T.C.

SELÇUK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

VERİ MADENCİLİĞİNİN WEB TABANLI UYGULAMALARDA İNSAN UYUMLULUKLARININ TESBİTİ ÜZERİNE BİR ÇALIŞMA

048229002005

Ayşe ONAT

YÜKSEK LİSANS TEZİ

BİLGİSAYAR MÜHENDİSLİĞİ ANA BİLİM DALI

Danışman: Prof. Dr. Ferruh YILDIZ

Konya 2008

T.C

SELÇUK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

VERİ MADENCİLİĞİNİN WEB TABANLI UYGULAMALARDA İNSAN UYUMLULUKLARININ TESBİTİ ÜZERİNE BİR ÇALIŞMA

Ayşe ONAT

YÜKSEK LİSANS TEZİ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

Konya, 2008

T.C.

SELÇUK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

VERİ MADENCİLİĞİNİN WEB TABANLI UYGULAMALARDA İNSAN UYUMLULUKLARININ TESBİTİ ÜZERİNE BİR ÇALIŞMA

Ayşe ONAT

YÜKSEK LİSANS SEMİNERİ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

Bu tez 29/02/2008 tarihinde aşağıdaki jüri tarafından oybirliği ile kabul edilmiştir.

Prof.Dr.FerruhYILDIZ	Prof.Dr. Şirzat KAHRAMANLI	Yrd.Doç.Dr.Mesut GÜNDÜZ
(Danışman)	(Üye)	(Üye)

ÖZET

Yüksek Lisans Tezi

VERİ MADENCİLİĞİNİN WEB TABANLI UYGULAMALARDA İNSAN UYUMLULUKLARININ TESBİTİ ÜZERİNE BİR ÇALIŞMA

Ayşe ONAT

Selçuk Üniversitesi Fen Bilimleri Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı

Danışman : Prof. Dr. Ferruh YILDIZ

2008, 57 Sayfa

Jüri :

Prof.Dr.FerruhYILDIZ
Prof.Dr. Şirzat KAHRAMANLI
Yrd.Doç.Dr. Mesut GÜNDÜZ

İnternetin her alanda giderek yaygınlaşması ve her kapsamda bilgi kaynağı olması günlük yaşantımızın büyük bir kısmını etkilemesine sebep olmuştur. Haber, ekonomi, kültür, eğitim, sağlık hizmetleri ve reklam gibi bir çok alanda bilgi kaynağı olan İnternet ortamında, kullanıcıların kendilerine ait gerekli ve yeterli verileri bulmasında zorluklar yaşanmaktadır. Bu sorunları çözmek için web madenciliği sayesinde gün geçtikçe daha iyi çözümler geliştirilmektedir.

Web madenciliği sayesinde müşterilerin ilgi alanları, ürün pazarlama stratejileri oluşturma, reklam alma, insanların birbirleriyle olan ilişkilerini gözden geçirme, insanların hangi sayfalarda daha çok gezdiğini belirleyerek o sayfalarda daha farklı sunumlar oluşturabilme gibi hususlarda kurumlara yardımcı olunur.

Bu çalışmada, web madenciliği yöntemi kullanılarak internet üzerinden insanların yaş, cinsiyet, yaşadığı yer, lisans düzeyi gibi özelliklerine bakılarak insanların birbirleriyle olan uyumluluklarının bulunması amaçlanmıştır. Bunun için veri madenciliğinde kullanılan Apriori Algoritması uygulanmıştır.

Anahtar Kelimeler : Veri Madenciliği, Web Madenciliği, Apriori Algoritması

ABSTRACT

Master Thesis

A STUDY ON DETERMINATION OF HUMAN SYNCRONIZATION OF DATA MINING IN WEB BASED APPLICATIONS

Ayşe ONAT

Graduate School of Natural and Applied Sciences
Department of Computer Engineering

Supervisor: Prof. Dr. Ferruh YILDIZ

2008, 57 Sayfa

Jury :

Prof.Dr.FerruhYILDIZ
Prof.Dr. Şirzat KAHRAMANLI
Yrd.Doç.Dr. Mesut GÜNDÜZ

Nowadays internet has been spreaded at every area and has become a knowledge source and so it affects our life in a positive way. The persons have some difficulties in finding necessary and enough information for users of the internet which is useful for news, economy, culture, education, health services and advertisements. To solve these problems by the means of web mining is developed better solutions by the day.

Web mining is help assistance to corporations about subject which are interest of customers, marketing product strategies, taking advertisements, looking over relations between people and each other, finding which pages are clicked too much and constituting different offers.

In this study, it has been intended to find the people harmony with the each other with using some attributes which are age, sex, the place which she lived and licence level. For this study Apriori Algorithm is applied which has used at web mining.

Keywords: Web Mining, Web Mining, Apriori Algorithm.

İÇİNDEKİLER

1. GİRİŞ 1
2. VERİ MADENCİLİĞİ VE KULLANILAN TEKNİKLER 4
2.1. Giriş 4
2.2.Veri Madenciliği 4
2.3. Veri Madenciliği ile Yapılabilecekler 7
2.4. Veri Madenciliği Modelleri 8
2.5. Veri Madenciliğinde Kullanılan Yöntemler 9
2.5.1. Sınıflama ve Regresyon Modelleri 9
2.5.2. Kümeleme Modelleri 10
2.5.3. Birliktelik Kuralları ve Ardışık Zamanlı Örüntüler 10
2.5.4. Bellek Tabanlı Yöntemler 10
2.5.5. Yapay Sinir Ağları(YSA) 11
2.5.6. Karar Ağaçları 11
2.6. Veri Madenciliği Süreci 12
2.6.1. Problemin Tanımlanması 12
2.6.2. Verilerin Hazırlanması 13
2.6.3. Modelin Kurulması ve Değerlendirilmesi 13
2.6.4. Modelin Kullanılması 13
2.6.5. Modelin İzlenmesi 14
2.7. Veri Madenciliği ve Diğer Disiplinler 14
3. WEB MADENCİLİĞİ 16
3.1. Web Madenciliğine Genel Bakış 16
3.1.1. Web İçerik Madenciliği 17
3.1.2. Web Yapı Madenciliği 18
3.1.3. Web Kullanım Madenciliği 19
4. BİRLİKTELİK KURALLARI 22
4.1. Birliktelik Kurallarında Kullanılan Değerler 22
4.2. Apriori Algoritması 22
4.2.1. Birleştirme Adımı 23
4.2.2. Budama Adımı 24

4.3. Apriori Algoritması Pseudocode 28
4.4. Olağan Elemankümelerden İlişkisel Kurallar Oluşturma 30
4.5. Apriori Algoritmasının Verimliliğini Artırma 31
4.5.1. Karmaşık Tabanlı Teknik 31
4.5.2. İşlem Azaltma 32
4.5.3. Bölme 32
4.5.4. Modelleme 33
4.5.5. Dinamik Elemanküme Sayımı 33
4.6. Aday Jenerasyonları Kullanmadan Olağan Elemankümelerin Çıkarımı 34
5. WEB TABANLI UYGULAMALARDA İNSAN UYUMLULUKLARININ TESBİTİ 35
5.1 Veritabanının Oluşturulması 36
5.2. Proje İçin Gerekli Formların Oluşturulması 39
5.2.1. Üye Kayıt Formu 39
5.2.2. Üye Giriş Formu 42
5.2.3. Profil Formu 42
5.2.4. Arama Formu 44
5.2.5. Mesaj Listeleme Formu 45
5.2.6. Mesaj Okuma Formu 46
5.2.7. Mesaj Gönderme Formu 47
5.2.8. Yorumlar Formu 48
5.2.9. Favoriler Formu 48
5.3. Apriori Algoritmasının Uygulanması 49
6. SONUÇLAR 54
7. KAYNAKLAR 56

1. GİRİŞ

Günümüz kuruluşlarında başarılı olabilmek için bilgi yönetimi oldukça önemli bir hale gelmiştir. İşletmelerdeki bilgi miktarı artmış ve veri tabanları da çok büyük boyutlara ulaşmıştır. Bu gelişmeler sonucunda bugün doğru ve detaylı bilgiye daha çabuk ulaşılabilmektedir. Fakat bu bilgi yığını bazı problemleri de beraberinde getirmiştir. Veri tabanı sistemlerinin artan kullanımı ve hacimlerindeki bu olağanüstü artış, organizasyonları elde toplanan bu verilerden nasıl faydalanılabileceği problemiyle karşı karşıya bırakmıştır. Bu büyük veri yığınlarının yönetilmesi, anlamlı hale getirilmesi, karar mekanizmalarında kullanılabilmesi, bilgileri ne şekilde değerlendirebileceğimiz ancak veri madenciliği yolu ile mümkün olabilmektedir.

Veri madenciliği büyük miktardaki veriden anlamlı bilginin çıkarılması ve bunlarla ilgili bir karar verme tekniği olup pazarlama, bankacılık, internet, ekonomi, kültür, sağlık hizmetleri, sigortacılık ve tıp sektörü başta olmak üzere birçok sektörde etkin şekilde kullanılmaktadır. Veri madenciliği yöntemlerinden biri de web madenciliğidir. Web madenciliği veri madenciliği tekniklerinin World Wide Web verileri üzerinde uygulanmasını konu alır.

Bu çalışmada, veritabanlarında bilgi keşfi süreçleri, veri madenciliği, web madenciliği ve Apriori algoritması hakkında bilgiler verilmiş, veri madenciliğinin en önemli kollarından biri olan web madenciliği yöntemiyle internet üzerinden insanların birbirleriyle olan uyumluluklarının bulunması amaçlanmıştır. İnsanların yaş, cinsiyet, ekonomik durum, lisans düzeyi gibi özelliklerine web madenciliği uygulanacaktır. İnsan uyumlulukları bulunurken veri madenciliğinde sıkça kullanılan Apriori algoritması uygulanmıştır.

Apriori algoritması, veri madenciliğinde sık geçen öğelerin keşfedilmesi için kullanılan en çok bilinen birliktelik-ilişki kuralı algoritmasıdır.

Apriori algoritmasını uygulayabilmek için Asp dilinde yazılmış web tabanlı bir uygulama yazılımı geliştirilmiştir. Uygulama yazılımı ile veritabanındaki veriler üzerinde Apriori algoritması uygulanmış, uygulama yazılımının çalışması esnasında algoritmanın her aşaması izlenmiş, girdiler ve sonuçlar adım adım gözlenmiştir.

Web verilerinden sıralı örüntülerin bulunması, ilginç kullanıcı bilgilerinin çıkarılması gibi birçok çalışma geçmiş yıllarda yapılmış ve farklı yaklaşımlar sunulmuştur. Bunlardan bazıları;

- Uğuz H. yaptıkları çalışmada, web sunucusunun kayıtlarını kullanarak web kullanım madenciliği yöntemiyle web sayfası ziyaretçilerinin en sık eriştiği sayfa çiftlerini, üniversite içi ve dışı kullanıcı erişim dağılımı gibi tanımsal ilişkileri tespit etmişlerdir (Uğuz H., 2003).
- Chen ve Syncara geliştirdikleri Web Mate adlı sistemlerinde, web sayfalarını inceleyerek, web içeriğinden kullanıcı ilgilerini belirlemeyi sağlamışlardır (Chen L.,1998). Böylece web üzerinden arama işlemlerinde kolaylık sağlamışlardır.
- Şakiroğlu A.M. yaptıkları çalışmasında, web erişim kayıt dosyalarından genetik algoritma yöntemiyle sıralı erişimleri tespit etmiştir (Şakiroğlu A.M, 2003).
- İşeri tarafından yapılan tez çalışmasında, geliştirdiği yazılım ile web günlüğünden zaman sınırlı bulanık bağıntı kuralları ve sıralı örüntülerin çıkarılmasını sağlamıştır (İşeri İ., 2005).
- Web madenciliğinin kullanımı özellikle alışveriş sitelerindeki karlılık oranını arttırmaktadır. Bugün birçok alışveriş sitesi bu yöntemi kullanarak müşterilerine daha fazla miktarda ürün satmaya çalışmaktadır. Bunun en bilinen ve en iyi örneği http://www.amazon.com/ sitesidir. Sistemde yeni üye olanlara üyelik bilgilerinden başka sorularda yöneltilir. Müşterinin karşısına belli sayıda ürün çıkarılarak bunların arasından hoşuna gidenleri ya da almak istediklerini seçmesi istenir; bu sayede kişinin istekleri ve ilgileri ortaya

çıkarılır. Amazon.com bu bilgileri kullanarak müşterilerine kişiye özel sepetler hazırlar. Kullanıcının daha önce ilgi gösterdiği ürünler ışığında kullanıcı siteye tekrar giriş yaptıktan sonra öneriler sunulmakta, ilgi göstermesi muhtemel ürünler sayfadaki çeşitli linklerle ön plana çıkarılmaktadır.

