Kütüphane Kullanıcılarının Erişim Örüntülerinin Keşfi

Discovery of Access Patterns of Library Users

Hidayet Takçı[*] ve İbrahim Soğukpınar[**]

Öz

Veri madenciliği çok büyük hacimli veriden anlamlı, ilginç, ve önemli bilginin otomatik veya yarı-otomatik yöntemlerle çıkarılması ile bir disiplin olup pazarlama, bankacılık, sigortacılık ve tıp sektörü başta olmak üzere bir çok sektörde etkin bir şekilde kullanılmaktadır. Veri madenciliği uygulamalarından biri olan web kullanım madenciliği sayesinde web üzerindeki faydalı bilginin keşfi ve analizi yapılmaktadır. Kullanıcı erişim örüntülerinin bulunması web içerik madenciliği konusuna girmekte olup veri madenciliği tekniklerinden özellikle bağlantı analizi ile yerine getirilmektedir. Bu çalışmada kütüphane sitesi web günlüklerine dayalı olarak kütüphane kullanıcılarının erişim örüntüleri bulunmaya çalışılmıştır. Bu çalışma yapılırken istatistiksel yöntemler kullanılmıştır.

Anahtar sözcükler: *Web madenciliği, Web kullanım madenciliği, Web günlük analizi, Örüntü keşfi ve analizi.*

Abstract

Data mining, which is a new technique interested in discovering knowledge huge amount of data, has been effectively used in many sectors especially in banking, assurancy, medicine and marketing. Web mining, which is an application of data mining, performs data mining functions on web data. With the help of web mining, which is seperated into two branches as web content mining and web usage mining, discovery and analysis of useful knowledge on web is made. Existance of user access patterns which is included in the subject of web content mining, is being done especially with the link analysis that is a technique of data mining. In this study, access patterns of library users is tried to be found based on web logs of library web site.

Keywords: *Web mining, Web usage mining, Web log analysis, Pattern Discovery and analysis.*

[*] Öğretim görevlisi, Gebze İleri teknoloji Enstiitüsü Bilgisayar Mühendisliği Bölümü, (htakci@bilmuh.gyte.edu.tr).

[**] Yard.Doç.Dr. Gebze İleri Teknoloji Enstiitüsü, (ispinar@bilmuh.gyte.edu.tr).

Giriş

Veri madenciliği ve web son zamanların geçerli iki araştırma sahasıdır. Bu iki sahanın doğal kombinasyonu web madenciliği olarak adlandırılmaktadır. Web madenciliği kabaca webden daha önce bilinmeyen, ilginç, önemli ve faydalı olan örüntülerin keşfi olarak tanımlanabilir (Etzioni, 1996).

Web madenciliği; web içerik ve web kullanım madenciliği şeklinde iki alt alana ayrılmaktadır. Otomatik tarama, bilgi alma ve kullanılabilir kaynakların milyonlarca web sitesi veya çevrim içi veri tabanlarından seçilmesi web içerik madenciliği, bir veya birçok web sunucu veya çevrim içi servisten kullanıcı erişim örüntülerinin analiz ve keşfi web kullanım madenciliği konusuna girmektedir.

Web içerik madenciliği, akıllı yazılım aracıları (web robotları, web örümcekleri vb.), makine öğrenimi ve/veya yapay zeka ile ilgilidir. Web içerik madenciliği, dokümanlardan bilgi çıkarmak, web kullanım madenciliği ise kullanıcı erişimlerinden kullanıcılar hakkında bilgi edinmek amacıyla kullanılmaktadır. Kullanıcı erişimlerine dayalı olarak kullanıcı davranışları bulunabilmektedir.

Kütüphane web sitesindeki sayfalara giriş isteklerinin bilinmesi, sitenin yeniden düzenlenmesi ve daha aktif hale getirilmesine yardımcı olacak bilgileri sağlayacaktır. Hangi sayfalara daha sık girildiği, hangi sayfaların birlikte ziyaret edildiği gibi bilgiler sitenin yeniden düzenlenmesinde faydalı olacak bilgileri sunmaktadır. Bu uygulamaların amacı, istatistiksel yöntemlerle kullanıcıların davranışlarını öğrenmek böylece site içeriği ve tasarımını bu bilgiler ışığında gözden geçirerek yenilemeler yapmaktır.

Bu makalede giriş bilgilerinin ardından ikinci bölümde web kullanım madenciliğinin ne olduğu, aşamaları, teknikleri anlatılmış, üçüncü bölümde web verileri üzerinde madenciliğin nasıl yapıldığı, dördüncü bölümde ise kütüphane web sitesi üzerinde yapılan veri madenciliği çalışmaları yer almıştır.

