum (y-\bar{y})^2\right).\frac{1}{n}} \tag{5.3}$$

İncelenen karakteri (neticeyi) etkileyen faktörlerle ilgili olarak iki durum söz konusudur. Bunlardan birinci durum, sebep değişkenleri arasında herhangi bir ilişki olmamasıdır (Sekil 5.5). İkinci durumda ise sebep değişkenlerinin birbirleriyle ilişkili oldukları durumdur (Şekil 5.6). Şekil 5.5 incelendiğinde sebep değişkenleri X_1 ve X_2 arasında herhangi bir ilişki olmadığı görülmektedir. X_1 ve X_2 değişkenlerinin *Y* değişkeni üzerinde yapmış olduğu etki Eşitlik 5.6'de verilen çoklu regresyon denklemi ile gösterilebilir;



Şekil 5.5. Birbirleriyle ilişkisi olmayan X_1, X_2 faktörleri ile sonuç *(Y)* arasındaki ilişkiler

$$Y = b_1 X_1 + b_2 X_2 \tag{5.4}$$

Y değişkeninde meydana gelen toplam varyasyon, varyasyona neden olan kaynaklara ayrıldığında;

$$\sigma_y^2 = b_1 \sigma_{x_1}^2 + b_2 \sigma_{x_2}^2 \tag{5.5}$$

eşitliği elde edilir. Burada,

σ_y^2: Eşitlik 3 ile elde edilen *Y* değişkenine ait varyansı belirtmektedir.

$\sigma_{x_1}^2$ ve $\sigma_{x_2}^2$: Eşitlik 5.2 ile elde edilen birinci ve ikinci *X* değişkenine ait varyansı, *b*'ler ise kısmi regresyon katsayılarını göstermektedir. 5.5 numaralı eşitliğin her iki tarafı σ_y^2'ye bölünürse;

$$\frac{\sigma_{y^2}}{\sigma_{y^2}} = b_1 \frac{\sigma_{X_1^2}}{\sigma_{y^2}} + b_2 \frac{\sigma_{X_2^2}}{\sigma_{y^2}} \tag{5.6}$$

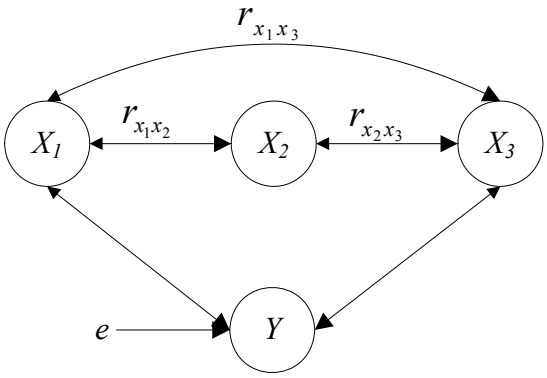
eşitliği elde edilir. Bu eşitliğin sol tarafı 1'e eşit olur. Eşitliğin sağ tarafı ise sırasıyla X_1 ve X_2 değişkenlerinin, Y değişkeninde meydana getirdikleri varyasyondaki paylarını göstermektedir. Buradan;

$$1 = b_1 \frac{\sigma_{x_1}}{\sigma_y} + b_2 \frac{\sigma_{x_2}}{\sigma_y} \tag{5.7}$$

eşitliği elde edilir. Bu eşitliğin sağında kalan birinci terim, X_1 değişkeninin *Y* değişkeni üzerinde yapmış olduğu etkiyi (Pyx_1), ikinci terim ise X_2 değişkeninin Y değişkeni üzerinde yapmış olduğu etkiyi (Pyx_2) göstermektedir. Her iki terim de ayni zamanda standardize edilmiş regresyon katsayılarıdır ve Eşitlik 1'deki Yol katsayısı tanımını vermektedirler. Yani;

$$P_{YX_1} = b_1 \frac{\sigma_{X_1}}{\sigma_Y};\ P_{YX_2} = b_2 \frac{\sigma_{X_2}}{\sigma_Y} \tag{5.8}$$

Buraya kadar anlatılanlardan, Yol katsayılarının aslında standardize edilmiş regresyon katsayıları olduğu görülmektedir ve P_{YX_1}; X_1 sebebinin bir standart sapma değişmesi ile Y neticesinde kendi standart sapması cinsinden meydana gelecek değişme miktarı olarak tanımlanır. Daha önce yapılan tanımlamaya da uygun olan bu tanım aynı zaman da regresyon katsayısının tarifine de uymaktadır. Eğer X_1 ve X_2 değişkenleri Şekil 5.4'de olduğu gibi bağımlı değişkenler ise değişkenler arasında birde kovaryans etkisi görülecektir.



Şekil 5.6. Birbirleriyle ilişkisi olan X_1, X_2 ve X_3 faktörleri ile sonuç *(Y)* arasındaki ilişkiler

Şekil 5.6'daki X_1, X_2 ve X_3 bağımsız değişkenlerinin Y bağımlı değişkeni ile olan ilişkileri çoklu regresyon denklemi,

$$Y = b_1X_1 + b_2X_2 + b_3X_3 + e \tag{5.9}$$

ile gösterilebilir. Y bağımlı değişkeninde gözlenen varyasyon, kaynaklarına parçaladığında bu kaynaklar içinde bu değişkenlerde gözlenen varyansın dışında, X_1, X_2 ve X_3 değişkenleri arasında karşılıklı ilişki olduğundan değişkenler arasında kovaryans da mevcuttur ve bu durum,

$$\sigma_y^2 = b_1\sigma_{x_1}^2 + b_2\sigma_{x_2}^2 + b_3\sigma_{x_3}^2 + 2Cov(X_1,X_2) + 2Cov(X_1,X_3) + 2Cov(X_2,X_3) \tag{5.10}$$

ile gösterilebilir. Bu durumda, X_1 değişkeni ile Y değişkeni arasında gözlenen korelasyon;

$$r(X_1,Y) = \frac{Cov(X_1, b_1X_1 + b_2X_2 + b_3X_3 + e)}{\sqrt{Var(X_1).Var(Y)}} \tag{5.11}$$

ile elde edilir. Eşitlik 5.11'deki X_1 değişkeninin diğer değişkenlerle olan kovaryansı parçalandığında 5.12 numaralı eşitlik elde edilir.

$$r(X_1,Y) = \frac{Cov(X_1,X_1)}{\sqrt{V(X_1).V(Y)}} + \frac{Cov(X_1,X_2)}{\sqrt{V(X_1).V(Y)}} + \frac{Cov(X_1,X_3)}{\sqrt{V(X_1).V(Y)}} + \frac{Cov(X_1,e)}{\sqrt{V(X_1).V(Y)}} \tag{5.12}$$

Burada; $Cov(X_1,X_1) = V(X_1), Cov(X_1,e) = 0$ 'dır.

X_1 ve X_2 değişkenleri arasındaki korelasyon;

$$r(X_1,X_2) = \frac{Cov(X_1,X_2)}{\sigma_{X_1}.\sigma_{X_2}} \tag{5.13}$$

ile hesaplanır. Eşitlik 13'de içler dışlar çarpımı yapıldığında,

$$Cov(X_1,X_2) = r(X_1,X_2)\ \ \sigma_{X_1}.\sigma_{X_2} \tag{5.14}$$

elde edilir. Bulunan bu değerler 5.12 numaralı eşitlikte yerine konulduğunda,

$$r(X_1,Y)=\frac{\sigma^2 X_1}{\sigma_{X_1}\cdot\sigma_Y}+\frac{r(X_1,X_2)\cdot\sigma_{X_1}\cdot\sigma_{X2}}{\sigma_{X_1}\cdot\sigma_Y}+\frac{r(X_1,X_3)\cdot\sigma_{X_1}\cdot\sigma_{X3}}{\sigma_{X_1}\cdot\sigma_Y}+0$$

$$=\frac{\sigma_{X_1}}{\sigma_Y}+r(X_1,X_2)\frac{\sigma_{X_2}}{\sigma Y}+r(X_1,X_3)\frac{\sigma_{X_3}}{\sigma Y} \quad (5.15)$$

eşitliği elde edilir. Eşitlik 5.15, Yol katsayıları cinsinden ifade edilirse,

$$r(X_1,Y)=P_{YX_1}+r_{X_1X_2}P_{YX_2}+r_{X_1X_3}P_{YX_3} \quad (5.16)$$

elde edilir. Görüldüğü gibi X_1 değişkeni ile Y değişkeni arasında üç yoldan etkileşim vardır.

X_1'in Y üzerindeki doğrudan etkisi (P_{YX_1})

X_1'in X_2 üzerinden olan dolaylı etkisi $(r_{X_1X_2}\cdot P_{YX_2})$

X_1'in X_3 üzerinden olan dolaylı etkisi $(r_{X_1X_3}\cdot P_{YX_3})$

Aynı yaklaşımla X_2 ve X_3 değişkenleri ile Y değişkenleri arasındaki etkileşimler bulunduğunda,

$$r(X_2,Y)=r_{X_1X_2}\cdot P_{YX_1}+P_{YX_2}+r_{X_2X_3}\cdot P_{YX_3} \quad (5.17)$$

$$r(X_3,Y)=r_{X_1X_3}\cdot P_{YX_1}+r_{X_2X_3}\cdot P_{YX_2}+P_{YX_3} \quad (5.18)$$

eşitlikleri elde edilir. 5.16, 5.17, 5.18 nolu eşitlikler matris notasyonuyla,

$$A=B\cdot C \quad (5.19)$$

şeklinde ifade edilebilir. Burada;

$$A=\begin{matrix}r(x_1,y)\\ r(x_2,y)\\ r(x_3,y)\end{matrix} \quad B=\begin{matrix}r_{X_1X_1} & r_{X_1X_2} & r_{X_1X_3}\\ r_{X_2X_1} & r_{X_2X_2} & r_{X_2X_3}\\ r_{X_3X_1} & r_{X_3X_2} & r_{X_3X_3}\end{matrix} \quad C=\begin{matrix}P_{yx_1}\\ P_{yx_2}\\ P_{yx_3}\end{matrix} \quad (5.20)$$

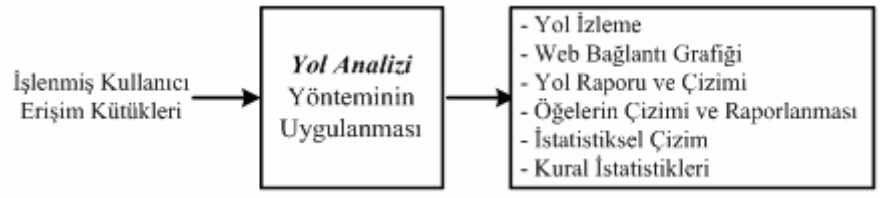
olup A; ilişkisini incelediğimiz değişkenler arasındaki korelasyon katsayılarından oluşan vektörü; B; üç denklemde de yer alan korelasyon katsayılarından oluşan vektörü ve C; Yol katsayılarından oluşan vektörü göstermektedir. Eşitlik 5.20'de, bir matrisin tersi ile kendisinin çarpımı birim matrisi vereceğinden, Yol katsayıları vektörü C; B matrisinin tersi ile A matrisinin çarpımına eşit olup,

$$C=B^{-1}.A \quad (5.21)$$

şeklinde hesaplanır.

5.3 Örüntü Analizi

Web kullanım madenciliği uygulamasının en son aşaması, örüntü analizidir. Örüntü analizinin amacı, örüntü keşfi işleminde ortaya çıkarılan ilişkisiz kurallar, örüntüler ya da istatistiklerin analiz edilerek, anlamlı bilgiler ve ilişkili ilginç kurallar, örüntüler ya da istatistiklerin ortaya çıkarılması işlemidir. Uygulamada örüntü analizi sonucu elde edilen bulgular Şekil 5.7'de gösterilmektedir.



Şekil 5.7. Örüntü analizi

Örüntü analizinde, bilgi sorgulama ve OLAP uygulamaları ile derinlemesine incelemeler yapılabilmektedir [153]. Örüntü analizi işlemleri için kullanılan birçok yazılım araçları herhangi bir veri tabanı (SQL, Oracle, v.b.) ile ilişkilidir. Uygulamamızda SAS yazılımının kendi içerisinde yer alan veri tabanı kullanılmıştır. Görsel tekniklerde, daha çok grafiksel örüntüler, farklı değerlerde yoğun ve dikkat çeken örüntüler, işaretlenmiş renkler göz önünde bulundurulmaktadır.

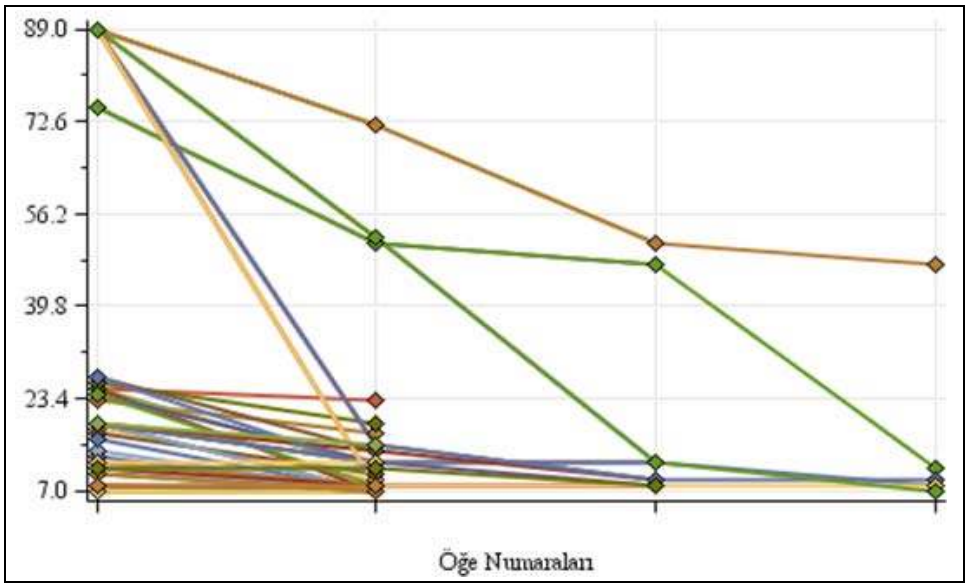
5.4 Çıkarılan İlginç Kurallar ve Örüntülerin İncelenmesi

Bu çalışmada, Fırat Üniversitesi Bilgi İşlem Daire Başkanlığında hizmet veren web sunucularından alınan web kullanıcı erişim kütük dosyaları kullanıldı. Erişim kütüklerinin boyutunun büyük olmasından dolayı bu uygulamada 2 günlük verilerin bulunduğu tek bir kütük dosyası kullanıldı. Seçilen dosya 14 -15 Ocak 2008 tarihlerindeki web kullanıcı erişim kütük dosyasıdır. Dosya içerisinde 146.178 satır bulunmaktadır. Bu kütük dosyası SAS Base yazılım ortamında işlenerek ön işlem aşamasından geçirilip, temizlenmiş erişim kütük verilerine *yol analizi* yöntemi uygulandı. Web kullanıcı erişim kütüklerine yol analizi yöntemi uygulandığı zaman yol tamamlama işleminde her zaman başarıya ulaşılamamaktadır. Örneğin, oturum tanımlama algoritması ile belirtilen zaman aşımı süresi içerisinde bir ziyaretçi siteden ayrılırsa ve siteye tekrar geri dönerse, yol tamamlama işlemi hatalı olabilir. Aynı zamanda, bir kullanıcı mevcut web sayfasındaki durum bağlantı adreslerinin yerine tarayıcıda tanımlanmış olan bağlantı adreslerini kullanarak bir site içerisinde gezinirse ya da bir siteye giriş yaparsa, bu durum yol tamamlama işleminde başarısızlıklara yol açabilir.

Yol analizi uygulamasının sonucunda web sitesinde yol izleme, web sayfaları arasındaki köprü bağlantıları, web sayfaları arasındaki ilişkiler, köprüler arasındaki direk ya da dolaylı bağlantı durumları ve web sitesine ait kural istatistikleri gibi anlamlı ve ilginç bilgiler çıkarıldı.

5.4.1 Yol İzleme

Yol analizi yönteminin yol izleme özelliği sonucu, ziyaretçilerin web sitesi içerisinde ilgi duyduğu web sayfaları arasında gezinirken bir sayfadan diğer bir sayfaya geçişteki ziyaretçi sayılarını ve ilgili web sayfalarını göstermektedir [82]. Yani, bu özellik ile web sitesi içerisindeki alt klasörlere ya da alt bölümlere yerleştirilmiş olan web sayfalarına olan ilgi ve gezinen kullanıcı sayıları kolayca tespit edilir. Web analizi için önemli olan yol izleme özelliği, kullanıcılar tarafından en çok izlenen yollar ve site içerisinde en derine gidilen alt sayfalar sayısal olarak büyükten küçüğe doğru tespit edilebilmektedir. Bu tespit sonucunda ziyaretçiler tarafından yüksek düzeyde ilgi gören web sayfaları, site içerisindeki alt kısımlardan alınarak, daha kısa yoldan ve kolay erişilebilen uygun noktalarda köprü bağlantısı verilebilir. Böylece, kullanıcıların o sayfaya ulaşması için izleyeceği uzun yollar bertaraf edilerek, yeni kısa yollar tanımlanmış olacaktır. Sonuçta, web sitesine erişim sağlayan kullanıcıların memnuniyeti artacaktır. Fırat Üniversitesi web sitesine ait kullanıcı erişim kütüklerinden yapılan yol izleme analizi sonucunda elde edilen öğelerin ilk on satırlık bölümü, Tablo 5.4'te, yol izleme analizinin grafiği ise Şekil 5.8'de gösterilmektedir.



Şekil 5.8. Yol izleme grafiği

Tablo 5.4. Yol izleme kuralları

Kural No	Kural	Öğe-1	Öğe-2	Öğe-3	Öğe-4
1	/sosyalbil/sonsite/index.htm==>/sosyalbil/sonsite/ilan.htm	89	72	.	.
2	/sosyalbil/sonsite/index.htm==>/sosyalbil/sonsite/index.htm	89	52	.	.
3	/sosyalbil/sonsite/index.htm==>/sosyalbil/sonsite/ilan.htm==>/sosyalbil/sonsite/index.htm==>/sosyalbil/sonsite/index.htm	89	72	51	.
4	/sosyalbil/sonsite/ilan.htm==>/sosyalbil/sonsite/index.htm	75	51	.	.
5	/sosyalbil/sonsite/index.htm==>/sosyalbil/sonsite/ilan.htm==>/sosyalbil/sonsite/index.htm==>/sosyalbil/sonsite/index.htm	89	72	51	47
6	/sosyalbil/sonsite/ilan.htm==>/sosyalbil/sonsite/index.htm==>/sosyalbil/sonsite/index.htm	75	51	47	.
7	/med/dahili/dahilitip.htm==>/med/dahili/dahiliabd.htm	25	23	.	.
8	/med/temel/temeltip.htm==>/med/temel/temelabd.htm	26	19	.	.
9	/med/cerrahi/cerrahitip.htm==>/med/cerrahi/cerrahiabd.htm	23	17	.	.
10	/veteriner/akadamik_database/picture/inc_default.asp==>/veteriner/akadamik_database/picture/cat.asp	19	15	.	.
11	/sosyalbil/sonsite/index.htm==>/sosyalbil/sonsite/anabilim.htm	89	15	.	.
12	/sosyalbil/sonsite/anabilim.htm==>/sosyalbil/sonsite/index.htm	19	14	.	.
13	/veteriner/akadamik_database/akademik_personel_ic.asp==>/veteriner/akadamik_database/akedemikgoster.asp27	27	14	.	.
14	/sosyalbil/sonsite/anabilim.htm==>/sosyalbil/aq.html.swf==>/sosyalbil/sonsite/anabilim.htm	19	12	12	.
15	/veteriner/akademik_personel.asp==>/veteriner/akadamik_database/akademik_personel_ic.asp	12	12	.	.
16	/sosyalbil/sonsite/anabilim.htm==>/sosyalbil/aq.html.swf	19	12	.	.
17	/sosyalbil/aq.html.swf==>/sosyalbil/sonsite/anabilim.htm	24	12	.	.
18	/sosyalbil/sonsite/index.htm==>/sosyalbil/sonsite/index.htm==>/sosyalbil/sonsite/ilan.htm	89	52	12	.
19	/veteriner/resim.asp==>/veteriner/akadamik_database/picture/inc_default.asp	11	11	.	.
20	/med/ogrenciler/ogrenciler.htm==>/med/ogrenciler/1sinif.htm	25	11	.	.
21	/veteriner/akadamik_database/akademik_personel_ic.asp==>/veteriner/akadamik_database/kimlik_goster.asp	27	11	.	.
22	/sosyalbil/sonsite/ilan.htm==>/sosyalbil/sonsite/index.htm==>/sosyalbil/sonsite/index.htm==>/sosyalbil/sonsite/ilan.htm	75	51	47	11
23	/med/image/galeri/galeri.htm==>/med/image/galeri/dekanlik/index.htm	17	10	.	.
24	/sosyalbil/sonsite/index.htm==>/sosyalbil/sonsite/anabilim.htm==>/sosyalbil/aq.html.swf==>/sosyalbil/sonsite/anabilim.htm	89	15	9	9
25	/sosyalbil/sonsite/anabilim.htm==>/sosyalbil/sonsite/index.htm==>/sosyalbil/sonsite/anabilim.htm	19	14	9	.
26	/sosyalbil/sonsite/index.htm==>/sosyalbil/sonsite/anabilim.htm==>/sosyalbil/aq.html.swf	89	15	9	.
27	/elkbilgi/lisansindex.asp==>/elkbilgi/dindex.asp	18	9	.	.
28	/sosyalbil/sonsite/genelbilgi.htm==>/sosyalbil/aq.html.swf==>/sosyalbil/sonsite/genelbilgi.htm	10	8	8	.
29	/sosyalbil/sonsite/index.htm==>/sosyalbil/sonsite/genelbilgi.htm==>/sosyalbil/aq.html.swf	89	8	8	.
30	/turkce/akademik.html==>/turkce/hoca.html	8	8	.	.

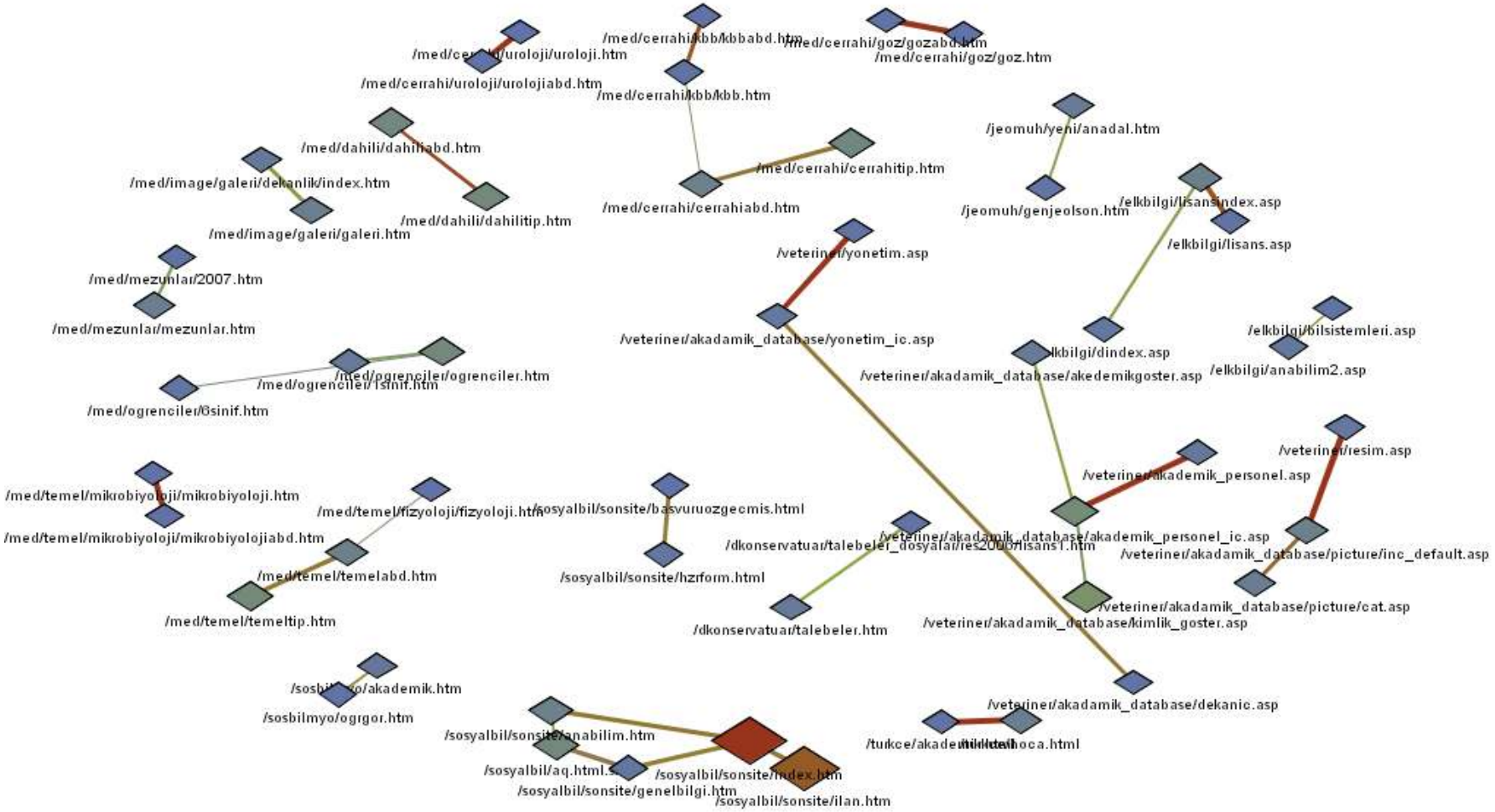
31	/sosyalbil/sonsite/index.htm==>/sosyalbil/sonsite/genelbilgi.htm==>/sosyalbil/aq.html.swf==>/sosyalbil/sonsite/genelbilgi.htm	89	8	8	8
32	/veteriner/yonetim.asp==>/veteriner/akadamik_database/yonetim_ic.asp	8	8	.	.
33	/sosyalbil/sonsite/genelbilgi.htm==>/sosyalbil/aq.html.swf	10	8	.	.
34	/veteriner/akadamik_database/yonetim_ic.asp==>/veteriner/akadamik_database/dekanic.asp	11	8	.	.
35	/veteriner/resim.asp==>/veteriner/akadamik_database/picture/inc_default.asp==>/veteriner/akadamik_database/picture/cat.asp	11	11	8	.
36	/sosyalbil/aq.html.swf==>/sosyalbil/sonsite/anabilim.htm==>/sosyalbil/sonsite/index.htm	24	12	8	.
37	/sosbilmyo/akademik.htm==>/sosbilmyo/ogrgor.htm12	8	.	.	
38	/sosyalbil/sonsite/anabilim.htm==>/sosyalbil/aq.html.swf==>/sosyalbil/sonsite/anabilim.htm==>/sosyalbil/sonsite/index.htm	19	12	12	8
39	/sosyalbil/aq.html.swf==>/sosyalbil/sonsite/genelbilgi.htm	24	8	.	.
40	/sosyalbil/sonsite/index.htm==>/sosyalbil/sonsite/genelbilgi.htm	89	8	.	.
41	/med/cerrahi/uroloji/uroloji.htm==>/med/cerrahi/uroloji/urolojiabd.htm	7	7	.	.
42	/med/cerrahi/goz/goz.htm==>/med/cerrahi/goz/gozabd.htm	7	7	.	.
43	/med/temel/mikrobiyoloji/mikrobiyoloji.htm==>/med/temel/mikrobiyoloji/mikrobiyolojiabd.htm	7	7	.	.
44	/med/cerrahi/kbb/kbb.htm==>/med/cerrahi/kbb/kbbabd.htm	8	7	.	.
45	/elkbilgi/lisans.asp==>/elkbilgi/lisansindex.asp8	7	.	.	
46	/sosyalbil/sonsite/genelbilgi.htm==>/sosyalbil/sonsite/index.htm	10	7	.	.
47	/sosyalbil/sonsite/hzrform.html==>/sosyalbil/sonsite/basvuruozgecmis.html	10	7	.	.
48	/elkbilgi/anabilim2.asp==>/elkbilgi/bilsistemleri.asp	11	7	.	.
49	/sosyalbil/sonsite/index.htm==>/sosyalbil/sonsite/index.htm==>/sosyalbil/sonsite/ilan.htm==>/sosyalbil/sonsite/index.htm	89	52	12	7
50	/dkonservatuar/talebeler.htm==>/dkonservatuar/talebeler_dosyalar/res2006/lisans1.htm	13	7	.	.
51	/jeomuh/yeni/anadal.htm==>/jeomuh/genjeolson.htm	14	7	.	.
52	/med/mezunlar/mezunlar.htm==>/med/mezunlar/2007.htm	16	7	.	.
53	/med/cerrahi/cerrahiabd.htm==>/med/cerrahi/kbb/kbb.htm	19	7	.	.
54	/med/temel/temelabd.htm==>/med/temel/fizyoloji/fizyoloji.htm	19	7	.	.
55	/med/ogrenciler/ogrenciler.htm==>/med/ogrenciler/6sinif.htm	25	7	.	.

5.4.2 Köprü Bağlantı Grafiği

Web sitesi ziyaretçilerinin takip ettikleri yolları modellemek için, ağırlıklandırılmış bir web bağlantı grafiği oluşturmak gerekmektedir. Web bağlantı grafiği, kullanıcı erişim kütüklerinden çıkarılan köprü bağlantısına ilişkin kurallara ait bilgi kümesinin grafiksel olarak gösterimidir. Web bağlantı grafiğinde her bir düğüm bir web sayfasını temsil etmektedir. Hedef web sayfası ile kaynak web sayfası arasındaki bağlantı geçiş sayısı, sayfalar arasındaki bağlantı ağırlığını vermektedir. Aynı şekilde, bu durum web bağlantı grafiğindeki düğümler arasındaki çizgilerin kalınlığına da yansımaktadır. Yani, bağlantı ağırlık değeri yüksek olan çizgiler daha kalın olarak çizilmektedir.