- İskandinav ülkelerinin portalı olan Jubii internet sitesi de web madenciliği yöntemini kullanmaktadır. http://www.jubii.dk/ sitesinin sayfalarındaki banner 'larının yerlerini ve içeriğini optimize edebilmek için müşteri profillerini ve davranışlarını belirlemişlerdir. Sayfalar yeni müşteri profillerine göre çalışmaya başladığında verilen reklamlara tıklama oranı % 30 'dan % 50 'ye çıkmış ve böylece reklam verenler verdikleri reklamlar sonucunda daha fazla ziyaretçi almaya başladıklarından Jubii 'nin kârı artmıştır. Ziyaretçi davranış modelleri Jubii'ye reklam gelirlerini artırma yönünde imkan sağlamıştır.

Bu çalışmayı bugüne kadar yapılan çalışmalardan ayıran unsur, kullanılacak olan verilerin elde ediliş yöntemlerinin farklı olmasıdır. Diğer çalışmalarda kullanıcıların girdikleri sayfalar hafızada tutulup o sayfalar doğrultusunda işlem yapılırken bu projede arama sırasında kullanılıp elde edilen verilere göre işlem yapılacaktır. Yani kullanıcı arama yaptığı anda arama için kullandığı özellikler veritabanına kaydedilerek bu veriler üzerinde Apriori Algoritması uygulanacaktır.

2. VERİ MADENCİLİĞİ VE KULLANILAN TEKNİKLER

2.1. Giriş

Günümüzde bilgisayar sistemleri her geçen gün hem gelişmekte hem ucuzlamakta hem de kullanımı gün geçtikçe yaygınlaşmaktadır. Bu gelişmeyle birlikte işletmelerde üretilen sayısal bilgi miktarı da artmaktadır. Şirketlerin bilgi sitemleri üzerinden ürettiği bilgi miktarının büyük artış gösterdiği ve firmaların veri tabanlarının boyutlarının çok fazla arttığı görülmektedir. Bunun yanı sıra veritabanları da daha fazla veriyi saklayabilecek şekilde geliştirilmektedirler. Bilgisayar sistemlerindeki gelişmeyle veriye ulaşım daha da kolaylaştırılmıştır. Fakat her ne kadar bilgisayar sistemleri ve veritabanları gelişip kullanımı da artsa, büyük veri yığınlarının yönetilmesi ve anlamlı hale getirilip yorumlanması sorunu ortaya çıkmıştır.

Özellikle veri tabanlarının bilgiyi sadece saklamak için dizayn edildiği düşünüldüğünde bilgisayar sistemleri ile üretilen bu veriler tek başlarına değersizdirler, yani herhangi bir anlam ifade etmezler. Bu veriler belli bir amaç doğrultusunda işlendiği zaman anlamlı hale gelebilmektedir. İşte ham veriyi anlamlı hale dönüştürme işi veri madenciliği ile yapılabilir.

2.2.Veri Madenciliği

Veri madenciliği; önceden bilinmeyen, geçerli ve uygulanabilir bilginin veri yığınlarından dinamik bir süreç ile elde edilmesi olarak tanımlanabilir. Bu süreçte kümeleme, veri özetleme sınıflama kurallarının öğrenilmesi, bağımlılık ağlarının bulunması, değişkenlik analizi ve anomali tespiti gibi farklı birçok teknik kullanılmaktadır.(http://www.bilgiyonetimi.org/cm/pages/mkl_gos.php?nt=538)

Veri madenciliği sayesinde büyük veri yığınlarından oluşan veritabanı sistemleri içerisinde bulunan veriler arasından gizli kalmış verilerin

keşfedilmesi sağlanır. Bu işlem, istatistik, matematik disiplinleri, modelleme teknikleri, veritabanı teknolojisi ve çeşitli bilgisayar programları kullanılarak yapılır. Veri madenciliği, problemi çözmek için gerekli bilgileri elde etmeye yarayan bir araçtır. Veri madenciliği, veriler arasındaki ilişkileri, kuralları bulmayı sağlar.

Günümüzde kullanılan veri madenciliğinin başlıca ilgi alanları;

Pazarlama alanında;

- Müşteri segmentasyonunda,
- Çeşitli pazarlama kampanyalarında,
- Mevcut müşterilerin elde tutulması için geliştirilecek pazarlama stratejilerinin oluşturulmasında,
- Pazar sepeti analizinde,
- Çapraz satış analizlerinde,
- Müşteri değerlendirmede,
- Müşteri ilişkileri yönetiminde,
- Çeşitli müşteri analizlerinde,
- Satış tahminlerinde.

Bankacılık alanında;

- Farklı finansal göstergeler arasındaki gizli korelasyonların bulunmasında,
- Kredi kartı dolandırıcılıklarının tespitinde,
- Müşteri segmentasyonunda,
- Kredi taleplerinin değerlendirilmesinde,
- Usulsüzlük tespitinde,
- Risk analizlerinde,
- Risk yönetiminde.

Sigortacılık alanında;

- Yeni poliçe talep edecek müşterilerin tahmin edilmesinde,
- Sigorta dolandırıcılıklarının tespitinde,
- Riskli müşteri tipinin belirlenmesinde.

Perakendecilik alanında;

- Satış noktası veri analizlerinde,
- Alışveriş sepeti analizlerinde,
- Tedarik ve mağaza yerleşim optimizasyonunda.

Borsa alanında;

- Hisse senedi fiyat tahmininde,
- Genel piyasa analizlerinde,
- Alım-satım stratejilerinin optimizasyonunda.

Telekomünikasyon alanında;

- Kalite ve iyileştirme analizlerinde,
- Hisse tespitlerinde,
- Hatların yoğunluk tahminlerinde.

Sağlık ve ilaç alanında;

- Test sonuçlarının tahmininde,
- Ürün geliştirmede,
- Tıbbi teşhisde,
- Tedavi sürecinin belirlenmesinde,

Endüstri alanında;

- Kalite kontrol analizlerinde,
- Lojistik,

- Üretim süreçlerinin optimizasyonunda.

Bilim ve mühendislik alanında;

- Veriler üzerinde modeller kurarak bilimsel ve teknik problemlerin çözümlenmesinde kullanılmaktadır (http://www.bilgiyonetimi.org /cm/pages/mkl_gos.php?nt=538).

2.3. Veri Madenciliği ile Yapılabilecekler

Veri madenciliğinin asıl amacı veri yığınlarından anlamlı bilgiler elde edip bu sayede bazı kararlar alarak o kararlara göre hareket etmeyi sağlar. Bu özelliği sayesinde yaygın bir kullanım alanı vardır. Veri madenciliği ile yapılabilecekler;

- Bir işletmenin, ülkede meydana gelen ekonomik değişimler sonucunda bu değişimlerden etkilenme oranının bulunmasında veya meydana gelebilecek risklerin tespitinde kullanılabilir.
- Müşterilerin ürün talepleri göz önüne alınarak farklı satış olanakları yaratılarak firmanın satışını artırmada kullanılabilir.
- Firmanın teknik anlamdaki kaynaklarının en uygun seçeneklerde kullanılmasını sağlamak için kullanılabilir.
- Bir ürün veya hizmetle ilgili bir kampanya programı için hedef kitle belirleyip bunun hedef kitleye nasıl sunulacağı kararına kadar olan süreçte veri madenciliği kullanılabilir.
- Geçmişte olan durumların analizi yapılarak gelecekle ilgili tahmin yürütülebilir.
- Hizmet sırasında hangi özelliklerin müşteriyi etkilediği, hangi müşterinin neden bunları tercih ettiği ortaya çıkarılabilir.
- Piyasada meydana gelen değişikliklere müşterinin ne şekilde cevap vereceği ve bu cevabın firmayı nasıl etkileyeceği gibi işlemlerin tespitinde kullanılabilir.

- Bir işletmeye ait müşterinin başka bir işletmeye geçmesi sonucunda onun neden geçtiği ve onu tekrar kendi işletmesine geçirebilmek için gerekli olan stratejilerin belirlenmesinde kullanılabilir.
- En fazla kâr getiren müşteriler belirlenerek onlara özel promosyonlar yapılabilir.

2.4. Veri Madenciliği Modelleri

Veri madenciliğinde kullanılan modeller iki çeşittir. Bunlar;

- Tahmin edici modeller
- Tanımlayıcı modeller (Zhong N., 1999).

Tahmin edici modellerde; bağımlı ve bağımsız değişken adı altında 2 çeşit değişken bulunmaktadır. Eldeki var olan yani bilinen veriler bağımsız değişken olarak adlandırılırken, istenilen soruların cevapları ise bağımlı değişken olarak adlandırılmaktadır. Çünkü sorunun cevabını bulabilmek için bağımsız değişkenlere ihtiyaç duyulmaktadır. Örneğin bir banka vermiş olduğu kredi verilerini kendisine ait veritabanlarında saklamaktadır. Bu verilerde bağımsız değişkenler olarak adlandırdığımız değişkenler kredi alan müşterinin özellikleri, bağımlı değişkenler ise kredinin geri ödenip ödenmediğidir. Böylece herhangi bir kredi talebi olduğunda müşteri özelliklerine göre verilecek olan kredinin geri ödenip ödenmeyeceğinin tahmini yapılabilmektedir. Bunun için aşağıdaki işlemler sırasıyla yapılır;

- Sonuçları bilinen verilerden bir model geliştirilir.
- Oluşturulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuç değerleri tahmin edilir.

Tanımlayıcı modellerde; ise karar vermeye yardımcı olabilecek örnekler tanımlanmaktadır. 25 yaş altı bekar kişiler ile, 25 yaş üstü evli kişiler

arasındaki ödeme performanslarını gösteren bir analiz örnek olarak verilebilir.

2.5. Veri Madenciliğinde Kullanılan Yöntemler

Veri madenciliğinde kullanılan yöntemler;

- Sınıflama ve Regresyon,
- Kümeleme,
- Birliktelik Kuralları ve Ardışık Zamanlı Örüntüler,
- Bellek tabanlı yöntemler,
- Yapay sinir ağları,
- Karar ağaçları(http://www.bilgiyonetimi.org/cm/pages/mkl_gos.php?nt=538)

Sınıflama ve regresyon modelleri, yapay sinir ağları ve karar ağaçları tahmin edici model olarak belirtilirken; kümeleme, birliktelik kuralları ve ardışık zamanlı örüntü modelleri tanımlayıcı modellerdir.

2.5.1. Sınıflama ve Regresyon Modelleri:

Mevcut verilerden hareket ederek geleceğin tahmin edilmesinde faydalanılan ve veri madenciliği teknikleri içerisinde en yaygın kullanıma sahip olan sınıflama ve regresyon modelleri arasındaki temel fark, tahmin edilen bağımlı değişkenin kategorik veya süreklilik gösteren bir değere sahip olmasıdır. Ancak çok terimli lojistik regresyon (multinomial logistic regression) gibi kategorik değerlerin de tahmin edilmesine olanak sağlayan tekniklerle, her iki model giderek birbirine yaklaşmakta ve bunun bir sonucu olarak aynı tekniklerden yararlanılması mümkün olmaktadır. Sınıflama ve regresyon modellerinde kullanılan başlıca teknikler,

- Genetik Algoritmalar,
- K-En Yakın Komşu,
- Naïve-Bayes,
- Çoklu Regresyon(Lojistik Regresyon)
- Faktör ve Ayırma Analizleri (http://www.bilgiyonetimi.org/cm/pages/mkl_gos.php?nt=538)

2.5.2. Kümeleme Modelleri:

Kümeleme modellerinde veritabanındaki özellikleri birbirinden farklı olan fakat birbirine çok benzeyen üyeler arasında kümeleme yapılmaktadır. Kümeleme analizinde; veri tabanındaki kayıtların hangi kümelere ayrılacağı veya kümelemenin hangi değişken özelliklerine göre yapılacağı konunun uzmanı olan bir kişi yada bilgisayar programları tarafından belirtilmektedir.

2.5.3. Birliktelik Kuralları ve Ardışık Zamanlı Örüntüler :

Bir markette müşterilerin alışverişleri sırasında müşterilerin aldığı ürünlerle ilgili bir analiz tutulmaktadır. Bunun sonucunda hangi ürünlerin daha fazla alındığının belirlenmesi yada herhangi bir ürünün yanında en çok hangi ürün satıldıysa o ürünün bulunarak her iki ürünün aynı reyona konulması müşteriye daha fazla ürünün satılmasını sağlama yollarından biridir. Satın alma eğilimlerinin tanımlanmasını sağlayan birliktelik kuralları ve ardışık zamanlı örüntüler, pazarlama amaçlı olarak pazar sepeti analizi adı altında veri madenciliğinde yaygın olarak kullanılmaktadır.

2.5.4. Bellek Tabanlı Yöntemler:

Bellek tabanlı veya örnek tabanlı yöntemler istatistikte 1950'li yıllarda önerilmiş olmasına rağmen o yıllarda gerektirdiği hesaplama ve bellek

yüzünden kullanılamamış ama günümüzde bilgisayarların ucuzlaması ve kapasitelerinin artmasıyla, özellikle de çok işlemcili sistemlerin yaygınlaşmasıyla, kullanılabilir olmuştur. Bu yönteme en iyi örnek en yakın *k* komşu algoritmasıdır (*k-nearest neighbor*)) (http://enm.blogcu.com/5187251). K en yakın komşu algoritmasında verilerin benzer özellikler arasında kümeleme işlemi gerçekleşmekte, ağırlık merkezi belirlenerek gerekli hesaplamalar yapılmaktadır. Bu hesaplar ağır olduğu için bellek miktarları arttıktan sonra kullanım alanı daha da yaygınlaşmıştır (Özekes S.).