I- Web Kullanım Madenciliği

Web kullanım madenciliği, bir veya birçok web sunucudan kullanıcı erişim örüntülerinin otomatik keşfinin ve analizin yapıldığı bir tip veri madenciliği etkinliğidir. Birçok kuruluş pazar analizleri için geliştirdikleri stratejileri ziyaretçi bilgilerine

dayanarak yerine getirir. Kuruluşlar günlük işlemlerle her gün yüzlerce megabayt (MB) veri toplamaktadırlar. 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 (Joshi, Anupam, Yesha ve Krishnapuram, 1999).

Web verilerinin analizi sonucunda; ziyaretçilerin sitede kalma süreleri, kuruluş için hizmet stratejileri, etkin kampanyalar ve diğerleri bulunabilir.

Web kullanım madenciliği; ilk işlem, örüntü keşfi ve örüntü analizi aşamalarından oluşur. Web kullanım madenciliği esnasında harmanlanacak veriler aşağıdaki tiplerde olabilir:

- **İçerik verisi**: Web dokümanlarında, genellikle metin şeklinde yer alan verilerdir. Herhangi bir web sayfası üzerinde yer alan veriler bu tip için bir örnektir.
- **Yapı verisi**: Web sitesinin bağlantı yapısı hakkındaki verilerdir. Web sitesinde yer alan sayfaların hangi alt dizinler içerisinde bulunduğunu gösteren verilerden oluşur.
- **Kullanım verisi**: Web sitesini ziyaret eden kullanıcıların oluşturdukları veri tipidir. Kullanım verisi genellikle hangi kullanıcı, ne zaman, hangi sayfaları ziyaret etti, ne kadar süre sitede kaldı gibi soruların cevaplarını içerir.
- **Kullanıcı profili**: Web sitesini ziyaret eden kullanıcı hakkındaki; kullanıcı kimlik verileri gibi bilgilerden oluşur.

Web Kullanım Madenciliği Aşamaları

1. Ön İşlem

Ön işlem web kullanım madenciliğinin ilk aşamasıdır. Bu aşamada web sunucu günlüklerindeki kullanım verisi ilişkisiz sahalardan arındırılır ve dönüşüme tâbi tutulur. Dönüşüm çeşitli şekillerde olabilmektedir ve burada dönüşüm için soyutlama kullanılmaktadır. Soyutlama bir çeşit istatistiksel özet çıkarmadır ve kullanıcı, sayfa görünümü, tıklama akışı (click stream), kullanıcı oturumu, sunucu oturumu gibi çeşitleri olabilmektedir.

Soyutlama konusunda bir örnek vermek gerekirse: Soyutlama tipimiz kullanıcı olsun, yani kullanıcı tabanlı olarak soyutlama yapmış olalım, o zaman aşağıdaki gibi değerler elde ederiz (Zaiane, Xin ve Han, 1998).

Kullanıcı No: 12345
İstek Listesi (A,B,C,D,E,A,B,D,A,C,E,B,D,F,G,H,...)

Burada A, B, C vs, kullanıcı tarafından istenen sayfaları temsil etmektedir, kullanıcı tabanlı soyutlama yapılmakta ve web günlüğü üzerindeki kayıtlar kullanıcı tabanlı olarak gruplara ayrılmaktadır.

2. Örüntü Keşfi

Örüntü keşfi, ön işlemden geçirilen verilere veri madenciliği tekniklerinin uygulandığı aşamadır. En sık kullanılan bazı veri madenciliği yöntemleri; istatistiksel yöntemler, eşleştirme kuralları, kümeleme, sınıflandırma ve sıralı örüntülerdir (Cooley, Mobasher ve Srivastava, 1997). Bu tekniklere kısaca göz atmak gerekirse:

İstatistik

İstatistiksel teknikler bir web sitesinin ziyaretçileri hakkında bilgi açığa çıkarmaya yarayan en güçlü araçlardır. Analizciler oturum dosyasını analiz ederken farklı değişkenler üzerinde farklı açıklamalı istatistiksel analiz tiplerini yerine getirirler.

Periyodik web sistem raporlarında bulunan istatistiksel bilgi analiz edilerek sistem performansını artırıcı, sistem güvenliğini genişletici, düzeltme işlemlerini kolaylaştırıcı ve pazarlama kararlarını destekleyici raporlar çıkartılabilir (Cooley ve diğerleri, 1997).

