ISSN:1306-3111
e-Journal of New World Sciences Academy
2008, Volume: 3, Number: 2
Article Number: A0072

Resul Daş
İbrahim Türkoğlu
Mustafa Poyraz
University of Firat
rdas@firat.edu.tr
Elazig-Turkiye

NATURAL AND APPLIED SCIENCES
COMPUTER ENGINEERING
Received: December 2007
Accepted: March 2008

BİR WEB SİTESİNE AİT KULLANICI ERİŞİM KAYITLARININ WEB KULLANIM MADENCİLİĞİ YÖNTEMİYLE ANALİZİ: FIRAT ÜNİVERSİTESİ ÖRNEĞİ

ÖZET

Web kullanım madenciliği, web sayfalarının kullanıcı erişim örüntülerini keşfetmek için web kayıt dosyalarını analiz etmektedir. Web sunucu erişim kayıtları web siteleri hakkında çok önemli bilgiler tutmaktadır. Web sunucu erişim kayıtlarından sitedeki sayfalar arasındaki bağlantılar, siteye erişen kullanıcıların profili, web sitesinin demografik özellikleri gibi bilgiler elde edilebilir. Bu çalışmada, Fırat Üniversitesi Web sunucusuna ait kullanıcı erişim kayıtları, 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 elde edilmiştir. Bu çalışmanın sonucunda elde edilen veriler, Fırat Üniversitesi web sitesinin etkililiğini arttırmak ve geliştirmek için kullanılacaktır.

Anahtar Kelimeler: Bilgi Keşfi, Web Madenciliği, Web Kullanım Madenciliği, Kullanıcı Erişim Kayıtları

ANALYZING OF THE USER ACCESS LOGS OF A WEBSITE USING WEB USAGE MINING METHOD: EXAMPLE OF FIRAT UNIVERSITY

ABSTRACT

Web usage mining is to analysis Web log files to discover user accessing patterns of Web pages. The web server access logs records very significant information about websites. Information about a web site such as links between the pages, the users' profile, and demographic properties can be obtained from the web server access logs. In this study, user access logs of the Web server of Firat University were analyzed with Web mining software by using Web usage mining method. The end of analysis, it was found useful information about general statistics like most requested pages, file access, entry pages access, file type access, file extension access within the visited pages etc. The outcomes of the study will be used in order to development of the website to increase its effectiveness.

Keywords: Knowledge Discovery, Web Mining, Web Usage Mining, User Access Logs

1. GİRİŞ (INTRODUCTION)

Internet, dünya üzerinde var olan en büyük bilgi erişim ağıdır. Internet kullanıcıları bilgi paylaşımı, veri aktarımı ve elektronik kaynaklara erişim işlemlerini Internet ağı üzerinden yapmaktadırlar. Bu durum neticesinde, Internet üzerindeki sunucularda kullanıcı kayıt dosyalarının kapasiteleri de hızlı bir şekilde artmaktadır. Yığınla biriken kullanıcı erişim kayıtlarının incelenmesi ve analiz edilmesi Web kullanım madenciliği çalışma alanına girmektedir.

Web erişim kayıtlarından sıralı örüntülerin bulunması, kullanıcı davranışlarının tespiti gibi birçok madencilik çalışmaları geçmiş yıllarda yapılmış ve birçok farklı yaklaşımlar sunulmuştur. Uğuz ve diğ., yaptıkları çalışmada, Internet erişim kayıtlarından web kullanım madenciliği yöntemi ile web sayfası ziyaretçilerinin en sık eriştiği sayfa çiftlerini, üniversite içi ve dışı kullanıcı erişim dağılımları gibi tanımsal ilişkileri tespit etmişlerdir [1]. Daş ve diğ. makale çalışmalarında, Proxy sunucusunda tutulan Internet kullanıcı erişim kayıtlarına Genetik Algoritma yöntemini uygulayarak en çok ziyaret edilen akademik veritabanı bilgisini tespit etmişlerdir [2]. İşeri, yaptığı tez çalışmasında, geliştirdiği yazılım ile web günlüğünden zaman sınırlı bulanık bağıntı kuralları ve sıralı örüntülerin çıkarılmasını sağlamıştır [3]. Şakiroğlu ve diğ., yaptıkları çalışmalarında, web erişim kayıt dosyalarından genetik algoritma yöntemiyle sıralı erişimleri tespit etmişlerdir [4]. Gezer ve diğ., çalışmalarında, İstanbul Üniversitesi Uluslar Arası Akademik İlişkiler Kurulu AB Eğitim Birimi web sunucu kayıt dosyalarına, *WUMprep* ve *WUMweb* yazılımlarını kullanarak, kullanıcı davranışlarını analiz etmişlerdir [5]. Carus ve Mesut, geliştirdikleri Web kullanım madenciliği yazılımı ile farklı formatlardaki erişim kayıt dosyalarından istatistiksel sonuçlar elde etmeyi başarmışlardır [6]. Belen ve diğ. yaptıkları çalışmada, kullanıcı ara yüzü ve veritabanı entegrasyonu olan 3 farklı web madenciliği tekniğini ve algoritmasını kullanan, istatistiksel analiz yapan bir kayıt araştırmacısı geliştirmişlerdir [7]. Geliştirilen WALA *(Web Access Log Analyzer)* adlı sistem, en çok ziyaret edilen sayfalar, en çok indirilen dosya tipleri ve birlikte ziyaret edilen sayfaların bilgilerini tespit etmektedir. Çalışmanın hedefi, web tasarımcıları ve web yöneticileri için bir çeşit karar destek sistemi olacak yeni bir yazılım geliştirmek olmuştur. Özakar ve diğ., çalışmalarında, İzmir İleri Teknoloji Enstitüsü sunuculardan alınan kayıt dosyalarındaki ham veriyi temizleyip, java sınıfları ile ilişkisel veritabanına aktarılmaya hazır hale getirmişlerdir. Veri hazırlama bölümünde geçersiz veri ayıklanıp, veri madenciliği uygulanabilecek formata çevrilmiştir [8]. Takci ve Soğukpınar, makale çalışmalarında, kütüphane sunucu kayıtlarından, kullanıcı davranışlarını analiz yapılmışlardır [9].

