WEB KULLANIM MADENCİLİĞİ UYGULAMASI

Aydın CARUS, Altan MESUT

Trakya Üniversitesi, Mühendislik-Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü
Edirne

e-posta: aydinc@trakya.edu.tr

ÖZET

Web kullanım madenciliğine yönelik .NET ortamında geliştirdiğimiz ve "Web Sunucusu Analizcisi - WSA" olarak isimlendirdiğimiz uygulama, web sitesi tasarımcılarına tasarladıkları sitenin kullanımının ve etkinliğinin değerlendirilmesinde yardımcı olacak bilgiler üretir. WSA veri olarak web sunucusu erişim günlük dosyalarını kullanır, bu verilerden elde ettiği bilgileri değerlendirerek sonuçları grafiksel raporlar olarak görüntüler ve yazıcıdan çıktılarının alınmasına imkân verir. WSA, W3C, IIS 4, IIS 5, IIS 6, NCSA gibi farklı günlük dosyası biçimlerini otomatik olarak tanıyabilmekte ve bunların hepsini bir rapor altında değerlendirebilmektedir. Ayrıca günlük dosyalarından elde edilen kullanışlı bilgiler, veri tabanına bir rapor ismi ile saklanır. Bu yönüyle önceden değerlendirilmiş verilerle yeni değerlendirilen veriler arasında kıyaslama yapma kolaylığı getirmektedir. İstenildiğinde bir raporun tamamı yerine belirli tarihler arasındaki kısmı alt rapor olarak kullanılabilmektedir. WSA, günlük dosyalarının incelenmesi sırasında günlükteki kayıt hatalarını da bulabilmektedir.

Anahtar Kelimeler: *Veri Madenciliği, Web Madenciliği, Web Kullanım Madenciliği, Web sunucu erişim günlükleri*

ABSTRACT

The web usage mining application that we developed in .NET platform and named as "Web Server Analyzer - WSA", produce information that helps web site designers to evaluate the usage and effectiveness of the site which they designed. WSA uses web server access logs as input data. By evaluating the information which it gets from that data, it shows the results as graphical reports, and allows getting their printer outputs. WSA can recognize different log file formats like W3C, IIS 4, IIS 5, IIS 6, NCSA, and evaluate all of them in one report. The useful information which retrieves from log files, are also stored in a database with a report name. It brings the easiness of doing comparison between previously evaluated data and currently evaluated data. The part of a report in a particular time period can get as a sub-report, instead of using the full report. WSA can also find record errors while doing log file examination.

Key Words: *Data Mining, Web Mining, Web Usage Mining, Web server access logs*

1. GİRİŞ

Veri madenciliği büyük miktardaki veriden anlamlı bilginin çıkarılması ile ilgili bir teknik olup pazarlama, bankacılık, sigortacılık ve tıp sektörü başta olmak üzere birçok sektörde etkin şekilde kullanılmaktadır.

Web madenciliği, WWW üzerinden kullanışlı bilgiyi keşfetme ve analiz etme işlemi, şeklinde geniş olarak tanımlanır. Bu geniş tanım bir yandan, milyonlarca siteden ve çevrimiçi (online) veritabanlarından veri ve kaynakların otomatik olarak aranması ve elde edilmesi işlemi olan Web İçerik Madenciliği'ni tarif ederken, diğer yandan, bir yada daha çok Web sunucusu veya çevrimiçi servisten kullanıcı erişim desenlerinin keşfi ve analizi işlemi olan Web Kullanım Madenciliği'ni tarif eder [1]. Daha sonradan bu iki kategoriye, Web sitelerinin bağlantı (link) yapılarını da kapsayan yapısal özetini üreten Web Yapı Madenciliği de eklenmiştir.

Verilerin dijital ortamda saklanmaya başlanması ile birlikte, yeryüzündeki bilgi miktarının sürekli arttığı günümüzde veri tabanlarında saklanan veri miktarı da benzer oranda artmaktadır. Yüksek kapasiteli işlem yapabilme gücünün ucuzlamasının bir sonucu olarak, veri saklama hem daha kolay olmuş, hem de verinin kendisi de ucuzlamıştır.

Günümüzde oldukça yaygınlaşan elektronik ticaret ve çevrimiçi alışveriş mekanizmalarının da artmasıyla birlikte, bu alanda birbirlerine rakip olan firmaların çalışmaları, veri madenciliğinin önemini ön plana çıkarmaktadır [2].