Web bağlantı grafiği, her kullanıcının ziyaretleri için ayrı ayrı işlem yapmaktadır. Kullanıcının web sitesine ilk girdiği sayfadan başlar, siteden ayrılmadan önce en son girmiş olduğu web sayfasına kadar devam eder. Kullanıcının en son girmiş olduğu web sayfası çıkış sayfasıdır. Bağlantı grafiğinde herhangi iki düğüm arasındaki yönlendirilmiş bir yolu tespit etmek için çıkış düğümünden giriş düğümüne kadar bir bağlantı eklenmektedir. Web sitesindeki bir sayfaya gelen tüm bağlantıların ağırlıklarının miktarı sayfadan ayrılan tüm bağlantıların ağırlıklarının miktarı ile aynı olmayabilir. Bu durumu çözmek için giriş/çıkış düğümüne ekstra gelen ağırlıklar ile ekstra çıkan ağırlıklar arasında dağıtım yapılarak bir denge sağlanabilir.

Önerilen yöntemin uygulamasında, Fırat Üniversitesi web sunucularındaki kullanıcı erişim kütük dosyaları kullanılarak oluşturulan köprü bağlantı grafiği Şekil 5.9'da gösterilmektedir. Web bağlantı grafiğinin oluşturulmasında düğüm veri tablosu ve bağlantı veri tablosu olmak üzere iki veri tablosu kullanıldı. Şekil 5.9'da da görüldüğü gibi, */veteriner/yonetim.asp* ve */veteriner/akadamik_database/yonetim_ic.asp* düğümleri arasındaki bağlantı çizgisi daha kalındır. Bu durum, bu düğümler arasındaki güven değerinin yüksek olduğunu göstermektedir. Bunun yanı sıra, */med/ogrenciler/ogrenciler.htm* ve */med/ogrenciler/1sinif.htm* düğümleri arasındaki bağlantı çizgisinin daha ince olduğu görülmektedir. Bu durum da, bu düğümler arasındaki güven değerinin daha düşük olduğunu belirtir. Bu kuralların her ikisi de benzer destek değerine sahiptir. Grafikteki düğümler arasındaki bazı bağlantılar gösterilmemiştir. Bunun sebebi ise grafik oluşturulurken web bağlantı eşik değerinin %10 olarak tanımlanmasıdır. Eşik değeri arttıkça oluşturulan web bağlantı grafiğinde düğümler arasındaki bağlantı çizgileri de azalmaktadır.



Şekil 5.9. Köprü bağlantı grafiği

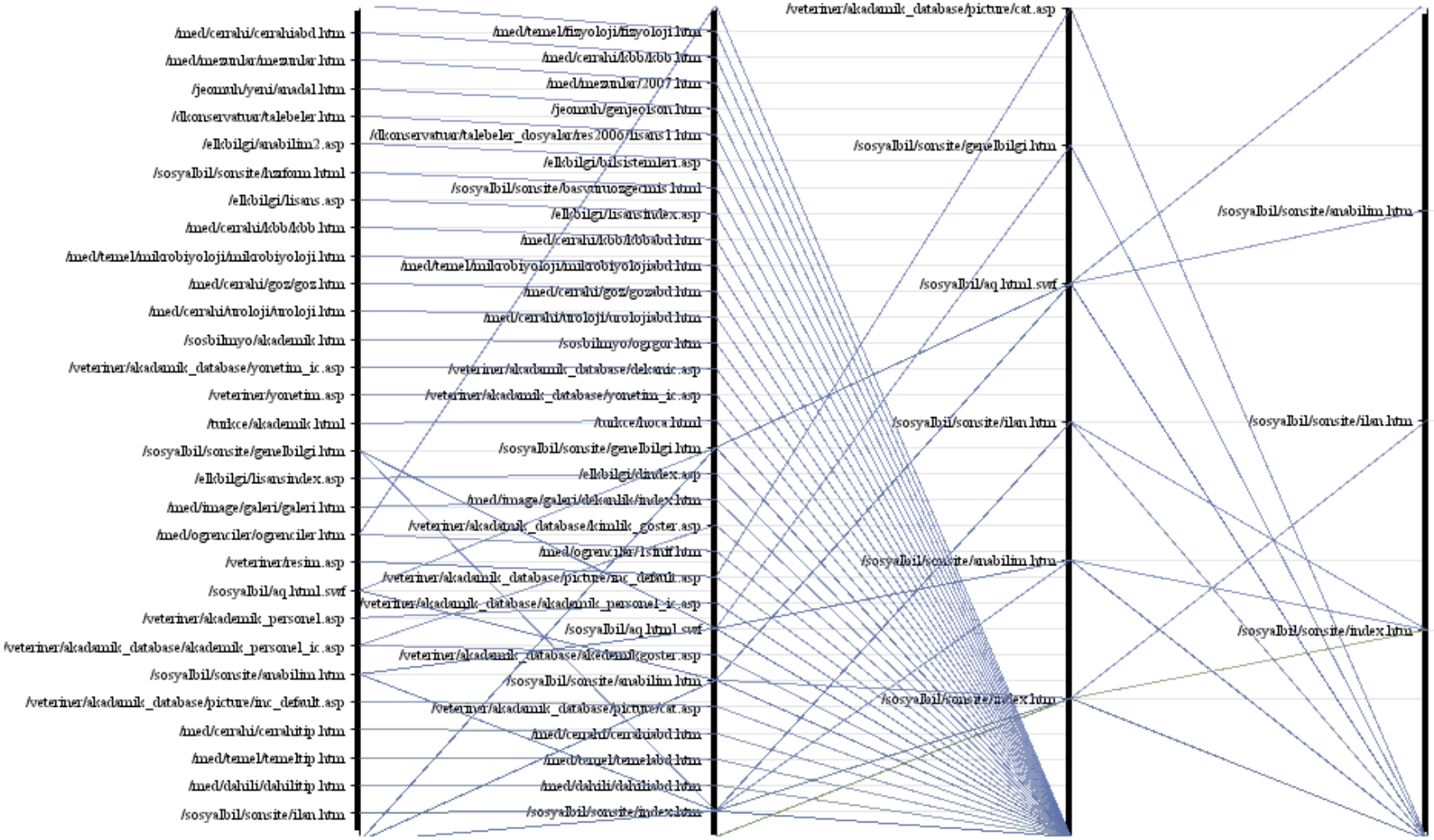
5.4.3 Yol Raporu ve Yol Çizimi

Bu bölümde, yol analizi ile elde edilen yol raporu bilgisi ve sayfalar arasındaki bağlantıları gösteren yol çizim grafiği verilmektedir. Tablo 5.5'te görüldüğü gibi, güven ve destek değeri en yüksek değere sahip olan kural */sosyalbil/sonsite/index.htm* → */sosyalbil/sonsite/ilan.htm* adreslerinin içermiş olduğu kuraldır. Bu kuralda, */sosyalbil/sonsite/index.htm* adresine giren bir ziyaretçi %80.899 olasılıkla */sosyalbil/sonsite/ilan.htm* adresini de ziyaret etmektedir. Çıkarılan bu kurallar analiz edildiğinde, web sitesi ziyaretçilerinin site içerisinde tercih ettiği web sayfaları görülebilmektedir. Bunun sonucunda ziyaretçi memnuniyeti ve site organizasyonunda standardizasyonu sağlamak için web sitesinde tercih edilen ve yoğun ilgi gösterilen sayfalar yeniden düzenlenir. Düzenleme işleminde en çok ve birlikte ziyaret edilen sayfalar, ardı ardına ziyaret edilen web sayfaları gibi bilgiler site tasarımında dikkate alınmalıdır. Tablo 5.5'deki kurallara göre, Şekil 5.10'daki yol çizimi oluşturuldu. Uygulamada maksimum zincir sayısı (öğe) 4 olarak seçildi. Bu değer web site yoğunluğuna göre yapılandırılabilir. Analizi yapılan kütükler için maksimum zincir sayısı 5'tir ancak değerler 4'e göre alınarak sonuçlara işlenmiştir.

Yol raporlama bilgileri elektronik ticaret yapan web sitelerinde müşteri gruplarını ayrıştırmada kullanılabilir. Bunun sonucunda, bilinçli ve ilgili ziyaretçi grupları tespit edilerek elektronik ticarette kazanımlar elde etmek için web madenciliği çalışmaları yapılmasında, yol raporlama bilgileri katkı sağlayacaktır.

Tablo 5.5. Yol bildirimi kuralları

Kural No	Zincir No	Sayma	Destek	Güven	Kural
1	2	72	59.504	80.899	/sosyalbil/sonsite/index.htm ==> /sosyalbil/sonsite/ilan.htm
2	2	52	42.975	58.427	/sosyalbil/sonsite/index.htm ==> /sosyalbil/sonsite/index.htm
3	3	51	42.149	70.833	/sosyalbil/sonsite/index.htm ==> /sosyalbil/sonsite/ilan.htm ==> /sosyalbil/sonsite/index.htm
4	2	51	42.149	68.000	/sosyalbil/sonsite/ilan.htm ==> /sosyalbil/sonsite/index.htm
5	4	47	38.843	92.157	/sosyalbil/sonsite/index.htm ==> /sosyalbil/sonsite/ilan.htm ==> /sosyalbil/sonsite/index.htm ==> /sosyalbil/sonsite/index.htm
6	3	47	38.843	92.157	/sosyalbil/sonsite/ilan.htm ==> /sosyalbil/sonsite/index.htm ==> /sosyalbil/sonsite/index.htm
7	2	23	19.008	92.000	/med/dahili/dahilitip.htm ==> /med/dahili/dahiliabd.htm
8	2	19	15.702	73.077	/med/temel/temeltip.htm ==> /med/temel/temelabd.htm
9	2	17	14.050	73.913	/med/cerrahi/cerrahitip.htm ==> /med/cerrahi/cerrahiabd.htm
10	2	15	12.397	78.947	/veteriner/akadamik_database/picture/inc_default.asp ==> /veteriner/akadamik_database/picture/cat.asp
11	2	15	12.397	16.854	/sosyalbil/sonsite/index.htm ==> /sosyalbil/sonsite/anabilim.htm
12	2	14	11.570	73.684	/sosyalbil/sonsite/anabilim.htm ==> /sosyalbil/sonsite/index.htm
13	2	14	11.570	51.852	/veteriner/akadamik_database/akademik_personel_ic.asp ==> /veteriner/akadamik_database/akedemikgoster.asp
14	3	12	9.917	100.000	/sosyalbil/sonsite/anabilim.htm ==> /sosyalbil/aq.html.swf ==> /sosyalbil/sonsite/anabilim.htm
15	2	12	9.917	100.000	/veteriner/akademik_personel.asp ==> /veteriner/akadamik_database/akademik_personel_ic.asp
16	2	12	9.917	63.158	/sosyalbil/sonsite/anabilim.htm ==> /sosyalbil/aq.html.swf
17	2	12	9.917	50.000	/sosyalbil/aq.html.swf ==> /sosyalbil/sonsite/anabilim.htm
18	3	12	9.917	23.077	/sosyalbil/sonsite/index.htm ==> /sosyalbil/sonsite/index.htm ==> /sosyalbil/sonsite/ilan.htm
19	2	11	9.091	100.000	/veteriner/resim.asp ==> /veteriner/akadamik_database/picture/inc_default.asp
20	2	11	9.091	44.000	/med/ogrenciler/ogrenciler.htm ==> /med/ogrenciler/1sinif.htm
21	2	11	9.091	40.741	/veteriner/akadamik_database/akademik_personel_ic.asp ==> /veteriner/akadamik_database/kimlik_goster.asp
22	4	11	9.091	23.404	/sosyalbil/sonsite/ilan.htm ==> /sosyalbil/sonsite/index.htm ==> /sosyalbil/sonsite/index.htm ==> /sosyalbil/sonsite/ilan.htm
23	2	10	8.264	58.824	/med/image/galeri/galeri.htm ==> /med/image/galeri/dekanlik/index.htm
24	4	9	7.438	100.000	/sosyalbil/sonsite/index.htm ==> /sosyalbil/sonsite/anabilim.htm ==> /sosyalbil/aq.html.swf ==> /sosyalbil/sonsite/anabilim.htm
25	3	9	7.438	64.286	/sosyalbil/sonsite/anabilim.htm ==> /sosyalbil/sonsite/index.htm ==> /sosyalbil/sonsite/anabilim.htm



Şekil 5.10. Yol çiziminin gösterilimi

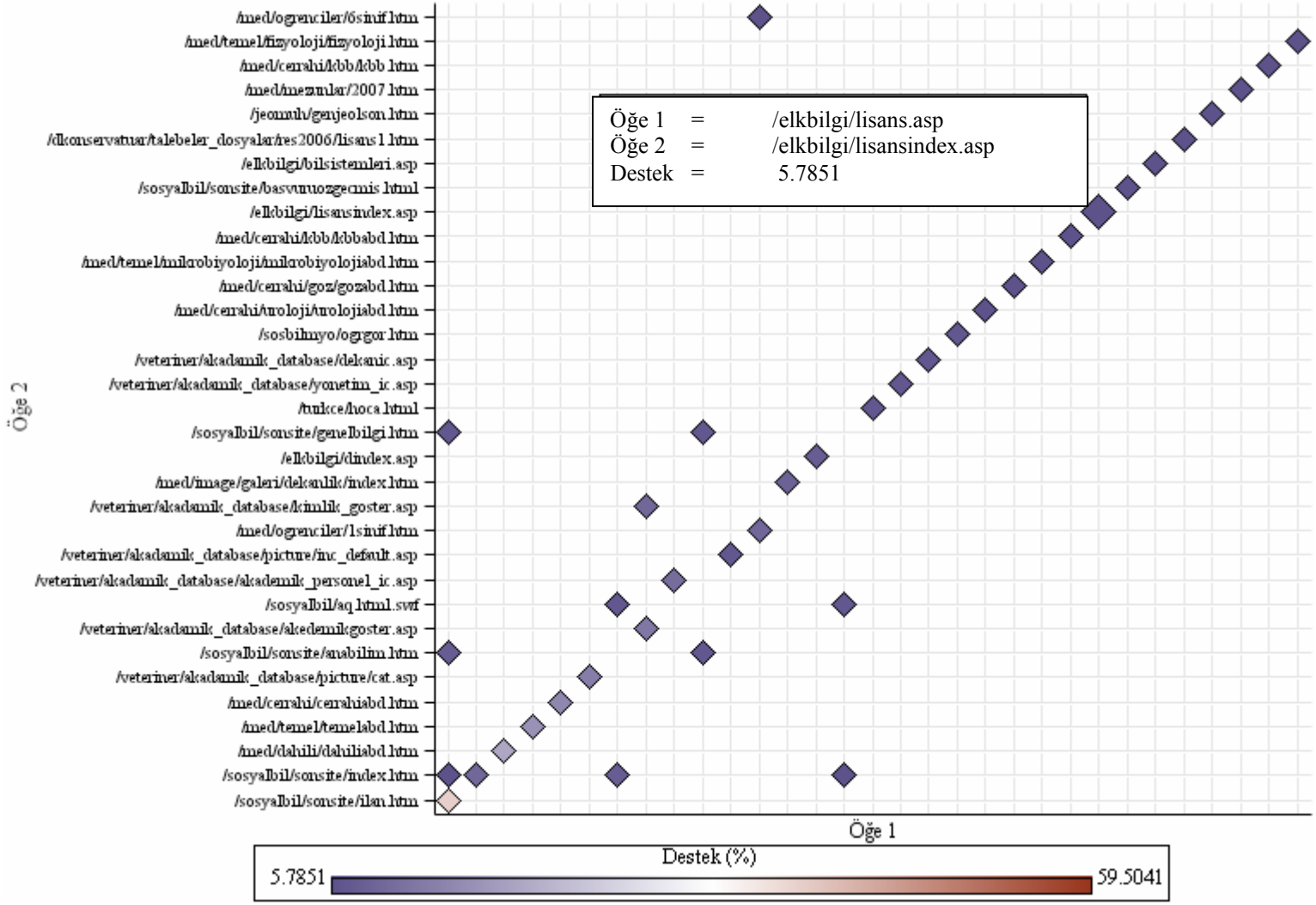
5.4.4 Öğelerin Raporlanması ve Çizimi

Öğelerin raporlanması işleminde, erişim kütüklerinde yer alan her web sayfasına ait işlem sayısını vermektedir. Tablo 5.6, en yüksek destek değerine ve işlem sayısına sahip olan öğelerin listesini göstermektedir. Dikkat edilirse en yüksek işlem sayısına sahip olan öğe */eemuh/default.asp* dosyasıdır. Yani, analiz edilen kütükler içerisinde bulunan kayıtlarda kullanıcının en çok gezindiği web sayfasıdır.

Tablo 5.6. Öğelerin raporlanması

Hedef Dosya	İşlem Sayısı	Destek Değeri (%)
/eemuh/default.asp	61	50
/eemuh/haber_detay_ic.asp	27	22.1311
/elkegitimi/ilanlar.asp	22	18.0328
/iletisim/default.asp	19	15.5738
/elkbilgi/lisansindex.asp	18	14.7541
/eemuh/foto/igallery.asp	18	14.7541
/fenbilimleri/duyurular.htm	16	13.1148
/cakaro/index.htm	15	12.2951
/dkonservatuar/talebeler.htm	13	10.6557
/eemuh/dersler_detay.asp	12	9.8361
/elkbilgi/dindex.asp	11	9.0164
/fenbilimleri/index.htm	10	8.1967
/fenbilimleri/genel.htm	10	8.1967

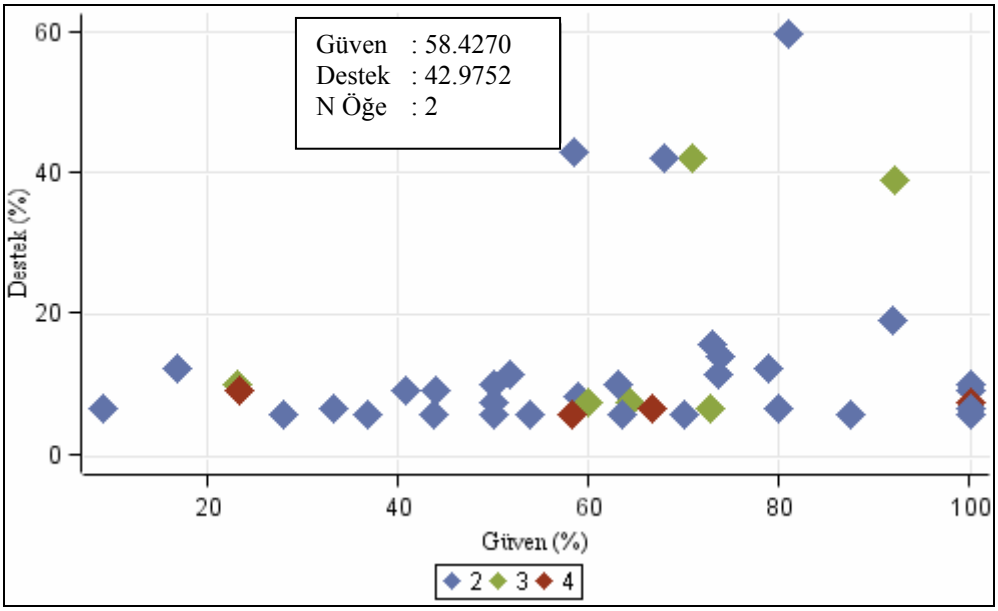
Öğelerin çizim grafiğinde, yol analizinde çıkarılan kurallara ait bilgilerin dağıtıksal olarak gösterilmektedir. Şekil 5.11'de görüldüğü gibi, seçilen öğenin üzerine fare ile tıklandığı zaman, çıkarılan kural ve destek değeri görülebilmektedir. Destek değeri büyük olan öğelere (web sayfaları) ait karoların renkleri de daha koyu olarak işaretlendi. Öğelerin gösterilimi grafiği analiz edildiğinde dikkat edilecek önemli nokta şudur; eğer öğelerin eksen üzerindeki yerleşimi doğrusal bir doğru gibi oluşturulmuşsa, yani öğeler dağıtık bir yapıya sahip değilse, kütük analizi yapılan web sitesinin o kıyasta gürültülü bir site olduğunu belirtmektedir. Bu vurgudan hareketle, Şekil 5.11'de aynı yapı görüldüğünden Fırat Üniversitesi web sitesinin çok gürültülü bir yapıda olduğu söylenebilir. Sitenin yeniden tasarımında, bu problemler dikkate alınarak site geliştirilmelidir.



Şekil 5.11. Öğelerin şekilsel gösterilimi

5.4.5 İstatistiksel Çizim

Öğelerin istatistiksel çiziminde, web sayfalarının yol analizlerinden elde edilen kuralların güven ve destek değerleri ile ilişkilendirilmiş dağılımsal gösterilimi belirtilmektedir. Şekil 5.12'de öğelere ait istatistiksel çizim grafiği gösterilmektedir. Şekil üzerindeki her karo yol analizindeki bir kuralı belirtmektedir. Yani, önce gelen ve sonucu olan iki öğe ile ilişkili bir kuralı belirtmektedir. Seçilen karo üzerinde ise o kurala ait güven ve destek değerleri görülebilmektedir. Mavi renkteki karolara ait zincir sayısı (köprü) iki, yeşil renkli karolara ait zincir sayısı üç ve kahve renkli karolara zincir sayısı dört olarak görülmektedir. Bu grafikteki her bir karo, güven ve destek değerleri ile tanımlanan bir yol analizi kuralını göstermektedir.



Şekil 5.12. İstatistiksel çizim

5.4.6 Kural İstatistikleri

Kullanıcı erişim kütüklerinden elde edilen kural istatistikleri tablosu, web sitesinde gerçekleşen işlemlere ait maksimum, minimum ve ortalama değerlerini sunmaktadır. Yapılan uygulama sonucunda elde edilen bulguların listesi Tablo 5.7'de gösterilmektedir.

Tablo 5.7. Kural istatistikleri

Etiketler	Minimum	Maksimum	Ortalama
Zincir Sayısı	2	4	2.40
Sayma	7.00	72.00	14.47
Destek	5.79 (%)	59.50 (%)	11.96 (%)
Güven	8.99 (%)	100.00 (%)	68.26 (%)

Kural istatistikleri tablosunda elde edilen bulgular incelendiğinde, yol analizi yöntemi ile en az iki, en fazla dört zincir sayısı olduğu görüldü. Çıkarılan diğer sayısal bilgiler Tablo 5.7'de gösterilmektedir. Ayrıca, zincir sayılarına göre tekrarlama sıklıklarına ait istatistiki bilgiler Tablo 5.8'de gösterilmektedir.

Tablo 5.8. Zincir sayılarına göre tekrarlama sıklıkları

Zincir Sayısı	Tekrarlama Sıklığı	Yüzde Oranı (%)	Toplam Tekrarlama	Toplam Yüzde(%)
2	39	70.91	39	70.91
3	10	18.18	49	89.09
4	6	10.91	55	100

5.5 Uygulama ve Bulguların Değerlendirilmesi

Web kullanıcı erişim kütük dosyalarının analiz edilmesi sonucunda, web sitesinin yapısı ve web ziyaretçilerinin davranışları hakkında önemli bilgi toplamak mümkündür. Bu çalışmada gerçekleştirilen yol analizi (iz analizi) uygulamasında bilimsel hedefe katkı sağlamak için yoğun teorik algoritmalar kullanılmış ve bu doğrultuda bir yöntem geliştirilmiştir. Geliştirilen yöntem Fırat Üniversitesi web sunucularındaki metin tabanlı kullanıcı erişim kütük dosyalarına uygulandı. Uygulama çalışmasında;

- Büyük boyuttaki yapısal olmayan web erişim kütüklerinin temizlenmesi ve gereksiz verilerden arındırılması için SAS yazılım ortamının kullanılarak geliştirilen bir ön işlem süreci önerildi. Önerilen süreçte, yüksek boyuttaki verilerle çalışmak için yüksek yapılandırmaya sahip bilgisayarlara ihtiyaç duyuldu. Ancak, aynı kütükler literatürdeki diğer temizleme yöntemlerinde ve farklı yazılım ortamlarında yapıldığında sistemin daha yavaş çalıştığı hatta zaman zaman sistemde sorunlar yaşandığı görüldü. SAS yazılım ortamını ile kendisine has veritabanı arasındaki güçlü ve doğrudan bir bağlantı bulunduğundan, ön işlem sürecindeki uygulamalarda hız ve performans açısından başarılı sonuçlar elde edildi. Ayrıca, SAS programlama kodlarındaki hazır fonksiyonların sağladığı kolaylıklar ile uygulamalara ve kullanıcıya büyük avantajlar sağlamaktadır.
- Ön işlem aşamasından geçirilmiş verilerden uygun alanlar ilişkilendirilerek, anlamlı ve ilginç bilgiler elde etmek için yol analizi yöntemi kullanıldı.

Önerilen ve gerçekleştirilen yol analizi uygulaması sonucunda elde edilen bulgular ve sonuçların değerlendirilmesinde web sitesine ilişkin aşağıda bilgiler çıkarıldı.

- Ziyaretçilerin web sitesinde gezinirken izledikleri yollar (web sayfaları) çıkarılarak, ziyaretçinin site içerisinde ilgi duydukları sayfalar tespit edildi. Site kullanımı için

benzer özellikleri taşıyan kullanıcıları kümelemek ve her bir kümedeki kullanıcılara uygun hizmet vermek mümkündür. Örneğin, katalog tarama hizmetini kullanan ziyaretçiler ile üniversite tanıtım bilgilerine göz atan kullanıcılar aynı nitelikte değildir. Birisi sadece ziyaret maksadıyla siteye uğramışken diğeri gereksinim duyduğu bilgiyi almak için siteye girmiştir.

- Web site içerisindeki alt klasörlere ya da alt bölümlere yerleştirilmiş sayfalarda gezinen ziyaretçi yoğunlukları tespit edildi. Tespit edilen bu sayfaların konumuna yönelik sitenin yapısında, yeniden iyileştirme çalışmaları yapılabilir. Yani, alt kısımlarda yoğun ziyaret edilen sayfalara site içerisinde hızlı erişim köprüleri konulabilir. Bu durum, sitenin iyileştirilebilmesi ve geliştirilebilmesi için web tasarımcılarına ve web yöneticilerine destek sağlayacaktır.
- Ziyaretçiler tarafından izlenilen yolları şekilsel olarak modellemek için web köprü bağlantı grafiği oluşturuldu. Ziyaretçi yoğunluğuna göre ağırlıklandırılmış köprü bağlantıları arasındaki çizgilerin kalınlığı o köprüler arasındaki geçişin yoğunluğuna göre değişim göstermektedir. Bu grafik ile sitedeki köprü bağlantıları dikkate alınarak web sitesinin yapısında iyileştirmeler yapılabilir.
- Ziyaretçilerin izlediği yollara ait tespit edilen güven ve destek değerleri dikkate alınarak yoğun gezinen sayfalardaki kullanıcı ilgileri çıkarılabilir. Örneğin, ticari web sitelerinde alışveriş yapan ziyaretçilerin ilgi duyduğu ürünlere göre sınıflandırma yapılabilir. Firma yeni ürünlerin satışına yönelik müşteri takibi yaparak, satışları arttırabilir.
- Web sitesinde en çok işlem gören ve gezinen web sayfaları bilgisi çıkarıldı. Bu bilgiler üniversite web sitesi içerisindeki bölümlere ait alt siteler arasında bir kıyaslama yapılmasına imkân vermektedir. Ziyaretçilerin sitelere olan ilgilerinin bulunması web sitesinin kalitesini arttırma da önemli bir ölçütü daha ortaya çıkarmış olacaktır.
- Yol analizi yöntemi ile yapılan web kütük uygulamasında elde edilen kural istatistikleri web sitesinin kullanımına yönelik genel istatistikî sonuçlar çıkarıldı. Bu sonuç bilgileri web sitesindeki köprü bağlantılarına ilişkin çıkarılan kuralların güven ve destek değerlerinin maksimum ve minimum sayıları, sayma sayıları ve zincir sayılarını vermektedir. Bu sonuçlar, analiz yapan kişiye web sitesinin genel kullanımına yönelik bir fikir oluşturacaktır.

Bu uygulama sonucunda, Fırat Üniversitesi web sunucularından alınan web kullanıcı erişim kütüklerinin analizi ile elde edilen bilgilerin kullanılabilirliği, hem web sitesi hem de kullanıcılar açısından ortaya konuldu. Web sitesindeki köprü bağlantıları incelenerek, sitenin

geliştirilebilmesi ve iyileştirilebilmesi için web tasarımcılarına ve web yöneticilerine destek sağlanılabileceği ortaya konuldu.

Web kullanım madenciliği uygulamaları sonucunda elde edilen bilgiler ile web kullanımına ilişkin derinlemesine analiz yapmak ve web kullanıcılarına daha iyi hizmet sunabilmek için kullanılmaktadır. Web kullanım madenciliği için birçok farklı yöntemler kullanılarak elektronik ticaret, biyoenformatik, bilgisayar güvenliği, web zekâsı, akıllı öğrenme, veritabanı sistemleri, finans, pazarlama, sağlık hizmetleri ve telekomünikasyon gibi birçok alanda ihtiyaçlar doğrultusunda daha fazla detaylı araştırmalar yapılabilir.

6. BİR WEB SİTESİNİN BAŞARIMINI ARTIRMAK İÇİN İSTATİSTİKSEL ANALİZ İLE WEB KULLANICI ERİŞİM KÜTÜKLERİNDEN BİLGİ ÇIKARIMI

Internet, elektronik ortamdaki verilerin uzaktaki mesafelere aktarılması ve kullanımının sağlanması açısından dünya üzerinde var olan en büyük bilgi paylaşım ortamıdır. Birçok kişi, kurum ve kuruluşlar gelişmiş Web sunucular kullanarak, bilgi paylaşımlarını ve veri aktarımlarını internet ağ sistemleri üzerinden yapmaktadırlar. İnternet kullanıcılarının, bu bilgi alışveriş işlemlerindeki tüm davranışları, Web sunucular tarafından kaydedilmektedir. Web kullanıcı erişim kütük dosyalarında kaydedilen bu veriler içerisinde, kullanıcı, sistem ve Web sitesi hakkında birçok önemli bilgiler bulunmaktadır.

