KIRIKKALE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
YÜKSEK LİSANS TEZİ

Kırıkkale Üniversitesi Web Sitesinin Kullanıcı Örüntülerinin Web Madenciliği ile Analizi

Kadir Can BURÇAK

EYLÜL 2012

Bilgisayar Mühendisliği Anabilim Dalında Kadir Can BURÇAK tarafından hazırlanan KIRIKKALE ÜNİVERSİTESİ WEB SİTESİNİN KULLANICI ÖRÜNTÜLERİNİN WEB MADENCİLİĞİ İLE ANALİZİ adlı Yüksek Lisans Tezinin Anabilim Dalı standartlarına uygun olduğunu onaylarım.

Prof.Dr. Hasan ERBAY
Anabilim Dalı Başkanı

Bu tezi okuduğumu ve tezin **Yüksek Lisans Tezi** olarak bütün gereklilikleri yerine getirdiğini onaylarım.

Prof. Dr. Hasan ERBAY
Danışman

Jüri Üyeleri

Başkan	: Doç. Dr. Necaattin BARIŞCI	_____
Üye (Danışman)	: Prof. Dr. Hasan ERBAY	_____
Üye	: Yrd. Doç. Dr. Taner TOPAL	_____

……/…../…….

Bu tez ile Kırıkkale Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu Yüksek Lisans derecesini onaylamıştır.

Doç. Dr. Erdem Kamil YILDIRIM
Fen Bilimleri Enstitüsü Müdürü

ÖZET

KIRIKKALE ÜNİVERSİTESİ WEB SİTESİNİN KULLANICI ÖRÜNTÜLERİNİN WEB MADENCİLİĞİ İLE ANALİZİ

BURÇAK Kadir Can
Kırıkkale Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı, Yüksek Lisans Tezi
Danışman: Prof. Dr. Hasan ERBAY
Eylül 2012, 56 sayfa

Bu tez çalışmasında web erişim günlükleri kullanılarak kullanıcı erişim örüntülerinin, web log dosyaları analizi ile bulunması amaçlanmaktadır. Web sunucu erişim kayıtlarından sitedeki sayfalar arasındaki bağlantılar, siteye erişen kullanıcıların istekleri, web sitesinde ziyaret edilen sayfaların tespit edilmesi, web sitesinin kullanımına ait çeşitli istatistiki bilgilerin elde edilmesi büyük önem arz etmektedir. Kırıkkale Üniversitesi web sunucusuna ait kullanıcı erişim kayıtlarından alınan değerler, web kullanım madenciliği metodu kullanılarak web madenciliği yazılımları ile analiz edilmiştir. Analiz sonucunda, sitede en çok erişilen sayfalar, dosya erişimleri, giriş sayfası erişimleri, dosya tipleri, dosya uzantıları ve genel istatistikler tablo ve grafiklerle gösterilmiştir. Elde edilen sonuçlar doğrultusunda, Kırıkkale Üniversitesi web sitesinin etkililiğini arttırmak ve geliştirmek için önerilerde bulunulmuştur.

Anahtar Kelimeler: Veri Madenciliği, Zeki Veri Madenciliği, Web Madenciliği, Bilgi Keşfi, Log Analizi

ABSTRACT

ANALYSIS OF USER PATTERNS OF THE WEB SITE OF KIRIKKALE UNIVERSITY WITH WEB MINING

BURÇAK Kadir Can
Kırıkkale University
Graduate School of Natural and Applied Sciences
Department of Computer Engineering, M.Sc. Thesis
Supervisor: Prof. Dr. Hasan ERBAY
September 2012, 56 pages

In this thesis study, it is intended to find the user access pattern by analyzing the web log files. It has a big importance that the connections between the pages from the web server access files, the prompts of the users that Access the web site,identifying the pages of the web site that were visited, obtaining various statistical information of the usage of the web site. It were analyzed the values that were taken from the access records that belong to the web server of Kırıkkale University with the web mining softwares using the web usage mining method. In the end of the analysis it were shown with tables and graphics, the most accessed pages in the web site, file accesses, file axtensions and general statistics. It has been made some suggestions to increase the effectuality and to improve the Kırıkkale University web site.

Keywords: Data Mining, Intelligent Data Mining, Web Mining, Knowledge Discovery, Log Analysis.

TEŞEKKÜR

Tezimin hazırlanması esnasında hiçbir yardımı esirgemeyen öğrencilerine büyük destek olan, bilimsel deney imkânlarını sonuna kadar bizlerin hizmetine veren, tez yöneticisi hocam, Sayın Prof. Dr. Hasan ERBAY 'a, tez çalışmalarım esnasında çok büyük fedakârlıklarla bana destek olan aileme teşekkür ederim.

İÇİNDEKİLER

ÇİZELGELER DİZİNİ

ŞEKİLLER DİZİNİ

1. GİRİŞ

Günümüzde bilgiye ulaşmamızı kolaylaştıracak en önemli araçlardan biri internettir. Bu da internet kullanımının hızlı bir şekilde artmasına neden olmaktadır. 1993 yılında, dünya çapında kullanıcılarının sayısı 900.000 iken, 2000 yılında bu sayı 304 milyona, Şubat 2002'de 544,2 milyona, 2004 yılı sonunda 934 milyona ulaşmıştır. 2008 yılı sonunda dünyadaki internet kullanıcı sayısı 1,463 milyara, 2011'de ise bu sayı 2 milyar 100 milyona ulaşmıştır.

Bilgisayarların yaşamımıza daha çok girmesiyle birlikte, artık her yaptığımız işlem sayısal ortamda kayıt altına alınmaya başlanmıştır. Hastanelerde, belediyelerde veya ticarette yaptığımız her işlem artık anında veri tabanında yerini alıyor. Hatta, bir mağazaya, alışveriş merkezine girerken ya da çıkarken, bazen de yolda yürürken kameraya çekilen görüntülerimiz bile veri tabanına kaydediliyor. Bütün bunlar bir yığın halinde depolanırken içlerinde çok önemli bilgiler gizlidir. Bu durum, eldeki verilerden işe yarar bilgiyi çıkarma zorunluluğunu doğurmuştur. Veri madenciliği eldeki veriden anlamlı bilgileri, ilişkileri çıkarmada kullanılan tekniklere verilen genel isimdir.

Veri madenciliğinin bir diğer uygulama alanı da internet üzerinde bulunan verilerdir. İnternet üzerinde bulunan veriler üzerinde işlem yapan veri madenciliği yöntemi web madenciliği olarak adlandırılır. Web madenciliği, veri madenciliği tekniklerinin kullanılarak web belgelerinden ve servislerinden otomatik olarak bilginin ayıklanması, ortaya çıkarılması ve tahlil edilmesidir.

Web madenciliğinin işi, bu bilgilerin farklı veri madenciliği teknikleri kullanılarak site sahibine yararlı bilgiler sunmasıdır. Bu sayede ticari amaçlı bir siteden elde edilen kar miktarı arttırılabileceği gibi, internet sayfaları farklı ilgi alanlarına göre düzenlenerek ziyaretçi memnuniyetinin artması sağlanmaktadır.

Bu tez çalışmasının amacı, web kullanım madenciliği teknikleri ile yapılan çalışmaları incelemek ve Kırıkkale Üniversitesinin bir yıllık web erişim kayıtlarından sitenin analizini yapmaktır. Web kullanım madenciliğinin ilk aşaması olan ön işlem

aşamasında sql sorgulama dili kullanılmış ve veriler üzerinde temizlik yapılmıştır. Elde edilen temizlenmiş veriler web uygulama yazılımlarından **nihuo** programıyla incelenmiş; sitede en çok erişilen sayfalar, dosya erişimleri, giriş sayfası erişimleri, dosya tipleri, dosya uzantıları ve genel istatistikler elde edilmiş, tablo ve grafiklerle gösterilmiştir.

Bu tez çalışması altı bölümden oluşmaktadır. Üçüncü bölümde veri madenciliği, dördüncü bölümde web madenciliği, beşinci bölümde Kırıkkale Üniversitesi web sitesinin analiz uygulaması ve son olarak altıncı bölümde sonuçlar ve öneriler sunulmuştur.

2. LİTERATÜR İNCELEMESİ

Veri madenciliği son zamanlarda akademik çevrenin ilgi odağı olmuştur ve bu konuyla ilgili farklı alanlarda birçok araştırma yapılmıştır. Bunlardan, Gezer ve arkadaşları [1] yapmış oldukları web kullanım madenciliği analiz çalışmasında, İstanbul Üniversitesi Uluslararası Akademik İlişkiler Kurulu AB Eğitim birimine ait web sitesi sunucu kayıt dosyalarını **wumprep** ve **wumweb** yazılımlarını kullanarak analiz yapmışlardır. Takci ve Soğukpınar [2] çalışmasında kütüphane kullanıcılarının veri tabanlarını kullanılarak, kullanıcıların web üzerindeki davranışları ile ilgili analiz yapmışlardır. Uğur ve Kınacı [3] yaptıkları yapay zekâ tekniği çalışmalarında, kategorilere ayrılmış web sitesindeki verilere yapay sinir ağları yöntemini uygulayarak web sayfalarını sınıflandırmışlardır. Bu çalışmada kullanıcı kayıt dosyalarındaki verilere **Apriori** algoritması uygulanarak kullanıcı erişim örüntülerinden kullanıcı bilgileri çıkarılmıştır. Öte yandan, Wang ve Lee [4] yaptıkları çalışmada kullanıcıların ilk başlangıç noktasından son çıkış noktasına kadar olan yerlerini kaydederek, sonraki ziyaret edilen web sayfalarının ziyaretinde doğru eğilimler gösterebilecek bir grafik **travers** algoritması geliştirmişlerdir. Cooley ve arkadaşları [5] yaptıkları makalede, web kullanım madenciliği için **webminer** adlı bir sistem geliştirmişlerdir. Bu sistemin amacı, otomatik olarak kullanıcı erişim kayıtlarından (access.log) birliktelik kuralları ile sıralı örüntüleri keşfetmektedir. Iocchi [6] çalışmasında, geliştirdiği **weboem** modeli, internet ortamında kaydedilmiş yarı yapısal bilgilerin çıkarılması için tasarlanmış bir bilgi modelidir. Bu model, internet ortamında kaydedilmiş dağınık bilgi yığınlarının büyük bir kısmından bilgi keşfi yapmaktadır. Oktay [7] Apriori ve **Tfpr** algoritmasını kullanarak en sık ziyaret edilen ve ziyaret edilme olasılığı en yüksek olan sayfaları bulmuştur.

3. VERİ MADENCİLİĞİ

Bilişim teknolojilerindeki hızlı gelişmeler dünyadaki veri miktarını arttırmıştır. Oldukça hızlı bir şekilde artan bu çok büyük hacimdeki verilerin saklanması için yıllardır kullanılan veritabanları tek başına yeterli olmamaya başlamış ve veri ambarları kavramının ortaya çıkmasına neden olmuştur.

Veri madenciliği, veri tabanlarında veya veri ambarlarında depolanan verilerde gizli bulunan öz bilgiyi keşfetme işlemidir. Bu amacına ulaşmak için yeni nesil hesaplama tekniklerini ve araçlarını kullanır. Örneğin, tezin sonuçlarını üretmede kullanacağımız **mssql** sorgulama dili bu araçlar arasında yer almaktadır. Teknikler arasında ise yukarıda literatür kısmında bahsettiğimiz algoritmalar vardır. Bu algoritmalardan bazıları ileri kısımlarda detaylandırılmıştır.

3.1. Giriş

Teknolojik cihazlara, internet teknolojilerinin entegre olması, internet kullanımını büyük ölçüde etkilemiştir. İnternet kullanıcısı sayısının artması web üzerindeki bilgi yığınına neden olmuştur. İlişkisi olmayan bilgilerden yeni anlamlı bilgilerin elde edilmesi veri madenciliğini doğurmuştur.

3.2. Veri Madenciliğinin Tanımı ve Tarihçesi

Veri madenciliği, istatistiksel ve matematiksel tekniklerle birlikte örüntü tanıma teknolojilerini kullanarak, depolama ortamlarında saklanmış bulunan veri yığınlarının elenmesi ile anlamlı yeni korelasyon, örüntü ve eğilimlerin keşfedilmesi sürecidir [9].

Veri madenciliğinin tarihi gelişimine baktığımızda, 1960'lı yıllarda istatistikçiler yeni bir algoritma keşfederek veri tabanı sistemlerini geliştirmiş, büyük sayıda metin dokümanlarının saklanmasını ve bilginin geri kazanılmasını sağlamışlardır.

1980'li yılların başında araştırmacılar makine öğrenimine çok farklı bir gözle bakmaya başlamışlardır. Makine öğrenimi, araştırmacıların yeni keşifler yapmasını sağlamıştır. Bunlar, objelerin karar ağaçlarıyla keyfi olarak sınıflandırılması için yapılan modellemelerdir. Aynı zamanda bilgisayar teknolojisinin de ilerlemesinden dolayı daha güçlü hale gelen bilgisayarlarda, yeni algoritmalar gerçek problemlerle uygulama olanağı bulmuştur. Üretim programlaması ve zaman tablosu planlaması gibi konuların bilgisayarlarla çözümlemesi oldukça zordur. Bu konuları tecrübeli planlamacılar daha kolay çözümlerler. Çünkü planlamacılar edindikleri tecrübelerle karmaşıklıkları nasıl gidereceklerini öğrenirler. Yapay zekada da öğrenme kapasitesinin rolü büyük olduğundan öğrenme algoritmaları önem kazanmıştır ve bu gibi nedenler veri madenciliğine olan ilgiyi artırmıştır [10]. Aşağıda bu faktörlerden bazılarına değinilecektir.

