Web Sunucu Loglarının Web Madenciliği Yöntemleri ile Analizi

Işıl Çinar[1], Muhammet Serkan Çinar[2], Hasan Şakir Bilge[3]

Özet- **Bu çalışmada web sunucusu loglarının web madenciliği yöntemleri ile analiz edilmesine ilişkin literatür taraması yapılarak bu alanda yapılan çalışmalarda kullanılan algoritmalar, elde edilen sonuçlar, karşılaşılan problemler ve yapılan çalışmaların eksiklikleri incelenmiştir. Web madenciliğinde kullanılacak olan log dosyalarının formatları ve bilgi çıkarımında kullanılan alanlar detaylı olarak ele alınmıştır. Web sunucusu loglarının analizi ile çıkarılan istatistiksel sonuçlar, saldırıların tespit edilmesi, saldırı çeşitleri ve saldırıda kullanılan yöntemlere ilişkin bilgi sağlanması amaçlanmıştır.**

Ayrıca çalışmada bahsedilen veri madenciliği yöntemleri, bir analiz aracı olan WEKA (Waikato Environment for Knowledge Analysis) yazılımı kullanılarak, internette aktif olarak çalışmakta olan bir web sunucusunun loglarına uygulanmış olup elde edilen analiz sonuçlarına ilişkin değerlendirmeler sunulmuştur.

***Anahtar Kelimeler*—Log Analizi, Veri Madenciliği, Web Madenciliği, Web Güvenliği**

Abstract- In this study, the literature review of analysing web server logs via web mining methods were performed and used algorithms in this area, obtained results, encountered problems and shortcomings of studies were examined. Log files formats used in Web mining and fields in log files used for extraction of information were discussed in details. It is intended to provide information related with statistical results derived from analysing web server logs, detection of attacks, variety of attacks and methods used in attacks.

Additionally data mining methods mentioned in this study were applied to a web server logs which obtained from actively used in Internet, with using WEKA software which is an analysing tool and obtained analyze results were examined.

Keywords—Log Analysis, Data Mining, Web Mining, Web Security

1. I.Çinar, Gazi Üniv.Bilg.Müh., e-posta: (isil.cinar@gazi.edu.tr).
2. M.S.Çinar,Hacettepe Üniv.Bilg.Müh., e-posta:(mscinar@hacettepe.edu.tr).
3. H.Ş.Bilge,Doç.Dr.,Gazi Üniv.Bilg.Müh.,e-posta:(bilge@gazi.edu.tr).

I. GIRIŞ

Günümüzde bilgisayar ve internet kullanımının hızla yaygınlaşmasıyla birlikte birçok işlem artık elektronik ortamdan gerçekleştirilmektedir. Bu durum sayısal ortamdaki verinin her geçen gün hızla büyümesine sebep olmakta; bu durum karmaşıklığı beraberinde getirmekte ve bu verilerin analiz edilmesi gerekliliği doğmaktadır.

Web sunuculardaki erişim ve hata kayıt dosyalarında kullanıcıların site içinde gezinirken yaptığı tüm hareket kayıtları tutulmaktadır. Böylece hareket kayıtlarının tutulduğu dosyalar çok hızlı büyümekte ve yer kazanmak için periyodik olarak bu dosyaların temizlenmesi gerekmektedir [1]. Çoğu zaman incelenmeden silinen bu veriler analiz edildiğinde çeşitli istatistiksel sonuçlar çıkarılabileceği gibi güvenlikle ilgili önlemler almak ve saldırıları tespit etmek de mümkün olabilmektedir. Bu veri dosyalarının hızla büyüyor olması beraberinde ham veriden anlamlı sonuçlar çıkarma ihtiyacını doğurmaktadır. Veri madenciliği, bu gibi durumlarda basit olmayan, önceden bilinmeyen veya tahmin edilemeyen ve yararlı olan desenlerin ya da bilginin çok büyük miktarlardaki veriden çıkarılmasını sağlamaktadır [2]. Veri madenciliğinde amaç, analiz edilerek bilgi çıkarılması zor olan büyük veri kümelerini işleyerek anlamlı, gizli ve faydalı olabilecek bilgiyi çıkarmak; bu bilgileri içeren bir model oluşturarak yeni gelecek veri nesnesi hakkında yorum yapmayı vetahminde bulunmayı sağlamaktır [3].

Veri madenciliği, çok büyük veri kümelerinin üretildiği ve dolayısıyla veritabanlarının oluştuğu ortamlarda kullanım alanı bulabilmektedir. Uygulama alanlarına örnek olarak bilimsel ve mühendislik verileri, sağlık verileri, iş verileri, alışveriş verileri, bankacılık ve finans verileri, eğitim sektörü verileri, internet (web) verileri, doküman verileri gösterilebilir [4].

Bu çalışmada veri madenciliği ile log analizi konusunda son yıllarda yapılan çalışmalar incelenerek kullanılan algoritmalar, karşılaşılan problemler, çalışmaların eksikleri ve elde edilen sonuçlar hakkında bilgi verilmiştir.

Çalışmanın ikinci bölümünde web madenciliği ve çalışma alanlarından bahsedilmekte, üçüncü bölümde web sunucusu log dosyalarında kullanılan formatlar açıklanmakta ve log dosyalarında kullanılan alanlarla ilgili detaylı bilgi verilmekte, bir sonraki bölümde literatürdeki çalışmalar incelenerek web sunucu loglarının analizi, çıkarılan istatistiksel sonuçlar ve log analizinde saldırıların tespit edilmesi, saldırı çeşitleri ve kullanılan yöntemlerle ilgili bilgi verilmektedir. Beşinci bölümde web sunucu logları üzerinde gerçekleştirilen çalışma ve analiz sonuçları anlatılmaktadır. Altıncı bölümde gelecekte yapılması planlanan çalışmalar özetlenmektedir. Son olarak çalışmadan elde edilen kazanımlar ve sonuçlar sunulmuştur.