Literatürde, Web kullanıcı erişim kütük verilerinin analiz edilerek, istatistiksel verilerin elde edilmesine yönelik çalışmalar bulunmaktadır [3, 15, 19-20, 25-42, 48-50, 55, 105, 126, 130-132]. Bu çalışmalarda özellikle bir Web sitesinin iyileştirilmesine, geliştirilmesine ve başarımın arttırılmasına yönelik yapılmış çalışmalar olduğu görülmüştür [3, 48-50, 52, 130-132].

Bu bölümde, Web kullanıcı erişim kütük verileri kullanılarak *istatistiksel analiz* yöntemi ile Web sitesinin kullanımına ve ziyaretçilerin davranışlarına yönelik anlamlı ve önemli bilgiler elde edildi. Web madenciliğinin etkin kullanımı ile bir Web sitesinin başarımını artırmaya yönelik çözüm önerileri oluşturuldu. Bu amaçla, Fırat Üniversitesi Web sunucularında tutulan Web kullanıcı erişim kütük dosyaları alınarak analiz edildi. Söz konusu erişim kütük dosyalarından Web sitesinin yapısı, kullanıcının Web sitesine erişimi ve kullanımına ilişkin bilgiler, kullanıcının kullandığı sistemine ait önemli bilgiler istatistiksel analiz yöntemi ile çıkarıldı. Ayrıca, Web sitesi başarımını arttırmak için HTTP durum kodları incelenerek, erişim kütüklerinden çıkarılan hata kayıtları analiz edildi. Yapılan bu analiz çalışmasından hareketle, Web sitesinde var olan kırık bağlantılar, engellenmiş sayfalar, kısıtlı erişimler, sunucu ve Web sitesi kaynaklı problemler tespit edilerek Web sitesinin geliştirilmesi, yeniden tasarlanması ve Web sisteminin başarımının arttırılması için çözüm önerileri oluşturuldu.

6.1 Problemin Tanımlanması

Web, 1989 yılında Avrupa Parçacık Fiziği Laboratuar'ından (European Particle Physics Laboratory-CERN) Tim Berners-Lee tarafından geliştirilmiştir. Web, internet üzerindeki sayısal nitelikteki kaynaklara ulaşabilmeyi sağlayan bir sunucu-istemci sistemidir. İstemci, kullanıcının sunucuya isteğini gönderdiği ve sonuçları gördüğü makinelerdir. Yani, istemciler Web

ziyaretçileridir. Web ziyaretçisi, internet ortamındaki HTTP isteğini kullandığı tarayıcı program aracılığıyla sunucuya gönderir. Sunucular ise istemcilerin HTTP yoluyla gönderdiği istekleri işleyen ve bu isteklere cevap veren yüksek donanımsal özelliklere sahip makinelerdir [133].

Bir Web sitesinin ve sunucu çalışma başarımının düşük olmasına neden olan birçok önemli etken bulunmaktadır. Bu temel unsurları aşağıdaki şekilde sıralanabilir:

- Web sayfalarındaki kırık köprülerin bulunması ya da yanlış köprü bağlantısından dolayı Web sayfalarının açılmaması
- Web sitesi ziyaretçileri için erişimlerin kısıtlanması ya da engellenmesi
- Ziyaretçi istekleri sonucunda ulaşılamayan veri tabanları ve sorgu ara yüzlerinin olması
- Web sitesi program kodlamalarında uzun kodlamalar, yanlış ve bilinçsiz yapılan yanlış sorgulamaların bulunması
- Web sitesindeki önemli olmayan, gereksiz ve ilgisiz bilgi yığınlarının olması
- Web sayfalarına yüksek boyutta resim ve müzik dosyalarının eklenmesi
- Web sitesindeki sayfalar arasında kodlama uyumsuzlukların olması
- Web sayfalarındaki dokümanlarda kullanılan dosyalar biçimlerinin uygun seçilmemesi

Son yıllarda Web madenciliği konusu araştırma konuları arasında Web site başarımının artırılması, site tasarımının geliştirilmesi ve iyileştirilmesi, Web sitelerinin etkin yönetimini gibi güncel ve vazgeçilmez çalışmalar bulunmaktadır. Son zamanlarda çeşitli organizasyonlar (ISO - Uluslararası Standartlar Organizasyonu, W3C) tarafından internet veri erişimi ve yönetimi ile ilgili çeşitli standartlar geliştirilmiştir. Bu standartlar, modelleri özelleştirilmiş internet programlama dillerini ve özel yapıları içermektedir. XML (Genişletilebilir Belirleme Dili) bu programlama dillerinden bir tanesidir.

6.2 HTTP - Köprü Aktarım Protokolünün İncelenmesi

HTTP (*Hyper Text Transfer Protocol – Köprü Aktarım Protokolü*), Internet üzerinde veri aktarmaya yarayan, dağıtık üst ortam bilgi sistemleri için kullanılan OSI (*Open System Interconnection*) referans modelinin uygulama katmanında yer alan bir iletişim protokolüdür. Sunucu ve istemciler arasındaki köprü metinlerinin iletişimini sağlamaktadır. İnternet'teki bir sunucuda bulunan ve metin, resim, ses ya da diğer sayısal bilgileri içeren Web sayfalarına ulaşmayı sağlar. Ayrıca, üst metin uygulamaları dışında da birçok iş için kullanılabilecek genel bir protokoldür. HTTP iletişimi, TCP/IP (*Transmission Control Protocol / Internet Protocol – Aktarım Kontrol Protokol / Internet Protokolü*) üzerinden yapılmaktadır. HTTP'ye ait önemli özellikleri şu şekilde sıralayabiliriz:

- HTTP, kullanıcıların sunuculara hangi iletileri gönderebilecekleri ve karşılığında hangi cevapları alacaklarını belirler. Tarayıcıya bir URL girilmesi, aslında sunucuya gönderilen bir HTTP komutudur. Bu komutla, istenen Web sayfası sunucu tarafından bulunup kullanıcıya iletilir [133].
- HTTP, bir istek yanıt protokolüdür. HTTP bağlantıları, genellikle kullanıcı aracıları tarafından başlatılır ve sunucu üzerindeki bir kaynağa uygulanacak istekten oluşur. HTTP isteği, GET ya da POST şeklindedir. Sunucu, her dosya için bir GET isteğinde bulunur. Web tarayıcı programlarının (Internet Explorer, Netscape Navigator, Mozilla, v.b.) GET isteğinin yanında, bağlantı hakkında başka bilgiler içeren HTTP üstbilgilerini de iletir. Vekil sunucular, geçit yolları ve tüneller söz konusu olduğunda, bu istek-yanıt süreci karmaşık bir hal alabilir. Vekil sunucu, bir URI'nın (Uniform Resource Identifier) mutlak biçimdeki isteklerini alır, iletinin tümünü ya da bir kısmını yeniden yazar ve yeniden biçimlendirilmiş isteği URI tarafından tanımlanmış sunucuya iletir. Bu yüzden vekil sunuculara *iletim aracısı* denilmektedir. Geçit yolu, diğer sunucular üzerinde bir katman işlevi görür ve istekleri asıl sunucunun protokolüne çevirir. Geçit yollarına da *alım aracısı* denilebilir. Tüneller ise iki bağlantı arasında iletileri değiştirmeden aktarım noktası olarak işlev görür. Yani, güvenlik duvarlarının üzerinden, iletişimin yapılması gerektiğinde tüneller kullanılır.
- HTTP, kullanıcı aracıları ile SMTP, NNTP, FTP diğer internet sistemlerine geçit yolları içeren sistemler arasında da kullanılır.
- HTTP, ön koşulsuz bir protokoldür. Çünkü her komut kendinden önceki komuta bakmaksızın birbirinden bağımsız olarak çalışır. Bu, Web sitelerinin kullanıcılara akıllı tepkiler verememesinin en önemli sebebidir. HTTP'deki bu eksiklik, ActiveX, Java ve Java Script gibi teknolojilerle giderilmeye çalışılmıştır.
- HTTP, tek başına güvenli bir protokol değildir. Bu nedenle SHTTP *(Secure Hypertext Transfer Protocol)* ve HTTPS *(Hypertext Transfer Protocol Secure)* adlı, verilerin şifrelenmesi temeline dayanan daha güvenli iki biçimi daha vardır [132].

6.2.1 HTTP Kütükleri

Kütük dosyaları, Web sunucular üzerinde bulunur ve sunucuya gelen istekleri gerçek-zamanlı olarak kaydeder. Her bir istek, bir kütük dosyasında tek bir satır olarak yer alır. Genellikle dört farklı şekilde HTTP kütükleri tutulmaktadır:

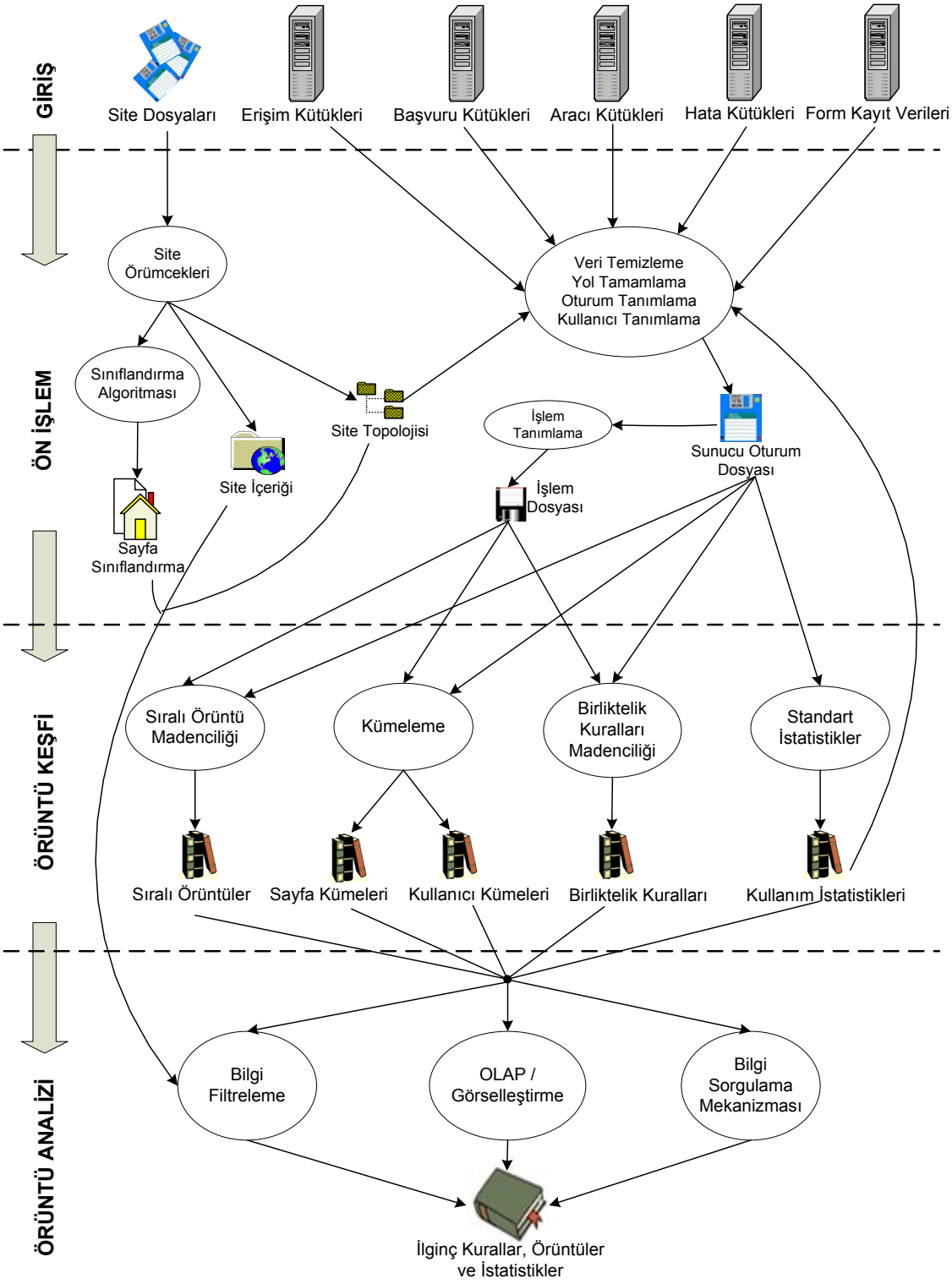
***Erişim Kütükleri*:** Web sunucusu tarafından işlem gören tüm istek ve yanıtlar hakkında bilgileri kaydeder. Internet kullanıcı davranışlarının saklandığı kütük dosyalarıdır.

***Başvuru Kütükleri*:** Kullanıcı isteklerinin tutulduğu kütük dosyalarıdır. Örneğin; isteğin nereden geldiğini ve sunucuya hangi sitelerin Web trafiğini yönelttiğini kaydeder (Örneğin: başka bir sayfadan bağlantı, bir taramanın sonucu v.b.).

***Aracı Kütükleri*:** Web sunucusundan istekte bulunan Web istemcilerinin kaydını tutan aracı kütüğüdür. Web istemcileri istek iletileriyle birlikte bir kullanıcı etken kodu gönderirler. Bu kod genellikle tarayıcı üreticisi tarafından belirlenmekte ve tarayıcının gerçek tipi hakkında bilgi içermektedir. Yani, Internet tarayıcısının adı, sürümü ve işletim sistemi hakkındaki bilgilerin tutulduğu kayıt dosyalarıdır.

***Hata Kütükleri*:** Sunucu üzerinde gerçekleştirilemeyen başarısız istekleri ya da meydana gelen hatalı işlemlerin kaydedildiği kütük dosyalarıdır. Örneğin bir istemci, sunucuda olmayan bir dosyaya erişmek isterken karşılaştığı kırık bağlantı durumu, sunucu hata kütüğüne bir hata mesajını otomatik olarak oluşturmaktadır.

Web sitesinin başarımlarının değerlendirilmesi, Web kullanıcı erişim kütüklerinin analizi ile mümkündür. Bu analiz sonucunda kapsamlı olarak bilgi çıkarılması, Web sunucularının otomatik olarak kaydettiği erişim kütük dosyalarının biçimlerine de bağlıdır. Bu nedenle, bu çalışmada kapsamlı ve çok yönlü bilgi elde edebilmek için ECLF biçimindeki Web kullanıcı erişim kütükleri tercih edildi. Kütük dosyalarından anlamlı ilginç kurallar, bilgiler ve istatistikler elde etmek için Şekil 6.1'de gösterilen yapı kullanıldı.



Şekil 6.1. Kütük dosyalarından bilgi çıkarımı

Bir istemci, Web sitesine herhangi bir istekte bulunduğu zaman, sunucu tarafından bu isteğe HTTP durum kodları ile cevap verilmektedir. Durum kodlarına bakılarak, yapılan isteğin amacı ile ilgili bilgi çıkarılabilmektedir. HTTP durum kodları 3 haneden oluşan sayısal değerlerdir. Bu değerlerin ilk rakamı bir ile beş (1-5) arasındaki olup, yapılan isteğe verilen cevabın sınıfını belirtmektedir. Son iki rakamın hiçbir sınıflandırma değeri yoktur. HTTP durum kodlarının sınıfları aşağıdaki gibi tanımlanmaktadır:

1xx (Bilgilendirici): Sunucuya istek geldiği zaman, sunucu tarafından yapılan işlemin devam ettiğini belirten geçici bir bilgilendirici yanıtı verildiğini gösteren durum kodlarıdır.

2xx (Başarı): İstemci tarafından yapılan isteğin sunucu tarafından başarılı bir şekilde alındığını, anlaşıldığını ve kabul edildiğini belirten durum kodlarıdır.

3xx (Yönlendirme): İstemci tarafından yapılan isteği tamamlamak için ek işlemlerin yapılması gerektiğini belirten durum kodlarıdır. Bu kodlar genellikle yeniden yönlendirme amacıyla kullanılmaktadır.

4xx (İstemci Hatası): İstemci tarafından yapılan isteğin hatalı olduğunu, yapılan isteğin engellendiği ya da yapılan işlemin tamamlanamadığını belirten durum kodlarıdır.

5xx (Sunucu Hatası): İstemci tarafından yapılmış tamamen geçerli olan bir isteğin işleme konulurken sunucudan kaynaklanan bir hatadan dolayı gerçekleştirilmediğini belirten durum kodlarıdır.

Uygulama sonuçlarının değerlendirilmesi için Tablo 6.1'de HTTP durum kodlarından bazıları verilmektedir. Yapılan uygulama sonucunda kütük dosyalarından çıkarılan hata kodları ve verilere bakılarak problemin çözümüne yönelik sistem analizi yapılmıştır. Çıkarılan HTTP durum kodlarının analiz edilmesinde ve yorumlanmasında, Tablo 6.1 önemli bir referans kaynağıdır.

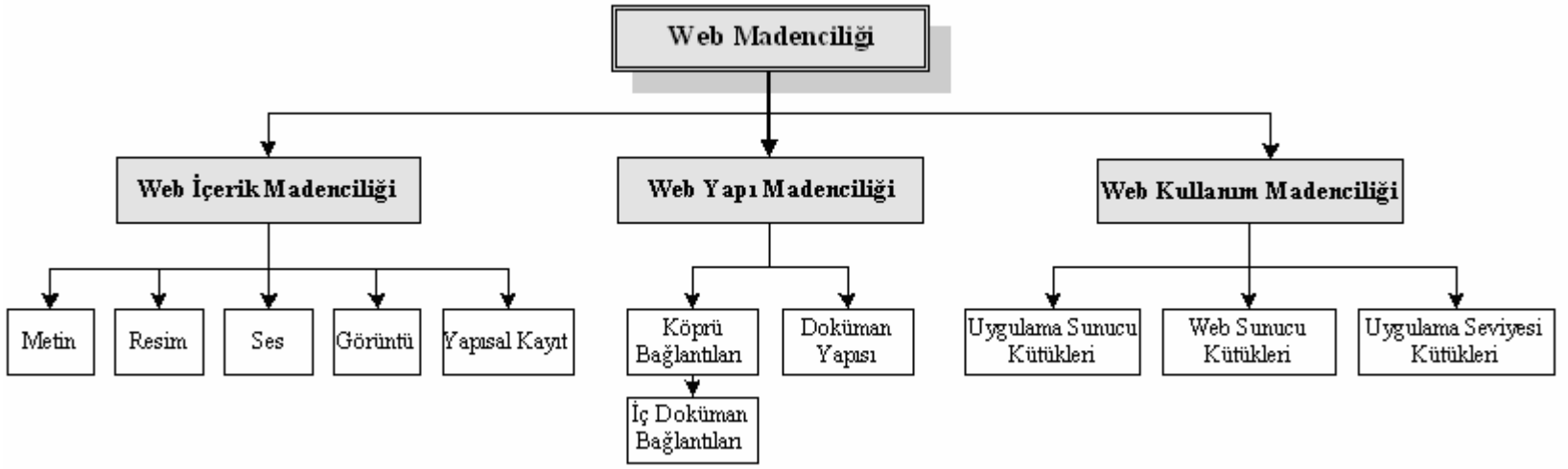
Tablo 6.1. HTTP Durum kodlarından bazıları

Kod	Durumun Sebebi	Açıklama
200	Başarılı	Sunucu, yapılan isteği, başarılı olarak işleme koymuştur.
202	Kabul edildi	Sunucu, yapılan isteği kabul etmiştir. Ancak, henüz işleme koymamıştır.
302	Bulunamadı	İstenen belge yeni bir yere taşınmış durumdadır.
304	Değiştirilmedi	İstenen sayfa, yapılan son istekten sonra değiştirilmemiştir.
400	Geçersiz istek	İstek, hatalı sözdizimi nedeniyle sunucu tarafından anlaşılamamıştır.
401	Kimlik doğrulama hatası	İstek yapılan sayfa, kullanıcıdan kimlik doğrulaması bekliyor, fakat istemci geçerli bir kullanıcı adı veya şifre sağlayamamıştır.
403	Yasak	Sunucu, yapılan isteği anladığı halde işlemi tamamlamayı reddetmektedir. Yani, istenen belgeye erişim yasaklanmıştır.
404	Bulunamadı	İstenen belge sunucu tarafından bulunamamaktadır. Adres yanlış olabilir ya da belge taşınmış olabilir. Ayrıca bu kod erişim izni olmayan istemcilere belgenin olmadığı mesajını vererek belgeyi korumak amacıyla da kullanılabilir.
405	Yönteme izin yok	İstekte belirtilen yönteme sunucu tarafından izin verilmemektedir.
406	Kabul edilemez	İstek yapılan sayfa, istenen içerik özellikleriyle yanıt verememektedir.
407	Vekil kimlik doğrulaması	401 koduna benzerdir. İstemcinin Vekil kullanarak kimlik doğrulaması yapmasını belirtir.
408	İstek zaman aşımı	Sunucu, yapılan isteği beklerken zaman aşımına uğradığını belirtir.
409	Çakışma	Sunucu, isteği gerçekleştirirken bir çakışmayla karşılaşmıştır.
411	Uzunluk gerekli	Sunucu, geçerli bir (içerik uzunluğu) başlık alanı olmadan isteği kabul etmez.
412	Önkoşul hatası	Sunucu, istemcinin belirttiği ön koşullardan birini karşılayamamaktadır.
413	İstek varlığı çok büyük	Sunucu, işleyemeyeceği kadar büyük veri olduğu için isteği işleyemiyor.
414	İstenen URI çok uzun	İstenen URI, sunucunun işleyemeyeceği kadar büyüktür.
415	Desteklenmeyen ortam türü	İstek, istenen sayfa tarafından desteklenmeyen bir biçime sahiptir.
417	Beklenti karşılanmadı	Sunucu, istek başlığı alanının gereksinimlerini yerine getiremiyor.
500	Dâhili sunucu hatası	Sunucu bir hatayla karşılaştı ve isteği tamamlayamadı. Bu durumda sunucu yöneticisi hata kütüğünü kontrol ederek hatanın neden kaynaklandığına bakmalıdır.
501	Uygulanmadı	Sunucu, istemci isteğini gerçekleştirecek işlevselliğe sahip değildir.
502	Bozuk ağ geçidi	Sunucu ağ geçidi veya Vekil gibi davranırken akış sunucusundan geçersiz bir yanıt aldığını belirtir.
503	Hizmet verilemiyor	Sunucu, aşırı yüklü veya onarım için servis kapalı olduğundan hizmet verilmemektedir. Bu genellikle geçici bir durumdur.
504	Ağ geçidi zaman aşımı	Sunucu ağ geçidi veya Vekil gibi davranırken akış sunucusundan zamanında istek alamadığını belirtir.
505	HTTP sürüm desteklenmiyor	Sunucu, istekte kullanılan HTTP protokol sürümünün desteklemediğini gösterir.

6.3 Web Kullanım Madenciliği Uygulaması

Web erişim kütüklerinden ziyaretçi davranışları hakkında bilgi kazanmak için standart istatistiksel teknikler kullanılmaktadır. Web kütük analizi için araştırmacı gruplar tarafından geliştirilen ticari yazılımlar kullanılmaktadır [6]. Bu yazılımlar, araştırmacıların ilgi alanlarına göre değişim göstermektedir. Yani, çalışma grupları Web madenciliğinde ihtiyaç duyulan yaklaşımları dikkate alarak yazılım üretmektedirler. Şekil 6.2'de görüldüğü gibi Web madenciliği çalışma alanlarına göre kullanılan veri sınıfları da farklılık göstermektedir. Bundan nedenle Web madenciliği alanında geliştirilen yazılım uygulamaları hem veri kümesi hem de yapı itibariyle birbirlerinden farklı olduğundan kullanım ve uygulama alanları olarak da farklılık göstermektedir. Ancak, genelde bütün uygulamaların geliştirilmesinde ise kullanılan örüntü keşfi yöntemleri veri madenciliği teknikleridir. Örneğin, elektronik ticaret yapan Web sitelerinin değerlendirilmesi ve analizi için geliştirilen Web madenciliği yazılımlarının odaklandığı nokta, daha çok ticari firmaların satışını arttırmaya yöneliktir. Geliştirilecek Web madenciliği yazılımı tüm alanları kullanabilir ama daha çok site içeriğine yönelik çalışmalar yapılması gerekir. Ancak, akademik Web sitelerinin değerlendirilmesi için geliştirilen uygulamalarda ise site kullanıcılarının memnuniyeti, uzaktan eğitim yapan Web materyallerinin geçerliliği, uygunluğu ve değerlendirilmesi, öğrenci davranışlarının değerlendirilmesi gibi konuları araştırmaya yöneliktir. Yani, yapılan analiz çalışmasında Web sunucu üzerinde tutulan kullanıcı erişim kütüklerinin kullanılması gerekmektedir. İstatistik ve yapay zekâ gibi matematik temelli bilimlerle ilgilenen araştırmacılar ise konuya daha teorik açıdan yaklaşarak uygulamalarda kısmen kullanılabilecek yaklaşımlar geliştirmektedirler.

Web sitelerinin gerekliliği ve yaygın kullanılması, birçok araştırmacıyı Web madenciliği alanında çalışma yapmaya yönlendirmiştir. Son yıllarda bu alanda yapılmış birçok çalışma olduğu görülmektedir [10-26, 45-52]. Şekil 6.2'de görüldüğü gibi, Web içerik madenciliği alanında yapılan çalışmalarda Web sayfaları içerisinde yer alan metin, resim, görüntü ve yapısal kayıtlarla ilgilenmişlerdir. Web yapı madenciliği alanında yapılmış çalışmalarda ise, Web sitesinin yapısı söz konusu olduğu için Web site ve sayfalar içerindeki köprü bağlantıları, kullanılan doküman yapıları incelenmektedir. Web kullanım madenciliğinde ise, araştırmacıların kullandıkları veriler, sunucular tarafından kaydedilen herhangi bir anlam ifade etmeyen karmaşık yapıdaki metin tabanlı kütük veri yığınlarıdır. Bu uygulama çalışmasında kullanım verileri olarak Fırat Üniversitesi Web sunucusu tarafından saklanan Web kullanıcı erişim kütük verileri kullanıldı.



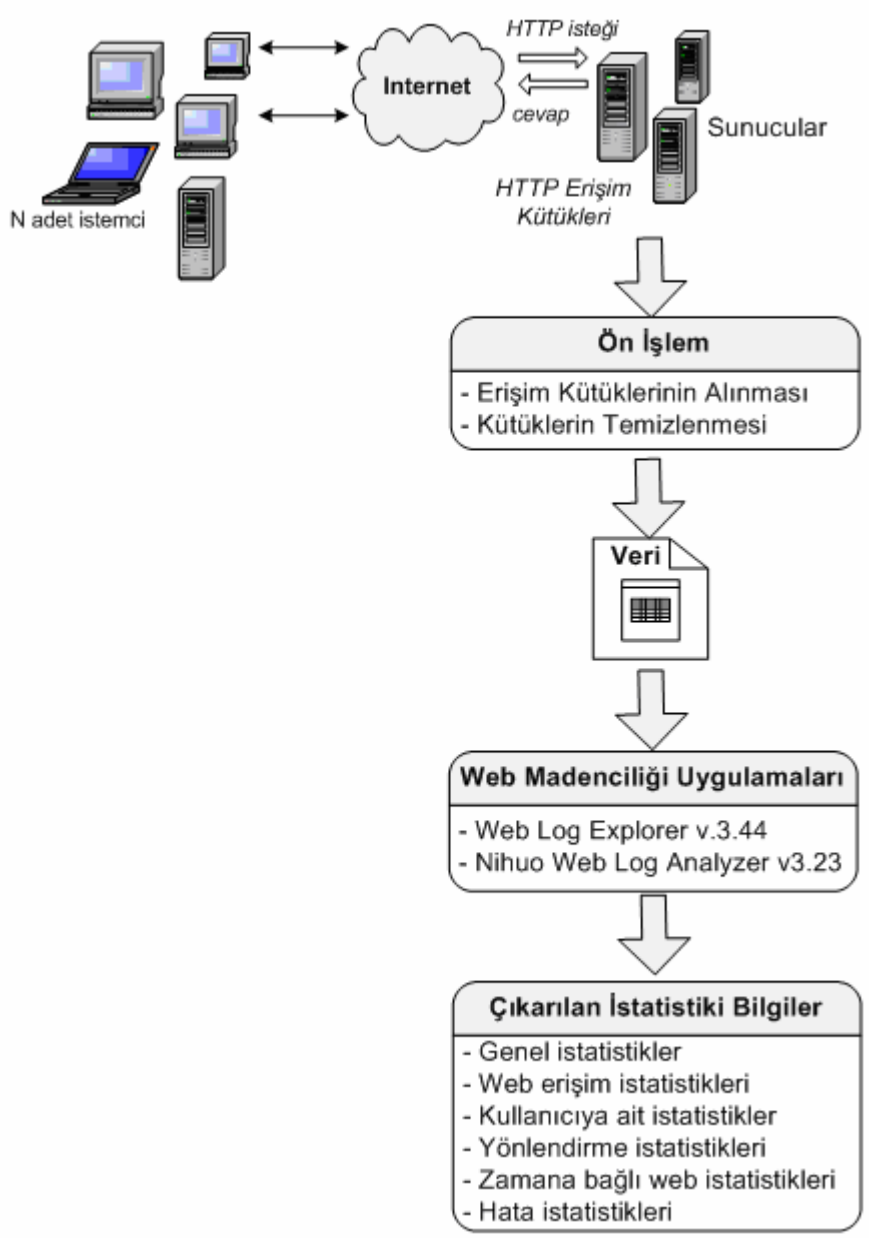
Şekil 6.2. Web madenciliğinde kullanılan verilerin sınıflandırılması

Web madenciliği uygulamaları için geliştirilen birçok farklı yazılım Tablo 6.2'de belirtilmiştir. Bu araştırma ve uygulama çalışmasında, Tablo 6.2'de belirtilen Web madenciliği yazılımlarının hepsi incelenerek, başarımları gözlemlendi. Çalışmada yüksek kapasitede kullanım kütükleri olduğu için Fırat Üniversitesi Web sitesine ait kullanıcı erişim kütüklerinin incelenmesi için başarım ve analiz açısından sadece Nihuo Web Log Analyzer (sürüm: 3.23) ve Web Log Explorer (sürüm:3.44) yazılımlarının sonuçları verildi.