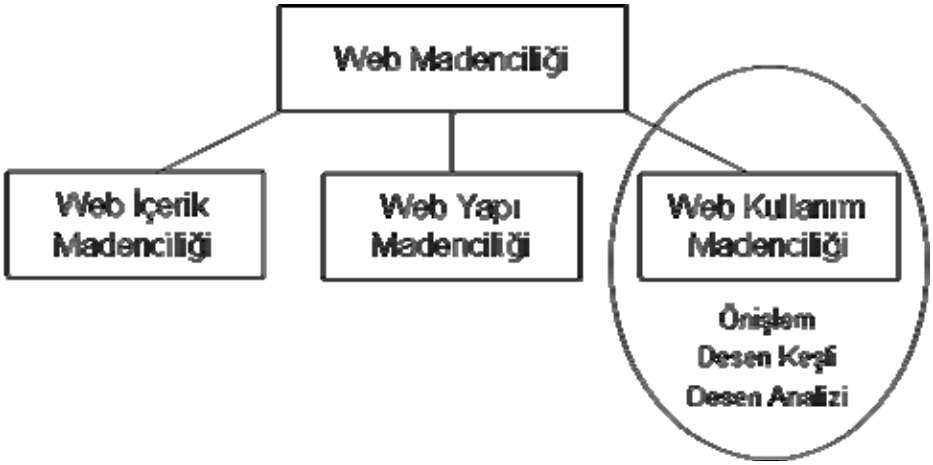
Veri Madenciliğinin gelecek yıllar için üstlenmiş olduğu amaç hakkında fikir sahibi olmak için dünyanın önde gelen araştırma ve danışmanlık firmalarından açıklanan rakamları dikkate aldığımızda, veri madenciliğinin gelecekte oldukça popüler bir konu olacağını görebiliyoruz. Örneğin, Gartner Group Araştırma Şirketi, gelecek on yıl içinde, hedef pazarlarda Veri Madenciliği kullanımının yüzde 80'lere ulaşacağı tahmininde bulunuyor. Öte yandan META Group ise Veri Madenciliği pazarının bu yıl 800 milyon dolara yükseleceği yönünde tahminlerde bulunuyor [3].

2. WEB KULLANIM MADENCİLİĞİ

Web kullanım madenciliği, bir veya birçok web sunucusundan kullanıcı erişim desenlerinin otomatik keşfinin ve analizin yapıldığı bir tip veri madenciliği etkinliğidir. Birçok kuruluş pazar analizleri için geliştirdikleri stratejileri ziyaretçi bilgilerine dayanarak yerine getirir. Kuruluşlar günlük operasyonlarla her gün yüzlerce MB veri toplamaktadır. Bu bilgilerin çoğu web sunucuların otomatik olarak tuttuğu günlük dosyalarından elde edilir. Günlük dosyaları, istemciden sunucuya gönderilen her bir isteğin bir kayıt olarak eklenmesi ile meydana gelir.

Günlük dosyalarının analizi, müşterilerin ilgi alanları, ürünler üzerinden pazar stratejileri oluşturma, promosyon kampanyalarının etkisi gibi hususlarda, kurumlara karar süreçlerinde yardımcı olur. Sunucu erişim kayıtlarının ve kullanıcı kaydı verilerinin analizi, aynı zamanda kurumun daha etkili bir sunumunun yapılabilmesi için Web sitesini nasıl daha iyi hale getirebileceği hakkında değerli bilgiler sağlar. İntranet teknolojilerini kullanan kurumlarda, bu tür analizler çalışma grubu iletişimi ve kurumsal altyapının daha iyi işletilmesine ışık tutabilir. Son olarak, WWW üzerinden reklam yapan kurumlar için kullanıcı erişim desenlerini analiz etmek, reklamların belirli bir kullanıcı grubuna yönlendirilmesine yardımcı olur.

Web madenciliği alanları ve web kullanım madenciliği aşamaları şeması Şekil 1'de verilmiştir.



Şekil 1. Web madenciliği alanları ve web kullanım madenciliği aşamaları

Web kullanım madenciliği; Ön işlem (pre-processing), desen keşfi (pattern discovery) ve desen analizi (pattern analysis) aşamalarından oluşur. Web kullanım madenciliği esnasında ayrıştırılacak veri aşağıdaki tiplerde olabilir [4].