II. WEB MADENCİLİĞİ

Web madenciliği, web sitelerinde gezinen kullanıcıların davranışlarını inceleyerek web sitelerinin yenilenmesi, geliştirilmesi, ziyaretçilerin ilgi alanlarının belirlenmesi, reklam alma, pazarlama stratejileri oluşturma, sayfa kullanım oranlarını belirleme gibi farklı birçok konuda ve alanda karar verilmesini sağlayan bilgileri sunmaktadır [5].

Web madenciliği terimi ilk kez Etzioni tarafından 1996'da ortaya atılmıştır. Etzioni "Web madenciliğinin veri madenciliği tekniklerini kullanarak World Wide Web'de bulunan dosya ve servislerden otomatik olarak örüntüler bulmak ve öngörülmeyen bilgiye ulaşmak" olduğunu iddia etmiştir [6].

Web madenciliğinde, bir web sitesi ve o sitenin ziyaretçileri hakkında bilgi açığa çıkarmak amacıyla *istatistik yöntemi*, daha çok alışveriş uygulamalarında sepet analizi amacıyla *ilişkilendirme kuralları*, belirli zaman aralıklarındaki oturumları incelemek, eğilim analizi, değişen nokta tespiti veya benzerlik analizleri gerçekleştirmek üzere *sıralı desen yöntemi,* benzer sayfa görüntülemesi yapan kullanıcıların tespiti veya arama motorları için benzer içerikli sayfaların bir arada gruplandırıldığı *kümeleme yöntemi,* bir veriyi daha önceden tanımlanmış sınıflara dağıtmak amacıyla *sınıflandırma* yöntemi kullanılmaktadır [7].

Araştırma alanlarına gore incelendiğinde, web madenciliği içerik madenciliği (content), web kullanım madenciliği (usage) ve web yapı madenciliği (structure) olmak üzere 3 alandan oluşmaktadır [8].

Web Içerik Madenciliği

Web kaynaklarından otomatik olarak bilgi arama tekniklerini içermektedir. Verinin değişik tiplerde oluşu ve yapısal olmayışı bu konudaki teknikleri karmaşıklaştırmıştır [9]. Amaç erişilebilir web dokümanları içerisindeki saklı bilgiyi ortaya çıkarmaktır [7].

İçerik; ziyaretçilerin web sayfalarında eriştikleri grafik, resim, metin, ses ve görüntü dosyaları gibi gerçek verilerdir [10].

Web Yapı Madenciliği

Web yapı madenciliği, web sitesi ve web sayfası ile ilgili bağlantı verisine bakarak bilgi üretmektedir. Web içerik madenciliği dokümanın içeriğine bakarken web yapı madenciliği dokümanlar arası bağlantılara yoğunlaşmaktadır [9]. Yapı; web sitesi içeriğinin organizasyonunu gösteren veridir. Web sitesinde bulunan yapı verisi, tasarımcı bakış açısını göstermektedir. 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 oluşturulan sitenin harita bilgisidir [11].

Web Kullanım Madenciliği

Bu çalışma alanı, kullanıcıların web sitelerinde dolaşırken gerçekleştirdikleri erişimler sonucu oluşturulan veriden bilgi üretmeyi hedeflemektedir. Kullanım bilgileri olarak kullanıcı IP adresi, sayfa referansları, bağlantı saat ve tarihleri, kullanıcının internet tarayıcısının adı ve sürümü gibi bilgiler yer almaktadır [10]. Web Kullanım Madenciliği kullanılarak kullanıcıların genel davranış biçimlerinin ortaya çıkarılması, site içerik ve yapısında yapılması gereken düzenlemelerin belirlenmesi, sistem iyileştirme çalışmalarının gerçekleştirilmesi ve bireysel taleplere yönelik değişen sitelerin oluşturulması hedeflenmektedir [9].

Bu çalışmada, web sunucularındaki kullanıcı erişim ve hata logları üzerinde web kullanım madenciliği işlemleri gerçekleştirilerek site güncellemesi, sistem iyileştirme, kişiselleştirme ve özellikle de web ve sistem güvenliği konularında çıkarımlarda bulunulmuştur.

III. WEB SUNUCUSU LOG DOSYALARI

Bir web sunucuda sunucu logları (erişim logları), hata logları ve çerez (cookie) logları olmak üzere 3 çeşit log bulunmaktadır.

World Wide Web konsorsiyum web sunucu logları için standart formatı belirlemektedir; ancak kaydedilen verinin miktarına göre farklı formatlar da bulunmaktadır. Örneğin; IIS, 6 farklı log dosyası formatını IIS tabanlı sitelerde ve servislerde bilgilerin izlenmesi ve analiz edilmesi için sağlamaktadır. Bu formatlar [10]:

- W3C Extended Log File Format
- W3C Centralized Logging
- NCSA Common Log File Format
- IIS Log File Format
- ODBC Logging
- Centralized Binary Logging formatlarıdır.

Genel olarak kaydedilen verinin miktarına gore "Common Log Format" ve "Combined Log Format" olmak üzere iki format mevcuttur. Bu çalışmada kullanılan erişim dosyasının kayıt formatı "Combined Log Format"dır. Erişim dosyasında bulunan alanlar ve örnek değer üzerinden detaylı açıklamaları aşağıda sunulmuştur:

IP Numarası (66.249.71.181): İsteği gerçekleştirentarayıcının IP adresidir. Bazen bu alanda gelen tarayıcının alan adı da bulunabilmektedir.