Tablo 6.2. Uygulamada kullanılan yazılımlar

Yazılım	Lisans	Web Adresi	Üretici Firma
e-Web Log Analyzer v2.02	Deneme sürümü	esoftys.com	Esoftys Inc.
Funnel Web Analyzer	Deneme sürümü	quest.com	Quest Software
Megaputer Web Analytics	Deneme sürümü	megaputer.com	Megaputer Intelligence
NetIQ Web Trends Log Analyzer	Deneme sürümü	netiq.com	NetIQ Software
Nihuo Web Log Analyzer v3.23	Deneme sürümü	nihuo.com	Exacttrend Software
Web Log Explorer v3.44	Deneme sürümü	exacttrend.com	Exacttrend Software
WMS Log Analyzer	Deneme sürümü	exacttrend.com	Exacttrend Software
Absolute Log Analyzer Pro v2.38	Deneme sürümü	bitstrike.com	BitStrike Software
Web Log Mixer	Deneme sürümü	bitstrike.com	BitStrike Software
WebTrends, Analytics&Statistics	Deneme sürümü	Webtrends.com	WebTrends Inc.
Web Log Storming v1.5.1	Deneme sürümü	datalandsoftware.com	Dataland Software
123Log Storming	Deneme sürümü	123logstorming.com	ZY Computing Inc.
Web Log Expert v3.1	Deneme sürümü	Weblogexpert.com	Alentum Software
AlterWind Log Analyzer v2.1	Deneme sürümü	alterwind.com	AlterWind Software
Deep Log Analyzer	Deneme sürümü	deep-software.com	Deep Software
Awstat	Deneme sürümü	awstats.sourceforge.net	Awstat Software
Analog	Deneme sürümü	analog.cx	Analog Software
Webalizer	Deneme sürümü	mrunix.net/Webalizer	Webalizer Software

Web kullanım kütüklerinden, anlamlı bilgilerin çıkarılması için kullanılan birçok yazılım araçları, veri tabanları (SQL, Oracle, v.b.) ile ilişkilidirler. Uygulamamızda kullanılan veriler, SAS yazılımının kendi içerisinde yer alan veri tabanı ilişkilendirildi. Uygulama çalışmasına ait akış şeması Şekil 6.3'de gösterilmektedir. Uygulama çalışmasında, sitedeki benzer sayfalar gruplandırılarak tespit edilen Web ziyaretçi bilgileri hedef ziyaret gruplarını analiz edildi. Ayrıca, sitedeki sayfa yapısı, sayfa içeriği, kırık köprüler v.b. birçok bulgular ile site gelişimine büyük oranda katkı sağlamaktadır. Web sitesinin bu şekilde organize edilmesi ile Web kullanıcıların siteyi etkin ve kolay kullanımı ortaya konuldu.



Şekil 6.3. İstatistiksel analiz ile bilgi çıkarımı için gerçekleştirilen uygulamanın prensip şeması

Bu uygulama çalışmasının amacı, istatistiksel analiz ile Web kullanıcı erişim kütüklerinden çıkarılan anlamlı ve yararlı bilgilerden yola çıkarak, Web sunucu ve Web sitesinin başarımını sağlamaktır. Bunun sonucunda ortaya konulan çözüm önerileri ile Web sitesinin geliştirilmesi ve etkin bir biçimde yönetimini sağlamak için Web site yöneticisine ve Web site tasarımcılarına destek sağlamaktır. Bu durum beraberinde siteye olan istek sayısının artmasını sağlayacaktır. Bu şekildeki çalışmalar ticari kuruluşlara uygulandığı zaman da, şirketin müşteri potansiyelini büyük oranda arttıracaktır.

6.4 Web Kullanıcı Erişim Kütüklerinden Çıkarılan İstatistikî Bilgiler

Fırat Üniversitesi Bilgi İşlem Daire Başkanlığı bünyesinde hizmet veren Web sunucularından alınan genişletilmiş Web kullanıcı erişim kütük dosyaları 1 Şubat 2008 – 30 Nisan 2008 tarihleri arasındaki üç aylık bir dönemi içermektedir. Alınan kütük dosyaları 90 adet ve toplamda 5.42 GB boyutundadır. Alınan bu Web kullanıcı erişim kütükleri hazırlanan yazılım ile veri ambarı oluşturmak amacıyla bir dönüşüm işleminden geçirilerek Web kütük veritabanına aktarıldı. Bu aşamada seçim, temizleme ve dönüşüm işlemleri yerine getirilmiş oldu. Oturum tanımlamaları sayesinde oturum bazında analiz imkânı sağlandı. Kütükler içerisindeki anlamsız ve bozuk satırlar temizlendi. Bunun sonucunda geriye kalan temiz kayıtlar toplamda 17.544.714 adet satırdır. Internet kullanıcı Web erişim kütük dosyalarından çıkarılan diğer önemli ve yararlı istatistiksel bilgiler genel hatlarıyla tablo halinde verildi.

Önerilen ve gerçekleştirilen Web madenciliği uygulama sonucunda önemli ve anlamlı istatistiksel bilgiler tespit edildi. Tablo 6.2'de belirtilen Web madenciliği yazılımları kullanılarak, sunuculardan alınmış olan metin tabanlı erişim kütük dosyaları çevrimdışı olarak analiz edildi. Ancak, deneme sürümleri kullanılan bu yazılımların büyük çoğunluğu, boyutu büyük ve birden fazla kütük dosyalarının analizinde problem yaşandı. Bu nedenle, boyutu büyük olan erişim kütük dosyalardan istatistiksel bilgi çıkarılmasında Web Log Explorer ve Nihuo Web Log Analyzer yazılımları daha başarılı ve yüksek başarım gösterdiği görüldü.

6.4.1 Genel İstatistikler

Web sunucularından alınan Web kullanıcı erişim kütük dosyalarına ait tarih bilgileri, kullanıcı istekleri, istek yapılan sayfalar, Web sitesi ziyaretçileri ve talep edilen bant genişliğine ait genel bilgiler Tablo 6.3 de gösterilmektedir.

Tablo 6.3. Erişim kayıtlarına ait genel istatistikî bilgiler

Tarih	Verilerin Zaman Aralığı	1/Şubat/2008 00:00:03 – 30/Nisan/2008 20:56:53
	Raporlama Tarihi	22/5/2008 10:34:23 AM
	Toplam Gün Sayısı	90
Web İstekleri	Toplam İstem Sayısı	7.183.176
	Günlük Ortalama	78936
Sayfalar	Sayfa Görüntüleme Sayısı	6.996.259
	Günlük Ortalama	76.881
Ziyaretçiler	Toplam Ziyaretçiler	427.891
	Günlük Ziyaretçiler	4.702
	Farklı IP Adresleri Sayısı	126.109
Bant Genişliği	Toplam Bant Genişliği	181.47 GB
	Her gün için bant genişliği	1.99 GB

6.4.2 Web Erişim İstatistikleri

Bu raporlama bilgisinde ziyaretçilerin Web sitesindeki alt klasörlere kaç defa erişim isteğinde bulunduklarını, gezindikleri sayfa adetleri ve bu sayfalarda kullandıkları bant genişliği boyutlarını göstermektedir. Elde edilen erişim istatistikleri bilgisi çok büyük olduğundan Tablo 6.4'de gösterilen liste, raporlama bilgisinden alınan ilk satırları göstermektedir. Bu rapor çoğunluğu siteye bilinçsizce gelmiş, yönlendirilmiş ziyaretçilerin ve arama işlemlerinin bir kısmı da arama robotlarının işaretçisi ile geldiğini göstermektedir. Erişim istek sayısı yüksek olan klasörlerdeki ziyaretçilerin varlığı, site içeriğinin yoğun şekilde dolaştığını veya taradığını göstermektedir.

Tablo 6.4. Web erişim istatistikleri

Klasör Adı	İstek Sayısı	Ziyaretçi	Sayfa	Bant Genişliği
/iletisim/	2.976.517	5.025	6	1.069.484.176
/metal/	2.112.571	9.645	41	1.374.268.831
/	285.292	183.251	138	99.973.848.356
/eemuh/	98.312	42.568	34	5.443.098.167
/fenbilimleri/	69.667	9.748	27	1.561.765.162
/med/	58.680	10.756	14	172.844.068
/kimmuh/kimya/	48.869	31.132	45	1.278.666.932
/kimmuh/cv/	46.354	26.653	11	657.940.229
/ogrencikonseyi/2008/	46.047	11.453	71	3.061.247.584
/veteriner/	45.542	17.375	74	1.129.581.347
/eemuh/foto/	44.560	1.579	13	1.469.297.183
/sosbilmyo/	42.740	5.471	58	178.769.995
/elkegitimi/	40.348	5.674	37	597.568.899
/yturel/	37.245	9	4	5.273.311.312
/makinetef/	35.311	13.689	54	312.813.292
/makinemuh/	34.326	9.827	46	2.553.833.402

6.4.2.1 Erişilen Dosya Tipleri

Web erişim kütükleri derinlemesine analiz edildiğinde, Web sitesinden istekte bulunulan bütün dosya tipleri çıkarılabilir. Analiz sonucunda erişim yapılan dosya tipleri ve istatistiki bilgiler Tablo 6.5'de gösterilmektedir. Çıkarılan bilgiler incelendiğinde, analiz edilen toplam erişim istekleri içerisinde en çok istekte bulunulan *asp* dosya tipidir. Dinamik bir işleve sahip olan bu dosyalara istekte bulunan kullanıcılar için kullanım yoğunluklarına göre iyi, daha iyi veya en iyi şeklinde gruplandırma yapılabilir. İkinci olarak resim ve grafik dosyaları (*.jpg, .gif*) en çok istekte bulunulan dosya tipleridir. Bu durum ise, Web sitesi içerisinde üniversitenin akademik ve idari birimlerin tanıtımında çok sayıda resim kullanıldığını göstermektedir. *Htm* uzantılı dosyalar ise Web sayfalarında en çok istem alan üçüncü grup dosyadır. Bilgi amaçlı olan bu dosyalara istekte bulunan kullanıcılara potansiyel iyi kullanıcı gözüyle bakılabilir. Bilgi

amaçlı dosyalar eğer iyi hazırlanmışsa siteye iyi kullanıcı kazandırmakta önemli bir işleve sahip olabilir. Bunların yası sıra site içerisinde yer alan *.css, .js, .pdf, .png, .html, swf, .doc* gibi diğer dosyalarda yapılan isteklerde etkin olarak kullanıldığı gözlemlenmektedir. Erişim yapılan dosyalara ilişkin yapılan isteklerin detaylı bilgileri Tablo 6.5'de verilmektedir.

Tablo 6.5 Erişim yapılan dosya tipleri

Dosya Tipi	İstek Sayısı	Ziyaretçi Sayısı	Sayfa	Bant Genişliği
.asp	6.115.674	325.126	3.725	21.899.763.288
.jpg	4.814.438	256.218	10.202	124.275.296.635
.gif	3.925.560	243.264	6.753	15.781.288.192
.htm	767.020	171.140	7.383	12.470.647.152
.css	337.944	126.189	371	628.771.581
.js	258.911	120.067	363	2.864.941.343
.pdf	218.557	52.821	1.302	43.370.540.995
.png	218.307	52.828	404	2.691.298.397
.html	183.888	54.302	1.827	3.021.607.172
.swf	175.597	70.231	315	27.152.660.037
.doc	150.063	72.004	2.307	19.452.038.483
.ico	124.040	84.500	11	200.438.885
.rar	98.036	3.693	108	149.231.968.027
.xml	29.373	2.336	395	83.708.232
.bmp	28.779	12.768	86	2.964.183.466
.xls	27.125	19.148	119	1.714.684.856
.cur	12.554	9.361	8	29.923.972
.txt	9.312	4.581	15	27.691.228
.aspx	6.226	4.380	24	168.177.382
.ppt	6.012	3.611	49	3.628.485.935
.pps	5.983	677	79	3.587.372.826
.asf	5540	1392	37	49.214.339.641

Web sitesinde gezinen bazı kullanıcılar ise sitede olmayan *exe* veya *dll* uzantılı dosyalara istekte bulunmaktadır. Bu tip kullanıcıların amacı Web sunucusu üzerindeki açıklardan faydalanarak sistemi çalışmaz ve kullanılmaz hale getirmektir. Bu tip dosyalara istek yapan kullanıcılar tehlikeli veya kötü amaçlı kullanıcı olarak tanımlanabilir.

6.4.2.2 En Fazla Giriş ve Çıkış Yapılan Sayfa Bilgileri

Web sitesinin en fazla giriş yapılan Web sayfaları Tablo 6.6'da gösterilmektedir. Tablonun ilk sırasında yer alan */kimmuh/kimya/CV.htm* dosyası yapılan madencilik işleminde en fazla giriş yapılan adres olarak çıkarılmıştır. Sayfaya olan yoğun ilgi sonucunda ziyaretçi sayısı ve kullanılan bant genişliği de en yüksek düzeyde olmuştur.

Tablo 6.6. En fazla giriş yapılan sayfalar

Giriş Noktası	İstek Sayısı	Ziyaretçi	Bant Genişliği
/kimmuh/kimya/CV.htm	28.940	28.940	787.861.324
/	24.163	24.163	4.838.375
/universiteevi/index.htm	18.929	18.929	140.785.826
/eemuh	15.836	15.836	5.655.484
/eemuh/haber_detay_ic.asp	8.662	8.662	134.357.366
/veteriner/hava_durumu.htm	7.863	7.863	18.123.853
/fenbilimleri	7.340	7.340	2.725.211
/eemuh/	6.551	6.551	420.920.690
/yapi/default.asp	6.330	6.330	84.201.599
/makinemuh	6.183	6.183	2.239.103
/makinetef	6.087	6.087	2.225.327
/veteriner	5.965	5.965	2.156.855
/ogrencikonseyi/2008/haberdetay.asp	5.952	5.952	463.659.099
/med	5.271	5.271	1.865.374
/kimmuh/cv/	5.197	5.197	1.038.331
/yapi	5.065	5.065	1.795.985
/elkbilgi	5.049	5.049	1.838.514

Ayrıca, tek başına görüntülenen sayfaların oranının yüksek olduğu görülmüştür. Bu sebeple en fazla giriş yapılan sayfaların aynı zamanda en fazla çıkış yapılan sayfalar olması da normaldir. Bunun dışında erişim örüntülerinin az olması sebebiyle bu Web sitesinin her bölümünden çıkış işlemlerinin mevcut olduğu görülmüştür.

Tablo 6.7. En fazla çıkış yapılan sayfalar

Çıkış Noktası	İstemler	Ziyaretçi	Bant Genişliği
/	23.599	23.599	4.724.847
/eemuh/	22.811	22.811	1.487.858.183
/universiteevi/index.htm	18.056	18.056	128.843.371
/kimmuh/cv/Örnek+cv.htm	11.760	11.760	272.098.867
/kimmuh/kimya/CV.htm	9.367	9.367	207.002.523
/veteriner/	6.887	6.887	226.526.741
/eemuh/haber_detay_ic.asp	6.700	6.700	108.155.277
/makinetef/	6.691	6.691	77.478.317
/ogrencikonseyi/2008/haberdetay.asp	6.612	6.612	517.514.500
/yapi/default.asp	6.161	6.161	87.971.037
/elkbilgi/index.asp	5.857	5.857	104.359.248
/insaatmuh/	5.800	5.800	102.586.644
/med/yaltFrame.htm	5.315	5.315	16.013.701
/fenbilimleri/index.htm	5.266	5.266	87.137.085
/kimmuh/cv/	5.156	5.156	1.030.131
/yapi/	5.148	5.148	2.153.167
/metal/	4.902	4.902	157.983.411

6.4.3 Kullanıcıya Ait İstatistikler

Bu bölümde, Fırat Üniversitesi Web sitesinde gezinen ziyaretçilere ait çıkarılan özel bilgiler yer almaktadır. Bu bilgiler alt başlıklar altında tablo şeklinde sunulmaktadır.

6.4.3.1 Ülkeleri

Ziyaretçilerin Web sitesine erişim yaptıkları ülkelere göre sınıflandırma yapıldığı zaman, Amerika Birleşik Devletleri'ndeki kullanıcıların sayısının diğer ülkelere göre çok daha yüksek olduğunu gösteren yanıltıcı bir bilgi görülmektedir. A.B.D.'li ziyaretçilerin profilleri incelendiğinde erişim hareketinde bulunan bu ziyaretçilerin büyük çoğunluğunun A.B.D.'li arama motorlarına (yahoo, google, msn, vs.) ait arama botları tarafından yönlendirildiği görülmektedir. Periyodik olarak siteleri ziyaret eden bu botlar sitenin içeriği, mimarisi ve köprü bağlantıları hakkında bilgi toplamaktadır. Bu ziyaretçilerin gerçek anlamda Fırat Üniversitesi Web sitesinin sürekli ve etkin bir kullanıcısı ya da üyesi olmadığı görülmüştür. Tablo 6.8'de ziyaretçilerin Web sitesine bağlandıkları ülke isimlerini ve diğer sayısal bilgiler gösterilmektedir. Bu tablo listesi, çıkarılan bilgi listesinden alınan ilk on dört ülke bilgilerini göstermektedir.

Tablo 6.8. Ülkelere göre kullanıcı istatistikleri

Ülke Adı	İstek Sayısı	Ziyaretçi	Sayfa	Bant Genişliği
Türkiye	4.514.183	332.167	9.061	57.060.574.039
Yerel Ağ (Firat Univ)	2.363.338	61.006	4.099	127.736.917.602
Amerika Birleşik Devletleri	49.471	2.164	1.113	5.459.876.050
Almanya	11.250	1.082	1.269	148.496.684
İngiltere	3.819	429	616	47.274.208
Bilinmiyor	2.478	212	233	31.839.274
Yunanistan	2.451	41	74	2.315.641
Fransa	1.355	258	376	102.917.165
Hollanda	1.324	294	421	96.201.155
Mısır	1.262	51	63	1.933.380
Meksika	1.185	188	271	38.074.611
İtalya	1.150	78	235	6.820.116
İsveç	1.102	210	224	9.955.688
Kanada	1.002	281	231	25.358.355

6.4.3.2 İşletim Sistemleri

Bu raporlama bilgisi, üniversite Web sitesinde gezinen ziyaretçilerin kullanmış oldukları bilgisayar işletim sistemlerinin kullanımına ilişkin istatistikî bilgiler sunmaktadır. Bilgilendirme amaçlı bu raporlama bilgisi ile ziyaretçilerin kullandıkları işletim sistemi

alışkanlıkları açıkça görülmektedir. Tablo 6.9'da görüldüğü gibi Microsoft firmasına ait işletim sistemleri ziyaretçilerin en yaygın kullandıkları işletim sistemleridir. Bu tablo bilgilerinden yola çıkılarak kullanıcıların Microsoft işletim sistemi ailesine olan ilgileri, birçok bakımdan yorumlanabilir. Özellikle işletim sistemindeki kullanım kolaylığı, görsellik ve donanım aygıtlarında tanıma uyumluluğu, yenilikleri sürekli güncelleyen bir yapıda olması ve yaygın kullanımı kullanıcıların işletim sistemini tercih etmelerini sağlamaktadır.

Tablo 6.9. İşletim sistemine göre kullanıcı istatistikleri

İşletim Sistemi	İstek Sayısı	Ziyaretçi Sayısı	Bant Genişliği (Byte)
Windows XP	5.735.781	309.020	66.681.891.002
Windows Vista	547.157	29.506	7.505.806.426
Win Server 2003	345.158	1.061	207.546.878
Bilinmiyor	161.539	53.294	2.972.164.368
Windows 98	131.801	6.160	111.838.817.740
Windows 95	15.786	28	5.742.745
Windows 2000	14.182	1.644	1.413.297.781
Mac OS X	8.984	269	26.357.714
Kullanıcı etmeni yok	2.361	93	98.493.506
Linux	1.931	536	59.338.033
Windows NT 4.0	164	25	268.678.464
Windows NT	28	19	668.177
Win 9x	21	8	475.661
Mac	10	4	88.780
Toplam	**6.964.903**	**401.667**	**191.079.367.275**

6.4.3.3 Web Tarayıcıları

Tablo 6.10'da, Web sitesinde gezinen ziyaretçilerin internette gezinirken kullanmış oldukları Web tarayıcılarını ve ilişkili oldukları işlemlerin sayısal bilgilerini göstermektedir.

Tablo 6.10. Web tarayıcılarına göre kullanıcı istatistikleri

Web Tarayıcı Adı	İstek Sayısı	Ziyaretçi Sayısı	Bant Genişliği (Byte)
MS Internet Explorer 6	4.026.132	205.362	46.133.403.231
MS Internet Explorer 7	2.229.627	114.056	23.735.930.828
Firefox	412.288	24.166	5.574.005.763
MS Internet Explorer 5	96.437	1.357	110.615.003.463
Opera 9	15.497	1.847	863.400.273
Safari	12.536	406	34.793.553
Netscape Communicator 4	4.165	989	89.920.999
Firefox 1.5	2.748	297	39.780.075
MS Internet Explorer 5.5	1.223	135	14.019.077
Mozilla	1.222	430	119.797.550
Netscape Communicator 4.5	434	42	805.306.102
Download accelerator 7	340	15	74.249.051
Netscape Communicator 4.7	74	30	1.863.958
Opera 8	38	12	746.395
MS Internet Explorer 4	9	8	190.832
MS Internet Explorer 3	6	5	245.213
Opera 6	5	5	208.294
Opera 7	5	4	216.310
Download accelerator 5	4	1	6.883.062
Opera 8.5	4	2	70.141
GetRight (Download Manager)	2	2	1.536.696
Toplam	**6.802.796**	**349.171**	**188.111.570.866**

6.4.3.4 Arama Motorları

Arama motorlarını kullanarak Fırat Üniversitesi Web sitesine gelen ziyaretçilerin kullandığı arama motorları bilgisi Tablo 6.11'de gösterilmektedir. Google arama motorunun kullanım oranının diğer arama motorları arasındaki üstünlüğü net bir şekilde görülmektedir.

Tablo 6.11. Arama motorlarına göre kullanıcı istatistikler

	Arama Motoru	İstekler	%	Ziyaretçiler	%	Sayfalar	%	Bant Genişliği	%
1	Google.com	75.015	98.96	66.179	98.97	1.490	89.11	3.03 GB	98.93
2	live.com	674	0.89	592	0.89	120	7,18	29.57 MB	0.94
3	MSN.com	59	0.08	41	0.06	32	1,91	1.90 MB	0.06
4	Yahoo.com	37	0.05	35	0.05	16	0.96	1.01 MB	0.03
5	Websearch.com	6	0.01	6	0.01	2	0.12	244.91 KB	0.01
6	t-online.de	5	0.01	4	0.01	4	0.24	137.64 KB	0.00
7	Altavista.com	3	0.00	3	0.00	3	0.18	107.22 KB	0.00
8	mysearch.com	2	0.00	2	0.00	2	0.12	64.25 KB	0.00
9	Alexa.com	1	0.00	1	0.00	1	0.06	479.13 KB	0.01
10	AOL.com	1	0.00	1	0.00	1	0.06	27.97 KB	0.00
11	search.com	1	0.00	1	0.00	1	0.06	41.59 KB	0.00
	Toplam	**75.804**		**66.865**		**1.672**		**3.06 GB**	

6.4.3.5 Anahtar Kelimeler

Bu kısımda, arama motorlarına anahtar kelimeler girerek Fırat Üniversitesi Web sitesine yönlendirilen ziyaretçilerin kullandığı anahtar kelimeler tespit edilmiştir. Bu durum, Web sitesinin içeriğinin kapsamlı bilgilerle dolu olmasına bağlıdır. Yani, akademik içerikli üniversite Web sitesindeki sayfalarda ders notları, kitap, doküman ve diğer önemli bilgilerin bulunması öğrencilerin siteye olan ilgisini arttıracaktır. Arama motorlarından anahtar kelime kullanarak siteye yönlendirilen ziyaretçilerin en çok *cv hazırlama* anahtar kelimesini kullanarak üniversite Web sitesine girmişlerdir. Arama motorlarından en çok ziyaretçi çeken diğer anahtar kelimeler Tablo 6.12'de verilmiştir. Diğer anahtar kelimeler ise, üniversitenin faaliyet alanları ve akademik birimleri ile ilgilidir. Tablo 6.12'de belirtilen anahtar kelimelerin listesi çıkarılan bilgilerden alınan ilk satırları göstermektedir.

Tablo 6.12. Arama motorlarında kullanılan anahtar kelimeler

Anahtar Kelimeler	İstemler	Ziyaretçi	Sayfa	Bant Genişliği
cv hazırlama	16.587	14.377	2	401.042.401
cv örneği	6.233	5.779	2	162.112.940
fırat haber	841	838	2	7.695.510
fırat üniversitesi	782	514	10	11.560.170
fırat tv	467	437	2	13.340.625
fırat	415	212	6	7.108.761
haberler	268	266	7	24.391.870
fırat üniversitesi tıp fakültesi	191	184	1	225.203
fırat üniversitesi veteriner fakültesi	171	87	2	3.098.075
veterinerlik fakültesi	156	82	2	2.895.456
tenis	155	138	5	242.509
firat haber	135	134	1	1.221.393
veteriner fakültesi	130	67	2	2.418.917
fırat öğrenci konseyi	126	11	3	9.375.222

6.4.4 Ziyaretçilerin Siteye Yönlendirilme İstatistikleri

Web sitesini kullanan ziyaretçilerin sayfaya geliş durumları hakkında bilgi elde etmek ve bu durumdan yola çıkarak Web sitesini ziyaret eden kullanıcıların siteye olan talepleri ile ilgili yorum ve analizler yapmak mümkündür. Tablo 6.13'de ziyaretçilerin siteye yönlendirilme durumları detaylı olarak incelendiğinde, 75.804 kişinin arama motorlarından yararlanarak Web sitemize giriş yaptığı gözlenmektedir. Bunun yanı sıra 312.976 kişi ise, direk olarak siteye giriş yapmış ve Web isteğinde bulunmuşlardır.

Tablo 6.13. Ziyaretçilerin siteye yönlendirilme durumları

Yönlendirme Tipi	İstekler	Ziyaretçi	Sayfa	Bant Genişliği
Harici yönlendirme	6.575.804	284.001	8.039	175.457.912.275
Yönlendirme yok	312.976	98.189	7.320	12.344.797.936
Örümcek	79.321	21.344	7.482	3.600.621.571
Arama motoru	75.804	66.865	1.532	3.289.089.044
Toplam	7.043.905	470.399	24.373	194.692.420.826

6.4.5 Web Aktivite İstatistikleri

Üniversitenin geniş bir alana hitap etmesi, siteye pek çok kullanıcıyı çekmektedir. Ancak bu kullanıcı kitlelerinin büyük çoğunluğunu üniversite öğrencileri ve akademik personel oluşturmaktadır. Web sayfalarında üniversite geneli ve akademik bölümler hakkında detaylı bilgi içeriğinin olması kullanıcıların bu sayfalarda uzun süre geçirmesine sebep olmaktadır. Özellikle akademik personellerin kişisel ilgi alanlarına ait dokümanlar, ders notları, öğrencilerin ilgisini çeken bilgi içeriğinin fazla olması sayfalardaki Web aktivitelerini arttırmaktadır. Web sitesinde bölümlerin Web sayfalarındaki bilgilendirmeler, ana sayfalarda verilen açıklama ve duyurular, üniversitenin faaliyet alanlarına yönelik akademik ve idari işler hakkındaki bilgilendirmeler kullanıcıların siteye ilgilerini ve daha uzun vakitler geçirmesini sağlamıştır. Tablo 6.14'de Web sunucularından alınan erişim kayıtlarının 3 aylık süre içerisindeki (Şubat, Mart, Nisan) Web aktiviteleri görülmektedir.

Tablo 6.14. Aylık Web aktiviteleri

Yıl, Ay	İstek Sayısı	Ziyaretçi	Sayfa	Bant Genişliği
2008 Şubat	3.432.723	148.528	7.753	26.346.267.347
2008 Mart	1.947.286	128.572	6.569	117.432.153.690
2008 Nisan	1.583.736	122.856	6.736	47.285.826.428

Tablo 6.15'de kullanıcıların 3 aylık süre içerisindeki günlük ortalama Web aktiviteleri verilmektedir. Belirtilen süre içerisinde en çok istek yapılan tarih 5 Şubat 2008 Salı günü

753.241 adet Web sunucusundan istek yapılmıştır. Tablo bilgileri kullanıcı istek sayısına göre sıralanmaktadır. Ancak, istenildiğinde tablo içerisindeki bu sıralama diğer kolonlara göre de sıralanabilir.

Tablo 6.15. Günlük Web aktiviteleri

Tarih	İstek Sayısı	Ziyaretçi Sayısı	Sayfa Adeti	Bant Genişliği
05/Şubat/2008 (Salı)	753.241	5.870	1.913	682.402.729
04/Şubat/2008 (Pazartesi)	423.551	5.349	1.847	559.132.441
01/Şubat/2008 (Cuma)	163.225	4.225	1.871	488.296.972
07/Şubat/2008 (Perşembe)	146.476	10.037	2.783	2.541.679.552
13/Şubat/2008 (Çarşamba)	142.123	6.781	1.956	1.798.383.319
01/Mart/2008 (Cumartesi)	139.641	2.909	1.681	730.028.288
23/Mart/2008 (Pazar)	131.331	8.652	2.831	1.641.836.323
14/Şubat/2008 (Perşembe)	127.436	6.660	2.331	550.686.173
11/Şubat/2008 (Pazartesi)	115.485	7.380	2.378	639.966.287
21/Mart/2008 (Cuma)	114.205	5.429	2.004	92.157.799.648
18/Mart/2008 (Salı)	112.817	5.270	1.879	536.719.639
25/Nisan/2008 (Cuma)	111.973	5.263	2.353	977.824.897
12/Şubat/2008 (Salı)	108.267	6.905	2.143	592.157.089
03/Mart/2008 (Pazartesi)	107.857	4.898	2.014	1.188.474.783
22/Şubat/2008 (Cuma)	105.167	4.985	1.919	847.973.554

Tablo 6.15'de, kullanıcıların 3 aylık süre içerisindeki yapılan istekleri ortalama saatlere göre sıralanmaktadır. Tablo 6.16 incelendiğinde toplamda en çok Web aktivitelerinin yapıldığı zaman sabah 8:00, öğlen 12:00 ve 13:00 saatleri dolaylarıdır.