Veri madenciliğine ilginin artması aşağıdaki faktörlerle açıklanabilir:

- 1980' lerde şirketler, müşterileri, rakipleri, ürünleri ile ilgili verilerden oluşan veri tabanları oluşturmuşlardır. Bu veri tabanları potansiyel altın madeni gibidir. Sayısı milyonları geçen bu verilere, veri tabanı sorgulama dili sql ya da başka yüzeysel sorgulama dilleri kullanılarak kolaylıkla ulaşılabilir olması veri madenciliğine olan ilginin artmasını sağlamıştır. Çünkü bu işlemi elle yapmak mümkün değildir.

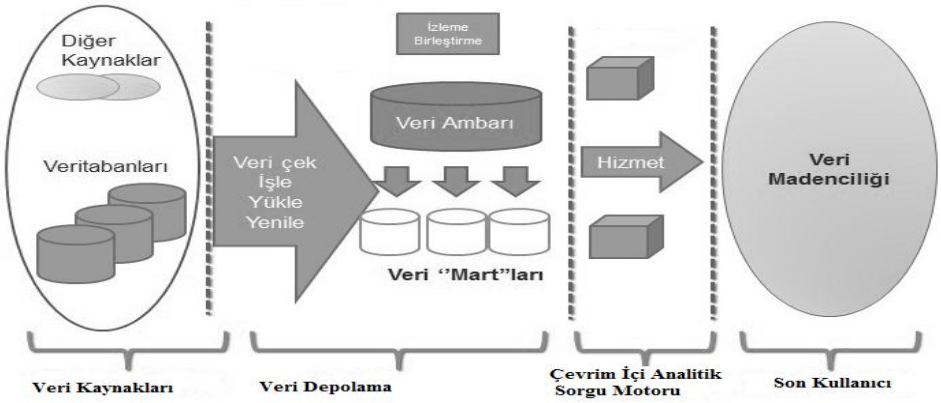
- Bilgisayarlarda ağ kullanımı gelişmeye devam etmektedir. Bu durumda veri tabanı ile bağlantı kurmak kolaylaşır. Böylece demografik verili dosya ile müşteri dosyası arasında bağlantı kurulabilir ve belirli popülasyon gruplarının kimliklerinin belirlenmesi sağlanabilir.

- Son birkaç yılda makine öğrenimi teknikleri oldukça gelişmiştir. Sinir ağları, genetik algoritmalar ve diğer basit uygulanabilir öğrenme teknikleri veri tabanlarıyla ilginç bağlantılar kurmayı kolaylaştırır.

- Müşteri ile hizmet veren arasındaki ilişki, kişisel bilgileri hizmet verenin masasındaki bilgisayardan merkezi bilgi sistemlerine gönderir. Depolanmış, ulaşılması kolay bilgiyi pazarlamacıların ve sigortacıların kullanmak istemeleri.

3.3. Veri Ambarı

Veri ambarı ilişkili verilerin sorgulanabildiği bir depodur. Aynı zamanda veri ambarı bir kurumun değişik birimleri tarafından toplanan bilgilerden değerli olanlarının, gelecekte analiz işlemlerinde kullanılması amacıyla veri tabanlarında depolanması işlemidir. Veri ambarı kullanıldığında, günlük işletimsel görevlerle yeterince meşgul olan veri tabanı kullanılmadan analiz işleminin yapılmasına olanak sağlar.



Şekil 3.1. Veri Ambarı Mimarisi

Veri Ambarı İşlevleri:

- Değişik platformlar üzerindeki işletimsel uygulamalara ait verilere erişim ve gerekli verilerin bu platformlardan alınması,

- Alınan verilerin temizlenmesi, tutarlı duruma getirilmesi, özetlenmesi, birleştirme ve birbirleriyle entegrasyonunun sağlanması,

- Dönüştürülen verilerin veri ambarı veya datamart (1 ile 10 GB arasında veri kapasiteli bölümsel ambar) ortamına dağıtımı,

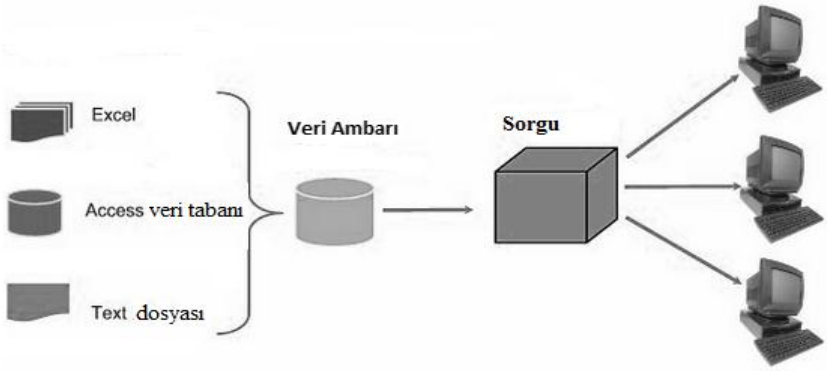
- Gönderilen verilerin bir veri tabanında toplanması,

- Depolanan bilgi ile metadata(veri hakkında veri) da bulunan ilgili bilgilerin veri kataloğunda saklanması ve son kullanıcılara sunulmasıdır.

3.4. Çevrim İçi Analitik Sorgu

Çevrim içi analitik sorgu, kullanıcı tarafından anlaşılabilecek şekilde gerçek boyutlara taşınmış işlenmemiş ham veri üzerinde çeşitli bilgi görüntüleri sunarak, analistler, yöneticiler ve çalışanların veriye hızlı, tutarlı ve etkileşimli bir biçimde erişmesini sağlayan bir yazılım teknolojisidir.

Çevrim içi analitik sorgu yapıları organizasyonel yapılarla uğraştığı için öznel bir yapıdadır ve aynı zamanda birçok sistemden de beslendiği için bütünleşik bir formda çalışmaktadır. Bu sorgu yapıları için en önemli özellik verilerin mutlaka zaman eksenli olarak tutuluyor olmasıdır. Çevrim içi analitik sorgu yapıları çok sık ekleme ve güncelleme işlemlerine tabii tutulmazlar. Bu yapı için güncelleme işleminin anlamı eski verinin silinmeden aynı kayıt için yeni verilerin giriş yapılmasıdır [11].



Şekil 3.2. Çevrim İçi Analitik Sorgu Yapısı

3.5. Veri Madenciliğinin Özellikleri

Veri tabanlarında veya veri ambarlarında depolanan verilerde gizli bulunan öz bilgiyi keşfedebilmek amacıyla insanlara yardımcı olacak yeni nesil hesaplama tekniklerine ve araçlarına ihtiyaç duyulmaktadır. Veri tabanlarında öz bilgi keşfinin konusu olan bu teknikler ve araçlar, veriyi anlamlı hale getirmek amacıyla yapılan değişik faaliyetlerin bütünüdür [12].

Veri madenciliği çoğu araştırmacı tarafından öz bilgi keşfi ile aynı anlamda kullanılmaktadır. Halbuki veri madenciliği, veri tabanlarında öz bilgi keşfi sürecinin

adımlarından birisidir. Veri tabanlarındaki öz bilgi keşfi aşağıdaki adımlarla ifade edilmektedir [13]:

- Verilerin temizlenmesi,
- Verilerin birleştirilmesi,
- Verilerin seçilmesi,
- Verilerin dönüşümü,
- Veri madenciliği algoritmasını uygulama,
- Örüntülerin değerlendirilmesi,
- Özbilginin sunumu.

3.6. Veri Madenciliğinin Uygulama Alanı

Veri madenciliği bankacılık, pazarlama, sigortacılık, sağlık gibi değişik alanlarda uygulanmaktadır. Veri madenciliğinin kullanılmasında sektör farkı gözetilmemekle beraber, geniş veri ambarlarının oluşturulmasına olanak veren, perakende satış, sigortacılık, sağlık gibi alanlarda yaygın şekilde kullanılmaktadır [14].

Veri madenciliğinin pazarlama alanındaki uygulamaları:

- Müşterilerin satın alma alışkanlıklarının belirlenmesi,
- Mevcut müşterilerin elde tutulması, yeni müşterilerin kazanılması,
- Pazar sepeti analizi,
- Müşteri ilişkileri yönetimi,
- Müşteri değerlendirme,
- Satış tahmini.

Veri madenciliğinin bankacılık alanındaki uygulamaları:

- Farklı finansal göstergeler arasında gizli ilişkilerin ortaya konulması,
- Kredi kartı dolandırıcılıklarının ve sahtekarlıkların belirlenmesi,
- Kredi kartı harcamalarına göre müşteri gruplarının belirlenmesi,
- Kredi taleplerinin değerlendirilmesinde.

Veri madenciliğinin sigortacılık alanındaki uygulamaları:

- Yeni poliçe talep edecek müşterilerin tahmin edilmesi
- Sigorta dolandırıcılıklarının tespiti
- Riskli müşteri gruplarının belirlenmesi

Veri madenciliğinin elektronik ticaret alanındaki uygulamaları:

- Saldırıların çözümlenmesi
- Web sayfalarına yapılan ziyaretlerin çözümlenmesi

3.7. Veri Madenciliği Modelleri

Veri madenciliği modelleri, gördükleri işlevlere göre dört ana başlık altında toplanabilir. Bu modeller **(1) değer tahmini modeli**, **(2) veri tabanı kümeleme modeli**, **(3) bağlantı analizi modeli** ve **(4) fark sapmaları modelidir.** Bu yöntemlerin uygulanmasında birçok teknik ve algoritmalardan yararlanılmaktadır. Kullanılan teknik ve algoritmalar genel olarak tahminleyici, tanımlayıcı veya her iki yaklaşımı da içerebilirler [14].

3.7.1. Değer Tahmini Modeli

Değer tahmini ya da tahminsel modeldeki öğrenme daha çok, bir insanın öğrenmesine benzemektedir. İnsan tüm yaşamı boyunca çevresini sürekli gözleyerek bir şeyler öğrenir. Tahminsel model de kendisine verilen veri tabanını inceleyerek, bu veri tabanındaki temel unsurları birbirine benzeterek tanımlamaya, onları isimlendirmeye ve sınıflamaya çalışmaktadır. Tıpkı bir çocuğun kadın ve erkek cinsiyetlerini sınıflandırması gibidir. Çocuk için ilk önce cinsiyet kavramı yoktur. Daha sonra anne, baba, teyze, hala, amca, kendinden büyük ve küçük erkek ve kız çocuklarını görür ve bir sınıflandırma yapar. Aslında tüm bunlar çocuk için bir veri tabanıdır. Bu veri tabanını inceleyen çocuk, kadınla erkek arasındaki temel farkları belirler daha sonra kendisine hiç tanımadığı kız çocuğu gösterildiğinde bir önceki deneyimine öğrenmesine dayanarak bunun kız olduğuna karar verir. Çocuğun yaptığı davranış tamamen bir sınıflandırma veya genelleme yapma işlemidir [15].

3.7.2. Bağlantı Analizi Modeli

Tahmini modelde kullanılan yazılım kendisine verilen veri tabanını bir bütün olarak düşünür ve öğrenmesini de bu bütünü temel alarak gerçekleştirir. Oysa bağlantı analizinde veri tabanındaki her bir kayıt veya kayıtlar grubu arasında bir bağlantı, ilişki yaratılmaya çalışılır. Bağlantı analizi bir veri tabanındaki kayıtlar ya da bir graf üzerindeki düğümler arasında çok rastlanan kuralları ortaya çıkarır. Çapraz satış, stok fiyat hareketleri ve hedef müşteri kitlesinin belirlenmesi gibi uygulamalar bağlantı analizinin en çok kullanıldığı alanlardandır [14]. Bağlantı analizi modeli üç ana başlık altında incelenebilir.

3.7.2.1. Birliktelik Kuralları

Birliktelik kuralı belirli türlerdeki veri ilişkilerini tanımlayan bir modeldir. Bu yönden de tanımlayıcı bir modeldir. Herhangi bir ürün alındığında bu ürünün yanında bir başka ürünün de satın alınması bir birliktelik kuralı oluşturur. Ürünlerin birlikte alınmaları söz konusu olunca, birliktelik kuralları daha çok perakendecilik sektöründe faaliyet gösteren işletmelerde uygulanmaktadır [16].

3.7.2.2. Örüntü Tanıma

Örüntü tanıma, daha önce belirlenmiş bir model diyebileceğimiz çok boyutlu bir örüntünün veri tabanındaki benzerlerini ya da 'en benzerini' aramaktır. Herhangi bir yazılı metni tanımak ya da o metnin çok benzerini bulmak örüntü tanımanın konusuna girer. Bunun dışında parmak izi, ses, yüz tanıma, kan hücrelerinin karşılaştırılması, el yazılarının tespiti gibi alanlarda da uygulanır. Dolayısıyla örüntüden kasıt el, yüz, resim ve ses gibi varlıkların sayısal ortamda sergiledikleri şekildir.

3.7.2.3. Ardışık Zaman Örüntüleri

Yukarıda örüntü sözcüğünün, herhangi bir çizim, ses, resim, parmak izi vs gibi bir şekil olduğundan söz edilmişti. Bu örneklere ek olarak, bir kimsenin yaptığı işler de örüntü olarak tanımlanabilir. Örneğin bir müşterinin süt, peynir ve ekmek satın alması bir örüntüdür. Bu noktadan hareket edilerek bir müşterinin birinci

gün A ürünü, onu izleyen gün veya günlerden birinde B ürünü ve daha sonraki bir günde de C ürünü alması ise yine bir örüntü oluşturacaktır. Ancak bu sefer birbirini izleyen, yani zaman içinde ardışık olan bir örüntü oluşturacaktır.