Web kullanım madenciliği için geliştirilen *Nihuo* [10], *Sarg* [11], *e-WebLog* [12] gibi birçok yazılım, Internet kullanıcı davranışlarının tutulduğu farklı türdeki erişim kayıt dosyalarından istatistiksel analizler yapabilmektedir. Ayrıca, Web kayıt dosyalarının analizi ile ilgili yapılmış *NetIQ* [13], *Web Trends* [14], *Funnel Web Analyzer* [15], *Megaputer Web Analyst* [16], *Web Log Mixer* [17] gibi birçok farklı yazılımlar da mevcuttur.

2. ÇALIŞMANIN ÖNEMİ (RESEARCH SIGNIFICANCE)

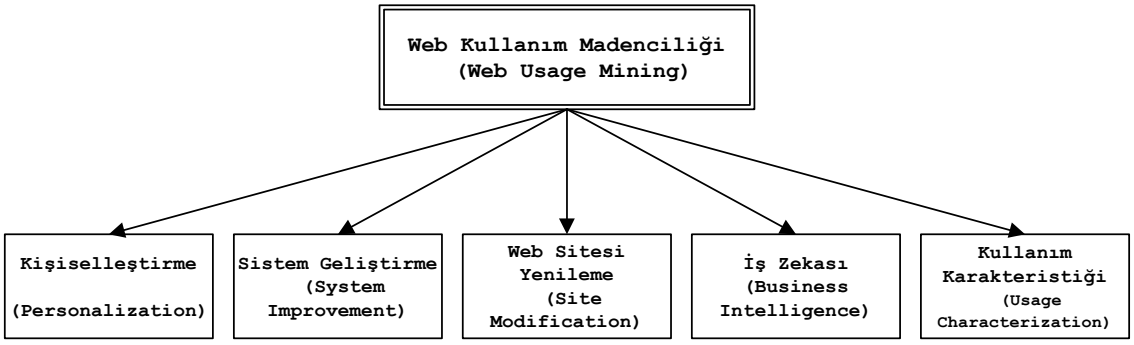
Bu çalışmanın amacı, Fırat Üniversitesi Web sunucusunda tutulan kullanıcı erişim kayıt dosyalarını Web kullanım madenciliği yöntemiyle analiz ederek, Web sitesinin gelişimine yönelik anlamlı ve ilginç bilgilerin çıkarılmasını sağlamaktır. Bu amaç doğrultusunda, F.Ü. Web

sitesine ait çıkarılan istatistikî bilgiler Web sitesinin gelişimi ve organizasyonuna büyük ölçüde ışık tutacaktır.

Bu makale çalışması 4 bölümden oluşmaktadır. Makalenin 1. bölümünde konu ile ilgili yapılmış çalışmalar irdelenmiş, 2. bölümde Web kullanım madenciliği ile ilgili teorik bilgi sunulmuştur. Ayrıca, bu konuda yapılmış uygulama ve çalışmalar da atıflarla belirtilmiştir. Makalenin 3.bölümünde Fırat Üniversitesi Web sunucusundan alınan kullanıcı erişim kayıt dosyalarının analiz uygulaması yapılarak önemli istatistiksel bilgiler çıkarılmıştır. 4. bölümde uygulama sonuçları tablolar halinde verilerek, çalışmasının değerlendirmesi yapılmış ve bu konuda öneriler sunulmuştur.

3. WEB KULLANIM MADENCİLİĞİ (WEB USAGE MINING)

Web kullanım madenciliği, Web server erişim kayıtlarından en yoğun ve ilginç kullanıcı erişim örüntülerini keşfetmek ve anlamlı verileri çıkarmayı amaçlar. Web kullanım madenciliği, Internet kullanıcı talepleri ile ilgili hizmetlerin yeterliliği, web sayfalarının kullanma durumlarını, kullanıcı oturumları ve kullanıcı davranışlarıyla üretilen erişim kayıtlarının analiz edilmesi konuları ile ilgilidir. Web kullanım verisi, web sunucu erişim kayıtları, Proxy sunucu kayıtları, tarayıcı kayıtları, kullanıcı profilleri, çerezler, fare klikleri ve sayfa kaydırmalar ve etkileşim sonuçları gibi verileri içerir [18]. Web kullanım madenciliği için temel uygulama alanları Şekil 1'de gösterilmiştir.