2.5.5. Yapay Sinir Ağları(YSA):

Yapay sinir ağları, insan beyninin özelliklerinden olan öğrenme yolu ile yeni bilgiler türetebilme, yeni bilgiler oluşturabilme ve keşfedebilme gibi yetenekleri herhangi bir yardım almadan otomatik olarak gerçekleştirmek amacı ile geliştirilen bilgisayar sistemleridir. Bu yetenekleri geleneksel programlama yöntemleri ile gerçekleştirmek oldukça zordur veya mümkün değildir. O nedenle, yapay sinir ağları bilim dalının, programlanması çok zor veya mümkün olmayan olaylar için geliştirilmiş adaptif bilgi işleme ile ilgilenen bir bilgisayar bilim dalı olduğu söylenmektedir (http://www.yapay-zeka.org/modules/wiwimod/index.php?page=ANN&back=WiwiHome). Yapay sinir ağlarında kullanılan öğrenme algoritmaları ağırlık hesaplamaları yaparak sonuca ulaşmaktadır. YSA'nın hem uygulama alanı geniştir hem de yüksek işlem gücü ve bellek gerektirmemektedir.

2.5.6. Karar Ağaçları:

Yapay sinir ağlarında veriden bir fonksiyon öğrenildikten sonra bu fonksiyonun insanlar tarafından anlaşılabilecek bir kural olarak yorumlanması zordur. Karar ağaçlarında ise ağaç oluşturulduktan sonra ağaçta kökten yaprağa doğru inilerek kurallar (IF-THEN kuralları) yazılmaktadır. Bu kurallar

uygulama konusunda uzman bir kişiye gösterilerek sonucun anlamlı olup olmadığı denetlenebilir. (http://www.bilgisayaransiklopedisi.com/bilgisayar/496/giris.html)

2.6. Veri Madenciliği Süreci

Veri madenciliğinde hangi algoritma kullanılırsa kullanılsın yapılan işin ve verilerin özelliklerinin bilinmemesi problemin çözümünde hiçbir fayda sağlamamaktadır. Bu yüzden sağlıklı sonuçlar elde edebilmek için veri özelliklerinin öğrenilmesi gerekmektedir. Bunun için izlenmesi gereken yol aşağıdadır;

1. Problemin Tanımlanması,
2. Verilerin Hazırlanması,
3. Modelin Kurulması ve Değerlendirilmesi,
4. Modelin Kullanılması,
5. Modelin İzlenmesi

2.6.1. Problemin Tanımlanması :

Veri madenciliği çalışmalarının başarılı olması için, projenin hangi amaçla yapılacağı, elde edilecek sonuçların başarı düzeylerinin nasıl ölçüleceği açıkça belirtilmelidir. Ayrıca yanlış tahminler sonucunda ortaya çıkan zararların, doğru tahminlerde ise kazançların tahminlerine bu aşamada yer verilmelidir. Aynı zamanda işletmede üretilen sayısal verilerin boyutlarının, proje için yeterlilik düzeyinin, işletme konusu hakkındaki iş süreçlerinin iyi analiz edilmesi gerekmektedir.

2.6.2. Verilerin Hazırlanması :

Bu aşamada verilerin iyi analiz edilmesi, veriler ile problemin arasında ilişki olması ve verilerin ne zaman ortaya çıktığının bilinmesi gerekmektedir. Verilerin hazırlanması toplama, birleştirme ve temizleme, dönüştürme şeklinde 3 e ayrılmaktadır.

- *Toplama:* Verilerin toplanması aşamasıdır. Bu aşamada toplanacak veriler ve bu verilerin hangi veritabanlarından alınacağı belirlenmektedir. Şirketin veritabanlarının yanı sıra başka veritabanlarından alınan verilerde kullanılabilmektedir.
- *Birleştirme ve Temizleme :* Bu aşamada hatalı olan verilerin belirlenerek veritabanından atılması işlemi gerçekleştirilmektedir. Daha sonraki işlemlerde sorun çıkarmaması için temizleme işleminin dikkatli yapılması gerekmektedir.
- *Dönüştürme :* Kullanılacak verilerin bazı kodlamalar kullanılarak tanımlanması veya gösterim şeklinin değiştirilmesi gerektiği durumlarda kullanılmaktadır.

2.6.3. Modelin Kurulması ve Değerlendirilmesi:

Tanımlanan problem için en uygun modelin bulunması gerekmektedir. Bu da birçok model oluşturup daha sonra bu modellerin denenmesi şeklinde uygulanmaktadır. Bu nedenle veri hazırlama ve model kurma aşamaları, en iyi olduğu düşünülen modele varılıncaya kadar yinelenen bir süreçtir.

2.6.4. Modelin Kullanılması:

Kabul edilen bir model doğrudan bir uygulama olarak kullanılabileceği gibi bazen bazı sistemlerin alt parçası olarakta kullanılabilmektedir. Modeller,

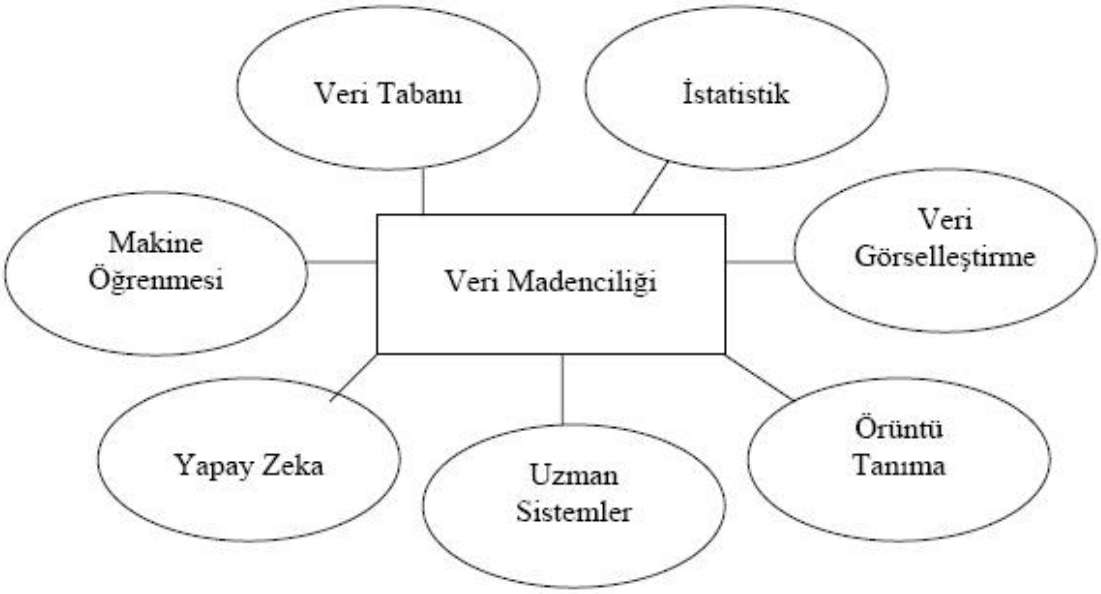
risk analizi, kredi değerlendirme gibi uygulamalarda doğrudan kullanılabileceği gibi, promosyon uygulamalarında alt parça olarak kullanılabilmektedir.

2.6.5. Modelin İzlenmesi:

Veriler üzerinde meydana gelen değişiklikler kurulan modellerin sürekli olarak izlenmesini ve gerekiyorsa yeniden düzenlenmesini gerektirmektedir. Tahmin edilen ve gözlenen değişkenler arasındaki farklılığı gösteren grafikler model sonuçlarının izlenmesinde kullanılan önemli yöntemlerden biridir.

2.7. Veri Madenciliği ve Diğer Disiplinler

Veri madenciliği, makine öğrenmesi, örüntü tanıma, veritabanı teknolojileri, istatistik, yapay zeka, veri görselleştirme gibi bir çok alanın kesişim noktası olarak doğmuş ve bu yönde gelişmesini sürdürmektedir. Bu yapı Şekil 2.1' de gösterilmiştir (Altintop, 2006).



Şekil 2.1 Veri Madenciliği ve Diğer Disiplinler

Bu disiplinlerin veri madenciliğinde kullanım şekilleri

- Makine öğrenmesi, örüntü tanıma ve istatik;örüntü keşfetme sırasında,
- Yapay zeka;örüntüleri yorumlama sırasında,
- Veritabanı; verileri depolama, süzme, temizleme ve sorgulama işlemi sırasında,
- Veri görselleştirme; raporlama sırasında kullanılır.

3. WEB MADENCİLİĞİ

3.1. Web Madenciliğine Genel Bakış

Internet dünya üzerindeki en büyük bilgi paylaşım ortamlarından birisidir. Günümüzde dünya üzerinde yaşayan birçok insan gerekli olan bilgi paylaşımlarını, birbirleriyle olan iletişimlerini Internet üzerinden yapmaktadırlar. Bu yüzden Internet üzerindeki verilerin boyutları da gün geçtikçe hızla artmaktadır. Biriken bu veriler üzerinde işlemler yapabilmek, bu verilere kolayca erişebilmek için kullanılan yöntemlerden bir tanesi de web madenciliğidir. Internet'ten bilgi çıkarımı ve bilgi keşfi işlemleri, web madenciliğinin önemli bir bölümüdür.

Web madenciliği ilk kez 1996 yılında Oren Etzoni tarafından ortaya çıkarılmıştır. Oren Etzoni' ye göre web madenciliği, veri madenciliği tekniklerinin kullanılarak web belgelerinden ve servislerinden otomatik olarak bilginin ayıklanması, ortaya çıkarılması ve tahlil edilmesidir (Etzioni, 1996). Çünkü Internet'te bulunan verilerin sürekli olarak değişmesi, güncellenmesi, silinmesi ve yeni bilgilerin eklenmesi web den bilgi çıkarımı işleminde karşılaşılan en önemli zorluklardan birisidir. İşte bu yüzden web le ilgili problemlerin çözümünde web madenciliği yöntemi kullanılmaktadır.

Web madenciliği, web kayıt dosyalarında yani veritabanlarında kayıtlı olan verilere ihtiyaç duyulduğunda yararlı bilgilerin veritabanlarından çıkarılması ve değerlendirilmesi sağlar.

Aynı zamanda web madenciliği sayesinde, ziyaretçiler web sayfalarını gezerken ziyaretçiyle ilgili birçok bilgiye ulaşılabilmektedir. Ziyaretçi web sayfasında gezerken sadece web sayfasında bazı linklere tıkladığını düşünür. Oysa arka planda ziyaretçinin hangi tür linklere tıkladığı tutulmaktadır. Ve bu linklere göre kullanıcının, web sitesine bir sonraki girişinde hangi sayfalara gireceği belirlenerek o sayfalarla ilgili düzenlemeler yapılabilmektedir. Bu

düzenlemeler arasında en yaygın olanı o sayfalara reklamlar koymaktır. Aynı zamanda üye kayıt formu gibi form doldurularak girilen web sayfalarında aslında kullanıcı farkında olmadan kendisiyle ilgili birçok bilgiyi de sanal ortama aktarmaktadır.

Web madenciliği içerik olarak web içerik madenciliği, web yapı madenciliği ve web kullanım madenciliği şeklinde 3 ana bölüme ayrılmaktadır. Bu bölümlerin birbiriyle olan farklılıkları Tablo 3.1 de detaylı olarak gösterilmiştir.

	Web Madenciliği		
	İçerik Madenciliği	**Yapı Madenciliği**	**Kullanım Madenciliği**
Veri	Metin belgeleri ve HTML	HTML Linkleri	server logs, browser logs...
Verinin Şekli	Yapısız ve karışık veya DB	Link Yapısında	Kullanıcı Etkileşimi
Gösterimi	İlişkisel ve sınıflandırmalı	Graph	İlişkisel Tablolar ve Graph

Tablo 3.1 Web madenciliği tekniklerinin temel farklarına göre basitçe kıyaslanması (http://www.bilyaz.com/bMakaleGetir.php?id=56)

3.1.1. Web İçerik Madenciliği

Web içerik madenciliği, temel olarak web kaynakları içerisinden bilginin bulunmasını sağlamaktadır. Web kaynağının veritabanı yapısında olması yapılan işlemleri daha da kolaylaştırmaktadır.

Web içerik madenciliği, veri madenciliği ile ilgili olduğundan dolayı web kaynakları içerisindeki verileri çıkarmak için veri madenciliği tekniklerini kullanır.

Web içerik madenciliği, sitelerin içeriğiyle ilgilenir. Web kaynakları içerisinde metin, resim, ses, görüntü ve linkler bulunmaktadır. Web içerik madenciliği, bu kaynaklar arasından saklı bilginin bulunması ve filtrelenmesini sağlamaktadır. Web kaynaklarından içeriklerine göre otomatik bilgi arama teknikleri tanımlamaktadır. Bu arama teknikleri 2 sınıfa ayrılırlar.