Çizelge 3.1. Müşteri Alış-Veriş Tablosu

Müşteri No	İşlem Zamanı	Ürün No
1	21- Ocak -2012	17-11
2	21- Ocak -2012	11
1	22- Ocak -2012	12-18-13
4	22- Ocak -2012	46
4	23- Ocak -2012	15-79-88-35
1	24- Ocak -2012	15
2	24- Ocak -2012	12
3	25- Ocak -2012	26
2	26- Ocak -2012	13

Çizelge 3.1'de tüm müşterilerin yaptıkları alışverişler satın aldıkları ürün kodları görülmektedir. Müşterilerin zamana göre alış-verişlerine bakıldığında bir dizi oluşturduğu görülmektedir. Burada tabloyu incelediğimiz zaman 2 numaralı müşterinin sırasıyla (11), (12), (13) numaralı ürünleri satın alması bir ardışık zaman örüntüsü oluşturmaktadır. Bu örüntünün, tablo içinde başka bir müşteri tarafından, tekrarı veya benzeri yoktur. Bu bir örüntü olarak değerlendirilebilir.

Çizelge 3.2. Oluşturulan Dizi Tablosu

Müşteri No	Ürün No
1	17-11-12-18-13-15
2	11-12-13
3	26
4	46-15-79-88-35

Çizelge 3.2'de görüldüğü gibi (11), (12), (13) dizisi hem 2 numaralı müşteri hem de 1 numaralı müşteri tarafından desteklenmektedir. 2 numaralı müşteri bunları sırasıyla alırken 1 numaralı müşteri (17), (15) numaralı ürünler arasında (11), (12), (13) numaralı ürünleri satın almıştır. (18) numaralı ürünü bunların arasında olması ardışıklığı bozmayacaktır.

3.8. Sınıflandırma Teknikleri ve Algoritmalar

Sınıflandırma en çok bilinen veri madenciliği tekniklerinden birisidir; örüntü tanıma, hastalık tanıları, dolandırıcılık tespiti, kalite kontrol çalışmaları ve pazarlama konuları sınıflandırma tekniklerinin bolca kullanıldığı alanlardır. Sınıflandırma tahminleyici bir modeldir; havanın bir sonraki gün nasıl olacağı ya da bir kutuda ne kadar mavi top olduğunun tahmin edilmesi aslında bir sınıflandırma işlemidir [16]. Veri madenciliği çerçevesinde ilerideki konularda istatistiksel yöntemlerin dışında sınıflandırma işleminde çeşitli teknik ve algoritmalara değinilmiştir.

Şekil 3.3. Modelleme Yaklaşımı

Modelleme tasarımı sürecinde, araştırmacı tarafından ilgili sınıflar, önceden belirlenen kriterlere göre ayrılarak, her sınıf için çeşitli örnekler verilir. Böylece sınıfların özellikleri belirlenmiş olur. Şekil 3.3'de gösterildiği gibi ilk aşama olan öğrenme süreci tamamlandıktan sonra yeni örnekler sistemde uygulanır. Bu örneklerin hangi sınıfa ait olduğu model tarafından belirlenir. Böylece verinin olağan kümelere yerleşmesi sağlanır.

3.9. Karar Ağaçları

Bu teknikte sınıflandırma için bir ağaç oluşturulur daha sonra, veri tabanındaki her bir kayıt bu ağaca uygulanır ve çıkan sonuca göre de bu kayıt sınıflandırılır. Temel olarak iki adımdan oluştuğu söylenebilir: Birincisi ağacın kurulması, ikincisi de verilerin teker teker ağaca uygulanarak sınıflandırmanın gerçekleştirilmesi şeklindedir.

Karar ağaçları oluşturulurken kullanılan algoritmanın ne olduğu önemlidir. Kullanılan algoritmaya göre ağacın şekli değişebilir. Değişik ağaç yapıları da farklı sınıflandırma sonuçları verecektir [17].

Karar ağacı temelli tipik uygulamalar:

- Bireylerin kredi geçmişlerini kullanarak kredi kararlarının verilmesi,
- İşletmeye en faydalı olan bireylerin özelliklerini kullanarak işe alma süreçlerinin belirlenmesi,
- Tıbbi gözlem verilerinden yararlanarak en etkin kararların verilmesi,
- Hangi değişkenlerin satışları etkilediğinin belirlenmesi,
- Üretim verilerini inceleyerek ürün hatalarına yol açan değişkenlerin belirlenmesidir.

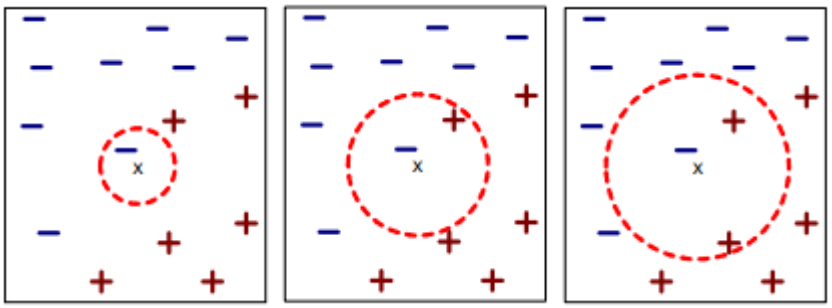
Karar ağacı temelli analizlerin yaygın olarak kullanıldığı sahalar:

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

- Gelecekteki olayların tahmin edilebilmesi için kurallar oluşturulması,
- Parametrik modellerin kurulmasında kullanılmak üzere çok miktardaki değişken ve veri kümesinden faydalı olacakların seçilmesi,
- Sadece belirli alt gruplara özgü olan ilişkilerin tanımlanması karar ağacı temelli analizlerdendir.

3.10. Mesafeye Dayalı Sınıflandırma Algoritmaları

Sınıflandırma yapılırken eldeki verilerin birbirlerine olan uzaklığı veya benzerliği kullanılarak yapılan sınıflandırma tekniğidir. Veriler arasındaki mesafe ölçülürken en çok kullanılan mesafe öklid mesafesidir. Mesafeye dayalı algoritmalardan en bilineni K-en yakın komşu algoritmasıdır. Burada bütün örnekler n-boyutlu uzayda bir noktaya karşı düşürülür. X' e uzaklığı en küçük olan K-en verisidir.



Şekil 3.4. En Yakın Komşu Tanımı

Şekil 3.4'de giriş parametresi olan K-en verisi, veri nesnelerinin kaç adet kümeye ayrılacağını belirler. Amaç bölümleme işlemi sonunda, elde edilen kümelerin, küme içi benzerliklerinin maksimum, kümeler arası benzerliklerinin minimum olmasını sağlamaktır.

3.11. Yapay Sinir Ağları

Sınıflama ve regresyon modellerinde kullanılan başlıca tekniklerden olan yapay sinir ağları biyolojik sinir ağlarından esinlenerek geliştirilmiş bir bilgi işleme sistemidir. Literatür incelemesinde belirtildiği gibi Uğur ve Kınacı [3] yaptıkları yapay zekâ tekniği çalışmalarında, kategorilere ayrılmış web sitesindeki verilere yapay sinir

ağları yöntemini uygulayarak web sayfalarını sınıflandırmışlardır. Yapay sinir ağıyla biyolojik sinir ağının bir modeli oluşturulmak istenmektedir. İşlem elemanlarının aynı doğrultu üzerinde bir araya gelmeleriyle katmanlar oluşmaktadır. Yapay sinir ağları **(1) girdi katmanı**, **(2) ara katman** (gizli katman) ve **(3) çıktı** katmanından oluşur.

4. WEB MADENCİLİĞİ

Bilgi ve belge yönetiminde, veri ve web madenciliği teknolojileri büyük önem taşımaktadır. Web madenciliği konusu, web içerik madenciliği, web kullanım madenciliği ve web yapı madenciliği olarak üç grupta incelenir. İlk grup, web içeriği olarak anılan world wide web genelinde kullanılan kaynaklardan bilgi veya kaynak keşfi sürecidir, ikinci grup web kullanım madenciliği olarak bilinen web erişim günlükleri veya kullanıcı işlemlerinden bilgi keşfi sürecini kapsar. Üçüncü grup ise web yapı madenciliğidir. Web yapı madenciliğinin amacı da web sayfaları arasındaki bağlantı verilerinden bilginin keşfi sürecidir. Burada web bilginin ana kaynağıdır. Web içeriği ve kullanımından güvenilir bilgi ve bilgi keşfini amaçlayan web madenciliği zorlu bir faaliyettir.

4.1. Web Madenciliği Tanımı

Web madenciliği; veri madenciliği teknikleri kullanılarak, web sunucularında bulunan kullanıcı kayıt dosyalarından, otomatik olarak öngörülemeyen bilgiye ulaşmaktır. Kısaca web de bulunan bilgilerin keşfedilmesidir.

Günümüzde birçok işlemin internet üzerinden yürütülmesi sonucu, çok büyük oranda veri yığınları internet ortamında oluşmuş durumdadır. İnternet üzerinde bir siteye bağlanan herkes bağlantı loglarını tutan sunucularda iz bırakır. Bu izler ip adresleri, tarayıcı kayıtları, çerez ’ler vb. dir.

Web üzerindeki veri yığınları:

- Web sayfaları
- Access Log dosyaları
- Kullanıcı kayıt bilgileri
- Oturum ve hareket bilgileri
- Site yapısı ve içeriği

Web madenciliği yukarıda sayılan çeşitli yapıdaki web sayfaları dokümanlarını ve kayıt bilgilerini incelemek, bunlardaki kalıpları keşfetmek için veri madenciliği tekniklerinin kullanılması olarak tanımlanabilir [19].

4.2. Web Veri Kaynakları

Web madenciliğinde veri kaynakları genellikle web sunucu kaynak dosyalarından oluşur. Web kullanım madenciliği esnasında harmanlanacak veriler aşağıdaki tiplerde olabilir:

- **İçerik verisi:** Web sayfalarında, metin şeklinde yer alan verilerdir. Örneğin bir web sayfasında bulunan menüler, haberler, resimler içerik verisine örnek gösterilebilir.

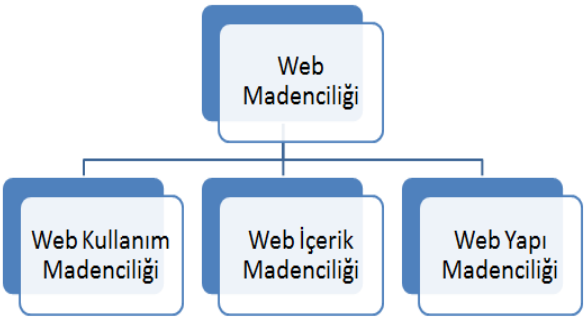
- **Yapı verisi:** Web sitesinde yer alan sayfaların hangi alt dizinler içerisinde bulunduğunu gösteren verilerden oluşur. Site haritası yapı verisine örnek verilebilir.

- **Kullanım verisi:** Kullanıcıların web sitesini ziyaretleri sırasında oluşturdukları veri tipidir. Web sitesini ziyaret eden kullanıcıların ne zaman, hangi sayfaları ziyaret ettiği, ne kadar süre sitede kaldığı gibi veriler kullanım verisidir.

- **Kullanıcı profili:** Web sitesini ziyaret eden kullanıcıların kimlik bilgileri, şifreleri, kullanıcı isimleri gibi bilgiler kullanıcı profili verisine örnektir.

4.3. Web Madenciliği Sınıflandırılması

Web madenciliği önceki konularda da belirtildiği gibi web içerik madenciliği, web kullanım madenciliği ve web yapı madenciliği olmak üzere üç kategoride literatürde yer almaktadır. Şekil 4.1 de web madenciliği yaklaşımı grafiksel olarak gösterilmiştir.



Şekil 4.1. Web Madenciliği Yaklaşımı

Web içerik madenciliği multi medya ve metin gibi web dökümanlarından yeni bilgilerin keşfini amaçlar. Web yapı madenciliği ise, sitenin yapısal dizaynını iyileştirmek için web sayfaları ve web siteleri arasındaki bağlantıları inceleyerek bir takım bilgiler üretir. Siteyi ziyaret eden kullanıcıların örüntüleriyle ilgili bilgi keşfi süreci de web kullanım madenciliğinin alanıdır [20].

4.3.1. Web İçerik Madenciliği

Web içerik madenciliği, internet üzerinde bilgi keşfi üzerine yoğunlaşmıştır. Ses, görüntü, video ve metin gibi web dökümanlarından otomatik olarak yeni bilgilerin keşfini amaçlamaktadır. Web içerik madenciliği, alt yapısında hazırlanan programlar web sayfalarını dolaşarak bilgi toplarlar. Örnek olarak arama robotları verilebilir. Bu robotlar web sayfalarını dolanarak site hakkında bilgi sahibi olurlar.

Web sayfalarının sahip olduğu içerikler genellikle metin tabanlıdır. Bundan dolayı web içerik madenciliği metin madenciliği ile de yakından ilgilidir. Çizelge 4.1'de web içerik madenciliğinde kullanılan veriler genellikle html sayfası ve metin belgeleri şeklindedir. Öte yandan web içerik madenciliği, veri madenciliği ile ilgilidir çünkü web dokümanları içerisindeki verileri çıkarmak için veri madenciliği tekniklerini kullanır. Veri içeriklerinin farklı türde olması verinin analizini zorlaştırmaktadır. Verilerin daha iyi analiz edilebilmesi için farklı web madenciliği yaklaşımları geliştirilmiştir.
Web içerik madenciliğinde kullanılan iki yaklaşım vardır

- **Bilgi Erişim Yaklaşımı**: Bu yaklaşım kullanıcılara gösterilen bilgileri filtrelemek ve bilgiye erişimi ilerletmek için kullanılan yöntemdir.

- **Veri Tabanı Yaklaşımı**: Geleneksel dosya tabanlı yaklaşımlardan farklı, önemli avantajlara sahip bu yaklaşım, ilgili bir veri havuzundan birden fazla uygulama programları ile veriyi modellemek ve veriyi bütünleştirerek daha karmaşık bir yapıya sokmak için kullanılan yöntemdir.