Eşleştirme Kuralları

Web etki alanında sıklıkla birbirini referans gösteren sayfalar eşleştirme kuralı üretimi uygulanarak tek bir sunucu oturumu şeklinde düzenlenebilir. Eşleştirme teknikleri bir işletimsel veri tabanında bulunan değerler arasındaki sıralı olmayan ilinti keşfinde kullanılır (Garofalakis, Rastogi, Seshadri ve Shim, 1999).

Kümeleme

Kümeleme analizi, kullanıcıları veya sayfaları benzer özelliklerine göre birlikte gruplara ayırır. Kullanıcının veya sayfaların kümelenmesi geliştirme ve gelecek pazarlama stratejilerinin çalıştırılmasını kolaylaştırabilir (Cooley ve diğerleri, 1997). Kullanıcıların kümelenmesi benzer navigasyon örüntüsüne sahip kullanıcı gruplarını

keşfetmede yardımcı olacaktır. Elektronik ticaret uygulamalarında müşterilere özel hizmet sunabilmek için gerekli olan pazar bölümlemesi kümeleme sayesinde yerine getirilebilmektedir. İlgili içeriğe sahip sayfa gruplarının keşfinde kullanılabilen sayfaların kümelenmesi, arama motorları ve web servis sağlayıcıları için de yararlı olmaktadır.

Sınıflandırma

Sınıflandırma bir veriyi daha önceden tanımlanmış sınıflara dağıtma tekniğidir. Web etki alanında, webmaster veya pazarlamacı sınıflandırma tekniğini kullanarak müşterilerinin hangi sınıf veya kategoride bir profile sahip olduğunu belirleyebilir. Sınıflandırma işleminde, verilen bir sınıf veya kategorinin özelliklerini en iyi biçimde açıklamak için seçim ve açığa çıkarma uygulamalarına ihtiyaç duyulur. Sınıflandırma; karar ağaçları, bayezian sınıflayıcıları, en yakın komşu ve destek vektör makineleri gibi denetlenen tümevarımsal öğrenim algoritmaları kullanılarak yapılabilir (Cooley ve diğerleri, 1997).

Sıralı Örüntüler

Sıralı örüntüler; oturumlar arasında örüntü bulmaya çalışır. Sıralı örüntü bulma işleminde, belirli zaman aralıklarında oturumlar incelenir ve karşılaştırmalar yapılır. Sıralı örüntülerin bulunması gelecekteki eğilimi tahmin edecek web pazarlamacıları için oldukça anlamlıdır. Böylece ilanlar belirli kullanıcı gruplarına yönlendirilebilecektir. Sıralı örüntüler için, eğilim analizi, değişen nokta bulma veya benzerlik analizleri gibi bazı geçici analiz tipleri kullanır (Cooley ve diğerleri, 1997).

3. Örüntü Analizi

Örüntü keşfi aşamasında ortaya çıkarılan kural veya örüntülerin analiz edilmesi işlemidir. Bilgi sorgulama ve OLAP (OnLine Analytical Processing-çevrim içi analitik işlem) uygulamaları ile derinlemesine analizler yapılabilmektedir (Zaiane ve diğerleri, 1998).

Aşağıda bazı örüntü analiz seçenekleri bulunmaktadır:

Görselleştirme teknikleri

Örüntü keşif aşamasında elde edilen sonuçların (özetler gibi) anlaşılabilmesi için görselleştirme tekniklerinden faydalanılır. Görselleştirmede daireler, düğümler ve kenarlar kullanılır.

OLAP teknikleri

OLAP, iş ortamında veri tabanlarının stratejik analizi için çok güçlü bir uygulama alanı olarak ortaya çıkmıştır. Stratejik analizin bazı önemli özellikleri şunlardır: 1) Çok büyük boyutlu veri. 2) Geçici boyutlar için açık destek. 3) Çeşitli bilgi tipleri için destek sağlama. 4) Uzun-sıra analizi, ki orada toplam trendler bireysel veri elamanlarından daha önemlidir. OLAP doğrudan ilişkisel veri tabanları üzerinde çalışabilir. OLAP kullanılırken analiz için veri küplerinden faydalanılır.

Veri ve Bilgi Sorgulama

Veri ve bilgi sorgulama için iki yol bulunmaktadır. Birincisinde bildiriler şeklinde bir dil kullanılarak veri elde edilirken, ikincisinde SQL'e (Structured Query Language-Yapısal Sorgu Dili) benzeyen diller kullanılarak bilgi sorgulanabilir.