Şekil 1. Web kullanım madenciliğinin başlıca uygulama alanları[19]
(Figure 1. Major application areas of web usage mining [19])

3.1. Temel Tanımlamalar (Basic Terms)

Web madenciliği uygulamalarında sıkça kullanılan temel terimler mevcuttur. Makale içerisinde de kullanılacak olan ve World Wide Web Konsorsiyumu'nun (W3C) önermiş olduğu bu önemli terimler Tablo 1'de gösterilmektedir.

3.2. Veri Kaynakları ve Tipleri (Sources and Types of Data)

Web kullanım madenciliğindeki en önemli aşamalardan biride uygun bir veri kümesi oluşturmaktır. Bu verinin oluşturulmasındaki temel veri kaynakları Web sunucu erişim kayıtları ve uygulama sunucu kayıtlarıdır[20]. Web madenciliğinde kullanılabilecek veriler, içerik, yapı, kullanıcı profili ve kullanım olmak üzere 4 farklı şekilde bulunmaktadır. Bu veriler sunucu (server), istemci (client) ve Proxy sunucu gibi farklı kaynaklardan elde edilebilir. Web madenciliğinde kullanılabilen veri çeşitlerini kısaca açıklayalım.

- **İçerik (Content):** Kullanıcıların Web sayfalarında eriştiği ve kullandıkları grafik, resim, metin, şekil, ses ve görüntü dosyaları gibi gerçek verilerdir. Bunların dışında bir Web

sitesi, tanımlayıcı kelimeler, anlamsal etiketler, doküman özellikleri gibi anlamsal ve yapısal veriler de içermektedir.

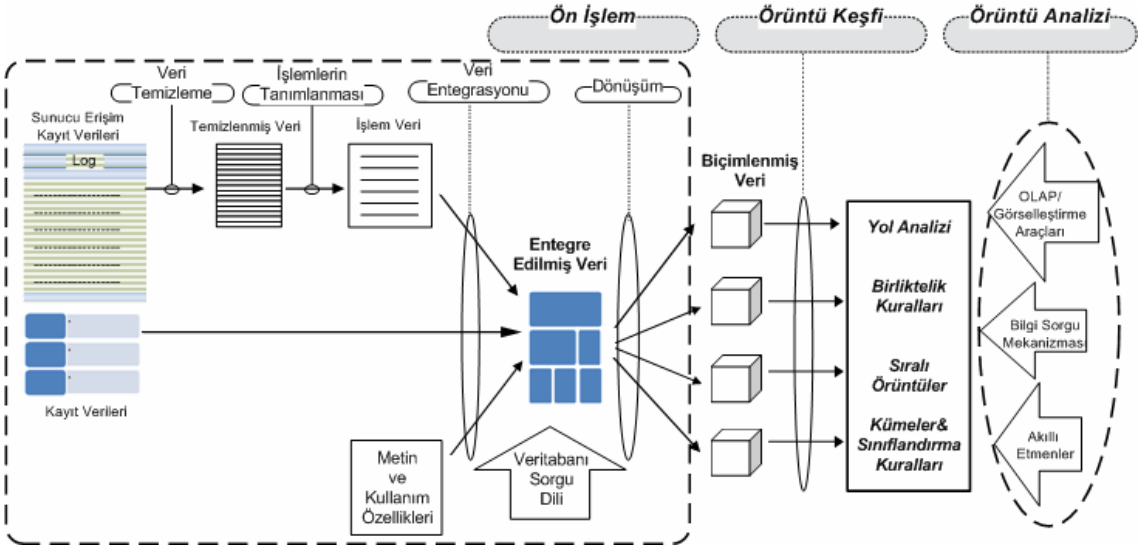
- **Kullanıcı Profili (User Profile):** Web sitesi kullanıcılarına ait demografik bilgilerinin bulunduğu verilerdir. Bir siteye kayıt olunduğunda alınan bilgiler de bu veriler içerisinde yer alır. Bu tür verilerin elde edilebilmesi için Internet kullanıcısının web sitesi ile etkileşim halinde olması gerekmektedir.
- **Yapı (Structure):** Web sitesi içeriğinin organizasyonunu gösteren veridir. Web sitesindeki yapı verisi, tasarımcı bakış açısını gösterir. Örneğin, web sitesinde bulunan sayfalar arasındaki bağlantı bilgisini, HTML (Hyper Text Markup Language) ve XML (Extensible Markup Language) dokümanları ağaç yapısını gösterir. Web sitesi yapı verisi, site haritalama araçları ile otomatik olarak oluşturulan sitenin harita bilgisidir.
- **Kullanım (Usage):** Internet kullanıcılarının web sitesinin kullanım bilgilerini gösteren verilerdir. Bu bilgiler içerisinde kullanıcı IP adresi, sayfa referansları, bağlantı saat ve tarihleri, kullanıcının Internet tarayıcısının adı ve sürümü gibi birçok bilgiler yer almaktadır. Sunucularda tutulan kullanıcı erişim kayıt dosyaları, kullanıcı davranışları ile ilgili bilgileri tutmaktadır. Kayıt dosyalardan istenilen bilgilerin elde edilebilmesi için, kayıt dosyası içindeki ilgili alanların seçilerek birbirleriyle ilişkilendirilmesi gerekmektedir. Yani, istenilen bilgilerin çıkarılması için amaca yönelik uygun alanların seçimi yapılmalıdır. Örneğin; arka arkaya ziyaret edilen Web sayfalarının sıklığını tespit etmek için erişim kayıt dosyasındaki *referer*, *request URL*, *date*, *time* alanları mutlaka seçilmelidir.