Çizelge 4.1. Web İçerik Madenciliği Veri Durumu

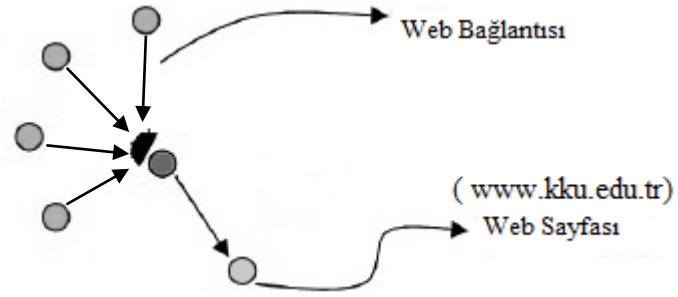
Web İçerik Madenciliği		
Veri	Verinin Şekli	Verinin Görseli
Html sayfası ve metin belgeleri	Yapısız,karışık	İlişkili ve sınıflandırmalı

Web içerik madenciliği uygulamaları:

- Sınıflandırılmış web dokümanlarını,
- Kümelenmiş web sayfalarını,
- Web site içeriklerinin karşılaştırılması,
- Doküman yapısının modellenmesini.

4.3.2. Web Yapı Madenciliği

Web yapı madenciliği, web sayfaları arasındaki bağlantıların ilişkilerine bakarak bilgi üretmektir. Örneğin hangi sitelerin, hangi sitelere bağlantı verdiği bilgisi bir grafik şekline dönüştürülerek en çok bağlantı alan siteleri analiz etmemize olanak sağlayabilir. Ayrıca web yapı madenciliği sitenin kendi içindeki bağlantı yoğunluğu hakkında site yöneticilerine faydalı bilgiler sunabilir. Şekil 4.2'de web sayfaları arasındaki bağlantı grafik şeklinde örneklendirilmiştir. Burada dökümanlar arasındaki oklar iki sayfa arasındaki ilişkiyi, noktalar ise sayfaları temsil etmektedir [20].



Şekil 4.2. Web Yapı Grafı

Bu grafikten yola çıkarak iki sayfa arasındaki en kısa yola ulaşılabilir. Bu bilgi web sayfaları arasındaki ilişkiyi belirlemek açısından son derece önemlidir. Çizelge 4.2'de görüldüğü gibi web yapı madenciliğinde kullanılan verilerin tipleri html linkleri şeklindedir.

Çizelge 4.2. Web Yapı Madenciliği Veri Durumu

Web Yapı Madenciliği		
Veri	Verinin Şekli	Verinin Görseli
Html linkleri	Link yapısında	Grafik

Bu tip sitenin içeriğine yönelik analizler site tasarımcılarına, sitenin geliştirilmesi açısından faydalı bilgiler sunabilir.

4.3.3. Web Kullanım Madenciliği

Web kullanım madenciliği, sunucu üzerinde tutulan kullanıcı erişim verilerinden bilgi keşfini amaçlar. Kullanıcıların siteyi ziyaretlerinin sonrasında bıraktığı erişim verilerini kullanarak yeni kullanıcı örüntüleri bulmayı hedefler. Web kullanım madenciliği, web sayfalarının kullanma durumlarının, kullanıcı oturum sürelerinin, günlere göre ziyaretçi dağılımın analiz edilmesi ile ilgili konuları içerir. Web kullanım madenciliğinin başlıca uygulama alanları Şekil 4.3 de gösterilmiştir. Burada

web kullanım madenciliği; sistemi geliştirme, sistemi kişiselleştirme, sistemi güncelleme gibi konularda programcıya önemli bilgiler sağlar.

Şekil 4.3. Web Kullanım Madenciliğinin Uygulama Alanları

Web sunucularda tutulan kullanıcı erişim kayıtları, tarayıcı kayıtları, kullanıcı profilleri, çerezler, proxy sunucu kayıtları, fare klikleri ve kullanıcıların web ile olan etkileşimlerinden oluşan tüm kayıtlar web kullanım verisini içermektedir [21]. Oluşturulan her bir veri dosyası, günlük olarak tutulur. Tutulan veriler kullanılan sunucuya göre farklılık gösterebilir. Bu veriler ip adres, sayfa erişim tarihi, tarayıcı sistemi bilgisi, işletim sistemi bilgisi gibi veri tipleridir. Bunlar sunucu ayarları ile değiştirilebilir. Çizelge 4.3'de görüldüğü gibi web kullanım madenciliğinde kullanılan veriler sunucu kayıt dosyalarıdır.

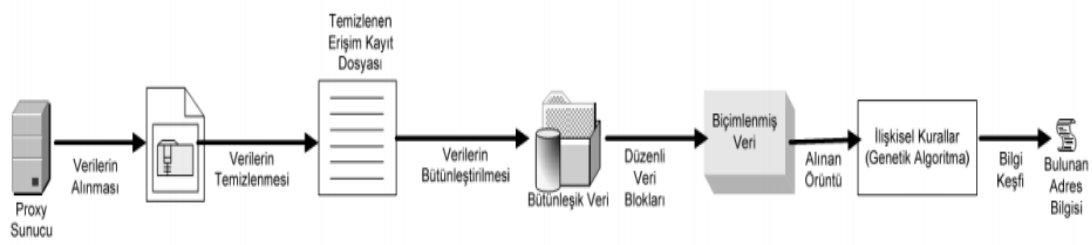
Çizelge 4.3. Web Kullanım Madenciliği Veri Durumu

Web Kullanım Madenciliği		
Veri	Verinin Şekli	Verinin Görseli
Log dosyaları	Kullanıcı etkileşimi	İlişkili tablolar

Gezer ve arkadaşları [1] yapmış oldukları web kullanım madenciliği analiz çalışmasında, İstanbul Üniversitesi Uluslararası Akademik İlişkiler Kurulu AB Eğitim birimine ait web sitesi sunucu kayıt dosyalarını incelemişlerdir. Web kullanım madenciliği günlük dosyalarının analizi ile elektronik ticaret alanında; müşteri ilgi alanlarının belirlenmesi, ürünler üzerinde yeni pazar stratejileri oluşturulması gibi hususlarda web sitesi yöneticilerine yardımcı olur.

4.4. Web Kullanım Madenciliği Aşamaları

Web kullanım madenciliği, bir veya birçok web sunucusundan alınan verilerle kullanıcı erişim desenlerinin keşfinin ve analizinin yapıldığı veri madenciliği etkinliğidir. Web kullanım madenciliğinin amacı, kullanıcının siteyi ziyaretinden sonra gerisinde bıraktığı erişim bilgilerinden yeni anlamlı veriler üretmektir. Bu veriler ikinci sınıf verilerdir, yani kullanıcının isteği dışında oluşan verilerdir. Kuruluşlar bu yolla her gün yüzlerce megabayt veri toplamaktadır. Bu bilgilerin çoğu web sunucuların otomatik olarak tuttuğu günlük dosyalarından elde edilir. Şekil 4.4'de verinin sunucudan alınıp, işlenmesi ve bilgi keşfi süreci gösterilmiştir.



Şekil 4.4. Web Kullanım Madenciliği Mimarisi

4.4.1. Ön İşlem

Web kullanım madenciliğinin ilk aşaması ön işlem aşamasıdır. Web sunucu üzerinde düzensiz olarak tutulan kullanıcı erişim dosyaları bir anlam ifade etmemektedir. Web sunucusu üzerinde tutulan kullanıcı erişim dosyalarından bilgi çıkarımı yapabilmek için gereksiz verilerden temizlenmesi ve belirli bir düzene sokulması gerekmektedir. Sunucular üzerinde tutulan kullanıcı erişim dosyalarının ilişkisiz verilerden temizlenmesi, belirli bir biçime getirilmesi ve veri tabanına aktarılması işlemine ön işlem süreci denir.

Genel olarak yapılan ön işlemler [22]:

- **Veri Ayrıştırma:** Kullanıcı erişim dosyalarından gereksiz ve ilişkisiz verilerin çıkarılması işlemidir. Bunlar **html** dosya içerisine gömülü robot istekleri ve başarısız isteklerdir.

- **Kullanıcı Kimliği**: Web sitesini ziyaret eden kişilerin web kayıt dosyaları üzerinden tespit edilmesidir. Kullanıcı tanımlama, benzer kullanıcılara ait olan aktiviteleri belirlemek için kullanılır.

- **Oturum Kimliği:** Kullanıcıların web oturumları içinde davranış ve faaliyet kayıtlarının kümelenmesidir.

- **Yol Tamamlama:** Erişim kayıtları içerisinde bulunan eksik referansları tamamlama işlemidir. Site içerisinde gezinti yapan bir kullanıcı tarayıcı üzerinden geri yaptığı zaman erişim kayıtlarında yer almayacağından yol eksik kalmış olacaktır.

4.4.2. Örüntü Keşfi

Ön işlemden geçirilen verilere veri madenciliği tekniklerinin uygulandığı aşamadır. En sık kullanılan veri madenciliği yöntemleri; **(1) istatistiksel yöntemler**, **(2) eşleştirme kuralları**, **(3) kümeleme**, **(4) sınıflandırma** ve **(5) sıralı örüntülerdir**.

4.4.3. Örüntü Analizi

Örüntü keşfi aşamasında ortaya çıkarılan kural veya örüntülerin analiz edilmesi işlemidir. Bazı örüntü analiz işlemleri:

- **Görselleştirme:** Web sayfalarını görselleştirmek için kullanılan örüntü analiz araçları geliştirme işlemidir. Bilgi keşfi aşamasında elde edilen sonuçların anlaşılabilmesi için görselleştirme tekniklerinden faydalanılır.

- **Veri ve Bilgi Sorgulama:** Analistlerin sorgu mekanizmasıyla, konu ile ilgili ve yararlı şablonlar çıkarabilmesini sağlar.

4.5. Web Madenciliği Kullanım Alanları

Günümüzde müşteri ilişkileri yönetimi yaygın bir uygulama alanıdır. Bir kurum eğer müşterileri ile öğrenen bir ilişki kurmak istiyorsa şunları yapabilmelidir. Müşterileri ne yapıyor onu fark etmeli, zaman içerisinde kendi ve müşterileri neler yapmıştır onu hatırlamalı, hatırdaki bilgilerden öğrenme ve müşterilerini daha karlı hale geçirecek uygulamalar için harekete geçmelidir [23].

Web madenciliğinin kullanım alanları:

- Sepet analizinde,
- Müşteriye özgü sayfa tasarımlarında,
- Risk analizi ve yönetiminde,
- Rekabet analizinde,
- Reklam hizmetlerinde,
- Elektronik Ticarette.

4.6. Log Dosyaları ve Türleri

Bir çok internet sunucusu çalışmalarını log dosyalarına kaydetmektedirler. Kullanılan log biçimine göre bu, istekte bulunulan sayfanın adı ve boyutu, istek sahibinin istemcisinin adı ve buna benzer birçok bilgi olabilmektedir. Elektronik ileti sunucusu durumunda da benzer bir log dosyası oluşturulmaktadır. Bu log dosyasında ise gönderenin adresi, iletiyi alanların adresleri, boyutu vs. yazılmaktadır. Aslında tüm internet sunucuları benzer özelliklere sahiptirler.

Log dosyaları inanılmaz fazla bilgi içermektedir, ancak kullanılan dosya biçiminden bilgileri algılayabilmek zordur. Bundan dolayı verileri özetlemeye ve incelemeye yarayacak bir araca gereksinim vardır. Sanal doku sunucuları için bu, en çok erişilen sayfalar, yöreler, toplam erişim sayısı ve birçok grafik anlamına gelmektedir [24].

Web kullanım madenciliği uygulamalarının asıl kaynağı web sunucularında oluşturulan web log dosyalarıdır. Dört çeşit log dosyası türü bulunmaktadır.

- **Erişim Kayıt Dosyaları**

  Bir internet sayfasını görüntülemek için web sunucusuna istek gönderen bir kullanıcı, web tarayıcısının bu istek kaydı, erişim kayıt dosyasına bir kayıt olarak kaydedilir. Erişim kayıt dosyasının formatı, bulunduğu işletim sistemine bağlı olarak farklılık gösterebilir.

- **Hata Kayıt Dosyaları**

  Web sunucunun üzerinde hata veren, gerçekleştirilemeyen işlemler için kaydedilen kayıt dosyalarıdır.

- **İstek Kayıt Dosyaları**

  İnternet kullanıcısının sayfa isteklerinin tutulduğu kayıt dosyalarıdır.

- **Etmen Kayıt Dosyaları**

  İnternet kullanıcısının kullandığı istemci bilgisayarın, işletim sistemi, web tarayıcısının adı, sürümü gibi bilgilerinin tutulduğu kayıt dosyalarıdır.

İstek ve etmen kayıt dosyasının web sunucusu üzerinde tutulup tutulmayacağı, kullanılan log dosya formatına bağlıdır

4.7. Apriori Algoritması

Veri madenciliğinde kullanılan ve veri kümeleri veya veriler arasındaki ilişkiyi çıkarmak için geliştirilmiş algoritmanın ismidir. Apriori algoritması, özellikle çok büyük ölçekli veri tabanlarındaki veriler üzerinde geliştirilmiştir. Algoritmanın asıl amacı, veri tabanında bulunan satırlar arasındaki bağlantıyı ortaya çıkarmaktır. Şekil 4.5'te apriori algoritması sanal kodu verilmiştir.

Algoritma yapı olarak her seferinde tek bir elemanı incelemekte ve bu elemanla diğer adaylarla münasebetini keşfetmeye çalışmaktadır. Algoritma, bu anlamda sığ öncelikli arama (breadth first search) yapısındadır. Algoritma adayları birer ağaç gibi düşünerek bu ağaç üzerinde arama işlemini gerçekleştirir.

