
GAZİ ÜNİVERSİTESİ
Gazi
1926

WEB KULLANIMI MADENCİLİĞİ TEKNİKLERİYLE BİR ÖNERİ SİSTEMİ UYGULAMASI GELİŞTİRİLMESİ

Şükrü Can ŞAYAN

YÜKSEK LİSANS TEZİ

YÖNETİM BİLİŞİM SİSTEMLERİ ANA BİLİM DALI

GAZİ ÜNİVERSİTESİ

BİLİŞİM ENSTİTÜSÜ

ARALIK 2019

Şükrü Can Şayan tarafından hazırlanan "WEB KULLANIMI MADENCİLİĞİ TEKNİKLERİYLE BİR ÖNERİ SİSTEMİ UYGULAMASI GELİŞTİRİLMESİ" adlı tez çalışması aşağıdaki jüri tarafından OY BİRLİĞİ ile Gazi Üniversitesi Yönetim Bilişim Sistemleri Anabilim Dalında YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Danışman: Dr. Öğr. Üyesi Tahsin ÇETİNYOKUŞ ...……
Endüstri Mühendisliği Ana Bilim Dalı, Gazi Üniversitesi
Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum.

Üye: Prof. Dr. Suat ÖZDEMİR ...……
Bilgisayar Mühendisliği Ana Bilim Dalı, Gazi Üniversitesi
Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum.

Üye: Dr. Öğr. Üyesi Abdullah YILDIZBAŞI ...……
Endüstri Mühendisliği Ana Bilim Dalı, Ankara Yıldırım Beyazıt Üniversitesi
Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum.

Tez Savunma Tarihi: 11/12/2019

Jüri tarafından kabul edilen bu tezin Yüksek Lisans Tezi olması için gerekli şartları yerine getirdiğini onaylıyorum.

……..…….

Doç. Dr. Aslıhan TÜFEKCİ
Bilişim Enstitüsü Müdürü

ETİK BEYAN

Gazi Üniversitesi Fen Bilimleri Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmasında;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmasında yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Şükrü Can Şayan

11.12.2019

WEB KULLANIMI MADENCİLİĞİ TEKNİKLERİYLE BİR ÖNERİ SİSTEMİ UYGULAMASI GELİŞTİRİLMESİ

(Yüksek Lisans Tezi)

Şükrü Can ŞAYAN

GAZİ ÜNİVERSİTESİ

BİLİŞİM ENSTİTÜSÜ

ARALIK 2019

ÖZET

Web hızla gelişen ve büyüyen bir ağdır. Bu ağ içerisinde çok büyük miktarda anlamlandırılmayı bekleyen veri bulunmakta ve bu veriler her bir ziyaretçi ile daha da genişlemektedir. Günümüzde bu büyüklükteki verilerin incelenip, anlamlandırılması veri madenciliği süreçleri ile mümkün olmaktadır. İş süreçlerinin web üzerine taşınması ile müşteri davranışlarının analizi ve bunları gelire dönüştürecek modellerin bulunması için de veri madenciliği teknikleri kullanılmaya başlanmıştır. Bu süreçte en sık kullanılan yöntemlerden biri web kullanım madenciliğidir. Bu yöntem genel olarak web sunucusu erişim kayıtları üzerinden yapılan çalışmalardır. Fakat bu çalışmalar genel olarak yüksek miktarda ön işleme faaliyeti gerektiren süreçlerdir. Bu çalışma ile bu alanda yer alan veri kaynağının değiştirilmesi ve süreçte yapılan bu iyileştirme ile de daha etkin bir analizin yapılabileceği gösterilmiştir. Veri kaynağı olarak web erişim kayıtları yerine Google tarafından geliştirilen Google Analytics uyarlanarak bütünleşmiş veri kaynağı olarak kullanılmıştır.

Bilim Kodu : 114611
Anahtar Kelimeler : Veri yönetimi
Sayfa Adedi : 85
Danışman : Dr. Öğr. Üyesi Tahsin ÇETİNYOKUŞ

DEVELOPING A SUGGESTION SYSTEM APPLICATION WITH WEB USE MINING TECHNIQUES (M. Sc. Thesis)

Şükrü Can ŞAYAN

GAZİ UNIVERSITY

INFORMATICS INSTITUTE
DECEMBER 2019

ABSTRACT

The web is a rapidly developing and growing network. There is a huge amount of data in this network waiting to be interpreted, and this data is expanding with each web visitor. Nowadays, it is possible to analyze data and get meaningful results with data mining processes. Today businesses moves their work processes to internet. Data mining techniques have been used in order to analyze customer behavior on web and to find models that will turn them into revenue. One of the most commonly used methods in this process is web usage mining. This method is generally done through web server access logs files. However, these studies are generally require a high amount of pre-processing job. In this study, it is aimed to change the data source in this field and to make a more effective analysis with this changed data source.

Science Code	:	114611
Key Words	:	Data management
Page Number	:	85
Supervisor	:	Asst. Prof. Dr. Tahsin ÇETİNYOKUŞ

TEŞEKKÜR

Çalışmalarım sırasında, yardım ve katkılarını sunan, beni süreç boyunca yönlendirip çalışmalarımda destek olan danışmanım Dr. Tahsin ÇETİNYOKUŞ'a analiz çalışmalarımda kendi verilerini kullanmama izin vererek analiz verisi sağlayan Mahmut KARACA ve Mehmet KARACA'ya ve bu süreçte her türlü desteği ve anlayışı göstererek, tüm süreçlerde yanımda olan Ezgi ŞAYAN'a teşekkürü bir borç bilirim.

İÇİNDEKİLER

Sayfa

ÇİZELGELERİN LİSTESİ

Çizelge **Sayfa**

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

HARİTALARIN LİSTESİ

Harita **Sayfa**

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış bazı simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Kısaltmalar	Açıklama
ARPANET	Gelişmiş araştırma projeleri ajans ağı
CERN	Avrupa nükleer araştırma örgütü
HTML	Hiper metin işaret dili (hyper text markup language)
HTTP	Hiper metin transfer protokolü (hyper text transfer protokol)
HTTPS	Güvenli hiper metin transfrer protokolü
WWW	Dünya çapında ağ (world wide web)

1.GİRİŞ

Yirminci yüzyıl içerisinde birçok bilimsel ve teknik gelişmeyi getirmiştir. Bu gelişmelerin belki de en önemlisi dünya çapında ağdır. Bugün internet olarak bildiğimiz bu araç, dünya üzerinde yer alan her bilgiyi veriyi birbirine bağlamakta ve muhtemelen yakın gelecekte daha güçlü araçlar ile bağlar olmaya devam edecektir. Bu gelişmiş ve karmaşık yapıyı daha iyi anlamak, bu yapıdan bilgi çıkarımı yapmak gitgide ödemi artan bir konu olmaktadır. İnterneti daha iyi analiz edip anlamak ve şekil verebilmek için birçok çalışma yapılmaktadır. Bu durum veri madenciliğine yeni bir dal olarak web madenciliğini eklemiştir.

Bu araştırmada, web madenciliğinin bir kolu olan web kullanım madenciliği ele alınmıştır. Web kullanım madenciliği, gelişen internet teknolojileri ile hızla değişen ve gelişmeye ayak uydurulması gereken bir alandır. Bu alanda yer alan problemlerin çözülmesi, web sitelerinin kullanım örüntülerinin tespitinde fayda sağlayacak ve kullanım analizlerini güçlendirerek internetin sağladığı veri ve bilgilerin kullanıcının daha kolay erişilebileceği hale getirilmesini destekleyecektir. Diğer yandan web yayıncılarının ve internet üzerinden iş yapan kurumların, müşterileri ile yüz yüze gelmeden onları ziyaret hareketlerinden tanımalarını sağlayacaktır. Bu alanda birçok veri işleme ve modelleme yöntemi bulunmakta ve akademik çalışmalar bu araç ve süreçlerin optimizasyonu ile oldukça ilgilenmektedir. Fakat bu çalışmalarda genel olarak sunucu erişim kayıtlarının kullanıldığı açık bir şekilde ortadadır. Bu araştırma ile, alanda yapılan optimizasyonu daha erken bir süreç olan veri toplama sürecine çekmenin faydaları incelenecektir.

Bu tez çalışması ile incelenen veri kaynağı değiştirme çalışması, gelişen web araçları ve değişen ihtiyaçlara, geleneksel metotlardan daha iyi cevap verecek bir yöntem geliştirilmesi, bu alanda yapılacak tüm çalışmalar için yeni yollar açacaktır. Bu çalışma içerisinde gösterilen yöntem ve araçlar ile, yazılım sektöründe olan araç deneyimi ile akademik çalışmalarda yer alan literatür bilgisini bir arada harmanlayarak, bütün aktörlerin faydasına bir çalışma yöntemi ortaya koymayı amaçlamaktadır.

Google günümüzde internetin en önde gelen şirketi ve sektöre yön veren bir fikir kaynağıdır. Google Analytics ürünü, web ziyaretleri hakkında, web sunucu kayıtlarına göre çok daha kaliteli veri toplamakta ve çalışmalarda yapılan ön işleme süreçlerinin birçoğunu gereksiz

kılmakta, bazı durumlarda ise ön işlemeye ihtiyaç olmadan sonuç alabilecek kalitede bilgiler sunabilmektedir. Bu çalışma ile Google Analytics kullanımı değil, Analytics ürünün veri kaynağı olarak kullanımının katkıları incelenmiştir.

Google Analytics kullanımı da elbette her yöntemde olduğu gibi bazı sınırlamalar getirmektedir. Örneğin platformun gizlilik seçenekleri kullanıcıların veri toplanmasına engel olmasına imkân sağlamaktadır. Fakat veri kaynağı değiştirmenin süreçlere genel katkısı ve bir ziyaret başına toplanabilen veri sayısındaki önemli artış dikkate alındığında bu durum oldukça önemsiz bir sınırlama olmaktadır.

Veri kaynağının geliştirilmesi, web kullanım madenciliği süreçlerine büyük ilerleme imkânı sağlayacak bir iyileştirmedir ve Google Analytics' in bu amaçla entegre sistemlerde kullanılması oldukça uygun bir çözümdür.

2. WEB

Günümüzde iletişim, çağlar boyunca hiç olmadığı kadar ihtiyaçtır. Internet erişiminin kolaylıkla erişilebilir ve karşılanabilir olması sonrasında internete erişen kullanıcı sayısında üstel bir artış gerçekleşmiştir. Bu durum sonuç olarak büyük sayılarda işletmenin internet dünyasından faydalanmasına imkân sağlamıştır(Musale and Chaudhari, 2017). Yıllar içerisinde ilerleyen iletişim ihtiyacı ile birçok gelişme sağlanmış ve insanlar arasındaki iletişim de mesaj iletme süresi sürekli olarak kısalmıştır. Bu hızlı değişime ek olarak, kurulan iletişimde medya çeşitliliği sürekli olarak artmıştır.

Telgraf ile sadece nokta ve tire sinyallerden oluşan elektronik iletişim, zamanla telefona, kablosuz telefonlara, internet ortamına, sesli ve görüntülü canlı yayınlara dönüşmüştür. Bu dönüşüm ile bant genişliği artmış ve iletişim kurulan mecraların kalitesi oldukça değişmiştir.

Bu süreç göz önüne alındığında son yıllarda gelişen web teknolojileri ile artık noktadan noktaya aracısız iletişim kurulabilmekte ve iletişimde mesafenin önemi kalmamaktadır. Artık uzay istasyonunda yer alan astronotlar dahi sosyal medya hesaplarını kullanarak anlık olarak bilgi paylaşabilmektir. Bu hızlı gelişimin en büyük sağlayıcısı şüphesiz internet yani dünya çapında ağdır.

2.1. Web'in Tarihçesi

İnternet, ARPANET temelli bir paket değişim teknolojisidir. Temel olarak parçalara bölünmüş bilgi paketlerinin, fiziksel ya da mantıksal katmanlardan geçerek farklı iş birimleri arasında değiş tokuş edilmesini konu almaktadır. ARPANET ilk olarak Amerikan Gelişmiş Savunma Sanayileri Dairesi çalışmaları sonucu ortaya çıkmış ilk bilgisayar ağıdır.

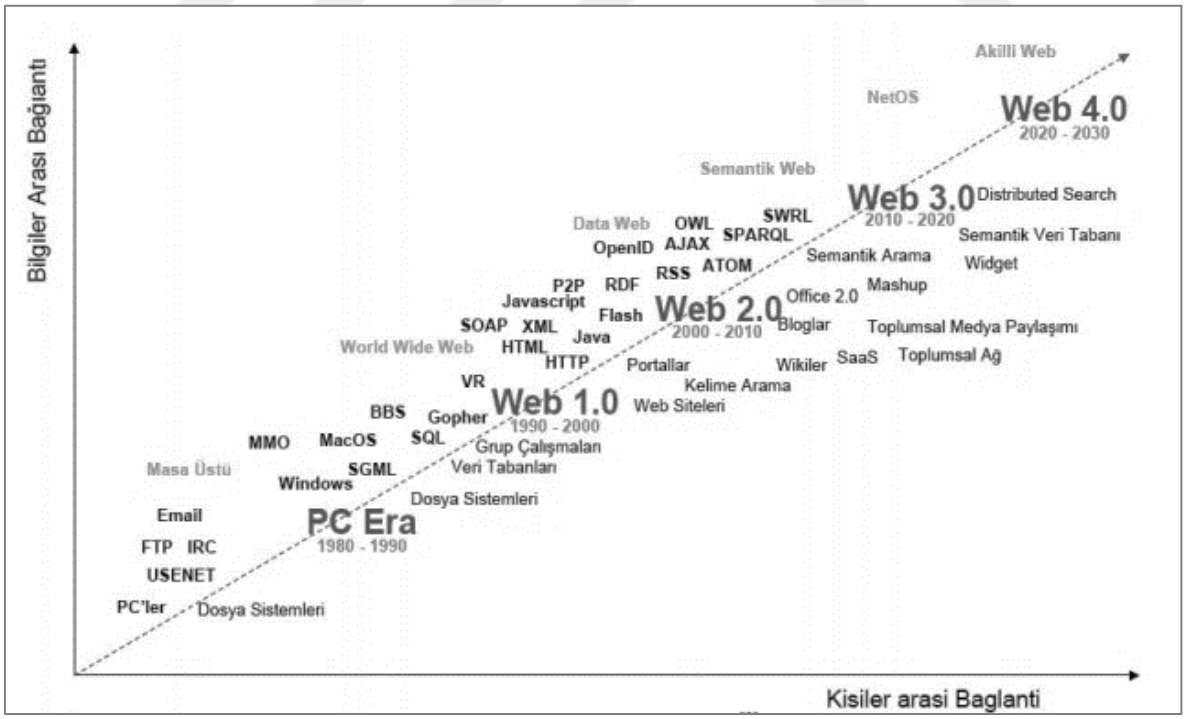
Web ise internet üzerine kurgulanmış ve geliştirilmiş bir teknolojidir. Web, Tim Berners-Lee tarafından CERN'de çalıştığı zaman diliminde 1989 yılında, İsviçre'de bulunmuştur. Web ilk geliştirildiği dönemde istemci ve sunucular arasında süregelen bir trafiktir. Temel olarak birbirine bağlantılı dokümanlardan oluşan bir yapıdır. Zamanla bu ilişkili dokümanlara, resim video gibi medya araçları da dahil olmuştur. Günümüzde ise bu medya araçları fazlasıyla gelişmiş, çok kanallı bir yapı sunmaktadır.

Web kendisi birçok protokolden oluşan sistemler ile çalışmaktadır. Bu protokolün en temel olanı HTTP ve bu protokolün güvenlik katmanı eklenmiş HTTPS versiyonudur. HTTP temel olarak zengin metin aktarım platformu olarak adlandırılabilir. Bu protokol ile istemciler sunuculardan bilgi talep etmekte ya da sunuculara bilgi göndermektedir. Günümüzde yeni oluşan protokoller ile bu iletişim gelişmektedir. Örneğin WEBRTC protokolü, herhangi bir sunucuya ihtiyaç duymadan istemci- istemci iletişimini de mümkün kılmaktadır.

İnternetin ülkemize gelişi ise 80'li yıllara rast gelmektedir. 1993 Yılında ülkemiz ilk olarak küresel ağa bağlanmıştır.

2.1.1. Web'in evreleri

Web gelişimini tarihi ve bant genişliği olarak incelemenin yanı sıra kavramsal evreler olarak incelemekte gerekmektedir. Son yıllarda webin kavramsal yapısı da bir hayli değişiklik göstererek çok kanallı bir iletişim aracına dönüşmesine katkıda bulunmuştur. Web 1.0 ve Web 3.0 arasındaki dönüşüm sürecini Şekil 2.1 ile göstermek mümkündür.



Şekil 2.1. Web'in gelişim evreleri (Aslan, 2007)

Web 1.0

Web 1.0, sadece web sitesini yayınlayanlar ve sitedeki bilgileri okuyanlardan ibarettir. Bu web anlayışının temelinde bir içerik yayınlayıcı bir de içerik okuyucu vardır. Web 1.0 ile site ziyaretçileri ile herhangi bir etkileşim kurulmamaktadır. Bu web siteleri bir kitap ya da dergi gibi sadece site yöneticisi tarafından oluşturulan yayının okunması ve bu konuda bilgi sahibi olunmasını hedefleyen sitelerdir. Bu siteler kullanıcı bilgisayarına indirilip okunduktan sonra herhangi bir geri dönüş almamakta, ticari bir kar elde etmek için diğer iletişim araçlarından yararlanmaktadır. Bu sitelerin sahipleri herhangi bir veri toplama yeteneğine sahip değildir.

Bu siteler bugün alışık olduğumuz sitelerin aksine statik olarak çalışan, herhangi bir karar yapısı içermeyen, arkasında bir web programlama dili içermeyen sistemlerdir.

Web 2.0

Web 2.0 teknolojisi ile birlikte web siteleri üzerinde etkileşimli, kullanıcıların görüş ve düşüncelerini paylaşabilmelerine olanak sunan, çeşitli paylaşımlar için web üzerinde uygun araçların geliştirildiği (Youtube, Facebook gibi) ortamlar oluşturulabilmektedir. Bu çalışmaların ilk örneğini, web sitelerinde yer alan iletişim formları ya da ziyaretçi formları ile görmek mümkündür. Bu yapılar gelişerek kullanıcıdan bilgi toplamanın ötesine geçmiştir. Bugün blog siteleri, sosyal medya kanalları gibi birçok sistemde artık site yöneticileri web için içerik üretilmesini değil, kullanıcıların içerik üretmesini sağlayacak yazılımları geliştirmektedir.

Web 2.0 teknikleri ile yapılan siteler daha çok kullanıcı merkezlidir. Kullanıcının katılımı ile zengin bilgi kaynakları oluşturmak Web 2.0'ın ana hedefleri arasındadır. Bu yeni yaklaşım ile kullanıcılara hareket özgürlüğü ve kullanım kolaylığı esas alınır (Aslan, 2007). RSS (Really Simple Syndication), Wiki, CMS (Content Management System) sistemleri ile kullanıcıların istedikleri içerikleri oluşturabilmeleri sağlanmıştır. Bu dönemin en önemli sonucu, artık teknik bilgi seviyesi neredeyse hiçe yakın olan kullanıcılar bile biraz emek ve para harcayarak internetten yayın yapabilir ve geri dönüş toplayabilir konuma gelmiştir.

Bu dönem aynı zamanda artık statik web sayfalarından veri tabanlarına geçildiği dönemdir. Bugüne kadarki her önemli internet uygulaması, Google'ın web taraması, Yahoo'nun dizini ve web taraması, Amazon'un ürün veri tabanı, eBay'in ürün ve satıcı veri tabanı, MapQuest'in

harita veri tabanları, Napster'ın dağıtılan şarkı veri tabanı gibi birçok veri tabanı kullanımı bu dönemde mevcuttur. Hal Varian bir konuşmasında "SQL yeni HTML'dir" şeklinde betimlemiştir. Veri tabanı yönetimi, Web 2.0 şirketlerinin temel bir yetkinliğidir, hatta bu uygulamalara yalnızca yazılımdan ziyade "bilgi" olarak değinilmektedir (O'reilly, 2005).

Bu dönemin web siteleri artık kullanıcıları kendisine çekmek zorunda kalan, gelirlerini kullanıcılara gösterdikleri web reklamları ile elde eden yapılar olmuştur. Bu nedenle kullanıcıya istediği hizmetin sunulması için bu ihtiyaçları analiz etme ihtiyacı doğmuştur.

Web 3.0

Günümüzde teknoloji gelişirken, yapay zekâlar artmakta ve makineler artık bizi daha iyi anlayabilmektedir. Bu gelişim anlamsal web olarak anılan Web 3.0'ı ortaya çıkartmıştır. Web 1.0 yayıncılar tarafından insanların okuması için üretilirken, Web 2.0 insanlar tarafından diğer insanların okuması için yayınlanmaktaydı. Web 3.0 ile ise, bu yapı daha da gelişmiştir. Web 3.0 ile içerik, insanlar ya da makineler tarafından, insanların ya da makinelerin okumaları için üretilmektedir.

Web 3.0, son birkaç yıldır popülerlik kazandıran uygulamalar ve hizmetler üzerine inşa edilmiştir. Arama motorları çok daha eksiksiz ve hedeflenen bilgileri üretebilmekte; kullanıcılar sosyal medya uygulamaları aracılığıyla arkadaşlarıyla ve işletmelerle daha sıkı bir bağlantı kurmaktadırlar (Nath, Dhar, & Basishtha, 2014). Örneğin, arama motorlarında "python" ve "programming" aramaları ile "python programming" aramasına çok farklı tepki verilebilmektedir. "Programming" aramasına arama motoru genel programlama eğitimi ile ilgili sonuçlar verirken, "python" aramasına Python programlama dili kaynaklarını göstermektedir. "Python programming" ifadesine ise doğrudan dilin kendi sayfası yanında Python programlama eğitimi ile ilgili kaynakları göstermekte, asıl istenen ifadeyi anlayabilmektedir. Bu süreç kişisel istekler incelenerek çıkartılmaktadır. "Python" araması yapan sıradan bir kullanıcıya "yılan" olan pythonun gösterilmesi de oldukça mümkündür. Bu işlem kişisel ilgiler hakkında toplanan verilerin incelenmesi ve analiz edilmesi ile mümkündür.

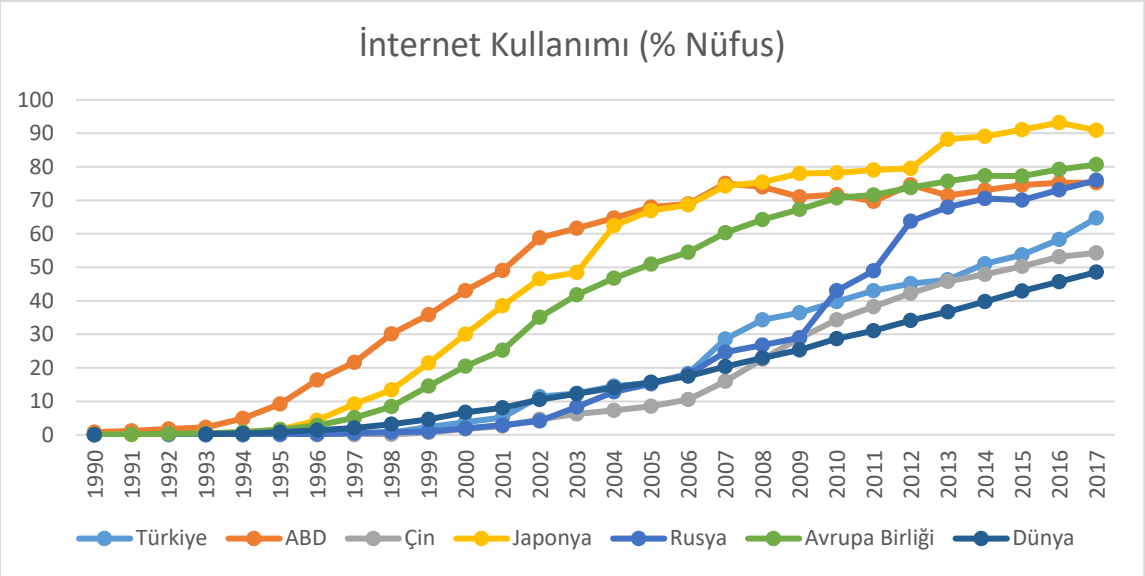
Web evreleri arasında şu şekilde bir karşılaştırma yapmak mümkündür:

Çizelge 2.1. Web 1.0, 2.0 ve 3.0'ın karşılaştırmaları (Nath ve diğerleri 2014).

Web 1.0	**Web 2.0**	**Web 3.0**
Sadece Okunabilir- Statik	Okunup yazılabilir -Etkileşimli	Okunup Yazılabilir- Zeki
Kurum Merkezli	Toplum Merkezli	Birey Merkezli
Düşük Taşınabilirlik	Orta Taşınabilirlik	Yüksek Taşınabilirlik
Uzmanlarca geliştirilmiş dışa kapalı	Kullanıcılarca geliştirilmiş dışa açık	Kullanıcılarca geliştirilmiş zeki
Söz dizimi tabanlı gezinti ve arama yeteneği	Gelişmiş söz dizimi gezinti ve arama yeteneği	Anlam tabanlı, içerik merkezli ileri seviyeli gezinti ve arama yeteneği
Düşük medya zenginliği	Orta medya zenginliği	Yüksek medya zenginliği
Noktadan noktaya mimari	Servis odaklı mimari	Ağ ve bulut odaklı mimari
Ayrık veri	Bağlanmış veri	Dünya çağında veri

2.2. Web'in Büyüklüğü

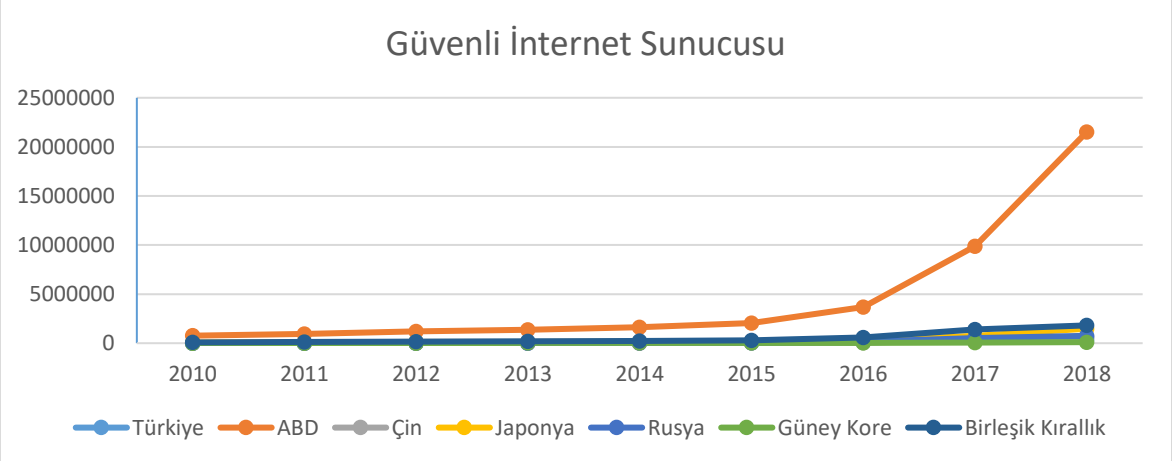
İnternet ve web'in bu kadar geniş kapsama yayılması ile günümüzde birçok cihaz bu ağa erişmekte ve bu ağ üzerinden her an veri alıp göndermektedir. Günümüzde, mobil telefonlar, televizyonlar, arabalar hatta buzdolabı ve kombi gibi beyaz eşyalar bile internet bağlantısı kurmakta web içeriklerini göstermekte ve bu içerikler aracılığıyla bilgi gönderilmesine olanak sağlamaktadır. Bu sebeple dünya üzerindeki internet aboneliği çok hızlı olarak yaygınlaşmakta ve gün geçtikçe internet aboneliğinin hızı artmaktadır. Aşağıda yer alan grafikte dünya üzerinde seçilen ülkelerin nüfusları içerisinde internet erişimi oranı Şekil 2.2.'de gösterilmektedir.