- İçerik verisi (Content data)
- Yapı verisi (Structure data)
- Kullanım verisi (Usage data)
- Kullanıcı görüntüsü (User profile)

3. GELİŞTİRİLEN UYGULAMA

Bir web sitesinin değerlendirilebilmesi için siteyi ziyaret eden kullanıcıların site tarafından günlük dosyalarının tutulması gerekmektedir, tutulan bu günlük dosyaları ise WUM, NCSA,W3C, IIS 4, IIS 5, IIS 6 gibi farklı biçimlerinde olabilmektedir. Bu günlük dosyalarında saklanan kayıtların anlamlı ve yararlı bilgi haline gelebilmesi için işlenmesi gerekmektedir.

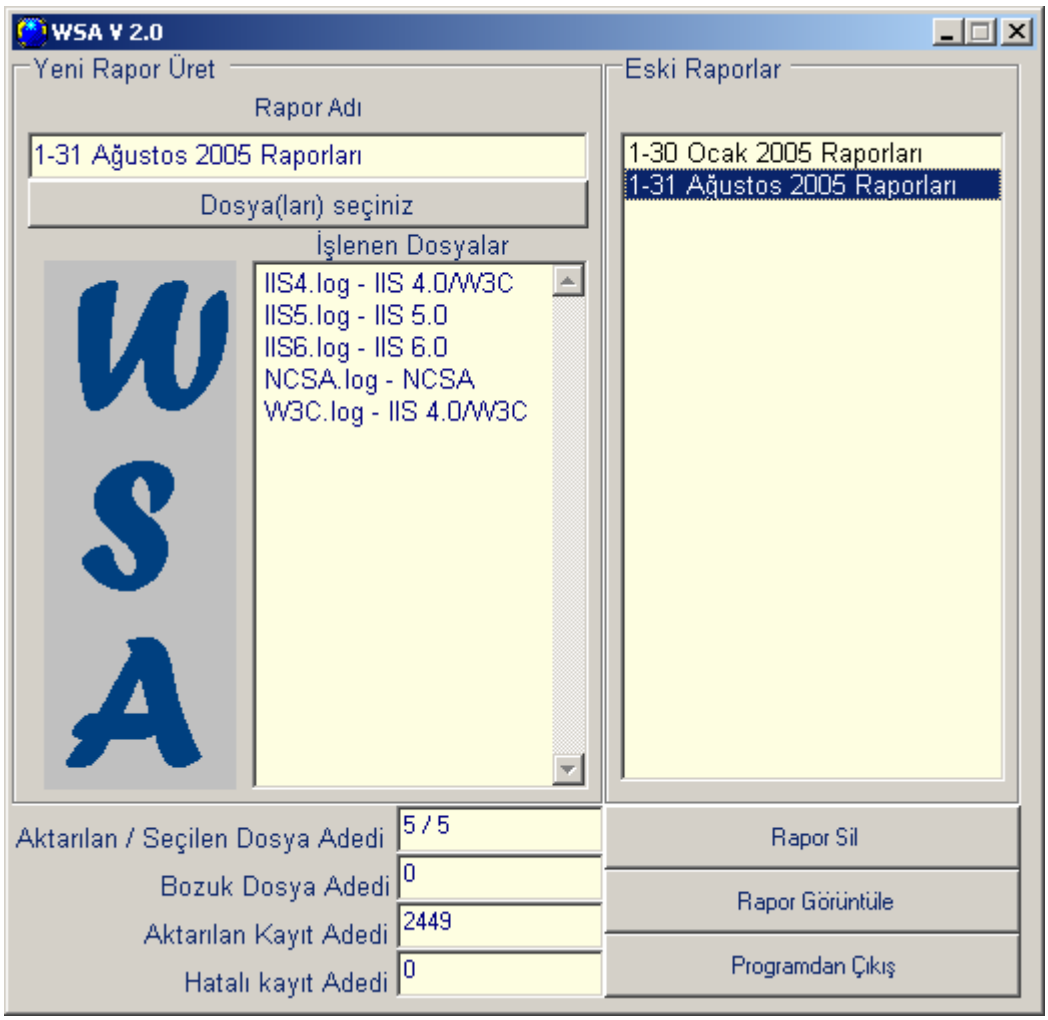
Web günlük dosyalarını işlemek, işlenen kayıtları değerlendirmek ve grafiksel raporlar üretmek amacıyla "Web Sunucusu Analizcisi - WSA" isimli bir yazılım, .NET ortamında tarafımızdan geliştirilmiştir. WSA yazılımı kolay kullanımlı bir arayüze sahip olarak hazırlanmıştır. WSA'nın ana sayfası Şekil 2'de görülmektedir.