Ağaç yapısında, k elemanlı bir aday listesinden k-1 elemana baktıktan sonra, alt frekans örüntüsü yetersiz olan elemanları budamakta ve kalan elemanların üzerinden arama yapmaya devam etmektedir.

$$
\begin{aligned}
&Apriori(T,\in)\\
&\quad L_1 \leftarrow \{\ l\arg e1-küme\}\\
&\quad k \leftarrow 2\\
&\quad While\ L_{k-1} \neq 0\\
&C_k \leftarrow \{c|c\in a\cup\{b\}\wedge a\in L_{k-1}\wedge b\in\cup L_{k-1}\wedge b\notin a\}\\
&\quad T'yi\ \ tara\ t\in T\\
&\quad\quad C_t \leftarrow \{c|c\in C_k\wedge c\subseteq t\}\\
&\quad\quad aday\ t\ alt\ kümlerini\ al\ \ c\in C_t\\
&\quad\quad\quad say[c]\leftarrow say[c]+1\\
&\quad\quad L_k \leftarrow \{c|c\in C_k\wedge say[c]\geq\in\}\\
&\quad\quad\quad \bigcup_{k}^{k\leftarrow k+1} L_k
\end{aligned}
$$

Şekil 4.5. Apriori Algoritması

Algoritmayı daha iyi anlamak için bir örnek üzerinden inceleyelim. Web sitesini ziyaret eden kullanıcıların URL kayıtlarının veri tabanında tutulduğunu düşünelim. Bu durumda ziyaret edilen sayfalar arasındaki ilişkileri keşfetme şansına sahip olabiliriz [25].

Örnek: Aşağıdaki küme içinde yazılmış numaralar kullanıcıların ziyaret ettikleri URL'leri, her bir küme ise kullanıcıların oturumları boyunca toplam ziyaret ettikleri sayfaları temsil etmektedir.

{1,2,3,4} ⟶ Burada kullanıcının oturumunda 1,2,3,4 numaralı URL'ler kaydedilmiştir.

{1,2}, {2,3,4}, {2,3}, {1,2,4}, {3,4}, {2,4}

Yukarıdaki her kümede, hangi URL'nin diğer hangi sayfalarla birlikte ziyaret edildiği görülmektedir. Apriori algoritmamızın ilk adımı, her URL'nin frekansını, yani kaç kere listede geçtiğini saymak olacaktır.

Çizelge 4.4. Bir Elemanlı Sayfa Frekans Tablosu

URL	Frekans
1	3
2	6
3	4
4	5

Çizelge 4.4'te, her URL'nin toplam ziyaret sayısı bulunmaktadır. Bu değere frekans veya destek ismi verilmektedir. Algoritmanın ikinci adımında, asgari destek değerini belirliyoruz. Bu belirleme işlemi, üretilen tabloya göre değişebilmektedir. Yukarıdaki örnek için asgari desteğimiz, bir numaralı URL'nin frekansı olan üç değerini alalım. Algoritmanın sıradaki adımı, URL'leri ikili gruplara ayırmak olacaktır. Burada ki amaç her elemanın diğer elemanlarla olan münasebetini bulmaktır. Yukarıdaki tabloda, frekansı düşük olan (daha seyrek olan, sık geçmeyen) elemanları eliyoruz. Bunların sonuç listesinde de yer almayacağını kabul ediyoruz. Bu defa tablomuzda sadece 2 elemanlı listeleri bulunduruyoruz.

Çizelge 4.5. İki Elemanlı Sayfa Frekans Tablosu

URL	Frekans
{1,2}	3
{2,3}	3
{2,4}	4
{3,4}	3

Çizelge 4.5'te bulunan değerler listelerdeki çiftlerden çıkmıştır. Örneğin {1,2} değeri, 3 yerde geçmektedir ve bunlar {1,2,3,4}, {1,2}, {1,2,4} dır. Dolayısıyla {1,2} ikilisi için 3, frekans değeri olarak hesaplanmış olur.

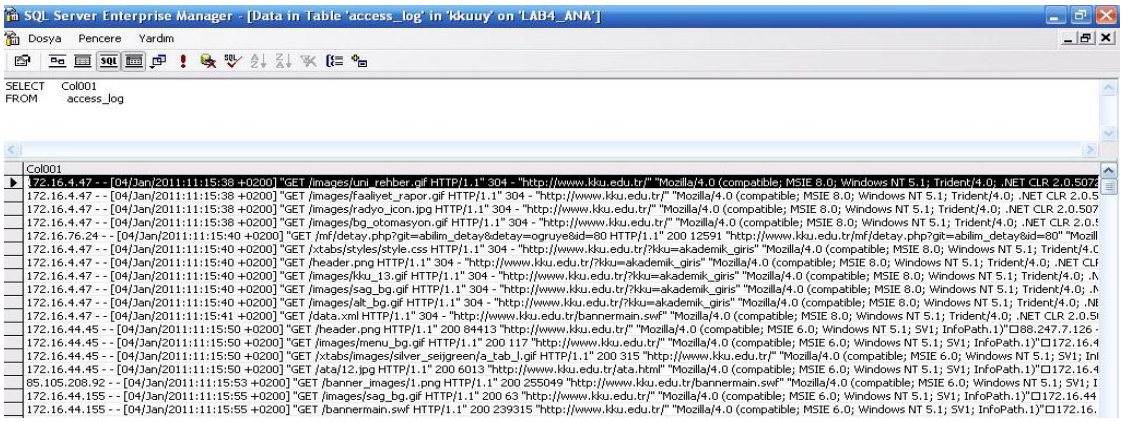
5.UYGULAMA

Kırıkkale Üniversitesi web sunucusu sistemindeki 04 Ocak 2011 - 21 Kasım 2011 tarihleri arasındaki on bir aylık web erişim günlük kayıt dosyasındaki verilerle analiz yapılmıştır. Web günlük dosyası özellikleri Çizelge 5.1 de verilmiştir. Çizelgeye göre toplam veri boyutu 39,1 GB, erişim sayısı ise 168.141.043 satırdır.

Çizel 5.1. Analiz Edilecek Dosya Özellikleri

Verinin Tarih Aralığı	04 Ocak 2011 - 21 Kasım 2011
Erişim (Satır) Sayısı	168.141.043
Toplam Veri Boyutu	39,1 GB

Günlük kayıt dosyasındaki bilgiler Şekil 5.1 de görüldüğü gibi mssql veri tabanına aktarılmıştır. Aktarılan bilgi işlenmemiştir ve gereksiz bilgilerden temizlenmesi gerekmektedir.



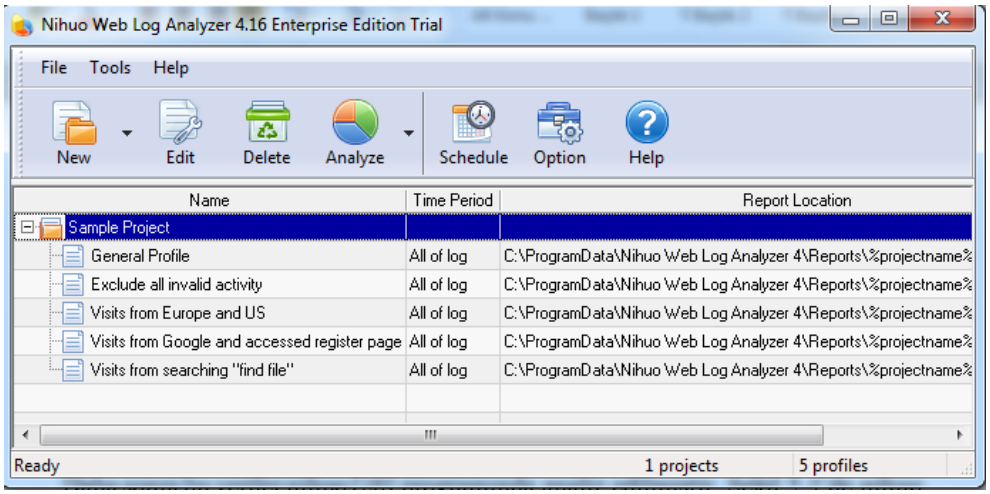
Şekil 5.1. Access_Log Dosyasından Sql Veri Tabanına Bilgi Aktarımı

Bundan dolayı veri ayrıştırılması ve temizlenmesi işlemi yapılarak web kayıt dosyasında bulunan gereksiz veriler uzaklaştırılmıştır. Veri temizlenmesi işleminde sql sorgulama dili kullanılmış ve gereksiz veriler arındırılmıştır. Böylece yeni bir log dosyası Şekil 5.2 görüldüğü gibi oluşturulmuş olur.

id	ip	tarih	metot	dosya	kod	istek	url	tarayıcı	işletim sistemi	mobil cihaz
2	172.16.4.47	04/Jan/2011:11:15:38	GET	/images/uni_rehber.gif	304	-	http://www.kku.edu.tr/""	Mozilla/4.0	Windows	Trident/4.0;
3	172.16.4.47	04/Jan/2011:11:15:38	GET	/images/faaliyet_rapor.gif	304	-	http://www.kku.edu.tr/""	Mozilla/4.0	Windows	Trident/4.0;
4	172.16.4.47	04/Jan/2011:11:15:38	GET	/images/radyo_icon.jpg	304	-	http://www.kku.edu.tr/""	Mozilla/4.0	Windows	Trident/4.0;
5	172.16.4.47	04/Jan/2011:11:15:38	GET	/images/bg_otomasyon.gif	304	-	http://www.kku.edu.tr/""	Mozilla/4.0	Windows	Trident/4.0;
6	172.16.76.24	04/Jan/2011:11:15:40	GET	/mf/detay.php?git=abilim_detay&	200	12591	http://www.kku.edu.tr/mf/detay.php?(	Mozilla/4.0	Windows	Triden"
7	172.16.4.47	04/Jan/2011:11:15:40	GET	/xtabs/styles/style.css	304	-	http://www.kku.edu.tr/?kku=akademi	Mozilla/4.0	Windows	Trident/4.0;
8	172.16.4.47	04/Jan/2011:11:15:40	GET	/header.png	304	-	http://www.kku.edu.tr/?kku=akademi	Mozilla/4.0	Windows	Trident/4.0;
9	172.16.4.47	04/Jan/2011:11:15:40	GET	/images/kku_13.gif	304	-	http://www.kku.edu.tr/?kku=akademi	Mozilla/4.0	Windows	Trident/4.0;
10	172.16.4.47	04/Jan/2011:11:15:40	GET	/images/sag_bg.gif	304	-	http://www.kku.edu.tr/?kku=akademi	Mozilla/4.0	Windows	Trident/4.0;
11	172.16.4.47	04/Jan/2011:11:15:40	GET	/images/alt_bg.gif	304	-	http://www.kku.edu.tr/?kku=akademi	Mozilla/4.0	Windows	Trident/4.0;
12	172.16.4.47	04/Jan/2011:11:15:41	GET	/data.xml	304	-	http://www.kku.edu.tr/bannermain.sw	Mozilla/4.0	Windows	Trident/4.0;
13	172.16.44.45	04/Jan/2011:11:15:50	GET	/header.png	200	84413	http://www.kku.edu.tr/""	Mozilla/4.0	Windows	SV1;
14	88.247.7.126	04/Jan/2011:11:15:50	GET	/iibf/imag"						
15	172.16.44.45	04/Jan/2011:11:15:50	GET	/images/menu_bg.gif	200	117	http://www.kku.edu.tr/""	Mozilla/4.0	Windows	SV1;
16	172.16.44.45	04/Jan/2011:11:15:50	GET	/ata"						
17	172.16.44.45	04/Jan/2011:11:15:50	GET	/xtabs/images/silver_seijgreen/a_	200	315	http://www.kku.edu.tr/""	Mozilla/4.0	Windows	SV1;
18	172.16.44.45	04/Jan/2011:11"								
19	172.16.44.45	04/Jan/2011:11:15:50	GET	/ata/12.jpg	200	6013	http://www.kku.edu.tr/ata.html""	Mozilla/4.0	Windows	SV1;
20	172.16.44.45	04/Jan/2011:11:15:50	GET	/at"						
21	85.105.208.92	04/Jan/2011:11:15:53	GET	/banner_images/1.png	200	255049	http://www.kku.edu.tr/bannermain.sw	Mozilla/4.0	Windows	SV1;
22	85.105.208.92	04/Jan/2011:11:15"								
23	172.16.44.155	04/Jan/2011:11:15:55	GET	/images/sag_bg.gif	200	63	http://www.kku.edu.tr/""	Mozilla/4.0	Windows	SV1;
24	172.16.44.155	04/Jan/2011:11:15:55	GET	/ima"						
25	172.16.44.155	04/Jan/2011:11:15:55	GET	/bannermain.swf	200	239315	http://www.kku.edu.tr/""	Mozilla/4.0	Windows	SV1;

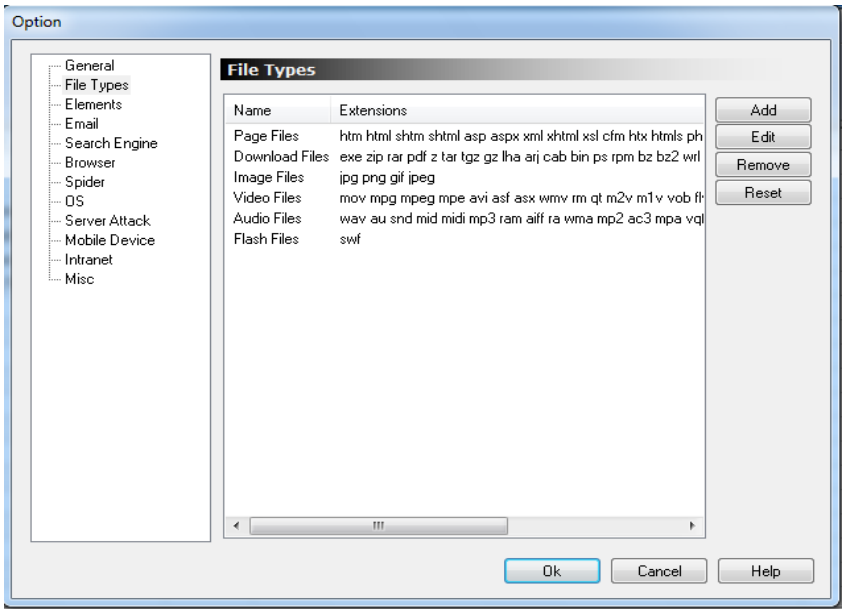
Şekil 5.2. Ayrıştırılmış Veri

Daha sonra bu veriler nihuo [26] programında analiz edilmiştir. Şekil 5.3'de nihuo programının genel görünümü gösterilmiştir.



