GENETİK ALGORİTMA YÖNTEMİYLE INTERNET ERİŞİM KAYITLARINDAN BİLGİ ÇIKARILMASI

Resul DAŞ [1], İbrahim TÜRKOĞLU [2], Mustafa POYRAZ [3]

[1] Fırat Üniversitesi, Enformatik Bölümü, 23119, ELAZIĞ, rdas@firat.edu.tr

[2] Fırat Üniversitesi, TEF, Elektronik Bilgisayar Eğitimi Bölümü, iturkoglu@firat.edu.tr

[3] Fırat Üniversitesi, Müh. Fak. Elektrik-Elektronik Mühendisliği, mpoyraz@firat.edu.tr

ÖZET

Internet kullanıcılarının davranış bilgileri, internet sunucularında ham veriler şeklinde tutulmaktadır. Bu kullanıcı erişim kayıt örüntülerinden yararlı bilginin keşfi ve analizi web madenciliği olarak tanımlanabilir. Bu çalışma da, kullanıcı erişim kayıt (log) dosyasındaki ham veriler düzenlenerek, genetik algoritma yöntemi ile bu verilerden istatistiksel bilgi çıkarımı yapılmıştır. Böylece, Internet kullanıcılarının en fazla kullandığı veritabanı adres bilgisi tespit edilmiştir.

Anahtar Kelimeler: Genetik Algoritma, Bilgi Çıkarımı, Web Madenciliği, Internet Erişim Kayıtları.

INFORMATION EXTRACTING FROM INTERNET ACCESS LOGS BY GENETIC ALGORITHM METHOD

ABSTRACT

The information on the behaviors of Internet users is saved on servers as raw data. The discovery and analysis of useful information from these user access logs patterns can be defined as Web Mining. In this study, raw data in user access logs files were disposed and statistical information extraction was performed from these data by genetic algorithm method. In this way, address link database which Internet users used most was determined.

Keywords: Genetic Algorithm, Information Extraction, Web Mining, Internet Access Logs.

1. GİRİŞ

Internet (World Wide Web) dünya üzerinde var olan en büyük bilgi paylaşım ortamıdır. Günümüzde birçok kişi, kurum ve kuruluşlar bilgi paylaşımlarını Internet üzerinden yapmaktadırlar. Böylece Internet üzerindeki veri miktarı da hızlı bir şekilde artmaktadır. Yığınla biriken bu verilere bilgisayar kullanıcılarının kolayca erişebilmesi ve bu verileri kullanabilmesi için web madenciliği yöntemleri kullanılmaktadır.

Web verilerinden sıralı örüntülerin bulunması, ilginç kullanıcı bilgilerinin çıkarılması gibi birçok çalışma geçmiş yıllarda yapılmış ve farklı yaklaşımlar sunulmuştur. Uğuz v.d. yaptıkları çalışmada, web sunucusunun sistem erişim kayıtlarına web kullanım madenciliği sistemini ve veritabanı yaklaşımı kullanılarak web sayfası ziyaretçilerinin en sık eriştiği sayfa çiftlerini, üniversite içi ve dışı kullanıcı erişim dağılımı gibi tanımsal ilişkileri tespit etmişlerdir [1]. Chen ve Syncara geliştirdikleri Web Mate adlı sistemlerinde, web sayfalarını inceleyerek, web içeriğinden kullanıcı ilgilerini belirlemeyi sağlamışlardır [2]. Böylece web üzerinden arama işlemlerinde kolaylık sağlamışlardır. Şakiroğlu v.d. yaptıkları bir makale çalışmalarında, web erişim kayıt dosyalarından genetik algoritma yöntemiyle sıralı erişimleri tespit etmişlerdir [3]. İşeri tarafından yapılan tez çalışmasında, geliştirdiği yazılım ile web günlüğünden zaman sınırlı bulanık bağıntı kuralları ve sıralı örüntülerin çıkarılmasını sağlamıştır [7]. Benzer şekilde yapılmış bu tür çalışmalarda akıllı bilgi çıkarım teknikleri kullanılmıştır [4].