Şekil 2.2. İnternet kullanımı (% nüfus) (Dünya Bankası, 2019a)

Grafikten görülebileceği üzere ülkelerde ve dünya üzerinde internet kullanımı hızla yükselmektedir. 2017 verilerine göre dünyada nüfusunun internet erişimine en çok sahip olduğu en büyük %98.87 oranı ile Andora'dır(Dünya Bankası, 2019a).

İnternetin dünya nüfusuna yayılması yanı sıra internette yer alan güvenli internet sunucu sayısı da büyük bir hızla yükselmektedir. İnternette veri alışverişinin güvenli bir şekilde, üçüncü taraf kişilerin okuyamayacağı şekilde şifrelenerek gönderilmesi için TLS/SSL sertifikaları kullanılmaktadır. Bu sertifika sayılarından faydalanılarak ölçülebilen sunucu sayısı Şekil 2.3' de yer alan grafikte gösterilmektedir.



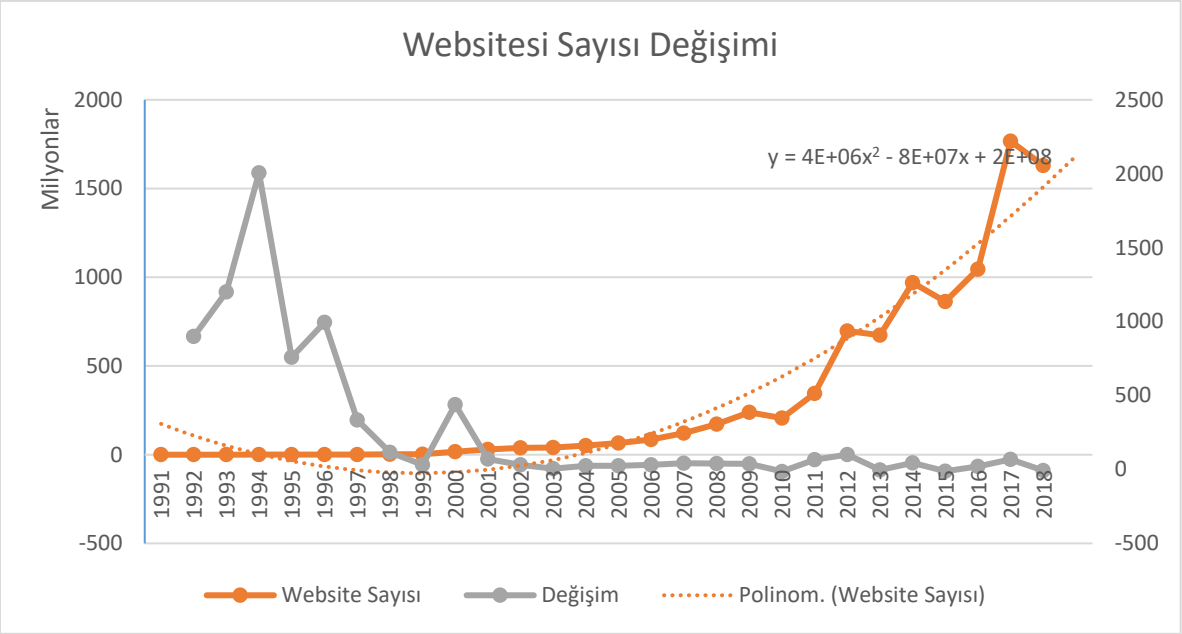
Şekil 2.3. Güvenli internet sunucusu sayısı(Dünya Bankası, 2019b)

İnternet sunucu sayılarına bakıldığında 2018 yılı itibariyle Türkiye bu listede 356.867 adet güvenli sunucu ile 20. sırada yer almaktadır (Dünya Bankası, 2019b). İlk 20 ülke Çizelge 2.2'de gösterilmektedir.

Çizelge 2.2 En çok güvenli internet sunucusu barındıran 20 ülke (Dünya Bankası, 2019b)

Sıra	Ülke	Güvenli İnternet Sunucusu Sayısı
1	ABD	21 517 004
2	Almanya	4 676 435
3	Birleşik Krallık	1 811 830
4	Hollanda	1 733 183
5	Japonya	1 476 696
6	Fransa	1 367 541
7	Kanada	1 147 057
8	Avusturalya	822 020
9	Rusya	749 930
10	İtalya	740 648
11	Danimarka	713 512
12	Güney Afrika	695 336
13	Çin	622 142
14	Polonya	616 219
15	İsviçre	580 292
16	İspanya	528 944
17	Singapur	477 674
18	Çekya	450 120
19	Brezilya	426 560
20	Türkiye	356 867

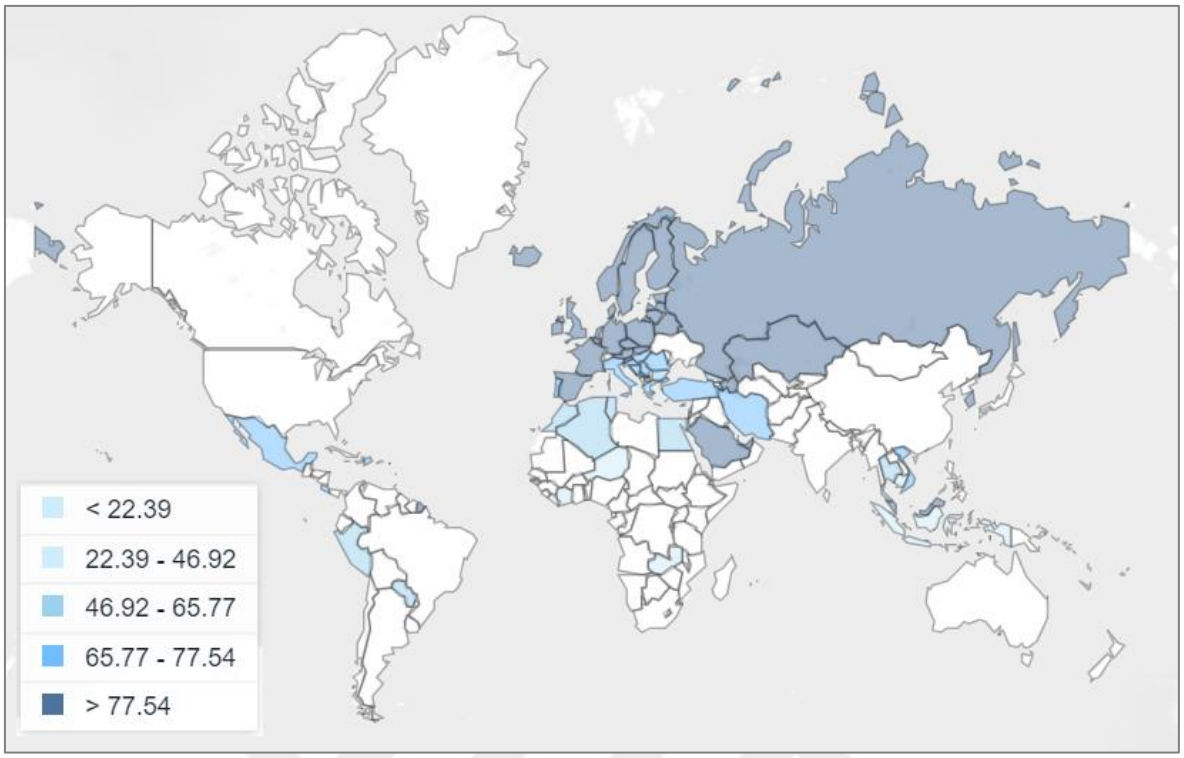
Web içeriği hızla büyümektedir. Webin temel öğesi olan web sitelerine bakıldığında bu durum çok daha açık bir şekilde görülmektedir. InternetLiveStats.com sitesi anlık olarak istatistiklere ve tahmin modellerine göre bilgi vermektedir. Bu bilgilere göre şu an dünya üzerinde yer alan web sitesi sayısı 1 686 737 759'dur (Internet Live Stats, 2019). Bu sayı worldometers.info adresinden alınan dünya nüfusu olan 7 704 159 356 sayısına göre hesaplandığında Dünya üzerinde her 4.56 kişiye bir web sitesi düşmektedir (Worldometers, 2019). Şekil 2.4'te yıllar itibariyle web sitesi sayısı ve değişim oranları gösterilmektedir.



Şekil 2.4. Web sitesi sayı ve değişiminin yıllara göre değişimi (Internet Live Stats, 2019)

İnternette anlık yaratılan veri sayısı giderek artmakta, Web 2.0 ve Web 3.0 ile kullanıcıların içerik oluşturucu hale gelmesiyle büyüme hızlanmaktadır. Şekil 2.3'de internetlivestats.com'dan alınan günlük veriler gösterilmektedir.

Dünya üzerinde bireysel internet kullanıcı sayıları hızla artmakta ve giderek nüfusun çok daha geniş kesimleri internete erişmektedir. Şekil 2.3'de yer alan haritada bireysel internet erişimine sahip kullanıcıların ülke nüfusuna oranı verilmiştir.



Harita 2.3. İnternet kullanan nüfus oranı (Dünya Bankası, 2019a)

2.3. Web'in Geleceği

Web giderek büyüyen bir hızda genişlemektedir. Her geçen gün webde yer alan web sitesi, kullanıcı sayısı ve ziyaretçi miktarı giderek artmaktadır. Bir diğer yandan ise yıllar önce sadece bilgisayarlarımızda yer alan web tarayıcıları televizyonlara, buzdolaplarına ve benzeri birçok cihaza yerleştirilmektedir. IOT dediğimiz internet cihazları ve bu cihazlara bağlı olan milyonlarca sensor her an yeni bilgiler oluşturarak webin erişimine açılmıştır.

Webin bu gelişimi ile birçok eski teknoloji artık kullanılmaz duruma gelmektedir. Günümüzde televizyon kanalları karasal yayın yapmayı bırakmakta ve web tabanlı içerik yayıncılığına geçmektedirler. Eskiden hayatımızda yer alan gündelik eşyalar birer birer web ortamına taşınmaktadır. Örneğin web tabanlı haritalar, günümüzde harita bulundurma ihtiyacını ortadan kaldırmıştır. Gazeteler bile bu duruma uyum sağlayarak, basılı yayınları azaltmaya ve ücretli web abonelikleri satmaya başlamışlardır.

Bu gelişim ve değişim sadece günlük araçları değiştirmekle kalmamış, birçok teknolojiyi de devre dışı bırakmaya başlamıştır. Google'ın geliştirdiği yeni scada teknolojisi ise oyunların bile harici cihazlara ihtiyaç olmadan web üzerinden oynanabilmesini mümkün kılmıştır. Günümüzde, ofis uygulamaları, fotoğraf düzenleme araçları, eposta takvim gibi

uygulamaların tümü web ortamlarına taşınmıştır. Artık kullanıcılar kişisel dokümanlarını bile tarayıcıları içerisinde düzenlemekte ve bu dosyaları bulut depolama ortamlarında barındırmaktadır. Web sadece veri iletme aracı olmaktan çıkmış ve artık bir iş platformuna dönüşmüştür. Bugün sadece web tarayıcısı olan bir bilgisayar, ortalama bir kullanıcının tüm işlemlerinde yeterli olabilmektedir.

Herhangi bir fiziksel mağazası olmayan işlemler web sayesinde satış yapmakta, fiziksel şubesi olmayan bankalar web ve mobil sistemlerle hizmet vermeye başlamışlardır. Günümüzde internet ve web olmadan ekonomik bir sürecin yürütülmesi neredeyse imkânsız hale gelmiştir. Yeni bir teknoloji olan blockchain ve kripto para ekonomisi ise sadece web ve internet kaynaklı bir ekonominin doğmasının yolunu açmaktadır.

Bu gelişme ve durumlar dikkate alındığında, webin giderek büyüyeceğini ve her geçen gün hayatımızın çok daha fazla noktasına erişim sağlayacağını söylemek mümkündür. Bu nedenle her geçen gün webi daha iyi takip edecek ve daha iyi analiz ederek çıkarımlar yapacak araçların inşa edilmesi gerekmektedir.

2.4. Web Analizi

Web ilk günlerinden, günümüze kadar gelirken oldukça hızlı bir şekilde gelişmiş ve genişlemiştir. Günümüzde birçok bilgi ve işlem yoğun sistemler web üzerinden çalışmakta ve büyük ekonomik işlemler web aracılığıyla iletilmektedir. Bu sistemlerin geliştirilebilmesi, sürekliliğinin sağlanması için analiz edilmesi ve bu analiz çıktılarına göre düzeltmelerin yapılması ve optimizasyon çalışmalarının yürütülmesi gereksinimi doğmuştur.

Zamanla gelişen web sunucuları ve çevre yazılımlar birtakım analizlerin yapılmasını olanaklı kılmaya başlamışlardır. Bu alanda yapılan çalışmalar sonucu çeşitli ölçüm parametreleri oluşmaya başlamış ve bu araçlar ile sitelerin karşılaştırmasını yapmak mümkün hale gelmektedir.

Örneğin bir sitenin performansı düşünüldüğünde ilk günlerden beri ilk akla gelen göstergenin günlük ziyaretçi sayısı olduğunu söylemek mümkündür. Bu göstergeler web sitelerini kendi arasında karşılaştırmayı sağladığı gibi, bu bilgiler sayesinde bazı tahmin ve

yatırım kararları yapmanın mümkün olduğu aşikârdır. Örneğin, web sitesinin günlük ziyaretçi sayısında yaşanan değişimler ile sunucu yatırımı gerekecek tarih konusunda tahminlerde bulunmak mümkündür. Ayrıca satışların mevsimselliğini hesaplamakta bu ölçütler ile kolayca hesaplanabilir bir gösterge olacaktır.

Teknolojinin gelişmesi ise şüphesiz ki web sitelerine erişim araçlarının da çeşitlenmesini sağlamıştır. Örneğin farklı internet tarayıcılarının kullanımı, farklı ekran çözünürlükleri, farklı mobil işletim sistemleri gibi birçok parametre, analiz edilebilecek göstergeler oluşturmaktadır. Bir diğer önemli değişim ise artık günümüzde web sitelerinin, oluşturulmuş statik sayfalar olmak yerine, uygulama yazılımları ile oluşan dinamik web içerikleri olmasıdır. Web 2.0 ile artık kullanıcıların da içerik üreterek web sayfaları oluşturmaya başlaması sonucu, web artık ölçülmesi oldukça karmaşıklaşan bir yapı haline dönüşmüştür.

Bütün bu gelişmeler, webin analizi ve anlamlı modellerin çıkarılması için web madenciliği alanının doğmasını sağlamıştır. Veri madenciliği ile birlikte, karmaşık web sitesi ve içeriklerin analizi kolaylaşmış ve gizli kalmış veri desenlerinin keşfi kolaylaşmıştır.

3. VERİ MADENCİLİĞİ

Bilgi sistemlerinin tarihsel gelişimi sürecine baktığımızda, sistemler ilk olarak veri kayıt üzerine kurulmuştur. Veriler kullanıcılar tarafından toplandıktan sonra bilgi sistemleri aracılığı ile kayıt araçlarına kaydedilmekteydi. Bilgi sistemlerinin gelişmesi sonucu bilgi değişimleri, kayıt sistemleri ile bilginin oluşturulması süreci hızlandı. Örneğin ilk web kamerasının, Cambridge Üniversitesi'nde yer alan araştırmacıların kahve makinesinin doluluğunu takip etmek için geliştirildiği birçok kaynakta yer almaktadır. Belki de bu icat ile ilk defa insan haricinde bir sistemin dijital veri elde etmesi sağlanmıştır.

Günümüzde, oluşturulan, aktarılan ve kullanılan veri boyutu oldukça yüksektir. Bu verileri anlık olarak üretilmekte, bazıları hiç kaydedilmeden kullanılarak silinmektedir. Bugün çevremizde yer alan cihazlar çok çeşitli sensörlere sahiptirler. Örneğin bir akıllı telefon global konum, hareket, ivme, ışık ve ses gibi birçok veriyi sensörleri yardımı ile kaydedebilmektedir. Bu verilerden ise sık geçtiğimiz konumlar, anlık trafik, günlük kalori harcaması gibi bilgileri üretebilmektedir.

Artan veri çeşitliliği ve veri üretim hızı artmakta, bu verilerin insanlar tarafından anlamlandırılabilmesi gün geçtikçe zorlaşmaktadır. Çeşitli veri madenciliği araçları da tam bu noktada görev almaktadır. Bilgisayarlar artan veri işleme kabiliyetleri ve günümüz bulut teknolojilerinin sunduğu geniş kayıt ve dağıtık işleme teknolojileri ile yığın verilerden anlamlı bilgiler çıkarma ve yeni şablonları bulma konusunda yardımcı olmaktadırlar.

Veri madenciliği araçlarının kullanımı, birçok farklı sektörde yeni çözümler getirmektedir. Örneğin finansal tahminler, dolandırıcılık tespiti, üretim planlamaları gibi birçok alanda hızlı çözümler getirmektedir. Veri madenciliği uygulamaları sağlık sektöründe doktoralardan daha iyi teşhis ve tanı işlemleri yapabilmektedir. Bugün satranç oynayan bilgisayarlar bile veri madenciliği araçlarını kullanarak kendilerini daha yetenekli hale getirmektedirler.

3.1. Veri Madenciliği Türleri

Veri madenciliği araçları birçok karmaşık problemin çözümünde kullanılmaktadır. Farklı türden sorunları çözmek için geliştirilmiş olan birçok farklı model bulunmaktadır. Bu modeller temel olarak iki grupta incelenebilir. Bunlar gözetimli ve gözetimsiz modellerdir.

3.1.1. Gözetimli modeller

Gözetimli modeller ilgili sürecin girdi ve çıktılarının belirli olduğu durumlarda kullanılan modellerdir. Modelin ürettiği sonuçlar, test için ayrılan veriler ile denetlenerek modelin performansı ölçülebilir. Bu modeller bir nevi hangi sonuçların hangi nedenlere bağlı olduğunu öğrenerek verilen yeni girdilere göre yeni çıktıları tahmin ederler. Bu modellere sınıflandırma modelleri ve regresyon modelleri örnek gösterilebilir.

3.1.2. Gözetimsiz modeller

Gözetimsiz modeller çıktıların belirsiz olduğu ve çıktıların model tarafından oluşturulduğu durumlardır. Gözetimli modeller bir test grubu ile denetlenebilirken, gözetimsiz modeller için böyle bir kontrol mümkün değildir. Kümeleme ve birliktelik modelleri bu gruba giren modellerdir. Bu modeller yeni verilerin en çok hangi gruba benzediğini, yeni gelenin benzer sonuçlar vereceği yapıları gösterirler.

3.2. Veri Madenciliği Süreci

Bilgi sistemleri, veri kayıt sistemlerinden, karar alma süreçlerine yardımcı olan sistemlere dönüşmektedir. Bugün birçok sistem içerisinde karar destek mekanizmaları ile çalışmakta tehditleri algılamakta ve karar alımlarda yardımcı olmaktadır. Etkili karar almak için güvenilir ve hatadan uzak bilgilere sahip olmak gerekir. Veri madenciliği süreci anlamlandırılamayan verilerin işlenerek kararlara yardımcı olacak bilgilere döndürülmesi işlemidir. İşlenmemiş yığın veriler ile güçlü ve etkili kararlar almak oldukça zordur.

Verilerin toplanmasından, bilginin ortaya çıkışına kadar birçok süreç görev almaktadır. Gelişen bulut ve veri tabanı sistemleri ile artık verilerimiz çok farklı kaynaklarda saklanabilmektedir. Bu yüzden veri kaynaklarının bir araya getirilmesi ile veri madenciliği işlemleri başlamaktadır.

Veriler çok farklı formatlarda saklanabilmektedir. Bu verilerin dönüştürülmesi ve aynı yapıya getirilmesi gerekmektedir. Ayrıca bu veriler üzerinde çeşitli dönüşümlerin yapılması, hatalı bilgilerin düzeltilmesi ve eksik bilgilerin tamamlanması gerekmektedir. Kalitesiz bir bilgi ile elde edilecek sonuçlar hatalı olacaktır. Bu süreçlerin bütünü önişleme olarak anılmaktadır.

Elde edilen veriler ile veri madenciliği modelleri çalıştırılmaktadır.

4. WEB MADENCİLİĞİ

Günümüzde web bilgi edinme sürecinde kullanılan ilk araçlardandır. Dünya çapında ağ, yani kısaca web, çeşitli ihtiyaçlarımıza göre faydalanılan çok farklı tiplerde verileri içerisinde saklayan, dünya çapındaki en geniş veri tabanı olarak kabul edilmektedir. Web verisi yapısal olmayan ham veriden oluşur. Eğer doğru şekilde madenciliğe tabi tutulabilirse, kişiler ve firmalar için sanal altın derecesinde değerlidir (Jayamalini and Ponnavaikko, 2017). Geniş, sürekli gelişen ve dinamik bir yapı olması sebebiyle güncel bilgiler çok süratli yayılmakta ve mevcut bilgilere de kolaylıkla erişim sağlanabilmektedir.

Her geçen gün daha fazla veri ve işlem web sitelerine taşınmakta, web kullanımı giderek yaygınlaşmaktadır. Bu durum webin daha iyi anlaşılmasını ve incelenebilmesini gerektirmektedir. Günümüzde dünya çapında ağın ölçüsüz büyümesinin yönetimi, şirketler, organizasyonlar, teknolojiler ve bireyler için artan bir trend haline gelmiştir. Bu verinin yönetilmesi için web madenciliği gerekli bir araçtır.(Kamakshi, Mehrotra, and Deep, 2018) Fakat şüphesiz ki bu kadar geniş bir ortamın elle incelenmesi ve anlaşılması, gelişim ve genişleme hızı karşısında oldukça zordur. Örneğin Netflix gibi platformlar, internet trafiğinin büyük bir kısmını oluşturmakta, sosyal medya siteleri ise milyonlarca kullanıcıya hizmet vermektedir.

Mevcut internet sitelerinin kendini geliştirmesi ve rakip siteler ile rekabet edebilmeleri, kullanıcılarını anlamaktan ve kullanıcı ihtiyaçlarına cevap verebilmelerinden geçmektedir. Bu süreçte bir diğer önemli konu ise şüphesiz bulunabilirlik konusudur. Bütün bu durumlar göz önüne alındığında web sitelerinin zenginleşmesi ve iyileşmesi için de web madenciliğinin kullanımı zorunluluktur. Web madenciliği bir veri madenciliği tekniği olmakla beraber, bilgileri özellikle web kaynaklarından ayırmak, içerisinden çıkartmak ve web kaynaklarını yönetmek için algoritmalar kullanır (Kamakshive diğerleri, 2018).

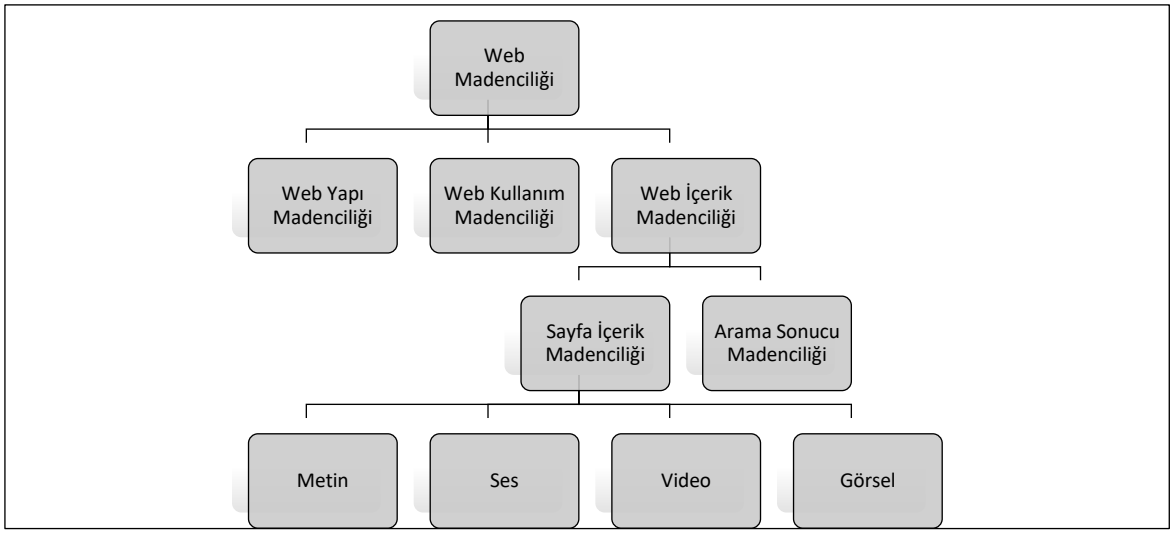
Web madenciliği ilk olarak 1996 yılında Oren Etzioni tarafından ortaya atılmıştır. Bu bildiride Etzioni'ye göre (1996) web madenciliği, VM tekniklerini kullanarak WWW'de bulunan dosya ve servislerden otomatik olarak bilginin ayıklanması, ortaya çıkartılması ve analiz edilmesidir. Web 'de bulunan verilerin sürekli olarak güncellenmesi, silinmesi ve yeni bilgilerin eklenmesi web'den bilgi çıkarım işleminde karşılaşılan en önemli zorluklardan birisidir.