Tablo 6.16. Saatlere göre Web aktiviteleri

Saat	İstekler	Ziyaretçi	Sayfa	Bant Genişliği
8	520.937	29.011	4.011	6.511.921.560
12	505.444	31.398	4.183	6.469.618.940
13	494.991	29.823	4.030	18.172.777.204
11	485.946	30.139	4.348	6.403.915.882
9	482.800	28.858	4.012	10.711.305.966
10	466.926	27.063	3.974	5.367.651.349
14	404.836	27.941	4.125	44.969.850.682
20	378.526	20.071	4.174	4.518.578.706
19	348.711	21.081	4.225	5.609.545.875
15	347.288	23.382	3.866	35.105.839.322
17	342.789	22.099	4.186	4.780.160.441
18	338.936	21.623	4.018	4.293.273.729
16	322.329	21.269	3.778	16.322.704.115
7	318.753	21.282	3.716	4.304.010.074
21	311.805	17.026	5.633	3.697.105.073
22	233.461	11.359	4.842	2.342.471.149
6	166.371	12.609	2.976	4.815.270.064
23	165.680	6.192	2.855	2.393.313.471
0	75.667	3.423	2.169	603.977.401
1	63.125	1.732	1.746	359.478.457
2	55.973	1.057	1.366	311.258.711
5	55.836	4.586	2.139	1.533.188.441
4	43.974	1.414	1.182	1.174.426.719
3	33.799	957	1.118	307.723.944

6.4.6 HTTP Durum Kodlarına Göre İstatistikler

Web sitesinde gezinen ziyaretçilerin sunucuya yapmış oldukları HTTP isteklerine, sunucu tarafından durum kodları ile cevap verilmektedir. Erişim kütüklerinden çıkarılan bu durum kodlarına göre yapılan isteklerin durumu analiz edilir.

Bu uygulama çalışmasında, incelenen erişim kütüklerinden çıkarılan HTTP durum kodları Tablo 6.17'de verilmektedir. HTTP kodları analiz edilerek, Web sitesini ve sunucu başarımını sınırlayan hatalar görülmektedir. Tespit edilen bu hataların analizinde en önemli referans kaynağı olarak HTTP durum kodları göz önünde bulundurulmaktadır [21].

HTTP durum kodları bilgilerinin çıkarıldığı Tablo 6.17 analiz edildiğinde, üç aylık süre içerisinde ziyaretçiler tarafından toplam 17.544.714 istek yapılmış ve bu isteklerin 4.609.136 tanesinde ise erişimin gerçekleştirilemediği görüldü. 200, 302, 304 kodlu erişim durum kodları dosyaya yapılan erişimleri göstermektedir.

Tablo 6.17. HTTP Durum kodlarına göre istatistikî bilgiler

HTTP Durum Kodları	Yapılan İstekler	Ziyaretçiler	Sayfalar	Bant Genişliği
Toplam Değerler	17.544.714	1.056.755	18.085	485.53 GB
200 - Başarılı	12.935.578 (73.73%)	489.427 (46.31%)	9.648 (53.35%)	411.54 (84.76%)
302 - Bulunamadı	35.242 (0.20%)	23.137 (2.19%)	219 (1.21%)	14.13 MB (0.00%)
304 - Değiştirilmedi	2.847.604 (16.23%)	107.727 (10.19%)	3.538 (19.56%)	577.50 MB (0.12%)
400 - Geçersiz istek	43 (0.00%)	26 (0.00%)	8 (0.04%)	-
401 - Kimlik doğrulama hatası	3 (0.00%)	3 (0.00%)	2 (0.01%)	7.62 KB (0.00%)
403 - Yasak	21.673 (0.12%)	6.414 (0.61%)	292 (1.61%)	7.75 MB (0.00%)
404 - Bulunamadı	1.164.182 (6.64%)	195.733 (18.52%)	3139 (17.36%)	1.77 GB(0.37%)
405 - Yönteme izin yok	8 (0.00%)	6 (0.00%)	3 (0.02%)	13.91 KB (0.00%)
406 - Kabul edilemez	16 (0.00%)	9 (0.00%)	4 (0.02%)	23.75 KB (0.00%)
500 - Dâhili sunucu hatası	4.208 (0.02%)	1.464 (0.14%)	298 (1.65%)	21.50 MB (0.00%)
501 - Uygulanmadı	39.177 (0.22%)	8.251 (0.78%)	77 (0.43%)	5.17 MB (0.00%)
502 - Bozuk ağ geçidi	161 (0.00%)	119 (0.01%)	2 (0.01%)	61.53 KB (0.00%)

6.5 Bulguların Değerlendirilmesi

Web sitesine ait kullanıcı erişim kütükleri analiz edilerek bir Web sitesinin kalitesi değerlendirilebilir. Web sitesinin kalitesini etkileyen birçok faktör vardır. Örneğin, sayfa içeriği, sunuş, kolay kullanım, kullanıcının sayfada bekleme süresi ve benzeri gibi göz ardı edilmeyecek faktörler bulunmaktadır. Web kullanıcı erişim kütük dosyaları derinlemesine analiz edildiğinde, herhangi bir işleve sahip olan dosyaları isteyen kullanıcılar, kullanım yoğunluklarına göre iyi, daha iyi, en iyi şeklinde gruplandırılabilirler. Bilgi amaçlı dosyaları isteyen kullanıcılara potansiyel iyi kullanıcı gözüyle bakılabilir. Bilgi amaçlı dosyalar eğer iyi hazırlanmışsa siteye iyi kullanıcı kazandırmakta önemli bir işleve sahip olabilir. Bazı kullanıcılar ise sitede olmayan bazı dosyalara istekte bulunmaktadır. Bu tip kullanıcıların amacı Web sunucu üzerindeki boşluklardan faydalanarak sistemi çalışmaz hale getirmektir. Bu tip dosyalara istekte bulunan kullanıcılar tehlikeli veya kötü kullanıcı olarak tanımlanabilir. Kullanıcılar, erişim yaptıkları dosya adetlerine göre puanlanabilir ve almış oldukları bu puanlara göre değerlendirilebilirler.

Bu çalışmada, Fırat Üniversitesi Web sunucularından alınan Web kullanıcı erişim kütükleri Web içerik, Web yapı ve Web kullanımı bakımından incelenerek, Web madenciliği uygulaması gerçekleştirildi. Önerilen ve gerçekleştirilen Web madenciliği uygulamalarında elde edilen sonuçlar ve bulguların değerlendirilmesi sonucunda Web sitesine ilişkin bilgiler aşağıda sunulmuştur.

1- Web içeriği ile ilgili tespit edilen sonuçlar

- Web sitesi içeriğindeki mevcut dosya yapısını, içeriğini ve bunlar arasındaki sıklık ilişkileri tespit edilebilmektedir.
- Web sitesine yönlendirme yapan arama motorlarının adları ve bu arama motorlarında kullanılan anahtar kelimeler tespit edilebilmektedir.

2- Web yapısı ile ilgili tespit edilen sonuçlar

- Web site mimarisindeki düzensizlikler ve bozukluklar görülebilmektedir.
- Kullanıcıların Web sitesinden talep ettikleri bağlantı erişimlerinde görüntülenemeyen dosyalar tespit edilebilmektedir.

3- Web kullanımı ile ilgili tespit edilen sonuçlar

- Web sitesini ziyaret eden kullanıcıların siteyi kullanımlarından dolayı oluşturdukları Web trafiği sonuç bilgileri kapsamlı ve detaylı olarak çıkarılabilmektedir. Bu bilgiler kapsamında aşağıda belirtilen maddeler altında detaylı olarak istatistikî bilgiler çıkarılabilmektedir.
  - Kullanıcıların siteye erişim yaptığı günler ve saatler
  - Kullanıcıların haftalık erişim eğilimler

  - Kullanıcıların siteye erişim yaptıkları ülkeler
  - Kullanıcıya ait Web tarayıcıları ve işletim sistemler
  - Kullanıcı ziyaret saatleri, süreleri ve sayılar
  - Kullanıcıların kullandıkları arama motorlar
  - Kullanıcıların talep ettikleri dosya tipler
  - Site içerisinde gezindiği Web sayfalar
  - Kullanıcıların siteye giriş ve çıkış yaptıkları sayfalar
- En sık kullanılan sayfa çiftleri tespit edilebilmektedir. Böylece, Web ziyaretçilerin sayfaları kullanım eğilimleri görülebilmektedir. Bu kapsamda, hedef ziyaretçi grupları sınıflandırılarak, Web sitesi tasarımında yeni bir organizasyon tanımlaması yapılabilir.

Gerçekleştirilen bu Web madenciliği uygulama çalışmasında, metin tabanlı erişim kütüklerinden çıkarılan bilgilerin Web kullanım madenciliğinin yanı sıra Web içerik ve Web yapı madenciliğine yönelik önemli bilgilerin çıkarılarak Web sitesinin gelişimine katkılar sağlanılabileceği görülmektedir. Bu uygulama çalışmasında Web madenciliğinin alt bölümleri bir bütün olarak ele alınmaya çalışılmış ve kapsamlı olarak istatistiksel bilgiler çıkarılmıştır. Yapılan bu istatistiksel analiz ile Web sitesinin tasarımı, güncellenmesi, gelişimi ve başarımı için hem Web site yöneticisine hem de Web site tasarımcısına büyük ölçüde destek sağlayacaktır.

6.6 Problemin Çözümüne Yönelik Öneriler

Birçok istemci, sunucu ve ağ sistemlerinden oluşan dağınık bir sistemin performansı üzerindeki iş yükünün özelliklerine bağlıdır. Bu nedenle, sistem başarımını değerlendirirken öncelikle sistem üzerindeki iş yükü belirlenmelidir. Bu iş yükü, sistemin belirli bir süre içerisinde çevresinden aldığı tüm girdilerin oluşturduğu bir kümedir. Web sunucu başarımını etkileyen birçok önemli girdiler kümesi bulunmaktadır. Bu girdiler genel olarak yazılım, donanım ve dış etkenlerden oluşmaktadır.

Web sitesi *içerik* açısından incelendiğinde, site içeriğinin amacına uygun olduğu ancak sayfa içeriklerindeki akademik bilgilerle zenginleştirildiğinde ve yeni güncel bilgiler eklendiğinde ziyaretçi sayısının artacağı tespit edilmektedir. Örneğin; Web sitesinde bulunan duyurular ve haberler sayfasına yeni bilgiler eklendiğinde bu sayfaların girilme sayı artmıştır. Üniversitede meydana gelen yeni gelişmeler eklendiğinde siteye ilgi artmış, merak uyandıran ve bilgi konulan her sayfanın kullanım oranı, önceki aylara göre mutlaka artmıştır. Siteye her yeni gelen ziyaretçinin büyük çoğunlu, Web sitesinde bilgi arayan öğrenci ve akademik kimselerdir.

Web sitesi *yapı* (*mimari)* açıdan incelendiğinde, site alt yapılanmasında mimari olarak düzensizlikler olduğu tespit edilmiştir. Bu düzensizliklerin giderilmesi durumunda, Web site ziyaretçileri açısından kullanım kolaylığı sağlayacaktır.

Web sitesi *kullanım* açısından incelendiğinde, ziyaretçilerin siteyi kullanımlarından dolayı oluşturdukları Web trafikleri ve Web aktivite bilgileri Bölüm 6.4 altında tablolar halinde verildi. Çıkarılan bu bilgiler ışığında Web site başarımını arttırmak için aşağıda belirtilen maddeler site tasarımcısı ve site yöneticisi tarafından mutlaka göz önünde bulundurulmalıdır.

- Genel istatistikler incelendiğinde Web sitesine erişim yapan kullanıcı sayısının bir üniversite Web sitesine göre az olduğu söylenebilir. Ziyaretçi istekleri dikkate alınarak Web sitesi içeriği zenginleştirilirse Web sitesi kullanım sayısı artacaktır.
- Site erişim istatistiklerine göre üniversite içerisindeki bölümlerin Web sayfalarında bir dengeleme olmadığı analiz edildi. Yani, üç aylık süre içerisinde birçok bölüm ve birimlerin Web sayfalarına dahi girilmediği görüldü. İlgili birim ve bölümler Web sayfalarını yeniden düzenleyerek, çalışır hale getirmelidirler.
- Web sitesinde en çok erişim yapılan ve en fazla bant genişliğini kullanan dosya biçimlerinden biride resim (jpg, gif) dosyalarıdır. Web sitesinin hızlı ve başarımlı çalışabilmesi sayfalardaki resim dosyaların boyutlarına da bağlıdır. Ne amaçla olursa olsun Web sitelerinde resim dosyalarının kullanılması zorunludur. Ancak, resim dosyalarını küçük boyutta ve yüksek kalitede seçilmesine özen gösterilmelidir. Üniversiteler akademik kurumlar olduğu için, zaruri olmayan Web sayfalarında ziyaretçilerin kullanımına metinsel sayfalar sunulabilir. Üniversitelerde önemli olan kullanıcılara geniş bilgi bankası sunmaktır.
- Web sitesine yabancı ülkelerden ziyaretçi erişimleri mevcuttur. Bu ziyaretçi grubuna daha iyi hizmet verebilmek için Web sitesinin İngilizce sürümünün bir an önce bitirilip hizmete sunulması gereklidir. Yabancı ülkelerden üniversite Web sitesini inceleyen kişilerin aynı zamanda site içerisinde akademik bilgi arayışı içersinde olabileceği düşünülerek, akademisyenlerin aynı zamanda hem kişisel Web sayfalarını hem de ders notlarının İngilizce olarak Web sayfalarına eklemeleri Web sitesinin gelişimine büyük katkı sağlayacaktır. Bu durum sonucunda Web sitesine ilgisi olan ziyaretçi sayısında da artış gösterecektir. Ticari kuruluşların Web sayfalarında farklı dilde hizmet sunmaları, o şirketin uluslar arası platformda da çalışmalar yaptığını ve önemli bir dış sermaye potansiyelinin olabileceği düşünülmelidir.
- Arama motorlarını kullanarak Web sitesine yönelen ziyaretçilerin kullandıkları anahtar kelimeler tespit edilebilmektedir. Aynı zamanda Web sitesi adresinin tüm arama motorlarında tanımlanması sitenin birçok aramalarda ön satırlarda çıkmasını

sağlayacaktır. Ayrıca, Web sitesinin zengin bir içeriğe sahip olmasının yanı sıra farklı birçok konuda bilgi ve ders notlarının eklenmesi siteye erişen ziyaretçi sayısında büyük artış gösterecektir.

- Günlere göre erişim adetleri tablosuna (Tablo 6.15) bakıldığında erişimlerin düzenli olmadı görülür. Ziyaretçiler periyodik olarak değil, ihtiyaç duyduğunda siteye başvuruyor. Ziyaretleri periyodik hale getirmek için sitenin içeriği zenginleştirilebilir. İlginç veya ilgili konulara sahip sitelere köprü verilip, bu köprüler bölüm ana sayfalarında ön plana yerleştirilebilir.
- Saatlere göre ziyaretçi erişimleri tablosu (Tablo 6.16) incelendiğinde, özellikle Web sitesinin ana sayfa güncelleştirmesi için en uygun zamanlamanın saat 22.00 ile 05.00 arasında yapılması daha uygun gözükmektedir.
- Çıkarılan durum kodlarına göre ziyaretçilerin toplam erişim isteklerinin %26,27'si hatalı olmuş, yani yapılan istekler bir şekilde karşılanamamıştır. Köprü bağlantısı kırık olan dosyaların ve site mimarisinin acilen gözden geçirilip aksaklıkları giderilmelidir. Site içerisinde yeni sürümlü ürünlerin eklenmiş olmasından dolayı ziyaretçi erişim problemi yaşamıştır. Teknoloji düzeyi yüksek olan sayfaların görüntülenmesi için site içerisinde uygun yazılımların indirilmesi için çalışma yapılmalıdır ya da sitenin yapısı orta seviye teknolojik donanıma sahip ziyaretçiler tarafından kullanılabilmesi için bu seviyeye çekilmelidir.
- Çıkarılan durum kodlarına göre 4.208 adet ziyaretçi isteğinin sunucu kaynaklı probleminden dolayı erişimlerinin gerçekleşmediği görüldü. Sunucular üzerindeki veri tabanlarına sorgulama yapıldığı zaman bu durum sunucunun işlemci ve bellek ünitesinde yüksek düzeyde performans isteyecektir. Sunucu başarımını arttırmak için gereksiz veri tabanı sorgulamalarını ve yanlış sorgulama kodlamalarını optimize etmek gerekir. Ayrıca, sunucu donanım yapılandırması olarak yükseltilmeye müsaitse hem işlemci hem de bellek ünitesinde yükseltilme yapılmalıdır. Bu durum Web site başarımını da beraberinde getirecektir. Çünkü Web sitesine yapılan isteklerin artması sonucunda sunucunun daha yüksek hızda ve kapasitede çalışmasını gerektirecektir.
- Kullanıcı ilgilerinin ziyaret sürelerine göre yoğun olduğu sayfalardaki sayfa içerikleri dikkate alınarak, diğer bölümlerde benzer sayfaların arttırılmasına gidilmelidir. Ders notları, faydalı ve gerekli güncel bölüm bilgilerinin olduğu sayfaların erişim sayısı diğer sayfalara göre daha fazladır. Web sayfalarının iyileştirilmesinde ziyaret sürelerini arttırmak sitenin asli görevi olan geniş tanıtım ve bilgilendirmenin yanı sıra öğretim elemanlarının ders notlarını eklemeleri sitenin gelişimi açısından fayda sağlayacaktır.

- Site mimarisindeki bozuklukların giderilmesi için hiyerarşik bir yapı sağlanmalıdır. Önemli ve güncel bilgilere ulaşım için hızlı erişim köprü bağlantıları konulmalıdır.
- En sık kullanılan sayfa çiftleri tespit edilerek, Web ziyaretçilerin sayfaları kullanım eğilimleri görülebilir. Bu kapsamda, hedef ziyaretçi grupları sınıflandırılarak, Web sitesi tasarımında yeni bir organizasyon tanımlaması yapılabilir. Elektronik ticaretle uğraşan Web sitelerinde ise, belirlenen sayfa çiftlerine göre yapılan müşteri gruplaması iyi değerlendirilip, hedef müşteri gruplarını müşteri portföyüne katmak için çalışmalar yapılmalıdır.

Çıkarılan bulgularda ve yapılan analizlerde arama motorları ve yönlendirmeler ile siteye gelen ziyaretçilerin tümünün bilinçli olarak siteye gelmediği görülmüştür. Yüksek tahmin ve gerçek sonuçlara yakın değerler elde edebilmek için bahsi geçen ziyaretçileri Web kullanım analizine dâhil etmemek gerekir. Bunun için ön işlem aşamasında da bu ziyaretçiler için filtreleme yöntemi kullanılmalıdır. Bu metotta ziyaretçiler örüntü analizinden önce IP adresi, ülke, bölge, erişim yaptığı saat, gün, hafta, ay, erişim isteğindeki hata veya başarı kodu, ilk erişim yaptığı sayfa, son erişim yaptığı sayfa, işletim sistemi, Web tarayıcısı, özel erişim yaptığı bir dosya, yönlendiren sayfa gibi kıstaslara göre ayıklanabilir.

7. BİRLİKTELİK KURALI YÖNTEMİ İLE WEB KULLANICI ERİŞİM KÜTÜKLERİNDEN BİLGİ ÇIKARIMI

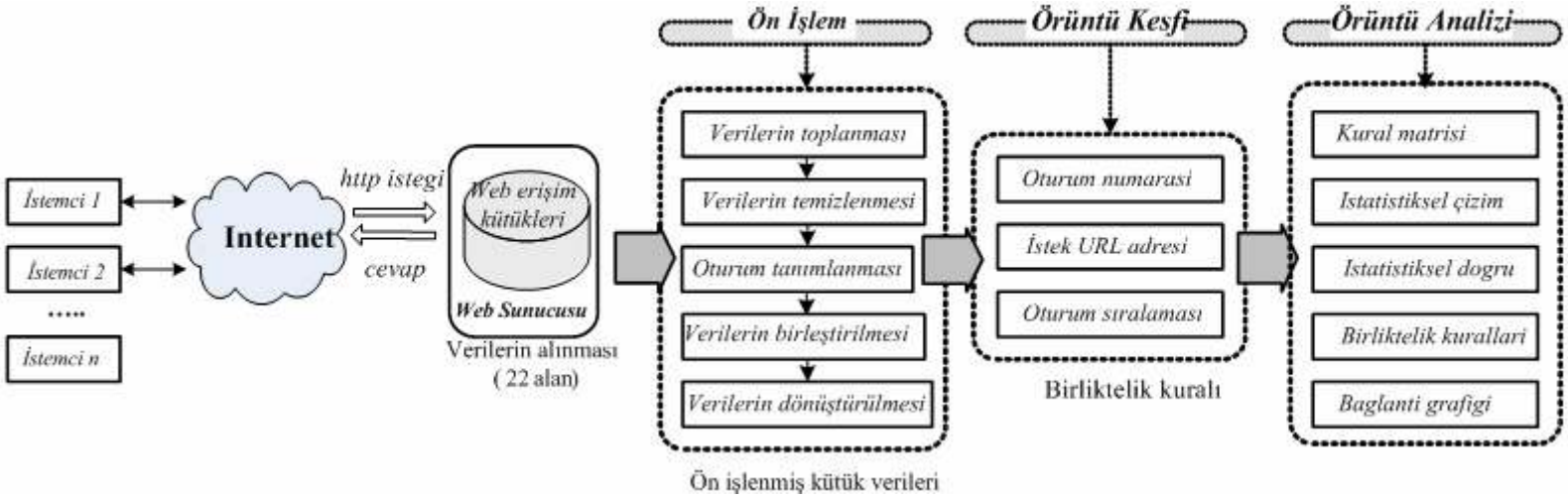
Bu bölümde, veri madenciliği tekniklerinden birliktelik kuralı kullanılarak web erişim kütüklerinden web sayfalarının birliktelikleri ile ilgili anlamlı sonuçlar çıkarılmıştır. Bu çalışmada, Bölüm 5'de kullanılan Fırat Üniversitesi web sunucularına ait web kullanıcı erişim kütükleri kullanılmıştır. Literatürde, birliktelik kuralları madenciliği ile ilgili yapılmış birçok çalışma bulunmaktadır [75, 96–98, 124–126]. Veri madenciliğinde yaygın olarak kullanılan birliktelik kuralı, bu çalışmada ise web sayfalarının birlikteliklerinin analizini yapmak için kullanılmıştır. Çalışmanın sonucunda elde edilen bulgular ve sonuçlarda birliktelik kurallarının yanı sıra istatistiksel bilgiler, kural matrisi ve köprü bağlantı grafiği bilgileri de çıkarılmıştır. İnternet kullanıcılarının web sitesi kullanım memnuniyetini arttırmak ve kullanıcı ilgilerinin sağlanması yönünde, sitenin geliştirilmesi ve iyileştirilmesine yönelik çözüm önerileri getirilmiştir**.**

Web kullanıcı erişim kayıtlarından birliktelik kuralı ile bilgi keşfi için önerilen yöntem, JAVA tabanlı SAS Base yazılım ortamında gerçekleştirildi [82].

Web kullanıcı erişim kütüklerinden birliktelik kurallarının ve istatistiklerin çıkarılması için geliştirilen yöntemin algoritma şeması Şekil 7.1'de gösterilmektedir. Yöntem; *ön işlem, örüntü keşfi* ve *örüntü analizi* olmak üzere üç ana bileşenden oluşmaktadır. Bu bileşenler üzerinden geliştirilen yöntemin ayrıntıları aşağıdaki bölümlerde sunulmuştur.

Bu bölümde, web kütük dosyalarının analiz edilmesi için önerilen yöntemin gerçekleştirilen uygulaması detaylı olarak açıklanmaktadır. Bu çalışmadaki amaç, bir web sitesinde meydana gelen köprü bağlantılarını tanımlayarak web tasarımcılarına ve web yöneticilerine web sitelerini geliştirmeleri ve iyileştirmeleri için destek sağlamaktır. Çalışmada kullanılan erişim kütük dosyaları, Fırat Üniversitesi Bilgi İşlem Daire Başkanlığı bünyesinde hizmet veren web sunucularına aittir. Önerilen yöntem için geliştirilen yazılım kodlamaları SAS Base yazılım ortamında gerçekleştirilmiştir.

Gerçekleştirilen yöntemin algoritma şeması Şekil 7.1'de her bir aşaması adım adım gösterilmektedir.



Şekil 7.1. Birliktelik kuralı ile bilgi çıkarımı için geliştirilen yöntemin prensip şeması

7.1 Ön İşlem Süreci

Bu uygulama çalışmasında kullanılan web kullanıcı erişim kütük verileri Bölüm 5'te kullanılan verilerin aynısıdır. Web kullanım madenciliği için ön işlem sürecinde izlenen aşamalar, yol analizi yönteminde nasıl kullanılıyorsa birliktelik kuralı yönteminde de bu kullanım aynen geçerlidir. Ancak, birliktelik kuralı madenciliği için ön işlem aşamasından geçirilmiş kütüklerden *oturum numarası*, *istenilen URL adresi* ve *oturum sırası* olmak üzere üç alanı belirten veriler alındı. Ön işlem süreci sonunda elde edilen ve Tablo 7.1'de gösterilen kullanım bilgilerine birliktelik kuralı yöntemi uygulandı.

Tablo 7.1. İşlenmiş erişim kütüklerinden bir kesit

Oturum numarası	İstek adresi	Oturum sıralaması
67e2a4da826149c5 2008-01-14 08:38:17	/default.asp?id=7	1
67e2a4da826149c5 2008-01-14 08:38:17	/?git=fakulteler	2
67e2a4da826149c5 2008-01-14 08:38:17	/eemuh/default.asp	3
67e2a4da826149c5 2008-01-14 08:38:17	/eemuh/dersler_detay.asp	4
67e2a4da826149c5 2008-01-14 08:38:17	/eemuh/default.asp	5
67e2a4da826149c5 2008-01-14 08:38:17	/eemuh/foto/igallery.asp	6
67eb39ebce6c73c3 2008-01-14 23:43:23	/default.asp	1
67eb39ebce6c73c3 2008-01-14 23:43:23	/?git=enstituler	2
67eb39ebce6c73c3 2008-01-14 23:43:23	/fenbilimleri/index.htm	3
67eb39ebce6c73c3 2008-01-14 23:43:23	/fenbilimleri/dersler.htm	4
67eb39ebce6c73c3 2008-01-14 23:43:23	/fenbilimleri/computer.htm	5
67cb8704bb370ee2 2008-01-14 13:44:04	/default.asp	1

7.2 Örüntü Keşfi

Çalışmanın bu aşamasında, Tablo 7.1'de görüldüğü gibi kullanıcı web erişim kütükleri ön işlem aşamasından geçirilmiş ve örüntü keşfi için hazır hale getirilmiştir. Bu verileri analiz etmek için birliktelik kuralı yöntemi kullanılmaktadır. Şekil 7.1'de görüldüğü gibi birliktelik analizi için *oturum numarası, istenilen URL adresi* ve *oturum sıralaması* alan bilgilerinin ilişkilendirilmesi yeterli olmaktadır. Bu alan bilgileri kullanılarak yapılan analiz uygulamasında elde edilen bulgular ve sonuçlar Bölüm 7.3'de açıklanarak, şekil ve grafiklerle verilmektedir.

7.2.1 Birliktelik Kuralı

Birliktelik kuralı keşfi popüler bir veri madenciliği yöntemidir. Birliktelik kuralı, büyük veri kümelerindeki öğeler arasındaki karşılıklı ve ilginç ilişkileri bulmayı amaçlamaktadır [146-147].

Web kullanım madenciliğinin içeriğindeki birliktelik kuralları, belirlenmiş bazı eşik değerlerini aşan bir destek değeri ile birlikte erişilen sayfaların kümelerine bakmaktadır. Bu sayfalar köprü bağlantıları yoluyla birbirlerine direk olarak bağlı olmayabilirler. Örneğin; Apriori algoritması kullanılarak birliktelik kurallarının keşfi için kullanıcılar arasında karşılıklı bir bağıntı olduğunu ortaya çıkarılabilir. Bu gruba sporla ilgili bir gereç hakkındaki web sayfası bilgilerine erişmek isteyen kullanıcılar ya da elektronik ürünlerin özelliklerini içeren bir web sayfasında gezinen kullanıcılar örnek olarak verilebilir.

Web kullanım madenciliği için birliktelik kurallarının keşfi oturum temeline dayanmaktadır. Her bir oturum, bir sepette konu ile ilgili var olan öğelerin oturumdaki web sayfalarının olayları ve bir işlemi olarak yorumlanmaktadır. Bir kullanıcı oturum süresi devam edene kadar ardı ardına sayfalarını takip etmesi, ardışık olan bu web sayfalarında gezinmiş olduğunu göstermektedir. Birliktelik madenciliği yöntemi web kullanım madenciliğine uygulandığı zaman, ilginç erişim örüntüleri ve web sayfaları arasında karşılıklı bağıntılar keşfedilebilir. Örneğin; büyük işlemlerin yapıldığı veritabanlarındaki öğe kümeleri arasında birlikte ziyaret edilen sayfaların bulunması probleminde birliktelik kuralları yöntemi etkin olarak uygulanabilir. Kullanıcı oturumları veya işlemler tanımlandığı zaman, sayfa birlikteliğini keşfetmek için derhal Apriori algoritması veya varyantları uygulanabilir [8, 75, 98, 125].