Bu çalışmanın amacı, Fırat Üniversitesi Bilgi İşlem Daire Başkanlığı bünyesindeki Internet sunucularında metin dosyası olarak tutulan kullanıcı erişim kayıtlarından yararlanarak, genetik algoritma yöntemi ile kampus Internet kullanıcılarının en çok kullandığı akademik veritabanı adres bilgisinin bulunmasıdır. Kullanıcı erişim kayıt dosyalarından akıllı bilgi çıkarım işleminde genetik algoritma yöntemi kullanılmıştır. Bu çalışma uygulaması ile düzenlenen kullanıcı erişim kayıtları içinde yer alan binlerce adres

bilgisi, bilinen veritabanı adres bilgileriyle karşılaştırılmış ve eşleşen bilgi kayıtlarına göre analiz işlemi yapılmıştır.

Makale 5 bölümden oluşmaktadır. Makalenin 2. bölümünde sistemi geliştirmede kullandığımız yöntemlerle ilgili teorik bilgi, 3. bölümünde yapılan uygulamanın aşamaları, 4. bölümünde uygulama sonuçları ve 5. bölümünde ise yapılan çalışmanın değerlendirmesi ve öneriler sunulmuştur.

II. TEORİK BİLGİ

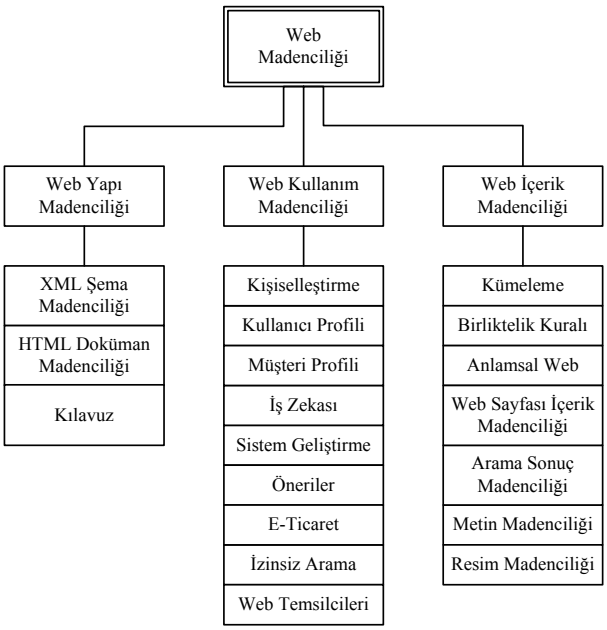
II.1 Web Madenciliği

Internet'ten bilgi çıkarımı ve bilgi keşfi işlemleri, web madenciliğinin önemli bir alanıdır. Web madenciliği, web kayıt dosyalarında ihtiyaç duyulan yararlı bilgilerin çıkarılması ve değerlendirilmesi işlemidir. Internet'te var olan verilerin sürekli olarak değişmesi, güncellenmesi ve yeni bilgilerin eklenmesi web den bilgi çıkarımı işleminde karşılaşılan bir zorluktur. Web sayfalarının bu dinamik yapısından dolayı web den bilgi çıkarımı, normal metin tabanlı dokümanlara göre daha zordur. Şekil 1'de görüldüğü üzere, web madenciliği genel olarak üç alt başlıkta kategorize edilebilir.

Web İçerik Madenciliği: Video, ses, görüntü, bağlantılı ve bağlantısız metinler içeren ve çoğu belli bir düzene sahip olmayan çoklu web dokümanlarından otomatik bilgi çıkarımı web içerik madenciliği ilgi alanına girmektedir. Web içerik madenciliği, bu verilerden anlamlı sonuçlar elde etmek için kullanılan akıllı programlardır. Bu programların amacı, web sayfalarında dolaşarak, bilgiler toplamaktır. Google, Lycos, Altavista gibi bilinen çeşitli arama motorları bu tekniklerden faydalanmaktadırlar [3].

Web Yapı Madenciliği: Web sayfaları arası ya da bir web sayfasındaki bağlantılar (grafik-yazı, grafik-grafik, resim-yazı vb.) arasındaki ilişkileri inceleyerek sonucunda bilgi üretir. Örneğin, önemli web sayfaları belirtilirse, Google arama motoru da tarama sonucunda o sayfaları bulduğunda önemli olarak işaretler. Web içerik madenciliği web sayfasının içeriği ile ilgilenirken, web yapı madenciliği ise doğrudan web sayfaları arasındaki bağlantıları inceler [3].