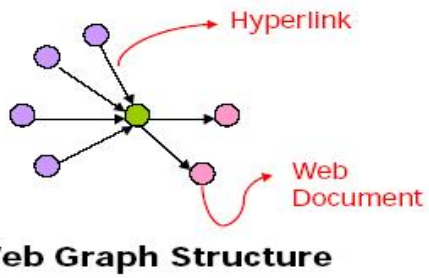
- **Information Retrieval Approach (IR):** Kullanıcı profili baz alınarak kullanıcılara gösterilen bilgileri filtrelemek ve bilgiye erişimi geliştirmek için kullanılan yöntemdir.

- **Database Approach:** Sınıflandırma, kümeleme gibi tekniklerden yararlanarak web'deki veriyi modeller. Özellikle arama motorları bu tekniği kullanırlar. Ayrıca web belgelerinin sınıflandırılmasında, farklı sunuculardaki aynı içeriğe sahip web sayfalarının bulunmasında ve web belgelerinin çeşitli konularda temsil edilmesinde kullanılır. (http://www.recepayaz.com/library/web_madenciligine_bir_bakis.doc)

3.1.2. Web Yapı Madenciliği

Web yapı madenciliğinin amacı web sitesi veya bağlantı verisine bakarak web sayfası hakkında bilgi üretmektir. Web içerik madenciliği dokümanın içeriğiyle ilgilenirken, web yapı madenciliği ise dokümanlar arası bağlantılarla ilgilenmektedir. Web yapı madenciliği, linklerin topolojisine dayanarak farklı siteler arasındaki benzerlik ve ilişki gibi bilgileri üretir (http://www.recepayaz.com/library/web_madenciligine_bir_bakis.doc). Ayrıca web yapı madenciliği sayfaların link tasarımlarının ortaya çıkmasında kullanılmaktadır. Araştırma hyperlink seviyesinde yapıldığından "Hyperlink Analysis" de denilmektedir. Şekil 3.1 'de web grafik yapısı görülmektedir.

Burada web sayfaları arasındaki oklar iki sayfa arasındaki bağlantıyı temsil etmektedir.



Şekil 3.1 Web sayfaları arasındaki link bağlantısı (Srivastava J.)

3.1.3. Web Kullanım Madenciliği

Web kullanım madenciliğinin amacı, site ziyaretçilerinin siteyi gezerken site içerisindeki hareketlerinin incelenmesini sağlamaktır.

Kuruluşlar, kullanıcının isteği dışında bu verileri toparlayarak istemcilerde, sunucularda ve Proxy sunucularında depolamaktadır. Toplanan bu veriler çok büyük boyutlardadır. Veri kaynakları olarak sunucu erişim kayıtları, istemci tarafında bulunan çerezler (cookies), kullanıcı profilleri yani kayıt bilgileri, sayfa özellikleri, içerik özellikleri, kullanılan veri, fare klikleri sayılabilir. Bu veriler günlük dosya şeklinde tutulmaktadır. Bu dosyalar, kullanıcının istemci makinesinden sunucu makinesine gönderdiği her bir isteğin kaydedilmesiyle oluşmaktadır.

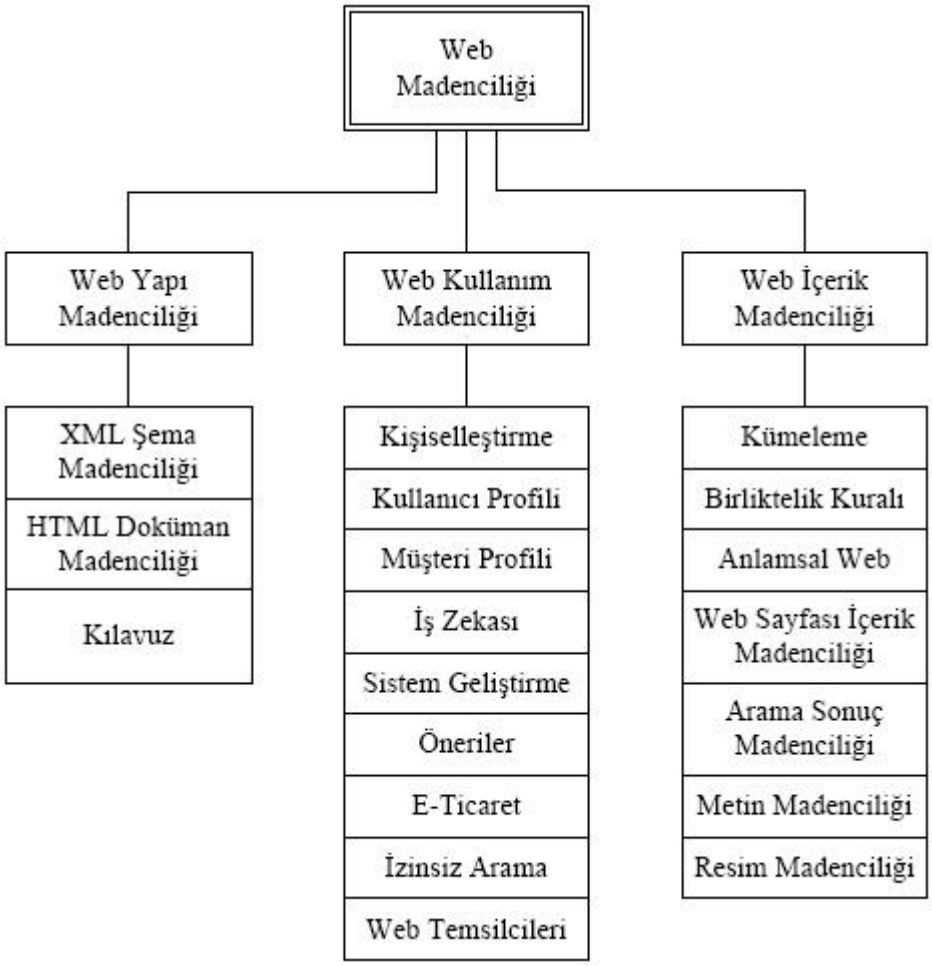
Günlük dosyalarının analizi, müşterilerin ilgi alanları, ürünler üzerinden pazar stratejileri oluşturma, promosyon kampanyaları gibi hususlarda, kurumlara karar verme süreçlerinde yardımcı olmaktadır. Sunucu erişim kayıtlarının ve kullanıcı kaydı verilerinin analizi, aynı zamanda kurumun daha etkili bir sunumunun yapılabilmesi için web sitesini nasıl daha iyi hale getirebileceği hakkında önemli bilgiler sağlar.

Web içerik madenciliği ve web yapı madenciliği gerçek veriyi kullanırken, web kullanım madenciliği ise kullanıcılar web de gezinirken elde edilen bilgilerinden sağlanan veriyi kullanmaktadır.

Web kullanım madenciliğinde kullanılacak veriler aşağıdaki tiplerde olabilir:

- **İçerik verisi**: Web dokümanlarında, genellikle metin şeklinde yer alan verilerdir. Herhangi bir web sayfası üzerinde yer alan veriler bu tip için bir örnektir.
- **Yapı verisi**: Web sitesinin bağlantı yapısı hakkındaki verilerdir. Web sitesinde yer alan sayfaların hangi alt dizinler içerisinde bulunduğunu gösteren verilerden oluşur.
- **Kullanım verisi**: Web sitesini ziyaret eden kullanıcıların oluşturdukları veri tipidir. Kullanım verisi genellikle hangi kullanıcı, ne zaman, hangi sayfaları ziyaret etti, ne kadar süre sitede kaldı gibi soruların cevaplarını içerir.
- **Kullanıcı profili**: Web sitesini ziyaret eden kullanıcı hakkındaki; kullanıcı kimlik verileri gibi bilgilerden oluşur (http://www.bilyaz.com/bMakaleGetir.php?id=56).

Web içerik madenciliğini, web yapı madenciliğini ve web kullanım madenciliğini Şekil 3.2'deki gibi özetleyebiliriz.



Şekil 3.2. Web Madenciliğinin Sınıflandırılması(Daş R.,2006)

4. BİRLİKTELİK KURALLARI

Birliktelik kuralları diye adlandırılan kurallar, verilen veri kümesinde elemanlar arasındaki önemli ilişkileri arar.

4.1. Birliktelik Kurallarında Kullanılan Değerler

- Support : Destek sayısı
  Support (X->Y) : X ve Y item setini içeren hareketlerin sayısı
  Support(A->B) = Support (A U B)

- Confidence : Güven değeri
  Confidence (X->Y) : X iteminin bulunduğu yerde X ve Y item setinin bulunma olasılığı
  Confidence (A->B) = Support (A U B) / Support (A)
- k : Küme sayısı
- Ck : Aday öğe kümesi
- Lk : Sık geçen öğe kümesi
- Λ : Birleştirme (join) işlemi
- A->B : Birliktelik-ilişki kuralı

4.2. Apriori Algoritması

Apriori algoritması, veri madenciliğinde sık geçen öğelerin bulunması için kullanılan algoritmalardan biridir.

Apriori, boolean ilişkisel kurallar için en sık işlem gören elemankümeyi çıkaran etkili bir algoritmadır. Apriori, en sık işlem gören elemankümeye ait çok az bilgiye ihtiyaç duymaktadır. Apriori algoritması, k+1 tane elemankümeyi

keşfetmek için k tane elemanküme kullanmaktadır. Burada level-wise (akıllı seviye) arama olarak da bilinen tekrarlı bir yordamı vardır. İlk önce 1-elemanküme bulunur. Bu L_1 şeklinde gösterilir. L_1, L_2 yi, L_2 ise L_3 ü bulmak için kullanılır. Bu işlem daha yüksek k-elemanküme bulunamayana kadar sürer. Her bir L_k yı bulmak için veritabanını tamamen taramak gerekir. Apriorinin verimliliğini arttırmak ve arama zamanını kısaltmak için kullanılan önemli bir özelliği vardır. Bu özellik;

- Apriori özelliğinin kullanımındaki olağan (frequent) elemankümenin tüm boş olmayan alt kümeleri de olağan olmalıdır. Bu kural şu görüşe dayanır: eğer bir elemanküme (I) minimum support eşiğini sağlamıyorsa I olağan değildir (P(I)<min_sup). Eğer elemanküme I'ya, A elemanı eklenirse , sonuç elemanküme ($I \cup A$), I'dan daha fazla frekansa sahip değildir. Bu yüzden, $I \cup A$ bir olağan elemanküme değildir (P($I \cup A$)<min_sup). Yani, eğer bir küme bir testi geçemediyse onun bütün süper kümeleri de o testten kalır. Bu özelliğe anti-monoton denir.

Algoritmada apriori özelliğinin nasıl kullanıldığı ve L_1 i bulurken L_{k-1} den nasıl faydalanıldığı işlemleri **birleştirme** (join) ve **budama** (prune) adı verilen iki aşamadan oluşan bir sistemdir(Han J., 2001).

4.2.1. Birleştirme Adımı:

L_k yı bulmak için, L_{k-1} kendisiyle birleştirilerek aday k-elemanküme bulunur. Bu aday küme C_k olarak gösterilir. I_1 ve L_2 de L_{k-1} deki elemankümeler olsun. $I_i[j]$ notasyonu I_i nin j inci elemanını belirtir. Örneğin $I_1[k-2]$, I_1 deki sondan 2. elemanı temsil eder. Apriori elemanların sözlüksel sıraya göre dizildiğini kabul eder. Birleştirme işlemi, L_{k-1} 'in üyeleri birleştirilebilir özelliğe sahipse gerçekleştirilebilir. Bunun için de üyelerin ilk k-2 elemanlarının ortak olması gerekir. L_{k-1} 'in üyeleri olan I_1 ve I_2 , ($I_1[1]=I_2[1] \wedge I_1[1]=I_2[1] \wedge \ldots \wedge I_1[k-2] = I_2[k-2] \wedge I_1[k-1] < I_2[k-1]$) şartı sağlanıyorsa birleştirilebilir. $I_1[k-1] < I_2[k-1]$ şartı kopyaların oluşmasını engeller. I_1 ve I_2 nin birleştirilmesiyle oluşan sonuç küme $I_1[1]I_1[2]\ldots I_1[k-1]I_2[k-2]$ dir (Han J., 2001).

