AKDENİZ ÜNİVERSİTESİ

SOSYAL BİLİMLER ENSTİTÜSÜ

Ömer UÇAN

DİJİTAL KÜTÜPHANELERDE VERİ MADENCİLİĞİ UYGULAMALARI: AKDENİZ ÜNİVERSİTESİ MERKEZ KÜTÜPHANESİ ÖRNEĞİ

İşletme Anabilim Dalı

Yüksek Lisans Tezi

Antalya, 2010

AKDENİZ ÜNİVERSİTESİ

SOSYAL BİLİMLER ENSTİTÜSÜ

Ömer UÇAN

DİJİTAL KÜTÜPHANELERDE VERİ MADENCİLİĞİ UYGULAMALARI: AKDENİZ ÜNİVERSİTESİ MERKEZ KÜTÜPHANESİ ÖRNEĞİ

Danışman

Doç. Dr. Can Deniz KÖKSAL

İşletme Anabilim Dalı

Yüksek Lisans Tezi

Antalya, 2010

İ Ç İ N D E K İ L E R SAYFA

ŞEKİLLER LİSTESİ

TABLOLAR LİSTESİ

ÖZET

Veri madenciliği, dijitalleşme sürecinde ortaya çıkan veri yığını içerisinden elde edilen sağlıklı bilgilerin, belirli işlemlerden geçirilerek karar vericiler için önemli bulguların ortaya konulmasını sağlayan bir süreçtir. Bu çalışmada kütüphanelerde oluşturulan veri yığınlarının arasından veri madenciliği tekniği ile elde edilebilecek bulgulara örnekler verilmiştir. Bu işleyiş içerisinde hangi adımların takip edilmesi gerektiği konusu üzerinde durulmuş, uygulama aşamasında Akdeniz Üniversitesi Merkez Kütüphanesi veritabanı üzerinde çalışmalar yapılmıştır.

Çalışmanın ilk bölümlerinde veri madenciliği kavramı üzerinde durulmuş ve bilimsel açıdan özellikleri anlatılmaya çalışılmıştır. Veri madenciliği sürecinin önemli bir bölümünü oluşturan veri hazırlık aşaması, detaylarıyla ortaya konulmuştur. Yine veri madenciliği araçları sayesinde dijital kütüphanelerde yapılan çalışmalar ile bu sürecin kütüphanelerde nasıl işlemesi gerektiğine dair bulgular bölüm sonunda paylaşılmıştır. Veri madenciliği sürecinde kullanılan modellemeler ve algoritmalar detaylı bir şekilde açıklanmış, kullanım alanları hakkında bilgiler verilmiştir.

Uygulama bölümü kütüphane içi kullanıcı ve yayın bilgileri yardımıyla yayın ödünç sirkülâsyonu verileri üzerinde çeşitli modellemeler uygulanmıştır. Modellerin kurulmasından önce veriler üzerinde ciddi bir ön-işleme süreci gerçekleştirilmiş, veriler analizler için en uygun duruma getirilmiştir. Birliktelik ve kümeleme modellerinde Apriori ve TwoStep algoritmaları uygulanmış ve kütüphane karar vericilerine yardımcı olabilecek anlamlı veriler elde edilmeye çalışılmış, grafiklerle zenginleştirilmiştir.

Anahtar Kelimeler: Veri madenciliği, Dijital Kütüphaneler, Apriori Algoritması, İki Adım Algoritması

ABTRACT

DATA MINING APPLICATIONS IN DIGITAL LIBRARIES : AKDENIZ UNIVERSITY CENTRAL LIBRARY

Data mining is a process of an information that is produced from data clusters that occurs during the period of digitalization in order to produce important findings for decision-makers. In this study, the examples of findings in terms of data mining among data clusters that were generated in libraries are included. It is emphasized that what steps should be followed and some studies took part that are completed in Central Library of the Akdeniz university in this process.

In the first chapter of this study data mining is emphasized and tried to explain the features in terms of scientific grounds. The preparation step which is an essential part of data mining is stated elaborately. Yet, the studies performed in digital libraries and the findings point that how the process to work are stated in the end of the chapter by means of data mining tools. Also the modelling and algorithms that are used during the process of data mining are emphasized in there.

Various modelling is applied on data of issue-loan-circulation by means of information of users and issue of the application section of library. An important pre-processing took part in data before setting up the models thus the data was gotten to suitable to analysis. Association and cluster models are processed by Aprioi and TwoStep algorithms in order to gather useful data to pass to decision-makers of library, also enriched with charts.