Web Kullanım Madenciliği: Bu metot ile veri madenciliği yöntemleri kullanılarak, web sunucularında tutulmuş olan erişim kayıtları verilerinden otomatik bilgi keşfi yapılmaktadır. Kullanıcı taleplerine vermiş olduğu hizmetlerin yeterliliği, web sayfalarının kullanma durumlarını, kullanıcıların oturumları ve davranışları tarafından üretilen verilerin incelenmesiyle gibi durumları inceler. Web içerik ve web yapı madenciliği web de birincil veriyi (gerçek veri) kullanırken, web kullanım madenciliği ise kullanıcılar web ile etkileşim halindeyken etkileşimlerinden sağlanan ikincil veriyi kullanır. 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 [8].



Şekil 1. Web Madenciliğinin Sınıflandırılması [8]

II.2 Web Kayıt Dosyaları

Web kayıt dosyaları sunucu platformundan bağımsız metin tabanlı dosyalardır. Dört çeşit sunucu kayıt dosyası vardır. Bunlar:

- Erişim Kayıt Dosyaları (Access Log)
- Hata Kayıt Dosyaları (Error Log)
- İstek Kayıt Dosyaları (Referrer Log)
- Etmen Kayıt Dosyaları (Agent Log)

Internet kullanıcı davranışlarını *erişim kayıt dosyaları*, sunucu üzerinde meydana gelen hatalı işlemleri *hata kayıt dosyaları*, kullanıcı isteklerini *istek kayıt dosyaları*, kullanıcının kullandığı Internet tarayıcısının adı, sürümü ve işletim sistemi hakkındaki bilgileri *etmen kayıt dosyaları* tarafından tutulmaktadır [3].

Bir Internet uygulamasında, web kayıt dosyaları içerisinde bilgi değişiklikleri (kayıt ekleme, kayıt güncelleme ve kayıt silme gibi) olabilir. Bu durumda, tüm veri tabanının defalarca taranıp sık kullanılan öğelerin bulunması hem çok vakit alıcı hem de çok gereksiz olacaktır. Bu nedenle, sadece değişen kayıtlardaki sık kullanılan öğe kümesini güncellemek ve buna göre ilginç örüntüleri keşfetmek için yeni algoritmalara ihtiyaç duyulmaktadır.

II.3 Genetik Algoritmalar

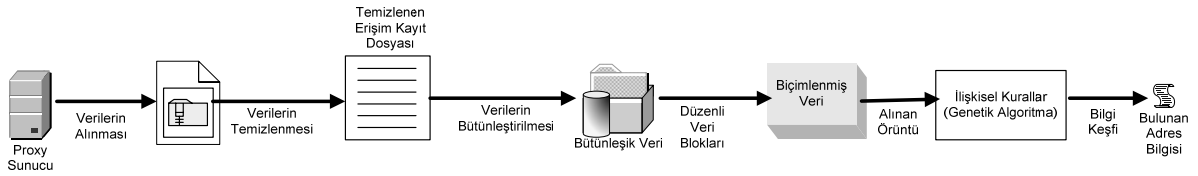
Genetik algoritmalar, değişik planlama teknikleri ile bir fonksiyonun optimizasyonu veya ardışık değerlerin tespitini içine alan birçok problem tipleri için çözüm arama yöntemidir. Genetik algoritmalar, en iyinin korunumu ve doğal seçilim ilkesine dayanarak, benzetim yoluyla bilgisayarlara uygulanan ve bilgisayar üzerinde oluşan bir evrim şeklidir. Bu metot uzun çalışmaların neticesinde ilk defa John Holland tarafından uygulanmıştır [5]. Genetik algoritmaların amacı, hem problemleri çözmek hem de evrimsel

sistemleri modellemektir. Genetik algoritmanın kullanım alanları her geçen gün artmakta olup, genetik algoritmanın temel işlemleri aşağıda adımlar halinde sıralanmıştır:

- Rasgele olarak *başlangıç popülasyonu* oluşturulur. Burada, oluşturulan popülasyon 0 – 1 aralığındadır. Daha sonra bireyler 0 yada 1'e yuvarlanır.
- Rasgele oluşturulan bireylerin her biri uygunluk fonksiyonunda yerlerine konularak değerlendirilir. Yani, bireyler amaç fonksiyonundan geçirilir. Bu işlem, bireylerin iyi olup olmadığını tespit etme işlemidir. *Uygunluk fonksiyonu,* belirlenen çözümlerin uygunluk derecelerinin ölçülmesini sağlayan bir fonksiyondur. Her problem için ayrı bir uygunluk fonksiyonunun belirlenmesi gerekmektedir.
- Bireylere *seçim* yöntemi uygulanır. Seçim işleminde amaç, seçilen uygunluk fonksiyonuna ve seçim yöntemine göre elimizdeki popülâsyondan yeni bir neslin bireylerinin seçilmesidir. Bu seçimde uygunluğu yüksek olan bireyin, yeni nesle aktarılma ihtimali de daha yüksek olacaktır. Böylece bireylerin (kromozomlar) en uygun olanı hayatta kalırken diğerleri de yok olmaya maruz kalacaktır.
- Bireylere *çaprazlama* (gen takası) yöntemi uygulanır. Çaprazlamanın ön adımı olarak çaprazlanacak bireyler eşleme süreciyle belirlenir. Eşleme sürecinde, seçilen kromozomların yeni nesil oluşturma işlemine *çaprazlama* denir. Bir problem çözüm uzayından kaç adet kromozomun çaprazlanacağı çaprazlama oranına göre belirlenmektedir.
- Bireylere *Mutasyon* yöntemi uygulanır. Çaprazlama sonucunda farklı çözümlere ulaşmak bazen zor olmaktadır. Yeni çözüm aramanın kolaylaştırılması ve aramanın yönünü değiştirmek amacı ile bir kromozomun bir elemanının değiştirilmesi işlemidir. Bir problem havuzunda kaç kromozomun mutasyona uğratılacağına mutasyon oranına göre karar verilmektedir.
- Yukarıdaki yöntemler uygulanarak değişime uğramış, yeni bireylere yer açmak için eski bireyler çıkartılarak sabit büyüklükte yeni bir popülasyon oluşturulması sağlanır.
- Sonuçta popülâsyonun hesaplanması sırasında en iyi birey bulunduğunda çözüm elde edilmiş olur. Genetik algoritma ile yapılan uygulamalarda her örnek için tek sonuç üretilir. Tek sonuçta bir kromozoma karşılık gelir.

III. INTERNET ERİŞİM KAYITLARINA GENETİK ALGORİTMA YÖNTEMİNİN UYGULANMASI

Internet sunucularında tutulan kullanıcı erişim kayıt dosyalarına web kullanım madenciliği kapsamında genetik algoritma yöntemini uygulayarak, kampus Internet kullanıcılarının en çok gezindiği akademik veritabanı adres bilgisinin tespiti yapılmıştır. bilgi çıkarımı yapılmıştır. Uygulamada kullanılan web madenciliği sisteminin yapısı Şekil 2'de gösterilmiştir.