Kullanılabilirlik Analizi

Bulunan veya ortaya konulan çözümlerin başarılı sonuçlar verebilmesi için kullanılabilir olmaları gerekmektedir. Veri analizlerinde de takip edilen yöntemin başarısı kullanılabilirlik analizleri ile yerine getirilir. Bu konunun şu an hedefi, kullanılabilirlik için sistematik bir yaklaşım geliştirme çabasıdır. İlk adımda, yazılım kullanılabilirliği için geliştirme metotları bir araya toplanır. Veri, hesaplanmış modeller oluşturmada kullanılır. Son olarak, çeşitli veri sunum ve görselleştirme teknikleri ile verinin anlaşılması sağlanır. Bu şekilde web kullanıcılarının davranışları bir model ile anlaşılabilir halde gösterilebilir.

II- Web Madenci (WebMiner) Tasarımı

Bugün birçok resmi ve özel kurum veya kuruluş, günlük işlemlerini web üzerine taşımış ve bu işlemler dolayısıyla büyük hacimli veriler toplanmaya başlamıştır. Bu veriler genellikle web sunucular tarafından otomatik olarak toplanmakta olup sunucu veya erişim günlüklerinde tutulmaktadır. Bu günlüklerin madenciliğinin yapılması ve analiz edilmesi, değerli örüntüleri ortaya çıkarabilir. Günlük dosyası analizleri sayesinde hedef kitleye ve özel kullanıcı öbeklerine (kümeler) hizmet verilebilmektedir.

Web madencisi, web madenciliği yapacak uygulamaların genel adıdır. Webmadencisi tasarımının amacı, web sitesini ziyaret eden kullanıcıların bilgilerini

toplayarak onların davranışlarını kestirmektir. Böylece web sitesinin daha etkin ve verimli bir hale getirilmesi ve hizmet kalitesinin dolayısıyla kullanıcı sayısının arttırması için gerekli bilgileri elde etmek mümkün olacaktır. Webmadencisi tasarımı yaparken eşleştirme kuralları ve sınıflandırma gibi temel veri madenciliği tekniklerinden faydalanılmaktadır, sistemin veri kaynağını ise, web sunucu üzerindeki günlük dosyaları oluşturmaktadır.

III- Kütüphanede Web Kullanım Madenciliği

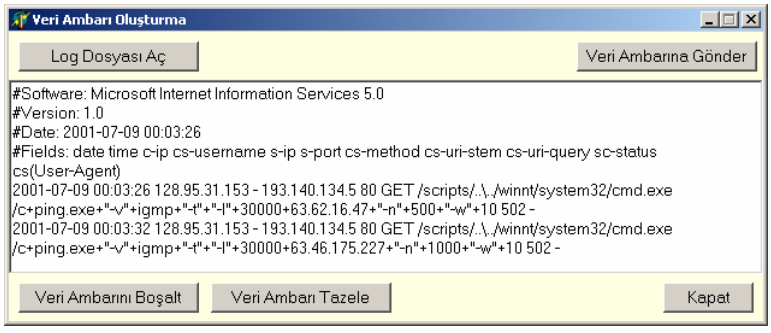
Web üzerinde hizmet veren bütün sitelerde web kullanım madenciliği yapılabilir ve her biri sistemin işleyişine yardımcı bilgiler üretir. Web kullanım madenciliği yapılabilecek yerlerden biri de kütüphane web siteleridir.

Sayısal yöntemlerle kütüphaneciliğin yapıldığı yerlerdeki ilk iş, arşivleme, tarama ve doküman işlemedir. Kütüphaneler artık bu hizmetlerin daha fazlasını yerine getirebilecek duruma gelmişler ve bir sonraki aşama olarak kütüphanelerdeki kullanılabilir bilginin madenciliği yapılmaya başlanmıştır (Grossman, 1998).