Tablo 1. Web madenciliğindeki temel terimler
(Table 1. Basic terms in web mining)

Terim	AÇIKLAMA
Web Tarayıcı	Internet üzerindeki sayfalarda gezinmeyi ve bilgi kaynaklarını aramayı sağlayan istemci yazılımlarıdır (Mozilla, Opera, Internet Explorer, vb.).
Kaynak	Yararlanılan ve özdeşliği olan her şeyi bu sınıfa koyulabilir.
Web Kaynağı	HTTP protokolleri kullanılarak ulaşılabilen herhangi bir Internet kaynağıdır.
Web Sunucu	Hazırlanan Web materyallerini Internet üzerinde erişim hizmetine açan yüksek hızda çalışan bilgisayarlardır.
Web Sayfası	Internet ortamında yayınlanabilecek farklı uzantılardaki dokümanlardır (.asp, html, php, vb).
Web Sitesi	Web sunucusu tarafından Internet'e sunulan veritabanları ve bunlarla bağlantılı belgeler, dosyalardır.
Kullanıcı	Web tarayıcısını kullanan kişi, Internet kullanıcılarıdır.
Web İsteği	Kullanıcının bir web kaynağına yapmış olduğu isteklerdir.
Ziyaret	Belli bir zamanda kullanıcı oturumu esnasında yapılmış olan web sayfalarının görüntülenmesidir.
Kullanıcı Oturumu	Bir kullanıcının sunucu üzerinde tanımlanmış ölçüde belli bir süre içerisinde kullanıcının ardı ardına görüntülediği web istekleridir.
URI	(Uniform Resource Identifier) Internet web sayfalarının fiziksel kaynağını tanımlayan karakterler zinciridir.
Oturum Tanımlama	Kullanıcının bir siteye oturum açmasından kapattığı zamana kadar geçen sürenin tanımlanarak sınıflandırılmasıdır.
Kullanıcı Tanımlama	Siteye erişen kullanıcıların browser ve kullanım özelliklerine göre sınıflandırılmasıdır.

4. BİLGİ KEŞFİ UYGULAMASI (APPLICATION OF KNOWLEDGE DISCOVERY)

Internet kullanıcılarına ait özel bilgilerin önemli olması ve gizliliğinin (internet bankacılığı, e-ticaret, kredi kartı bilgileri, vb) ön plana çıkması ile Internet'te sunulan bilgilerdeki mahremiyete karşı katı kuralların artmasına sebep olmuştur. Bu nedenle Web kullanım madenciliğinin çok daha fazla etkin, verimli ve güncel kullanılması Internet ortamına büyük kazanımlar sağlayacaktır. Internet Web kayıtlarının tutulduğu metin tabanlı kayıt (log) dosyalarından anlamlı ve gerekli olan bilgilerin çıkarılması için web kullanım madenciliği işlem basamaklarından geçirilmesi gerekmektedir. Bu işlem basamakları *Ön işlem*, *Örüntü Keşfi*, *Örüntü Analizi* şeklinde sıralanmaktadır [33 ve 34].



Şekil 2. Web kullanım madenciliği uygulamasının mimari yapısı
(Figure 2. A general architecture system for web usage mining)

4.1. Ön İşlem (Pre-Processing)

Sunucularda tutulan kayıt (log) dosyalarındaki metinsel verilerin formatı farklı türdeki kayıt dosyalarına göre birbirinden farklıdır. Metin tabanlı verilerden sağlıklı bilgi çıkarımı yapabilmek, bilgilerin doğruluk hata oranını azaltmak için kayıt dosyalarındaki gereksiz verilerin ayıklanması gerekmektedir. Sunucu üzerinde karmaşık ve düzensiz bir biçimde bulunan erişim kayıt dosyalarındaki verilerin, analiz değeri olmayan ilişkisiz sahalarından arındırılması, ayıklanması ve belirli bir düzene getirilmesi işlemi ön işlem aşamasıdır. Bu işlem çok karışık olmakla beraber, içeriğe göre farklılık gösterebilir. Burada önemli olan veri kaynağında toplanan verilerin işlenmesi ve gereksiz verilerden temizlenmesidir. Ayrıca, modelleme, sınıflandırılma ve filtreleme işlemleriyle bir sonraki evre olan Örüntü Keşfi aşamasına uygun hale getirmektir. Metinsel biçimdeki kullanıcı kayıt verileri MS Excel programı kullanılarak artık verilerden ayıklanabilir. İstenildiğinde, bu veriler herhangi bir veri tabanına da kolaylıkla aktarılabilir.

Kullanılan Erişim Kayıt Dosyalarının Yapısı: Web kullanım madenciliği uygulamasının ana veri kaynağı sunucular üzerinde biriken kayıt (log) dosyalarıdır. Farklı işletim sisteminin yüklü olduğu web sunucuların (Apache/IIS) tuttuğu kayıt dosyalarının formatı birbirinden farklıdır. Bu web kayıt dosyaları sunucu platformundan

bağımsız metin tabanlı dosyalardır. Erişim kayıt (access log), hata kayıt (error log), istek kayıt (referrer log), etmen kayıt (agent log) olmak üzere dört çeşit sunucu kayıt dosyası vardır [21 ve 35]. Yaptığımız uygulama çalışmasında, Internet kullanıcı davranışlarının analizi için Web erişim kayıt dosyaları kullanılmıştır. Web sunucuları ile Proxy sunucularının sakladığı kullanıcı erişim kayıt dosyalarının biçimleri birbirinden farklı olabilir. Örneğin, Linux işletim sistemi üzerinde çalışan Proxy sunucusu ile Windows Server 2003 işletim sistemi üzerinde çalışan bir IIS (Internet Information Server) Web sunucusunun tuttuğu erişim kayıt dosyasının biçimi birbirinden farklı olabilir. Microsoft IIS Web sunucusunda Common Log Format, Extended Common Log Format), NCSA Common Log Format gibi farklı formatlarda kullanıcı erişim kayıt (log) dosyaları tutulmaktadır.