WSA yazılımı kullanılarak bir sitenin değerlendirilme işleminin yapılabilmesi için öncelikle günlük dosyalarına ihtiyaç duyulmaktadır, bu günlük dosyalarının değerlendirilmesi için bu dosyalardan elde edilecek kayıtlar işlenerek öncelikle veritabanına aktarılmaktadır. Bu aktarılma işlemi sırasında günlük dosyasının biçimini belirlemeye yönelik algoritmalar çalıştırılmakta ve NCSA, W3C, IIS4, IIS5, IIS6 biçimindeki günlük dosyaları yazılım tarafından otomatik olarak tanınmaktadır. Günlük dosyalarının işlenip aktarılması sırasında hatalı günlük dosyaları ve hatalı kayıtlar tespit edilmekte, hatalı dosya ve kayıt sayıları işlem sonunda ekrana gösterilmektedir.

WSA yazılımının farklı günlük dosyası biçimlerini tanıyabilmesi için kullanılan algoritmada;

- IIS5 ve IIS6 biçimlerini tanıyabilmek için günlük dosyası başlıklarından,
- IIS4 ve ona denk olan W3C biçimlerini tanıyabilmek için günlük dosyası kayıtlarındaki ":" ve boşluk karakterlerinin konumlarından,
- NCSA biçimini tanıyabilmek için günlük dosyası kayıtlarındaki "|" karakterinin konumlarından,

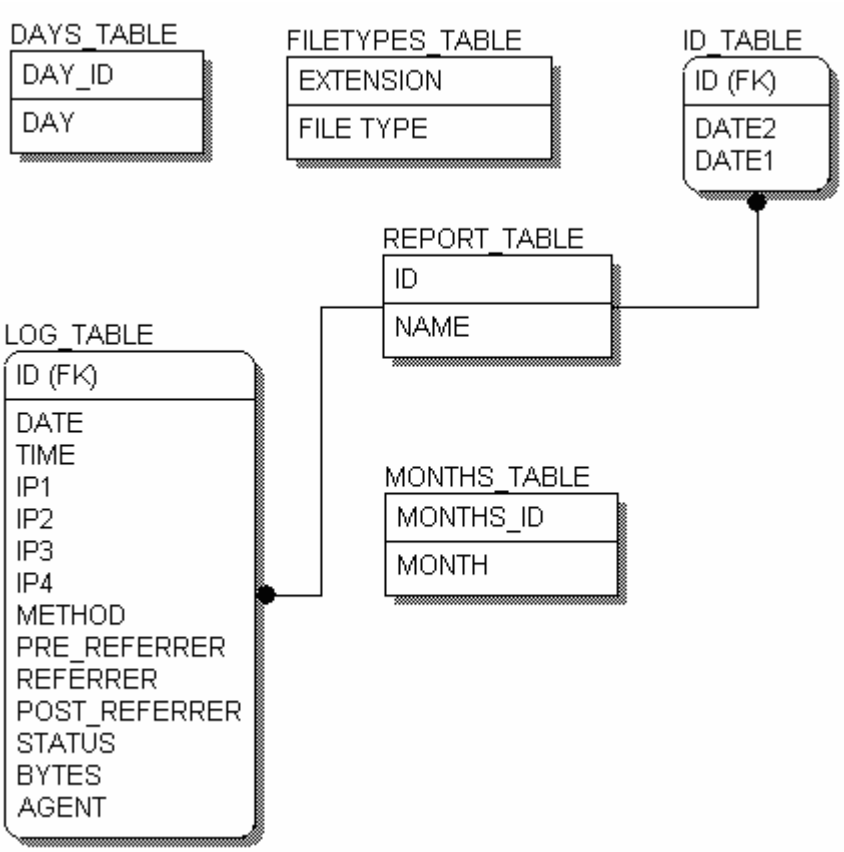
faydalanılmaktadır.



Şekil 2. WSA Yazılımı Arayüzü

Biçimi tespit edilen günlük dosyalarından desen keşfi ve desen analizi aşamalarında kullanılmayacak olan fazla verileri temizleyebilmek için, günlük dosyasının biçimine özgü alan ayıraçlarından faydalanılmıştır. Fazlalıkları atılmış olan veriler Şekil 3'te varlık ilişki diyagramı verilen veri tabanına aktarılmaktadır.

Veri tabanında birçok alan daha etkin sınıflama ve kümeleme yapabilmek için parçalara ayrılarak veri tabanında saklanmaktadır. Örneğin 15 karakter uzunluğundaki IP adresleri 4 adet 1 byte uzunluğunda nümerik alana bölünerek veri tabanında saklanmaktadır. IP adresinin bu şekilde bölünmüş olarak tutulması veri tabanında yer kazancı sağlamakta ve sorguları daha etkin kılmaktadır.