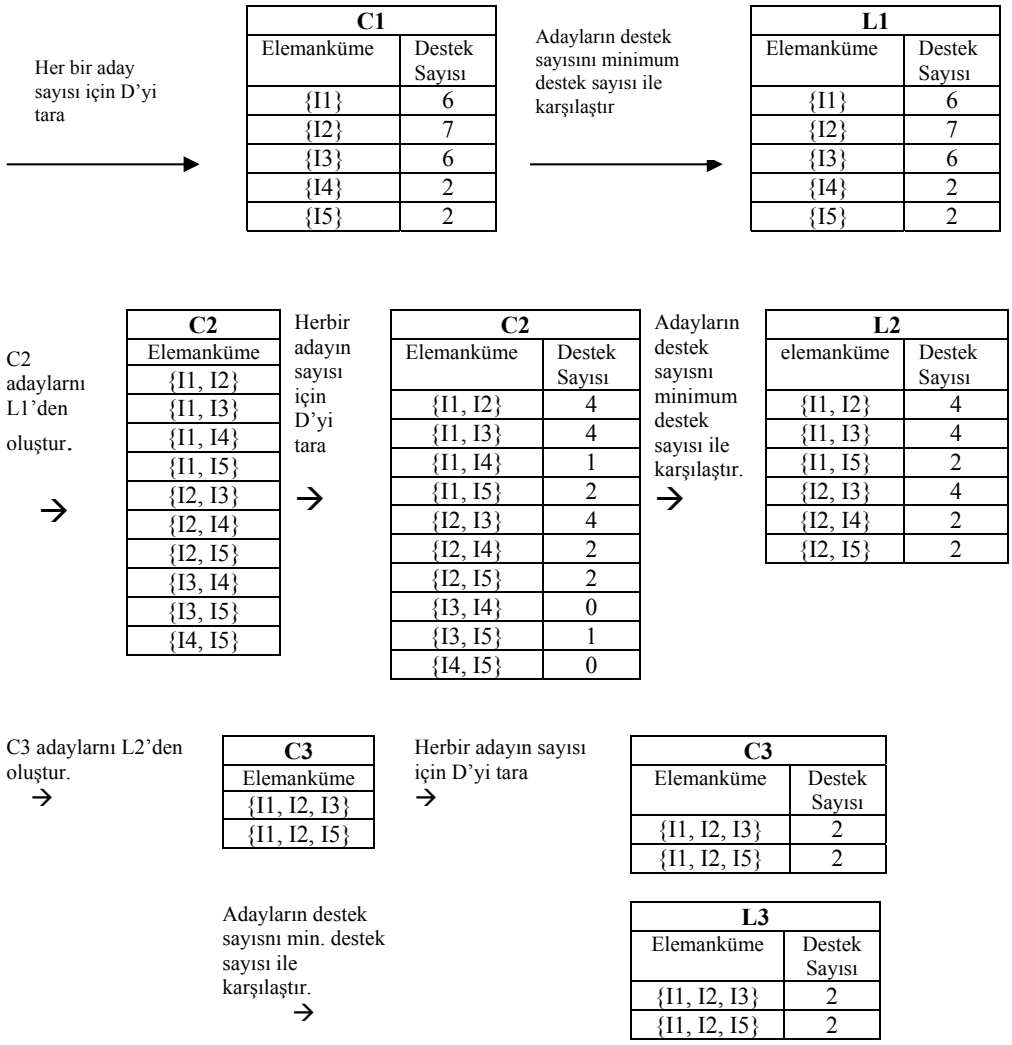
4.2.2. Budama Adımı:

C_k , L_k nın süper kümesidir. Üyeleri olağan (frequent) olabilir veya olmayabilir. Fakat tüm olağan k-elemankümeleri C_k nın içindedir. C_k nın içindeki her bir adayın sayısını belirlemek için yapılan veritabanı taraması, L_k nın belirlenmesiyle sona erer. Örneğin: tanım olarak, minimum support sayısından daha fazla sayıya sahip olan tüm adaylar olağandır ve bu yüzden L_k ya aittir. C_k ne derece büyükse, o kadar ağır bir hesaplamayı gerektirir. C_k nın boyutunu küçültmek için Apriori özelliği şu şekilde kullanılır. Olağan olmayan herhangi bir (k-1)-elemanküme, k-elemankümenin alt kümesi olamaz. Bundan dolayı eğer aday k-elemakümenin herhangi bir (k-1)-altkümesi L_{k-1} in içinde değilse, aday da olağan olamayacağından C_k nın içinden çıkarılabilir. Bu alt küme testi, tüm olağan elemankümelerin karmaşık ağacı oluşturularak yapılabilir(Han J., 2001).

TID	Item_ID Listesi
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

Tablo 4.1 İşlem verileri

Apriori algoritması ile ilgili somut bir örnek aşağıda verilmiştir. Bu örnekte Tablo 4.1 deki işlem veritabanı (D) kullanılmıştır. Bu veritabanında 9 tane işlem vardır (|D|=9). D'deki olağan elemankümeleri bulmaya yarayan apriori algoritmasını göstermek için Tablo 4.2 kullanılmıştır.



Tablo 4.2 Minimum Destek Sayısı = 2

1. Algoritmanın birinci adımında, her eleman, aday 1-elemankümelerin kümesi olan C1'in adayıdır. Algoritma her bir elemanın hareket sayısını bulmak için tüm işlemleri sırayla tarar.

2. Minimum işlem destek sayısının 2 olarak kabul edilip olağan 1-elemankümelerin kümesi olan L_1 belirlenir. Bu, minimum destek eşiğini geçen 1-elemankümelerden oluşur.

3. Olağan 2-elemankümelerin kümesi olan L_2'yi bulmak için algoritma L_1►◄ L_1' i kullanır. Böylece 2-elemankümelerin aday kümesi olan C_2 oluşturulur. C_2, 2-elemankümelerden oluşur.

4. Daha sonra, D'deki işlemler taranır ve C_2'deki her bir aday elemankümenin destek sayısı toplanır. Bu Tablo 4.2 de orta sıradaki orta tabloda gösterilmiştir.

5. Artık olağan 2-elemankümelerin kümesi L_2 oluşturulabilir. L_2, C_2'deki minimum desteğe sahip 2-elemankümelerden oluşur.

6. Aday 3-elemankümelerin kümesi, C_3'ün oluşturulması Tablo 4.3'te ayrıntılarıyla gösterilmiştir. $C_3=L_2$►◄L_2={{I1,I2,I3},{I1,I2,I5}, {I1,I3,I5},{I1,I3,I4},{I2,I3,I5},{I2,I4,I5}} olsun. Apriori özelliğine göre olağan elemankümelerin tüm alt kümeleri de olağan olmalıdır. Böylece son 4 adayın olağan olmadığı belirlenir. Bu yüzden, L_3'ü oluşturmak için yapılacak bir sonraki D taramasında, onların sayısı boş yere tutulmayarak C_3'den çıkarılırlar. Şuna dikkat edilmelidir ki apriori algoritması akıllı seviye (level-wise) arama stratejisi kullandığından, bir aday l-elemanküme verildiğinde, ihtiyaç olan tek şey onun (k-1)-alt kümelerini kontrol etmektir.

7. L_3'ü oluşturmak için D'deki işlemler sırasıyla taranır. Böylece C_3 içindeki aday 3-elemankümeler minimum destek değerini alır.

8. Algoritma, 4-elemankümelerin kümesi olan C_4'ü bulmak için L_3►◄L_3 'ü kullanır. Birleşimin sonucu {{I1,I2,I3,I5}} olmasına rağmen, alt kümesi olan {{I2,I3,I5}} olağan olmadığı için bu elemanküme budanır. Böylece $C_4=\Phi$ dir ve algoritma, tüm olağan elemankümeleri bularak sonlanır.

1- Join: $C_3 = L_1 ►◄ L_1 = \{\{I1,I2\}, \{I1,I3\}, \{I1,I5\}, \{I2,I3\}, \{I2,I4\}, \{I2,I5\}\} ►◄ \{\{I1,I2\}, \{I1,I3\}, \{I1,I5\}, \{I2,I3\}, \{I2,I4\}, \{I2,I5\}\} = \{\{I1,I2,I3\}, \{I1,I2,I5\}, \{I1,I3,I5\}, \{I2,I3,I4\}, \{I1,I3,I5\}, \{I2,I4,I5\}\}$.

2- Budama apriori özelliğini kullanır.

- {I1,I2,I3} ün 2 elemanlı alt kümeleri {I1,I2},{I1,I3} ve {I2,I3}. 2 elemanlı alt kümelerin hepsi L_2'nin üyesidir. Onun için {I1,I2,I3} ü C_3'ün içinde tutulur.

- {I1,I2,I5} ün 2 elemanlı alt kümeleri {I1,I2},{I1,I5} ve {I2,I5}. 2 elemanlı alt kümelerin hepsi L_2'nin üyesidir. Onun için {I1,I2,I5} i C_3'ün içinde tutulur.

- {I1,I3,I5} ün 2 elemanlı alt kümeleri {I1,I3},{I1,I5} ve {I3,I5}. {I3,I5} L_2'nin üyesi değildir. Olağan da değildir. Onun için {I1,I3,I5} i C_3'den çıkartılır.

- {I2,I3,I4} ün 2 elemanlı alt kümeleri {I2,I3},{I2,I4} ve {I3,I4}. {I3,I4} L_2'nin üyesi değildir. Olağan da değildir. Onun için {I2,I3,I4} ü C_3'den çıkartılır.

- {I2,I3,I5} ün 2 elemanlı alt kümeleri {I2,I3},{I2,I5} ve {I3,I5}. {I3,I5} L_2'nin üyesi değildir. Olağan da değildir. Onun için {I2,I3,I5} i C_3'den çıkartılır.

- {I1,I3,I5} ün 2 elemanlı alt kümeleri {I1,I3},{I1,I5} ve {I3,I5}. {I3,I5} L_2'nin üyesi değildir. Olağan da değildir. Onun için {I1,I3,I5} i C_3'den çıkartılır.

- {I2,I4,I5} ün 2 elemanlı alt kümeleri {I2,I4},{I2,I5} ve {I4,I5}. {I4,I5} L_2'nin üyesi değildir. Olağan da değildir. Onun için {I2,I4,I5} i C_3'den çıkartılır.

3- Budama işleminden sonra $C_3 = \{\{I1,I2,I3\}, \{I1,I2,I5\}\}$

Tablo 4.3 C_3 ün oluşturulması

4.3 Apriori Algoritması Pseudocode

Apriori algoritmasının pseudocode'unu ve ilgili prosedürleri aşağıda göstermektedir. 1. adımda apriori olağan 1-elemankümeleri (L_1) bulmaktadır. 2.-10. adımlarda, aday C_k'yı ve daha sonra L_k'yı bulmak için L_{k-1}'i kullanmaktadır. Apriori_gen prosedürü adayları oluşturup olağan olmayan alt kümelere sahip olanları elemek için Apriori özelliğini kullanmaktadır. İlk olarak adaylar oluşturulmuş, veritabanı taranmıştır (4. adım). Her bir işlemin üye altkümelerini bulmak için subset_function kullanılmıştır (5. adım). Daha sonra bu adayların her birinin sayısı toplanmıştır (6.-7. adım). Sonuç olarak tüm bu minimum support noktasına ulaşan adaylar, olağan elemankümelerin kümesi olan L'yi oluşturmaktadır. Daha sonra olağan elemankümelerden ilişkisel kuralları çıkaracak bir prosedür çağrılır.

Apriori_gen prosedürü, birleştirme ve budama isimli iki işlemi yerine getirir. Birleştirme kısmında, potansiyel adayları bulmak için L_{k-1}'i kendisiyle birleştirmektedir (1.-4. adım). Budama kısmı ise olağan olmayan alt kümelere sahip adayları elemek için apriori özelliğini çalıştırmaktadır (5.-7. adım). Olağan alt küme olup olmadığını anlamak için kullanılan test has_infrequent_subset prosedürü tarafından yerine getirilmektedir. (Han J.,2001)

Input: Veritabanı: D işlemleri; minimum destek eşiği: min_sup
Output:L, D'deki olağan elemankümeler
Method:

1 L_1=find_frequent_1-itemsets(D);
2 For (k=2;$L_{k-1}\neq\Phi$;k++){
3 C_k=apriori_gen(L_{k-1},min_sup);
4 for each transaction t$\in$D { //D'yi sayılar için tara
5 C_t=subset(C_k,t); //aday olan t alt kümelerin al
6 for each candidate c$\in C_t$

```
7               c.count++;
8           }
9           Lk={c∈Ck|c.count≥min_sup}
10      }
11      return L=UkLk;
```

procedure apriori_gen(L_{k-1}:frequent(k-1)-itemsets; min_sup:minimum support threshold)

```
1           for each itemset I1∈Lk-1
2                   for each itemset I2∈Lk-1
3                       if (I1[1]=I2[1])∧ (I1[2]=I2[2])∧…∧ (I1[k-2]=I2[k-2])∧
            (I1[k-1]<I2[k-1]) then {
4                           c=I1►◄I2; //birleştirme adımı: adayları oluştur
5                           if has_infrequent_subset(c,Lk-1)        then
6                             delete c;//budama adımı
7                           else add c to Ck;
8                   }
9           return Ck;
```

procedure has_infrequent_subset(c:candidate k-itemset; L_{k-1}:frequent(k-1)-itemset);

```
1           for each (k-1)-subset s of c
2                   if s∉Lk-1 then
3                           return TRUE;
4           return FALSE;           (Han J.,2001)
```

4.4. Olağan Elemankümelerden İlişkisel Kurallar Oluşturma

D veritabanındaki işlemlerden olağan elemankümeleri bulunduktan sonra güçlü ilişkisel kurallar bulunacaktır. Güçlü ilişkisel kurallar minimum destek ve minimum güven değerlerini sağlar. Bu özellik F 4.1 güven eşitliği ile sağlanır. Burada şartlı olasılık destek sayısı terimiyle ifade edilir.

$$\text{Güven}(A \Rightarrow B) = P(A|B) = \frac{\sup port_sayıası(A \cup B)}{\sup port_sayıası(A)} \qquad \text{(F 4.1)}$$

Destek_sayısı(A∪B), A∪B elemankümelerini içeren işlem sayısını, destek_sayısı(A) ise A elemankümelerini içeren işlem sayısını ifade eder. Bu eşitliğe dayanarak ilişkisel kurallar şu şekilde oluşturulabilir;

- Her bir olağan elemanküme I için, I'nın boş olmayan alt kümeleri oluşturulur.
- Her bir boş olmayan alt küme için F 4.2 kuralı uygulanır.

$$\text{“}s \Rightarrow (I\text{-}s)\text{” eğer } \frac{\sup port_sayıası(I)}{\sup port_sayıası(s)} \geq \min_conf \text{ (F 4.2)}$$

Kurallar olağan elemankümelerden çıkarıldığı için, her biri otomatik olarak minimum destek noktasını sağlayacaktır. Olağan elemankümeler tablolarda sayılarıyla birlikte tutulabilecek böylece onlara erişim daha kolay bir hale gelecektir.