4.1. Web Madenciliği Türleri

Web madenciliği, farklı amaçları gerçekleştirmek için farklı modeller sunmaktadır. Temel olarak web madenciliğini üç temel kategoride incelemek mümkündür.

- Web İçerik Madenciliği
- Web Yapı Madenciliği
- Web Kullanım Madenciliği

Bu farklı madencilik gruplarının her biri, webi başka açılardan ele almakta ve bu alanlarda bilgi keşfi yapılması için aracı olmaktadır. Web içerik madenciliği, metin, görsel, ses ve video gibi farklı veri tiplerini incelemek için kullanılan madencilik türüdür. Web yapı madenciliği ise link yapısının elde edilmesi ve kullanıcı aramaları için daha iyi sonuçların elde edilmesi süreçlerinde kullanılmaktadır. Web kullanım madenciliği ise web kayıtlarından kullanıcının sık kullandığı yapıları içeren kıymetli bilgilerin çıkarılmasında kullanılan metotlardır (Jayamalini and Ponnavaikko, 2017). Web madenciliği türlerini ve yapısı Şekil 4.1'de gösterilmektedir.



Şekil 4.1 Web madenciliği türleri (Jayamalini and Ponnavaikko, 2017)

4.1.1. Web içerik madenciliği

Web temel olarak birbirine bağlı içeriklerden oluşan dünya çapında bir ağdır. Bu ağın içerisinde çok farklı türden içerikler bulunmaktadır. Kullanıcılar web sitelerini içerikleri için ziyaret etmektedir. Web içerik madenciliği web sayfalarında bulunan, metin, görsel, ses ve

video gibi değerli bilgilerin keşfedilmesi için bir yöntemdir. Web içerik madenciliği web içeriğinde yer alan önemli bilgileri tespit eder (Jayamalini and Ponnavaikko, 2017).

Web 1.0'dan Web 3.0 a doğru giden yolda, artık web siteleri ilk dönem örnekleri gibi statik yapılar değil, içeriklerin daha çok kullanıcılar tarafından oluşturulabildiği karmaşık bir yapıya dönüşmüştür. Internet uzmanları geleceğin interneti hakkında farklı fikirlere sahiptirler. Önde gelen bilgi teknolojileri uzmanları Web 3.0 kavramını anlamsal web ve kişiselleştirme olarak tanımlamaktadır (Nath, Dhar, and Basishtha, 2014). Artık içerik kullanıcılar tarafından oluşturulmakta ve webe yüklenmektedir. Wordpress, Blogger gibi sistemler dünya üzerindeki her insana bir internet sayfası sahibi olma imkânı vermiş ve interneti sınırları aşacak bir yapı haline getirmiştir. Bugün sadece internet bağlantısına sahip bir bilgisayar ile bütün kullanıcılar kendi ilgisini çeken içeriği internette yayınlayabilmektedir.

İnternet içeriklerinin tüm dünyadan farklı dillerde oluştuğu düşünüldüğünde bu içeriği, tümünü okumadan anlamlandırabilmek teknolojilere ihtiyaç artmıştır. Bu web yapılarını daha iyi anlamak için günümüzde sitelerin kendi verilerini işaretleyerek dış dünyanın daha iyi anlamasını sağlayacak Schema.org gibi yapılar inşa edilmiştir. Bing, Google ve Yahoo bir araya gelerek Schema.org'u oluşturmuş ve geliştirmek ve anlatmak için anlaşmışlardır. Tüm sözlük bilgilerini tanımlamak tabi ki oldukça gerçekdışı bir durumdur. Shema.org genel geçer web üzerindeki ifadeleri tanımlamayı ve popüler alanları tanımlamaya aktarmaya çalışmaktadır (Ronallo, 2012). Örneğin aşağıda yer alan kod örneği yer alan sayfa kendini bir kişi bilgisi olarak etiketlemekte ve içerik analizcilerine veri sunmaktadır.

```
<script type="application/ld+json">{
        "@context": "http://schema.org",
        "@type": "Person",
        "name": "John Doe",
        "jobTitle": "Graduate research assistant",
        "affiliation": "University of Dreams",
        "additionalName": "Johnny",
        "url": "http://www.example.com",
        "address": {
                "@type": "PostalAddress",
                "streetAddress": "1234 Peach Drive",
                "addressLocality": "Wonderland",
                "addressRegion": "Georgia"
        }
}</script>
```

4.1.2. Web yapı madenciliği

Web yapısı ve çalışması yıllar içinde gelişmiş ve daha karmaşık web sitelerinin yapılarak daha fazla sorunun çözülmesi sağlanmıştır. İlk dönem web siteleri sadece HTML ile işaretlenmiş metinler iken, günümüz siteleri daha karışık yollar içeren uygulamalar haline dönüşmüşlerdir. Web yapı madenciliği asıl çalışma alanı bu farklı dokümanlar arası ilişkileri incelemek ve bu ilişkileri çözümlemektir. Web link analizi teknikleri temel olarak, web arama motorları ve sosyal network analizlerinde kullanılmaktadır (Jayamalini and Ponnavaikko, 2017).

Web siteleri birbirleri ile link adını verdiğimiz yapılar ile bağlanmıştır. Web yapı madenciliği bu yapılar ile ilgilenmektedir. Web yapı madenciliği ile birbiri ile alakalı sitelerin bulunması ve sitenin tüm web içerisinde nerede olduğunun konumlandırılması mümkündür. Ayrıca arama motorları da bu yapıları izleyerek çalışmakta, hangi sitelerin diğerlerinden daha iyi olduğunu, arama sonuçlarında hangisinin daha önde yer alması gerektiğini bu şekilde hesaplamaktadırlar.

4.1.3. Web kullanım madenciliği

Web siteleri boyutuna göre gün içerisinde çok çeşitli sayılarda ziyaretçi alabilmektedir. Ufak siteler 1-2 ziyaretçi alırken, sosyal medya platformları haber siteleri gibi siteler binlerle ölçülmektedir. Bu ziyaretçiler bazen tek ziyaret yapmakta, bazen gün içinde birçok sefer ziyaret etmektedir. Web sitelerin optimizasyonu, geliştirilmesi ya da kişileştirilmesi için bu bilgilerin kullanılması gerekmektedir. Web kullanım madenciliği, Web madenciliğin bir kolu olmakla beraber, web sunucuları, proxy sunucuları ve web istemcileri tarafından oluşturulan kayıt bilgilerinden ilginç bilgilerin çıkarılması ile uğraşır. Web kullanım madenciliği çeşitli kayıtlardan, faydalı verilerin çıkarılarak, çok kullanılan sayfaların, birlikte erişilen sayfaların ve bu sayfalara erişen kullanıcıların amaç be tiplerini bulmak için çalışır. Web kullanım madenciliği kullanıcı ziyaret şablonlarının otomatik olarak, bir ya da daha fazla sunucudan keşfine evrilmektedir. Web kullanım madenciliği web kayıtlarından başlayarak veri temizleme, entegrasyon, ön işleme, örüntü çıkarma ve analizi ile başlayıp, keşfedilen şablonların uygulanması ile son bulur(Azitha, Surjandari, and Laoh, 2018).

Web kullanım madenciliği çok çeşitli amaçlara hitap etmekte ve çeşitli uygulama alanları bulmaktadır. Örneğin sistem geliştirmeleri, adaptasyon, kişiselleştirme ve öneri, reklamcılık bu alanlardandır (Ganibardi and Ali, 2018).

Ayrıca ziyaret verileri, ticari sonuç oluşturan e-ticaret, pazaryeri gibi sitelerde doğrudan geliri etkileyebilecek bir faktördür. Bu sebeple bu verilerin incelenmesi ve bu verilerden fayda sağlayacak bilgilerin çıkarılması önem arz etmektedir. Bu çıkarılan bilgiler kullanıcı davranışlarını anlamayı kolaylaştıracaktır (Herath and Jayaratne, 2017).

4.2. Literatür Araştırması

Veri analizi, webden toplanan verilerin analizinde kullanılmıştır. Bu çalışmalar Web Madenciliği olarak 1996 yılında Etzioni tarafından isimlendirilmiştir. Bu alandaki çalışmalar Roberts Armstrong ve çalışma arkadaşlarının çabalarından günümüze yayılan araştırmalardan beslenmektedir(Rao and Kumari, 2014). Bu alanda birçok çalışma yapılmış ve süreçte farklı alanlar iyileştirilmeye çalışılmıştır. Bu alandaki çalışmaları bazı alt başlıklara bölmek mümkündür. Rao ve Kumari çalışmalarında bu başlıkları şu şekilde belirlemişlerdir(Rao and Kumari, 2014):

- Kişiselleştirme
- Sistem İyileştirme
- Kullanım karakterizasyonu
- Website güncelleştirme
- İş zekası

Bu alanda yapılan çalışmalardan, tezin mevcut konusunu içeren şu örnek çalışmaları göstermek mümkündür.

2000 yılında Martin Arlitt “Characterizing Web User Session” isimli çalışması ile ziyaretleri oturumlara ayırma konusunda farklı ölçümleri paylaşmıştır(Arlitt, 2000).

Behrouz ve William 2003 yılında yaptıkları “Using Genetic Algorithms for Data Mining Optimization in an Educational Web-Based System” isimli çalışmalarında öğrenci

başarılarını tahminlemek için web tabanlı bir sistemin kayıt dosyalarını kullanmıışlardır (Minaei-Bidgoli and Punch, 2003).

2004 yılında yayınlanan “Discovery of Web Robot Sessions based on their Navigational Patterns” adlı çalışmalarıyla Pang-Ning Tan ve , Vipin Kumar, robots.txt erişimi, head isteği sayısı ve belirlenmemiş yönlendiren içeren isteklerin sayısının web botlarının tespitinde önemli göstergeler olduğunu göstermişlerdir (Tan and Kumar, 2004).

Domenech ve Javier, 2007 yılında yaptıkları “A tool for web usage mining” sınıflandırma algoritmalarını kullanmış ve web kayıt dosyalarını daha iyi sınıflandıran bir uygulama önermişlerdir (Domenech and Lorenzo, 2007).

2007 yılında, G. Castellano, A. M. Fanelli, M. A. Torsello “Log Data Preparation For Mining Web Usage Patterns” isimli çalışmalarında LODAP (LOg DAta Preprocessor) isimli bir uygulama önermiş ve bu uygulama ile ziyaretleri oturumlara bölmenin analiz yeteneğini arttırdığını göstermişlerdir(Castellano, Fanelli, and Torsello, 2007).

Renata Ivancsy ve Sandor Juhasz “Analysis of Web User Identification Methods” isimli makaleleri ile 2007 yılında çerez tabanlı kullanıcı tanımlama metodu göstermişlerdir. Bu çalışmalarında, çerez takibi yapmak için ufak bir resim dosyasını web sayfalarına ekleyerek bir çözüm yoluna gitmişlerdir(Ivancsy, Juhasz, and Technology, 2007).

Chitraa ve Davamani, 2010 yılında "A Survey on Preprocessing Methods for Web Usage Data" çalışmalarıyla ön işleme metodları hakkında bilgi vermiş web sunucu loğlarının en iyi veri kaynağı olduğunu fakat çok fazla gereksiz bilgi olduğu için temizlenmesi gerektiğini belirtmişlerdir (Chitraa and Davamani, 2010).

2013 yılında Varnagar, Madhak, Kodinariya ve Rathod “Web usage mining: a review on process, methods and techniques” isimli çalışmalarında ön işleme işlemlerini derlemişlerdir (Varnagar, Madhak, Kodinariya, and Rathod, 2013).

Sukumar, Robert ve Yuvaraj, 2016 yılında yaptıkları “Review On Modern Data Preprocessing Techniques in Web Usage Mining (WUM)” çalışmalarında, literatürde olan ön işleme yöntemlerini aktararak bu çalışmalar için örnek pseudo kodlar sunmuşlarıdır (Sukumar, Robert, and Yuvaraj, 2016).

2017 yılında Hao, Zhaoxiang ve Bingbing "A User Clustering Algorithm on Web Usage Mining" isimli makalelerinde iki algoritmanın birleşimi ile çalışan yeni bir sınıflandırma algoritması önermişlerdir (Hao, Zhaoxiang, and Bingbing, 2017).

Kundu ve Garg, 2017 yılında yaptıkları "Web Log Analyzer Tools: A Comparative Study to Analyze User Behavior" isimli çalışmalarında kullanılmakta olan web kayıt analizi uygulamalarını derlemiş ve bu uygulamalar arasında karşılaştırma yapmışlardır (Kundu and Garg, 2017). İlgili çalışmada, bu tez içerisinde kullanılan Google Analytics ürününden de bilgi verilmiştir.

Sathya ve Devi, 2017 yılında "Apriori Algorithm on Web Logs for Mining Frequent Link" isimli çalışmaları ile Apriori algoritması hakkında güzel bir örnek sunmuşlardır (Sathya and Devi, 2017).

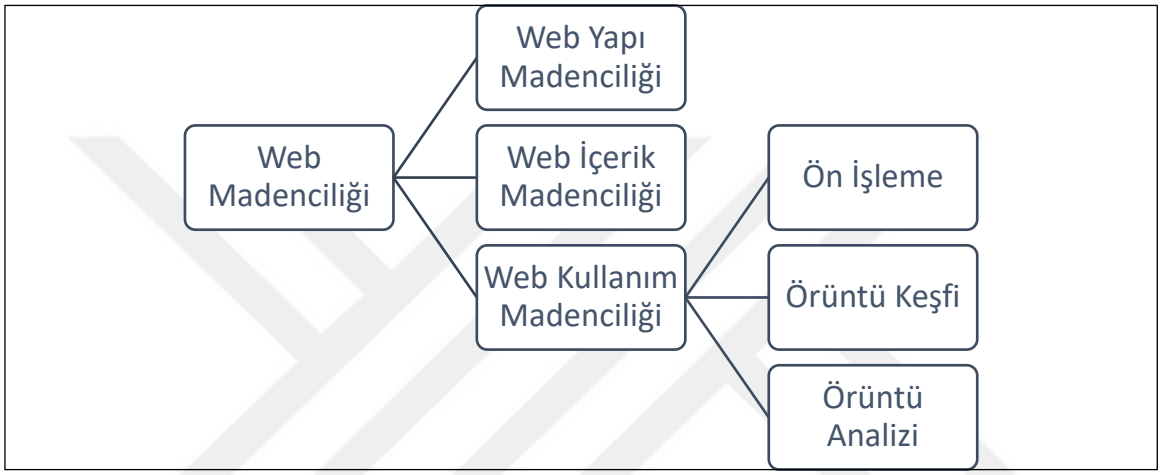
Deshpande, Deshpande ve Thakare 2018 yılında yaptıkları "Web User Identification: Analysis of Heuristic Solutions" çalışmalarıyla kullanıcı tanımlama alanında yapılan çalışmaları özetlemişlerdir (Deshpande, Deshpande, and Thakare, 2018).

Mevcut literatür tarandığında, yapılan çalışmalarda genel olarak daha verimli ön işleme metotları geliştirilmeye çalışılarak önerilerde bulunulmuş ve mevcut algoritmalardan kaynaklı modelleme sıkıntıları üzerinde çalışılmıştır. Bu çalışmalarda genel olarak kullanılan veri kaynağı ise web sunucu erişim kayıtlarıdır.

4.3. Web Kullanım Madenciliği Süreci

Web kullanım madenciliği, veri madenciliğinin bir kolu olan web madenciliğinin bir dalıdır. Web kullanım madenciliğinde yer alan veri madenciliği işlemi, web kullanım kayıtlarından bilgi çıkarılması sürecidir (Singh, 2017). Web siteleri kullanıcılar tarafından kullanıldıklarında ve ziyaret edildiklerinde bu süreçler için kayıtlar oluşturmaktadır. Bu kayıtların doğru araçlarla analizi web kullanım madenciliğinin temelini oluşturmaktadır. Web kullanım madenciliği bazı durumlarda web kayıt dosyası madenciliği olarak anılmaktadır (Kumar, 2017). Fakat bu kısmen yanlış bir kullanımdır. Web kullanım madenciliğinde tek veri kaynağı olarak web server kayıtlarının kullanılması artık eskimekte olan ve güvenilirliği tartışılır bir yöntemdir.

Web kullanım madenciliği süreci diğer madencilik süreçleri ile benzerlik göstermektedir. Web kullanım madenciliği, ön işleme, örüntü keşfi ve örüntü analizi adlarında üç alt sürece sahiptir (Kumar, 2017). Bu süreçlerin öncesinde ise asıl çıktı analizlerin kalitesini doğrudan etkileyen bir süreç bulunmaktadır. Veri toplama süreci birçok çalışmada doğrudan web sunucu erişim kayıtları kullanıldığı için önemsenmemektedir. Fakat bu sürecin optimizasyonu diğer süreçler kadar önem arz eden bir süreçtir. Bu süreçlerin bir gösterimi Şekil 4.2 de gösterilmektedir.



Şekil 4.2. Web kullanım madenciliği hiyerarşik yapısı (Sathya and Devi, 2017)

Web kullanım madenciliğinde de verilerin kaynak sistemlerden toplanması ve temizlenerek işlemeye uygun hale getirilmesi gerekmektedir. Web kullanım madenciliği süreçlerinde, genel olarak anonim veriler kullanılmaktadır. Toplanan verilerin web kullanım madenciliği süreçlerinde kullanılan özel ön işleme süreçlerinden geçirilmesi gerekmektedir. Web kullanım madenciliğinde ön işleme önde gelen bir süreçtir. Web erişim kayıtları, ham veridir ve bilgi çıkarımı için doğrudan kullanılması mümkün değildir. Bu sebeple ham verilerin anlaşılabilir verilere dönüştürülmesi sürecinde çeşitli tekniklerin kullanılması gerekmektedir (Kumar, 2017).

Düzenlenmiş ve işlenmiş veriler için uygulanabilecek birçok örüntü keşif aracı bulunmaktadır. Örüntü keşfi, sık kullanılan yol keşfi (frequent pattern discovery) gibi veri madenciliği tekniklerinin, ilginç örüntülerin keşfedilmesi için web sunucu kayıtlarına uygulanması sürecidir (Jayamalini and Ponnavaikko, 2017). Bu süreç üç farklı ana grubu içermektedir.

- Eşleşme: birlikte kullanılan sayfaların arasında yer alan ilişkileri tanımlar.

- Kümeleme: kullanıcı, sayfa ve süreçlerin gruplamasının keşfeder.
- Ardışıklık Analizi: erişilen sayfaların sıralaması ile ilgili örüntüleri keşfeder.

Örüntü analizi aşaması ise en son gelen aşamadır. Bu aşamada önceki süreçlerde elde edilen bilgilerin kullanımı ve bu bilgiler ile web sitelerin geliştirilmesi, önlemelerin alınması ve süreçlerin oluşturulması amaçlarına yer verilir.

Web kullanım madenciliği süreçlerini temel olarak dört ana aşamaya bölmek mümkündür. Bu aşamaların şu şekilde göstermek mümkündür.

- Veri toplama
- Ön işleme
- Örüntü keşfi
- Örüntü analizi

Bu dört aşamanın her birinde farklı süreçler işlemekte ve asıl amaç için farklı çıktılar üretilmektedir. Literatürde yer alan çalışmalar incelendiğinde bu süreçte en çok düzenleme ve çalışma gerektiren sürecin ön işleme olduğu görülmektedir. Eğer çizimle göstermek gerekirse bu süreç zaman ve emek açısından Şekil 4.3' de gösterildiği şekilde bir dağılım elde edilmektedir.

Şekil 4.3. Geleneksel yöntemlerle yapılan çalışmalarda süre ve emek kullanımı

Fakat web kullanım madenciliğinde asıl katma değer ve bilgi sağlayan ve web sitesinin amaçlarında fayda sağlayan süreç örüntü analizi ve bu örüntülerin web sitesi içerisinde değerlendirilmesi olmalıdır. Fakat bir çalışmada olması gereken emek ve süre dağılımının Şekil 4.4. de yer aldığı gibi hedeflenmesi gereklidir.

Şekil 4.4. Hedeflenmesi gereken çalışmalarda süre ve emek kullanımı

4.3.1. Veri toplama

HTTP ile aktarılan web istek ve cevapları durumsuz bir protokolden geçmektedir. Web sunucusu iki istek arasında doğrudan ilgiyi bilmemekte sadece gelen isteğe uygun cevabı vermektedir.

Web kullanım madenciliği kullanıcı ziyaretlerini inceleyerek bu ziyaretlerden bilgi keşfi yapmak üzerine çalıştığı için öncelikle farklı kullanıcı verilerinin toplanması gerekmektedir. Web yapısı gereği, web sitesi kullanımları oldukça anonim süreçlerdir. Bir web sitesi ziyaret edildiğinde, web sitesi kişi hakkında oldukça kısıtlı bilgiye erişmektedir. Temel olarak bir web ziyaretinde, yapılan isteği tanımlayan bilgiler, tarayıcı bilgileri ve ziyaretçinin ip bilgisi sunucuya iletilmektedir. Bu durumu geliştirmek için cookie adı verilen çerez dosyaları kullanılmaya başlanmış böylece web sitelerinin kullanıcıları takip edebileceği bazı izler oluşturulmaya başlanmıştır.

Web kullanım madenciliğinde kullanılabilecek veri kaynakları Jayamalini ve Ponnavaikko ve kendi çalışmalarında 3 çeşitle göstermişlerdir (Jayamalini and Ponnavaikko, 2017). Bunlar şu şekildedir:

- Web Sunucu Kayıt Dosyaları
- Uygulama Sunucusu Kayıtları
- Uygulama Seviyesi Verileri

Fakat web gibi esnek bir ortamda veri kaynaklarını sınırlamak doğru bir karar değildir. Web birçok bilginin ve verinin birbirleri ile bağlantılı olduğu bir sistemdir. Bu sayede farklı veri kaynaklarının kullanılması da mümkündür. Bu kaynaklara ek olarak aşağıda yer alan kaynaklarda örnek gösterilebilir.

- İstemci Verisi: İstemci verileri, web sitesini ziyaret eden tarafından kaydedilen kullanım verileridir. Java veya Javascript ile verileri kaydeden ajanlar yardımıyla uzakta yer alan bu verilere erişilmektedir. Bu ajanlar kullanıcı ziyaret verisini toplamaktadır. Bu yöntem server tarafından toplanan ziyaret verisinden çok daha güvenilirdir ve ip ataması ve web önbelleği gibi sorunlara karşı dirençlidir. (Jain, Rawat, and Bhandari, 2017)
- Aracı Verisi: Aracı verisi iki farklı kaynaktan toplanabilmektedir. Bunlar proxy sunucuları ve paket koklaycılardır. Proxy sunucusu web sitesine kullanıcı adına bağlanarak bu sunucudan verileri alarak gönderen bir ajan uygulamadır. proxy sunucusu üzerinde yer alan erişim kayıtları web sitesine yapılan istekleri ve bu isteklerin cevaplarını barındırırlar (Jainve diğerleri, 2017). Paket koklayıcılar ise ağ üzerinde akan trafik akışını inceleyen yazılım ya da donanım ürünleridir. Bu araçlar TCP/IP paketlerini inceleyerek web kullanım madenciliği süreçlerinde veri kaynağı olmaktadır (Jainve diğerleri, 2017).
- Üçüncü Parti Araçlar: Bu araçlar web sitelerinde yer alan ziyaret verilerin ne kullanıcı bilgisayarlarında ne de sunucu ortamında saklandığı araçlardır. Bu araçlar kullanım verilerini kendilerine ileten çeşitli yöntemler kullanarak web kullanım verilerini kayıt altına almaktadırlar.

Yapılan çalışmalarda bu veri kaynakları kullanılarak çalışmalar için veri toplanmaktadır. Fakat bu veri kaynaklarından sadece birinin seçilmesine gerek yoktur. Farklı kaynaklardan toplanan verilerin bir araya getirilmesi ile daha anlamlı sonuçların üretilmesi mümkündür. Birçok durumda bu veri kaynaklarının sadece tek bir tanesinin kullanılması yetersiz bir veri ile çalışılması demektir. Bu yüzden farklı verilerin bir araya getirilmesi gerekmektedir. Çerezler, kullanıcılar hakkında, oturum kimlik değeri de olmak üzere faydalı bilgiler sağlamaktadır. Bu bilgiler kullanıcı web sitesini ne zaman ziyaret etse web sunucusuna gönderilmektedir. Bu verilerin çalışmalarda kullanılmak üzere kayıtlarda saklanması faydalı olacaktır (Kapusta, Munk, and Halvoník, 2017). Kapusta, Munk ve Halnovik yaptıkları çalışmada çerez kullanımı ile oturum süre metodunu karşılaştırmışlardır. Yaptıkları çalışma sonucunda şu sonuca erişmişlerdir: Yapılan deneyler oturum süre aralığı ve çerez kullanımı arasında, çıkarılan kuralların kalitesini etkileyecek önemli bir değişim göstermemektedir (Kapustave diğerleri, 2017). Çerez içerisinde gelen hazır verileri kullanmak tekrar hesaplama yapmaktan daha kolay bir yöntemdir.

Yapılan incelemelerde görüldüğü üzere web kullanım madenciliği için genel olarak web erişim kayıt dosyaları kullanılmaktadır. Bu durumu göstermek için IEEE Xplore dijital kütüphanesi üzerinde yer alan 2017 ve 2018 yılı "web usage mining" anahtar kelimeli çalışmalar incelenmiş ve bu konuda toplanan verilerin çıktısı Çizelge 4.1 de gösterilmiştir.