Şekil 2. Web Kullanım Madenciliği Mimarisi

Verilerin Alınması: Fırat Üniversitesi Bilgi İşlem Daire Başkanlığı bünyesinde Proxy sunucusunda kaydı tutulan erişim kayıt dosyası üzerinde uygulama yapılmıştır. Şekil 3'de, sunucu üzerinde tutulan erişim kayıt dosyasının metin şeklindeki düzensiz biçimi görülmektedir.

```
CP_IMS_HIT/304 253 GET http://img.sabah.com.tr/i/topbar_kaydet.gif - NONE/- image/gif011623
4612 10.6.2.20 TCP_IMS_HIT/304 254 GET http://img.sabah.com.tr/i/yazar_yukari.gif - NONE/-
0 1335 GET http://anket.memurlar.net/images/common/member6.gif - NONE/- image/gif0116236177
200 6040 GET http://img245.imageshack.us/my.php? - DIRECT/38.99.76.207 text/html01162361772
10.6.2.20 TCP_IMS_HIT/304 253 GET http://img.sabah.com.tr/i/y/t/0002.gif - NONE/- image/gif
254 GET http://img.sabah.com.tr/i/tumhisseler_hdr.gif - NONE/- image/gif01162361772.626
rswebclubhouse.com/club/chat_member.php? - DIRECT/193.239.90.199 text/html01162361772.685
6.2.20 TCP_IMS_HIT/304 253 GET http://www.sabah.com.tr/i/anket_icin_tiklayiniz.gif - NONE/-
M_HIT/200 1110 GET http://www.sabah.com.tr/2006/11/01/gny/im/0647977A493AB745A63DE345e.gif
/302 615 GET http://ad.e-kolay.net/getad.a2? - DIRECT/83.66.160.10 text/html01162361772.940
_MISS/200 44604 GET http://www.internethaber.com/news_detail.php? - DIRECT/89.106.24.67 tex
http://ankt.memurlar.net/images/piechart.aspx? - DIRECT/209.85.10.99 image/gif01162361773.0
l01162361773.086    650 10.6.2.95 TCP_MISS/200 381 GET http://kpss.osym.gov.tr/default.aspx
_MISS/200 312 GET http://ads.sabah.com.tr/adserver/adlog.ads? - DIRECT/213.74.5.114 image/g
.239 text/html01162361773.200     54 10.1.3.23 TCP_MISS/200 381 GET http://kpss.osym.gov.tr
10.6.2.20 TCP_IMS_HIT/304 253 GET http://www.sabah.com.tr/i/_spacer.gif - NONE/- image/gif0
p://kpss.osym.gov.tr/default.aspx - DIRECT/193.140.115.113 text/html01162361773.344     98
```

Şekil 3. Erişim Kayıt Dosyasından Bir Kesit.

Verilerin Temizlenmesi: Karmaşık ve düzensiz bir biçimde bulunan erişim kayıt dosyasındaki verilerin ayıklanarak, belirli bir düzende tablo haline getirilmesi için *Squid Analysis Report Generator (SARG)* programı

kullanılmıştır [11]. Bu program kullanılarak Internet kullanıcı erişim kayıt dosyası çağrılmıştır.

Pedro Lineu Orso tarafından C programlama dilinde yazılmış olan SARG programı, Linux ve Unix tabanlı işletim sistemlerinin bulunduğu sunucularda çalışmaktadır [11]. Bu program Şekil 4. de görüldüğü gibi, sunucu üzerindeki metin tabanlı dosyaları alıp, belli bir düzende tablo haline dönüştürerek HTML formatında oluşturulmasını sağlamaktadır.

SITE	BAGLANTI	BYTE	%BYTE	ICERI-CACHE-DISARI		HARCANAN ZAMAN	MILISANIYE	% ZAMAN
www.visual-prolog.com	77	18.18M	54.29%	0.04%	99.96%	00:02:40	160.80K	1.13%
www.sciencedirect.com	371	3.32M	9.92%	3.15%	96.85%	00:02:06	126.45K	0.89%
www.esbuilder.com	16	3.03M	9.06%	0.37%	99.63%	00:00:36	36.88K	0.26%
www.gensym.net	127	1.92M	5.76%	0.80%	99.20%	00:00:34	34.02K	0.24%
ieeexplore.ieee.org	3	1.35M	4.04%	0.00%	100.00%	00:00:22	22.32K	0.16%
mail.google.com	554	877.30K	2.62%	0.04%	99.96%	00:03:46	226.40K	1.59%
journals.tubitak.gov.tr	1	868.40K	2.59%	0.00%	100.00%	00:00:02	2.02K	0.01%
www.gensym.com	40	584.62K	1.75%	0.00%	100.00%	00:00:20	20.87K	0.15%
www.pcai.com	96	348.68K	1.04%	39.99%	60.01%	00:00:24	24.02K	0.17%
www.adobe.com	96	323.55K	0.97%	79.94%	20.06%	00:00:17	17.57K	0.12%
www-s.tucows.com	36	216.32K	0.65%	63.57%	36.43%	00:00:15	15.54K	0.11%
www.google.com.tr	70	212.90K	0.64%	9.31%	90.69%	00:00:17	17.57K	0.12%
www.elsevier.com	33	182.97K	0.55%	54.35%	45.65%	00:00:01	1.30K	0.01%
www.softpedia.com	63	171.34K	0.51%	16.15%	83.85%	00:00:10	10.95K	0.08%
www.programsdb.com	27	155.88K	0.47%	9.37%	90.63%	00:00:09	9.64K	0.07%
www.pdc.dk	10	155.29K	0.46%	0.00%	100.00%	00:00:03	3.53K	0.02%
b.mail.google.com	58	149.42K	0.45%	0.00%	100.00%	03:42:16	13.33M	93.58%
www.brothersoft.com	85	147.01K	0.44%	8.83%	91.17%	00:00:11	11.09K	0.08%
g-images.amazon.com	4	121.73K	0.36%	100.00%	0.00%	00:00:00	53	0.00%
dchublist.com	1	121.33K	0.36%	0.00%	100.00%	00:00:05	5.05K	0.04%