Şekil 5.3. Nihuo Programının Genel Görünümü

Program açıldıktan sonra, yeni proje oluştur bölümünden projenin ismi belirlenir. Sonra seçenekler sekmesinden istenilen dosya tipleri belirlenir. Örneğin Şekil 5.4'de görüldüğü gibi html, aspx, php gibi dosya tiplerinin hangisinin analiz edilmesini istiyorsanız ekle, sil bölümünden ayarları yapılabilir.

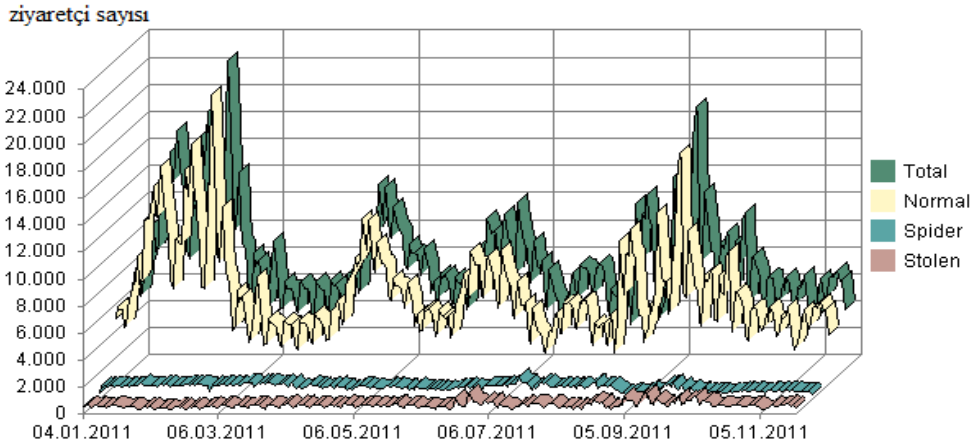


Şekil 5.4. Nihuo Programı Ayarları

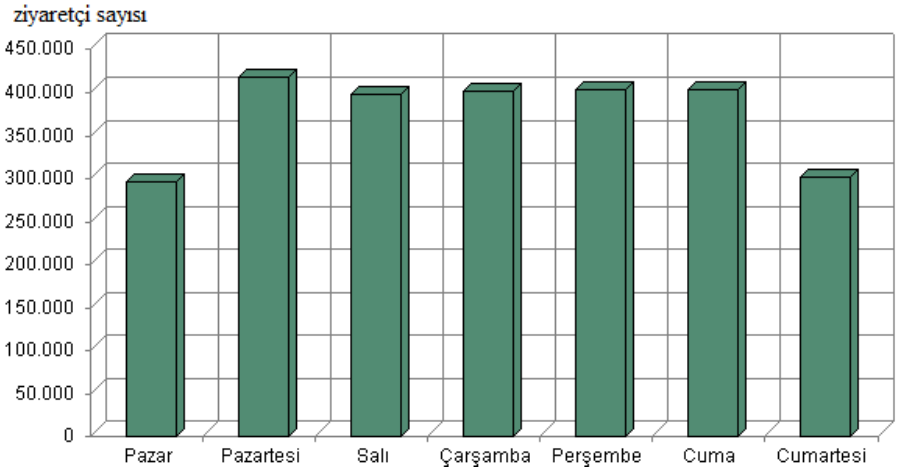
İstenilen şekilde ayarlar yapıldıktan sonra işlenecek log dosyasının yolu programa bildirilir. Sonra analiz sekmesi bölümünden veriler analiz edilir.

5.1. Aylara ve Günlere Göre Ziyaretçi Örüntüleri

Şekil 5.5'te Kırıkkale Üniversitesinin web sitesinin günlük ziyaretçi sayıları grafiksel olarak gösterilmiştir. Site ziyaretinin en çok ve en az olduğu tarih aralığını görmemiz mümkündür. Grafiğe göre şubat ayı toplam 24000 ziyaretçi ile en çok ziyaret edilen ay olmuştur. Aylara göre ziyaretçi sayılarındaki farklılıklar, dönemsel ders kayıtlarının olması, sınav sonuçlarının ilanı, hafta sonları gibi nedenlerden kaynaklanmaktadır. 2011-2012 öğretim yılı Kırıkkale Üniversitesi akademik takviminde şubat ve eylül ayında kayıt yenilemelerinin olduğu gözlenmektedir. Şekil 5.5'te şubat ve eylül ayında ziyaretçi sayılarının fazla olması kayıt yenilemelerinden kaynaklandığının göstergesidir.



Şekil 5.5. Aylara Göre Toplam Ziyaretçi Sayısı



Şekil 5.6. Haftanın Günlerine Göre Toplam Ziyaretçi Dağılımı

Ay içinde ziyaretçilerin siteyi en çok ziyaret ettiği gün Şekil 5.6'da görüldüğü gibi **pazartesi** günüdür. Kullanıcı sayısının artması sunucudan isteklerin artmasına neden olmuştur. Bundan dolayı sunucu bant genişliğinin en büyük olduğu değer pazartesini gününde olmaktadır. Sunucu band genişliği genişliğinin günlere göre dağılımı Çizelge 5.2'de verilmiştir.

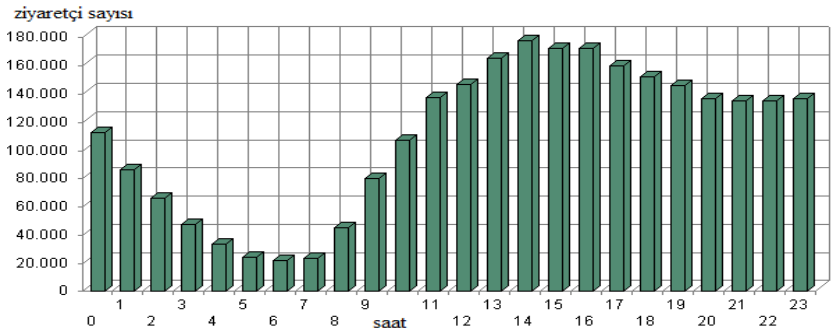
Çizelge 5.2. Haftalık Ziyaretçi Dağılımı

Günler	Tıklama	Sayfa	Ziyaret	Ortalama Oturum Süresi	Band Genişliği
Pazar	14.630.882	1.329.996	294.091	4:06	684,37 GB
Pazartesi	28.495.226	2.171.628	416.361	5:42	1,03 TB
Salı	27.010.189	2.059.688	396.151	5:40	961,40 GB
Çarşamba	28.240.113	2.159.714	400.927	5:58	969,63 GB
Perşembe	27.378.772	2.092.000	401.550	5:45	1.000,73 GB
Cuma	27.004.573	2.062.227	402.122	5:36	991,59 GB
Cumartesi	15.381.288	1.355.096	299.865	4:11	717,22 GB
Ortalama	24.020.149	1.890.049	373.009	5:22	910,71 GB

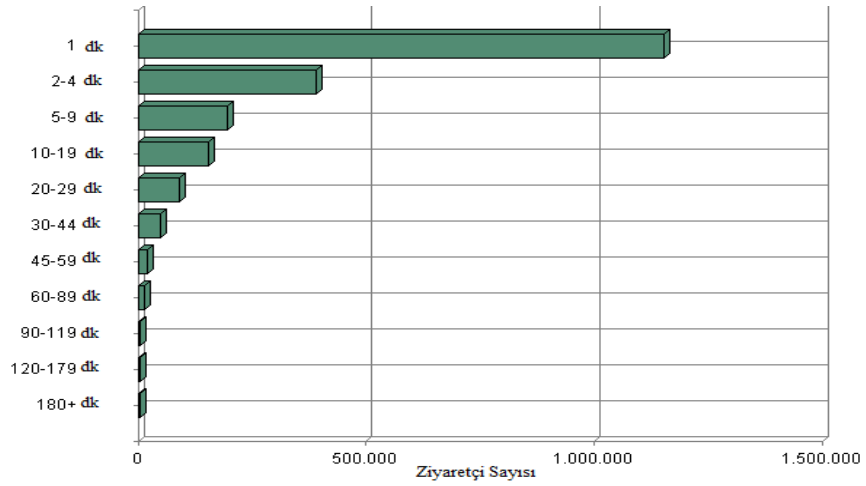
Çizelge 5.2'de ilk sütun haftanın günlerini, ikinci sütun toplam tıklama sayısını, üçüncü sütun ziyaret edilen sayfa sayısını, dördüncü sütun toplam ziyaretçi sayısını, beşinci sütun ziyaretçilerin ortalama oturum süresini, altıncı sütun sunucu bant genişliğini göstermektedir.

5.2. Ziyaret Derinliği ve Ziyaret Saatleri

Ziyaret derinliği web sitesini ziyaret eden kullanıcıların bir oturumda ziyaret ettiği sayfa sayılarıdır. Web sitesini ziyaret eden kullanıcıların siteyi en fazla 14:00 da en az ise 06:00 saat dilimlerinde ziyaret etmişlerdir. Şekil 5.8'de görüldüğü gibi kullanıcıların site üzerinde geçirdikleri süreler ise genellikle 1dakikadan fazla olmamıştır. Oturum süresini artırmak için, kullanıcıyı sitede tutmaya yönelik çalışmalar yapılabilir.

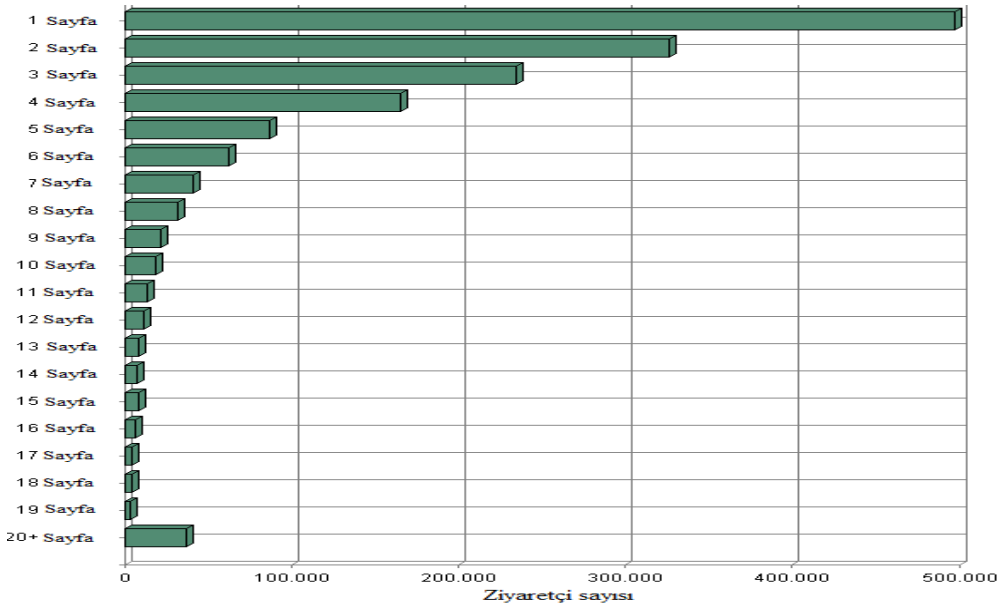


Şekil 5.7. Kullanıcıların Ziyaret Saatleri



Şekil 5.8. Kullanıcıların Ziyaret Süreleri

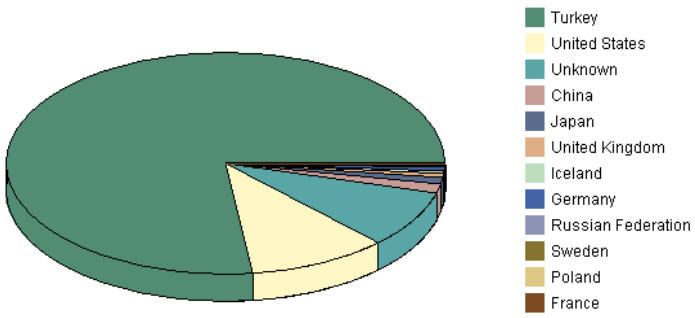
Şekil 5.9'a bakıldığında web sitesini ziyaret eden kullanıcılar en çok **bir sayfayı** ziyaret edip oturumunu sonlandırmışlardır. Yukarıda da değinildiği gibi zaten kullanıcılar genellikle sitede bir dakika kalmışlardır. Buda ziyaret derinliğinin kısalmasına neden olmuştur. Çünkü ziyaretçilerin bir oturumda ziyaret ettiği sayfa sayıları arttıkça, oturum süreleri de doğal olarak artacaktır.



Şekil 5.9. Ziyaret Derinliği

5.3. Ülke Dağılımları

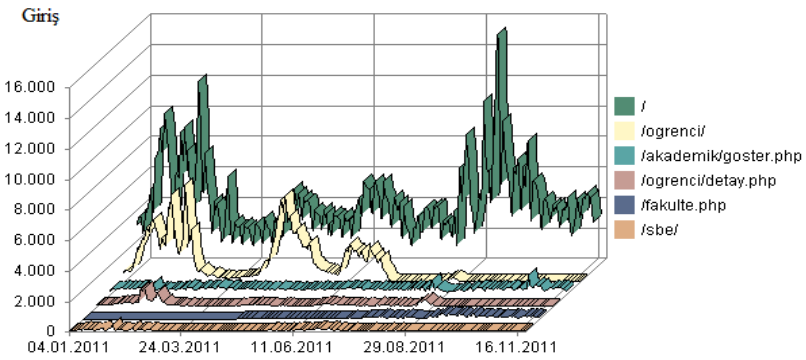
Web sitesini ziyaret eden kullanıcıların ülke dağılımını veren grafik Şekil 5.10'da verilmiştir. Burada ziyaret eden kullanıcıların ip adresleri log dosyasından elde edilerek sınıflandırılmaktadır. Öğrenci değişim programlarının yaygınlaşmasıyla, diğer ülkelerden Kırıkkale Üniversitesi web sitesine girişlerin arttığı söylenebilir.