Tablo 4.1 de gösterilen işlem verilerine göre bir örnek aşağıda verilmiştir.

I={I1,I2,I5} olağan elemankümeyi içeren veri düşünüldüğünde I'dan çıkarılabilecek ilişkisel kurallar; {I1,I2}, {I1,I5}, {I2,I5}, {I1}, {I2} ve {I5} I'nın boş olmayan alt kümeleridir. Böylece I'nın ilişkisel kuralları aşağıdaki gibi oluşturulmuş ve her biri güven değerleriyle birlikte verilmiştir.

$I1 \wedge I2 \Rightarrow I5$ Güven değeri=2/4=50%

$I1 \wedge I5 \Rightarrow I2$ Güven değeri=2/2=100%

$I2 \wedge I5 \Rightarrow I1$ Güven değeri=2/2=100%

$I1 \Rightarrow I2 \wedge I5$ Güven değeri=2/6=33%

$I2 \Rightarrow I1 \wedge I5$ Güven değeri=2/7=29%

$I5 \Rightarrow I1 \wedge I2$ Güven değeri=2/2=100%

Eğer güven değeri eşiğinin %70 olduğu söylenseydi sadece 2. 3. ve 6. kural güçlü olacaktı.

4.5. Apriori Algoritmasının Verimliliğini Artırma

Apriorinin önerilen birçok varyasyonu, orijinal algoritmanın verimliliğini geliştirmeye odaklanır. Bunlardan bir kaçı;

4.5.1. Karmaşık Tabanlı Teknik:

Karmaşık tabanlı (hash-based) teknik aday k-elemankümelerin boyutunu küçültmek için kullanılabilir (k>1). Örneğin, aday 1-elemanküme olan C_1'den, olağan 1-elemankümeyi elde etmek için veritabanındaki her bir işlemi tararken, her bir işlem için 2-elemanküme oluşturulabilir ve hash tablosunun ayrı bir kısmına yazılabilir. Aynı zamanda bu alandaki işlem sayısı da tespit edilebilir. Bu sayının destek eşiğinden düşük olduğu 2-elemanküme olağan olamayacağından aday kümeden çıkarılacaktır. Bu teknik sayesinde aday sayısı yeterince azaltılmaktadır. (Han J.,2001)

4.5.2. İşlem Azaltma:

Olağan k-eleman kümeyi içermeyen bir işlem olağan (k+1)-elemankümeyi de içeremez. Bu yüzden bu tip işlemler işaretlenebilir veya göze alınmayabilir. k'dan daha büyük bir j değeri için j-elemankümeyi elde etmek için yapılan veritabanı taramasında bu işleme gerek yoktur. (http://www3.itu.edu.tr/~sgunduz/courses/verimaden/slides/d7.pdf)

4.5.3. Bölme:

Bölme tekniği, olağan elemankümeleri çıkarmak için iki veritabanı taramasına ihtiyaç duyulduğu zamanlarda kullanılır. İki aşamadan oluşur. 1. aşamada, algoritma D'deki işlemleri n tane birbirinden tamamen ayrı parçaya böler. Eğer D'deki işlemler için minimum destek noktası min_sup ise, bölümlerden her biri için minimum elemanküme destek sayısı *min_sup* * *bölümdeki işlem sayısı* olacaktır. Her bir bölüm için içindeki tüm olağan elemankümeler bulunur. Bunlar lokal olağan elemankümelerdir. Prosedür özel bir veri yapısı kullanır. Bu yapı, her bir elemanküme için, elemankümedeki elemanları içeren işlemlerin TID'larını kaydeder. Bu ona bir veritabanı taramasıyla her bir lokal olağan k-elemankümeleri bulmasına izin verir (k=1,2,…).

Lokal bir olağan elemanküme, tüm veritabanına göre olağan oladabilir olmayadabilir. D'ye göre potansiyel olağan olan bir elemanküme, bölümlerin en az birinde bulunmalıdır. Bu yüzden, tüm lokal olağan elemankümeler D'ye göre birer adaydır. Tüm bölümlerdeki lokal olağan elemankümelerin birikimi sonucu D'ye göre global aday elemankümeler oluşacaktır. 2. aşamada her bir aday için gerçek desteğin hesaplandığı ikinci bir D taraması gerçekleştirilir. Bölüm boyutları ve bölüm sayısı ana belleğe kolay bir şekilde sığacak kadar uygundur. Bu alan her bir aşama için sadece 1 kez okunur. (Özekes S.)

4.5.4. Modelleme:

Modelleme yaklaşımının ana fikri verilen verinin (D), rastgele bir modelini seçmek (S) ve D yerine S'nin içinde olağan elemankümeleri aramaktır. Burada verimlilik için kesinlikten bir nebze ödün verilmiştir. S örneği, olağan elemankümelerin aranması esnasında, ana belleğe sığacak boyuttadır. Ve S'deki işlemlerin sadece bir kere taranması yeterlidir. D'nin yerine S'nin içinde olağan elemankümeleri aramamız, global olağan elemankümelerden bazılarını kaybedilmesine sebep olabilir. Bu ihitimali düşürmek için S'ye ait olağan elemankümeleri ararken, normalden düşük bir support eşiği kullanılır. (L^s S'ye ait elemankümelerdir.) Veri tabanının geri kalan kısmı L^s'deki elemankümelerin gerçek frekanslarını hesaplamak için kullanılır. Bir mekanizma, L^s'nin tüm global olağan elemankümeleri içerip içermediğine karar verir. Eğer L^s, gerçekten tüm global olağan elemankümeleri içeriyorsa, D'nin sadece bir defa taranması gerekir. Aksi halde, birinci seferde kaybedilen olağan elemankümeleri bulmak için ikinci bir deneme gerekir. Modelleme yaklaşımı verimliliğin en önemli olduğu, sık sık tekrarlanan işlemlerin yoğun olduğu sistemlerde çok faydalıdır. (Han J.,2001)

4.5.5 Dinamik Elemanküme Sayımı:

Dinamik elemanküme sayım tekniği başlangıç noktaları işaretlenerek bölümlere ayrılmış veritabanlarında önerilir. Bu varyasyondaki aprioriden farklı olarak, yeni aday elemankümeler herhangi bir başlangıç noktasına eklenebilir. Bu metod bir yandan alt kümeleri olağan olduğu anlaşılan aday elemankümeyi eklerken, diğer yandan da öteki elemankümelerin supportunu hesaplar. Bunun için dinamiktir. Sonuç algoritma, aprioriden birkaç tane daha fazla veritabanı taramasına ihtiyaç duyar. (Han J.,2001)

4.6. Aday Jenerasyonları Kullanmadan Olağan Elemankümelerin Çıkarımı

Görüldüğü gibi bir çok durumda Apriori algoritmasının aday oluşturma-test etme metodu aday kümelerin önemli biçimde küçülmesine ve performansın artmasına sebep olmaktadır. Bununla birlikte iki dezavantajı vardır;

- Çok fazla sayıda aday küme oluşturmaya ihtiyaç duyar. Örneğin, 10^4 tane olağan 1-elemanküme varsa apriori algoritması 10^7 tane aday 2-elemanküme oluşturmak ve bunların gerçekleşme frekanslarını hesaplamak zorunda kalacaktır. $\{a_1,\ldots,a_{100}\}$ gibi 100 elemanlı bir olağan numunenin keşfi için, toplamda 2^{100} (yaklaşık 10^{30}) den fazla aday oluşturacaktır.

- Veritabanını tekrar tekrar taramaya ve büyük aday kümelerini, numune eşleştirerek kontrol etmeye ihtiyaç duyabilmektedir.

5. WEB TABANLI UYGULAMALARDA İNSAN UYUMLULUKLARININ TESBİTİ

Veri madenciliğinin web tabanlı uygulamalarda kullanılmasının web madenciliği olarak adlandırıldığından Bölüm 3.1 de bahsedilmişti. Web madenciliği sayesinde kullanıcının siteye yaptığı ziyaret esnasında toplanan tüm veriler bilgiye dönüştürülerek kullanıcıya daha da iyi imkanlar sunulabilir. Günümüzde pek çok web sitesi üzerinden alışveriş, Internet üzerinde arama yapma, ücretsiz e-posta, web sohbeti, web üzerinden kendilerine uygun insanlarla tanışma, web üzerinden mesajlaşma gibi pek çok hizmet sunumaktadır. Bu şekildeki web sayfaları yayınladıkları reklamlardan, siteye üyeliklerin paralı olmasından, online satışlardan para kazanmaktadırlar. Kuruluşlar daha da ileriye gidebilmek, rakiplerinin bir adım önüne geçebilmek ve parasal anlamda daha çok kazanabilmek için web madenciliği yöntemini tercih etmektedirler.

Web sitesi üzerinden toplanan verilere Veri Madenciliği (Data Mining) yöntemleri uygulandığında verideki gizli ilişkiler ve ziyaretçilerin sitedeki davranış modelleri keşfedilebilir. Yani web madenciliği aşağıdaki soruların cevaplarını vermeyi sağlar.

- Ziyaretçiler web sayfasında nasıl hareket ederler? Ziyaretçilerin sitedeki tüm hareketleri tutulur. Bu sayede ziyaretçinin isteklerinin ne doğrultuda geliştiği bulunur.
- Ziyaretçilerin davranış modelleri nasıldır? Hangi ziyaretçiler hangi sayfalara girmiş?
- Bir sonraki adımda ziyaretçi neler yapabilir? Ziyaretçinin bir önceki siteye gelişindeki verilerden yola çıkarak ziyaretçinin bir sonraki adımda hangi sayfaya gidebileceği tahmin edilir.
- Ziyaretçi hangi durumlarda sitede daha çok kalabilir? Ziyaretçinin ihtiyaçlarına özel durumlar sunularak sitedeki üyeliğinin devam ettirilmesi veya sitede daha çok gezinmesi sağlanabilir.

- Ziyaretçi sayısı nasıl artırılabilir? Sitede birtakım özel promosyonlar düzenlenerek ziyaretçi sayısının artırılması sağlanabilir.

Bu tür durumlar sonucunda sitede ne kadar çok ziyaretçi olursa siteye verilen reklamlardan, siteye parayla üye olan kullanıcılardan gün geçtikçe artan gelir sağlanmaktadır.