Şekil 3. Veri tabanı varlık ilişki diyagramı

İşlenip aktarılan verilere bir rapor ismi verilmekte ve bu veriler bu rapor ile ilişkili olarak veri tabanında saklanmaktadır. Raporu üretilmiş verilerin veri tabanında saklanıyor oluşu, farklı tarih aralıklarında bir raporun alt raporlarının üretilebilmesine ve belli dönemlere ait raporların birbirleriyle kıyaslanmasına imkân vermektedir.

WSA yazılımının işlem aşamalarının daha iyi anlaşılabilmesi için, siteye bağlanan bilgisayarların işletim sistemi bilgilerini raporlama görevi aşağıda adım adım verilmiştir:

Örnek olarak seçtiğimiz IIS5 biçimindeki bir günlük dosyasında, bir kayıt satırı aşağıdaki gibidir:

```
2002-01-06 13:45:24 65.116.145.138 - 193.255.141.93 80 GET
/dersler/grafik/Notes/default.html - 200
Mozilla/4.0+(compatible;+MSIE+5.0;+Windows+98;+DigExt)
```

Bu satırda 10 adet boşluk karakteri ile ayrılmış 11 adet bilgi bulunmaktadır. Bu bilgiler sırasıyla; tarih, saat, istemci IP numarası, kullanıcı adı, sunucu IP numarası, sunucu port numarası, yöntem, istek yapılan dosya, sorgu, durum kodu ve agent bilgileridir. Bu kayıt içinde kullanıcı adı ve sorgu alanları

‘-’ ile gösterilmiş yani bu bilgiler saklanmamıştır. İhtiyacımız olan işletim sistemi bilgisinin yer aldığı alan agent alanıdır.

Kullanılmayacak olan fazla veriler temizlenerek veri tabanına aktarıldığında yukarıdaki günlük dosyası satırından aşağıdaki veri tabanı kaydı elde edilmektedir:

ID	Date	Time	IP1	IP2	IP3	IP4	Method	Pre_referrer	Referrer	Post_referrer	Status
5	06.01.2002	13:45:24	65	116	145	138	GET	/dersler/	/grafik/Notes/default.	html	200

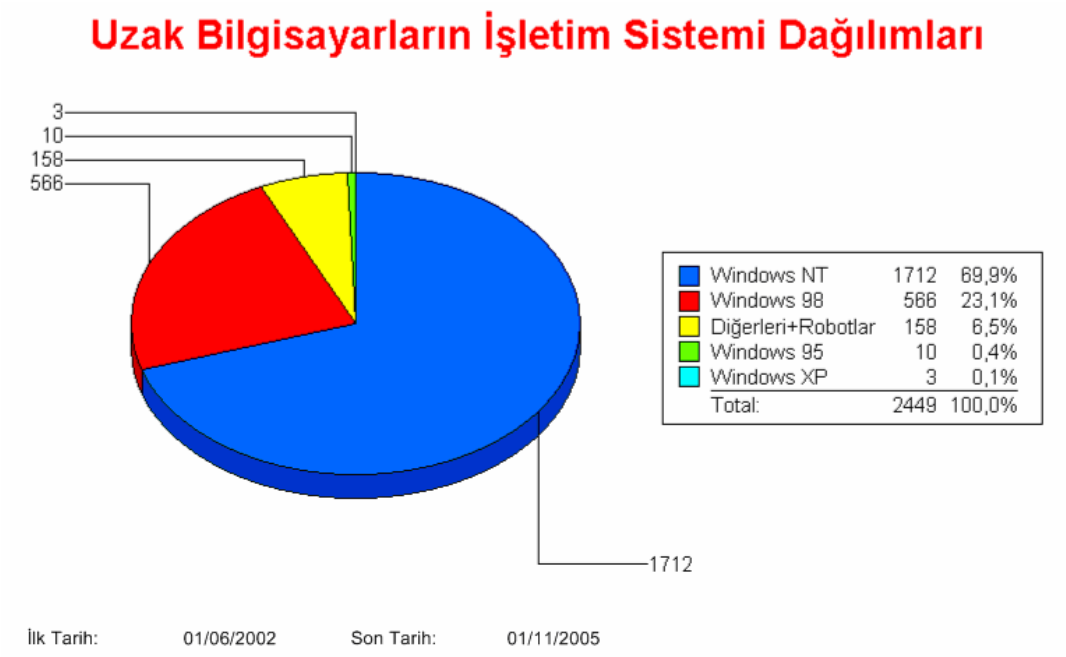
Bytes	Agent
0	Mozilla/4.0+(compatible;+MSIE+5.0;+Windows+98;+DigExt)