Şekil 4. SARG Programı ile Düzenlenmiş Kullanıcı Kayıtları.

Verilerin Bütünleştirilmesi: HTML biçimdeki kullanıcı kayıt verileri, Şekil 5.de görüldüğü gibi MS Excel programı kullanılarak artık verilerden ayıklanmıştır. MS Excel dosyası (XLS) biçimine dönüştürülmüş verilerden, istenilen bilgilerin çıkarılabilmesi için bu veriler MATLAB programı kullanılarak veritabanına aktarılmıştır. Daha sonra MATLAB programında genetik algoritma yöntemi kullanılarak yazılan program ile istenilen bilginin çıkarımı yapılmıştır. Bu uygulama da, kampus ağındaki Internet kullanıcıları tarafından en çok kullanılan akademik veritabanı bilgisi bulunmuştur.

A	B	C
1	http://www.sciencedirect.com	www.firat.edu.tr
2	http://isiknowledge.com	www.hurriyet.com.tr
3	http://www3.interscience.wiley.com/cgi-bin/home	http://www.sciencedirect.com
4	http://www3.interscience.wiley.com/journalfinder.html	download.windowsupdate.com
5	http://www.acs.org	http://ieeexplore.ieee.org/Xplore/DynWel.jsp
6	http://pubs.acs.org/about.html	rad.msn.com
7	http://pubs.acs.org/journals/query/subscriberSearch.jsp	www.google-analytics.com
8	http://taylorandfrancis.metapress.com	www.internethaber.com
9	http://journalsonline.tandf.co.uk	www.google.com
10	http://link.springer.de	http://www.sciencedirect.com
11	http://www.springerlink.com	www.sabah.com.tr
12	http://www.blackwell-synergy.com	www.firat.edu.tr
13	http://ieeexplore.ieee.org/Xplore/DynWel.jsp	download.windowsupdate.com
14	http://site.ebrary.com/lib/firat	http://link.springer.de
15	http://www.ulakbim.gov.tr/cabim/vt/	www.ankara.edu.tr
16	http://www.engineeringvillage2.org	www.sabah.com.tr
		http://www.sciencedirect.com
		http://www.ulakbim.gov.tr
		http://www.basbakanlik.gov.tr
		http://www.birses.net

Şekil 5. MS Excel Programı ile Düzenlenmiş Web Kayıtları

Tekrarlı örüntülerin bulunması: Uygulamanın bu aşamasında, düzenlenmiş veritabanına genetik algoritma yöntemini uygulayarak kullanıcılar arasında en çok ziyaret edilen web sayfası adresinin bulunması

amaçlanmıştır. MATLAB programında kodlanarak, uygulaması yapılan genetik algoritma metodunun adımları aşağıda sıralanmıştır:

1.Adım – Kodlama: Internet kullanıcıları tarafından en çok ziyaret edilen web sayfasının bulunması amacıyla program kodlaması yapılmıştır. Bu kodlama işlemini yaparken, aranan sayfa için ikili kod verilmiştir. Uygulamada en çok ziyaret edilen tek web sayfası arandığı için 8 bitlik kodlama yapılmıştır. Bir sayfanın kodlanmasında, bulunması muhtemel 256 sayfayı gösterebilecek 8 bitlik ikili kodlama kullanılmıştır. Birlikte en çok ziyaret edilen ilk 5 sayfa aranacak olsaydı, 40 bitlik kodlama işlemi yapma durumunda olacaktık. Bu arama uzayında arama yapmak üzere oluşturulacak popülasyonun büyüklüğü 10 kromozom olarak belirlenmiştir.