Tarih ve Saat (26/Aug/2012:06:51:59 +0300): İsteğin yapıldığı saat ve tarihtir.

Istek Yapılan URL (GET /11001-arnavutkoy-balikcisi HTTP/1.1): İlk kısımda istek türü (Get ve Post) ikinci kısımda istenilen dosya (11001-arnavutkoy-balikcisi) ve son kısımda protokol verilmektedir (HTTP/1.1).

Durum Kodu (200): Bu kod sunucunun verdiği yanıtı göstermektedir. '200' kodu hiç bir sorun oluşmadığını göstermektedir. Bu kodlar RFC2616 teknik belgesinin 10. kısmında belgelendirilmiştir.

Boyut (18512): Gönderilmiş olan dosyanın Byte cinsinden boyutudur.

Başvurulan Yer: İsteğe nereden ulaşıldığı konusunda bilgi vermektedir.

Tarayıcı Bilgisi (Mozilla/4.0 (compatible; MSIE 6.0; Windows 98)): İstekte bulunan tarayıcı bilgilerini vermektedir.

İncelenen erişim dosyası alanları arasında web log dosyalarında en sık kullanılan ortak alanların IP adresi, tarih, saat, URL ve user-agent bilgileri olduğu belirlenmiştir [12].

IV. LİTERATÜRDEKİ ÇALIŞMALAR

A. Web Sunucu Loglarının Analizi, Web Madenciliği Uygulamaları, Çıkarılan İstatistiksel Sonuçlar

Web sunucu logları analizi konusunda literatürdeki çalışmalar incelendiğinde veri madenciliği ve özellikle web madenciliği konusunda detaylı bilgilerle karşılaşılmıştır. Analizlerin daha çok istatistiksel sonuçlar üretmeye yönelik yapıldığı görülmüştür.

B.Özakar ve ark., çalışmasında [9] web içerik ve web kullanım madenciliği tekniklerinin birleştirilmesi ile oluşturulmuş bir veritabanından nasıl yararlanılabileceğinden bahsetmektedir. Çalışmada İzmir Yüksek Teknoloji Enstitüsü (IYTE) web madenciliği sistem mimarisi çıkarılarak IYTE web sunucusu üzerinde biriken web erişim, hata ve kullanıcı dosyalarındaki veriyi site içerik verisi ile bir araya getirerek analiz etmek ve kullanım örüntülerini tespit etmek amaçlanmıştır. Bu çalışmada; üniversite içinden ve dışından erişim sağlayan kullanıcıların dağılımı, kullanım sürelerinin günlere ve saatlere göre dağılımı, sayfalara göre kullanıcı erişimlerinin dağılımı, ulaşılamayan sayfalar gibi tanımsal sorgulara yer verilmiştir.

Başka bir çalışmada [7] ise özellikle web kullanım madenciliği konusu üzerinde durularak 1 aylık süre için NTR web sunucu logları, web log keşif programıyla analiz edilmiştir. Burada kullanılan web log analizcisi, bilinen veri madenciliği tekniklerini erişim loglarına uygulayarak çeşitli web kullanım erişim modellerini oluşturmuştur. Bu analizler istatistiksel tanımlamalar ve birliktelik kurallarını içermektedir. Analiz sonucunda, kullanıcıların erişim modelleri, band genişliği tüketimi, haftalık ziyaretçi sayısı gibi bilgiler alınmıştır. Bu çalışma alanında yapılan araştırmada, ana zorlayıcı noktanın web robotları tarafından gerçekleştirilen girdilerin insan girdilerinden ayrılması konusu olduğu belirtilmiştir [7].

Bir başka çalışmada [13] ise e-ticaret alanında web log madenciliği çalışması yapılmıştır. Yapılan çalışmada Apriori algoritması kullanılmış ve daha önce 2005 yılında geliştirilen "Improved AprioriAll" algoritması ile performans yönünden karşılaştırması yapılarak; E-web Miner adı verilen çalışmanın zaman bakımından ve sonuçların doğruluğu bakımından diğer algoritma ile karşılaştırıldığında çok daha başarılı olduğu görülmüştür. E-web Miner çalışmasında veritabanı taramaları azaltılmış, aday kümesi çok daha küçük tutulmuştur. Sonuçta maliyet etkin ve çok daha hızlı çalışan bir algoritma elde edilmiştir.

B. Log Analizinde Saldırıların Tespit Edilmesi, Saldırı Çeşitleri ve Kullanılan Yöntemler

Log analizi ile elde edilecek istatistiksel verilerin yanında diğer bir önemli çalışma alanı da saldırıların tespit edilmesi konusudur.

Vigna ve ark. web sitelerinin ve web uygulamalarının hızla büyümesiyle uygulamaların daha karışık bir hale geldiğini ve buna paralel olarak web korsanlığı (hacking) faaliyetlerinin de arttığını belirtmektedirler [14]. Log analizinde saldırıların tespit edilmesine yönelik gerçekleştirdikleri çalışmada [14] gelen isteklerin görüntülenmesi ve analiz edilmesi için web

saldırı tespit aracı geliştirerek bu aracı güvenlik mimarisine entegre etmişlerdir. Çalışmada web uygulamalarına yönelik saldırılar 5 ana başlık altında ele alınmıştır. Aşağıda bu başlıklar hakkında detaylı bilgiye yer verilmektedir:

Zararlı Web Böceği Senaryosu: Bu senaryoda robots.txt dosyası okunarak her bir istek için URL ve user-agent header kontrol edilmektedir.

Model Eşleştirme Senaryosu: Bu senaryo URL içinde geçen bilinen saldırı örneklerini tespit etmek için kullanılmaktadır.

Tekrarlayan hatalı erişim dosyası: Bu senaryoda belirli bir subnetten veya istemciden gelen tekrarlayan hatalı erişim senaryosu kontrol edilmektedir.