Bir birliktelik kuralı, belirlenmiş biçimdeki bir deyimin ifade edilmesidir. Örneğin, bir kural tanımlamasını ifade ederken (*A kümesi* $\Rightarrow$ *B kümesi*) şeklinde yazılabilir. İki kısımdan oluşan bu yapıda, ilk kısım bir nevi şart ifadesini ikinci kısım ise sonuç durumunu belirtmektedir. Birliktelik kuralı, bir kural hakkındaki belirsizlik derecesini açıklamak için *Güven* ve *Destek* değerleri olmak üzere iki sayısal ifadeye sahiptir.

Destek, A ve B öğelerini birlikte içeren işlemlerin sayısının toplam işlemlerin sayısına oranıdır. Hem A hem de B öğelerini birlikte içeren bir işlemin olması muhtemeldir. $A \Rightarrow B$ kuralı için destek, bu iki öğe kümesinin birlikte gerçekleşmesi olasıdır. $A \Rightarrow B$ kuralı için *destek* değerinin hesaplanması denklem (7.1)'de gösterilmektedir.

$$\frac{\text{A ve B yi birlikte içeren islemler}}{\text{Bütün islemler}} \tag{7.1}$$

Destek değeri dönüşümlüdür. Yani, A=>B kuralı için destek değeri B=>A kuralı için destek değeri ile aynıdır.

Güven, sonuç kısmındaki bütün öğeleri içeren işlem sayısının, şartlı kısmındaki bütün öğeleri içeren işlemlerin sayısına oranıdır. Eğer, bir işlem A öğesini içeriyorsa aynı zamanda B öğesini de içeriyor olması muhtemel bir durumdur. $A \Rightarrow B$ bir birliktelik kuralı için *güven,* A öğesi ile beraber B öğesini de içeren bir işlemin olası durumudur. $A \Rightarrow B$ kuralı için *güven* değerinin hesaplanması denklem (7.2)'de gösterilmektedir.

$$\frac{\text{A ve B ögesini içeren islemler}}{\text{A ögesini içeren islemler}} \tag{7.2}$$

Birliktelik kural madenciliğinin sonuçlarında, birlikte ziyaret edilen web sayfaları gruplarının çiftleri bulunabilir. Birliktelik kuralları, etkili pazarlama stratejileri geliştirmek ve onların web sitelerini daha iyi kullanmak için organizasyonlara yardım etmektedir [82].

Birliktelik kuralı genellikle veri tabanındaki veriler arasındaki ilişkileri tespit etmeye çalışır. Web kullanım madenciliğinde, birliktelik kuralı algoritmaları iki amaç için kullanılır: İlk olarak, çevrimiçi satın alma verisini analiz edilir, çevrimiçi müşteriler tarafından hangi ürünlerin beraber satın alındığı belirlenir (geleneksel süpermarket sepet veri analizine benzer). Çevrimiçi alışveriş veri tabanları, her bir müşterinin bir ürün alt kümesine seçildiği, önceki müşteri verilerini içerir. Bu veri, ürün seçimini yapan müşteriye yeni ürünlerin tavsiyesini dinamik olarak üretmekte kullanılabilir. İkinci olarak, web oturumlarındaki sayfa erişimleri çözümlenir. Web siteleri aynı zamanda, sörf oturumu sırasında gözden geçirme desenine bağlı olarak yerler arası ilişki kurmak için bağlantıların dinamik olarak değişim setlerini gösterir. Ayarlanmış birliktelik kuralı algoritmasının kullanımı ile sık sık birlikte ziyaret edilen ilgili sayfalar bulunabilir. Bu web sayfaları kendi aralarında bir köprü bağlantısına sahip olmayabilirler.

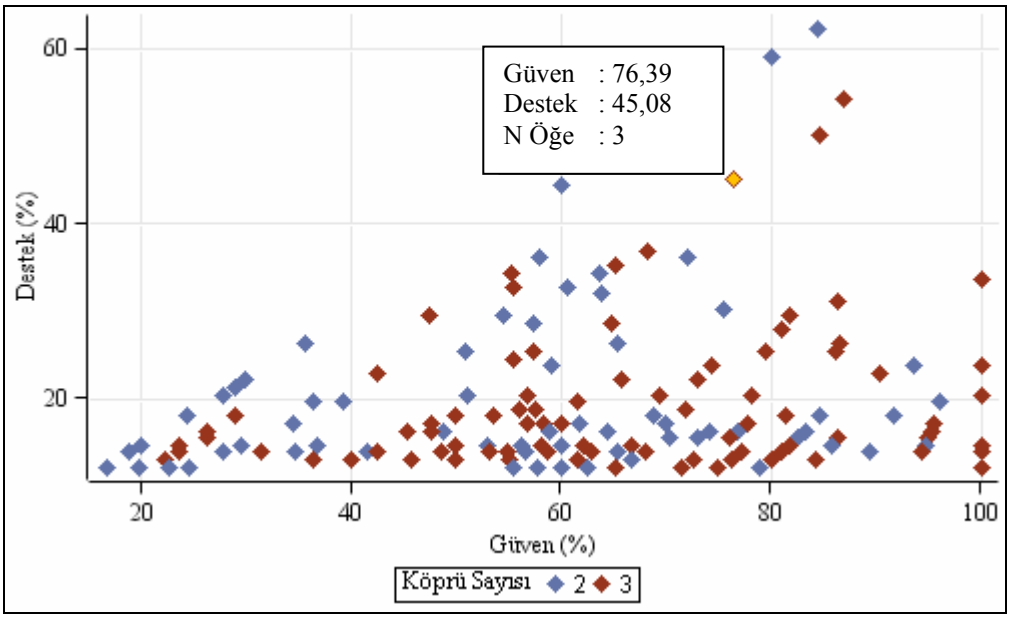
Birliktelik kuralı yöntemi ile yapılan madencilik uygulamalarıyla bir web sitesinin mimarisindeki bozuklukları tespit etmek, geçmişteki eksiklikleri öngörmek mümkündür. Ayrıca, web sitesine ait sunucudaki kütük dosyalarının birliktelik kuralı yöntemiyle incelenmesi sonucunda, beraber kullanılan web sayfaları tespit edilerek, bu sayfalar aynı sunucuya konulabilirler. Böylece, web sitesinde ziyaret eden kullanıcılar sayfalar arasında daha hızlı gezinebileceklerdir.

7.3 Uygulamada Çıkarılan İlginç Kurallar ve Örüntüler

Gerçekleştirilen uygulamanın analiz beş alt kısımdan oluşmaktadır:

7.3.1 İstatistiksel Çizim

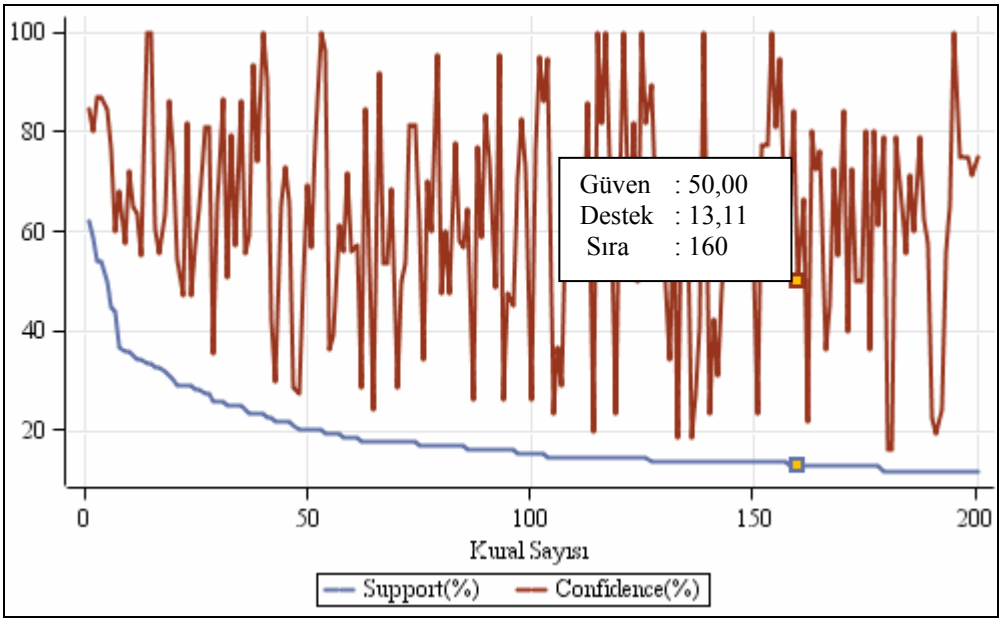
Dağıtık şekilde gösterilen istatistiksel çizim grafiğinde, X ekseni güven değerini, Y ekseni ise destek değerini belirtmektedir. Şekil 7.2'de gösterilen istatistiksel çizim grafiğinde görüldüğü gibi hem birlikteliklerin bulunduğu öğeler hem de sıralı olan öğeler renklendirilmiştir.



Şekil 7.2. İstatistiksel çizim

7.3.2 İstatistiksel Doğru Çizgisi

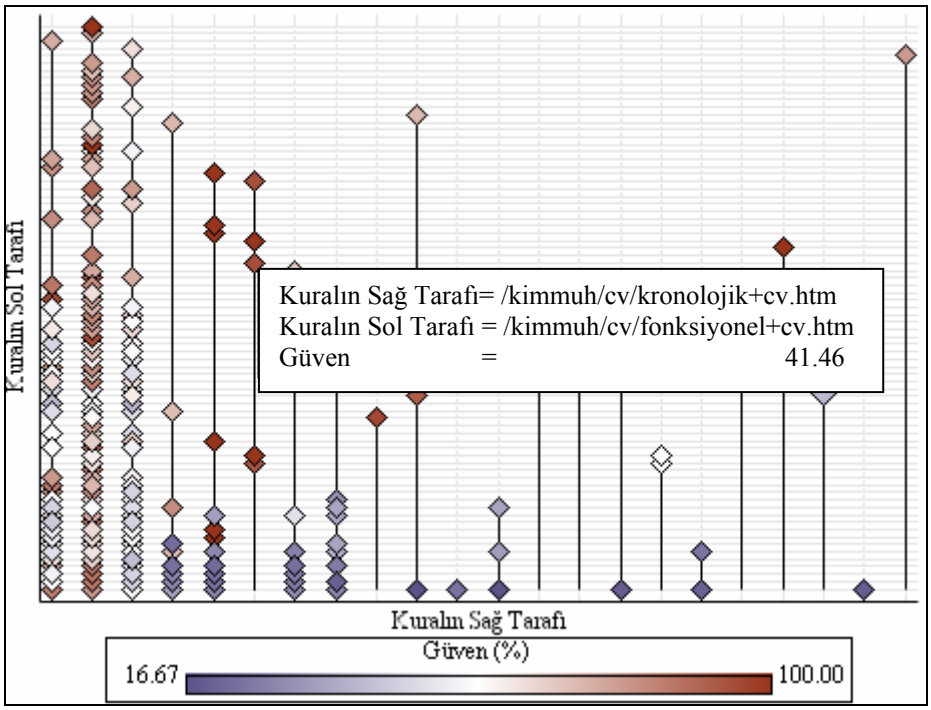
İstatistiksel doğru çizgisinde, Y ekseninde her bir kural için destek ve güven değerlerinin çizgisel grafiği gösterilmektedir. X ekseni ise kural listesi belirtilmektedir. İstatistiksel doğru çizgisi kural sıralama numaraları ile her bir kural için güven ve destek değerlerini grafiksel olarak göstermektedir. Destek yüzdesi, sayfaya istekte bulunan ziyaretçilerin toplam istek sayısına oranıdır. Güven yüzdesi ise, kuralın sağ tarafında bulunan web isteklerinin, kuralın sol tarafında bulunan web istekleriyle beraber olma oranıdır. Örneğin, $A \Rightarrow B$ kuralı dikkate alındığında, destek için bir kullanıcının A ve B öğelerinin her ikisine de sahip olması gerekir. Aynı kural için güven ise, A öğesine sahip olan bir kullanıcının B öğesine de sahip olması durumudur. Gerçekleştirilen uygulamada, ortaya konulan istatistiksel doğru çizgisi Şekil 7.3'de gösterilmektedir.



Şekil 7.3. İstatistiksel doğru çizgisi

7.3.3 Kural Diyagramı

Birliktelik kuralları, sol tarafta bir şart ifadesi ile birlikte sağ tarafta bir ya da daha fazla sonuçları bulunduran deyimler bütünüdür. Şekil 7.4'de görüldüğü üzere, kural matrisi X ekseninde kuralın sol tarafındaki öğeler ve Y ekseninde ise kuralın sağ tarafındaki öğeler gösterilmektedir. Çizimde bir veri göstergesi bir işlem kuralına benzerlik göstermektedir. Kural matrisinde, birliktelik kurallarının sağ ve sol tarafındaki öğeler temeline dayanan kuralları belirtilerek, şekilsel olarak çizilmektedir. Çıkarılan birliktelik kurallarına ait güven değerleri Şekil 7.4'deki grafik üzerinde renkli noktalarla gösterilmektedir.



Şekil 7.4. Kural diyagramı

7.3.4 Kurallar Tablosu

Kurallar tablosu, kullanıcı erişim kütük dosyalarındaki verilerden elde edilen birliktelik kuralları hakkında bilgilerin bulunduğu tablodur. Fırat Üniversitesi web sunucularına ait kullanıcı erişim kütüklerinden birliktelik kuralı keşfi için SAS Base yazılım ortamı kullanıldı. Uygulama çalışmasında kullanılan web öğelerinin birliktelik kuralları için en uzun zincir sayısı 3, işlem frekans değeri ise %2 olarak seçilmiştir. Sonuçlarda görüldüğü gibi, veri kümesinden 200 tane birliktelik kuralı üretilmiştir. Tablo 7.2'de keşfedilmiş birliktelik kurallarının listesi gösterilmektedir. Bu tablo, 11 kolon ve 200 satırdan oluşmaktadır. Birliktelik kurallarına ait detaylı bilgilerin bulunduğu bu tablonun sütunları zincir uzunluğu, işlem sayısı, destek, güven, üretilen birliktelik kuralları, zincir öğeleri, kural sıralaması gibi önemli bilgileri içermektedir. Kurallar tablosundaki değerlerin ve öğelerin bulunduğu her bir sütunun adlandırılması SAS program kodlarındaki etiketlendirme kısmında değiştirilebilmektedir.

Tablo 7.2'de belirtilen zincir sayısı, oluşturulan kurallardaki maksimum sayfa istek sayısıdır. İşlem sayısı, çıkarılan kurallardaki öğelere ilişkin gerçekleşen işlemlerin sayısıdır. Destek değeri, kural için nitelenen işlemlerin toplam işlem sayısına oranıdır. Güven değeri, kuraldaki bütün öğeleri içeren işlem sayısının, kuralın sol tarafındaki bütün öğeleri içeren işlemlerin sayısına oranıdır. Kurallar sütunu ise çıkarılan kurallara ait tam metinleri

içermektedir. Kural sıralaması, her bir kuralı tanımlamak için kullanılan sayısal değerdir. Yani, çıkarılan bütün kuralların sayısal olarak sıralanmasıdır.

Birliktelik kuralı madenciliğinin sonucunda, birlikte ziyaret edilen web sayfa çiftleri bulundu. Yani, iki ya da daha fazla sayfa birlikte gösterilme eğiliminde ise, bu sayfalar pozitif birliktelik kuralı ile tanımlanmaktadır. Örneğin, Tablo 7.2'de görüldüğü gibi, keşfedilen ilk kural */sosyalbil/sonsite/index.htm* $\Rightarrow$ */sosyalbil/sonsite/ilan.htm* şeklindedir. Bu kuralın güven değeri %84.44 ve destek değeri %62.30 olarak bulunmuştur. Bu kuraldaki zincir sayısı ise 2'dir.

Tablo 7.2 (a) Birliktelik kuralları tablosu

Zincir Sayısı	İşlem Sayısı	Destek (%)	Güven (%)	Kurallar	Öğeler Kümesi – 1
2	76	62.30	84.44	/sosyalbil/sonsite/index.htm ==> /sosyalbil/sonsite/ilan.htm	/sosyalbil/sonsite/index.htm
2	72	59.02	80.00	/sosyalbil/sonsite/index.htm ==> /sosyalbil/sonsite/index.htm	/sosyalbil/sonsite/index.htm
2	66	54.10	86.84	/sosyalbil/sonsite/ilan.htm ==> /sosyalbil/sonsite/index.htm	/sosyalbil/sonsite/ilan.htm
3	66	54.10	86.84	/sosyalbil/sonsite/index.htm ==> /sosyalbil/sonsite/ilan.htm ==> /sosyalbil/sonsite/index.htm	/sosyalbil/sonsite/index.htm
3	61	50.00	84.72	/sosyalbil/sonsite/index.htm ==> /sosyalbil/sonsite/index.htm ==> /sosyalbil/sonsite/index.htm	/sosyalbil/sonsite/index.htm
3	55	45.08	76.39	/sosyalbil/sonsite/index.htm ==> /sosyalbil/sonsite/index.htm ==> /sosyalbil/sonsite/ilan.htm	/sosyalbil/sonsite/index.htm
2	54	44.26	60.00	/sosyalbil/sonsite/index.htm ==> /kimmuh/cv/?rnek+cv.htm	/sosyalbil/sonsite/index.htm
3	45	36.89	68.18	/sosyalbil/sonsite/ilan.htm ==> /sosyalbil/sonsite/index.htm ==> /sosyalbil/sonsite/index.htm	/sosyalbil/sonsite/ilan.htm
2	44	36.07	72.13	/eemuh/default.asp ==> /eemuh/default.asp	/eemuh/default.asp
2	44	36.07	57.89	/sosyalbil/sonsite/ilan.htm ==> /sosyalbil/sonsite/ilan.htm	/sosyalbil/sonsite/ilan.htm
3	43	35.25	65.15	/sosyalbil/sonsite/ilan.htm ==> /sosyalbil/sonsite/index.htm ==> /sosyalbil/sonsite/ilan.htm	/sosyalbil/sonsite/ilan.htm
2	42	34.43	63.64	/kimmuh/cv/?rnek+cv.htm ==> /sosyalbil/sonsite/index.htm	/kimmuh/cv/?rnek+cv.htm
3	42	34.43	55.26	/sosyalbil/sonsite/index.htm ==> /sosyalbil/sonsite/ilan.htm ==> /sosyalbil/sonsite/ilan.htm	/sosyalbil/sonsite/index.htm
2	41	33.61	100.00	/metal/saat.asp ==> /metal/saat.asp	/metal/saat.asp
3	41	33.61	100.00	/metal/saat.asp ==> /metal/saat.asp ==> /metal/saat.asp	/metal/saat.asp
2	40	32.79	60.61	/kimmuh/cv/?rnek+cv.htm ==> /kimmuh/cv/?rnek+cv.htm	/kimmuh/cv/?rnek+cv.htm
3	40	32.79	55.56	/sosyalbil/sonsite/index.htm ==> /sosyalbil/sonsite/index.htm ==> /kimmuh/cv/?rnek+cv.htm	/sosyalbil/sonsite/index.htm
2	39	31.97	63.93	/eemuh/default.asp ==> /sosyalbil/sonsite/index.htm	/eemuh/default.asp
3	38	31.15	86.36	/sosyalbil/sonsite/ilan.htm ==> /sosyalbil/sonsite/ilan.htm ==> /sosyalbil/sonsite/index.htm	/sosyalbil/sonsite/ilan.htm
2	37	30.33	75.51	/yapi/default.asp ==> /sosyalbil/sonsite/index.htm	/yapi/default.asp
2	36	29.51	47.37	/sosyalbil/sonsite/ilan.htm ==> /kimmuh/cv/?rnek+cv.htm	/sosyalbil/sonsite/ilan.htm
2	36	29.51	54.55	/kimmuh/cv/?rnek+cv.htm ==> /sosyalbil/sonsite/ilan.htm	/kimmuh/cv/?rnek+cv.htm
3	36	29.51	81.82	/eemuh/default.asp ==> /eemuh/default.asp ==> /eemuh/default.asp	/eemuh/default.asp
3	36	29.51	47.37	/sosyalbil/sonsite/index.htm ==> /sosyalbil/sonsite/ilan.htm ==> /kimmuh/cv/?rnek+cv.htm	/sosyalbil/sonsite/index.htm
2	35	28.69	57.38	/eemuh/default.asp ==> /kimmuh/cv/?rnek+cv.htm	/eemuh/default.asp
3	35	28.69	64.81	/sosyalbil/sonsite/index.htm ==> /kimmuh/cv/?rnek+cv.htm ==> /sosyalbil/sonsite/index.htm	/sosyalbil/sonsite/index.htm
3	34	27.87	80.95	/kimmuh/cv/?rnek+cv.htm ==> /sosyalbil/sonsite/index.htm ==> /sosyalbil/sonsite/ilan.htm	/kimmuh/cv/?rnek+cv.htm
3	34	27.87	80.95	/kimmuh/cv/?rnek+cv.htm ==> /sosyalbil/sonsite/index.htm ==> /sosyalbil/sonsite/index.htm	/kimmuh/cv/?rnek+cv.htm
2	32	26.23	35.56	/sosyalbil/sonsite/index.htm ==> /eemuh/default.asp	/sosyalbil/sonsite/index.htm
2	32	26.23	65.31	/yapi/default.asp ==> /sosyalbil/sonsite/ilan.htm	/yapi/default.asp

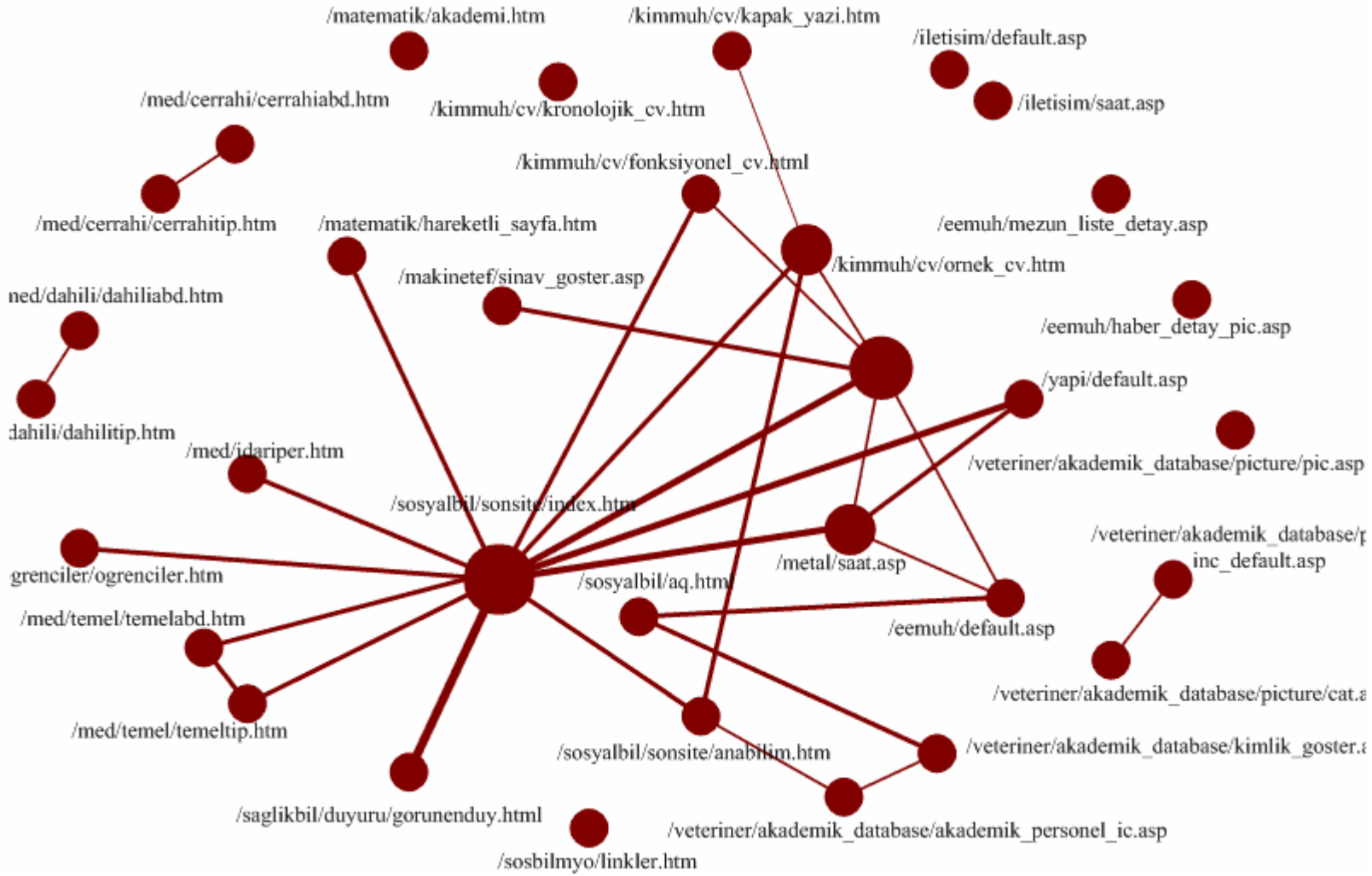
Tablo 7.2 (b) Birliktelik kuralları tablosu

Öğeler Kümesi – 2	Öğeler Kümesi – 3	Kural	Sol Taraf	Sağ Taraf
/sosyalbil/sonsite/ilan.htm		1	/sosyalbil/sonsite/index.htm	/sosyalbil/sonsite/ilan.htm
/sosyalbil/sonsite/index.htm		2	/sosyalbil/sonsite/index.htm	/sosyalbil/sonsite/index.htm
/sosyalbil/sonsite/index.htm		3	/sosyalbil/sonsite/ilan.htm	/sosyalbil/sonsite/index.htm
/sosyalbil/sonsite/ilan.htm	/sosyalbil/sonsite/index.htm	4	/sosyalbil/sonsite/index.htm ==> /sosyalbil/sonsite/ilan.htm	/sosyalbil/sonsite/index.htm
/sosyalbil/sonsite/index.htm	/sosyalbil/sonsite/index.htm	5	/sosyalbil/sonsite/index.htm ==> /sosyalbil/sonsite/index.htm	/sosyalbil/sonsite/index.htm
/sosyalbil/sonsite/index.htm	/sosyalbil/sonsite/ilan.htm	6	/sosyalbil/sonsite/index.htm ==> /sosyalbil/sonsite/index.htm	/sosyalbil/sonsite/ilan.htm
/kimmuh/cv/?rnek+cv.htm		7	/sosyalbil/sonsite/index.htm	/kimmuh/cv/?rnek+cv.htm
/sosyalbil/sonsite/index.htm	/sosyalbil/sonsite/index.htm	8	/sosyalbil/sonsite/ilan.htm ==> /sosyalbil/sonsite/index.htm	/sosyalbil/sonsite/index.htm
/eemuh/default.asp		10	/eemuh/default.asp	/eemuh/default.asp
/sosyalbil/sonsite/ilan.htm		9	/sosyalbil/sonsite/ilan.htm	/sosyalbil/sonsite/ilan.htm
/sosyalbil/sonsite/index.htm	/sosyalbil/sonsite/ilan.htm	11	/sosyalbil/sonsite/ilan.htm ==> /sosyalbil/sonsite/index.htm	/sosyalbil/sonsite/ilan.htm
/sosyalbil/sonsite/index.htm		12	/kimmuh/cv/?rnek+cv.htm	/sosyalbil/sonsite/index.htm
/sosyalbil/sonsite/ilan.htm	/sosyalbil/sonsite/ilan.htm	13	/sosyalbil/sonsite/index.htm ==> /sosyalbil/sonsite/ilan.htm	/sosyalbil/sonsite/ilan.htm
/metal/saat.asp		14	/metal/saat.asp	/metal/saat.asp
/metal/saat.asp	/metal/saat.asp	15	/metal/saat.asp ==> /metal/saat.asp	/metal/saat.asp
/kimmuh/cv/?rnek+cv.htm		16	/kimmuh/cv/?rnek+cv.htm	/kimmuh/cv/?rnek+cv.htm
/sosyalbil/sonsite/index.htm	/kimmuh/cv/?rnek+cv.htm	17	/sosyalbil/sonsite/index.htm ==> /sosyalbil/sonsite/index.htm	/kimmuh/cv/?rnek+cv.htm
/sosyalbil/sonsite/index.htm		18	/eemuh/default.asp	/sosyalbil/sonsite/index.htm
/sosyalbil/sonsite/ilan.htm	/sosyalbil/sonsite/index.htm	19	/sosyalbil/sonsite/ilan.htm ==> /sosyalbil/sonsite/ilan.htm	/sosyalbil/sonsite/index.htm
/sosyalbil/sonsite/index.htm		20	/yapi/default.asp	/sosyalbil/sonsite/index.htm
/kimmuh/cv/?rnek+cv.htm		22	/sosyalbil/sonsite/ilan.htm	/kimmuh/cv/?rnek+cv.htm
/sosyalbil/sonsite/ilan.htm		21	/kimmuh/cv/?rnek+cv.htm	/sosyalbil/sonsite/ilan.htm
/eemuh/default.asp	/eemuh/default.asp	23	/eemuh/default.asp ==> /eemuh/default.asp	/eemuh/default.asp
/sosyalbil/sonsite/ilan.htm	/kimmuh/cv/?rnek+cv.htm	24	/sosyalbil/sonsite/index.htm ==> /sosyalbil/sonsite/ilan.htm	/kimmuh/cv/?rnek+cv.htm
/kimmuh/cv/?rnek+cv.htm		25	/eemuh/default.asp	/kimmuh/cv/?rnek+cv.htm
/kimmuh/cv/?rnek+cv.htm	/sosyalbil/sonsite/index.htm	26	/sosyalbil/sonsite/index.htm ==> /kimmuh/cv/?rnek+cv.htm	/sosyalbil/sonsite/index.htm
/sosyalbil/sonsite/index.htm	/sosyalbil/sonsite/ilan.htm	27	/kimmuh/cv/?rnek+cv.htm ==> /sosyalbil/sonsite/index.htm	/sosyalbil/sonsite/ilan.htm
/sosyalbil/sonsite/index.htm	/sosyalbil/sonsite/index.htm	28	/kimmuh/cv/?rnek+cv.htm ==> /sosyalbil/sonsite/index.htm	/sosyalbil/sonsite/index.htm
/eemuh/default.asp		29	/sosyalbil/sonsite/index.htm	/eemuh/default.asp
/sosyalbil/sonsite/ilan.htm		30	/yapi/default.asp	/sosyalbil/sonsite/ilan.htm

7.3.5 Köprü Bağlantı Grafiği

Köprü bağlantı grafiği, bulunan birliktelik sonuçlarının düğümler ve köprüler kullanılarak, şekilsel olarak gösterilmesidir. Bir düğümün rengi ve büyüklüğü, kurallar veri kümesindeki işlemlerin sayısını göstermektedir. Boyutları geniş olan düğümlerin işlem sayıları kendisinden daha küçük olan düğümlerin işlem sayılarına göre daha büyüktür. Bir köprü çizgisinin kalınlığı ve rengi bir kuralın güven seviyesini gösterir. Web kütük verilerinden çıkarılan köprü bağlantı grafiğindeki köprü çizgilerinin kalın olması, o kuralların daha yüksek güven değerine sahip olduğunu belirtmektedir.