Günlük dosyası kaydında yer alan sunucu IP numarası (193.255.141.93) ve sunucu port numarası (80) bilgileri, günlükteki diğer kayıtlarda da aynı değere sahip olduğu için, bu alanların saklanmasına ve bu alanlarla ilgili rapor alınmasına gerek yoktur. Bu nedenle, veri tabanı kaydına bu alanlar dâhil edilmemiştir.

Raporlamada Crystal Reports programından faydalanılmaktadır. Agent alanı içinde yer alan işletim sistemi bilgisinin keşfedilmesi amacıyla aşağıdaki program kodu işletilmektedir;

```
if {ID_table.tarih1}<={LOG_Table.Dates} and  {ID_table.tarih2}>={LOG_Table.Dates} and
{ID_table.ID}={LOG_Table.ID} then
    if LowerCase ({LOG_Table.Agent}) like "*winnt*" or LowerCase ({LOG_Table.Agent})
    like "*windows nt*" or ({LOG_Table.Agent}) like "*windows+nt*" then "Windows NT"
    else
    if LowerCase ({LOG_Table.Agent}) like "*win98*" or LowerCase ({LOG_Table.Agent})
    like "*windows 98*" or ({LOG_Table.Agent}) like "*windows+98*" then "Windows 98"
    else
    if LowerCase ({LOG_Table.Agent}) like "*win95*" or LowerCase ({LOG_Table.Agent})
    like "*windows 95*" or ({LOG_Table.Agent}) like "*windows+95*" then "Windows 95"
    else
    if LowerCase ({LOG_Table.Agent}) like "*winxp*" or LowerCase ({LOG_Table.Agent})
    like "*windows xp*" or ({LOG_Table.Agent}) like "*windows+xp*" then "Windows XP"
    else "Diğerleri+Robotlar"
```

Yapılan bu işlemlerden sonra Şekil 4’te görülen işletim sistemi dağılımları raporu oluşturulmaktadır. Diğer raporların üretilmesinde de benzer yöntemler kullanılmıştır.



Şekil 4. İşletim sistemi dağılımları raporu

4. SONUÇLAR

Bu çalışmadaki amacımız web kullanım madenciliği özellikleri olan bir web günlük dosyası analizcisi geliştirmekti. Geliştirdiğimiz WSA yazılımı, web sunucu günlük dosyalarını ham veri kaynağı olarak kullanarak, verilecek tarih aralıklarındaki verileri işleyip farklı özelliklerde günlük dosyalarını değerlendirebilecek ve site hakkında yeni kararlar alabilmeye yardımcı olacak birçok grafiksel çıktıyı üretebilecek özelliklere sahiptir.

Geliştirilen WSA yazılımı aşağıdaki özelliklerinin bir ya da birkaçı ile diğer istatistiksel analiz yazılımlarından farklılık göstermektedir:

- Günlük dosyası biçimini otomatik tanıyabilmesine imkân veren algoritmalara sahiptir,
- Değerlendirilen günlük dosyaları kayıtlarının veri tabanında saklanması sayesinde hızlı rapor üretme yeteneğine sahiptir,
- Farklı günlük dosyası biçimlerini ortak bir rapor olarak değerlendirebilme yeteneğine sahiptir,
- Rapor görüntüleme ekranlarındaki tarih aralığı verebilme olanağı sayesinde görüntülenen raporun farklı tarih aralığı için çok kısa sürede tekrar sorgulanabilmesi olanağına sahiptir.

5. KAYNAKLAR

[1] R. Cooley, B. Mobasher, J. Srivastava. Web Mining: Information and Pattern Discovery on the World Wide Web. Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), Kasım 1997.

[2] A. Vahaplar, M. M. İnceoğlu**.** "Veri Madenciliği ve Elektronik Ticaret". Türkiye'de Internet Konferansları VII, Kasım 2001.

[3] M. Karakaş. "Veri Madenciliği Üzerine", http://www.bilgiyonetimi.org

[4] H. Takci, İ. Soğukpınar, "Kütüphane Kullanıcılarının Erişim Desenlerinin Keşfi", Akademik Bilişim 2002, Şubat 2002, Selçuk Üniversitesi, Konya