Geliştirilen Uygulama

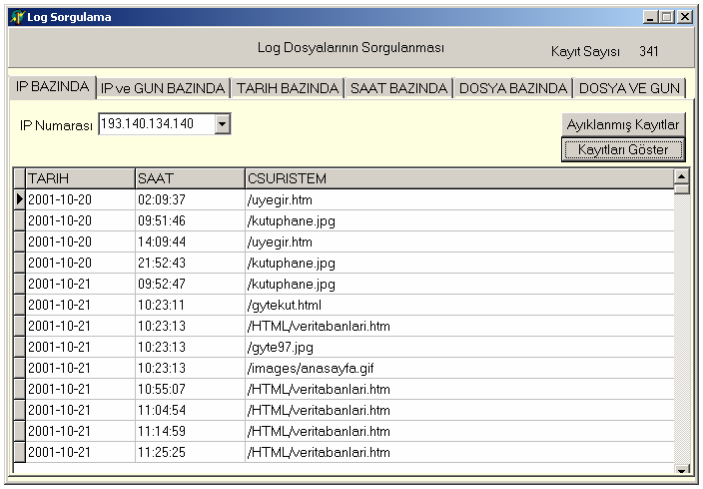
Web günlüklerindeki verileri kullanarak işimize yarayacak bilgileri elde etmek amacıyla bir WEBMINER tasarlanmıştır. Hazırlanan uygulama ile özellikle istatistik tabanlı analizler yapılmaya çalışılmıştır. Yazılım sayesinde elde edilen verilerin EXCEL gibi harici uygulamalar kullanarak etkinliğin arttırılması amaçlanmıştır.

Geliştirilen uygulama, verileri günlük dosyalarından alıp günlük veri tabanına aktarmakta ve verilerden çeşitli istatistiksel özetler çıkarmaktadır. Bunun yanı sıra eşleştirme kuralları ve sınıflandırma teknikleri için de veri sağlamaktadır. Yazılımın ekran görüntüleri aşağıdadır:



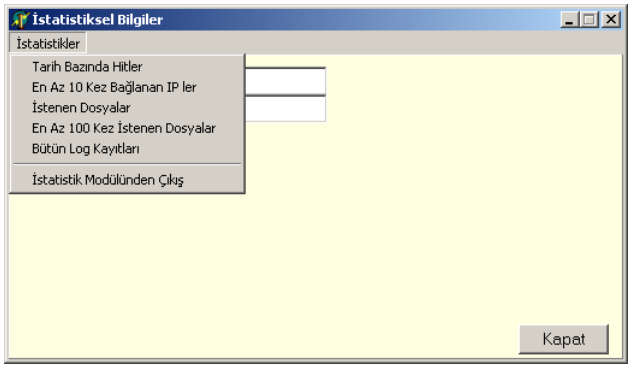
Şekil 1- Veri Ambarı Oluşturma Ekranı

Web günlük veri tabanına atılacak günlük dosyası **Log Dosyası Aç** düğmesiyle açılır ve istenirse bu dosya bir dönüşüm işleminden geçirilerek veri tabanına aktarılır. Aynı pencere üzerindeki **Veri Ambarı Boşalt** düğmesi veri ambarını silmeye, **Veri Ambarı Tazele** düğmesi ise veri ambarını güncellemeye yaramaktadır.



Şekil 2 - Log Dosyalarının Sorgulanma Ekranı

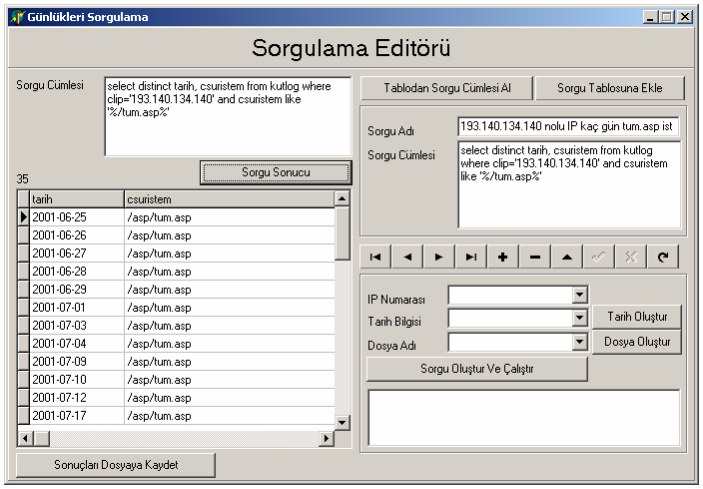
Bir önceki aşamada veri ambarına atılan kayıtlar üzerinde çeşitli sorgulamalar yapmak, sorgulamaları derinleştirerek soyutlamalar elde etmek mümkündür. (Bkz. Şekil 2) Soyutlamalar, kullanıcı, sayfa görünümü, kullanıcı oturumu, sunucu oturumları şeklinde olabilir**. Ayıklanmış Kayıtları Göster** düğmesi kullanıcıya analiz değeri olan kayıtları gösterir. Bu aşamada web sitesinden istenen resim dosyalarına ait kayıtlar silinir.