Çerez Çalma Senaryosu: Çerez çalma senaryosunda bir çerez, oturum ID'si olarak birden fazla kullanıcı tarafından kullanılmaktadır [14]. Çerez çalma genellikle javascript koduyla XSS saldırısı yaparak gerçekleştirilmektedir [15].

Bellek Taşması Senaryosu: Bu atakla bellek taştığı durumda komşu hafıza alanlarının üzerine yazılmakta bu durum ise hatalara veya çökmelere neden olmaktadır.

Bir başka çalışmada ise XSS ataklarını da içeren önemli web atakları 3 ana başlık altında sıralanmaktadırlar:

Karmaşık HTTP Saldırıları: Saldırıda genellikle 80 nolu HTTP portu veya diğer HTTP bağlantıları kullanılmaktadır. İstekler HTTP ile işlendiği için; saldırganlar web sunucusuya giriş yetkisini kazanmak için bu HTTP isteklerini bozmakta veya değiştirmektedir [16].

SQL Enjeksiyon Saldırıları: SQL enjeksiyonu, kullanıcınıngirdilerinden SQL cümleleri oluşturarak web sitelerini istismar etmek için kullanılan bir saldırı tekniği olarak tanımlanmaktadır. Bu saldırı ile saldırgan, veritabanının kontrolü ele geçirebildiği gibi, sistem üzerinde de çeşitli komutlar çalıştırabilmektedir [15].

Yapılan çalışmada web saldırıları tespit aracı geliştirilmiştir. Bu araçta paket yakalama, ön işleme, saldırı tespit sistemi, denetim logları ve çıktı modülüne yer verilmiştir [16].

Yine saldırıların tespitine yönelik yapılan bir başka çalışmada ise üniversite yerel alan ağı için ağ saldırılarını tespit etmek üzere web log madenciliği kullanılmıştır. Çok büyük log dosyalarındaki anormal verileri tespit etmek için "sapan (outlier) veri tespiti" kullanılmıştır. Bu çalışmada sapan veri madenciliği yöntemi ile genel veriden farklı davranış gösteren sıradışı veri modelleri tespit edilmeye çalışılmıştır. Anomali bulmadaki genel yaklaşım; normal davranış üzerine bir profil geliştirerek bu profilden farklı olan anomalilerin bulunmasıdır [17].

Bilgi Güvenliği Akademisi (BGA) tarafından yapılan çalışmada ise saldırı olarak değerlendirilebilecek durumlar saldırı yapılan sunucuya özel bazı dizin/dosyaların istenmesi,sunucuya bağlantı kuran ip adresleri ve bağlantı sayıları, dizin/klasör gezinme denemelerini bulma, SQL enjeksiyonu denemelerini yakalama başlıkları altında ele alınmıştır [18].

Bu çalışmamızda WEKA kullanılarak öncelikle veri setinden çıkarılacak istatistiksel sonuçlar incelenmiş ardından, incelenen veri setinde herhangi bir saldırı örüntüsü olup olmadığını tespit etmek amacıyla WEKA'da bulunan çeşitli algoritmalardan faydalanılmıştır.

V. WEB SUNUCUSU LOGLARI ÜZERİNDE YAPILAN ANALİZ ÇALIŞMASI

Çalışmada çeşitli restoranlarda rezervasyon yapılabilen bir web sitesine ait web erişim log dosyası ve hata günlüğü dosyaları üzerinde analiz gerçekleştirilmiştir. Web Madenciliği ile ilgili uygulamalar birçok veri madenciliği uzmanının da tercih ettiği WEKA yazılımı kullanılarak gerçekleştirilmiştir.

WEKA başlangıçta bir proje olarak başlayıp bugün dünya genelinde birçok kişi tarafından kullanılmaya başlanan bir veri madenciliği uygulaması geliştirme programıdır. WEKA, java platformu üzerinde geliştirilmiş açık kaynak kodlu bir programdır. WEKA programında veri işleme, veri sınıflandırma, veri kümeleme, veri ilişkilendirme gibi özellikler bulunmaktadır. Projenin amacına göre programın içerisinden sınıflandırma, kümeleme, ilişkilendirme gibi çeşitli algoritmalar arasından uygun olanı seçilerek veriler üzerine uygulanmakta ve en doğru sonucu veren algoritma seçilebilmektedir [19].

WEKA kullanılarak analiz edilecek olan web log erişim dosyasının özellikleri Tablo 1'de gösterilmiştir.

TABLO I
ERİŞİM DOSYASI ÖZELLİKLERİ

Alanlar	**Değerler**
Dosyanın Adı	access_websitem.log.1
Dosyanın Boyutu	313.413.409 KB
Tarih Aralığı	01.10.2012-29.11.2012
Erişim Sayısı	575.576

Hata günlüğü dosyasının özellikleri Tablo 2'de gösterilmiştir.

TABLO 2
HATA GÜNLÜĞÜ DOSYASI ÖZELLİKLERİ

Alanlar	Değerler
Dosyanın Adı	error_websitem.log
Dosyanın Boyutu	86.6 KB
Tarih Aralığı	01.10.2012-29.11.2012
Hata Sayısı	430

A. Önişleme Adımları

Web log dosyaları üzerinde web madenciliği konusunda yapılan çalışmaların önemli bir kısmını önişleme adımı oluşturmaktadır. Web log dosyalarında gereksiz bilginin çok büyük yer tutması, önişlemeyi zorunlu kılmaktadır. Bu işlem özellikle web site yöneticileri ve kullanıcıları tarafından web log dosyalarının yönetimini kolaylaştırmaktadır.

Önişleme aşamasında veri temizleme, veri filtreleme, veri birleştirme gibi farklı teknikler kullanılmaktadır.