Çizelge 4.1. 2017 ve 2018 yılı çalışmaları veri kaynağı tercihi

	Web Sunucu Kayıtları	Web sunucu Kayıtları + Çerez	Proxy Kayıtları	E-Devlet Site Kayıtları	Biyometrik Sensor	Servis Sağlayıcı	Toplam
2017	17	1	1	1	1		21
2018	6					1	7

Kayıt dosyaları farklı sunucu tiplerinde değişiklik gösterse de genel kabul edilen standart W3C tarafından belirlenen "*the common logfile"* formatıdır. Bu format şu bilgileri içermektedir (W3C, 1995).

remotehost rfc931 authuser [date] "request" status bytes

- Remotehost: Uzak bilgisayarın(ziyaretçi) ip adresi
- Rfc931: Uzak bilgisayarın oturum kullanıcısı
- Authuser: HTTP oturum uygulamaları için oturum açan kullanıcının adı
- [date]: Ziyaret tarih ve saati
- "Request": Ziyaretçiden gelen istek
- Status: Üretilen cevabın durum kodu
- Bytes: İstek için üretilen cevabın bayt uzunluğu

Bu bilgiler her bir istek için erişim kayıtlarına eklenmektedir. Fakat bu veriler yeterli olmadığı için birçok sunucu "combined" ya da "*extented*" adı verilen birleşik formatı tercih etmektedir. Bu format common log format'a benzemekle birlikte tarayıcı bilgileri, yönlendiren bağlantı, istemci işletim sistemi gibi bilgileri de içeren ek alanlar bulundurmaktadır (Jainve diğerleri, 2017).

- “Referer": Ziyaretçiyi web sayfasına yönlendiren diğer web sayfası
- "User-agent": Kullanıcın ziyaret ettiği tarayıcı bilgisini gösteren metin

Bu veriler her ne kadar kullanıcıları tanımlıyor gibi gözükse de tam olarak yeterli olamamaktadır.

- RFC931: Uzak sunucu adı, uzak bilgisayar oturum adı aslında oldukça güvensiz bir bilgidir. Bu bilgi Apache web server tarafından özellikle ayarlanmadıkça belirlenmemektedir. Çoğu zaman “-” değerini göstermektedir.
- Authuser: Kullanıcı değeri HTTP protokolü ile oturum açma işlemlerinde güvenilir bir veridir. Fakat bu protokol, yeterli bir güvenlik katmanı sağlamaması ve kaba kuvvet saldırılarına açık olması sebebiyle genel olarak kullanılmamakta bunun yerine çerez ya da token tabanlı güvenlik mekanizmaları kullanılmaktadır. Bu bilgi yerine çoğu zaman “-” değeri gelmektedir.
- Referer: Referer bilgisi gönderilmesi zorunlu olmayıp, tarayıcı sistemlerin isteği üzerine gönderilmektedir.
- User-agent: Tarayıcı bilgisi gönderilmesi zorunlu olmayan tarayıcı tarafından gönderilen bir bilgidir. İsteğe göre bu bilgi değiştirilerek gönderilebilmektedir.

4.3.2. Ön işleme

Mevcut web kullanım madenciliği çalışmalarında kullanılan web sunucu kayıtları düşük oranda yapılandırılmış, metin tabanlı dosyalardır. Madencilik uygulamalarında, madencilik uygulamaları, eğer kaynak veri hata ve eksiklikler içeriyorsa faydasız sonuçlar üretecektir (Saste, Bedekar, and Kosamkar, 2017). Bu nedenle web gibi anonim bir ortamda faydalı çıktılar üretecek kaynak veriye iyi bir şekilde ulaşmak ve verinin kalitesinin yüksek olmasını sağlamak gerekmektedir. Bu metin dosyaların anlamlı verilere dönüştürülmesi için çeşitli ön işleme metotlarından geçmeleri gerekmektedir. Ön işleme süreci zaman alıcı ve grift bir süreçtir (Jainve diğerleri, 2017).

Bu ön işleme metotlarına şu örnekler verilebilir.

- Bot kayıtlarının temizlenmesi
- Anlamsız kayıtların temizlenmesi
- Kullanıcı Tanımlama
- Oturum Tanımlama
- Akış Tamamlama

Bu yöntemlerin bir veya birden fazlası bir arada kullanılarak verilerin kalitesinin arttırılması sağlanmaktadır. IEEE Xplore üzerinde 2017 yılı için yapılan çalışma taramasında bu ön işleme süreçlerinin analizi yapılmış ve sonuçlar Çizelge 4.2'de verilmiştir.

Çizelge 4.2. Ön işleme yöntemlerinin kullanımı dağılımı

Veri Temizleme	Kullanıcı Tanımlama	Oturum Tanımlama	Akış Tamamlama
%50	%41,6	%37,5	%12,5

Bu yöntemlerin kullanımları ve bu kullanımlar sırasında karşılaşılan sorunlar aşağıda yer alan bölümde verilmiştir.

Veri temizleme

Veri madenciliği süreçlerinde, işlemlerde hataya sebep olabilecek, eksik ve hatalı verilerin, modelleme süreçlerinden önce veri setlerinden çıkarılması gerekmektedir. Bu durum web kullanım madenciliği içinde geçerli bir süreçtir. Faydasız verilerin, web kayıtlarından elenmesi ön işleme sürecinin ilk işidir(Jainve diğerleri, 2017) .

Web siteleri her türlü kullanıcı ve diğer erişimlere açık yayınlanmaktadır. Bu nedenle web sitelerinde de istenmeyen ve alakasız ziyaretler gerçekleşebilmekte ve aslında analiz konusu ile alakalı olmayan birçok kayıt erişim kayıtlarına yazılmaktadır. Bu hatalı kayıtların verilerinin temizlenmesi gerekmektedir. Temel olarak 3 farklı temizleme gereksinimi olduğu söylenebilir:

- Bot Kaynaklı Kayıtlar
- İlgisiz Kayıtlar
- Hatalı İşlemler

Bot kaynaklı kayıtlar

Web üzerinde yer alan sitelerin tümünü incelemek ve bilmek mümkün olmadığı için yıllar içerisinde arama motorları ve indeksleme yazılımları oluşturulmuştur. Popüler arama motorları web sayfalarını verilen linkler ile gezmekte ve bu sitelerden tanımlama verilerini çıkartarak, kendi kullanıcılarının aramalarında bu bilgilere göre sonuç vermektedir. Ayrıca web üzerinde yer alan çeşitli iyi ya da kötü niyetli yazılımlar web sitelerini gezerek bu sitelerden telefon numarası, e-posta adresi gibi verileri çıkartmaktadırlar. Web kullanım madenciliği kullanıcı erişim örüntüleri üzerine madencilik yapmak için kullanıldıkları için insan kaynaklı olmayan tüm etkilerin web erişim kayıtlarından çıkarılması gerekmektedir (Sukumarve diğerleri, 2016).

Bu araçlar normal ziyaretçilerden farklı olarak erişebildikleri tüm içeriği gezmek ve işlem yapmaya programlanmışlardır. Bu yüzden bir web sitesini ziyaret ettiklerinde, ziyaretçi gibi aradıkları bilgiye odaklanmak yerine sitedeki tüm web sayfalarını gezmektedirler. Bu nedenle bot trafikleri web kullanım madenciliği analizlerinde istenmeyen veriler oluşturmaktadır. Günümüzde oldukça yüksek sayıda çeşitli botlar web sitelerini ziyaret etmektedir. Bu botlardan en bilinen 10 tanesi Çizelge 4.3’de listelenmektedir.

Çizelge 4.3. En çok karşılaşılan 10 web botu (Jackson, 2019)

Bot	Agent Adı	Agent İmza Değeri
GoogleBot	Googlebot	Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)
Bingbot	Bingbot	Mozilla/5.0 (compatible; Bingbot/2.0; +http://www.bing.com/bingbot.htm)
Yahoo	Slurp	Mozilla/5.0 (compatible; Yahoo! Slurp; http://help.yahoo.com/help/us/ysearch/slurp)
DuckDuckGo	DucDuckBot	DuckDuckBot/1.0; (+http://duckduckgo.com/duckduckbot.html)
BaiduSpider	Baudispider	Mozilla/5.0 (compatible; Baiduspider/2.0; +http://www.baidu.com/search/spider.html)
Yandex	Yandexbot	Mozilla/5.0 (compatible; YandexBot/3.0; +http://yandex.com/bots)
Sogou Spider	Sogou Spider	Sogou Pic Spider/3.0(http://www.sogou.com/docs/help/webmasters.htm#07)
Exabot	Exabot	Mozilla/5.0 (compatible; Konqueror/3.5; Linux) KHTML/3.5.5 (like Gecko) (Exabot-Thumbnails)
Facebook	facebot	facebookexternalhit/1.0 (+http://www.facebook.com/externalhit_uatext.php)
Alexa Crawler	ia_archiver	ia_archiver (+http://www.alexa.com/site/help/webmasters; crawler@alexa.com)

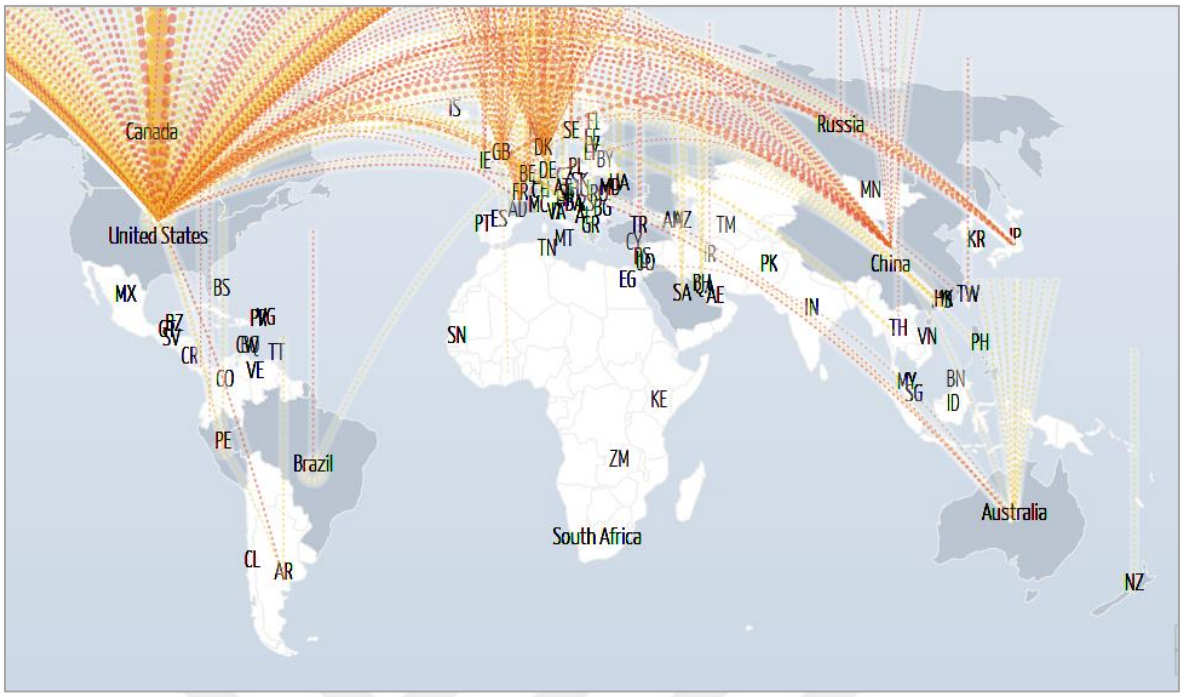
Bot isteklerinden temizleme için User-Agent ifadeleri kullanılabileceği gibi, bu botların ip adreslerininde kullanılması mümkündür. Sukumar, Robert ve Yuvaraj kendi makalelerinde aşağıda yer alan algoritmanın kullanarak verilerini botlardan temizlemişlerdir (Sukumarve diğerleri, 2016).

Adım 1: Kısmen temizlenmiş kayıt dosyasından her bir kaydı incele
Adım 2: Eğer
(kayıt ip adresi, Robot IP adresi olarak belirlenen aşağıdaki aralıklarda yer alıyorsa)
(
66.231.160.0 dan 66.231.191.255 e
66.102.0.0 dan 66.231.15.255 e
66.249.64.0 dan 66.249.95.255 e
72.14.192.0 dan 72.14.255.255 e
74.125.0.0 dan 74.125.255.255 e
209.285.128.0 dan 209.85.255.255 e
216.239.32.0 dan 216.239.63.255
)
Ya da
(ilgili ip adresi robot.txt dosyasına istek yapmışsa)
Ya da
(User Agent değeri listede bulunan ifadeler ile eşleşiyorsa
(
..www.googlebot.comlbot.htrnl ..
)
)
İse
Bu kaydı kayıt dosyasından sil

Adım 3: Adım 1 ve 2'yi dosyanın sonuna kadar uygula
Adım 4: Süreci bitir

Bu alanda yaşanabilecek bir başka problem ise ilgili web sitesinin bir devre dışı bırakma saldırısından etkilenebiliyor olması durumudur. Bilişim alanında en çok görülen siber saldırıların başında DoS (Denial of Service) ve DDoS (Distributed Denial of Service) saldırıları gelmektedir. DoS saldırısı, kısaca, belli bir sunucunun belli bir şekilde hizmet bekleyen kullanıcılara hizmet verememesini sağlamak amacıyla, o bilgisayarın işlem yapmasını engellemek, bir başka deyişle hedef bilgisayarı bilişim sisteminin içerisine girmeksizin kilitlemektir (Özocak, 2012). Kötü niyetli kişilerce web sitesini sabote etmek için çok sayıda istek gönderilmesi de mümkündür. Bu sebeple bu verilerin temizlenmesi gerekmektedir.

Digital Attack Map isimli Google tarafından geliştirilen servis dünya üzerinde gerçekleşen DDOS saldırılarını kayıt altına alarak raporlamaktadır. Örneğin 7 Eylül 2014 tarihinde Amerika ve Çin arasında gerçekleşen ddos saldırılarını gösteren harita Harita 4.1'de yer almaktadır.



Harita 4.1. 7 Eylül 2014 ddos saldırısı

Bu haritada yer alan kırmızı çizgiler web sitelerine yapılan saldırıları işaret etmektedir. Bu büyük saldırılar dikkate alındığında web erişim kayıtlarının ddos saldırılarından temizlenmesi doğru olacaktır.

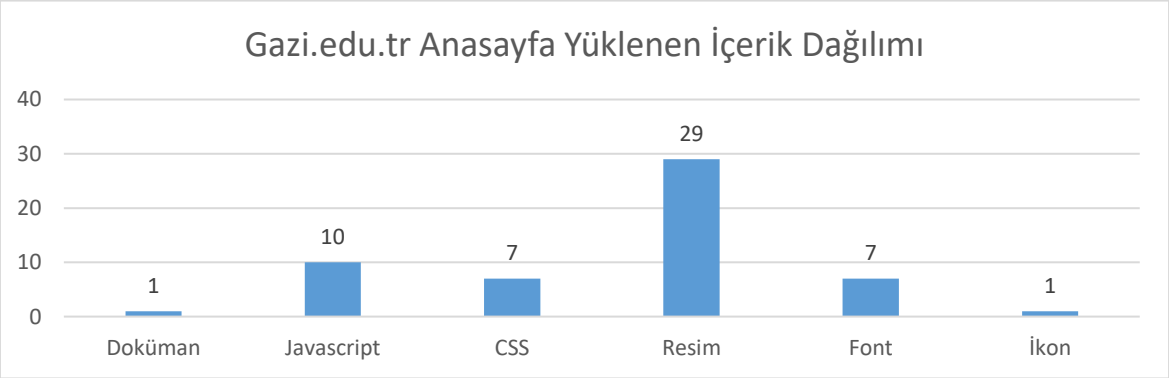
İlgisiz kayıtlar

İlk geliştirilen web siteleri birbiri ile alakalı dokümanlar şeklinde yayınlanmaktaydı. Günümüzde kullandığımız web siteleri çok daha karmaşık yapılar içermektedir. http protokolü web sunucusundan elde edilecek her dosya için ayrı bağlantılar oluşturulmasını gerektirmektedir. Bu durumun bir sonucu olarak, grafik ve script dosyalarının ek olarak indirilmesi, bir sayfanın görüntülenmesi sonucunda genellikle birçok erişim kaydı oluşturmasını gerektirir(Cooley, Mobasher, and Srivastava, 1999).

Mevcut geldiğimiz teknoloji düzeyinde web siteleri çok sayıda görsel öğeyi içerisinde barındırmaktadır. Bu içerikleri şu şekilde listelemek mümkündür.

- Stil dosyaları
- Script dosyaları
- Çeşitli formatlarda görsel dosyaları
- Video ve ses kayıtları
- Java appletler
- Flash/Silverlight gibi çeşitli medya araçları
- Fontlar
- Tanımlama dosyaları

Bu kaynaklar çoğu zaman web sunucuları tarafından sunulmakta ve bu işlemlerin kayıtları da erişim kayıtlarında yer almaktadır. Bu kayıtlar çoğu zaman asıl web içeriğinden daha fazla sayıda veri oluşturmaktadır. Örneğin Gazi Üniversitesi web sitesi için 1 adet istek yapıldığında, tarayıcı yazılımı toplam 55 adet istek göndermektedir. Bu isteğin dağılımı Şekil 4.5. ‘de verilmiştir.



Şekil 4.5. Gazi.edu.tr ana sayfa yapılan istek sayıları

Bu grafikten görüleceği içere oluşturulan içeriklerden %45’i resimlerden oluşmaktadır. Yapılan istekler sadece %1,8 oranında asıl isteği oluşturmakta gerisi yardımcı istekler olarak erişim kayıtlarına eklenmektedir. Bu istekler modern sitelerde CDN denen yapılara taşınmaya başlanmış olsa da bu kullanım henüz genel bir çerçeveye yayılmamıştır.

Bu alanda yaşanabilecek bir diğer durum ise asıl içeriğe bir faydası olmayıp, web sitesi süreçlerine yardımcı olmak için oluşturulmuş bağlantılardır. Örneğin birçok site web

sayfalarında yer alan verileri anlık olarak güncellemek için AJAX teknolojisinden faydalanmaktadır. AJAX teknolojisi mevcut sayfayı tekrar yüklemeden, sayfada yer alan bir bilgiyi güncellemek için kullanılmaktadır. Bu istekler düzenli aralıklarla yapılabildiği gibi kullanıcı isteğiyle de gönderilebilmektedir. Kullanıcı tarayıcısında açık kalan bir sayfanın yüksek sayıda isteği göndererek erişim kayıtlarına isteğe bağlı ziyaretler gibi veri kaydetmesi oldukça beklenebilir bir durumdur. Bu veriler de kontrol edilerek temizlenmelidir.

Hatalı istekler

Web sayfaları yayınlanırken, oluşturulan sayfalarda hatalar oluşabilmektedir. Örneğin veri tabanı sunucusunda yaşanabilecek kesintiler, servis entegrasyonu hataları, yetkisiz erişim ya da bulunamama durumlarında web sunucu yazılımları isteğin sonucunu bildiren çıktılar üreterek durum kodu hanesinde farklı içerikler göndermektedir. Bu kodlara ait liste Çizelge 4.4. ‘de verilmiştir.

Çizelge 4.4. HTTP protokolü durum kodları ve grupları

Kod Grubu	Anlamı
100-101-102-103	Bilgi Verilmesi
200-201-202-203-204-205-206-207-208-226	Başarılı cevap kodları
300-301-302-303-304-305-307-308	Yönlendirme kodları
400-401-02-403-404-405-406-407-408-409-410-411-412-413-414-415-416-417-418-421-422-423-424-425-426-428-429-431-451	İstemci hata kodları
500-501-502-503-504-505-506-507-508-510-511	Sunucu hata kodları

Hata kodları tablosundan da görülebileceği üzere oluşabilecek birçok hata durumu bulunmaktadır. Bu isteklerin de erişim kayıtlarından çıkarılması ve temizlenmesi gerekmektedir.

Kullanıcı tanımlama

Web siteleri özellikle kişisel bilgilerin girilerek oturum belirlenmeyen web sitelerde kullanıcıları tanıması imkansızdır. Web kullanım madenciliği çalışmalarında bu kullanıcı

işlemlerini tanımlayıp ayrıştıracak yöntemler geliştirilmektedir. Bu yöntemler ile farklı kullanıcıların istekleri gruplanarak, bu isteklere göre analiz çalışmaları yapılmaktadır.

Kullanıcıların belirlenmesi, kullanıcıların, giriş bilgilerini kullanarak oturum açmadıkları sistemlerde oldukça zor bir süreçtir. Gelen kullanıcıların belirlenebilmesi ve takip edilebilmesi için çerez dosyalarından faydalanılmaktadır. Mevcut çalışmalara bakıldığında bu süreçte çoğu zaman istek gönderen bilgisayarın ip adresi ve tarayıcı bilgisinden yararlanıldığı görünmektedir. Çizelge 4.5'de örnek olarak farklı çalışmalarda kullanılan kullanıcı tanımlama yöntemlerine örnekler verilmiştir.

Çizelge 4.5. Farklı çalışmalarda kullanıcı tanımlama yöntemlerine örnekler

Çalışma	Kullanılan Yöntem
Apriori Algorithm on Web Logs for Mining Frequent Link (Sathya and Devi, 2017)	Her ip yeni bir kullanıcı, İp aynı fakat zaman sapmalı ise yeni bir kullanıcı
Predicting Demographic Attributes From Web Usage: Purpose and Methodologies (Sasteve diğerleri, 2017)	Ip adresi, tarayıcı bilgisi ve yönlendirme kaynağı
Web Usage Mining Tool By Integrating Sequential Pattern Mining with Graph Theory (Musale and Chaudhari, 2017)	Her ip yeni bir kullanıcı
Mining Information Search Pattern on Website: A Case Study of Educational Institution (Azithave diğerleri, 2018)	Her ip yeni bir kullanıcı

Bu çalışmalarda genel olarak ip adresi ve tarayıcı bilgisi birleştirilerek oluşturulan ifadelerin tek bir kişiyi işaret ettiği varsayılmaktadır. Fakat ip adreslerinin ve tarayıcı bilgilerinin kullanımı da kullanıcı oturumlarının takip edilmesinde oldukça eksik bir yöntemdir. Bu kullanım sırasında yaşanabilecek hata kaynaklarını şu şekildedir:

- Mobil Cihazlar: Günümüzde gelişen mobil cihazlar ile kullanıcılar aynı web sitesine çok farklı ip adreslerinden ve farklı internet tarayıcılarından bağlanabilmektedir. Kullanıcılar evlerindeki ya da iş yerlerindeki kablosuz modemlerden, mobil hareket ettikleri süreçlerde ise operatörlerinden aldıkları ip adresleri ile web sitesini ziyaret

etmekte gün içerisinde birden çok ip ile web erişim kayıtlarında yer alabilmektedirler.

- Önbellek: Önbellekler veri iletişimini hızlandırmak ve kullanıcılara daha hızlı erişim sağlatmak için kullanılan araçlardır. Bu mekanizmalarda, tüm web sayfası istekleri web erişim kayıtlarında yer almamaktadır. Bir web isteği oluşturulduğunda işlemci ilk önce içeriği, önbellekte aramaktadır. Eğer bu istek önbellek içerisinde var ise server üzerinden tekrar veri getirilmemekte hazır olan istek kullanılmaktadır (Deshpandeve diğerleri, 2018). Bu hataları gidermek için akış tamamlama yöntemleri kullanılmaktadır.
- Dağıtık Sistemler: Gelişen teknoloji ile web yazılımları gelişmiş ve daha karmaşık sistemler kurmak mümkün olmuştur. Ayrıca web 2.0 gelişimi ile sosyal medya siteleri gibi yazılımların artık tek bir sunucudan yayınlanması imkânsız hale gelmiştir. Geniş ölçekli web sitelerinde, içeriğin birden çok web ve uygulama sunucusundan gelmesi artık genel bir durum olmuştur. Bazı durumlarda ise birden çok sunucunun artan kaynakları tek sunucunun üzerine düşen yükü azaltmak için kullanılmaktadır (Liu, 2011). Bu durumda kullanıcıların çok farklı ortamlardan toplanması ve eşleştirilmesi gerekliliği doğmaktadır.
- Dinamik Ip Adresi: Ip adresleri tek başına genel olarak kullanıcıları kayıtların her birine atamak için yeterli olamamaktadır. Bu sürecin temel nedeni servis sağlayıcıların kullanıcılara zaman içerisinde değişen ip adresleri atamaları ve bu adreslerin kullanıcı internette gezdikçe değişebilmesidir (Liu, 2011). Bu durum belki küçük sitelerde bir sıkıntı oluşturmasa da, geniş ölçekli sitelerde farklı kullanıcıların aynı ip adresi ile ziyareti durumunda yanlış bir ilişkilendirmeye sebep olacaktır.
- Proxy Sunucuları: Proxy sunucuları vekil olarak kullanıcı adına ilgili web sitesine bağlanarak bilgileri alarak kullanıcıya geri ileten yazılımlardır. Proxy sunucuları bütün istekleri web sunucusuna göndermeyip öncelikle proxy önbelleğinde aramaktadır. Bu durumda proxy sunucuları isteği sunucuya göndermek yerine hazır kaynakları hızlıca kullanıcıya sunmaktadır (Deshpandeve diğerleri, 2018). Proxy sunucularının bir diğer özelliği ise tüm gelen istekler için tek bir bağlantı sağlamasıdır. Bu durumda birçok kullanıcı sadece tek bir ip adresinden sunucuya erişiyor görünmektedir.
- Tek Tip İşletim Sistemi ve Tarayıcılar: Günümüzde bir işletme, kurum ya da okul benzeri bir alanda, yüzlerce bilgisayar bulunabilmektedir. Artık gelişmiş güncelleme

ve sistem yönetim yazılımları sayesinde, aynı networke sahip bütün cihazların aynı versiyonda işletim sistemi ve tarayıcı kullanması sistem yöneticileri tarafından sağlanabilmektedir. Bu durumda bu bilgisayarlar aynı ip çıkışını kullanabilecekleri gibi, tanımlama süreçlerinde kullanılan tarayıcı isim bilgisini de paylaşabilmektedirler. Böyle bir durumda bir üniversitenin tüm bilgisayarlarının aynı tanımlama imzasına sahip olması oldukça mümkündür. Bu durum aynı şekilde mobil cihazlar içinde geçerlidir. Aynı marka ve model cihazların aynı tanımlama bilgisiyle erişim sağladıklarını söylemek yanlış olmayacaktır.

Bütün bu hata kaynakları incelendiğinde web kayıtları içerisinde kullanıcı tanımlama yöntemleri oldukça hataya açık bir süreçtir.