Bu projede yapılan işlemler;

5.1. Veritabanının Oluşturulması

Projede oluşturulan tablolar;

- *kullanici* isimli tablo; kullanıcının sisteme üye olurken kullandığı tablo

  id : Otomatik Sayı (Primary Key)
  kullanici_adi : String
  sifre : String
  aktivasyon_kodu : String
  aktivasyon : String (Evet/Hayır)

- *kullanici_bilgileri* isimli tablo;kullanıcının sahip olduğu tüm özellikleri tutan tablo.

  kullanici_id : Sayı (*kullanici* isimli tablodaki id alanıyla bağlantılı) (Foreign Key)
  ad : String
  soyad : String
  cinsiyet : String
  yasadigi_yer : String
  yas : Sayı
  dogum_tarihi : Tarih/Saat
  meslek : String

medeni_durum : String
burc : String
egitim_durumu : String
boy : Sayı
kilo : Sayı
sac_rengi : String
goz_rengi : String
kendini_nasil_anlatir : String
giris_tarihi : Tarih/Saat
kayit_tarihi : Tarih/Saat
goruntulenme_sayisi : Sayı

- *mailkutusu* isimli tablo; kullanıcıların birbirleriyle mesajlaşmalarında kullanılan tablo.

  kullanici_id : Sayı (*kullanici* isimli tablodaki id alanıyla bağlantılı) (Foreign Key)
  msj_id : Otomatik Sayı
  mesaj : String
  kimden : String
  konu : String
  tarih : Tarih/Saat

- *favorilerim* isimli tablo; kullanıcıların birbirlerini favorilerine eklemek istediklerinde kullanılan tablo.

  ekleyen_id : Sayı (*kullanici* isimli tablodaki id alanıyla bağlantılı) (Foreign Key)
  eklenen_id : Sayı
  id : Otomatik Sayı
  tarih : Tarih/Saat

- *profil* isimli tablo; kullanıcı farklı bir kullanıcının sayfasına baktığında diğer kullanıcının sayfasına kimlerin baktığını görmesini sağlayan tablo.

  kim : Sayı (*kullanici* isimli tablodaki id alanıyla bağlantılı) (Foreign Key)

  kime : Sayı

  id : Otomatik Sayı

  tarih : Tarih/Saat

- *yorumlar* isimli tablo; kullanıcıların birbirleriyle ilgili yorum yazmalarını sağlayan tablo.

  ekleyen_id : Sayı (*kullanici* isimli tablodaki id alanıyla bağlantılı) (Foreign Key)

  eklenen_id : Sayı

  id : Otomatik Sayı

  tarih : Tarih/Saat

  yorum : String

- *arkadas_daveti* isimli tablo; kullanıcıların birbirleriyle görüşmelerine, arkadaş olmalarına imkan veren tablo.

  ekleyen_id : Sayı (*kullanici* isimli tablodaki id alanıyla bağlantılı) (Foreign Key)

  eklenen_id : Sayı

  id : Otomatik Sayı

  tarih : Tarih/Saat

  onay : String(Evet/Hayır)

- *apriori_kayitlari* isimli tablo; apriori algoritmasının kullanılabilmesi için oluşturulan tablo.

  kullanici_id : Sayı (*kullanici* isimli tablodaki id alanıyla bağlantılı) (Foreign Key)

  ozellik : String

5.2. Proje İçin Gerekli Formların Oluşturulması

Projede oluşturulan formlar;

5.2.1. Üye Kayıt Formu

Üye Kayıt Formu 3 aşamadan oluşmaktadır. İlk bölümde rumuz, ad, soyad, e-posta, şifre, cinsiyet bilgileri girilmektedir. Bu bilgiler girildikten sonra kişinin girmiş olduğu email adresine sistem tarafından oluşturulan aktivasyon kodu gönderilmektedir. Bu aktivasyon kodu sayesinde kişinin üye olduktan sonra bu kodu aktif ederek diğer bilgilerini güncellemesi sağlanmaktadır. Aynı zamanda aktivasyon kodu mail olarak gönderilip, maildeki aktivasyon kodunu aktif hale getirerek sisteme giriş yapılması sayesinde hackerların siteye sürekli boş kayıt yapması engellenmektedir. Aktivasyon kodu aktif edilmemiş kullanıcılar 20 gün sonra sistemden otomatik olarak silinmektedir. Böylece veritabanındaki gereksiz kayıtlar engellenmiş olur.



Şekil 5.2.1.1 Üye Giriş Formu 1.Bölüm

İkinci bölümde kullanıcıya mail olarak gönderilen aktivasyon kodunun girilmesi sağlanmaktadır. Girilen aktivasyon kodu veritabanındaki ile aynıysa aktif hale getirilip kullanıcının istediği işlemleri yapması sağlanır.



Şekil 5.2.1.2 Üye Giriş Formu 2.Bölüm

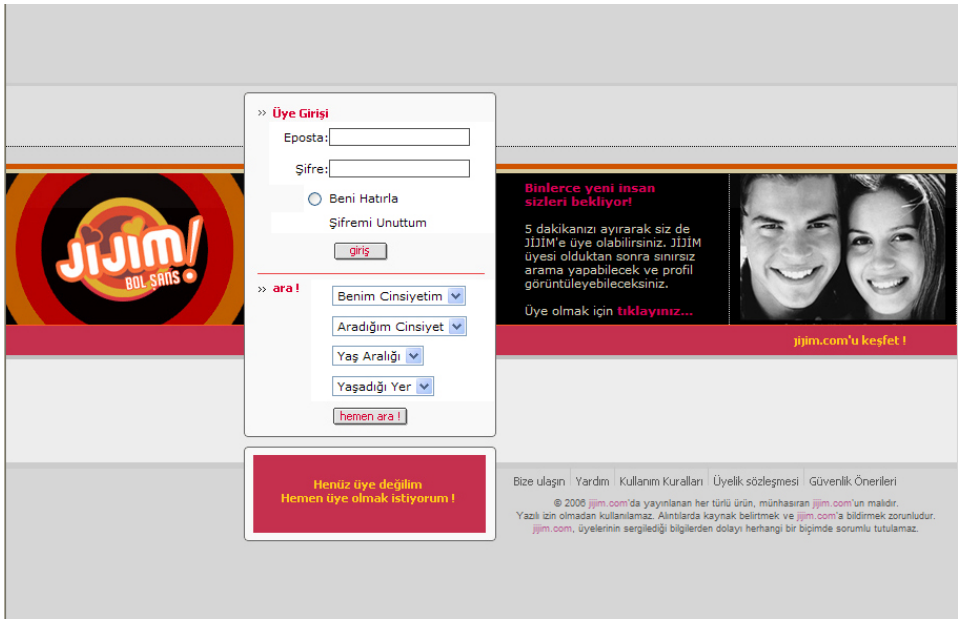
Üçüncü bölümde ise yaşanılan yer, cinsiyet, doğum tarihi, yaş, meslek, medeni durum, eğitim durumu, boy, kilo, saç rengi, göz rengi ve kullanıcının kendisini anlatan bilgiler girilmektedir.



Şekil 5.2.1.3 Üye Giriş Formu 3.Bölüm

5.2.2. Üye Giriş Formu

Kullanıcının e-posta adresi ve şifresiyle sisteme giriş yapabileceği bölümdür. Giriş yapıldıktan sonra e-posta adresi ve şifre doğruysa Bölüm 5.2.3' te anlatılan profil sayfası bölümüne geçiş yapılır.



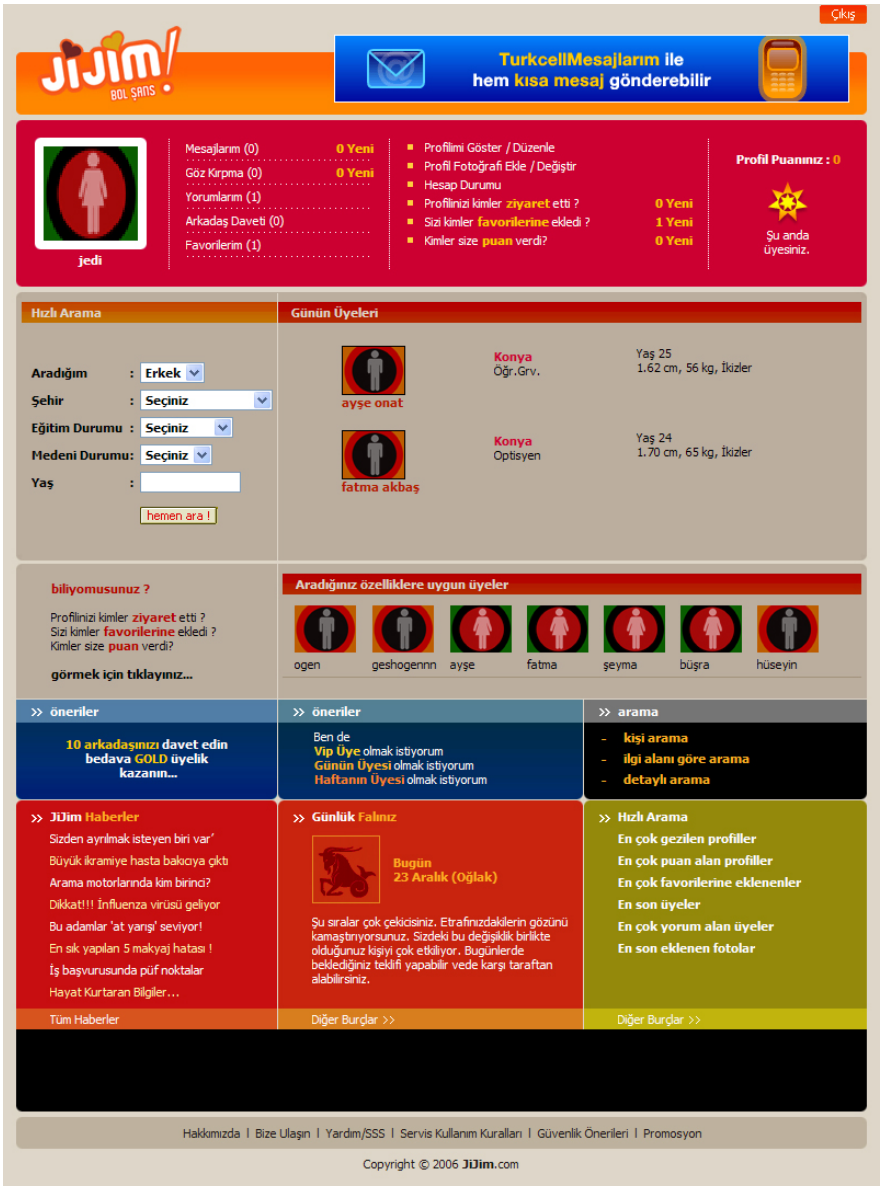
Şekil 5.2.2 Üye Giriş Formu

5.2.3. Profil Formu

Profil formu, kişinin kendisiyle ilgili tüm bilgilerini görebildiği formdur. Bu formda yapılabilecek özellikler;

- Kullanıcı istediği özellikteki kişileri arayabilmektedir.
- Kullanıcı istediği kişilerle http://www.yahoo.com veya http://www.hotmail.com sitelerinde olduğu tarzda mesajlaşabilmektedir.
- Kullanıcı istediği kişileri favorilerine ekleyebilmektedir.
- Kullanıcı kimlerin profil sayfasını ziyaret ettiğini görebilmektedir.
- Kullanıcı kendisine yazılan yorumları okuyabilmektedir.

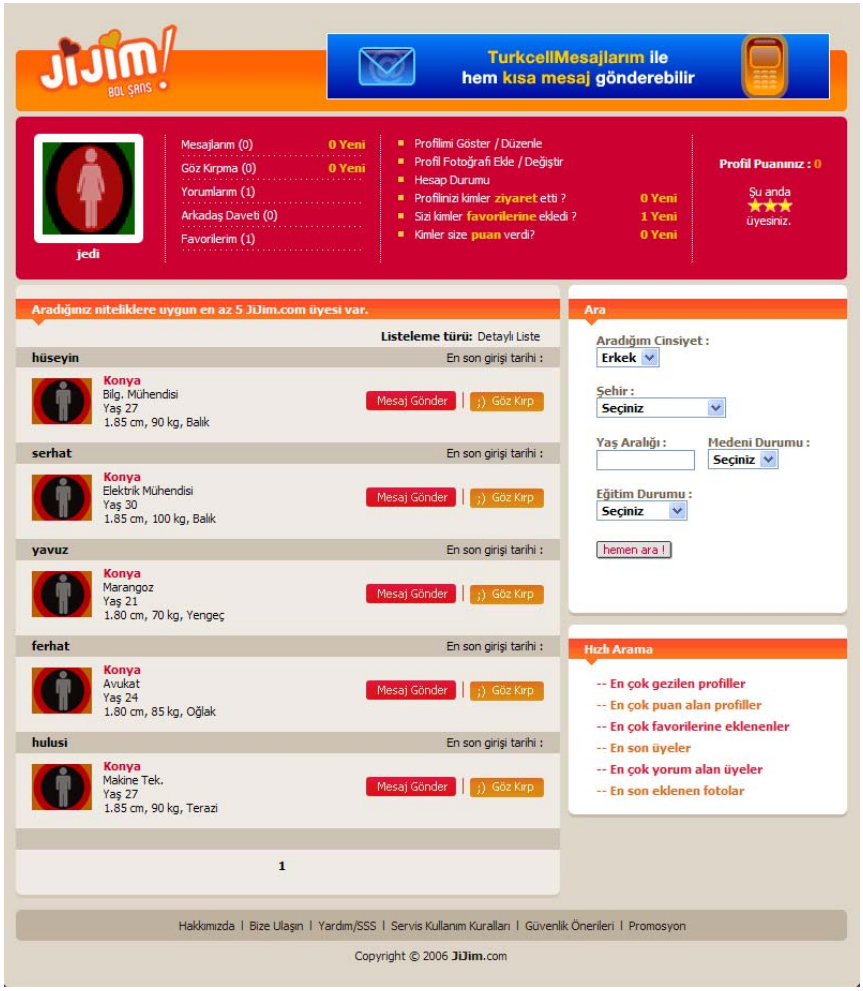
- Kullanıcı kendisiyle ilgili daha önceden belirtmiş olduğu özellikleri değiştirebilmektedir.
- Kullanıcı istediği kişilere davet göndererek onlarla arkadaş olabilmektedir.
- Kullanıcı kendi profil sayfasında listelenen başka kullanıcıları da kendi ekranında görebilmektedir.



Şekil 5.2.3 Profil Formu

5.2.4. Arama Formu

Kullanıcı istediği özelliklerde kişileri arayabilir. Aranan kişinin cinsiyeti, yaşı, hangi şehirde yaşadığı, medeni durumu ve eğitim düzeyi aranan özelliklerdendir. Bu özelliklere göre arama yapıldığında bu özellikleri barındıran diğer kullanıcılar ekranda listelenir. Bu listelemede kullanıcı diğer kullanıcıların profillerini inceleyebilir ve onlara mesaj atabilir.