Tablo 2. Firat web sunucu kayıtlarından örnek bir web isteği satırı (Table 2. A sample Web request from Firat's Web server logs)

2007-04-29 00:04:58 W3SVC894523 10.1.1.18 GET /inc/strbkgde.gif - 80 - 88.231.140.143 Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1; +.NET+CLR+1.1.4322) 404 64

Tablo 3. Genişletilmiş kayıt dosyası alanlarının tanımlanması (Table 3. Description of Extended log file fields)

Alan Bilgisi	**AÇIKLAMA**
2007-04-29	Bu alan kullanıcının istek yaptığı tarih bilgisini gösterir.
00:04:58	Bu alan kullanıcının istek yaptığı saat bilgisini gösterir.
W3SVC894523	Kayıt dosyasının tutulduğu klasör adının bilgisini verir.
10.1.1.18	Yayın yapan Web sunucusunun IP veya DNS adres bilgini gösterir.
GET /inc/strbkgde.gif	Kullanıcının yapmış olduğu http isteğiyle ilgili detaylı bilgi verir. Bu bölümdeki ilk kısım istek türünü (GET veya POST), ikinci kısım bağlantı yapılmak istenilen dosyayı ve son kısım bağlantı isteğinde kullanılan protokolün adını (HTTP/1.1) vermektedir.
-	RFC931 veya kimlik tanımlamalarıdır. Özel tanımlamalar yapılmadığı sürece bilgi bulunmaz.
80	Kullanıcı tarafından Web sunucusuna yapılan istek için bağlanılan port numarasını gösterir.
-	Web sitesini kullanan yetkili isimlerin listesidir. Sistem parola korumalıysa ve kimlik denetlemesini başarıyla geçmiş ise bu alanda kullanıcının adı gözükecektir.
88.231.140.143	Sunucuya istekte bulunan bilgisayarın IP veya DNS adres bilgisidir.
Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+1.1.4322)	Kullanıcı Internet tarayıcısının adı, versiyonu ve işletim sistemi hakkında bilgilerin tutulduğu alandır.
404	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.
64	İstenen sayfanın boyutunu byte cinsinden gösteren alandır.

4.2. Örüntü Keşfi (Pattern Discovery)

Örüntü keşfi, ön işlem sürecinden geçirilmiş, düzenli ve işlenebilir formattaki verilerden gerekli ve istenilen bilgiyi ortaya çıkarma işlemidir. Web kullanım madenciliğinde örüntü keşfi için birçok yöntem ve algoritma bulunmaktadır [21, 22, 23, 24, 25 ve 26]. İstatistiksel analiz, ilişkilendirme kuralları, yol analizi, kümeleme, sınıflandırma, sıralı örüntüler gibi teknikler kullanılmaktadır. Srivastava ve diğ. yaptıkları makale çalışmasında, örüntü keşfi yöntemlerini analiz ederek örneklerle bu teknikler hakkında detaylı bilgi vermişlerdir [22].

4.3. Örüntü Analizi (Pattern Analysis)

Örüntü analizi, Şekil 2'de görüldüğü gibi Web kullanım madenciliği uygulamasının son işlem aşamasıdır. Örüntü analizinin amacı; örüntü keşfi aşamasında elde edilmiş ilginç olmayan kurallar ya da örüntülerin filtrelenmesidir [22, 23 ve 24]. Genellikle örüntü analiz işlemi web madenciliği uygulamaları tarafından elde edilir. SQL, MySQL gibi veritabanı uygulamaları ve On-Line Analytical Processing (OLAP) yaygın olarak kullanılan bilgi sorgulama mekanizmalarıdır. Görsel teknikler olarak daha çok grafiksel örüntüler, farklı değerlerle yoğun ve dikkat çeken örüntüler, işaretlenmiş renkler göz önünde bulundurulmaktadır [26].

Örüntü analizi konusunda yapılmış birçok çalışma ve uygulamalar mevcuttur. Örneğin; Iocchi makale çalışmasında, Web kullanım madenciliği uygulaması ile kullanımı kolay ve anlaşılır kullanıcı ara yüzü sayesinde kullanıcının istekleri ve seçimleri doğrultusunda örüntü analizi yapılabilmektedir [24]. Örüntü analizinde önemli olan konulardan biri de, ilginç örüntülerinin nasıl öğrenileceğidir.

Web kayıt dosyalarının temizlenip, istatistikî bilgilerin elde edilmesini sağlayan birbirinden farklı Nihuo [10], Sarg [11], eWebLog [12] NetIQ [13], WebTrends [14] gibi birçok farklı program bulunmaktadır. Ancak, bu programların bir çoğunda, büyük boyuttaki kayıt dosyaları yada birden fazla kayıt dosyaları ile çalışma yapıldığında programlar kilitlenip istenilen sonucu verememektedirler. Bu noktada, en etkin ve verimli şekilde çalışarak istenilen bilgileri veren program olarak Nihuo Web Kayıt analizi programıdır.