Web ziyaretlerinde kullanıcıların web sitesine ziyaretleri belirli mantıksal gruplara bölünmektedir. Bu çalışmanın amacı daha anlamlı ve grup veriler elde etmek ve bu sayede analiz yeteneğini arttırmaktır. Kullanıcıların web sitelerine girişleri farklılık arz etmektedir. Örneğin bir kullanıcı bir web sitesini her gün ziyaret edebileceği gibi, haftalık olarak da ziyaret edebilir. Bazı durumlarda ise bir web sitesini gün içerisinde bir den fazla sefer ziyaret edebilir.

Oturum tanımlama

Oturum, bir kullanıcının bir ziyaretinde siteye yaptığı ziyaretlerin toplamıdır. Kullanıcılar belirlendikten sonra kullanıcının ziyaret kayıtları zamana göre sıralanarak kullanıcının oturumları belirlenebilir. Kullanıcıların oturum bilgileri iki farklı yöntemle bulunabilmektedir. Bunlar Zaman bazlı metot ve içerik bazlı metotdur (Jainve diğerleri, 2017).

Oturumlar ziyaretçi web sitesine tek bir ziyaretinde eriştiği tüm sayfaları içerir. Oturum tanımlama her bir kullanıcının siteye yaptığı tüm ziyaretleri mantıklı parçalara böler. Genel olarak her bir oturum arasında 30 dakika olduğu kabul edilerek işlem yapılır (Sasteve diğerleri, 2017). Oturum tanımlama sürecinde yaşanan çeşitli sıkıntılar mevcuttur. Bu sıkıntıları şu şekilde listelemek mümkündür:

- Hatalı Kullanıcı Tanımlama: Oturum tanımlama kullanıcı tanımlama sürecinden sonra gelmekte ve doğrudan oturum tanımlama sürecine bağlı olarak kalitesi

değişmektedir. Kullanıcı tanımlama hatalı olduğunda, ya da oturum tanımlayıcısının bir oturum içinde değişmesi durumunda, hatalı sonuçlar üretilecektir.

- Önbellek: Tarayıcılar ziyaret edilen sayfaları önbelleklerinde tutarak kaydetmektedirler. Örneğin tarayıcının geri tuşu birçok zaman bir önceki sayfayı tamamen eski kayıtlardan geri yüklemektedir. Böyle durumlarda bu istekler sunucuya erişemediği için bir oturum içerisinde gezilen sayfanın kaydedilememesine sebep olmaktadır.

Akış tamamlama

Akış tamamla bir diğer önemli ön işleme sürecidir ve genelde oturum tanımlama işleminden sonra gerçekleştirilir. Tarayıcı yada proxy önbelleği, önbelleğin devreye girdiği durumlarda, sayfa ya da sayfa elemanlarının web erişim kayıtlarından eksik olmasına yol açar (Liu, 2011). Bu durumu gidermek için oturum tanımlama işlemi sonrasında her bir oturum için bağlantılar incelenerek mümkün olmayan akışlar tespit edilerek bu akışların yerine eksik akışlar eklenir. Bu ön işlemeyi yapabilmek için site topolojisine hâkim olmak gereklidir. Bu süreçte karşılaşılabilecek önemli problemler şunlardır:

- Hatalı Kullanıcı ve Oturum Tanımlama: Bu ön işleme çalışması için oturum tanımlamalarının doğru yapılmış olması gerekir. Oturumların tam ayrılamamış ya da eksik ayrılmış olması hatalı tamamlama süreçlerine sebebiyet verecektir.
- Yüksek Önbellek: Tarayıcı ya da proxy araçlarının çok fazla kaynağı önbellekte tutması sebebiyle tamamlama işlemi yapılamayacak kadar eksik olabilecektir. Bu durumlarda eksik sayfaların telafisi oldukça zor olacaktır.
- Paylaşılan Bağlantılar ve Yer İmleri: Günümüzde sosyal medya ve diğer çeşitli araçlarda kullanıcılar web sayfası bağlantılarını paylaşabilmekte ve aynı zamanda kullanıcılar web sayfalarını yer imi olarak kaydedebilmektedir. Kaydedilen yer imleri arasında gezinme işlemi ya da paylaşılan bağlantıları kullanarak web sitesinde olduğu konumu hızla değiştirecek kullanıcılar için bu tamamlama işlemi oldukça zordur.
- Değişen Link Yapıları: Web 2.0 ve Web 3.0'ın hayatımıza girmesi ile web siteleri dinamik olarak şekillenmekte ve sayfa içerikleri ve sayfaların birbirlerine olan bağlantıları sürekli olarak değişmektedir. Böyle bir durumda web sitesi içeriğinin, kullanıcı ziyareti ve analiz çalışması arasında yer alan bir zaman diliminde değişmesi durumunda, arada yer alan sayfaların takibi oldukça zor olacaktır.

Veri kaynakları arası entegrasyon

Yapılan web kullanım madenciliği çalışmalarına bakıldığında genel olarak kullanılan veri kaynağı web sunucusu kayıt dosyalarıdır. Bu dosya içerisinde yer alan veriler oldukça kısıtlı olduğu için geniş analizlerde bu verileri genişletilmesi gerekmektedir. Yüksek etkinlikte bir örüntü keşfi yapmak için farklı veri kaynaklarının bir araya getirilmesi ve entegre ziyaret akışı verisinin oluşturması gereklidir. Bu durum özellikle e-ticaret uygulamalarında, kullanıcı demografik bilgileri, ürün puan ve kategorileri gibi uygulama veri tabanı kaynakları olarak kendini gösterir (Liu, 2011). Örneğin bir e-ticaret sitesi için hangi ziyaretçinin hangi ürünleri aldığı bu verilere eklenerek daha faydalı analizler yapılabilir. Bu durumda bir başka kaynaktan e-ticaret verilerinin çıkarılması ve web kayıt dosyasındaki bilgiler ile eşleştirilmesi gerekmektedir.

4.3.3. Örüntü keşfi

Diğer veri madenciliği süreçlerinde olduğu gibi, web madenciliği süreçlerinde de model oluşturmak ve bu oluşturulan modele göre sonuçlar almak çalışmanın en önemli noktasıdır. Bu süreçte çeşitli modelleme araçları ve algoritmaları kullanılarak kullanıcılar gruplanmakta, isteklerin birliktelikleri incelenerek sonuçlar çıkarılmaktadır.

Örüntü ve bilgi keşfi, bilinen araçlar ile yapılabileceği gibi, bu işlemler için özelleşmiş çeşitli araçlarda bulunmaktadır. Bu yazılımlar genel olarak web log analiz araçları olarak bilinmektedir. Web kayıt analiz araçları text, xml gibi veri kaynaklarını kullanarak web sitelerini analiz etmektedir. Bu araçlara örnek olarak şu liste verilebilir (Kundu and Garg, 2017):

- Webalizer
- Piwik
- Open Web Analytics
- Deep Log Analyzer
- Fire Stats
- Go Access
- Web Forensik
- AW Log Analyzer
- Web Log Expert

Kundu ve Garg yaptıkları " Web Log Analyzer Tools: A Comparative Study to Analyze User Behavior" isimli çalışmalarında Google Analytics ürününü kayıt analiz aracı olarak belirtmişlerdir. Fakat Google Analytics web sunucusu tarafından tutulan kayıtları analiz eden bir araç değildir.

Web kullanım madenciliğinde uygulanan bir çok teknik vardır (Jainve diğerleri, 2017). Bunlar şu şekilde sayılabilir:

- İstatiksel Analiz
- Birliktelik Kuralları
- Kümeleme
- Sınıflandırma

İstatiksel analiz

İstatiksel analiz en kolay ve hızlı cevap veren analiz yöntemidir. Bu analiz yöntemi karmaşık kurallar vermez. Bu yöntemle erişim kayıtları ortalama, medyan, sıklık ve benzeri istatiksel yöntemlere göre analiz edilirler (Jainve diğerleri, 2017). Web log analiz araçları genel olarak bu yöntemleri kullanmaktadır.

Birliktelik kuralları

Birliktelik kuralı ve istatiksel korelasyon çalışmaları farklı gruplarda yer alan ve birlikte yer alan sayfaların bulunması, birlikte satın alınan ürünlerin tespitinde kullanılır. Bu araçlar ile web site içeriklerinin daha iyi organize edilmesi ve çapraz satış yaratılması mümkün olmaktadır (Liu, 2011). Bu çalışmalar genel olarak sitelerde "bu ürünü alanlar bu ürünü de aldı" ya da "bu yazıyı da sevebilirsiniz" tarzı ifadeler ile karşımıza çıkmaktadır. Bu analiz şekli web site analizlerinde sıkça kullanılmaktadır.

Kümeleme

Kümeleme veri madenciliğinde sık kullanılan gözetimsiz bir madencilik metodudur. Kümeleme veri madenciliği metodu kayıtları benzer karakteristik özelliklerine sahip olma durumlarına göre gruplayan algoritmalardır (Liu, 2011). Web kullanım madenciliğinde genel olarak sayfalar, kullanıcılar, ziyaret akışları gruplanabilmektedir. K-means bu süreçlerde en sık kullanılan metottur (Herath and Jayaratne, 2017).

Sınıflandırma

Sınıflandırma farklı veri gruplarının belirli sınıflara önceden ayrılması ile yapılan gözetimli bir analizdir. Sınıf karakteri seçilen özellikler ve özelliklerin değerlerine göre seçilmektedir (Jainve diğerleri, 2017). Önceden seçilen parametrelere göre kullanıcılar belirli sınıflara ayrılır ve bu sınıflara göre işlemler yapılır. Web analizi uygulamala kullanıcıları çeşitli parametrelere göre sınıflarlar. Örneğin verilen kullanıcı alışverişlerine bakarak, her bir kullanıcı için belirli dönemde yapılan alışveriş tutarı hesaplanabilir (Liu, 2011).

4.3.4. Örüntü analizi

Literatürde yapılan çalışmalarda, daha çok ön işleme süreçleri ve bu süreçlerin optimizasyonları yer almaktadır. Fakat veri madenciliği süreçlerinin en önemli çıktıları örüntü analizleri ve bu süreçler ile alınan kararların uygulanması ile ortaya çıkmaktadır.

Örüntülerin analizi sonrasında bu bilgilerin kullanılabileceği çeşitli uygulamalar mevcuttur. Bu uygulamalar şu şekilde sayılabilir (Jainve diğerleri, 2017):

- Web içeriğini kişiselleştirmek
- Önceden yükleme ve tasarım
- Tasarım kararlarının desteklenmesi
- Müşteri deneyiminin ve memnuniyetinin arttırılması
- İş zekası
- Zenginleştirilmiş elektronik öğrenme

4.4. Web Kullanım Madenciliğinde Geliştirme Alanları

Web kullanım madenciliği oldukça geniş bir kullanım alanına sahiptir. Günümüzde her geçen gün, daha fazla işlem internet üzerinde yapılmaya başlamaktadır. Web kullanımını anlayarak bu kullanımı kişiselleştirmek hem işletmelerin hem de kullanıcıların ihtiyaçlarını karşılamaktadır.

Mevcut teknolojik imkanlar genişlerken, web kayıt standardı olan “common log format” ve “extended log format” toplanabilecek tüm verileri toplamakta geride kalmakta ve teknolojik ihtiyaçlar standartların çok önünde yer almaktadır.

Mevcut çalışmalarda ve literatürde yer alan metotlar, zaten eksik olan veri toplama yöntemlerini daha iyileştirmeye çalışmakta, fakat teknolojik gelişmelere asla yetişememektedirler. Örneğin günümüzde internete çok farklı cihazlardan bağlanılabilmektedir. Ayrıca birçok aracı internet servisi tarayıcılara entegre olarak sunulmakta ve yeni gelişen arayüz uygulama çatılarında statik sayfa yayınları yerine API destekli yapılara geçilmektedir.

İnternet sitelerine erişim metotlarımızın bu kadar geliştiği bir ortamda veri toplama ve birleştirme yöntemlerinin de geliştirilmesi ve madencilik süreçlerinin asıl önemli kısımlarına yönelmek zorunluluğu doğmuştur.

5. UYGULAMA

Literatürde yer alan web kullanım madenciliği çalışmaları incelendiğinde bu çalışmaların web sunucu erişim kayıtlarından elde edilen verilere ve bu verilerin ön işleme faaliyetlerine odaklandığı görülmektedir. Fakat bir veri madenciliği çalışmasında, asıl önemli olan sürecin modelleme ve bu modelleme ile elde edilen modelin kullanımı olduğu yadsınamaz bir gerçektir.

İncelenen çalışmalarda, birçok veri toplanmış ve bu veriler temizleme ve ön işleme süreçlerinden geçirilmiştir.

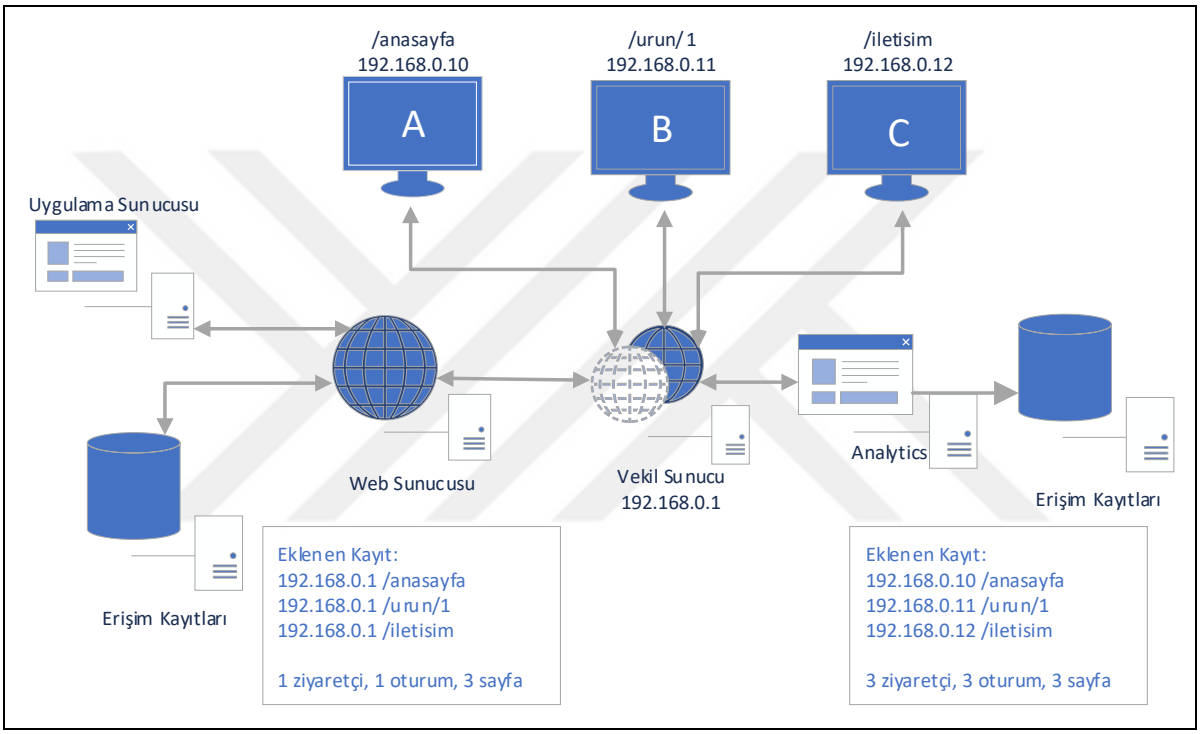
Web kullanım madenciliği sürecinde yaşanan problemleri iyileştirmek ve daha iyi bir ön işleme yapılması amacıyla süreçte yer alan yöntemlerin değiştirilmesi bir zorunluluktur. Bu çalışma içerisinde bu alanlarda çalışmalar yaparak sürecin iyileştirilmesi için yöntemler aranmıştır. Bu sürecin iyileştirilmesi için yeni bir veri kaynağı ve inceleme aracı kullanılmasının çalışmaya yaratacağı farklar gösterilmek istenmiştir. Bu sebeple veri kaynağı olarak web erişim kayıtları yerine Google tarafından geliştirilen Google Analytics kaynağı bütünleşmiş veri kaynağı olarak kullanılmıştır.

5.1. Google Analytics

Google Analytics, Google bünyesinde yer alan 2005 yılında urchin.js'in satın alınması ile devreye alınmış bir takip yazılımıdır. Bu yazılım web sitesi sahiplerinin sitelerine gelen ziyaretçilerin takip ve basit analizlerini yapmak için oldukça elverişli bir araç olmuştur. Zaman içerisinde yazılım Google tarafından geliştirilerek, e-ticaret, tarayıcı bilgisi gibi çeşitli alanları da içerisinde bulunduran bir çözüm aracı haline gelmiştir. (Şayan, Çetinyokuş 2019) Adsense, Adwords gibi diğer Google servisleri ile entegre olan servis, web sitesi sahipleri için yetenekli analizler oluşturabilmektedir.

5.1.1. Çalışma yapısı

Analytics veri toplama yapısı, web kayıt dosyalarından daha farklı çalışmaktadır. Web erişim kayıtları, kullanıcılar sunucu ile etkileşime geçtiklerinde, etkileşimi başlattıkları ip adresi için kaydedilmektedir. Veriler sunucu tarafında kaydedilmektedir. Analytics ise bu sürece farklı olarak web sitelerine eklenen bir javascript dosyası ile veri toplamaktadır. Analytics ürünün çalışma yapısı Şekil 5.1 de gösterilmektedir.



Şekil 5.1. Analytics çalışma yapısı

Bir web sayfası yüklenme işlemi Şekil 5.1 de görülebileceği gibi bir istemciden gelen istekler web sunucuları tarafından web erişim kayıtlarına dönüştürülmekte ve bu istekler erişim kayıt dosyalarına yazılmaktadır. Fakat Şekil 5.1 de örnek olarak gösterildiği gibi arada bir vekil sunucu ya da benzeri bir aracı olduğu durumda bu istekler tek kaynaktan geliyor gibi görünebilmektedir. Fakat Analytics çalışma yapısında, istemciler, ip adresleri yerine kendilerine özel atanmış olan istemci kimlikleri ile kaydedildikleri için istemci sayısı ve ziyaret sayısı doğru olarak kaydedilebilmektedir.

Günümüzde, istemci ve sunucu arasında yer alan web trafiği, vekil sunucu, güvenlik duvarı, internet servis sağlayıcı gibi pek çok ağ katmanı tarafından değişikliğe uğratılabilmekte ve bu durum, literatürde yer alan ön işleme metotlarında sıklıkla kullanılan, kullanıcı ip ve

tarayıcı imza bilgilerinde değişiklikler yapabilmektedir. Google Analytics yapısında kullanıcılara verilen benzersiz imza, istek paketi ile birlikte iletildiği için, ön işleme süreçleri oldukça kısalmaktadır.

5.1.2. Google analytics veri örneği

Google Analytics web erişim kayıtlarından çok daha geniş miktarda veriyi, her sayfa yüklemesinde sunucularına göndermektedir. Şekil 5.2'de Google Analytics tarafından toplanarak sunucuya gönderilen bilgilerden bir bölüm görünmektedir.

Şekil 5.2 Google Analytics gönderilen veriler

Şekil 5.2'de gösterilen verilen tek bir sayfa ziyaretinde Google Analytics tarafından kaydedilen verilerin bir bölümünü göstermektedir. Sayılar ile işaretlenmiş satırlar aşağıda yer alan verileri içermektedir.

1. İlgili web sitesi için Google Analytics tarafından verilen benzersiz site numarası
2. İlgili web sitesi tarafından kullanıcı için verilen kullanıcı id değeri
3. Google Analytics'e web sitesi tarafından bildirilen kullanıcı userId bilgisi
4. Sayfada gösterilen ürünlere ait ürün adı, fiyatı, markası ve liste konumu bilgisi
5. Google Analytics tarafından ilgili kullanıcı ve ilgili site için verilen benzersiz takip numarası
6. Google Analytics sunucularına gönderilen kayıt tipi
7. Kullanıcı tarafından kullanılan tarayıcının java desteği (0:java desteği yok, 1: java desteği var)
8. Kullanıcı tarayıcısının dili
9. Kullanıcının ziyaret ettiği sayfa bağlantısı
10. Kullanıcı ekranının renk derinliği değeri
11. Kullanıcı ekranının çözünürlüğü
12. Kullanıcısın ziyaret ettiği sayfanın başlık bilgisi

Bu veri setinde gönderilen verilerden bir bölümünü erişim kayıtları ile almak mümkün iken özellikle ekran renk derinliği, ekran çözünürlüğü, tarayıcı dili(genelde kullanıcının dili ile aynı), tarayıcı java desteği gibi bilgileri elde etmek mümkün değildir.

5.1.3. Veri güvenliği

Web erişim kayıtları, web sunucularında yer alan erişim dosyalarına yazılmakta ve bu dosyalarda saklanmaktadır. Bu dosyalara erişim için sunucu erişimi gerekmektedir. Birçok web sitesi için bu erişimin yetkisiz bir şekilde sağlanması mümkün değildir. Kullanıcıların ip adreslerinin, bu kişilere erişilmesini ve bu kişilere karşı yapılacak kötü niyetli çalışmalarının yolunu açmaktadır. Bu sebeple bu adreslerin, analiz süreçlerinde doğrudan kullanılması doğru bir yöntem değildir.

Analytics ip adreslerini anonimleştirerek saklamakta ve herhangi bir arayüz üzerinde ya da raporlama araçlarında analistlere vermemektedir. Bu sayede kullanıcılar sadece takip kodları ile modelleme yazılımlarına dahil edilebilmektedir.

5.1.4. Analytics API

Google Analytics tarafından kaydedilen veriler, uygulama arayüzünden incelenebilmekte ve uygulama programlama arayüzü ile veri halinde alınabilmektedir.

Google Analytics API, http protokolü üzerinden çalışmakta ve modern dillerde kütüphaneler sunmaktadır. Örneğin aşağıda verilen adres ile istenen özellikteki verilere ulaşılabilmektedir.

```
https://www.googleapis.com/analytics/v3/data/ga
&start-date=30daysAgo
&end-date=yesterday
&metrics=ga:sessions,ga:pageviews
&dimensions=ga:dimension1
&sort=-ga:pageviews
&filters=ga:sessions>2
&start-index=100
&max-results=100
&access_token=
```

Yukarıda verilen bağlantı ile ziyaret tarihi son 30 günden sonra ve dünden önce, oturum sayısı 2'den çok olan kullanıcıları sayfa gösterimlerini çoktan aza sıralayarak ilk 100 kullanıcıyı atladıktan sonra geriye kalan kullanıcılardan 10 tanesini deimension1 de yer alan bilgiye göre gruplayarak oturum sayıları ve sayfa gösterim sayıları ile raporlamak mümkün olmaktadır.

Bu sorgu ile Google Analytics çok hızlı bir biçimde Çizelge 5.1'de yer alan tabloyu oluşturabilmektedir.

Çizelge 5.1. Google analytics api örnek çıktı tablosu

1 Nolu Boyut	Oturum Sayısı	Sayfa Gösterim Sayısı
visitor-1570380202	5	6
visitor-1570741742	3	6
visitor-1571055362	3	6
visitor-1564166918	4	5
visitor-1565249830	4	5
visitor-1566845853	3	5
visitor-1567027537	3	5
visitor-1568237591	5	5
visitor-1568414955	5	5
visitor-1568749024	5	5

5.2. Google Analytics Kullanımının Sağladığı Faydalar

Web kullanım madenciliği çalışmalarında genel olarak kullanılan erişim kayıtlarının detaylı analizler yapmak için yetersiz olduğu ve kullanıcı ayrıştırması işlemlerini sağlayamadığı kolaylıkla anlaşılabilen bir durumdur. Bu sebeple bu çalışma da veri kaynağı değiştirilmeye çalışılmış ve madencilik süreçleri için Google Analytics ürününden faydalanmanın yolları aranmıştır.

5.2.1. Ip adresi bağımsız kayıt

Analytics ürününde kullanıcılar ip adresleri ile değil kullanıcı tarayıcıları takip edilmektedir. Bu takip işlemi için Web Sitesi için bir takip kodu oluşturulmakta ve bu takip kodu web sitesine eklenmektedir. Örnek bir takip kodu şu şekildedir.

```
<!-- Global site tag (gtag.js) - Google Analytics -->
<script async src="https://www.googletagmanager.com/gtag/js?id=UA-XXXXXXXX-
X"></script>
<script>
window.dataLayer = window.dataLayer || [];
function gtag(){dataLayer.push(arguments);}
gtag('js', new Date());
gtag('config', 'UA-XXXXXXXX-X');
</script>
```

Eklenen bu kod ile Analytics kayıt işlemlerini yönetecek olan “ga” fonksiyonu window nesnesine yüklenmekte ve bu sayfa ziyareti Analytics sunucularına bildirilmektedir. Ayrıca Analytics hizmeti, bir başka Google hizmeti olan Tag Manager ile de web sayfalarına izlenebilmektedir.

Bu kod eklendiğinde, Analytics tarafından, siteye yapılan ilk ziyarette ilgili tarayıcı içerisine bir çerez dosyası oluşturulmakta ve bu dosya içerisinde Analytics tarafından verilen benzersiz kullanıcı kimliği yer almaktadır. Bu ziyaretten sonra yapılan tüm Analytics ziyaretleri bu kod ile Analytics sunucularına gönderilmektedir. Bu durumun temel olarak şu faydalarını göstermek mümkündür.

Her tarayıcıya farklı bir takip kodu verilmesi sayesinde aynı ip adresini kullanan her bir farklı bilgisayar farklı bir kimlikle kaydedilmektedir. Böylece tek bir vekil sunucusu ya da aynı ağı kullanan kullanıcılar tek bir kişi olarak değil, farklı bilgisayarlar olarak görünmektedir.

Dinamik ip adresine sahip olup, her modem açılışında havuzdan yeni bir ip adresi alan bilgisayarlar bu yöntem sayesinde sürekli aynı takip kodu ile takip edilmekte, böylece tek bir kişinin erişim kayıtlarında farklı kullanıcılar gibi öngörülmesi engellenmiş olmaktadır.

5.2.2. UserId desteği

Artık günümüzde kişilerin internete erişebilen birden çok aracı, bu araçlarda ise interneti gezinmesini sağlayan birden çok tarayıcısı olabilmektedir. UserId özelliği Analytics’ e sonradan eklenmiş, birden çok cihazda aynı kullanıcıyı takip etmeyi sağlayan bir yöntemdir.