Köprü bağlantı grafiği, web sunucularda biriken kullanıcı erişim kütüklerindeki veriler kullanılarak oluşturulmaktadır. Sitenin web sayfaları düğüm olarak, bu düğümler arasındaki çizgiler de köprü olarak sunulmaktadır. Düğümler arasındaki köprüler, kaynak web sayfasından hedef web sayfasına kadar yapılan işlemlerin sayısına bağlı olarak belirtilmektedir. Bir köprü bağlantı grafiği kurallar veri kümesindeki verilerin grafiksel sunumudur. Köprü bağlantı grafiğinin, gerekli olan köprüleri kullanarak sonuçları yorumlamak için çok anlamlı, ilginç ve kolay bir görünümü vardır. Şekil 7.5'de birliktelik madenciliği analizinin bir grafiği görülmektedir. Grafikte görünen en geniş düğüm */sosyalbil/sonsite/index.htm* web sayfasıdır. Çünkü web sayfasını kullanan kullanıcılar arasındaki en popüler web sayfasıdır.



Şekil 7.5. Köprü bağlantı grafiği

7.4 Bulguların Değerlendirilmesi

Birliktelik kuralı analizi, web sitesinde kullanıcılar tarafından ziyaret edilen birbirleriyle ilişkili web sayfa çiftlerini tespit ederek web sitesinin geliştirilmesi için web tasarımcılarına yardım sağlayabilmektedir. Çıkarılan birliktelik kuralları, üniversite web sitesini anlama ve kavrama da çok yararlı bilgiler içermektedir. Fırat Üniversitesi web sitesinin yeniden yapılandırılmasında ve organizasyonunda, çıkarılan birliktelik kurallarından faydalanmak site tasarımında yararlı olacaktır.

Bu çalışmada gerçekleştirilen birliktelik kuralları madenciliği uygulamasında önerilen yöntem ile yüksek düzeyde anlamlı sonuçlar alındı. Bu uygulamada, kullanılan web kullanıcı erişim kütükleri Bölüm 5'te kullanılan veritabanı ile aynıdır. Uygulamanın ön işlem sürecindeki veri alanların seçimi işleminde, birliktelik kurallarında istenilen girdi verilerine göre uygun alanlar seçildi. Önerilen ve gerçekleştirilen birliktelik kuralları uygulaması sonucunda elde edilen bulgular ve sonuçların değerlendirilmesinde web sitesine ilişkin aşağıda bilgiler çıkarıldı.

- Ziyaretçilerin web sitesinde gezinirken birlikte kullandıkları sayfalar çıkarılarak, ziyaretçilerin sitede ilgi gösterdikleri sayfa grupları tespit edildi. Önerilen bu yöntem ile yüksek düzeyde destek ve güven değerlerine sahip web sayfası çiftlerinin tespitinin yapıldı. Site kullanımı için benzer özellikleri taşıyan sayfaları kümelemek ve her bir kümedeki kullanıcılara uygun hizmet vermek mümkündür.
- Ziyaretçilerin birlikte gezindiği sayfalara ilişkin çıkarılan kurallara ait güven ve destek değerleri dikkate alınarak yoğun gezinilen sayfalardaki kullanıcı ilgileri tespit edilebilir. Örneğin, ticari web sitelerinde alışveriş yapan ziyaretçilerin ilgi duyduğu ürünlere göre sınıflandırma yapılabilir. Bu ziyaretçinin bir sonraki ziyaretinde alabileceği ürün tahmini yapılarak kullanıcılar yönlendirilebilir. Bu çalışma ile şirketler ürünlerin satışına yönelik müşteri analizi yaparak, satışları ciddi boyutta arttırabilir.
- Birliktelik kuralı yöntemi ile yapılan web kütük uygulamasında çıkarılan kuralları web sitesinin kullanımına yönelik bilgiler içermektedir. Bu sonuç bilgileri web sitesindeki köprü bağlantılarına ilişkin çıkarılan kuralların güven ve destek değerlerini, işlem sayıları ve zincir sayılarını vermektedir. Bu sonuçlar, analiz yapan kişiye web sitesinin genel kullanımına yönelik bir fikir oluşturacaktır.
- Ziyaretçilerin gezindiği sayfaları şekilsel olarak modellemek için web köprü bağlantı grafiği oluşturuldu. Ziyaretçi yoğunluğuna göre ağırlıklandırılmış köprü bağlantıları arasındaki çizgilerin kalınlığı o köprüler arasındaki geçişin yoğunluğuna göre değişim

göstermektedir. Bu grafik ile sitede birlikte kullanılan köprü bağlantı grupları dikkate alınarak web sitesinin yapısında iyileştirmeler yapılabilir.

Birliktelik kuralları uygulaması sonucundaki bulgular, web kullanıcı erişim kütüklerinin analizi ile elde edilen bilgilerin kullanılabilirliği, hem web sitesi hem de kullanıcılar açısından ortaya konuldu. Web sitesindeki köprü bağlantıları incelenerek, sitenin geliştirilebilmesi ve iyileştirilebilmesi için web tasarımcılarına ve web yöneticilerine destek sağlanılabileceği ortaya konuldu. Ayrıca, Bölüm 5'te aynı web kütükleri için yapılan yol analizi uygulamasındaki bulgular ile bu bölümdeki sonuçlar web sitenin gelişimi için farklı noktalara işaret etmektedir. Web tasarımcılar için bu iki yöntemdeki sonuçlar ayrı ayrı irdelenerek web sitesinin yeniden oluşturulmasında yeni fikirler ve yardımlar sağlayacaktır. Bu pozitif durum, web sitesini ziyaret eden kullanıcılara da direk olarak yansıyacak ve kullanıcı memnuniyetinin artmasını sağlayacaktır.

8. SONUÇLAR VE DEĞERLENDİRME

İnternetin hızlı bir şekilde yaygınlaşması ve kullanıcı sayısının artmasıyla beraber bu alanda yapılan bilimsel çalışmalarında hızlıca arttığı görülmektedir. Web kullanım madenciliği internet web sitesinde gezinen ziyaretçilerin web üzerinde bıraktıkları izlerden yola çıkarak örüntü keşfi ve analizi, web sitesi üzerinde gerçekleşen olayların ve kullanıcı davranışlarının incelenmesi gibi konularda kanıtların sağlanabilmesi için araştırma gruplarının ve ticari şirketlerin çalıştıkları önemli bir araştırma alanıdır. Tüm bu çalışmalarda web sunucular üzerinde kaydedilen ve saklanan metin tabanlı kullanıcı erişim kütükleri kullanılmaktadır.

Web sunucu üzerinde kayıtları tutulan erişim kütük dosyaları, karmaşık, düzensiz ve herhangi bir anlam ifade etmeyen metinlerdir. Bu erişim kütük dosyalarındaki veriler içerisinden, analiz değeri olmayan ilişkisiz alanların arındırılması, ayıklanması ve belirli bir düzene getirilerek veri tabanına aktarılması gibi işlemler karmaşık ve çok zor olan bir ön işlem sürecidir. Bu süreçte önemli olan veri kaynağından alınan verilerin işlenirken değerlerin ve bilgilerin orijinalliğini koruması, anlamlı ve doğru sonuçların üretilebilmesi için başarılı veri tablolarının oluşturulmasıdır. Daha sonra ön işlem aşamasından geçirilmiş bu verilerden uygun alanlar ilişkilendirilerek, anlamlı bilgiler, ilginç kurallar ve istatistiksel bilgilerin çıkarılması gerekmektedir.

Bu tez çalışmasında, öncelikle web madenciliği kavramına değinilmiş, web kullanım madenciliğinin uygulama aşamaları ve örüntü keşfi yöntemleri detaylı olarak alt bölümler halinde verilmiştir. Tezin uygulamalı çalışmalar bölümünde ise, öncelikle metin tabanlı web erişim kütüklerinin temizlenmesi ve ön işlem süreci için geliştirilen başarımı yüksek bir yöntem önerilmiştir. Geliştirilen ve önerilen yöntemde kullanıcı erişim kütüklerinin temizlenmesi ve ön işlem süreci başarılı bir şekilde gerçekleştirilmiştir. Ön işlem aşamasından geçirilmiş web kullanıcı erişim kütüklerine birbirinden farklı üç yöntem uygulanarak, anlamlı ve kullanılabilir bilgilerin elde edilebildiği gösterilmiştir. Uygulama sonucunda elde edilen anlamlı bulgular ve sonuçların web sitesi açısından kullanılabilirliği gösterilmiştir. Uygulamalarda Fırat Üniversitesi web sunucusundan alınan gerçek erişim kütükleri kullanılmıştır. Yapılan bu uygulamalar ile web kullanım madenciliğinin sadece teorik değil aynı zamanda pratikte uygulanabilir bir bilimsel araştırma alanı olduğu bir kez daha ispatlanmıştır.

8.1 Sonuçların Değerlendirilmesi

1. Web kullanım madenciliğinin ön işlem süreci için SAS Base yazılım ortamında JAVA tabanlı SAS kodlarıyla yeni bir yöntem geliştirildi. Önerilen yöntemde, erişim

kütüklerindeki gürültülü ve bozuk alanların giderilmesi, karmaşık yapının düzenlenmesi, SAS yazılım veri tabanına aktarılması ve istenilen alan bilgilerinin ilişkilendirilmesi gibi işlemlerde ve ayrıca büyük boyutlu kütük dosyalarının kısa sürede işlenebilmesi yönünden tatmin edici sonuçlar alındı.

2. SAS veri tabanına aktarılmış düzenli erişim kütüklerinden hazır fonksiyonlarla istenilen sorgulamaların kolayca yapılarak, veri analizi yapılmıştır. ECLF biçimindeki web erişim kütüklerinde bulunan 22 alan bilgisi bu aşamada birbirlerinden ayrıştırıldı ve veri tabanına aktarıldı. Web erim kütüklerinden istenilen anlamlı bilgilerin keşfi için bu alan bilgilerinden birbirleriyle ilişkili olanlar veri tabanından alınarak, istenilen bilgilerin çıkarılabilmesi için kullanılacak alan verileri giriş verileri haline getirildi.
3. Literatürde veri madenciliği yöntemlerini kullanarak web kullanıcı erişim kütüklerinden anlamlı ve ilginç bilgilerin çıkarılmasına yönelik çalışmalar bulunmaktadır. Ancak, daha çok tarım ve zirai alanda yaygın olarak kullanılan hiyerarşik bir çoklu regresyon analizi olan yol analizi ilk defa web erişim kütüklerine uygulanmıştır.
4. Ön işlem aşamasından geçirilmiş web kullanıcı erişim kütüklerine *Yol Analizi* yöntemi uygulanarak, web sayfalarında gezinen ziyaretçilerin web sitesinde izlediği yollar çıkarıldı. Ziyaretçilerin izlediği yollara ait tespit edilen güven ve destek değerleri dikkate alınarak yoğun gezinilen sayfalardaki kullanıcıların ilgi alanlarına yönelik bilgiler tespit edildi. Böylece, hem kullanıcı memnuniyetini, hem de web sitesinin kullanılabilirliğini arttıracak temel bilgiler çıkarıldı.
5. *İstatistiksel Analiz* ile üç aylık bir zaman dilimi içerisindeki web kullanıcı erişim kütükleri incelenerek, web sitesinin içeriği, yapısı ve kullanımına yönelik anlamlı istatistikî bilgiler çıkarıldı. Çıkarılan bu anlamlı bilgiler ışığında web sitesinin geliştirilmesi ve iyileştirilmesi için raporlama bilgileri de çıkarıldı. Bu bilgilerin analizi ile Fırat Üniversitesi web sitesinin ve sunucusunun başarımını artırmak için çözüm önerileri oluşturuldu.
6. Birliktelik kuralı yöntemi ile yapılan web kütük uygulamasında, web sitesinde aynı oturumda birlikte kullanılan web sayfaları için birliktelik kuralları bilgisi çıkarıldı. Bu bilgiler web sitesindeki sayfalar arası köprü bağlantılarına ilişkin çıkarılan kuralların güven ve destek değerlerini, işlem sayıları ve zincir sayılarını vermektedir.
7. Bölüm 5 ve Bölüm 7'de aynı web erişim kütüklerine farklı yöntemler (yol analizi ve birliktelik kuralı) uygulanarak, web sitesinin geliştirilmesine ve iyileştirilmesine katkı sağlayacak bilgiler ve öneriler iki yöntemde de birbirini desteklemiştir.

8.2 Öneriler

1. Bu tez çalışmasında web kullanım madenciliğinin ön işlem süreci için geliştirilen ve önerilen yöntemin tüm aşamalarının içinde bulunacağı, kullanımı kolay bir kullanıcı yazılım arabirimi geliştirilebilir.
2. İleriye dönük büyük web madenciliği yazılım projeleri geliştirebilmek için veri madenciliği uzmanları, bilgisayar ve web programlama uzmanlarının yanı sıra sistem yöneticilerinden oluşan üç kollu bir araştırma ekibinin birlikte çalışabileceği bir organizasyonun oluşturulması gereği görülmektedir.
3. Büyük ve özel ticari şirketler, kamu kurum ve kuruluşlar, üniversiteler ve diğer eğitim kurumları, web sitelerini etkin kullanan tüm organizasyonlar kendi web sitelerini profesyonel anlamda geliştirmek, iyileştirmek ve bu anlamda kendi kurumlarına kalite, ciddiyet ve ekonomik kazanç sağlamak için web madenciliği analizi yapmak zorundadırlar. Yapılan uygulamalar, web sitelerindeki madencilik çalışmalarının web sitesinin geliştirilmesi, iyileştirilmesi ve yeniden organizasyonu için büyük bir ihtiyaç olduğunu göstermektedir.
4. *Yol Analizi* yöntemi uygulanarak, ziyaretçilerin izlediği yollara ait tespit edilen güven ve destek değerleri ile ziyaretçilerin site içerisinde ilgi duydukları sayfalar belirlenerek, site kullanımı için benzer özellikleri taşıyan kullanıcılar kümelenebilir ve her bir kümedeki kullanıcıların isteğine yönelik daha uygun hizmet verilebilir.
5. Web madenciliği alanında yapılacak bilimsel çalışmalar ile internet suçlarına yönelik izlemeler yapılabilir, güvenlik açıklarını kapatmaya yönelik önlemler alınabilir.

8.3 Yayınlar

Bu tez çalışması 2 adet yurt dışı uluslararası dergi yayını, [123, 124], 2 adet yurtiçi uluslararası dergi yayını [105, 134] ve 2 adet ulusal dergi yayını [1, 18] olmak üzere toplam 6 adet makale ile sonuçlanmıştır.

KAYNAKLAR

1. Daş, R., Türkoğlu, İ., Poyraz, M. (2007). Web Kayıt Dosyalarından İlginç Örüntülerin Keşfedilmesi", Fırat Üniversitesi Fen ve Mühendislik Bilimleri Dergisi, 19(4), 493-503, Elazığ.

2. Etzioni, O. (1996). The World Wide Web: Quagmire or gold mine? Communication of the ACM, 39 (11), 65–68.

3. Gezer, M., Erol, Ç. ve Gülseçen, S., (2007). Bir Web Sayfasının Web Madenciliği ile Analizi, AB–2007 Akademik Bilişim Konferansı, 31 Ocak–2 Şubat 2007, Kütahya.

4. Dean,N.(Ed.)., (2000). OCLC Researchers measure the World Wide Web." *Online Computer Library Center (OCLC)* Newsletter.

5. Daş, B., (2006). Tip–2 Bulanık Mantık Tabanlı Veri Madenciliği Yöntemi ile Zaman Serisi Tahmini, Fırat Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 76s. Elazığ.

6. Cooley, R., Mobasher, B., Srivastava, J. (1997). Web Mining: Information and Pattern Discovery on the World Wide Web", In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), 558 – 567, USA.

7. Chen L., Sycara K. (1998). WebMate: A Personal Agent for Browsing and Searching", The Second International Conference on Autonomous Agents, ACM.

8. Srivastava, J., Cooley, R., Deshpande, M., and Tan,P. (2000). Web Usage Mining: Discovery and Applications of Usage Patterns From Web Data. SIGKDD Exploartions. (2)1, 1–12.

9. Cooley, R., Mobasher, B., and Srivastava, J., (1999). Data Preparation for mining World Wide Web Browsing Patterns, Knowledge and Information Systems 1, 1–27.

10. Araya, S., Silva, M., and Weber, R., (2004). A Methodology for web usage mining and its applications to target group identification", Fuzzy sets and systems 148, 139–152.

11. Internet: Open Directory Project, http://dmoz.org, Erişim tarihi: Aralık 2006.

12. Uğur, A., Kınacı, A.C. (2006). Yapay Zekâ Teknikleri ve Yapay Sinir Ağları Kullanılarak Web Sayfalarının Sınıflandırılması, XI. Türkiye'de Internet Konferansı, (21–23 Aralık 2006) Bildirileri, Ankara.

13. Uğuz, H., Kodaz, H., Çomak, E., Baykan, Ö.K. (2003). Apriori Algoritması Kullanılarak Web Kullanım Madenciliği Yönteminin Web Log Kayıtlarına Uygulanması. IJCI Proceeding of International Conference on Signal Processing, ISSN 1304–2386, (1)2.

14. Uğuz, H., Kodaz, H., Saraçoğlu, R., Baykan, Ö.K. (2003). Genetik Algoritmalar Kullanılarak Web Kullanım Madenciliği Yönteminin Sistem Log Kayıtlarına Uygulanması, International XII. Turkish Symposium on Artificial Intelligence and Neural Networks – TAINN 2003, T–1, s. 45 – 47.

15. İşeri, İ. (2005). Web Günlüğünden Zaman Sınırlı Bulanık Bağıntı Kuralları ve Sıralı Örüntülerin Çıkarılması", Fırat Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 50 s, Elazığ.

16. Şakiroğlu, A.M., Tuğ, E., Bulun, M. (2003). Web Log Dosyalarından Genetik Algoritma Yöntemiyle Sıralı Erişimlerin Tespit Edilmesi, Türkiye Bilişim Derneği 20. Bilişim Kurultayı.

17. Tuğ, E., Şakiroğlu, A.M., Arslan, A. (2006). Automatic discovery of the sequential accesses from web log data files via a genetic algorithm. Knowledge-Based Systems 19, 180–186.

18. Daş, R., Türkoğlu, İ., Poyraz, M. (2006). Genetik Algoritma Yöntemiyle Internet Erişim Kayıtlarından Bilgi Çıkarılması. Sakarya Üniversitesi Fen Bilimleri Enstitüsü Dergisi, Cilt (10)2, 67–72, Sakarya.

19. Carus, A. ve Mesut, A. (2005). Web Kullanım Madenciliği Uygulaması, II. Mühendislik Bilimleri Genç Araştırmacılar Kongresi–MBGAK 2005, 17–19 Kasım 2005, İstanbul.

20. Belen, E., Özgür, Ç., Özakar, B. (2003). WALA: Web Erişim Kütük Araştırmacısı (Web Access log Analyser). (inet-tr'03) IX. Türkiye'de Internet Konferansı, (11–13 Aralık 2003) Bildirileri, İstanbul.

21. Özakar, B., Püskülcü, H. (2002). Web içerik ve web kullanım madenciliği tekniklerinin entegrasyonu ile oluşmuş bir veri tabanından nasıl yararlanılabilir?

22. Iocchi, L. (1999). The Web OEM approach to Web Information Extraction. Journal of Network and Computer Applications, (22), 259–269.

23. Internet: Amazon Web Sitesi, http://www.amazon.com, Erişim tarihi: Şubat 2007.

24. Habegger, B., Quafafou, M., (2004). Web services for information extraction from the Web, Web Services, IEEE International Conference on 6–9 July 04 Page(s):279–286

25. Takcı, H. ve Soğukpınar, İ., (2002). Kütüphane Kullanıcılarının Erişim Örüntülerinin Keşfi, Bilgi Dünyası, 3(1), 12–26.

26. Takcı, H., Soğukpınar, İ. (2002). Kütüphane Kullanıcılarının Erişim Desenlerinin Keşfi. Akademik Bilişim'02 Konferansı, Selçuk Üniversitesi, Konya, (6–8 Şubat 2002).

27. Internet: eWebLog Analyzer, http://www.esoftys.com, Erişim tarihi: Ocak 2008.

28. Internet: Funnel Web Analyzer, http://www.quest.com/, Erişim tarihi: Ocak 2008.

29. Internet: Megaputer Web Analyticst, http://www.megaputer.com/products/wa/index.php3, Erişim tarihi: Ocak 2008.

30. Internet: NetIQ Web Trends Log Analyzer, http://www.netiq.com, Erişim tarihi: Ocak 2008.

31. Internet: Nihuo Web Log Analyzer (NWLA), http://www.nihuo.com ve http://www.loganalyzer.net, Erişim tarihi: Ocak 2008.

32. Internet: Web Log Mixer, http://www.bitstrike.com/, Erişim tarihi: Mayıs 2007.

33. Internet: WebTrends Marketing Web Analytics and Web Statistics, http://www.webtrends.com, Erişim tarihi: Ocak 2008.

34. Internet: Web Log Storming, http://www.datalandsoftware.com, Erişim tarihi: Ocak 2008.

35. Internet: 123Log Storming, http://www.123logstorming.com, Erişim tarihi: Ocak 2008.

36. Internet: Web Log Expert v3.1, http://www.weblogexpert.com, Erişim tarihi: Ocak 2008.

37. Internet: AlterWind Log Analyzer v2.1, http://www.alterwind.com, Erişim tarihi: Ocak 2008.

38. Internet: Deep Log Analyzer, http://www.deep-software.com, Erişim tarihi: Ocak 2008.

39. Internet: Awstat, http://www.awstats.sourceforge.net, Erişim tarihi: Ocak 2008.

40. Internet: Analog, http://www.analog.cx, Erişim tarihi: Ocak 2008.

41. Internet: Webalizer, http://www.mrunix.net/webalizer, Erişim tarihi: Ocak 2008.

42. Internet: SARG, http://sarg.sourceforge.net, Erişim tarihi: Ocak 2008.

43. Guo, D. (2006). Collector Engine System: A Web Mining Tool for E-Commerce. Innovative Computing, Information and Control, First International Conference, (2)1, 632 – 635.

44. Guo, L., Xiang, X., Shi, Y. (2004). Use Web Usage Mining to Asist Online E-Learning Assessment. Proceedings of the IEEE International Conference on Advanced Learning Technoligies (ICALT'04), China.

45. Laffey, J., Ai, J., (2007). Web Mining as a Tool for Understanding Online Learning, Journal of Online Learning and Teaching, 3(2).

46. Luan, J. (2002). Data Mining and Its Applications in Higher Education. New Directions for Institutional Research, 2002, 113 (Spring), pp. 17–36.

47. Carlos G. Marquardt, Karin Becker, Duncan D. Ruiz. (2004). A Pre-processing Tool for Web Usage Mining in the Distance Education Domain. Proceedings of the International Database Engineering and Applications Symposuim (IDEAS'04), IEEE Computer Society.

48. Velezquez, J.D., Yasuda, H., Aoki, T., Weber, R. (2004). A New similarity measure to understand visitor behavior in the web site", IEICE Trans, Inform. Systems E87-D (2), 389–396.

49. Spiliopoulou, M., Pohle, C., Faulstich, L. (2000). Improving the effectiveness of a Website with Web usage mining, Lecture Notes in Computer Science, Berlin: Springer-Verlag, vol.1836, pp. 142–162.

50. Azhar, A. (2005). Web usage mining using appriori algorithm: UUM-learning care portal case", In Proceeding at ICKM 05, UPM.

51. Ed H. Chi, Peter Pirolli, Kim Chen, James Pitkow, (2001). Using Information Scent to Model User Information Needs and Actions on the Web, In Proc of ACM CHI 2001 Conference on Human Factors in Computing Systems, pp. 490–497, ACM Press, April 2001. Seattle, WA.

52. Spiliopoulou, M., Pohle, C. (2001). Data mining for measuring and improving the success of Websites", Journal of Data Mining and Knowledge Discovery, 5(2), 85–114.

53. Woon, Y.K., Ng, W.K., Lim, E.P., (2002). Online and incremental mining of separately-grouped web access logs, Proceedings of the 3rd International Conference on Web Information Systems Engineering (WISE'02), IEEE Computer Society.

54. Wang Bin, Liu Zhijing, (2003). Web Mining Research, Proceedings of the Fifth International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'03), IEEE Computer Society.

55. Xiaozhe Wang, Ajith Abraham, Kate A. Smith, "Intelligent web traffic mining and analysis", Elsevier Science, Journal of Network and Computer Applications 28, 147–165, 2005.

56. Mortazavi-Asl, B. (2001). Discovering and Mining user web-page traversal patterns. Yüksek Lisans Tezi, Simon Fraser University.

57. Oosthuizen, C., Wesson, J., Cilliers, C. (2006). Visual Web mining of Organizatioanl Web Sites. Proceedings of the Information Visualization (IV'06), IEEE Computer Society.

58. Pal, S., Talwar, V., Mitra, P. (2002). Web Mining in soft computing framework: relevance, state of the art and future directions", IEEE Transactions on Neural Networks 13 (5), 1163–1177.

59. Junjie Chen, Wei Liu, "Research for Web Usage Mining Model", International Conference on Computational Intelligence for Modeling Control and Automation – Intelligent Agents, Web Technologies and Internet Commerce, (CIMCA-IAWTIC'06), 2006.

60. Khasawneh, N., Chan, C.C. (2005). Web Usage Mining Using Rough Sets", IEEE Annual Meeting of the North American Fuzzy Information Processing Society – (NAFIPS'05)

61. Kim, Y., Lee, K., (2005). Detecting tables in Web documents. Engineering Applications of Artificial Intelligence 18, 745–757.

62. Kosala, R., Blockeel, H. (2000). Web mining research: a survey", SIGKDD: SIGKDD explorations: newsletter of the special interest group (SIG) on knowledge discovery & data mining, ACM 2(1), 1–15.

63. Liu, L., Chen, J., Song, H. (2002). The Research of Web Mining. Proceedings of the 4th World Congress on Intelligent Control and Automation, June 10–14, Shanghai/China.

64. Chen, J., Liu, W. (2006). Research for Web Usage Mining Model. International Conference on Computational Intelligence for Modelling Control and Automation – Intelligent Agents, Web Technologies and Internet Commerce, (CIMCA-IAWTIC'06).

65. Cooley, R., (2000). Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data, PhD thesis, University of Minnesota, 170s.

66. Desikan, P., Srivastava, J., "Mining Information from Temporal Behaviour of Web Usage", AHPCRC Technical Report TR-2003-121, 2003.

67. Zhang, F., Chang, H.Y. (2002). Research and Development in Web Usage Mining System-Key Issues and Proposed Solutions: A Survey. Proceedings of the First International Conference on Machine Learning and Cybernetics, 986–990, Beijing.

68. Facca, F.M. and Lanzi, P.L., (2005). Mining interesting knowledge from web logs: a survey, Elsevier Science, Data & Knowledge Engineering 53, pp: 225–241.

69. Gündüz, Ş., Adalı, E. (2004). Web kullanıcılarının davranışları için örüntü bulma ve modelleme. İstanbul Teknik Üniversitesi, Mühendislik Dergisi, (3)6, 15–24, İstanbul.

70. Huiying, Z., Wei, L., (2004). An Intelligent Algorithm of Data Pre-processing in Web Usage Mining. Proceedings of the 5th World Congress on Intelligent Control and Automation, June 15–19, IEEE, Hangzhou, China.

71. Zhou, B., Hui, S.C., Fong, A.C.M. (2005). Discovering and Visualizing Temporal-Based Web Access Behavior, International Conference on Web Intelligence (WI'05) - IEEE/WIC/ACM.

72. Zhong, N., Liu, J., Yao, Y. (2003). Web Intelligence, (ISBN: 3–540–44384–3), Springer, 440 p.

73. Joachhims, T., Freitag, D., Mitchell, T. (1997). Webwatcher: A tour guide for the world wide web. In The 15th International Conference on Artificial Intelligence, Nagoya, Japan.

74. Leieberman, H. (1995). Letizia: An agent that assists web browsing. In Proc. Of the 1995 International Joint Conference on Artificial Intelligence, Montreal, Canada.

75. Moore, J., Han, E.H.(Sam), Boley, D., Gini, M., Gross, R., Hastings, K., Karypis, G., Kumar, V., Mobasher, B. (1997). Web page categorization and feature selection using association rule and principal component clustering. In 7th Workshop on Information Technologies and Systems.