Web log dosyalarındaki ön işleme süreci tüm veri madenciliği sürecinin %80 ini kapsamaktadır. Önişleme adımı daha sonraki adımlar olan örüntü keşfi ve örüntü analizinin daha etkili ve kolay yapılmasını sağlamaktadır [12].

Madencilik işlemlerinin kolaylıkla yapılabilmesi için verinin etkinliğini artırmak üzere verinin önişlenmesi gerekmektedir. Veri ön işlemede yapılması gereken ana işlem gürültülü ve ilgisiz verinin temizlenmesi ve verinin azaltılmasıdır [20].

Aye yaptığı çalışmada veri temizleme algoritmasıyla analiz edilen verideki gereksiz olan kısımları çıkarmıştır. Log girdileri bölünmesi gereken çeşitli alanlar içermektedir. Log dosyasındaki tek bir satırdan alanların bölünerek çıkarılması işlemine "alan ayrıştırma" adı verilmektedir. Burada en çok kullanılan ayırıcı karakterler ',' veya 'boşluk' karakteridir. Aye yaptığı çalışmada bu işlemi gerçekleştirmek için ayrıştırma algoritması oluşturmuştur [20].

Bu çalışmada alan ayrıştırma işlemini gerçekleştirmek için log dosyası excel ortamında alanlara göre ayrılarak yüklenmiştir. Bu alanlar standart web log dosyasında bulunan temel alanlara göre belirlenmiştir.

- Identity ve user alanlarında veri tutulmadığı için bu alanlar silinmiştir.
- Zamanla ilgili sütunlar birleştirilmiştir.

Dönüştürme işlemi tablo alanları için uygun veri tipleri seçilerek tamamlanmıştır.

Veriler üzerinde, gürültü temizleme işlemini daha hassas gerçekleştirebilmek ve analiz etmek amacıyla çeşitli sorgular yazmak üzere Excel'den Access veritabanına veriler aktarılmıştır.

Veri temizleme aşamasında ise analiz edilen log dosyasına özel olarak analiz amacı için önem arzetmeyen nesneler (stil dosyaları, grafik dosyaları, ses dosyaları vb.) temizlenmektedir [20].

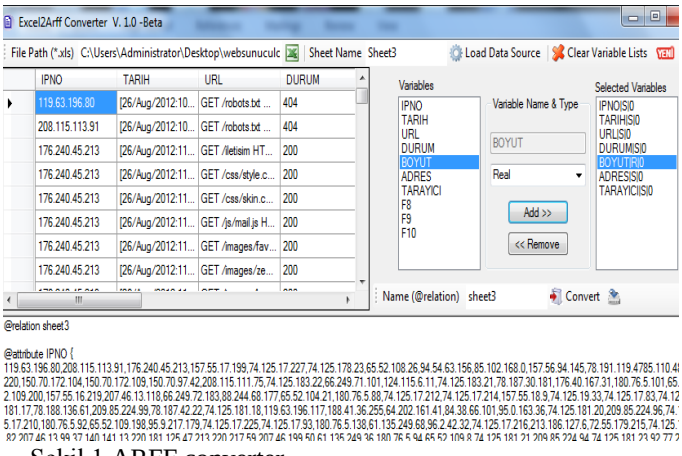
Bununla birlikte bazı özel resim sitelerinde veya haber sayfalarında kullanıcıların ilgisinin analiz edilmesi gibi bir durumda özellikle bu dosyalara odaklanmak da gerekebilmektedir. Burada önemli olan web log madenciliğinin amacının doğru olarak belirlenmesidir [21].

Yapılan çalışmada resim ve diğer bağlı olan dosyalar olmadan kaç ziyaret olduğunu tespit etmek amacıyla .gif, .jpg, .css, .js, .png, javascript uzantılı dosyalar silinmiştir.

Veride "boyut" alanındaki değerler sürekli olduğu için bu alan 4 sınıfa ayrılarak her bir aralığa yeni atamalar yapılmıştır.

Önişlemden geçirilen veri setinin WEKA programına yüklenebilmesi için ARFF formatına dönüştürülmesi gerekmektedir. ARFF, CSV, C4.5 formatında bulunan dosyalar WEKA programına eklenebilmektedir. Herhangi bir text dosyasındaki veriler WEKA ile işlenememektedir [19].

Excel dosyasını ARFF formatına dönüştürmek için Şekil 6.6'da görülmekte olan Excel2ARFF Convertor adlı program kullanılmıştır.



Şekil 1.ARFF converter

Bu programda dönüştürülmek istenen işlenmiş olan log dosyası yüklendikten sonra değişkenlerin tipleri belirlenmekte ve son olarak dönüşüm işlemi tamamlanmaktadır.

Gürültülü Verinin Temizlenmesi

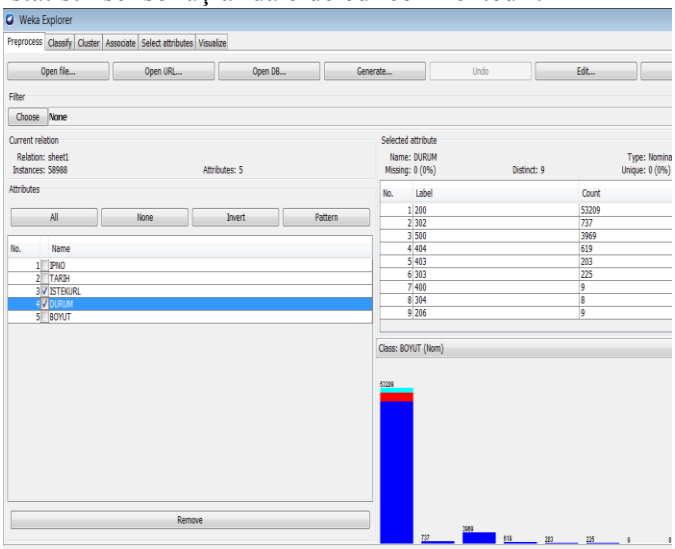
ARFF dönüştürücü programı ile dönüşüm yapıldığı sırada dosyanın içeriğine ARFF formatına uymayan

yabancı karakterler eklendiği görülmüştür ve bu karakterler temizlenmiştir. Örneğin bazı satırların başına veya sonuna dönüşüm sırasında /par ifadesi eklendiği görülmüştür.