Bu yöntem için uygulama sunucusundan yararlanılarak, oturum açmış kullanıcılar belirli bir kimlik değişkeni ile oturum açtıkları tüm cihazlarda takip edilmektedir. Bu takip işlemi ile, kullanıcıları ayrıştırma için dolaylı yöntemler yerine web uygulaması ile doğrudan erişimli veriler kullanılabilmektedir.

5.2.3. Geniş veri toplama yeteneği

Web erişim kayıtları ile genişletilmiş kayıt formatı ile veri toplandığında 9 farklı kategoride veri tutulabilmekte ve bu veriler kullanıcı tarayıcısından gönderildikleri şekilde anlamlandırılabilmektedir. Bu toplanan 9 veri ile mevcut çalışmalarda genel olarak link yapısı üzerinde çalışmalar yapılmaktadır.

Google Analytics ise tarayıcı içerisinde çalıştırılan takip kodu ile çok farklı alanlarda veri toplayabilmektedir. Bu alanlara şu örnekler verilebilir.

- Kullanıcı
- Oturum
- Trafik Kaynakları
- E-ticaret
- Platform ve cihazlar
- Sistem
- Sayfa Etkileşimleri

Analytics içerisinde 699 farklı boyut ve metrik yer almaktadır. Bu veriler ile tarayıcı ekran büyüklüğü, sayfa içerisinde tıklanan buton sayısı ve satın alınan ürün sayısı gibi değişkenleri içeren verilerin kullanılması mümkün olmaktadır.

5.2.4. Veri sorgulama ve API desteği

Analytics sunulduğu API desteği ile uygun şekilde yetkilendirilmiş araçlara, toplamış olduğu verileri API üzerinden sunmakta ve bu verilerin kendi sunucularından elde edilmesini sağlamaktadır. Modern diller için hazır kütüphaneler aracılığıyla veriler formatlanmış ve temizlenmiş olarak elde edilebilmektedir. Örneğin aşağıda rest api üzerinden bilgi getiren bir bağlantı ve sonucu yer almaktadır.

5.3. Ön İşleme

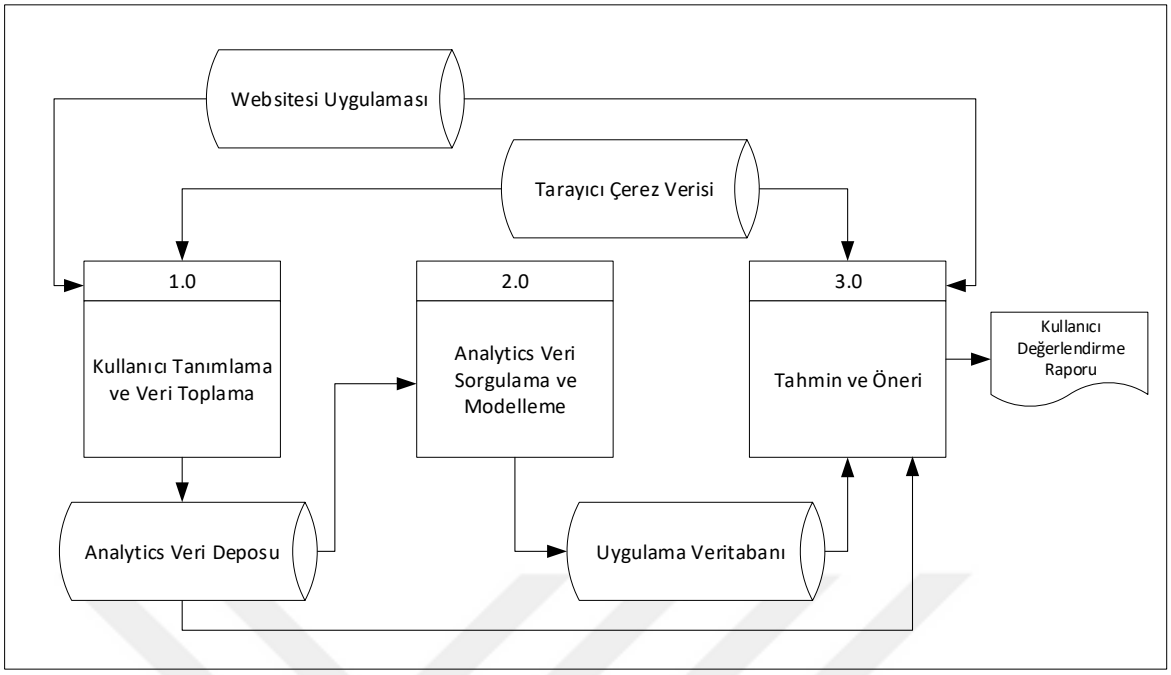
Web kayıtlarından elde edilen veriler, birçok hatalı ve gereksiz kaydı içermektedir. Bu kayıtların işlemesinin geleneksel yöntemlerle yapılması oldukça uzun süren ve hatalara oldukça açık ve çok fazla kabul gerektiren bir süreçtir. Literatürde yer alan bu ön işleme süreçlerini iyileştirmek yerine, daha etkin bir çözüm olan verinin iyileştirilmesini seçmek en doğru çözüm olacaktır.

5.4. Google Analytics ile Entegre Veri Madenciliği Uygulaması

Bu çalışma kapsamında, veri kaynağı olarak web erişim kayıtları yerine Google Analytics ürünü ile veri toplayarak bu sistem üzerinden modelleme yaparak öneride bulunan bir yazılım geliştirilmiştir. Bu uygulama sayesinde web erişim loğlarının sunucu üzerinden indirilmesi gereksinimi ortadan kaldırılmış ve UserId özelliği ile kullanıcıların çapraz cihazlardan kayıt altına alınarak veri entegrasyonunun otomatik olarak gerçekleştirilmesi sağlanmıştır.

5.4.1. Sistem mimarisi

Geliştirilen yazılım temel olarak 3 bölümden oluşmaktadır. Bu yapıyı oluşturan veri akış diyagramı Şekil 5.3 de gösterildiği şekildedir.



Şekil 5.3. Sistem veri akış diyagramı

Uygulama içerisinde 3 temel süreç ve 4 veri kaynağı bulunmaktadır.

Kullanıcı tanımlama ve veri Toplama süreci web tarayıcı içerisinde gerçekleşmektedir. Tarayıcı içerisinde yer alan tanımlayıcı bilgileri ve uygulama sunucusu tarafından alınan tanımlama bilgileri bir araya getirilerek Analytics sunucularına gönderilmektedir. Bu gönderim işlemi Analytics javascript kütüphanesi ile sağlanmaktadır.

Analytics veri sorgulama ve modelleme süreci geliştirilen uygulama içeresinde iki farklı kısımda ilerlemektedir. Bu işlemlerden bir tanesi oluşturulacak modelin planlanması ve buna uygun Analytics sorgularının oluşturulması işlemidir. Diğer işlem ise oluşturulan sorguya uygun verilerin Analytics API ile toplanması ve modelleme işleminin yapılmasıdır. Bu süreçler tamamen geliştirilen uygulama içerisinde ilerlemekte ve Analytics veri kaynağından toplanan bilgiler işlenerek uygulama veri kaynağına model bilgisi yazılmaktadır.

Tahmin ve Öneri süreci kullanıcı tarayıcısından başlayarak, uygulama sunucusu içerisinde tamamlanmaktadır. Web sitesi ziyaret edildiğinde, web sitesi veri kaynağı ve tarayıcı çerez kaynağındaki veriler alınarak uygulama sunucusuna gönderilmekte, ilgili tanımlama bilgilerine göre Analytics veri kaynağından toplanan veriler, uygulama veri kaynağından alınan model ile değerlendirilmektedir.

Uygulama temel olarak 4 farklı veri kaynağı kullanmakta ve bu kaynaklar farklı ortamlarda yer almaktadır.

Web sitesi uygulaması veri kaynağı, web sitesinin çalışmasını ve kullanıcı, içerik ve benzer siteye özel bilgilerin saklandığı veri kaynağıdır. Bu veri kaynağından kullanıcının mevcut bir kullanıcı mı yoksa yeni bir ziyaretçi mi olduğu bilgisi alınmakta, mevcut kullanıcı ise kullanıcı id verisi elde edilmektedir.

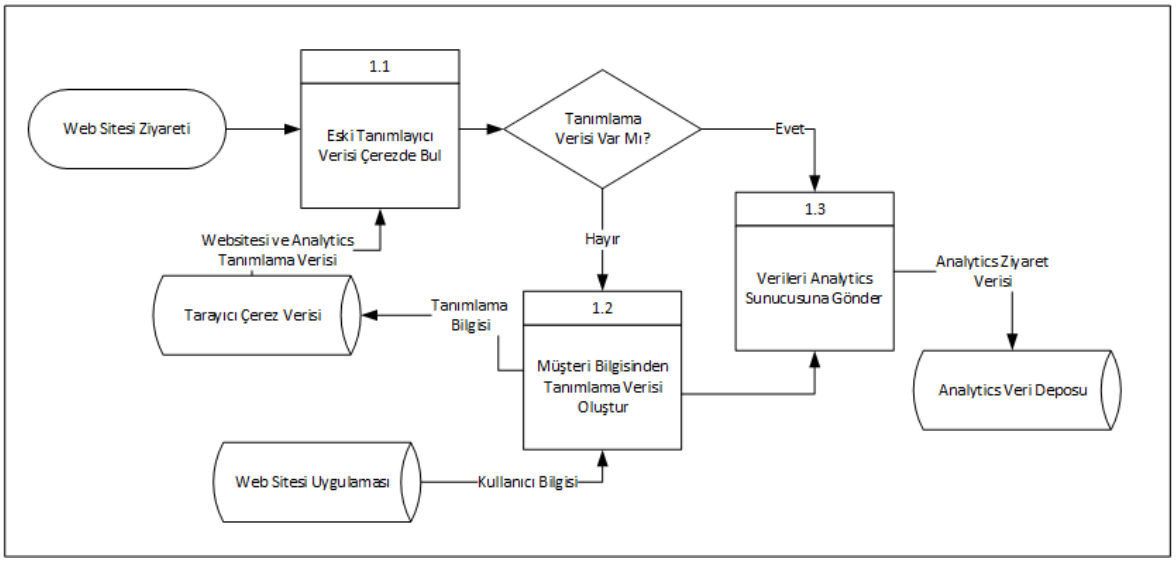
Tarayıcı çerez verisi tek bir noktada tutulmamakta web sitesini ziyaret eden her kullanıcının bilgisayarında saklanmaktadır. Bu veri Analytics tarafından kullanıcıyı tanımlamak için kullanılmaktadır.

Analytics veri deposu, Analytics kütüphanesi tarafından kayıt edilen ve sunulan API ile okunan kullanıcı ziyaret verilerinin saklandığı veri deposudur.

Uygulama veri tabanı, geliştirilen uygulama bilgilerinin ve oluşturulan sorgular ile modellerin saklandığı veri sistemidir.

Kullanıcı tanımlama ve veri toplama süreci

Kullanıcıların tanımlandığı ve verilerin kaydedildiği alt süreçtir. Üç alt süreçten oluşmakta ve üç veri kaynağı ile etkileşime geçmektedir. Bu süreç web sitesi ziyareti gerçekleştiğinde otomatik olarak gerçekleşmektedir. Bu bölümde gerçekleştirilen görevlerin büyük kısmı Analytics javascript kütüphanesi tarafından gerçekleştirilmekte ve web sitesi yazılımı bu sürece kullanıcı bilgisi sağlamaktadır. Süreç akış diyagramı Şekil 5.4'de gösterilmektedir.



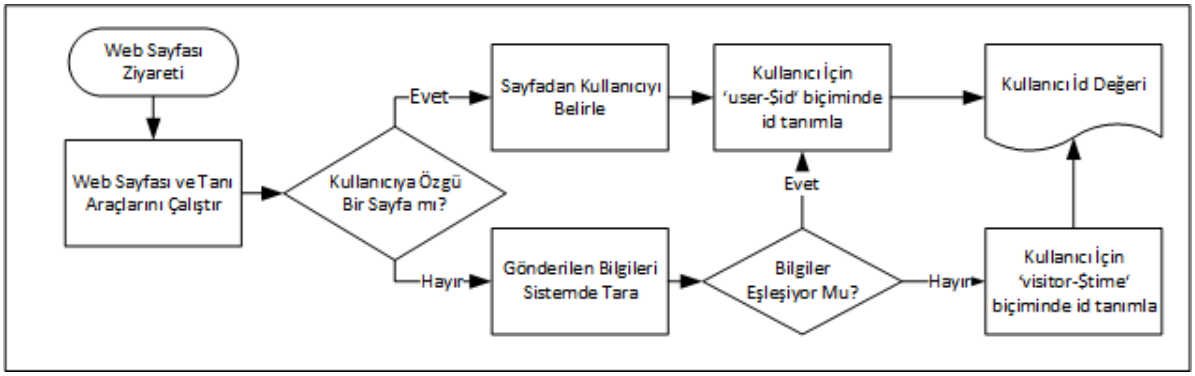
Şekil 5.4. Kullanıcı tanımlama ve veri toplama süreci veri akış diyagramı

1.1. Eski Tanımlayıcı Verisi Çerezde Bul

Google Analytics tanımlama bilgileri ve web sitesi kullanıcı tanımlama bilgisi *Tarayıcı Çerez Verisi* veri deposunda saklanmaktadır. Bu süreç içerisinde bu veriler tespit edilmektedir. Ziyaretçinin ilk ziyaretinde bu veriler mevcut değildir. Eğer tanımlama verisi yoksa *1.2 Müşteri Bilgisinden Tanımlama Verisi Oluştur* süreci işletilmekte, tanımlama verisi var ise *1.3 Verileri Analytics Sunucusuna Gönder* süreci işletilmektedir.

1.2. Müşteri Bilgisinden Tanımlama Verisi Oluştur

Mevcut kullanıcı için tanımlama verisi olmadığında işleyen bu süreç *Web Sitesi Uygulaması* veri deposundan tanımlama bilgileri almakta ve bu bilgileri daha sonraki ziyaretlerde kullanmak üzere *Tarayıcı Çerez Verisi* veri deposuna eklemektedir. Tanımlama verisi oluşturulurken Şekil 5.5'de yer alan süreç kullanılmaktadır.



Şekil 5.5. Kullanıcı id değeri tanımlama süreci

Tanımlama verisi Tarayıcı Çerez Verisi veri kaynağında olmasa da web sitesi uygulaması sayfa isteğinden bu veriyi oluşturabilmektedir. Örneğin kişiye özel bir sayfa veya bir formdan gönderilen telefon numarası gibi kişisel bir veri ile kullanıcı tanımlanabilmektedir. Tanımlanabilen kullanıcılar için tanımlama verisi 'user-' ifadesi ile id değerleri birleştirilerek, tanımlanamayan kullanıcılar için ise 'visitor-' ifadesi anlık Unix Time değeri birleştirilerek oluşturulmaktadır.

1.3. Verileri Analytics Sunucusuna Gönder

Bu süreç Analytics tarafından sağlanan javascript kütüphanesi ile sağlanmaktadır. Analytics izleme kodu web sitelerine javascript olarak eklenmekte verilen bilgilere göre *Analytics Veri Deposu* kaynağına tarayıcı ile gönderilmektedir. Aşağıda web sitesi HTML kodlarına eklenen Analytics kütüphanesi gösterilmektedir.

```
<!-- Global site tag (gtag.js) - Google Analytics -->
<script async src="https://www.googletagmanager.com/gtag/js?id=UA-XXXXX"></script>
<script>
  window.dataLayer = [{'userID': 'user-74'}];
  function gtag(){dataLayer.push(arguments);}
  gtag('js', new Date());
  gtag('config', 'UA-XXXXXX');
</script>
```

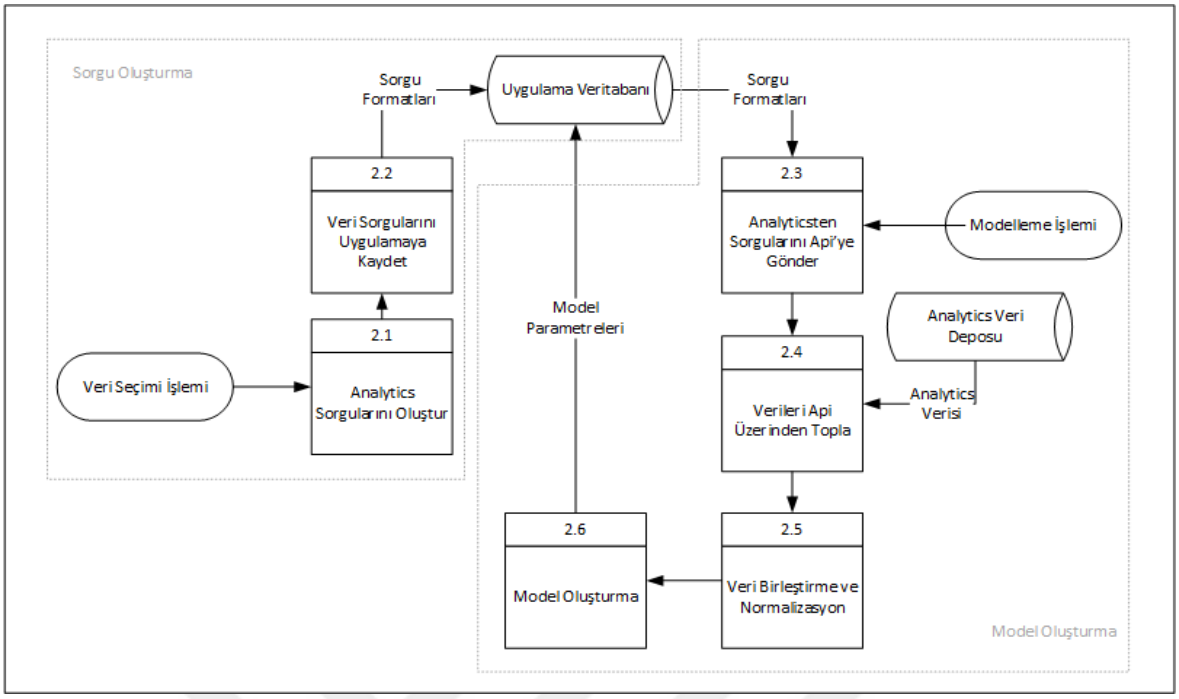
Bu kod içerisinde yer alan **UA-XXXXXX** ifadesi Google Analytics tarafından her web sitesi için oluşturulan kimlik değeridir. Kod içerisinde yer alan **{'userID': 'user-74'}** ifadesi ise *1.2 Müşteri Verisinden Tanımlama Verisi Oluştur* süreci ile oluşturulmuş olan tanımlama ifadesini göstermektedir. Bu veri kütüphane yardımı ile Analytics UserID özelliğine gönderilmekte ve böylece farklı tarayıcılar ve mobil cihazların verileri tek bir kullanıcıya ait olacak şekilde belirlenmektedir. Bu sayede, bir kullanıcının farklı cihazlarda ya da konumlarda, farklı kullanıcılar olarak değerlendirilmesi engellenmiş olmaktadır.

Analytics veri sorgulama ve modelleme süreci

Veri sorgulama ve modelleme süreci, uygulama sunucusunda gerçekleşen iki bölümlü bir süreçtir. Bu bölümleri şöyle sıralamak mümkündür.

- Sorgu Oluşturma
- Model Oluşturma

Sorgu oluşturma süreci, Analytics API kaynaklarından toplanacak verilerin sorgu haline getirilmesi için hazırlık aşamasıdır. Model oluşturma süreci ise toplanmış verilerin mevcut veri madenciliği algoritmaları ile modellenmesi ve bu modellerin daha sonra kullanılmak üzere saklanması sürecidir. Sorgu oluşturma ve modelleme süreci veri akış diyagramı Şekil 5.6’ da gösterildiği şekildedir.



Şekil 5.6. Analytics veri sorgulama ve modelleme süreci veri akış diyagramı

Sorgu oluşturma işlemi 5.1.3 Analytics API bölümünde aktarılan veri sorgulama dili formatına göre hazırlanan sorgular ile yapılmaktadır. Uygulamaya tanımlanmış olan Analytics hesap detayları, seçilen parametrelere göre veriler toplanarak uygulamaya kaydedilmektedir.

2.1. Analytics Sorgularını Oluştur Süreci

Analytics sorgularının oluşturulması için, istenen boyut ve ölçütlere, filtrelere ve hesap bilgilerine ihtiyaç vardır. Bu bilgiler sağlanarak her bir hesap için farklı sorgular tanımlanabilmektedir. Bu süreç analiz yapacak kullanıcının sorgularını uygulama ekranlarından seçerek, hiç kod yazmadan Analytics API ile toplanmasını istediği veri setlerini oluşturmasını sağlamaktadır.

2.2. Veri Sorgularını Uygulamaya Kaydet Süreci

Bu süreçte kullanıcı tarafından oluşturulan sorgular, uygun JSON formatına dönüştürülerek *Uygulama Veri tabanı* veri kaynağına kaydedilmektedir. Uygulama verileri MongoDB üzerinde nesneler olarak saklamaktadır. Uygulama içerisinde oluşturulmuş bir sorgu verisi aşağıda yer aldığı şekilde kaydedilmektedir.

```
{
   "_id": {
      "$oid": "5c8802347ba92131ee6aca3a"
   },
   "accountId": "599c8331bde0c679db78f030",
   "name": "Kullanım ve Eticaret",
   "description": "kullanım ve eticaret verilerini toplar",
   "queries": [{
         "name": "Kullanım Verileri",
         "dimensions": {
            "ga:dimension1": "Özel Filtre 1"
         },
         "metrics": {
            "ga:sessions": "Sessions",
            "ga:avgSessionDuration": "Avg. Session Duration",
            "ga:pageviews": "Pageviews",
            "ga:transactionRevenue": "Revenue"
         },
         "filters": [{
               "item": "ga:visits",
               "type": ">",
               "value": "3"
            },{
               "item": "ga:dimension1",
               "type": "=~",
               "value": "(user|visitor)-"
            }]
      },{
         "name": "Eticaret verileriş",
         "dimensions": {
            "ga:dimension1": "Özel Filtre 1"
         },
         "metrics": {
            "ga:transactions": "Transactions",
            "ga:revenuePerTransaction": "Avg. Order Value"
         },
         "filters": [{
               "item": "ga:transactions",
               "type": ">",
               "value": "0"
            },{
               "item": "ga:dimension1",
               "type": "=~",
               "value": "(user|visitor)-"
            }]
      }],
   "end-date": "1daysAgo",
   "start-date": "500daysAgo"
}
```


Bu kaydedilmiş veri yapısı içerisinde yer alan "**599c8331bde0c679db78f030**" ifadesi yine aynı veri tabanında yer alan Analytics hesap kaydı için uzak anahtar değeridir. Her bir sorgu içerisinde bir den fazla sorgu grubu olabilmektedir. Bu sorgu grupları örnek veri içerisinde yer alan **queries** anahtarı ile işaret edilen ayrı gruplardır. Her bir grup kendi boyut, ölçüt ve

filtrelerine sahip olabilmektedir. Örneğin yukarıda yer alan veride sayfa ziyaret bilgileri ziyaret sayısı 3'ten büyük kullanıcılar için toplanırken, e-ticaret verileri sadece ticaret işlemi yapmış kullanıcılar için toplanmaktadır. Bu sayede yapılacak API sorgu sayısı azaltılmış olmaktadır.

2.3. Analytics Sorgularını API'ye Gönder Süreci

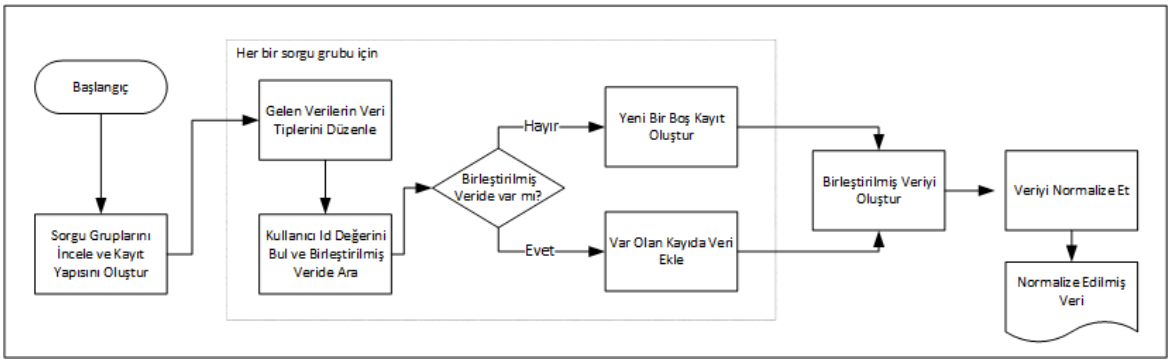
Uygulama kullanıcısı tarafından yapılan istek ile oluşturulmuş ve JSON formatında *Uygulama Veri tabanı* veri kaynağında yer alan sorgu formatları Analytics sorgularına bu süreç içerisinde dönüştürülmektedir. Bu süreçte oluşturulan sorgu formatları, sorgu yapılacak hesap verileri, sorgu tarih aralıkları ve istenen verileri içermektedir. Bu süreç uygulama yazılımı içerisinde gerçekleşmektedir. Oluşturulan sorgular *2.4 Verileri API Üzerinden Topla* sürecine gönderilmektedir.

2.4. Verileri API Üzerinden Topla Süreci

Bu süreç uygulama yazılımı içerisinde tamamlanmaktadır. *2.3 Analytics Sorgularını API'ye Gönder* sürecinden elde edilen formatlar ile *Analytics Veri Kaynağı* veri deposundan veriler getirilmektedir. Analytics API ile tek bir istekte 10.000 satır veri alınabilmektedir. Bu nedenle, bu süreç getirilecek veri miktarını 10.000 satırlık parçalara bölerek gerektiği kadar planlamakta ve her bir istekte gelen verileri biriktirerek toplam veriye erişmektedir.

2.5. Veri Birleştirme ve Normalizasyon Süreci

Bu süreç *2.4 Verileri API Üzerinden Topla* sürecinden gelen verilerin bir araya getirilmesinden sorumlu olan süreçtir. Bu süreç eksik bilgileri oluşturmakta ve farklı sorgu gruplarından gelen verileri tek bir veri grubuna çevirmekte ve bu veri üzerinde normalizasyon işlemleri yapmaktadır. Bu süreç Şekil 5.7'de verilen akışa göre işlemektedir.