Şekil 5.10. Ülkelere Göre Ziyaretçi Dağılımı

5.4. Günlük Giriş ve Çıkış Sayfaları

Kırıkkale Üniversitesi web sitesini ziyaret eden kullanıcıların ana sayfaya uğradıktan sonra en çok **/öğrenci/** sayfasına giriş yapmışlardır. Kullanıcıların bu sayfayı ziyaretlerinin ardından genellikle siteden çıkış yapmışlardır. Çünkü Şekil 5.11'i incelediğimizde ziyaretçilerin en çok /öğrenci/ sayfasından sonra oturumunu sonlandırmıştır.



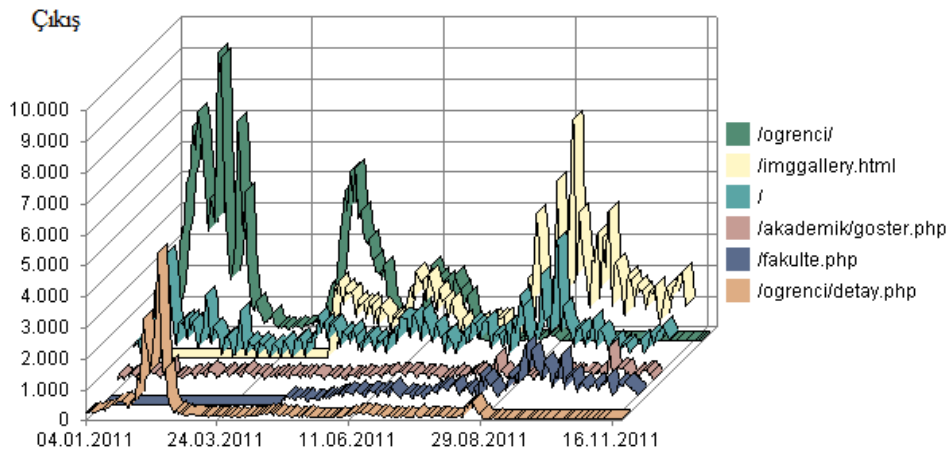
Şekil 5.11 Giriş Sayfası Grafiği

Web sitesi içerisinde en yoğun olarak kullanılan ilk 10 sayfa Çizelge 5.3 verilmiştir.

Çizelge 5.3. Web Sitesi Günlük Giriş Tablosu

	Sayfa	Giriş
1	/	1.429.082
2	/ogrenci/	345.688
3	/akademik/goster.php	118.996
4	/ogrenci/detay.php	38.473
5	/fakulte.php	21.263
6	/sbe/	18.826
7	/fen_edebiyat/detay.php	18.682
8	/mf/detay.php	13.977
9	/kutuphane/	12.692
10	/yardim/pop3/	10.297

Ziyaretçilerin oturumu sonrasında en son sayfa çıkış sayfası olarak düşünülebilir. Şekil 5.12 de görüldüğü gibi en çok oturumun sonlandığı sayfa **/öğrenci/** sayfasıdır.



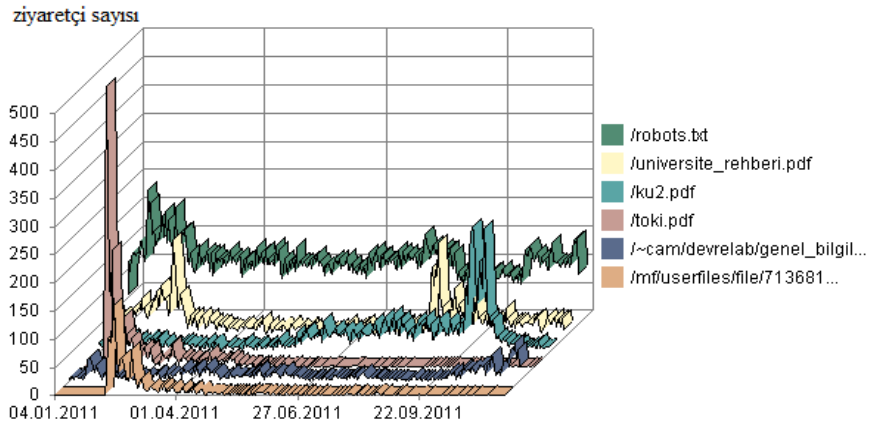
Şekil 5.12. Çıkış Sayfası Grafiği

Çizelge 5.4. Web Sitesi Günlük Çıkış Tablosu

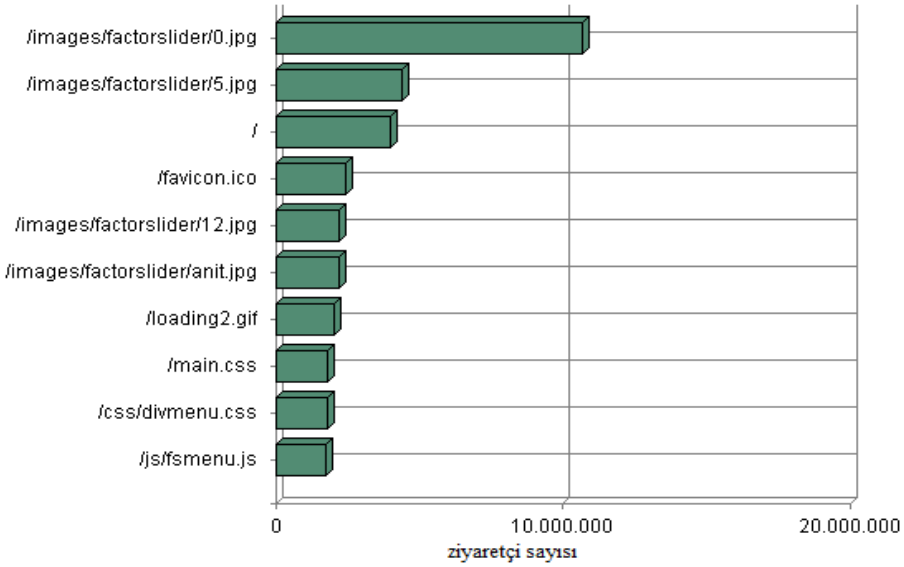
	Sayfa	Çıkış
1	/ogrenci/	441.716
2	/imggallery.html	431.926
3	/	410.779
4	/akademik/goster.php	142.962
5	/fakulte.php	111.092
6	/ogrenci/detay.php	85.857
7	/data.xml	64.581
8	/enstitu.php	40.332
9	/sbe/	32.724
10	/fen_edebiyat/detay.php	23.954

5.5. Günlük İndirilen Dosyalar

Web sitesinde en çok indirilen dosyalar sırasıyla **/universite_rehberi.pdf** ve **robots.txt** dosyasıdır. Şekil 5.13 de ayrıntılı olarak grafikle gösterilmiştir. Ayrıca Şekil 5.14 incelendiğinde **images/factorslider/0.jpg, /favicon.ico** gibi görsel dosyaların indirilme linki olmadığı halde sunucudan istekte bulunulduğu grafikle anlaşılmaktadır.



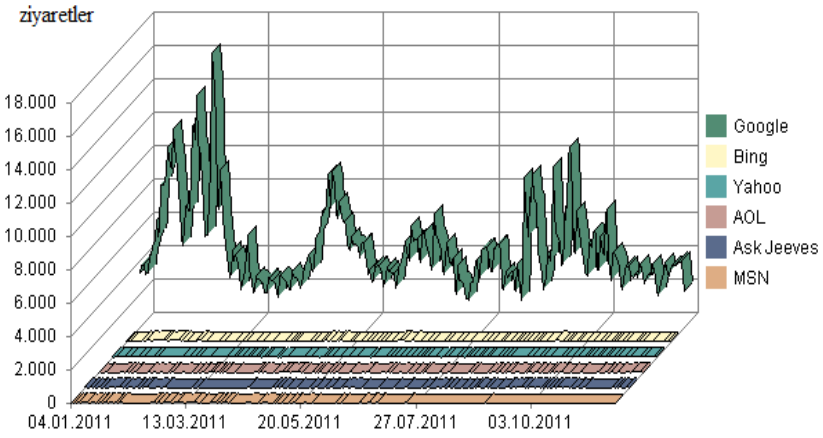
Şekil 5.13. Günlük İndirilen Dosyalar Grafiği



Şekil 5.14. Günlük İndirilen Dosyalar

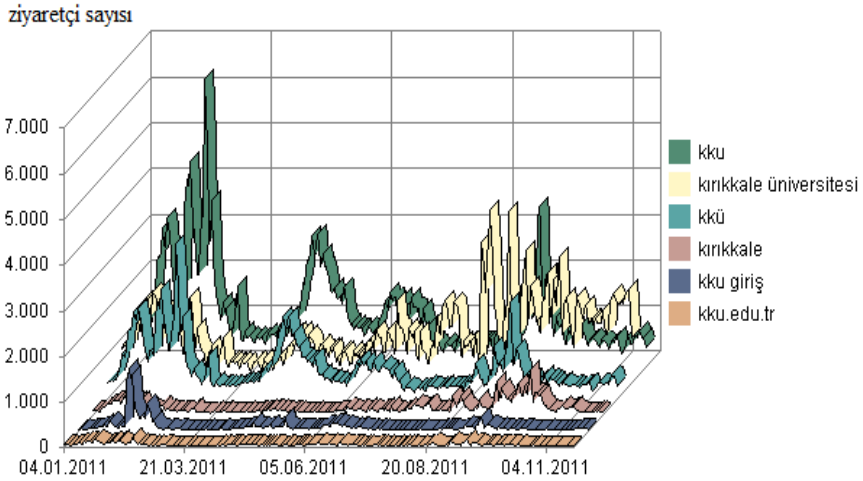
5.6. Arama Motorları ve Aranan Kelime Dizisi

Ziyaretçilerin Kırıkkale Üniversitesi web sitesinin arama motorlarından bulma oranları Şekil 5.15 de verilmiştir. Siteye en çok Google arama motorlarından ulaşılmıştır. Bu da kullanıcıların en çok tercih ettiği arama motorunun google arama motoru olduğunu göstermektedir.



Şekil 5.15. Arama Motorları Dağılımı

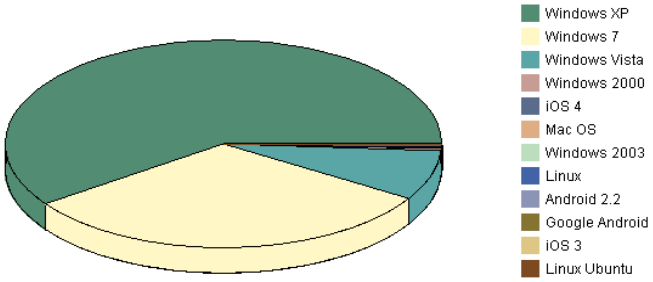
Ziyaretçilerin arama motorlarına yazdıkları kelime dizisi Şekil 5.16'da gösterilmiştir. Buna göre en çok "kku" kelime dizisi kullanılarak Kırıkkale Üniversitesi web sitesine ulaşılmıştır. Buda ziyaretçilerin, sitenin alan adı uzantısını tam olarak bilmediğini göstermektedir.



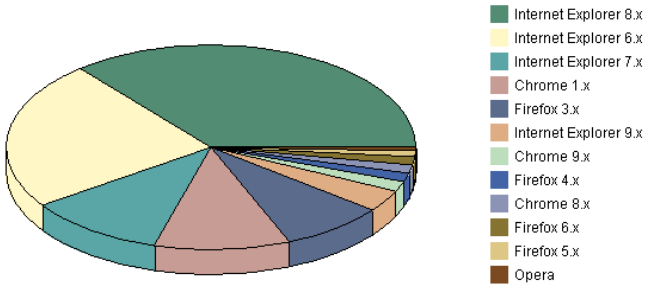
Şekil 5.16. Kelime Dizisi Grafiği

5.7. Site Ziyaretçilerinin Kullandığı İşletim Sistemleri ve Tarayıcı Dağılımı

Ziyaretçilerin kullandığı işletim sistemlerinin dağılımı Şekil 5.17'te verilmiştir. Siteye giren ziyaretçilerin çoğunluğunda Windows XP işletim sisteminin kullanıldığı görülmüştür. Kullanıcılar genellikle Microsoft tabanlı yazılımları tercih etmiştir. Microsoft tabanlı işletim sistemlerinin tarayıcı programı internet explorerdir. Bu gösteriyor ki ziyaretçiler genellikle işletim sisteminin içinde bulunan tarayıcıları kullanmışlardır.