Hataların diğer sebepleri arasında ARFF formatında tek tırnak (‘), noktalı virgül (;), boşluk( ) gibi karakterlerin bulunması yer almaktadır. Bu karakterler Excel ortamında temizlendikten sonra yeniden ARFF formatına dönüşüm sağlanmıştır.

B. WEKA’dan Elde Edilen Sonuçlar

WEKA’ya ARFF uzantılı dosya yükledikten sonra aşağıdaki önişlem adımı görüntülenmektedir. Bu adımda çeşitli filtreleme işlemleri yapılabileceği gibi çeşitli istatistiksel sonuçlar da elde edilebilmektedir.



Şekil 2. WEKA’da Önişlem Penceresi

Erişim log dosyasından elde edilen erişim ve ziyaretçi bilgilerine ait istatistiksel sonuçlar Tablo 3’de gösterilmiştir.

TABLO 3
ERİŞİM LOG DOSYASI İSTATİSTİKLER

Erişim Bilgileri	**Değerler**
Toplam Veri	575.576
Önişlemden Sonra Veri Sayısı	58.988
Günlük Ortalama Erişim	983.13
Ziyaretçi Başına Ortalama Erişim	8.02
Başarılı İstek Sayısı	53.209

Ziyaretçi Bilgileri	**Değerler**
Toplam Ziyaretçi	7347
Ortalama Günlük Ziyaretçi	122.45

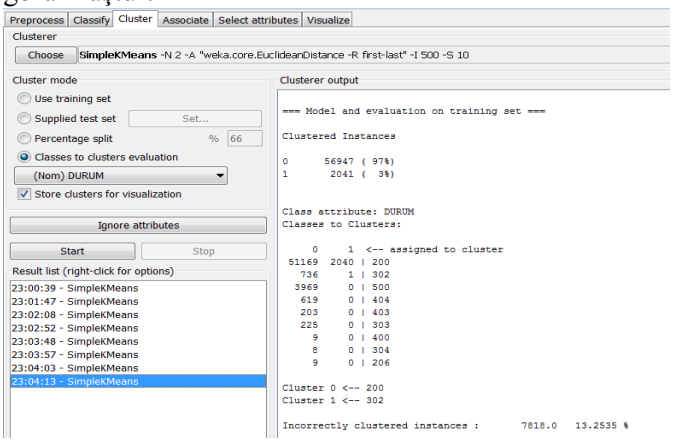
Hata log dosyasından elde edilen istatistiksel sonuçlar Tablo 4’de gösterilmiştir.

TABLO 4
HATA LOG DOSYASI İSTATİSTİKLER

Alanlar	**Değerler**
Toplam Veri	430
Alınan Hatalar	Dosya bulunamadı- 420 kez
	Sunucu ayarlarıyla istemci reddedildi- 1 kez
	İstek içerisinde geçersiz URI- 9 kez

Kümeleme, verideki benzer kayıtların gruplandırılmasını sağlayan veri madenciliği tekniğidir. Kümeleme yönteminde her kayıt var olan kümelerle karşılaştırılmakta ve kendisine en yakın kümeye atanmaktadır. Bu yeni gelen kayıtla kümeyi tanımlayan değer de değişmektedir. Optimum çözüm bulununcaya kadar kayıtlar yeniden atanmakta ve küme merkezleri ayarlanmaktadır [22].

Durum alanı için kümeleme algoritmalarından K-Means çalıştırılmış; 200 (OK-Sayfa görüntülendi) ve 302 (Found- Bulundu) alanları için 2 küme oluştuğu görülmüştür.



Şekil 3. Durum Alanı için K-Means Kümeleme Çıktısı

Her iki durum kodu da erişimin sağlandığı sayfada bir sorun olmadığında verilen durum kodudur ve incelenen logun verilen zaman aralığında herhangi bir saldırıya uğramadığı bilindiğinden 404, 401, 500 gibi durum kodları ile bir küme oluşmadığı görülmektedir.

Bir başka inceleme ise Apriori algoritması kullanılarak gerçekleştirilmiştir. Apriori algoritması, genel anlamda birliktelik kuralı çıkarımında kullanılan bir algoritmadır. Algoritmanın amacı, veri tabanında bulunan satırlar arasındaki bağlantıyı ortaya çıkarmaktır.