Şekil 5.7. Veri birleştirme ve normalizasyon süreci veri akış diyagramı

2.6. Model Oluşturma Süreci

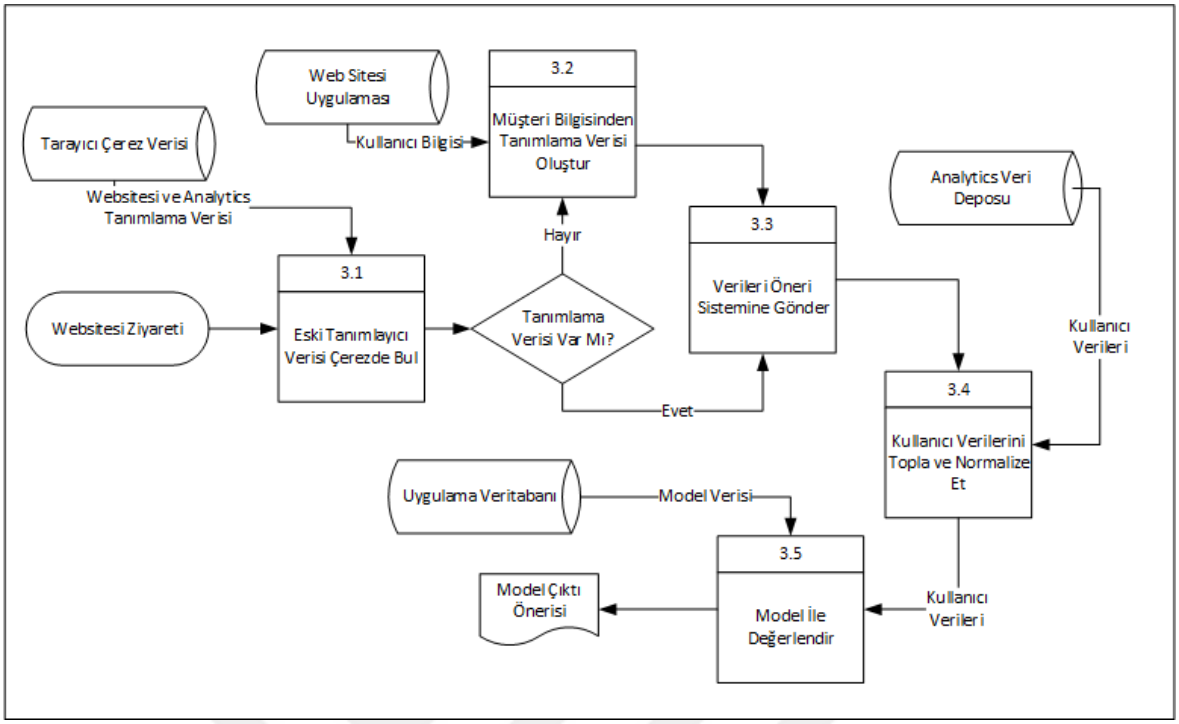
Bu süreç uygulama içerisinde gerçekleşmektedir. *2.5 Veri Birleştirme ve Normalizasyon* süreci ile birleştirilmiş veriler sistemde yer alan veri madenciliği modelleri ile işlenmektedir. İşlenen modeller uygulama ekranları üzerinden incelenmekte ve Tahmin ve Öneri süreçlerinde kullanılmak üzere *Uygulama Veri Tabanı* veri kaynağında saklanmaktadır. Veri tabanında saklanması için modelin JSON formatına çevrilmesi yine bu süreç içerisinde yapılmaktadır. Veri tabanına kaydedilmiş örnek bir model yapısı şu şekildedir. Bu örnek veri içerisinde yer alan **clusters** grubu karakter sayısı uzunlu sebebiyle kısaltılmıştır.


```
{
   "_id": {"$oid": "5c91412fce13b64e3a53d9da"},
   "name": "Raspberry Takip",
   "user": {"$oid": "590e4176edea0b63c3117d17"},
   "accountId": { "$oid": "599c8331bde0c679db78f030"},
   "queryId": "5c8802347ba92131ee6aca3a",
   "nearest": null,
   "cluster": "3",
   "model": {
      "clusters": [
0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,0,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,0
,1,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0
……
……
……
2,2,2,2,2,0,0,0,2,0,0,0,0,0,2,2,2,0,2,2,0,2,2,0,2,2,2,0,0,2,0,0,0,2,0,2,0,2,0,2,0,0,0,2,0,2,2,2,0,2,2,0,0,2,2,2,0
,2,0,0,0,0,0,0,2,2,0,2,2,0,0,0,2,2,2,0,0,2,0,0,0,0,0,0,2,2,0,2,0,2,0,2,0,2,0,0,0,0,0,2,0,0,0,0,0,0,         ],
      "centroids": [{
         "centroid": [
            0.05421796482412145,
            0.03987437185929663,
            0.008671482412060576,
            0.00003545477386934673 36,
            0.0032788944723618095,
            0.006658291457286431
         ],
```

```
            "error": 0.01169396721555347,
            "size": 3184
          }, {
            "centroid": [
              0.07194610778443113,
              0.5317514970059878,
              0.032679640718563434,
              0.004296407185628741,
              0.013652694610778445,
              0.009880239520958085
            ],
            "error": 0.05368932845387084,
            "size": 668
          },{
            "centroid": [
              0.13023166023166016,
              0.2713513513513516,
              0.11405405405405401,
              0.13208494208494206,
              0.19984555984556035,
              0.5589189189189185
            ],
            "error": 0.2949468016278828,
            "size": 259
          }],
        "converged": true,
        "iterations": 23
    },
    "modelName": "kmeans",
    "limits": {
        "sessions": [0,140 ],
        "avgSessionDuration": [0, 484.25],
        "pageviews": [0,1181],
        "transactions": [0,36 ],
        "revenuePerTransaction": [0,325 ],
        "transactionRevenue": [0,5038 ]
    }
}
```


Tahmin ve öneri süreci

Tahmin ve öneri süreci, web sitesini ziyaret eden kullanıcıların bilgilerinin mevcut modellere göre karşılaştırılması ve bu kullanıcılara uygun tahminlerin çıkarılıp önerilerin geliştirildiği süreçtir. Bu süreç kullanıcı tarayıcısından başlayan bir tetikleme işlemi ile uygulama sunucusunda tamamlanmaktadır. Bu sürece ait veri akış diyagramı Şekil 5.8 de gösterilmiştir.



Şekil 5.8. Tahmin ve öneri süreci veri akış diyagramı

Bu süreç Kullanıcı Tanımlama ve Veri Toplama süreci gibi başlamaktadır. Elde edilen tanımlayıcı bilgisi uygulama sunucusuna gönderilmektedir.

3.1. Eski Tanımlayıcı Verisi Çerezde Bul

Tarayıcı Çerez Verisi kaynağında, kullanıcı tanımlayıcı id değeri aranarak bu veriye göre işlem yapılmaktadır. Veri mevcudiyeti kontrol edilerek tanımlama verisi bulunamadığı taktirde *3.2 Müşteri Bilgisinden Tanımlama Verisi Oluştur* süreci çalıştırılmaktadır. Tanımlama verisi veri kaynağında bulunduğunda *3.3 Verileri Öneri Sistemine Gönder* süreci çağırılmaktadır.

3.2. Müşteri Bilgisinden Tanımlama Verisi Oluştur

Mevcut kullanıcı için tanımlama verisi olmadığında işleyen bu süreç *Web Sitesi Uygulaması* veri deposundan tanımlama bilgileri almakta ve bu bilgileri daha sonraki ziyaretlerde kullanmak üzere *Tarayıcı Çerez Verisi* veri deposuna eklemektedir.

Bu süreç 1.2. Müşteri Bilgisinden Tanımlama Verisi Oluştur süreci ile aynı akışta çalışmaktadır.

3.3. Verileri Öneri Sistemine Gönder Süreci

Ziyaretçi tarayıcısında oluşturulan tanımlama bilgileri uygulama sunucusuna gönderilmektedir. Bu istek uygulama sunucusu tarafından bir görsel dosya olarak cevaplanmaktadır. Bu yöntem çeşitli güvenlik yazılımlarının analiz isteğini engellemesini önlemek için kullanılmaktadır.

3.4. Kullanıcı Verilerini Topla ve Normalize Et Süreci

Bu süreç uygulama yazılımı içerisinde gerçekleşmektedir. *3.3 Verileri Öneri Sistemine Gönder* süreci ile alınan kullanıcı tanımlayıcı verisi kullanılarak kullanıcı verileri toplanmaktadır. *Uygulama Veri Tabanı* veri kaynağından alınan Analytics sorgu formatları kullanılarak tekrar Analytics sorguları oluşturulmakta ve sadece ilgili ziyaretçi için sorgular düzenlenmektedir. Oluşturulan sorgular Analytics Veri Deposu veri kaynağında sorgulanarak ilgili kullanıcı için olan veriler elde edilmekte ve bu veriler model oluştururken elde edilen özellikler ile normalize edilmektedir.

3.5. Model ile Değerlendir Süreci

3.4 Kullanıcı Verilerini Topla ve Normalize Et süreci ile Analytics Veri Deposu veri kaynağından elde edilen kullanıcı verisi *Uygulama Veri Tabanı* veri kaynağından alınan daha önce kaydedilmiş model ile test edilerek sonuç raporu oluşturulmaktadır. Bu şekilde her bir ziyaretçinin oluşturulan modeller ile test edilmesi ve özelliklerine göre önerilerin hazırlanması mümkün olmaktadır.

Bu yapı ile kullanıcı ziyareti şu süreci doğurmaktadır.

1. Web sayfasında yer alan Analytics kodu kullanıcı için benzersiz bir kimlik oluşturarak bu veriyi kullanıcı tarayıcısında kaydetmektedir.
2. Google Analytics kodu, kullanıcı işlemlerini Analytics sunucularına kullanıcının bilgilerini ve tanımlama bilgileri ile göndererek kayıt altına almaktadır.

5.4.2. Yazılım dili

Günümüzde web uygulamaları ve tarayıcıları oldukça gelişmiştir. Bu sebeple analiz uygulaması web tabanlı olarak geliştirilmiştir. Bu sayede uygulamanın her sistemden kolayca erişilebilir ve platform bağımsız olması amaçlanmıştır. Geliştirilen uygulamada, sunucu yazılım dili olarak node.js kullanılmıştır. Uygulama arayüzlerinin geliştirilmesinde ise HTML işaretleme dili kullanılmış ve Angular 7.0 uygulama çatısı tercih edilmiştir.

Node.js ve express.js framework

Node.js eş zamansız olaylara dayalı Javascript çalışma ortamı olan Node.js, ölçeklenebilir ağ uygulamaları oluşturmak için tasarlanmıştır. Bu eş zamansız çalışma sistemi, işletim sistemlerinin kullanıldığı günümüzün daha yaygın eşzamanlılık modelinin aksine çalışacak şekilde geliştirilmiştir. Thread tabanlı ağ iletişimi nispeten yetersiz ve kullanımı karmaşıktır. Node.js kullanıcıları, süreçlerinde bekleme ve kilitlenmeler olmadan çalışırlar.

Node.js paket kütüphanesi npm üzerinde yapılan "machine learning" aramasında 440 adet paket listelenmektedir. Google tarafından geliştirilen makine öğrenmesi platformu olan Tensorflow'da javascript diline uyarlanmıştır. Bu durum da göz önüne alındığında javascript tabanlı node.js dili bu uygulamalar için oldukça ideal görünmektedir. Bu dilin yazılım süreçlerinde kullanılmak için seçilmesinin diğer nedenleri şu şekilde sayılabilir.

- Açık kaynaklı ve ücretsiz
- Bütün işletim sistemlerinde çalışma
- Hızlı öğrenme eğrisi
- Geniş kütüphane desteği(350,000 üzerinde paket)(Brown, 13.01.2017)

Express.js ise günümüzde Node.js ile en çok kullanılan web frameworklerinden birisidir. Haftalık indirme sayısı npm.js verilerine göre 9 milyonun üzerindedir(npmjs.com, 2019). Bu framework hızlıca uygulanabilir, routing ve veri işlemlerini hızlandıran bir yapıdır. Middleware mantığıyla çalışan framework oldukça esnek bir şekilde geliştirilebilmektedir.

MongoDB

Geliştirilen uygulama içerisinde veri tabanı çözümü için MongoDB kullanılmıştır. MongoDB ölçeklendirmeye ve esnekliğe uygun, sorgulama ve indeksleme özelliklerine

sahip nesne veri tabanıdır. İlişkisel bir veri tabanı değil nesneleri kaydeden bir veri tabanıdır. Geliştirilen uygulama da aşağıda verilen özellikleri sebebiyle tercih edilmiştir (mongodb.com, 2019).

- Verileri esnek bir şekilde JSON benzeri dokümanlarda saklar ve böylece her bir kaydın veri yapısı kolaylıkla değiştirilebilir.
- Node.js gibi javascript tabanlı bir dilde verilerin uygulama kodunda kullanılması sırasında tip dönüşümleri gibi süreçlere gerek kalmadan kullanılmasını sağlar.
- Tamamen ücretsiz olarak kullanılmaktadır.

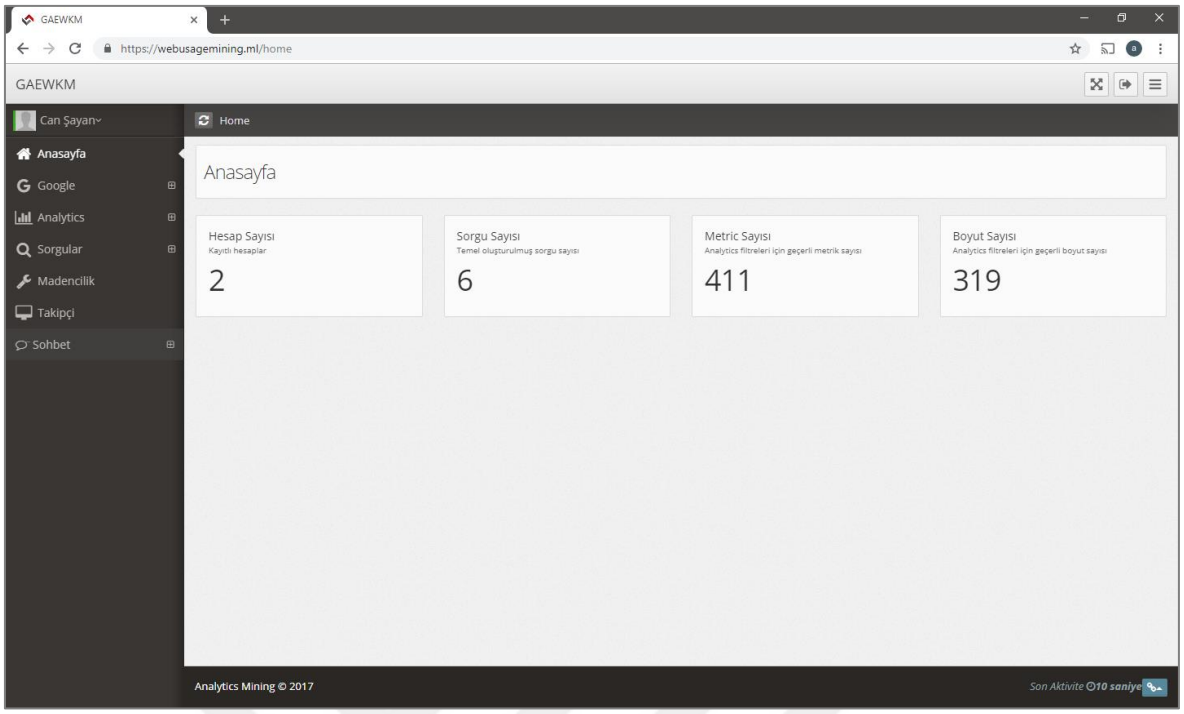
Angular 7.0

Angular, Google tarafından geliştirilen web arayüz platformudur. Typescript dili ile yazılarak javascript, html ve css diline derlenerek tarayıcı içerisinde istenen arayüzlerin oluşturulmasını sağlamaktadır. Sunucu etkileşimi için ek özellikler sağlamakta ve bu hızlı geliştirme fırsatı sağlamaktadır (angular.io, 2019). Angular'ın geliştirilen uygulama da tercih edilmesinin nedenleri şu şekilde sayılabilir.

- Farklı platformlara uygun olarak çalışmakta, web, mobil, native web gibi farklı ortamlarda başarılı çalışmaktadır.
- Hızlı ve performanslı bir platformdur.
- Geliştirme süreçlerini kısaltmak için mevcut Node.js modüllerinin kullanılmasına imkân sağlamakta ve birçok ek özelliği içerisinde getirerek geliştirme süreçlerini kolaylaştırmakta ve güvenliğini arttırmaktadır.

5.4.3. Arayüz tasarımı

Uygulama arayüz tasarımları HTML, CSS Angular 7.0 uygulama çatısı ile yapılmıştır. Arayüz tasarımında açık kaynaklı Bootstrap Framework kullanılmıştır. Bootstrap Framework arayüzlerin tüm mobil ve masaüstü cihazlara uygun olarak geliştirilmesine yardımcı olmakta ve geliştirme süreçlerini hazır eklentiler ve arayüz formatları ile hızlandırmaktadır. Geliştirilen uygulamanın ana sayfası görüntüsü Şekil 5.9’da yer almaktadır.



Şekil 5.9. Geliştirilen uygulama ana sayfa ekran görüntüsü

Ekran yerleşimi

Uygulama ekran yerleşimi temel olarak 3 bölümden oluşmaktadır. Bunlar şu şekilde sayılabilir.

- Başlık Çubuğu: Uygulama adı, arama ve çıkış yap ve tam ekran gibi genel işlevleri barındırır.
- Menü: Uygulama içerisinde gezinmeyi sağlayan menünün yer aldığı bölümdür.
- Sayfa İçeriği: Seçilen menü aksiyonuna göre yönetim elemanlarının gösterildiği alandır.

Bu alanların uygulama ekranında yerleşimleri Şekil 5.10'da gösterilmektedir.

Şekil 5.10. Uygulama ekran yerleşimi

Uygulama arayüzleri

Uygulama arayüzleri tasarlanan sistem mimarisinin kullanıcı denetimine ihtiyaç duyulan tüm süreçlerini içerecek şekilde planlanmış ve menü içerisinden erişilebilir kılınmıştır. Ekran tasarımında yer alan menü 5 menü elemanı ve alt elemanlarından oluşmakta ve ilgili tüm süreçlere erişim sağlamaktadır. Alt menü elemanları gruplanarak ana menü altına yerleştirilmiştir. Uygulama menüsü aşağıda verilen hiyerarşik yapıda konumlandırılmıştır.

- Ana sayfa
- Google
    - Google İle Bağlan
- Analytics
    - Analytics Hesapları
    - Yeni Hesap
- Sorgular
    - Sorgular
    - Yeni Sorgu
- Madencilik
- Takipçi

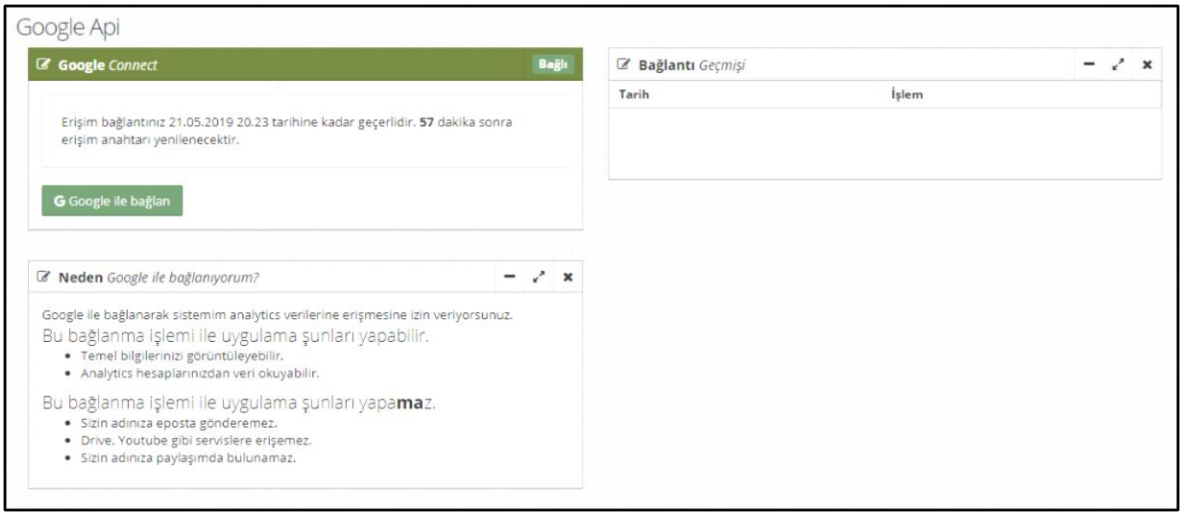
Ana sayfa arayüzü

Ana sayfa arayüzü, uygulamada oturum açıldıktan sonra kullanıcıyı karşılayan ve kullanıcıya özet bilgi veren ekrandır. Bu ekranda 4 temel bilgi sunulmaktadır.

1. Hesap Sayısı: Sistemde kayıtlı Google Analytics hesabı sayısı
2. Sorgu Sayısı: Sistemde ilgili hesaplardan veri toplama işlemini gerçekleştiren sorguların sayısı
3. Metrik Sayısı: Sistem tarafından tanınan ve Google Analytics içerisinde kullanılabilen metrik ifadeleri sayısı
4. Boyut Sayısı: Sistem tarafından tanınan ve Google Analytics içerisinde kullanılabilen boyut ifadeleri sayısı

Google ile bağlan arayüzü

Uygulamanın Google Analytics verilerine erişebilmesi için, ilgili Analytics hesabı yöneticisi tarafından erişim izninin verilmesi ve erişim anahtarının oluşturulması gerekmektedir. Bu süreç tarayıcı vasıtasıyla gerçekleşmektedir. Google ile bağlan arayüzü Şekil 5.11'de gösterilmektedir.

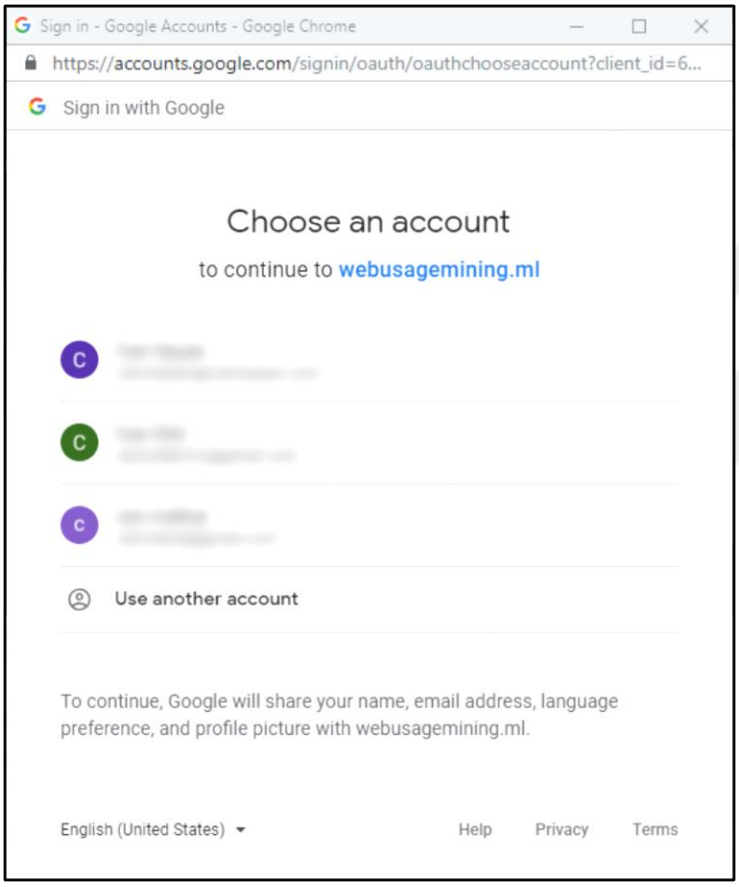


Şekil 5.11. Google ile bağlan arayüzü

Bağlantı durum ve bağlan bölümünde mevcut yetkilendirmenin durumu ve süresi gösterilmektedir. Yetkilendirme anahtarları 60 dakikalık ömre sahiptir ve bu süre sonrasında ilk hesap erişimi sırasında uygulama yazılımı tarafından otomatik olarak yenilenmektedir. Bu ekranda yer alan "Google ile bağlan" düğmesi, Google Hesap bağlantısı sürecini

başlatmaktadır. Bu süreç uygulama sunucusu tarafında yer alan Google Javascript Kütüphanesinin ilgili metotları tarafından sağlanmaktadır.

Google ile bağlan düğmesi yeni bir pencere açarak Google Erişim sayfasına yönlendirmekte ve Google tarafından izin ekranı kullanıcıya sunulmaktadır. Google uygulama izin verme arayüzü Şekil 5.12'de gösterilmektedir.



Şekil 5.12. Google izin ekranı

Ekranın ikinci kısmı ise daha önce yapılmış Google Analytics hesap erişimlerinin listesini göstererek bilgi vermektedir. Bu ekran vasıtasıyla uygulamanın Google Analytics verilerine erişimleri izlenebilmektedir.

"Neden Google ile bağlanıyorum" başlıklı bölüm ise bu bağlanma sürecinin ayrıntılarını ve istenen erişim izni hakkında kullanıcılara bilgi vermektedir.