2.Adım – Uygunluk Fonksiyonu: Uygunluk fonksiyonu olarak Internet kullanıcılarının en çok girmiş olduğu web sayfasının tespiti amaçlanmıştır. Bunun için program içerisine dâhil edilen düzenli erişim kayıt dosyasında o kromozomun (web sayfasının) kaç defa tıklanmış olduğunun tespit edilmesidir. Her bir kromozomun metin sütunu içerisinde kaç defa tıklanmış olduğunu bulmak, bize uygunluk fonksiyonunu verir. Adres sayfası olarak da, kromozomlar arasındaki uygunluk fonksiyon değeri en büyük olan alınır.

Uygunluk fonksiyonunda kullanılan parametrelerin anlamları aşağıda belirtilmiştir.

UF(x) = Temel Uygunluk

T =Veri tabanı dosyasındaki işlemlerin toplam sayısı

$K(t_i)$ = En fazla ziyaret edilen veritabanı adresinin toplam işlemler içindeki bulunma oranı

$$M = \sum_{i=1}^{T} K(t_i)$$

UF1 = Aranan sayfa adresi = M / T

M = Bütün $K(t_i)$ oranlarının toplamı

3.Adım – Seçim: Uygunluk fonksiyonundan gelen bireyin bir sonraki nesile aktarılmasına karar vermek için Rulet Tekeri yöntemi kullanılmıştır. İlk olarak tüm kromozomların amaç fonksiyonlarının toplamı bulunur. Her bir kromozomun seçilme olasılıkları ve birikimli olasılık değerleri bulunduktan sonra 1'den 10'a kadar 0 – 1 Aralığında rasgele sayılar atanır. Bu sayılar birikimli olasılık değerleriyle karşılaştırılır. Bunun sonucunda istenilen kromozomlar seçilir.

4.Adım – Çaprazlama: Popülâsyondaki tüm elemanlar çaprazlama işlemine tabi tutulmuştur. Çaprazlama işlemine, tek noktalı çaprazlama yöntemi uygulanmıştır. Nokta olarak ise bireylerin 4.geninden sonrası seçilmiştir.

5.Adım – Mutasyon: Popülâsyonda çeşitliliği sağlayan en önemli faktörlerden biri olan mutasyon işlemi için 1. kromozomun 5.geni dikkate alınarak yapılmıştır.

IV. UYGULAMA SONUÇLARI

Proxy sunucusu üzerinde tutulan Internet kullanıcı erişim kayıt dosyasının günlük bilgi kaydının sıkıştırılmış boyutu yaklaşık olarak 250 MB büyüklüğündedir. Proxy sunucu üzerindeki erişim kayıt dosyası büyüdükçe, SARG programı ile sıkıştırılıp yedeği alınmaktadır. Bu metin kayıt dosyası yüz binlerce satır karakterlerden oluştuğu için dosyanın herhangi bir metin programı ile açılması oldukça güç, bilgilerin anlaşılması da zordur. Bu nedenle verilerdeki kodlar ve numaralar programlarla analiz edilip, anlamlı veriler ortaya çıkarılmaktadır.

- Uygulamada, Proxy sunucusundan dosyaların alınması ve düzenli tablo haline getirilmesi işlemlerinde C++ programı ile yazılmış olan SARG programı kullanılmıştır.
- HTML dosyası biçiminde düzenli tablo halinde bulunan kullanıcı verileri, MS Excel programı ile artık verilerden temizlenmiştir.
- Kampus ağı Internet kullanıcılarına açık olan akademik veritabanı adres bilgileri, kütüphane sayfasından alınarak düzenli MS Excel dosyasında yeni bir sütun bilgisi olarak eklenmiştir.
- Genetik algoritma yöntemiyle MATLAB da yazılan program, bu düzenli ve temizlenmiş MS Excel dosyasını kendi veritabanına aktarmıştır. Daha sonra program, web sayfası adres bilgilerini MATLAB veritabanından okumuştur.
- Genetik algoritma yöntemi kullanılarak yazılan programda, Internet kullanıcı erişim kayıtlarından Internet kullanıcılarının en çok kullandığı akademik veritabanı adresi bulunmuştur.