Şekil 5.2.4 Arama Formu

Kullanıcı arama yaptığı zaman arama sırasında kullandığı özellikler veritabanına kaydedilmektedir. Kaydedilen bu verilere Apriori algoritması uygulanarak kullanıcının profil sayfasında daha önceden aradığı özelliklere uygun kişiler görüntülenmektedir. Bu sayede kullanıcının profil sayfasını açar

açmaz daha önceden aradığı özelliklerde kişiler görmesi onun siteye daha çok bağlanmasını sağlayacaktır. Üyeliğini sürekli devam ettirerek sitenin bu durumdan para kazanması sağlanacaktır.

5.2.5. Mesaj Listeleme Formu

Kullanıcı kendisine gelen ve kendisinin gönderdiği mesajları görebildiği formdur. Bu formda mesajın kimden geldiği, mesajın konusu, mesajın geliş tarihi ve mesajın durumu(okundu/okunmadı/cevaplandı) görüntülenmektedir.



Şekil 5.2.5 Mesaj Listeleme Formu

5.2.6. Mesaj Okuma Formu

Kullanıcı kendisine gelen ve kendisinin gönderdiği mesajları mesaj okuma formu sayesinde görebilmektedir. Bu formda mesajın kimden geldiği, mesajın konusu, mesajın içeriği, mesajın geliş tarihi ve mesajın durumu (yeni/eski) görüntülenmektedir. Aynı zamanda istenildiği takdirde mesaj silinebilir.



Şekil 5.2.6 Mesaj Okuma Formu

5.2.7. Mesaj Gönderme Formu

Kullanıcının istediği kişilere mesaj gönderebilmesini sağlayan formdur. Bu formda mesajın kime gönderileceği, mesajın konusu, mesajın içeriği, mesajın gönderiliş tarihi görüntülenmektedir.



Şekil 5.2.7 Mesaj Gönderme Formu

5.2.8. Yorumlar Formu

Bu formda kullanıcı, kendisinin başka kişilerle ilgili yazdığı yorumları ve kendisine yazılan yorumları görebilmektedir.



Şekil 5.2.8 Yorumlar Formu

5.2.9. Favoriler Formu

Kullanıcı kendisinin en çok beğendiği kişileri favorisi olarak belirleyebilmektedir. Favorilerim formunda kişinin favorisi olarak belirlediği kişilerle, hangi kullanıcılar kendisini favorisi olarak eklediyse o kullanıcılar listelenmektedir.



Şekil 5.2.9 Favoriler Formu

5.3. Apriori Algoritmasının Uygulanması

Arama işlemi sonucunda yapılan aramaların veritabanındaki apriori_kayitlari isimli tabloya kaydedilmiştir. Bu tablodaki verilere göre Apriori algoritması uygulanmaktadır. A kişisine ait tablo kayıtları Tablo 5.3.1 de gösterilmiştir.

	Durumlar
1	Kadın, 22, Bekar
2	22, Konya
3	22, Lisans
4	Kadın, 22, Konya
5	Kadın, Lisans
6	22, Lisans
7	Kadın, Lisans
8	Kadın, 22, Lisans, Bekar
9	Kadın, 22, Lisans

Tablo 5.3.1 *apriori_kayitlari* tablosundaki A kişisine ait veriler

Apriori Algoritmasının adım adım gösterimi;

minimum_destek=2 olarak alınmıştır.

İlk olarak tabloda bulunan her elemanın kaçar tane olduğu sayılır.

C1	
Kadın	6
22	7
Lisans	6
Konya	2
Bekar	2

minimum_destek=2 seçeneğine göre sayısı 2 den az olanlar silinir.

L1	
Kadın	6
22	7
Lisans	6
Konya	2
Bekar	2

L1'e göre ikili kombinasyonlar oluşturulur.

C2
{Kadın, 22}
{Kadın, Lisans}
{Kadın, Konya}
{Kadın, Bekar }
{22, Lisans}
{22, Konya }
{22, Bekar}
{Lisans, Konya}
{Lisans, Bekar}
{Konya, Bekar}

Kombinasyonları oluşturulan her elemanın kaçar tane olduğu sayılır.

C2	
{Kadın, 22}	4
{Kadın, Lisans}	4
{Kadın, Konya}	1
{Kadın, Bekar }	2
{22, Lisans}	4
{22, Konya }	2
{22, Bekar}	2
{Lisans, Konya}	0
{Lisans, Bekar}	1
{Konya, Bekar}	0

Kombinasyonları oluşturulan değerler arasında min_sup=2 seçeneğine göre sayısı 2 den az olanlar silinir.

L2	
{Kadın, 22}	4
{Kadın, Lisans}	4
{Kadın, Bekar }	2
{22, Lisans}	4
{22, Konya }	2
{22, Bekar}	2

L2'e göre üçlü kombinasyonlar oluşturulur.

C3	
{Kadın, 22, Lisans}	2
{Kadın, 22, Bekar }	2

Kombinasyonları oluşturulan değerler arasında minimum_destek=2 seçeneğine göre sayısı 2 den az olanlar silinir.

L3	
{Kadın, 22, Lisans}	2
{Kadın, 22, Bekar }	2

I={Kadın, 22, Bekar} olağan elemankümeyi içeren veriyi düşünüldüğünde I'dan çıkarılabilecek ilişkisel kurallar; {Kadın, 22}, {Kadın, Bekar}, {22, Bekar}, {Kadın}, {22} ve {Bekar} I'nın boş olmayan alt kümeleridir. Böylece I'nın ilişkisel kuralları aşağıdaki gibi hesaplanır.

Kadın$\wedge$22$\Rightarrow$Bekar Güven değeri=2/4=50%

Kadın$\wedge$Bekar$\Rightarrow$22 Güven değeri=2/2=100%

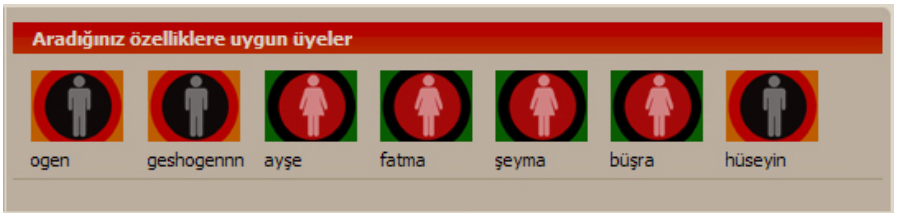
22$\wedge$Bekar$\Rightarrow$Kadın Güven değeri =2/2=100%

Kadın$\Rightarrow$22$\wedge$Bekar Güven değeri =2/6=33%

22$\Rightarrow$Kadın$\wedge$Bekar Güven değeri =2/7=29%

Bekar$\Rightarrow$Kadın$\wedge$22 Güven değeri =2/2=100%

Güven değerinin eşiğinin %70 olarak alırsak sadece 2. 3. ve 6. kural güçlü olacaktı. Kullanıcının profil sayfasında listelenen diğer kişileri gösterirken oluşturacağımız yeni sorgu 2.,3. ve 6. kurala göre olacaktır. Bu özelliklere göre listelenecek kişiler, kullanıcının profil sayfasında görüntülenmiştir. Şekil 5.3.1' de profil sayfasında aranan özellikteki kişiler listelenmiştir. Böylece kullanıcı sayfaya her girdiğinde daha önceden aradığı özelliklerde kişileri karşısında görünce üyeliğini devam ettirmek isteyecektir. Böylece web sitesinin sahibi de parasal anlamda kazanç sağlanacaktır.



Şekil 5.3.1 Apriori Algoritması'nın uygulanması sonucu profil sayfasında listelenen kullanıcılar

6. SONUÇLAR

Internet kullanımının gün geçtikçe daha da yaygınlaşması Internet sunucuları üzerinde tutulan verilerin de hızlı bir şekilde artmasına, sunucularda veri yığınlarının oluşmasına neden olmuştur. Web tabanlı olarak saklanan bu verilerin üzerinde birtakım algoritmalar geliştirilerek faydalı bilgilerin ortaya çıkarılması Web Madenciliği teknikleri sayesinde gerçekleştirilmektedir.

Bu çalışmada ve bugüne kadar yapılan çalışmalar sayesinde, veri madenciliği tekniklerini kullanılarak Internet sitelerinin kullanımını daha da arttırabilmesi ve aranılan bilgiye daha çabuk ulaşılabilmesi sağlanmıştır.

Bu çalışmada veri madenciliği, web madenciliği, web madenciliğinin alt bölümleri, Apriori algoritması ve bunların uygulama alanları incelenmiştir. Apriori algoritması sık olarak kullanılan veriler arasından yeni bilgiler oluşturmayı sağlamaktadır. Veri madenciliğinde sıkça kullanılan Apriori algoritmasıyla bir Internet sunucusunda kayıtlı olan kullanıcıların yaş, cinsiyet, yaşadığı şehir, lisans düzeyi gibi özelliklerine göre uyumlulukları bulunmuştur. Böylece kişi sürekli aradığı özelliklerde farklı bir kullanıcıyı sayfasında gördüğü zaman siteye olan üyeliğini devam ettirmektedir. Bu sayede sunucu sahibinin daha çok reklam alması ve daha çok kâr etmesi sağlanmıştır.

Bu çalışmanın bugüne kadar yapılan çalışmalardan farkı ise veri toplanması işlemidir. Yani kullanılacak olan verilerin elde ediliş yöntemlerinin farklı olmasıdır. Diğer çalışmalarda kullanıcıların girdikleri, gezdikleri sayfalar hafızada tutulup o sayfalar doğrultusunda işlem yapılırken bu projede arama sırasında kullanılan verilere göre işlem yapılmıştır. Yani kullanıcı arama yaptığı anda arama için kullandığı özellikler veritabanına kaydedilmiştir. Veritabanına kaydedilen bu veriler üzerinde uygulanan Apriori Algoritması sonucunda %70 oranında başarı sağlanmıştır. Yani kişilerin yaptıkları aramalar sonucunda kendi profillerinde listelenen “Aradığınız Özelliklere Uygun Üyeler” başlıklı kısımda %70 oranında doğru sonuçlar elde edilmiştir.

Gelecekte, web madenciliği teknikleri, firmaların yeni stratejiler geliştirmesinde, kâr amaçlı uygulamalarda kullanabilmesinde ve rakiplerinin bir adım daha önüne geçilebilmesinde büyük kolaylık sağlamaya devam edecektir. Çünkü veri kendi başına kullanıldığında değersizdir ama üzerinde birtakım işlemler yapıldığında daha faydalı hale dönüşebilir.

7. KAYNAKLAR

Altıntop Ü., *"Internet Tabanlı Öğretimde Veri Madenciliği Tekniklerinin Uygulanması"*, Y.Lisans Tezi, 2006

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

Daş R., Türkoğlu İ., Poyraz M., *"Genetik Algoritma Yöntemiyle Internet Erişim Kayıtlarından Bilgi Çıkarılması"*, SAÜ Fen Bilimleri Enstitüsü Dergisi 10. Cilt, 2.Sayı, s. 67-72, 2006

Etzioni O., *"The World Wide Web: Quagmire or Gold Mine"*, in Communications of the ACM, 39(11):65-68, 1996.

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

Özekes S., *"Veri Madenciliği Modelleri ve Uygulama Alanları"*, İstanbul Ticaret Üniversitesi Dergisi

Srivastava J., *"Web Mining :Accomplishments & Future Directions"*

Şakiroğlu A.M., Tuğ E., Bulun M., *"Web Log Dosyalarından Genetik Algoritma öntemiyle Sıralı Erişimlerin Tespit Edilmesi"*, Türkiye Bilişim Derneği 20. Bilişim Kurultayı, 2003.

Uğuz H., Kodaz H., Saraçoğlu R., Baykan Ö.K.,*"Genetik Algoritmalar Kullanılarak Web Kullanım Madenciliği Yönteminin Sistem Log Kayıtlarına Uygulanması"*, International XII. Turkish Symposium on Artificial Intelligence and Neural Networks – TAINN 2003, T-1, s. 45 - 47, 2003.

Zhong N., Zhou L., *"Methodologies for Knowledge Discovery and Data Mining"*, Third Pacific-Asia Conference, Pakdd-99, Beijing, China, Proceedings, *Springer Verlag,* 1999.

Han J., Kamber M., *"Data Mining: Concepts and Techniques"*, 2001

http://www.recepayaz.com/library/web_madenciligine_bir_bakis.doc, Ocak 2008

http://www.bilyaz.com/bMakaleGetir.php?id=56, Ocak 2008

http://enm.blogcu.com/5187251, Ocak 2008

http://www.yapay-zeka.org/modules/wiwimod/index.php?page=ANN&back=WiwiHome, Ocak 2008

http://www.bilgisayaransiklopedisi.com/bilgisayar/496/giris.html, Ocak 2008

http://www3.itu.edu.tr/~sgunduz/courses/verimaden/slides/d7.pdf, Ocak 2008