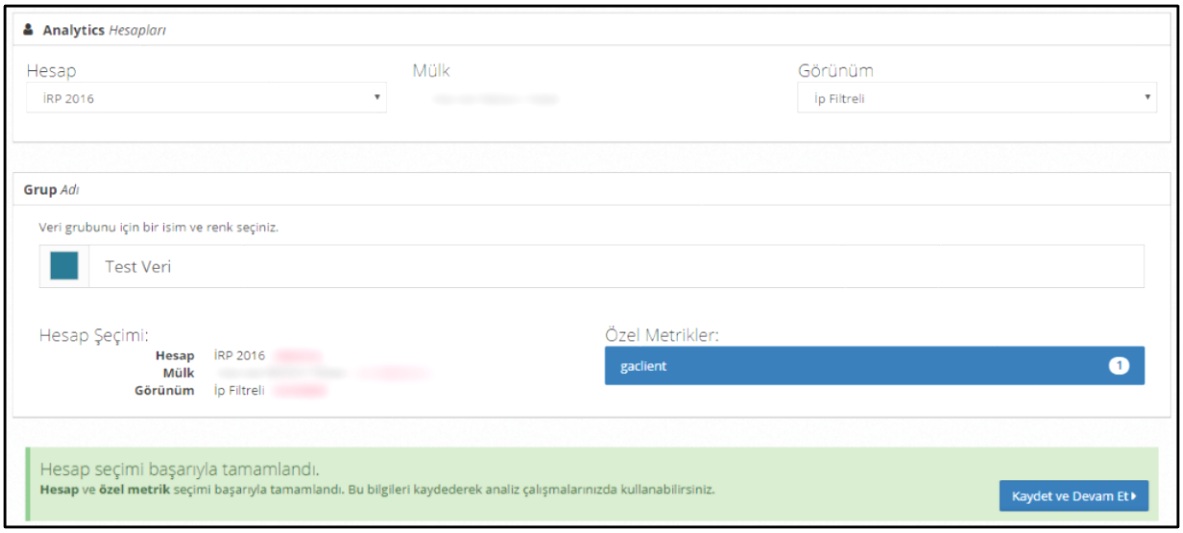
Analytics hesapları arayüzü

Analytics hesapları ekranı sisteme eklenen Google hesabının erişim sahibi olduğu Google Analytics hesaplarının sisteme tanımlanması ve bu tanımlı hesapların listelenmesini sağlamaktadır. Ekran yerleşimi Şekil 5.13'de gösterilmektedir.

Şekil 5.13. Hesaplar arayüzü

Yeni analytics hesabı arayüzü

Hesaplar arayüzünde yer alan "Yeni Hesap Ekle" düğmesi ya da menü ile ulaşılabilen yeni hesap ekleme arayüzü, sisteme yeni hesaplar, mülkler ve görünümler tanımlanmasını sağlamaktadır. Bu arayüz ile sistemde erişim sağlanmış Google Hesabına bağlı, Google Analytics hesap, mülk ve görünümlerine erişilebilmekte ve bu hesap sisteme tanımlanmaktadır. Bu arayüz Şekil 5.14'de aşağıda gösterilmektedir.

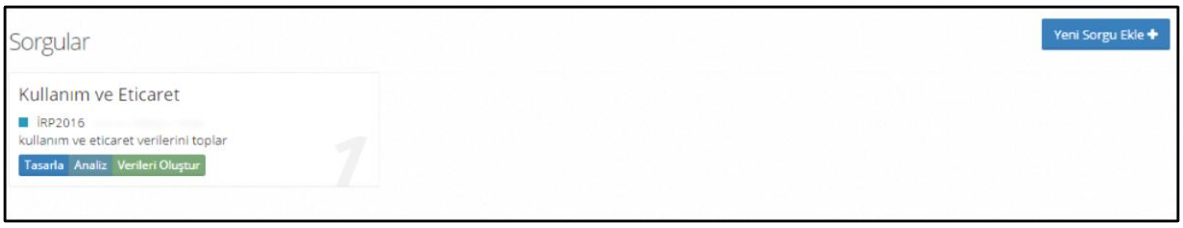


Şekil 5.14. Yeni hesap ekleme arayüzü

Oluşturulan hesaplara, isim ve takip rengi bu ekran vasıtasıyla tanımlanmaktadır.

Sorgular arayüzü

Sistem içerisinde, Google Analytics üzerinden veri toplama amacıyla oluşturulan sorgular bu ekran vasıtasıyla yönetilmektedir. Sorgular arayüzü tasarımı Şekil 5.15'de gösterilmektedir.



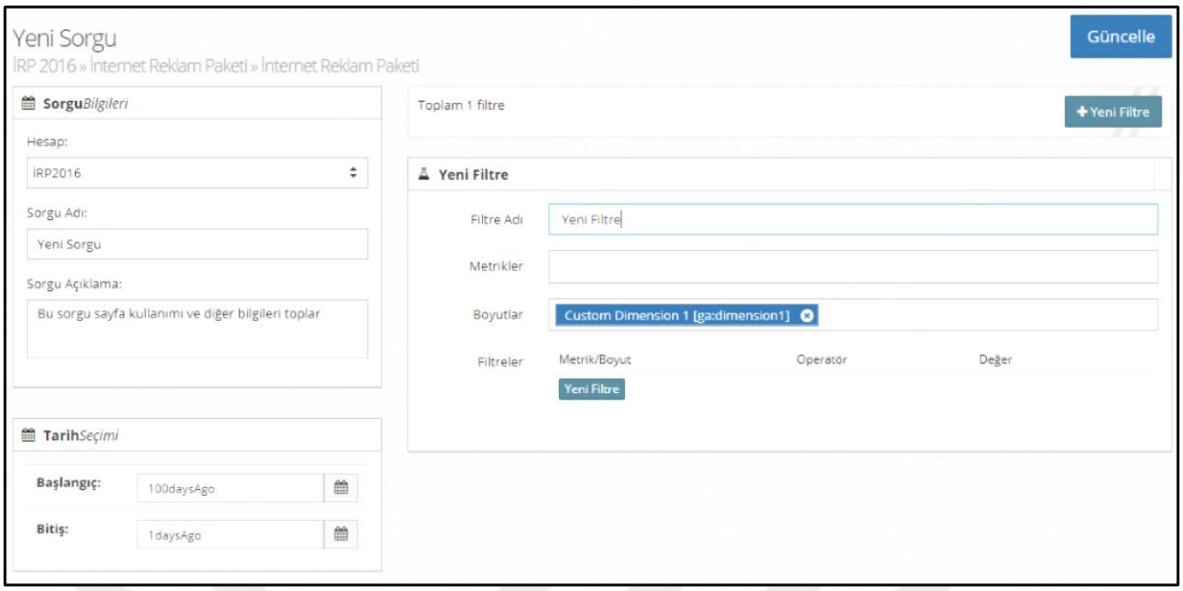
Şekil 5.15. Sorgular arayüzü

Sorgular arayüzünde sistemde mevcut bulunan sorgu tarifleri yer almakta ve bu tarifler üzerinden işlem yapılabilmektedir. Bu arayüzde her bir sorgu için üç işlem bulunmaktadır. Bu işlemler şunlardır:

- Tasarla: Bu işlem sorgu üretme ve düzenleme ekranını açarak, mevcut sorgunun yapısının ve veri tarihlerinin değiştirilmesini sağlamaktadır.
- Analiz: Analiz sorgu için Analytics sunucularından toplanmış verilerin basit istatiksel analizlerini göstermektedir.
- Verileri Oluştur: Verileri oluştur işlemi Sistem Mimarisi kısmında yer alan "2.3 Analytics Sorgularını API'ye Gönder Süreci", "2.4 Verileri API Üzerinden Topla Süreci" ve "2.5 Veri Birleştirme ve Normalizasyon Süreci" süreçlerini gerçekleştirmektedir. Bu aksiyon kullanıldığında sunucu yazılımı gerekli sorguları ve hesap erişimi bilgilerini bir araya getirerek Analytics Api üzerinden veri toplayarak sunucu yazılımı içerisine kaydetmektedir.

Sorgu tasarım arayüzü

Analytics Api üzerinden veri toplanabilmesi için Analytics sorgularının hazırlanması gerekmektedir. Bu işlem sorgu tasarım arayüzü vasıtasıyla yapılmaktadır. Sorgu tasarım arayüzü Google Analytics bölümünde tanımlanan sorgulama yapısına göre gerekli sorgu yapısını oluşturmaktadır. Tasarım ekranı görüntüsü Şekil 5.16'de gösterilmektedir.



Şekil 5.16. Sorgu tasarım ekranı

Bu arayüz sistemin en önemli arayüzlerinden birisidir. Bu arayüz ile oluşturulan sorgular ile, herhangi bir yazılım dili ya da algoritması kullanmadan, büyük bir kolaylıkla Analytics sunucularından toplanmak istenen veriler için tasarım yapılabilmektedir. Bu arayüz içerisinde üç temel bölüm bulundurmaktadır:

- Sorgu Bilgileri: Sorgu bilgileri kutusu ile sorgunun çalışacağı Analytics hesabı seçilebilmekte ve sorgu adı ve açıklaması belirlenebilmektedir.
- Tarih Seçimi: Google Analytics' in veri kaynağı olarak kullanımında önemli faydalardan biri de sunucu kayıtlarındaki gibi tarih verilerini yorumlamadan farklı tarihler için sorguların çalıştırılabilmesidir. Bu alanda yer alan seçim aracı ile aylar, yıllar gibi tarih aralığı seçimi yapılabilmektedir. Ayrıca yine Google Analytics veri yeteneklerinin kullanılması ile son 100 gün ya da son 20 günden önceki 50 gün gibi daha ayrıntılı analiz süreleri kolayca belirlenebilmektedir.
- Filtreler: Filtreler alanı ile Google Analytics içerisinde yer alan metrik ve boyutlar seçilerek, farlı veri kombinasyonları hiç kod yazmadan oluşturulabilmektedir. Ayrıca bunlar içinde kolaylıkla filtreleme yapılabilmektedir. Örneğin sadece mobil cihazlar, sadece Türkiye, hatta sadece Ankara gibi filtreler tanımlanabilmektedir.

Kayıtlı modeller arayüzü

Uygulama içerisinde toplanan veriler için birçok farklı modelleme ve madencilik yürütülmesi mümkündür. Model planlama ve listeleme işlemleri kayıtlı modeller arayüzü vasıtasıyla yapılmaktadır. Şekil 5.17'de modeller ekranı gösterilmektedir.

Şekil 5.17 Kayıtlı modeller arayüzü

Yeni bir model, hesap, sorgu ve modelleme algoritması seçerek planlanmaktadır. Mevcut modellerin listelendiği sayfada yer alan incele işlemi ile modelleme arayüzüne erişilmektedir.

Modelleme arayüzü

Modelleme arayüzü önemli arayüzlerden bir diğeridir. Modelleme arayüzü ile farklı parametreler ile toplanan verilerin modellenmesi sağlanmakta, veriler modele uygun olarak sunucu tarafında yeniden ayarlanmaktadır. Şu an sistem içerisinde yer alan K-Means kümeleme algoritmasının çalışmasını ve modelleme çıktılarını gösteren arayüz Şekil 5.18'de yer almaktadır.

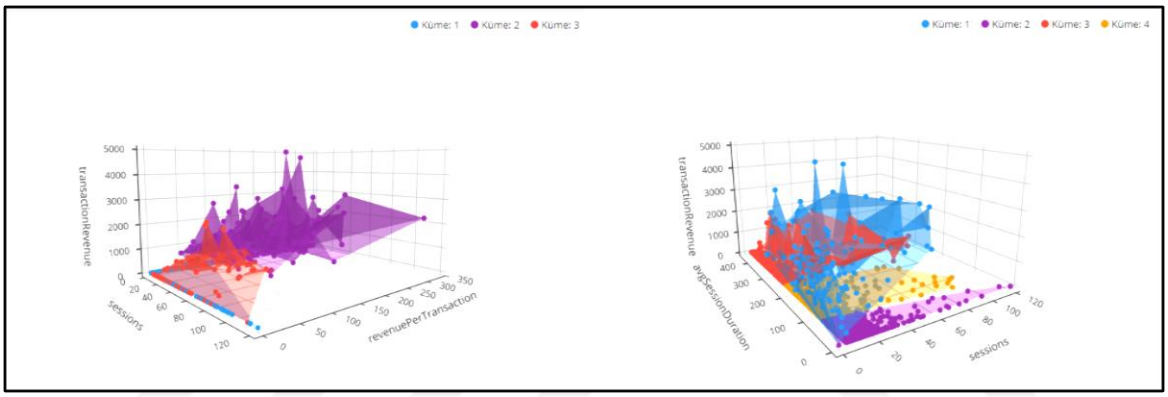
Şekil 5.18. Modelleme arayüzü

Bu arayüz ile sadece küme parametresi değiştirilerek, sunucu yazılımının tüm modeli tekrar kurarak yorumlaması sağlanabilmektedir. Bu arayüz şu bileşenlerden oluşmaktadır.

- Parametre Seçimi: En üstte yer alan parametre alanında ilgili veri madenciliği algoritması için gerekli olan parametreler seçilmektedir. K-Means algoritması için küme sayısı seçilmektedir.
- Model Özeti: Bu alanda model ile erişilen sonuçlara dair özet bilgilere ulaşılmaktadır. K-Means algoritması için seçilen veri tipleri için küme ortalama ve toplam değerleri, seçilen hedef için en iyi kümeyi göstermektedir. Seçilen kümeye takipçi eklenebilmesi için gereken aksiyon bu alanda yer almaktadır.

- Grafikler Alanı: Model sonuçlarının grafikler vasıtasıyla görselleştirildiği alandır. K-Means algoritması için kullanılan veri tiplerinden seçilen üç tanesinin dağılımları 2 ve 3 boyutlu grafikler olarak verilmekte ve görsel olarak incelenmesi sağlanmaktadır.
- Veri Özeti: Modelleme sürecinde kullanılan verinin özetinin yer aldığı, en küçük, en büyük, toplam ve ortalama gibi istatistiki değerlerinin gösterildiği bölümdür.

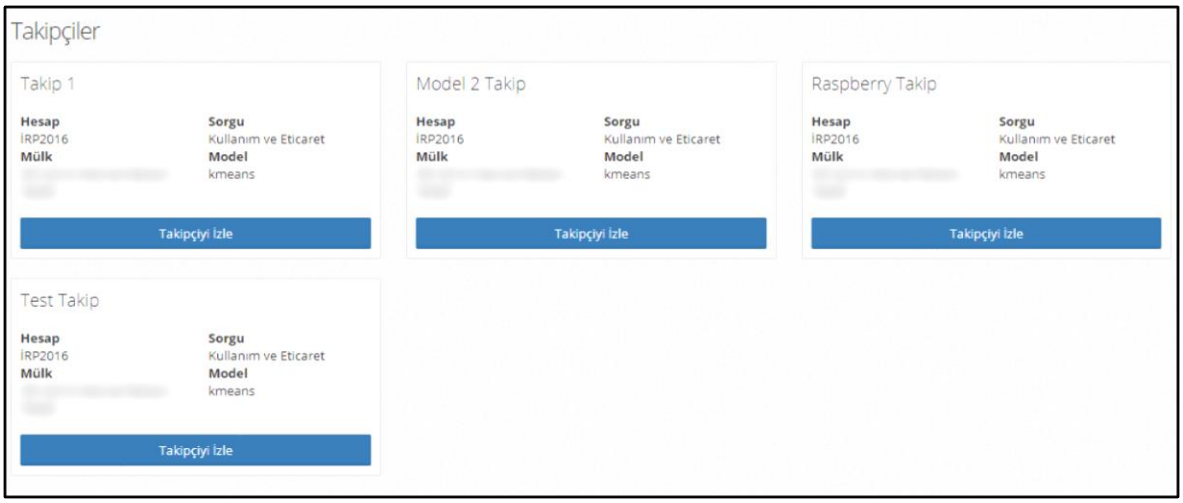
Bu arayüzde yer alan grafikler ile analistin kolayca sonuçları yorumlaması sağlanabilmektedir. Örneğin Şekil 5.19'da K-Means algoritması ile 3 ve 4 kümeye bölünen verilerin görselleştirilmiş grafikleri yer almaktadır.



Şekil 5.19. K-Means ile 3 ve 4 kümeye bölünmüş veri görselleştirmesi

Takipçiler arayüzü

Geliştirilen uygulama, ziyaretçiler hakkında analiz yapıp bu ziyaretçileri değerlemek için kaydedilen modelleri takip etmektedir. Takipçiler arayüzü sistem içerisine tanımlanmış takipçi kodların listelendiği arayüzdür. Takipçiler arayüzü Şekil 5.20'de gösterilmektedir.



Şekil 5.20. Takipçiler arayüzü

Bu arayüzde her bir takipçi için yer alan "Takipçiyi İzle" düğmesi ile takipçi izleme arayüzüne ulaşılmaktadır.

Takipçi izleme arayüzü

Sistemde kaydedilen takipçilerin izlenmesi ve takip edilen sayfalara gelen kullanıcıların değerlendirmelerinin izlendiği arayüzdür. İlgili arayüz Şekil 5.21'de gösterilmektedir.

Şekil 5.21. Takip arayüzü

Bu arayüz iki bölümden oluşmaktadır. Özet kısmında ilgili takipçi de kullanılan model ve hesap bilgileri gösterilmektedir.

Aktif ziyaretler bölümü ise canlı olarak takipçi kodu eklenen siteyi ziyaret eden kullanıcıları göstermektedir. Örneğin Şekil 5.18 de görünen ekranda "user-1" olarak işaret edilen kullanıcı K-Means kümeleme algoritması ile kurulan modelde istenen 3 numaralı kümeye girerek siteyi ziyaret ederken "visitor-1539850917" ile işaret edilen ziyaretçi bu değerlendirme neticesinde 2 numaralı kümede yer almaktadır.

6. SONUÇ

Yapılan bu tez çalışması ile, Web Kullanım Madenciliğinde yaşanan ön işleme sorunları ve bu sorunların, analiz çalışmalarına yansıtabileceği sonuçlar incelenmiştir. Geliştirilen uygulama ile Google Analytics veri kaynağından ara katmanda filtreleme ve kod yazım yetenekleri de kullanılarak elde edilen veriler, ön işleme sürecindeki zorlukları en aza indirerek yazılıma aktarılmış ve bu yazılım ile anlık önerilerin alınabildiği bir öneri sistemi geliştirilmiştir. Bu yöntem literatürde yer alan diğer çalışmalara göre ön işlemlerini kısaltarak, modelleme ve analiz süreçlerine öncelik verilmesini sağladığı için önemli bir geliştirmedir. Elde edilen bu sonuç ile, farklı veri kaynakları kullanımının web kullanım madenciliğine sağlayabileceği yeni imkanlar gösterilebilmiştir.

Mevcut veri toplama teknikleri genel olarak web sunucu erişim kayırlarından oluşmakta ve bu verilerin işlenmesi oldukça zaman ve emek harcanmasına sebebiyet vermektedir. Ayrıca bu kayıtlardan elde edilen veri miktarı oldukça kısıtlı olarak oluşmakta ve yeterli bir analize imkân vermemektedir. Önerilen veri kaynağı, mevcut araştırma yöntemlerinin erişemediği birçok veriyi de veri toplama ve veri madenciliği süreçlerine dahil edebildiği için önceki çalışmalardan daha geniş bir veri kaynağına sahip çalışma gerçekleştirilebilmesini sağlamaktadır.

Geliştirilen uygulama ile elde edilen veri hazırlama hızı ve analiz yeteneğinin, modelleme süreçlerinde de daha iyi sonuçlar elde edilebilmesini sağlayacağı düşünülmektedir. Elde edilen daha geniş veri yapısı ile, literatürde yapılan analizlerin gerçekleştirilmesi ve bu yöntemlerde de ilerlenme sağlanması imkânı artmaktadır. Mevcut yazılım içerisinde yer alan modelleme katmanının geliştirilerek daha faydalı sonuçlar alınması da mümkündür.

KAYNAKLAR

İnternet: angular.io. Angular Features and Benefits. URL: https://angular.io/features, Son Erişim Tarihi: 20.05.2019

Arlitt, M. J. A. S. P. E. R. (2000). Characterizing web user sessions. *28*(2), 50-63.

Azitha, R., Surjandari, I., and Laoh, E. (2018). *Mining Information Search Pattern on Website: A Case Study of Educational Institution. 2018 5th International Conference on Information Science and Control Engineering (ICISCE)*.

İnternet: Brown, P. State of the Union: npm. URL: https://www.linux.com/news/event/Nodejs/2016/state-union-npm, Son Erişim Tarihi: 20.05.2019

Castellano, G., Fanelli, A., and Torsello, M. (2007). *Log data preparation for mining web usage patterns. IADIS International Conference Applied Computing*.

Chitraa, V., and Davamani, D. S. A. (2010). A survey on preprocessing methods for web usage data.

Cooley, R., Mobasher, B., and Srivastava, J. (1999). Data Preparation for Mining World Wide Web Browsing Patterns. *Knowledge Information Systems, 1*(1), 5-32.

Deshpande, D., Deshpande, S., and Thakare, V. (2018). *Web User Identification: Analysis of Heuristic Solutions. 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*.

Domenech, J. M., and Lorenzo, J. (2007). *A tool for web usage mining. International Conference on Intelligent Data Engineering and Automated Learning*.

Dünya Bankası. (2019a). *Individuals using the Internet (% of population)*. Retrieved from: https://data.worldbank.org/indicator/IT.NET.USER.ZS

Dünya Bankası. (2019b). *Secure Internet servers*. Retrieved from: https://data.worldbank.org/indicator/IT.NET.SECR

Ganibardi, A., and Ali, C. A. (2018). *A Genetics Clustering-Based Approach for Weblog Data Cleaning. 2018 Sixth International Conference on Enterprise Systems (ES)*.

Hao, S., Zhaoxiang, S., and Bingbing, Z. (2017). *A user clustering algorithm on web usage mining. 2017 First International Conference on Electronics Instrumentation and Information Systems (EIIS)*.

Herath, D., and Jayaratne, L. (2017). *A Personalized Web Content Recommendation System for E-Learners in E-Learning Environment. 2017 National Information Technology Conference (NITC)*.

İnternet: Internet Live Stats. Total number of Websites. URL: https://www.internetlivestats.com/total-number-of-websites/, Son Erişim Tarihi: 15.05.2019

Ivancsy, R., Juhasz, S. J. W. A. o. S., Engineering, and Technology. (2007). Analysis of web user identification methods. *2*(3), 212-219.

İnternet: Jackson, B. Web Crawlers and User-Agents - Top 10 Most Popular. URL: https://www.keycdn.com/blog/web-crawlers, Son Erişim Tarihi: 22.05.2019

Jain, S., Rawat, R., and Bhandari, B. (2017). *A Survey Paper on Techniques and Applications of Web Usage Mining. 2017 International Conference on Emerging Trends in Computing and Communication Technologies (ICETCCT)*.

Jayamalini, K., and Ponnavaikko, M. (2017). *Research on Web Data Mining Concepts, Techniques and Applications. 2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET)*.

Kamakshi, Mehrotra, D., and Deep, V. (2018). *Spatial Mining of Web-Log to Observe Reachability of Website*

Kapusta, J., Munk, M., and Halvoník, D. (2017). *Quality of Extracted Sequential Rules by Session Identification Using STT and Cookies. 2017 European Conference on Electrical Engineering and Computer Science (EECS)*.

Kumar, M. (2017). *Analysis of Visitor's Behavior from Web Log Using Web Log Expert Tool. 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*.

Kundu, S., and Garg, L. (2017). *Web Log Analyzer Tools: A Comparative Study to Analyze User Behavior. 2017 7th International Conference on Cloud Computing, Data Science and Engineering-Confluence*.

Liu, B. (2011). *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*: Springer Science and Business Media

Minaei-Bidgoli, B., and Punch, W. F. (2003). *Using genetic algorithms for data mining optimization in an educational web-based system. Genetic and evolutionary computation conference*.

İnternet: mongodb.com. What Is MongoDB*. URL: https://www.mongodb.com/what-is-mongodb, Son Erişim Tarihi: 20.05.2019

Musale, V., and Chaudhari, D. (2017). *Web Usage Mining Tool By Integrating Sequential Pattern Mining with Graph Theory. 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM)*.

Nath, K., Dhar, S., and Basishtha, S. (2014). *Web 1.0 to Web 3.0-Evolution of the Web and its various challenges. 2014 International Conference on Reliability Optimization and Information Technology (ICROIT)*.

İnternet: npmjs.com. express - npm. URL: https://www.npmjs.com/package/express, Son Erişim Tarihi: 20.05.2019

Özocak, G. J. B. (2012). DDoS Saldırısı ve Failin Cezai Sorumluluğu. *28*, 23.

Rao, V. M., and Kumari, V. V. (2014). An Extensive Literature Survey on Comprehensive Research Activities of Web Usage Mining. *International Journal of Computer Engineering and Technology, 5*(12), 275-289.

Ronallo, J. J. C. L. J. (2012). HTML5 Microdata and Schema. org. *16*.

Saste, A., Bedekar, M., and Kosamkar, P. (2017). *Predicting Demographic Attributes From Web Usage: Purpose and Methodologies. 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*.

Sathya, M., and Devi, P. I. (2017). *Apriori Algorithm on Web Logs for Mining Frequent Link. 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*.

Singh, S. P. (2017). *Analysis of Web Site Using Web Log Expert Tool Based on Web Data Mining. 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*.

Sukumar, P., Robert, L., and Yuvaraj, S. (2016). *Review On Modern Data Preprocessing Techniques in Web Usage Mining (WUM). 2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*.

Tan, P.-N., and Kumar, V. (2004). Discovery of web robot sessions based on their navigational patterns. In *Intelligent Technologies for Information Analysis* (pp. 193-222): Springer.

Varnagar, C. R., Madhak, N. N., Kodinariya, T. M., and Rathod, J. N. (2013). *Web usage mining: a review on process, methods and techniques. 2013 International Conference on Information Communication and Embedded Systems (ICICES)*.

W3C. (1995). Logging Control In W3C httpd. In *The Common Logfile Format*.

İnternet: Worldometers. Worldometers - Gerçek Zamanlı Dünya İstatistikleri. URL: https://www.worldometers.info/tr/, Son Erişim Tarihi: 15.05.2019

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, adı	:	Şayan, Şükrü Can
Uyruğu	:	T.C.
Doğum yeri ve tarihi	:	11/07/1988 Ankara
Medeni hali	:	Bekar
Telefon	:	0 (312) 497 76 01
Fax	:	
e-posta	:	sukrucansayan@gazi.edu.tr

Eğitim Derecesi	**Okul/Program**	**Mezuniyet Yılı**
Lisans	Gazi Üniversitesi, Endüstri Mühendisliği Bölümü	2010
Lise	Etimesgut Anadolu Lisesi	2006

İş Deneyimi, Yıl	**Çalıştığı Yer**	**Görev**
2014-devam ediyor	Hazine ve Maliye Bakanlığı	Uzman Yardımcısı
2012-2014	Sistemim Bilişim Çözümleri	Bilişim Uzmanı
2011-2011	e-PTT AVM	Yazılım Geliştirici
2010-2011	Frappe Ajans	Yazılım Geliştirici

Hobiler

Web, Bilgisayar teknolojileri, yazılım, iot, blockchain

GAZİLİ OLMAK AYRICALIKTIR…