76. Mobasher, B., Cooley, R., Srivasta, J. (1999). Creating adaptive web sites through usage-based clustering of URLs. In Knowledge and Data Engineering Workshop.

77. Ngu, D.S.W., Wu, X. (1997). Sitehelper: A localized agent that helps incremental exploration of the world wide web. In 6th International World Wide Web Conference, Santa Clara, CA.

78. Perkowitz, M., Etzioni, O. (1998). Adaptive web sites: Automatically synthesizing web pages. In Fifteenth National Conference on Artifical Intelligence, Madison, WI.

79. Perkowitz, M., Etzioni, O. (1999). Adaptive web sites: Conceptual cluster mining. In Sixteenth International Joint Conference on Artifical Intelligence, Stockholm, Sweden.

80. Buchner, A.G., Baumgarten, M., Anand, S.S, Mulvenna, M.D., Hughes, J.G. (1999). Navigation pattern discovery from internet data. In WEBKDD, San Diego, CA.

81. Buchner, A., Mulvenna, M.D. (1998). Discovery internet marketing intelligence through online analytical web usage mining. SIGMOD Record, 27(4): 54-61.

82. http://support.sas.com/documentation/, (last accessed: 20.01.2008)

83. Internet: SPSS, http://www.spss.com/clementine/, Erişim tarihi: Mayıs 2007.

84. Zaiane, O.R., Win, M., Han, J. (1998). Discoverying web Access patterns and trends by appyling olap and data mining technology on web logs. In Advances in Digital Libraries, pages 19-29, Santa: Barbara, CA.

85. Catledge, L., Pitkow, J. (1995). Characterizing browsing behaviors on the world wide web. Computer Networks and ISDN Systems, 27(6).

86. Pitkow, J. (1997). In search of reliable usage data on the www. In Sixth International World Wide Web Conference, pages 451-463, Santa Clara, CA.

87. Pitkow, J., Kehoe, C.M (1995). Results from the thrid www user survey. The World Wide Web Journal, 1(1).

88. Liu, B. (2007). Web Data Mining: Exploring Hyperlinks, Contents and Usage Data, (ISBN 13: 978-3-540-37881-5), 532p, Springer.

89. Gündüz, Ş. (2003). Recommendation Models for Web Users: User Interest Model Click-Stream Tree. Doktora tezi, İstanbul Technical University, Institute of Science and Technology, 140s.

90. Liu, H., Keselj, V., (2007). Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users' future requests. Data & Knowledge Engineering, (61)2, 304–330.

91. Gündüz, Ş., Veri madenciliği ders notları, www.cs.itu.edu.tr/~gunduz/courses/verimaden Erişim Tarihi: 2007.

92. Ye, Nong (Ed). (2003). The Handbook of Data Mining, Lawrence Erlbaum Associates, Publishers, Mahwah, New Jersey, London.

93. Nabiyev., V.Vasif. (2003). Yapay Zekâ. Seçkin Yayınevi, Ankara.

94. Koutsoupias, N. (2002). Exploring web access logs with correspondence analysis, Proc. The 2nd Hellenic Conference on Artificial Intelligence, Companion Volume, pp. 229-236.

95. Özekes, S., (2003). Veri madenciliği modelleri ve uygulama alanları, İstanbul Ticaret Üniversitesi Dergisi, 3, 65-82.

96. Zaki, M. J. (1999) Parallel and Distributed Association Mining: A Survey, IEEE Concurrency, special issue on Parallel Mechanisms for Data Mining, 7(5), 14-25.

97. Shen, L., Cheng, L., Ford, J., Makedon, F., Megalooikonomou, V., Steinberg, T. (1999). Mining the most interesting web access associations", Proc. The 5th International Conference on Knowledge Discovery and Data Mining (KDD'99), pp.145–154.

98. Tan, P., Kumar, V. (2002). Mining indirect associations in web data, LNAI 2356, Springer, pp. 145–166.

99. Mobasher, B., Dai, H., Luo, T., Sun, Y., Zhu, J. (2000). Combining Web usage and content mining for more effective personalization, Proc. The International conference on E-Commerce and Web Technologies.

100. Yao, Y.Y., Hamilton, H.J., Wang, X. (2000). PagePrompter: an intelligent web agent created using data mining techniques", Technical Report, CS–2000–08, Department of Computer Science, University of Regina.

101. Karypis, G., Han, E., Kumar V. (1999). Chameleon: Hierarchical Clustering Using Dynamic Modeling, IEEE Computer, 68-75.

102. Ramkumar, G.D., Swami, A., (1998). Clustering Data Without Distance Functions, IEEE Bulletin ofthe Technical Committee on Data Engineering, 21(1), 9-14.

103. Internet: WWW Consortium (W3C), http://www.w3.org/, Erişim tarihi: Ocak 2008.

104. http://www.microsoft.com/technet/prodtechnol/windowsserver2003/library/iis/ (Erişim tarihi: 20.01.2008)

105. Daş, R., Türkoğlu, İ., Poyraz, M. (2008). Analyzing of the user access logs of a website using web usage mining method: Example of Firat University", e-Journal of New World Sciences Academy (NWSA), Natural and Applied Sciences, 3(2), 310–320.

106. Michalski, R.S. ve Stepp, R.E. (1983). Learning from observation: Conceptual clustering. R. Michalski, J. Carbonell ve T. Mitchell (eds.). Machine learning: An artificial intelligence approach içinde (c.1, s. 331–363). San Mateo, CA: Morgan Kauffmann Inc.

107. Zhong, N. ve Ohsuga, S. (1994). Discovering concept clusters by decomposing databases. Data and Knowledge Engineering, 12: 223–244.

108. Sever, H., Oğuz, B. (2003). Veri Tabanlarında Bilgi Keşfine Formel Bir Yaklaşım: Kısım II: Eşleştirme Sorgularının Biçimsel Kavram Analizi ile Modellenmesi.. Bilgi Dünyası 4(1), 15–44.

109. Foskett, D.J. (1997). Thesaurus. K.S. Jones ve P. Willet (eds.). Readings in information retrieval içinde (s. 111–134). New York, NY: Morgan Kaufmann Publishers, Inc.

110. Y.Fu, K.Sandhu, M.Shih: A generalization-based approach to clustering of web usage sessions, Proc. WEBKDD'99:Workshop on Web Usage Analysis and User Profiling (1999) pp.21-38.

111. Nasraoui, O., Frigui, H., Joshi, A., Krishnapuram, R. (1999). Mining web access logs using relational competitive fuzzy clustering, Proc. the 8th International Fuzzy Systems Association World Congress.

112. Cadez, I., Heckermain, D., Meek, Smyth, P., White, S. (2000). Visualization of navigation patterns on a web site using model-based clustering, Proc. KDD'00, pp.280-284.

113. Chen, M.S., Park, J.S., Yu, P.S. (1998). Efficient data mining for path traversal patterns, IEEE Transactions on Knowledge and Data Engineering, 10, 209–221.

114. Nanopoulos, A., Manolopoulos, Y. (2000). Finding generalized path patterns for web log data mining, LNCS 1884, pp.215–228.

115. Bunchner, A.G., Baumgarten, M., Anand, S.S., Mulvenna, M.D., Hughes, J.G. (1999). Navigation pattern discovery from internet data. Proc. WEBKDD'99, Workshop on web usage analysis and user profiling, pp.25–30.

116. Chen, R., Sivakumar, K., kargupta, H. (2001). Collective mining of Bayesian networks from distributed heterogeneous data, Knowledge and Information Systems,

117. Borges, J., Levene, M. (1999). Data mining of user navigation patterns, Proc. WEBKDD'99: Workshop on Web usage analysis anad user profiling, pp.92–111.

118. Borges, J., Levene, M. (1999). Heuristic for mining high quality user web navigation patterns, Research Note RN/99/68, Department of Computer Science, University College, London.

119. Deshpande, M., Karypis, G. (2000). Selective Markov models for predicting web page accesses. Technical Report #00-056, University of Minessota.

120. Sahinler, S., Gorgulu, O. (2000). Path Analysis and an application", M.K. University, Journal of Agriculture Faculty, 5(1–2): 87–102.

121. Sirali, R., Kayaalp, T. (1995). Trakya bölgesi arılarının bal verim özelliği ve bu özelliği etkileyen bazı faktörlerin Path analizi yöntemi ile saptanması. Harran Üniversitesi, Ziraat Fakültesi Dergisi, 1 (2): 211-217.

122. Sarukkai, R.R. (1999). Link prediction and path analysis using Markov Chains", In the Proc. of the 9th World Wide Web Conference.

123. Das, R., Turkoglu, I. (2008). Creating meaningful data from web logs for improving the impressiveness of a website by using path analysis method, (Uluslararası hakemli dergiye gönderildi).

124. Das, R., Turkoglu, I. (2008). Discovering of interesting patterns to improve the performance of website usage, (Uluslararası hakemli dergiye gönderildi).

125. Agrawal, R., Imielinski, T., and Swami, A.N. (1993). Mining association rules between sets of items in large databases, ACM SIGMOD'93 International Conference on Management of Data, pp. 207–216, Washington.

126. Battioui, C., (2007). Data mining techniques to analyze a library database, SAS Institute Inc., paper 076-31.

127. Kösehan, Y., Leblebicioğlu, K. (2003). Mayın Tarlası Oluşturma Problemine Genetik Algoritma Yaklaşımı, KHO Savunma Bilimleri Dergisi, Vol. 2, s.34–56.

128. Emel, G.G., Taşkın, Ç., (2002). Genetik Algoritmalar ve Uygulama Alanları", Uludağ Üniversitesi, İktisadi ve İdari Bilimler Fakültesi Dergisi, Cilt XXI, Sayı 1, s.129–152.

129. Gui-Rong Xue, Hua-Jun Zeng, Zheng Chen, Wei-Ying Ma, Chao-Jun Lu. (2002). "Log Mining to Improve the Performance of Site Search", Proceedings of the Third International Conference on Web Information Systems Engineering (Workshops) - (WISEw'02), p.238.

130. Drott, M. (1998). Using web server logs to improve site design", Proc. ACM Conference on Computer Documentation, pp 43–50.

131. Ayaz, R. (2007). Web Madenciliğine Bir Bakış, Yıldız Teknik Üniversitesi, Bilgisayar Mühendisliği, İstanbul.

132. Velezquez, J.D., Yasuda, H., Aoki, T., Weber, R. (2004). A New similarity measure to understand visitor behavior in the web site", IEICE Trans, Inform. Systems E87-D (2), 389–396.

133. Baykal, N. (2005), Bilgisayar Ağları, SAS Bilişim Yayınları, Ankara.

134. Das, R., Turkoglu, I., Poyraz, M. (2007). Analyzing of system errors for increasing a web server performance by using Web Usage Mining", Istanbul University - Journal of Electrical & Electronics Engineering (IU-JEEE), 7(2), 379–386, Istanbul.

135. Srivastava, J., Desikan, P., Kumar, V. (2005), Web Mining: Concepts, Applications and Research Directions, Studies in Fuzziness and Soft Computing 180, 275–307.

136. Cooley, R., Mobasher, B., Srivastava, J. (1999). Grouping Web page references into transactions for mining World Wide Web browsing patterns, Journal of Knowledge and Information Systems, 1(10), 1-13.

137. Di Guo, (2006). Collector Engine System: A Web Mining Tool for E-Commerce. Innovative Computing, Information and Control, First International Conference, (2)1, 632-635.

138. Puntheeranurak, S., Tsuji, H., (2005). Minings Web logs for a Personalized Recommender System, IEEE Computer Society.

139. Albanese, M., Picariello, A., Sansone C., Sansone, L. (2004). A Web Personalization System Based on Web Usage Mining Techniques. 17–22, New York, USA.

140. Eirinaki, M., Vaziargiannis, M. (2003). Web mining for web personalization, ACM Transactions on Internet Technology (TOIT) 3(1), 1–27.

141. Jian-Gua Liu, Wei-ping Wu. (2004). Web Usage Mining for Electronic Business Applications. Proceedings of the Third International Conference on Machine Learning and Cybernetics, (26–29 August 2004), Shanghai.

142. Habegger, B., Quafafou, M.(2004). Building web information extraction tasks. Web Intelligence (WI), IEEE/WIC/ACM International Conference (20–24 Sept. 2004) 349 – 355.

143. Jianhan Zhu, Jun Hong, and John G. Hughes, "Using Markov Chains for Link Prediction in Adaptive Websites", In Proc of ACM SIGWEB Hypertext 2002.

144. Khasawneh, N., Chan, C.C. (2005).Web Usage Mining Using Rough Sets. IEEE Annual Meeting of the North American Fuzzy Information Processing Society – (NAFIPS'05).

145. Khasawneh, N., Chien-Chung, C. (2006). Active User-Based and Ontology-Based Web Log Data Preprocessing for Web Usage Mining, Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'06), IEEE Computer Society.

146. Michael J.A.Berry, Gordon Linoff, (1997). Data Mining Techniques, Published by John Wiley & Sons, Inc. USA.

147. Mitra, S., Acharya, T. (2003). Data Mining: Multimedia, Soft Computing and Bioinformatics, A John Wiley & Sons, Inc. Publication, USA.

148. Mobasher, B., Cooley, R., Srivastava, J. (2000). "Automatic Personalization based on Web Usage Mining", Communications of the ACM, 43(8), 142–151.

149. Zhou, B., Hui, S.C., Chang, (2004). An Intelligent Recommender System using Sequential Web Access Patterns. Proceeedings of the 2004 IEEE Conference on Cybernetics and Intelligent Systems, (1–3 December 2004), Singapore.

150. Spiliopoulou, M., Mobasher, B., Berendt, B., Nakagawa, M. (2003). A framework for evaluation of session reconstruction heuristic in web usage analysis, INFORMS Journal on Computing, 15(2), 171–190.

151. Yew-Kwong Woon, Wee-Keong Ng, Ee-Peng Lim. (2002). Online and Incremental Mining of Separately-Grouped Web Access Logs, Proceedings of the 3rd International Conference on Web Information Systems Engineering (WISE'02), IEEE Computer Society.

152. Zhang, X., Edwards, J., Harding, J., (2007). Personalised online sales using web usage data mining, Computers in Industry.

153. Zaiane, O. R., Xin, M., and Han, J. (1998). Discovering Web access patterns and trends by applying OLAP and data mining technology on Web logs", In Proc. Advances in Digital Libraries Conference (ADL'98), Santa Babara, CA.

ÖZGEÇMİŞ

Resul DAŞ

Fırat Üniversitesi Enformatik Bölümü 23119, Elazığ

Tel: +90-424-237 00 00 # 3152

E.posta: rdas@firat.edu.tr , resuldas@gmail.com

1975 Pertek'de doğdu.

1989 – 1993 Elazığ Merkez Teknik Lisesi, Elektrik bölümünden mezun oldu.

1995 – 1999 Fırat Üniversitesi Teknik Eğitim Fakültesi, Bilgisayar Öğretmenliği bölümünden mezun oldu.

1999 – 2000 T.C. M.E.B Aydın Anadolu Teknik ve Teknik Lisesi'nde bilgisayar öğretmeni olarak görev yaptı.

2000 – 2002 Fırat Üniversitesi, Fen Bilimleri Enstitüsü, Elektronik ve Bilgisayar Eğitimi Anabilim Dalında "*Video Konferans Sistemi Tasarımı*" konulu tez çalışması ile Yüksek Lisans derecesini aldı.

2000 – … Fırat Üniversitesi Rektörlüğü, Enformatik Bölümü'nde öğretim elemanı olarak görev yapmaktadır. Aynı zamanda, Bilgi İşlem Daire Başkanlığı bünyesindeki ağ altyapısı bilişim hizmetleri ile Cisco Ağ Akademisi eğitmenliği görevlerini yürütmektedir.

2002 – ... Fırat Üniversitesi, Fen Bilimleri Enstitüsü, Elektrik – Elektronik Mühendisliği Anabilim Dalında doktora tez çalışması devam etmektedir.

EK – 1

YOL ANALİZİ UYGULAMASININ SAS PROGRAM KODLARI

```
%let Rules = WORK.RULES;
%let pathScore = WORK.PATHSCORE;
%let restrictDs = WORK.RESTRICT;
%let transitionDs = WORK.PROBABILITY;
%let funnelDs = WORK.FUNNELCOUNTS;
%let recountDs = WORK.RECOUNTRULES;
%let rulesDs = WORK.RULES;
%let freqDs = WORK.FREQDS;
%let sortPathData = N;
*------------------------------------------------------------*;
* Kurallar veri kümesini oluşturma;
*------------------------------------------------------------*;
data WORK.SCORINGRULES;
 length   RULEID                 8
          SIZE                   8
          COUNT                  8
          SUPPORT                8
          CONF                   8
          ITEM1                  $ 120
          ITEM2                  $ 120
          ITEM3                  $ 120
          ITEM4                  $ 120
          RULE                   $ 495
          ;
 infile   cards  dsd;
 input    RULEID
          SIZE
          COUNT
          SUPPORT
          CONF
          ITEM1 $
          ITEM2 $
          ITEM3 $
          ITEM4 $
          RULE $
          ;
 label RULEID="Rule Identifier"
          SIZE="Chain Size"
          COUNT="Count"
          SUPPORT="Support"
          CONF="Confidence"
          ITEM1="Item 1"
          ITEM2="Item 2"
          ITEM3="Item 3"
          ITEM4="Item 4"
          RULE="Rule"
          ;
 format RULEID 8.
          SIZE 3.
          COUNT 12.
          SUPPORT 8.4
          CONF 8.4
          ;
```

```
  informat ITEM1 $CHAR120.
        ITEM2 $CHAR120.
        ITEM3 $CHAR120.
        ITEM4 $CHAR120.
        RULE
        ;
%macro PathScoring;
*------------------------------------------------------------*;
* Yol Prosedürünün Puanlanması;
*------------------------------------------------------------*;
%if &sortPathData eq Y %then %do;
proc sort data=&pathData out=WORK.PATH_SORTEDDATA(keep=
request_file
date_time
sequence
referer_url
);
by date_time sequence;
quit;
run;
%end;
%if &sortPathData eq Y %then %do;
proc path data=WORK.PATH_SORTEDDATA %end;
%else %do;
proc path data=&pathData %end;
out=&rulesDs freqout=&freqDs
pctsup = 5
items=4
;
id date_time;
target request_file
;
sequence sequence / min=0 max=1000;
REFERRER referer_url / BACKUP_LIMIT=5
;
Score
Rules = work.scoringRules
out=&pathScore
numrules=200 sortby SUPPORT include all_id;
Transition
out=&transitionDs
;
Funnel
out=&funnelDs
;
run; quit;
*------------------------------------------------------------*;
* Transpose Kuralları;
*------------------------------------------------------------*;
proc sort data=&pathScore out=_temp(keep= date_time) nodupkey;
by date_time;
run;
proc sort data=&pathScore(where=(ruleid ne .)) out=_temp2(keep=ruleid) nodupkey;
by ruleid;
run;
proc sort data=&rulesDs out=_temp3(keep=rule ruleid);
by ruleid;
run;
data _temp4;
merge _temp2(in=_a) _temp3;
```

```
by ruleid;
if _a then output;
run;
proc sql;
create table _temp5 as select
_temp.date_time,
_temp4.ruleid, _temp4.rule, 0 as _scoreId from _temp, _temp4;
run;
proc sort data=&pathScore(where=(ruleid ne .)) out=_temp2;
by date_time ruleid;
run;
proc sort data=_temp5;
by date_time ruleid;
run;
data _temp6;
merge _temp5 _temp2(in=_b);
by date_time ruleid;
if _b then _scoreId=1;
run;
proc transpose data=_temp6 out=&em_score_output(drop=_name_) prefix=RULE;
by date_time;
var _scoreId;
id ruleid;
idlabel rule;
run;
proc datasets lib=work nolist;
delete _temp _temp2 _temp3 _temp4 _temp5 _temp6;
run;
%mend PathScoring;
%Pathscoring;
```

EK – 2

BİRLİKTELİK KURALI UYGULAMASI İÇİN SAS PROGRAM KODLARI

```
* %let EM_SCORE_OUTPUT=;
* %let SEQDATA =;
* data &EM_SCORE_OUTPUT;
* set &SEQDATA;
* run;
*-------------------------------------------------------------*;
* Kurallar veri kümesini oluşturma;
*-------------------------------------------------------------*;
data WORK.RULESUBSET;
  length   NITEMS                 8
           COUNT                  8
           SUPPORT                8
           CONF                   8
           ISET1                  $ 120
           ISET2                  $ 120
           ISET3                  $ 120
           index                  8
           _LHAND                 $ 200
           _RHAND                 $ 200
           ;
  infile   cards  dsd;
  input    NITEMS
           COUNT
           SUPPORT
           CONF
           ISET1 $
           ISET2 $
           ISET3 $
           index
           _LHAND $
           _RHAND $
           ;
  label    NITEMS="Chain Length"
           COUNT="Transaction Count"
           SUPPORT="Support(%)"
           CONF="Confidence(%)"
           ISET1="Chain Item 1"
           ISET2="Chain Item 2"
           ISET3="Chain Item 3"
           index="Rule Index"
           _LHAND="Left Hand of Rule"
           _RHAND="Right Hand of Rule"
           ;
  format   NITEMS 8.
           COUNT 8.
           SUPPORT 6.2
           CONF 6.2
           ;
  informat ISET1 $CHAR120.
           ISET2 $CHAR120.
           ISET3 $CHAR120.
           _LHAND
           _RHAND ;
*-------------------------------------------------------------*;
```

```
* Kural Haritası Oluşturma ve Çıkış Veri Kümeleri;
*------------------------------------------------------------*;
    MACRO: EM_SCORESEQ
    PURPOSE: Score a transaction dataset to determine the rules
    PARAMETERS:
       INDATA          = dataset to score
       OUTPUT          = scored data
       RULES           = datasets containing the rules
       TARGETVAR       = TARGETVAR variable (e.g. PRODUCT)
       IDVARS = string of  the id variables (e.g. CUSTOMER)
       SEQVAR          = sequence variable (TIME) (e.g. TIME)
       PREFIX          = prefix used for rule variables created (optional)
       NODEID          = node identifier (optional)
       RULEMAP         = dataset containing the variable-label mapping (optional)
       TIMINGWINDOW
    USAGE:
%em_scoreseq(INDATA=SAMPSIO.ASSOCS,output=SCOREASSOC,RULES=RULE,
TARGETVAR=PRODUCT,IDVARS=CUSTOMER,
                SEQVAR=TIME, RULEMAP=, TIMINGWINDOW=1);
    AUTHOR:   Dominique Latour
    AMENDMENTS: Initial Version  30Jan04 SASDZL
                        05Apr05 SASDZL Fix S0293063 check if timingWindow is < 1
                        07Jul05 SASDZL Fix S0309411 set timingWindow to -1
*/
%macro scoreseq(indata=, output=SCORESEQ, rules=, TARGETVAR=, idvars= , seqvar=,
prefix=RULE, nodeid=ASSOC, RULEMAP=RULEMAP, TIMINGWINDOW=-1);
%if (&indata eq ) or (&TARGETVAR eq ) or (&idvars eq ) or (&seqvar eq ) %then %goto exit;
%let dsid=%sysfunc(open(&rules));
%if &dsid eq 0 %then %goto exit;
%let rulenum  = %sysfunc(varnum(&dsid, RULE));
%let transnum = %sysfunc(varnum(&dsid, TRANSPOSE));
%let lhandnum = %sysfunc(varnum(&dsid, _LHAND));
%let rhandnum = %sysfunc(varnum(&dsid, _RHAND));
%let lhandlen = %sysfunc(varlen(&dsid, &lhandnum));
%let rhandlen = %sysfunc(varlen(&dsid, &rhandnum));
%let dsid = %sysfunc(close(&dsid));
%if &timingWindow < 1 %then
   %let timingWindow = -1;
*------------------------------------------------------------*;
* Kuralların çıkarılması;
*------------------------------------------------------------*;
data RULESUBSET;
  set &rules;
%if ^&rulenum %then %do;
   %let rulelen = %eval(&lhandlen + &rhandlen + 5);%put &rulelen;
   length RULE $&rulelen;
   RULE = trim(_LHAND)!!' ==> '!!trim(_RHAND);
%end;
%if &transnum %then %do;
   where TRANSPOSE=1;
%end;
run;

*------------------------------------------------------------*;
* Kuralların sayısının tanımlanması ve sıralanması;
*------------------------------------------------------------*;
proc sql;
  reset noprint;
  select count(*) into :_num from RULESUBSET;
quit;
```

```
%let tempDs        = ASSOCRULE;
%let chainDs       = ASSOCHAIN;
%let sorteddata    = SORTEDDATA;
*-------------------------------------------------------------*;
* Veri kümelerinin sıralanması;
*-------------------------------------------------------------*;
proc sort data=&indata out=&sortedData;
  by &idvars &seqvar;
run;
*-------------------------------------------------------------*;
* Veri kümesinde zincir oluşturma;
*-------------------------------------------------------------*;
data &chainDs(keep=_chain &idVars);
  length _tempchain _chain $2000;
  set &sortedData;
  by &idVars &seqVar;
  retain _chain " _tempchain " _ptime 0 _chainsize 0;
  if
   %let stringvar = &idvars;
   %do %while(&stringvar ne ");
      %let var = %scan(&stringvar, 1, ");
      %if &var ne %then %do;
         first.&var and
      %end;
      %let pos = %sysfunc(index(string, ' '));
      %if &pos %then %let stringvar = left(substr(string, pos));
         %else %let stringvar=";
   %end;
   first.&seqVar then do;
        _chain = ";
      _tempchain = trim(&targetVar);
         _chainsize = 1;
     end;
   else do;
       if ((&seqvar - _ptime)<= &timingWindow) or (&timingWindow = -1) then do;
          _tempchain = trim(_tempchain)!!' ==> '!!trim(&targetVar);
          _chainsize + 1;
      end;
      else do;
        if _chainsize > 1 then do;
             if _chain = " then
           _chain = trim(_tempchain);
              else
           _chain = trim(_chain)!!'|'!!trim(_tempchain);
        end;
          _tempchain = trim(&targetVar);
        _chainsize = 1;
      end;
     end;
     _ptime = &seqvar;
  if
   %let stringvar = &idvars;
   %do %while(&stringvar ne ");
      %let var = %scan(&stringvar, 1, ");
      %if &var ne %then %do;
         last.&var and
      %end;
      %let pos = %sysfunc(index(string, ' '));
      %if &pos %then %let stringvar = left(substr(string, pos));
      %else %let stringvar=";
```

```
   %end;
   last.&seqVar then do;
         if _chainsize > 1 and _tempchain ne '' then do;
           if _chain = '' then
            _chain = trim(_tempchain);
              else
            _chain = trim(_chain)!!'|'!!trim(_tempchain);
        end;
        output;
     end;
run;

*------------------------------------------------------------*;
* Kural değişkenlerinin Haritalanması;
*------------------------------------------------------------*;
data &rulemap(keep=map rule);
  length map $8;
  set RULESUBSET;
  map = "&prefix"!!left(trim(put(_n_, 12.0)));
run;

proc transpose data=RULESUBSET(keep=RULE) out=&tempDs(drop=_NAME_ _LABEL_);
  var rule;
run;
data &output;
  set &chainDs;
  length _tempchain tempsnipet $200;
  if _n_=1 then set &tempDs;
  %let endcol = col%trim(&_num);
  array &prefix{&_num};
  array col{&_num} col1-&endcol;
  do i=1 to &_num;
    &prefix{i}=0;
  end;
  if _chain ne '' then do;
    _tempchain=_chain;
    do while(index(_tempchain, '|'));
      snipet = substr(_tempchain, 1, index(_tempchain, '|')-1);
      do i=1 to &_num;
        tempsnipet = snipet;
        do while(index(trim(tempsnipet), trim(col{i})));
          pos = index(trim(tempsnipet), trim(col{i}));
          &prefix{i}+1;
 tempsnipet = left(substr(substr(tempsnipet, pos), index(substr(tempsnipet, pos), ' ==> ')+6));
        end;
      end;
      _tempchain = substr(_tempchain, index(_tempchain, '|')+1);
    end;
    if _tempchain ne '' then do;
      snipet = _tempchain;
      do i=1 to &_num;
        tempsnipet = snipet;
        do while(index(trim(tempsnipet), trim(col{i})));
          pos = index(trim(tempsnipet), trim(col{i}));
          &prefix{i}+1;
 tempsnipet = left(substr(substr(tempsnipet, pos), index(substr(tempsnipet, pos), '==> ')+6));
        end;
      end;
    end;
  end;
```

```
  output;
  drop i _chain col1--&endcol _tempchain pos tempsnipet snipet;
run;
*--- ----------------------------------------------------------*;
* Sonuç veri kümesinde Kural değişkenlerinin etiketlenmesi;
*----------------------------------------------------------*;
%let library = WORK;
%let member = %scan(&output, 1, '.');
%if %index(&output, .) %then %do;
   %let library = %scan(&output, 1, '.');
   %let member = %scan(&output, 2, '.');
%end;
proc datasets library=&library nolist;
  modify &member;
%let dsid=%sysfunc(open(&RULEMAP));
%if &dsid %then %do;
   %let rulenum = %sysfunc(varnum(&dsid, RULE));
   %let mapnum  = %sysfunc(varnum(&dsid, map));
   %do %while( ^%sysfunc(fetch(&dsid)));
      %let rule  = %sysfunc(getvarc(&dsid, &mapnum));
      %let label =%trim(%bquote(%sysfunc(getvarc(&dsid, &rulenum))));
      label &rule = "&label";
    %end;
%end;
%let dsid=%sysfunc(close(&dsid));
run;
quit;
*--- ----------------------------------------------------------*;
* Geçici veri kümelerinin temizlenmesi;
*----------------------------------------------------------*;
/*proc datasets lib=WORK nolist;
  delete &tempDs &chainDs &sorteddata RULESUBSET;
run;*/
%exit: %mend scoreseq;
%scoreseq(indata=&EM_SCORE_OUTPUT, output=&EM_SCORE_OUTPUT,
rules=WORK.RULESUBSET, targetvar=request_file, prefix=RULE, idvars= date_time,
seqvar=sequence, timingWindow=.);
```