Şekil 3 - İstatistiksel Bilgiler

Bir başka modül istatistik işlemleri içindir (Bkz. Şekil 3). İstatistik modülü sayesinde, web sitesine en sık bağlanan kullanıcılar, web sitesinden en sık istenen dosyalar gibi bilgiler bulunabilir. Ayrıca incelenmek istenen kullanıcıların seçimi konusunda istatistiksel analiz ile faydalı bilgiler elde edilmektedir. İstatistik işlemleri

doğrudan bir veri madenciliği seçeneği olmamasına rağmen yararlı bilginin yüzde seksenini sağlamaktadır.



Şekil 4 - Sorgulama Editörü

Analizler yapılırken veri tabanını çeşitli şekillerde inceleme imkanı sağlayan SQL kullanılmıştır. SQL komutlarından bazıları veri tabanından istatistiksel bilgileri doğrudan sağlayabilmektedir. Örneğin, toplamlar, ortalamalar, dizi içerisindeki değerlerin minimum ve maksimum olanlarını bulma, verileri belirli bir kritere göre sıralama, baştan veya sondan en büyük veya en küçük değerleri bulma SQL ifadeleri ile mümkün olabilmektedir. Esnek bir şekilde SQL ifadeleri oluşturmaya, geliştirmeye yarayan bir modülün geliştirilmesi derinlemesine analizleri kolayca yapabilme olanağı getirmektedir. Uygulama sırasında ekranın sağ ve sol tarafındaki panellerde yapılan iki işlem ile de SQL ifadesi oluşturulabilmekte ve çalıştırılabilmektedir (Bkz. Şekil 4). Sorgu cümlelerinin veri tabanına kaydedilmesi ve gerektiği yerde çağrılması ve ifadenin rahatça geliştirilmesi büyük kolaylık sağlamaktadır. Ayrıca Şekil 4'de görünen pencere üzerinden sonuçların dış ortama özellikle bir metin dosyasına gönderilmesi yapılabilmektedir. Bu veriler, başta EXCEL olmak üzere başka uygulamalarda da kullanılabilmektedir.

Kütüphane Hizmetleri için Veri Madenciliği Deneyleri

Kütüphanede Verilen Hizmetler Arasındaki Eşleştirme Kuralları

Bu uygulamada istatistiksel analizler sonucunda genellikle birlikte istendiği görülen katalog tarama (/asp/tum.asp) ile çevrim içi veri tabanları (/html/veritabanlari.htm) sayfaları arasında ne derece bir uyumun olduğu ölçülmeye çalışılmıştır. Bu şekilde

uyum dereceleri destek ve güvenilirlik bilgileriyle birlikte web sitesindeki bütün sayfalar için de ölçülebilir.

Eşleştirme miktarını bulmak için yerine getirilmesi gereken işlem adımları şu şekildedir:

İlk aşamada analizi yapılacak web günlük dosyaları seçilir (bu uygulamada 86 günlük aralık). Seçilen bu dosyalar hazırlanan yazılım sayesinde veri ambarı oluşturmak amacıyla bir dönüşüm işleminden geçirilerek web günlük veri tabanına aktarılır. Bu aşamada seçim, temizleme ve dönüşüm işlemleri yerine getirilmiş olur.

Bir sonraki aşamada veri tabanına atılan kayıtlar üzerinde SQL ifadeleri yardımıyla özetler oluşturulur. Oturum tanımlamaları sayesinde oturum bazında analiz imkanı sağlanır.

Sıklıkla birlikte kullanılan bu hizmetler arasında bir uyumun varlığı SQL gibi bir sorgu dili/makinesi ile incelendiğinde aşağıdaki sonuçlara ulaşılmıştır:

İstek yapılan toplam gün sayısı= 59
Toplam istek sayısı=1464
Veri tabanları (/html/veritabanlari.htm) hizmetinin istendiği gün sayısı=50
Veri tabanları için istek sayısı=70
Katalog tarama (/asp/tum.asp) hizmetinin istendiği gün sayısı=35
Katalog tarama için istek sayısı=222
Veri tabanları ve katalog tarama hizmetinin birlikte istendiği gün sayısı=32

Bu veriler ışığında;

- Katalog tarama ile veri tabanları hizmeti arasında yüksek seviyede bir uyum olduğu saptanmıştır.
- Katalog tarama hizmetinin kullanıldığı her 35 günden 32 sinde veri tabanları hizmeti de beraber (aynı oturumda) kullanılmaktadır. Her ikisi arasında %91 oranında bir uyum vardır.
- Aradaki yüksek uyum dolayısıyla bu iki hizmetin verildiği sayfaları birbirine bağlamak gerekmektedir.