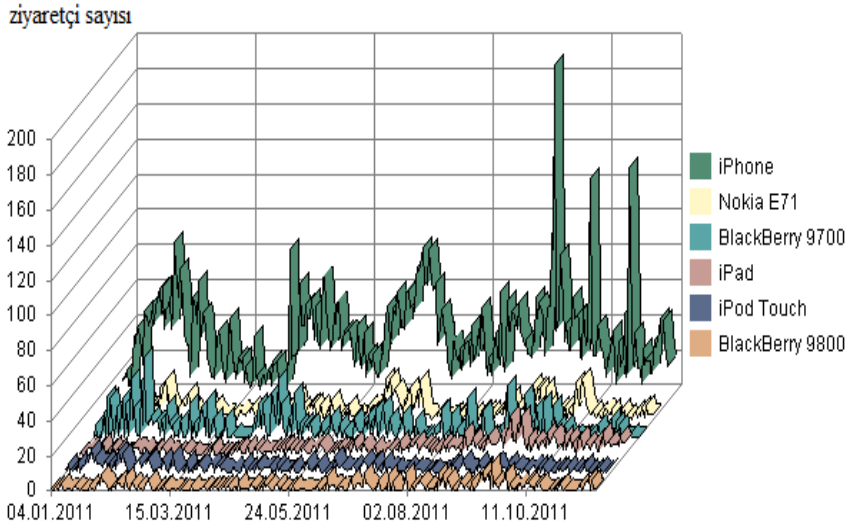
Şekil 5.17. Ziyaretçilerin Kullandığı İşletim Sistemleri



Şekil 5.18. Tarayıcı Dağılımı

5.8. Ziyaretçinin Kullandığı Mobil Aygıtlar

Kırıkkale Üniversitesi web sitesini ziyaret eden kullanıcıların kullandıkları mobil aygıt kullanım dağılımı Çizelge 5.5'te gösterilmiştir. Grafiğe göre en çok tercih edilen mobil aygıt iphonedur. Mobil aygıtların kullanımı zaman içerisinde artmaktadır. Bunun sonucunda, ileriki yıllarda kurumsal sitelerin, mobil teknolojileri desteklemesi öngörülmektedir.



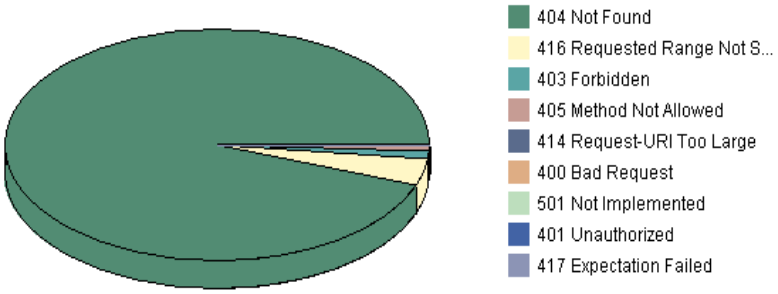
Şekil 5.19. Ziyaretçilerin Kullandığı Mobil Aygıtlar

Çizelge 5.5. Mobil Aygıt Kullanım Oranı Tablosu

	Mobil Aygıt	Ziyaretçi Sayısı	Bandwidth
1	iPhone	12.183	18,85 GB
2	Nokia E71	1.778	3,61 GB
3	BlackBerry 9700	1.607	759,40 MB
4	iPad	1.265	3,66 GB
5	iPod Touch	816	1,24 GB
6	BlackBerry 9800	618	819,99 MB
7	Nokia N97	610	1,16 GB
8	Samsung i9000	594	1,37 GB
9	Nokia N95	560	1,09 GB
10	BlackBerry 9000	205	118,54 MB
	Ara Toplam	21.560	34,64 GB
	Toplam	2.611.067	6,23 TB

5.9. Günlük Hatalar

Kırıkkale Üniversitesi web sunucusunun istekte bulunan ziyaretçilere vermiş olduğu hatalardır. Şekil 5.20'de görüldüğü gibi, sistem en çok 404 dosya bulunamadı hatasını vermiştir.



Şekil 5.20. Sunucu Hataları Grafiği

5.10. Genel İstatistikler

Çizelge 5.6 de siteye ait genel istatistikler verilmiştir. Burada haftanın en aktif günü, haftanın en durağan gününü, yıl içerisindeki en aktif günü, gün içerisindeki en aktif saat dilimi gibi bilgilere ulaşılabilir.

Çizelge 5.6. Siteye Ait Genel İstatistikler

Haftanın en aktif günü	**Pazartesi**
Haftanın en durağan günü	**Pazar**
En aktif gün	**Çarşamba, 16 Şubat, 2011**
En aktif gündeki tıklama sayısı	**1.605.821**
En aktif gündeki ziyaret sayısı	**22.731**
En aktif gündeki bant genişliği	**73,49 GB**
En durağan gün	**Cumartesi, 16 Temmuz, 2011**
En durağan gündeki tıklama sayısı	**125.199**
En durağan gündeki ziyaret sayısı	**3.245**
En durağan gündeki bant genişliği	**5,93 GB**
Gün içerisindeki en aktif saat dilimi	**14:00 - 14:59**
Gün içerisindeki en durağan saat dilimi	**06:00 - 06:59**

6. LOG ANALİZ SONUÇLARI VE DEĞERLENDİRİLMESİ

Web sitelerin de kullanıcıları tanıyabilmek ve onlara özgü uygulamalar sunmak kullanıcı memnuniyeti açısından büyük önem taşımaktadır. Web günlük dosyalarının analiz edilmesi, birbiriyle alakasız olan verilerden yeni anlamlı bilginin elde edilmesine yardımcı olacak ve bu sayede istenilen kullanıcılara yani hedef kitleye daha iyi hizmet verilmesi sağlanacaktır.

Bu tez çalışması altı bölümden oluşmaktadır. Çalışmanın üçüncü bölümünde veri madenciliği, dördüncü bölümde ise web madenciliği ve web kullanım madenciliği hakkında bilgi verilmiştir. Uygulama bölümünde ise web kullanım madenciliğini içeren Kırıkkale Üniversitesi web sitesinin log analiz uygulamasına yer verilerek site hakkında bir yıllık istatistiki bilgiler elde edilmiştir.

Kırıkkale Üniversitesi web sitesi log analizi ile toplam ziyaretçi sayısına, günün saatlerine göre ziyaretçi trafiğine, haftanın günlerine göre ziyaretlere, site ziyaretçilerinin kullandığı işletim sistemlerine, ülke dağılımlarına, tarayıcı dağılımına, tarayıcı dillerine, ziyaretçinin kullandığı mobil aygıtlara, günlük arama robotu faaliyetlerine, günlük giriş sayfalarına, günlük çıkış sayfalarına, indirilen dosya dağılımına, günlük aranan kelime dizisine, hatalara ve genel istatistiklere yer verilmiştir.

Bu çalışma Kırıkkale Üniversitesinin web sitesinin daha iyi bir görünüm ve işlerlik kazanması için web sitesi yöneticilerine önemli bilgiler sunmaktadır.

Kırıkkale Üniversitesi web sitesi analizi sonucunda;

- Ziyaretçilerin günlük kullandığı giriş sayfaları **/ (kök dizin)** sayfasıdır. Bu sayfa dizinin ismi açık yazılması web sitesinin kullanılabilirliği açısından faydalı olacaktır. Çünkü bu dizin üniversitenin web sitesinin ana sayfa dizinini kastetmektedir. En çok girilen ikinci giriş sayfası **/ogrenci/** sayfasıdır. Bu durum kullanıcıların genellikle öğrenci olduğunu göstermektedir. Ayrıca bu sayfalarda yoğunluğun olması alt sayfaların

genellikle kullanılmadığını işaret etmektedir. Sayfa isimlendirmelerine dikkat edilmesi arama robotları açısından ana dizinde anlamlı çıkmasını ve sağlayacaktır.

- Web sitesi kullanıcılarının siteyi en çok ziyaret ettikleri saat 14:00'dır. Burada sayfa istekleri yoğun olacağından site sanal sunucularla desteklenmelidir.

- **Pazartesi** günü web sitesinin en çok ziyaret edilen günü olması mesainin başladığı gün açısından mantıklıdır. Doğal olarak web sitesinin en az ziyaret edildiği gün Pazar günü olarak bulunmuştur. Site bakımının pazar günleri yapılması uygun olacaktır.

- **Eylül** ve **Ocak** aylarının en çok ziyaret edildiği aylardır. Bu aylarda yoğunluğun çok olması öğrenci ders kayıtları ile ilişkilendirilebilir.

- **/ universite_rehberi.pdf** en çok indirilen dosyadır. Bu gibi dosyaların ana sayfada gösterilmesi ulaşılmasını kolaylaştıracaktır.

- Ülkeler dağılımına baktığımızda ilk sırada Türkiye ikinci sırada Amerika Birleşik Devletleri gelmektedir. Bu yüzden site İngilizce olarak ayrı bir sayfada dizayn edilmelidir.

- Web sitesini ziyaret eden kullanıcılar site üzerinde fazla zaman geçirmemişlerdir. Kullanıcılar sitede genellikle **bir dakika** kalmışlar ve oturumlarını sonlandırmışlardır. Bu da site içinde sayfa derinliğinin azalmasına neden olmuştur. Web sitesinde kullanıcının ilgisine yönelik içeriklerin sunulması ziyaretlerin derinliğini artıracaktır.

- Sitenin analizinden çıkan diğer bir sonuç ana sayfa da bulunan görsel dosyaların sunucu tarafından her kullanıcı için bir istekmiş gibi algılaması ve cevap vermesi sunucu bant genişliğinin artmasına neden olmuştur. Sitenin tasarımında ajax, master page gibi yapıların kullanılması bunu önleyecektir.

- Günümüz de mobil aygıtların önemi artmıştır. Kırıkkale Üniversitesi web sitesini ziyaret eden kullanıcılar, en çok iphone markalı mobil aygıtlarla giriş yapmışlardır. Sitemizin **iphone** marka mobil aygıta uygun yazılımla desteklenmesi kullanımını artıracaktır.

- Dosya hatalarından en çok **404** hatasıyla karşılaşılmıştır. Dosya isimleri kurallara uygun yazılmalı, Türkçe karakter içermemelidir.

KAYNAKLAR

[1] Gezer, M., Erol, Ç., Gülseçen, S. "Bir Web Sayfasının Web Madenciliği ile Analizi",AB-2007 Akademik Bilişim Konferansı, (31 Ocak – 2 Şubat 2007), Kütahya.

[2] Takcı, H., Soğukpınar, İ. " Kütüphane Kullanıcılarının Erişim Desenlerinin Keşfi." Akademik Bilisim'02 Konferansı, Selçuk Üniversitesi, Konya, (6–8 Şubat 2002).

[3] Ugur, A., Kınacı, A. C. Yapay Zeka Teknikleri ve Yapay Sinir Ağları Kullanılarak Web Sayfalarının Sınıflandırılması. XI. Türkiye'de Internet Konferansı, (21 – 23 Aralık 2006) Bildirileri, Ankara.

[4] Wang, Y., Anthony, J. , Lee, T., Mining Web navigation patterns with a path traversal graph Original Research Article Expert Systems with Applications, Volume 38, Issue 6, June 2011, Pages 7112-7122

[5] Cooley, R., Mobasher, B., Srivastava, J. "Web Mining: Information and Pattern Discovery on the World Wide Web", Tools with Artificial Intelligence, Ninth IEEE International Conference on 3-8 November 1997, 558 – 567, USA.

[6] Iocchi, L. "The Web OEM approach to Web Information Extraction." Journal of Network and Computer Applications, (22), 259-269. 1999.

[7] H. , M. Oktay, " Web kullanım madenciliğinde birliktelik kurallarının uygulanması." Yüksek Lisans Tezi. Yıldız Teknik Üniversitesi, İstanbul, 2009.

[8] Kantardzic Mehmed: "Data Mining:Concepts, Models, Methods and Algorithms", John Wiley&Sons, 2003

[9] Akpınar, H. "Veri tabanlarında bilgi keşfi ve veri madenciliği." İ.Ü. işletme Fakültesi Dergisi. Nisan-2000, 1-22.

[10] Adriaans, P. ve D, Zantinge. 1996. “Data Mining.” Addison Wesley Longman, England, 158 P.

[11] http://www.iszekam.net/?tag=/oltp+ve+olap+farkliliklari (Erişim tarihi: 16.4.2012)

[12] Jeeffry W. S.,”Data Mining and the Search for Security: Challenges for Connecting the Dots and Database,”Government Information Quarterly,No:21, 2004, s. 463.

[13] Jeeffry W. S.,”Data Mining and the Search for Security: Challenges for Connecting the Dots and Database,”Government Information Quarterly,No:21, 2004, s. 463.

[14] Silahtaroğlu G., “Kavram ve Uygulamalarıyla Temel Veri Madenciliği.” Papatya Yayıncılık, İstanbul, 2008, s. 29.

[15] Cabena Peter ve ark. , Discovering Data Mining: From Concept to Implementation, USA, International Business Machines Corporation, 1998, s.12

[16] Dumhan Margaret H. , Data Mining Introductory and Advanced Topics, Prentice Hall, Pearson Education Inc. , New Jersey, 2003, s. 8

[17] Lori Bowen Ayre, “Data Mining for Information Professionals.” June, 2006.

[18] Elmas Ç., Yapay Zeka Uygulamaları. Seçkin Yayıncılık, Ankara, 2007

[19] Oren Etzioni, “The World Wide Web: Quagmire or Gold Mine”, in Communications of th ACM, 39(11) 39-68, 1996.

[20] http://www.bilyaz.com/index.php/22-web-madenciligi-siniflandirmasi.html/ (Erişim tarihi: 19.3.2012)

[21] http://alper-ozen.blogcu.com/web-kullanim-madenciligi/10882134(Erişim tarihi: 27.3.2012)

[22] http://www.bilyaz.com/index.php/223-web-kullanim-madenciligi.html/(Erişim tarihi: 18.4.2012)

[23] http://binedir.com/blogs/veri-madenciligi/archive/2012/05/31/ver-madenc-l-1.aspx (Erişim tarihi: 8.3.2012)

[24] http://linuxfocus.org/Turkce/September2001/article213.shtml (Erişim tarihi: 27.4.2012)

[25] Özkan Y.,Veri Madenciliği Yöntemleri. Papatya Yayıncılık, İstanbul, 2008

[26] http://download.cnet.com/Nihuo-Web-Log-Analyzer-32-bit/3000-10248_4-10211931.html (01.03.2012)