5. UYGULAMA SONUÇLARI (EXPERIMENTAL RESULTS)

Bu çalışmada, SAS Enterprise Guide 4.1 [27] yazılımı (lisans numarası: 291468) kullanılarak metin tabanlı Fırat Üniversitesi Web erişim kayıt dosyaları ön işlem aşamalarından geçirilmiştir. Örüntü analizi işlemleri için Nihuo Web Log Analyzer [10] programı kullanılarak kullanıcı erişim kayıt dosyaları analiz edilmiştir. Nihuo Web Log Analyzer programı 2 ve 3 boyutlu grafikler şeklinde 80'den fazla istatistikî bilgiler sunabilen Web kullanım madenciliği programıdır [10]. Analiz edilen erişim kayıt dosyalarından aşağıda belirtilen bazı önemli bilgiler çıkartılmıştır.

- Günlük ve aylık olarak site ziyaretçi sayılarına ait detaylı istatistikî bilgiler
- IP ve ülke bazında ziyaretçi istatistikleri
- Web sitesinde ziyaret edilen sayfalara ait bilgiler
- Aylık ve günlük olarak toplam bant genişliğine ait raporlar
- Web sitesinde ve her web sayfada kullanıcıların harcadıkları süre miktarları
- Günlük erişilen dosya tipleri

Uygulama çalışmasında kullanılan kullanıcı erişim kayıt verileri 30 Mart 2007 Cuma/07:35:04-20 Haziran 2007 Çarşamba/23:59:54 zaman aralığını kapsamaktadır. 83 günlük bu süre içerisinde 287 farklı IP adresinden toplam 5229 kez web sunucuya bağlanılmış ve 6,86 GB'lık kayıt dosyası birikimi elde edilmiştir. Bu kayıt dosyalarının analizinde ortaya çıkarılan önemli bilgiler Tablo 4, Tablo 5, Tablo 6, Tablo 7, Tablo 8 ve Tablo 9'da verilmiştir.

Tablo 4. Kullanıcı kayıt dosyalarının genel özeti
(Table 4. A general summary of user acess log files)

No	Kullanıcı Kayıt Dosyalarından İstenen Bilgi	Çıkarılan Bilgi
1	Sunucu üzerinde en yoğun talep olan gün	Çarşamba
2	Sunucu üzerinde en az talep olan gün	Pazar
3	83 gün içinde kullanıcının en yoğun bağlandığı gün	18 Haziran 2007 Pazartesi
4	83 gün içinde kullanıcının en az bağlandığı gün	19 Ocak 2008 Cumartesi
5	Kullanıcının en yoğun bağlandığı gündeki bant genişliği	9.605,67 MB
6	Toplam günler içerisinde kullanıcıların sunucuya en yoğun bağlandıkları zaman aralığı	13:00 - 13:59
7	Toplam günler içerisinde kullanıcıların sunucuya en az bağlandıkları zaman aralığı	03:00 - 03:59

Tablo 5. En çok erişilen sayfalar
(Table 5. Most accessed pages)

No	URL Adresleri	Bağlantı Sayısı	Yüzde Oranı (%)
1	/default.asp	1697	32,58
2	/metal/saat.asp	1179	22,64
3	/perweb/default.asp	188	3,61
4	/kimmuh/kimya/CV.htm	67	1,29
5	/ogrotomasyon/	65	1,25
6	/perweb/goster.asp	61	1,17

Tablo 6. Günlük en çok indirilen dosyalar
(Table 6. Daily most downloaded files)

No	Günlük İndirilen Dosya Erişimleri	İndirme Sayısı	Yüzde Oranı(%)
1	/robots.txt	150	12,5
2	/duyuru/yazokulu.rar	40	4,1
3	/duyuru/tanitim.rar	39	3,6
4	/perweb/personel/yayinlar/fua_69/69_22586.pdf	20	2,4
5	/fenbilimleri/Dergi/17-1/ icolak rbayindir.pdf	9	1,08

Tablo 7. Günlük giriş sayfası erişimi
(Table 7. Daily access of access)

No	Giriş Sayfası Adresi	Bağlantı Sayısı	Yüzde Oranı(%)
1	/	629	59,86
2	/kimmuh/kimya/CV.htm	43	4,10
3	/bilmuh/robot/forum/calendar week.asp	30	2,88
4	/ogrotomasyon/	24	2,34
5	/perweb/default.asp	21	2,06

Tablo 8. Günlük dosya tipi erişimi
(Table 8. Daily access of file type)

No	Dosya Tipleri	Bant Genişliği	Yüzde Oranı(%)
1	Resim Dosyaları	71.498,59 MB	45,06
2	Bilinmeyen Dosyalar	33.133,15 MB	20,88
3	İndirilen Dosyalar	31.626,58 MB	19,93
4	Sayfa Dosyaları	21.639,63 MB	13,64
5	Video Dosyaları	739,11 MB	0,47

Tablo 9. Günlük dosya uzantısı erişimi
(Table 9. Daily access of file extension)

No	Dosya Uzantısı	Bant Genişliği	Yüzde Oranı(%)
1	.jpg	67.518,34 MB	42,55
2	.rar	23.929,68 MB	15,08
3	.pdf	6.566,21 MB	4,14
4	.asp	5.156,71 MB	3,25
5	.htm	4.093,96 MB	2,58

Yapılan Web kullanım madenciliği uygulaması sonucunda alınan verilere göre şu sonuçlar önemle belirtilebilir.