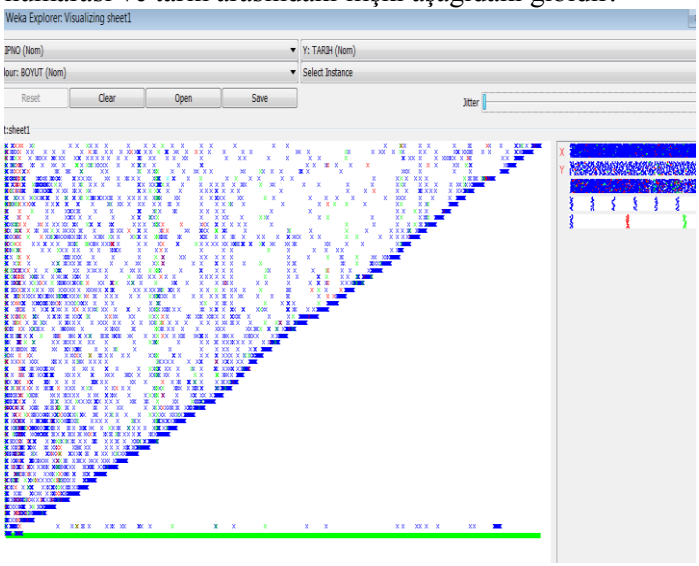
Bu algoritmada sık geçen öğe kümelerini bulmak amacıyla birçok kez veritabanı taranmaktadır. İlk

taramada tek elemanlı minimum destek metriğini sağlayan sık geçen öğe kümeleri tespit edilmekte; sonraki taramalarda önceki taramada tespit edilen sık geçen öğe kümeleri aday kümeler olarakyeni sık geçen öğe kümelerini üretmek için kullanılmaktadır. Aday kümelerin destek değerleri tarama yapılırken hesaplanmakta ve aday kümelerden minimum destek metriğini sağlayan kümeler o taramada üretilen sık geçen öğe kümelerini oluşturmaktadır. Belirlenen sık geçen öğe kümeleri bir sonraki tarama için aday küme olmaktadır. Bu döngü yeni bir sık geçen öğe kümesi bulunmayana kadar devam etmektedir [1].

WEKA'da bu algoritma kullanılarak minimum destek değeri 0.1 verilerek, güven değeri 0.9 olarak belirlenmiş ve İstekURL, Durum ve Boyut alanları arasında birliktelik kuralları oluştuğu görülmüştür. Birliktelik kuralından çıkarılan örnek sonuçlar:

- Iphone'dan bağlanılan restoran_bilgi sayfasında; boyut değerinin 1 (<50.000) olduğu görülmüştür; bant genişliğinin çok kullanılmadığı sonucuna varılmıştır.
- Anasayfa olan index.php sayfasına yapılan erişimlerde sayfanın başarıyla görüntülendiği görülmüştür. Oluşan 10 birliktelik kuralında da DURUM=200 ve/veya BOYUT=1 olması ve IP numarasına göre birliktelik kuralının oluşmamış olması bir saldırı yaklaşımının bulunmadığını göstermektedir.

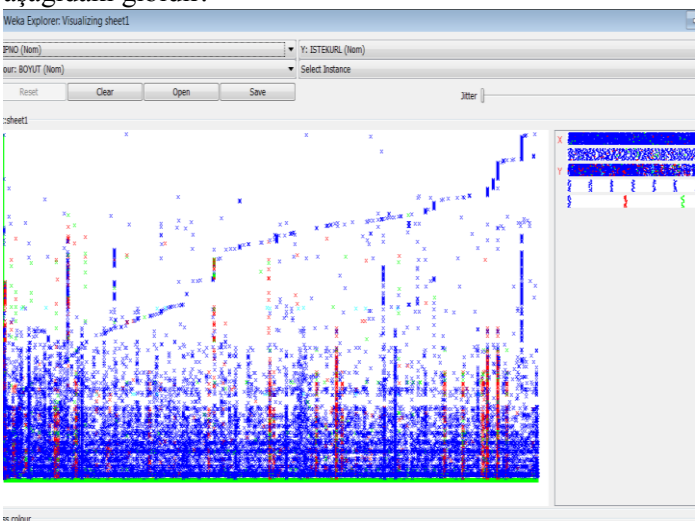
WEKA'da Visualize penceresinden çeşitli sonuçlar elde edilebilmektedir. Aşağıda erişim dosyasındaki çeşitli alanların birbirleriyle ilişkileri gösterilmiştir. IP numarası ve tarih arasındaki ilişki aşağıdaki gibidir:



Şekil 4. IP Numarası-Tarih Arasındaki İlişki

Burada x ekseni IP numarasını, y ekseni ise TARIH'i göstermektedir. x eksenine paralel çizgi şeklinde görülen kayıtlar aynı tarihte çok farklı IP adreslerinden sayfaya erişildiğini göstermektedir. Bu tarihin 24 Kasım öğretmenler günü olduğu görülmüştür.

IP numarası ve istek yapılan URL arasındaki ilişki aşağıdaki gibidir:



Şekil 5. IP Numarası-İstek Yapılan URL Arasındaki İlişki

Bu görüntüde de x eksenine doğru yoğunluk olduğu görülmektedir. Bu durum bize bazı sayfaların diğerlerine göre çok daha sık ziyaret ediliğini göstermektedir.

VI. SONUÇ VE ÖNERİLER

Bu çalışma ile web günlük iz bilgileri ile ilgili detaylı bilgi verildikten sonra web sunucu logları üzerinde web madenciliği konusunda yapılabilecekler hakkında literatürdeki çalışmalardan faydalanarak açık kodlu WEKA programı kullanılarak uygulama ile bir prototip oluşturulmuştur.

Veri madenciliğinde bulunan veri ön işleme, veri tanımlama, veri madenciliği tekniklerinin uygulanması ve sonuçların sunulması adımlarından oluşan web madenciliği sistem mimarisi oluşturulmuştur. Erişim ve hata loglarında yer alan çeşitli alanlar ve bu alanlar arasındaki ilişkiler ile ilgili çıkarımlar elde edilmiştir.

Bir sonraki çalışmada web logları, hata logları, kullanıcı logları birbirleriyle ilişkilendirilerek veri madenciliği yöntemleri ile istatistiksel analizlerin yanı sıra saldırı örüntülerinin tespit edilmesi amaçlanmaktadır.

REFERANSLAR

[1] M.Karabatak, M.C.İnce,"Apriori Algoritması ile Öğrenci Başarı Analizi", http://www.emo.org.tr/ekler/24f4c5eef7ec01c_ek.pdf.