V. SONUÇ

Internet kullanımının yaygınlaşması, Internet sunucuları üzerinde tutulan verilerin de hızlı bir şekilde artmasına neden olmuştur. Web kayıt dosyaları olarak saklanan bu metin tabanlı verilerin analiz edilerek faydalı bilgilerin çıkarılması ve yorumlanması Web Madenciliği teknikleriyle gerçekleştirilmektedir.

Bu çalışmada, Fırat Üniversitesi Proxy sunucusundan alınan Internet kullanıcı erişim kayıtlarına web kullanım madenciliği uygulanarak, akıllı bilgi çıkarımı için genetik algoritma yöntemi kullanılmıştır. Sonuçta, kampus Internet kullanıcılarının en çok kullandığı akademik veritabanı adres bilgisinin tespiti yapılmıştır.

Yazılan bilgisayar programı geliştirilerek, Internet kullanıcıları arasında en çok ziyaret edilen web sayfa grubu, web sayfaları içerisinde ulaşılamayan web adreslerinin (kırık bağlantılar) tespiti, kullanıcıların en çok zaman geçirdiği Internet sayfaları gibi bilgiler bulunabilir.

TEŞEKKÜR

Makale uygulamamızda kullanmış olduğumuz *Internet kullanıcı erişim kayıt dosyalarını*, tarafımıza sağlayan Fırat Üniversitesi Bilgi İşlem Daire Başkanlığı'na teşekkür ederiz.

KAYNAKLAR

[1]. Uğuz, H., Kodaz, H., Saraçoğlu, R., Baykan, Ö.K., "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, s. 45 - 47, (2003).

[2]. Chen L., Sycara K., "WebMate: A Personal Agent for Browsing and Searching", The Second International Conference on Autonomous Agents, ACM., (1998).

[3]. Şakiroğlu, A.M., Tuğ, E., Bulun, M. "Web Log Dosyalarından Genetik Algoritma Yöntemiyle Sıralı Erişimlerin Tespit Edilmesi", Türkiye Bilişim Derneği 20. Bilişim Kurultayı, (2003).

[4]. Cooley, R., Mobasher, B. and Srivastava, J. "Web Mining: Information and Pattern Discovery on the World Wide Web", Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55455, USA, (1997).

[5]. Nabiyev., V.Vasif, Yapay Zekâ Kitabı, Seçkin Yayınevi, Ekim 2003, Ankara.

[6]. Kösehan, Y., Leblebicioğlu, K., "Mayın Tarlası Oluşturma Problemine Genetik Algoritma Yaklaşımı", KHO Savunma Bilimleri Dergisi, Vol. 2, s.34–56, (2003).

[7]. İşeri, İ., "Web Günlüğünden Zaman Sınırlı Bulanık Bağıntı Kuralları ve Sıralı Örüntülerin Çıkarılması", Fırat Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, Elazığ, (2005).

[8]. Sushmita Mitra, Tinku Acharya, "Data Mining: Multimedia, Soft Computing and Bioinformatics" A John Wiley & Sons, Inc. publication, USA, (2003).

[9]. Nong Ye, "The handbook of Data Mining", Lawrence Erlbaum Associates publishing Company Inc. London, (2003).

[10]. Michael J.A.Berry, Gordon Linoff, "Data Mining Techniques", published by John Wiley & Sons, Inc. USA, (1997).

[11]. Internet: SARG, http://sarg.sourceforge.net, *Erişim tarihi:* Aralık 2006.

[12]. J.Srivasta, R.Cooley, M.Deshpande and P.Tan, "Web Usage Mining: Discovery and Aplications of Usage Patterns From Web Data" SIGKDD Exploartions. 1(2), 1–12, (2000).

[13]. Bulut, B., "Veri Madenciliği Yöntemlerinin İncelenmesi ve Uygulamaları", Fırat Üniversitesi, Fen Bilimleri Enstitüsü, Y.L. Semineri, Elazığ, (2006).

[14]. Emel, G.G., Taşkın, Ç., "Genetik Algoritmalar ve Uygulama Alanları", Uludağ Üniversitesi, İktisadi ve İdari Bilimler Fakültesi Dergisi, Cilt XXI, Sayı 1, s.129-152, (2002).

[15]. Ye, Nong (Ed), "The Handbook of Data Mining", Lawrence Erlbaum Associates, Publishers, Mahwah, New Jersey, London, (2003).