- Kullanıcıların Web sunucusuna yoğun talep gösterdiği gün ve saatlerde ikinci bir sunucu ile destek verilebilir. Özellikle otomasyonlara ait web sitelerini, hem sistem güvenliği hem de sunucu performansı açısından farklı sunucularda hizmet vermesi sağlanabilir.
- Boyutu büyük olan sıkıştırılmış (rar, zip) ses, görüntü ve resim dosyalarının farklı sunucu üzerinde hizmet veren bir FTP sunucuna aktarılabilir. Bu durum web sunucusu üzerindeki yoğunlaşan download bant genişliği miktarını azaltacak ve web sunucusundaki aşırı yüklenmelerden kaynaklanan kilitlenmeleri önleyecektir.
- 82, 87 ve 88 ile başlayan IP adreslerinden web sunucusuna ataklar olduğu tespit edilmiştir. Ancak, sistem üzerinde alınan güvenlik önlemleri sayesinde bu ataklar başarısız kalmıştır. Atak tespitlerine göre güvenlik duvarları üzerinde geliştirilen stratejiler yeniden gözden geçirilebilir.

6. SONUÇ (CONCLUSION)

Üniversitelerin Web sayfaları, ulusal ve uluslar arası alanda kurumun dışarıya açılmasını sağlayan en önemli araçlardan birisidir. Bu nedenle hazırlanan web sayfalarının akademik içeriği, tasarımı ve W3C (World Wide Web Consortium) standartlarına uygun geliştirilmesi büyük önem taşımaktadır [30 ve 31]. Bu durum üniversite Web sitesinin ziyaretçi sayısını ve kullanımını büyük oranda etkileyecektir. Web sitesi ziyaretçi sayısının artması da, Internet Web sunucularında tutulan kayıtlarının hızlı bir şekilde artmasını beraberinde getirmektedir. Sunuculardaki metin tabanlı Web kayıt verilerinin analiz edilip, yararlı ve gerekli bilgilerin çıkarılması ve yorumlanması Web Madenciliği teknikleriyle gerçekleştirilmektedir. Özellikle, Internet kullanıcı davranışlarının incelenmesi ve analizi Web kullanım madenciliğinin araştırma alanına girmektedir.

Bu makale çalışmasında, Fırat Üniversitesi Web sunucusunda tutulan metin tabanlı kullanıcı erişim kayıtları Web kullanım madenciliği yöntemleriyle analiz edilmiştir. Yapılan uygulama çalışmasında, Web sitesinin kullanım durumu ile ilgili sitede en çok erişilen sayfalar, dosya erişimleri, giriş sayfası erişimleri, dosya tipleri, dosya uzantıları gibi detaylı istatistiki bilgiler çıkarılmıştır. Uygulamalı bu çalışma ile Fırat Üniversitesi web sitesindeki yeni Web teknolojileri gözden geçirilmiş ve çıkarılan bilgiler ışığında Web sitesinde görülen eksiklikler ve hatalar tespit edilmiştir. Elde edilen bu bilgiler, Web sitesinin etkililiğini arttırmak ve geliştirmek için kullanılabilir. Böylece sunucuların, hem performans açısından hem de kullanıcıların kullanım kolaylığı açısından rahatlılığı ortaya çıkacaktır.

BİLGİLENDİRME (ACKNOWLEDGEMENTS)

Bu çalışmada, kullanmış olduğumuz *Web sunucusu kullanıcı erişim kayıt dosyalarını* tarafımıza sağlayan Fırat Üniversitesi Bilgi İşlem Daire Başkanlığı'na teşekkür ederiz. Çalışma, Fırat Üniversitesi Bilimsel Araştırma Projeleri (FÜBAP) Birim'inin 1526 numaralı projesi ile desteklenmiştir.

KAYNAKLAR (REFERENCES)