- Veri ambarında var olan verilere örüntü keşif seçenekleri uygulanarak veri içindeki örüntüler bulunabilmektedir. Örneğin; veri ambarındaki veriler sınıflayıcılardan geçirilerek bazı kurallar elde edilebilmektedir.

Kullanıcıların Kümelenmesi

Web kullanım madenciliğinde önemli bir konu, web kullanıcılarının kümelenmesidir. Kümelemede kullanıcılar genel özelliklerine dayalı olarak gruplara ayrılırlar. Kullanıcı kümeleme özellikle kişiselleştirme tipi uygulamalar için gerekmektedir. Kullanıcıları web sitesi erişimlerine göre kümeleyerek onlara özel hizmetler sunmak mümkün hale gelecektir. Örneğin, kütüphaneye alınan yeni kitaplara ilişkin haberleri içeren duyuru, kişiselleştirme sayesinde herkes yerine sadece ilgili kişilere yapılabilir.

Kullanıcılar, kabaca iyi kullanıcılar ve iyi olmayan kullanıcılar gibi iki gruba ayrılabilir. Erişim yapılan dosyaların niteliği bir bakıma erişimi yapan kullanıcıların da niteliğini ortaya koymaktadır. Örneğin, katalog tarama hizmetini kullanan kullanıcıyla site tanıtım bilgilerine göz atan kullanıcı aynı nitelikte değildir. Birisi sadece ziyaret maksadıyla siteye girmişken diğeri belirli bir kaynağı aramak üzere katalog taramak için siteye uğramıştır.

Web günlük dosyaları üzerinde yapılan uzman denetimleri sayesinde web sitesinden üç tipte dosyanın istendiği ortaya çıkmıştır. Bu dosya tipleri şunlardır.

Herhangi bir işlevi olan dosyalar (genellikle *.asp uzantılı dosyalar ile veritabanlari.htm)

- Bilgi amaçlı dosyalar (genel.htm veya personel.htm gibi dosyalar ve resimler)
- Sitede olmayan dosyalar (*.exe ve *.dll uzantılı ve siteye saldırı amaçlı dosyalar)

Herhangi bir işleve sahip olan dosyaları isteyen kullanıcılar, kullanım yoğunluklarına göre iyi, daha iyi, en iyi şeklinde gruplandırılabilir.

Bilgi amaçlı dosyaları isteyen kullanıcılara potansiyel iyi kullanıcı gözüyle bakılabilir. Bilgi amaçlı dosyalar eğer iyi hazırlanmışsa siteye iyi kullanıcı kazandırmakta önemli bir işleve sahip olabilir.

Bazı kullanıcılar ise sitede olmayan bazı dosyalara istekte bulunmaktadır. Bu tip kullanıcıların amacı, web sunucu üzerindeki boşluklardan faydalanarak sistemi çalışmaz hale getirmek olabilir. Bu tip dosyaları isteyen kullanıcılar, tehlikeli veya kötü kullanıcı olarak tanımlanabilir.

Kullanıcılar, erişim yaptıkları dosya adetlerine göre puanlanabilir ve almış oldukları bu puanlara göre değerlendirilebilirler. Geliştirilen uygulama ile elde edilen verilerden bazıları aşağıdadır:

	İlk.htm	*.asp	*.htm	*.jpg	*.gif	*.exe	*.dll	*.ida	Toplam
Bilg. 1	161	166	326	55	58	0	0	0	766
Bilg. 2	0	0	0	0	0	1063	210	0	1273
Bilg. 3	0	0	0	0	0	843	167	0	1010
Bilg. 4	0	0	0		0	600	120	122	842
Bilg. 5	77	361	204	141	117	1	0	10	901

Tablo 1 - Dosya Kullanım Sıklıkları

Bu verilere dayanarak puanlama yapıldığında 2, 3 ve 4 numaralı bilgisayarların saldırganlara ait olduğu ortaya çıkmaktadır. Yoğunluk testi ile de aynı bilgisayarların saldırgan kullanıcılara ait olduğu bulunabilmektedir.

Kullanım sıklıklarından elde edilen katsayılar ile dosya niteliklerinden elde edilen birim puanlarının belli bir işlemden geçirilmesi sonucu elde edilen puanlarla kullanıcılar benzer gruplar halinde toplanarak kümelenmeleri mümkün hale getirilmiş olur.