[2] S.Özdemir,"Veri Madenciliği Ders Notları", Gazi Üniversitesi, 2012.

[3] C.Çoşkun ve A.Baykal, "Veri Madenciliğinde Sınıflandırma Algoritmalarının Bir Örnek Üzerinde Karşılaştırılması", Akademik Bilişim'11 - XIII. Akademik Bilişim Konferansı Bildirileri, 2011.

[4] Ö.Akgöbek ve F.Çakır, "Veri Madenciliğinde Bir Uzman Sistem Tasarimi", Akademik Bilişim'09 - XI. Akademik Bilişim Konferansı Bildirileri, 2009.

[5] T. Özseven ve M. Düğenci, "LOG PreProcessing: Web Kullanım Madenciliği Ön İşlem Aşaması Uygulma Yazılımı", Akademik Bilişim'11 - XIII. Akademik Bilişim Konferansı Bildirileri, 2011.

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

[7] İ.Haberal, "Veri Madenciliği Algoritmaları Kullanılarak Web Günlük Erişimlerinin Analizi", Başkent Ünv. Fen Bilimleri Enst. Bilgisayar Müh. Tez Çalışması, 2007.

[8] J.Duan ve ark., "Research on web log mining analysis", Instrumentation & Measurement, Sensor Network and Automation (IMSNA) 2012 International Symposium on, 515 - 519, 2012.

[9] B. Özakar, ve H.Püskülcü, "Web içerik ve web kullanım madenciliği tekniklerinin entegrasyonu ile oluşmuş bir veri tabanından nasıl yararlanılabilir?", 2002.

[10] D.S. Sisodia, S. Verma, "Web usage pattern analysis through web logs: A review ", Computer Science and Software Engineering (JCSSE), 2012 International Joint Conference on, 49-53, 2002.

[11] R Daş ve ark., "Bir Web Sitesine Ait Kullanıcı Erişim Kayıtlarının Web Kullanım Madenciliği Yöntemiyle Analizi: Fırat Üniversitesi Örneği", e-Journal of New World Sciences Academy Natural and Applied Sciences, 310-320, Mart 2008.

[12] T.Hussain ve ark., "Web usage mining: Asurvey on preprocessing of web log file", Information and Emerging Technologies (ICIET), 2010 International Conference on , 1-6, 2010.

[13] M.P. Yadav ve ark., "An efficient web mining algorithm for Web Log analysis: E-Web Miner", Recent Advances in Information Technology (RAIT), 607 – 613, 2012.

[14] G.Vigna ve ark., "A stateful intrusion detection system for World-Wide Web servers", Computer Security Applications Conference, 34 - 43, 2003.

[15] Cyber-Security, http://www.cyber-security.org.tr/Madde/180/Web%20Uygulamalar%C4%B1na%20Y%C3%B6nelik%20Sald%C4%B1r%C4%B1%20Analizleri%20-%202, 21.03.2009.

[16] M.Auxilia ve ark., "Anomaly detection using negative security model in web application", Computer Information Systems and Industrial Management Applications (CISIM) , 481-486, 2010.

[17] Z.Sun ve ark., "Application of web log mining in local area network security", Electronic and Mechanical Engineering and Information Technology (EMEIT), 3897-3900, 2011.

[18] Bilgi Güvenliği Akademisi, "Web Sunucu Loglarından Saldırı Analizi", http://www.bga.com.tr/makaleler-2/web-sunucu-loglarindan-saldiri-analizi/.

[19] M.Dener ve Ark., " Açık Kaynak Kodlu Veri Madenciliği Programları:WEKA'da Örnek Uygulama", Akademik Bilişim'09 - XI. Akademik Bilişim Konferansı Bildirileri, 2009.

[20] T.T.Aye, "Web log cleaning for mining of web usage patterns ", Computer Research and Development (ICCRD), 2011 3rd International Conference on , 490-494, 2011.

[21] S.Ma ve ark., "The Study on the Preprocessing in Web Log Mining", Knowledge Acquisition and Modeling (KAM), 2011 Fourth International Symposium on , 315-317, 2011.

[22] K.Kayaalp, "Asenkron Motorlarda Veri Madenciliği ile Hata Tespiti", http://tez.sdu.edu.tr/Tezler/TF01106.pdf, Süleyman Demirel Ünv.,Fen Bilimleri Ens., 2007.

Işıl Çinar, Gazi Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği bölümünde yüksek lisans yapmaktadır. Bilişim güvenliği, veri madenciliği, insan bilgisayar etkileşimi alanlarında çalışmalar yapmaktadır.

Muhammet Serkan Çinar, 2011 yılında Dokuz Eylül Üniversitesi Bilgisayar Mühendisliği bölümünden yüksek lisans derecesi almıştır. Şu anda Hacettepe Üniversitesi Bilgisayar Mühendisliği bölümünde doktora öğrencisidir. IPTV tabanlı uzaktan eğitim, yazılım mühendisliği, bilişim güvenliği ve semantik web alanlarında çalışmalar yapmaktadır.

Hasan Şakir Bilge, 1992 yılında Yıldız Teknik Üniversitesi Bilgisayar Mühendisliği bölümünden mezun oldu. 1997 yılında Kırıkkale Üniversitesi Elektrik-Elektronik Mühendisliği Anabilim Dalı'nda yüksek lisansını, 2003 yılında Başkent Üniversitesi Elektrik-Elektronik Mühendisliği Anabilim Dalı'nda doktorasını tamamladı. Gazi Üniversitesi Bilgisayar Mühendisliği bölümünde 2003 yılında yardımcı doçent unvanını ve 2012 yılında doçent unvanını aldı.