1. Uğuz, H., Kodaz, H., Saraçoğlu, R. ve 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, ss:45-47.

2. Daş, R., Türkoğlu, İ. ve Poyraz, M., (2006). Genetik Algoritma Yöntemiyle Internet Erişim Kayıtlarından Bilgi Çıkarılması, Sakarya Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 10(2), ss:67-72, Sakarya.
3. İseri, İ., (2005). Web Günlüğünden Zaman Sınırlı Bulanık Bağıntı Kuralları ve Sıralı Örüntülerin Çıkarılması, Fırat Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, ss:50.
4. Şakiroğlu, A.M., Tuğ, E. ve Bulun, M., (2003). Web Log Dosyalarından Genetik Algoritma Yöntemiyle Sıralı Erişimlerin Tespit Edilmesi, Türkiye Bilişim Derneği 20. Bilişim Kurultayı.
5. 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.
6. Carus, A. ve Mesut, A., (2005). Web Kullanım Madenciliği Uygulaması, II. Mühendislik Bilimleri Genç Araştırmacılar Kongresi–MBGAK 2005, 17-19 Kasım 2005, İstanbul.
7. Belen, E., Özgür, Ç. ve Özakar, B., (2003). WALA: Web Erişim Kütük Araştırmacısı (Web Access log Analyzer), (inet-tr'03) IX. "Türkiye'de Internet" Konferansı, İstanbul, 11-13 Aralık 2003.
8. Özakar, B. ve Püskülcü, H., (2002). Web içerik ve web kullanım madenciliği tekniklerinin entegrasyonu ile oluşmuş bir veri tabanından nasıl yararlanılabilir?.
9. Takcı, H. ve Soğukpınar, İ., (2002). Kütüphane Kullanıcılarının Erişim Örüntülerinin Keşfi, Bilgi Dünyası Cilt:3, Sayı:1, ss:12-26, Nisan 2002.
10. Internet: Nihuo Web Log Analyzer (NWLA), http://www.nihuo.com ve http://www.loganalyzer.net, Erişim tarihi: Ocak 2008.
11. Internet: SARG, http://sarg.sourceforge.net, Erişim tarihi: Ocak 2008.
12. Internet: eWebLog Analyzer, http://www.esoftys.com, Erişim tarihi: Ocak 2008.
13. Internet: NetIQ Web Trends Log Analyzer, http://www.netiq.com, Erişim tarihi: Ocak 2008.
14. Internet: WebTrends Marketing Web Analytics and Web Statistics, http://www.webtrends.com, Erişim tarihi: Ocak 2008.
15. Internet: Funnel Web Analyzer, http://www.quest.com/, Erişim tarihi: Ocak 2008.
16. Internet: Megaputer Web Analyticst, http://www.megaputer.com /products/wa/index.php3, Erişim tarihi: Ocak 2008.
17. Internet: Web Log Mixer, http://www.bitstrike.com/files /weblogmixer_setup.exe, Erişim tarihi: Ocak 2008.
18. Lizhen Liu, Junjie Chen, Hantao Song. (2002). The Research of Web Mining, Proceedings of the 4th World Congress on Intelligent Control and Automation, June 10-14, Shanghai/China.
19. Mobasher, B., Cooley, R., Srivastava, J. (2000). Automatic Personalization based on Web Usage Mining, Communications of the ACM, Volume:43, No:8, pp:142-151.
20. Liu, B., (207). Web Data Mining: Exploring Hyperlinks, Contents and Usage Data, (ISBN-13: 978-3-540-37881-5), pp:532, Springer.
21. Daş, R., Türkoğlu, İ., ve Poyraz, M., (2006). Genetik Algoritma Yöntemiyle Internet Erişim Kayıtlarından Bilgi Çıkarılması", 10(2), ss:67-72, Sakarya.
22. Srivasta, J., Cooley, R., Deshpande, M., and Tan, P., (2000). Web Usage Mining: Discovery and Applications of Usage Patterns From Web Data, SIGKDD Explorations. 1(2), pp:1–12.
23. Cooley, R., (2000). Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data, PhD thesis, University of Minnesota.

24. Iocchi, L., (1999). The Web OEM approach to Web Information Extraction, Journal of Network and Computer Applications, Volume:22, ss:259-269.
25. Cooley, R., Mobasher, B., and Srivastava, J., (1997). Web Mining: Information and Pattern Discovery on the World Wide Web", Tools with Artificial Intelligence, Ninth IEEE International Conference on 3-8 November 1997, pp:558–567, USA.
26. Cooley, R., Mobasher, B., and Srivastava, J., (1999). Data Preparation for mining World Wide Web Browsing Patterns, Knowledge and Information Systems 1, pp:1-27.
27. http://support.sas.com/documentation/, (last accessed: 20.01.2008)
28. 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, pp:139–152.
29. 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.
30. Internet: Configuration File of W3C, http://www.w3.org/ Daemon/User /Config/, Son Erişim tarihi: Ocak 2008.
31. Internet: Extended Log file Format, http://www.w3.org/TR/WD-logfile.html, Erişim tarihi: Ocak 2007.
32. Internet: Hypertext Transfer Protocol Overview, http://www.w3.org/Protocols, ftp://ftp.isi.edu/in-notes/rfc2616.txt, Erişim tarihi: Aralık 2007.
33. 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.
34. Feng Zhang, Hui-You Chang., (2002). Research and Development in Web Usage Mining System-Key Issues and Proposed Solutions: A Survey, Proc. of the First International Conference on Machine Learning and Cybernetics, pp: 986–990, Beijing, 4-5 November.
35. Tuğ, E., Şakiroğlu, A.M., and Arslan, A., (2006). Automatic discovery of the sequential accesses from web log data files via a genetic algorithm, Knowledge-Based Systems 19, ss:180-186.
36. Gündüz, Ş. ve Adalı, E., (2004). Web kullanıcılarının davranışları için örüntü bulma ve modelleme, İstanbul Teknik Üniversitesi, Mühendislik Dergisi, İstanbul, 3(6), ss:15-24.
37. Habegger, B. and Quafafou, M., (2004). Web services for information extraction from the Web, Web Services, IEEE International Conference on 6-9 July 04 pp:279–286.
38. Huiying, Z. and 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, IEEE, Hangzhou, China, June, pp:15-19.
39. Junjie Chen, Wei Liu., (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).
40. Oosthuizen, C., Wesson, J., and Cilliers, C., (2006). Visual Web mining of Organizational Web Sites, Proceedings of the Information Visualization (IV'06), IEEE Computer Society.