İlk.htm	*.asp	*.htm	*.jpg	*.gif	*.exe	*.dll	*.ida
2	4	2	1	1	-1	-1	-1

Tablo 2 - Her Dosya için Birim Puanlar

	İlk.htm	*.asp	*.htm	*.jpg	*.gif	*.exe	*.dll	*.ida	PUAN

Bilg. 1	322	664	652	55	58	0	0	0	1751
Bilg. 2	0	0	0	0	0	-1063	-210	0	-1273
Bilg. 3	0	0	0	0	0	-843	-167	0	-1010
Bilg. 4	0	0	0	0	0	-600	-120	-122	-842
Bilg. 5	154	1444	408	141	117	-1	0	-10	2275

Tablo 3 - Her Bilgisayar için Hesap Edilmiş Puanlar

Tablo 2 ve 3'deki sonuçlara göre 1. ve 5. bilgisayarlar benzer göz atma etkinliği gösterirken, 2., 3. ve 4. bilgisayarlarda kendi aralarında benzer göz atma etkinliği göstermişlerdir.

Kullanıcıların kümelenmesi yapılırken en yakın k komşu algoritması tekniklerinden faydalanılmıştır.

Sistemin Taşınabilirliği

Tablo 2 ve Tablo 3'de görüldüğü gibi, Gebze Yüksek Teknoloji Enstitüsü (GYTE) Kütüphanesi web sitesindeki verilere dayanılarak uyum kurallarının bulunması ve kümeleme işleminin yapılması sağlanmıştır. Sistem günlük verilerine dayalı olarak yerine getirildiğinden diğer web sitelerine de taşınabilir. Sistemin çalışma prensibi, bütün web sitelerinde aynı olmakla birlikte kullanılan nesneler birbirinden farklı olabilmektedir.

Eşleştirme kuralları için en iyi örneğin uygulama pazar sepeti analizi olduğu söylenebilir. Pazar sepeti analizinde, sepette yer alan ürünler arasındaki eşleştirme kuralları incelenerek iki ürünün beraber alınma sıklığı bulunmaktadır. Eşleştirme kurallarını, kütüphane web sitesine uygulamak istendiğinde bu sefer ürünlerin yerini web servisleri, bilgi veya kütüphane hizmetleri almaktadır.

Sonuç

Kütüphane kullanıcılarına daha iyi hizmet verebilmenin en önemli yolu onları tanıyabilmektir. Kullanıcıları tanıyabilmenin en iyi yöntemi ise hiç şüphesiz onların web sitesi üzerinde gezinirken web günlüklerine bıraktıkları denetleme verilerinin veri madenciliği yöntemleri ile analizidir.

Bu çalışmada örüntü keşfi ve daha sonrasında yapılan analizler için kullanılan en önemli yöntem istatistik, en önemli araç ise EXCEL olmuştur. Ayrıca kullanıcıların kümelenmesi için en yakın k komşu algoritmasından esinlenilmiştir.

Kaynakça

Cooley, R., Mobasher, B. ve Srivastava, J. (1999) *Data preparation for mining world wide web browsing patterns*. Knowledge and Information Systems, 1(1):5-32

Cooley, R.; Mobasher, B. ve Srivastava, J. (1997) Web mining: Information and pattern discovery on the World Wide Web. *Proceedings of the Ninth IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97)* içinde (s. 558-567). Minnesota. IEEE Computer Society Press.

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

Garofalakis, M. N., Rastogi, R., Seshadri, S. ve Shim, K. (1999). *Data Mining and the Web: Past, Present and Future* [Çevrim içi], Elektronik adres: *http://citeseer.nj.nec.com/231213.html* [10.07.2002[.

Grossman, R.L. (1998). Data mining challenges for digital libraries. *ACM Computing Surveys (CSUR),* 28(4): 1-2.

Joshi, K. P., Anupam, J., Yesha, Y., ve Krishnapuram, R. (1999). Warehousing and mining Web logs. *Proceedings of the second international workshop on Web information and data management i*çinde (s. 63-68). ACM Press. New York, NY.

Zaiane, O.R., Xin, M. ve Han, J. (1998). Discovering Web access patterns and trends by applying OLAP and data mining technology on Web logs. *Proceeding of the Advances in Digital Libraries Conference (ADL'98)* içinde (s.19:29). Santa Barbara, CA.

Yardımcı Kaynakça

Connolly, T.M. ve Begg, C. E. (1999). *Database Systems: a Practical Approach to Design Implementation, and Management.* Harlow, England: Addison-Wesley.

Cooley, R. (2000). *Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data*. Yayınlanmış doktora tezi. Minnesota Üniversitesi, Minnesota.

Cooley, R., Mobasher, B. ve Srivastava, J. (1999) *Data preparation for mining world wide web browsing patterns*. Knowledge and Information Systems, 1(1):5-32