Keywords : Data Mining, Digital Libraries, Bibliomining, Apriori Algorithm, TwoStep Algorithm

GİRİŞ

Bilginin değerini her geçen gün arttırdığı günümüz dünyasında teknolojinin gelişmesiyle birlikte bilgi, veri yığınları şeklini almaya başlamıştır. Bir yandan her türlü bilgiye erişimi kolaylaştıran bu durum, diğer taraftan kirlenen veriler içinde gerçek bilgiye ulaşılmasında engeller çıkarmaktadır. Bu yüzden veri madenciliği gün geçtikçe önemini arttıran bilimsel bir araçtır. Özellikle kullanıcı odaklı işlemlerin günümüz işletmelerince ön planda tutulması, bu amaçla kullanıcı işlemleri sonucu oluşan birçok verinin kayıt altına alınması büyük veri yığınlarının ortaya çıkmasına neden olmaktadır. Son yıllarda kullanıcı tercihlerinin belirlenmesi konusunda düzensiz bir şekilde tutulan müşteri bilgileri, işlemsel kayıt dosyaları ya da sipariş bilgileri önemli birer örnektir.

Kullanıcı tercihlerinin takibi ile başlayan ve bu tercihlerden anlamlı sonuçları çıkarılmasına kadar uzanan veri madenciliği sürecinin en önemli bölümü sağlıklı verilerin elde edilmesi aşamasıdır. Verileri sistematik bir düzen içerisinde ele alarak verilerin temizlenmesi, uyarlanması ve dönüştürülmesi ana unsurları oluşturmaktadır. Özellikle hangi verilerin hangi modellemelerde kullanılabileceğinin önceden kestirilmesi ve bu doğrultuda verilerin hazırlanması titizlikle gerçekleştirilmelidir. Veriler içinde bulunan hatalar genellikle sistemli bir şekilde tekrarlanan kirli verilerdir. Bu nedenle özelleştirilmiş yazılımlarla bu hataların tespiti veri madenciliği açısından doğru bir adım olacaktır. Yine benzer yazılımlar verilerin dönüştürülmesi için kullanılmalıdır. Veriler üzerinden modellemelerde kullanılabilecek yeni verilerin elde edilmesi dönüştürme ve uyarlama işlemleri sırasında gerçekleştirilmelidir. Bu verilerin, temizlenmiş veri üzerinden gerçekleştirilmesi gerekliliği unutulmamalıdır.

Veri madenciliğinin düzenleyici ve kestirimci işlevi birçok alanda kullanıldığı gibi kütüphanelerde de kullanılmaktadır. Kütüphaneler pek çok işletmenin ya da organizasyonun hedeflediği gibi kullanıcı memnuniyetini ön planda tutmak istemektedirler. Kullanıcıların teknolojiye bakış açısı değiştikçe dijitalleşme süreci kaçınılmaz bir hale gelmiş ve kütüphane kavramının tekrar baştan oluşmasına neden olmuştur. Günümüzde kütüphaneler sadece bina ve içindeki yayınlardan oluşmamaktadır. Kütüphaneler, yayınlarının internet aracılığıyla kütüphane dışından erişimine olanak sağlamış ya da yine sadece internet ortamından belirli

konulara odaklanmış özel yayın araçları haline gelmiştir. Bu durum dijitalleşme sürecinde kaçınılmaz bir gelişmedir. Büyük ölçekli kütüphaneler bu süreci iyi yönetirken orta ve küçük ölçekli kütüphaneler ise bu sürecin henüz başlarında bulunmaktadır. Doğaları gereği bu durum kütüphane karar vericileri için temkinli davranmayı gerektiren bir süreç olmasına rağmen aynı zamanda kaçınılmazdır. İnsanoğlunun bilgiye istediği yerde ve zamanda rahatlıkla erişebileceğinin farkına varması, ister istemez bu sürecin gün geçtikçe daha da hızlanması için küçük veya büyük tüm şirketleri ve organizasyonları bu sürecin içinde yer almaya zorlamaktadır.

Kütüphaneleri dijital ortamda farklı işlevlerle görmek mümkündür. Geleneksel kütüphane anlayışını sürdüren kütüphaneler olduğu gibi, tamamen özelleştirilmiş kütüphaneler de ortaya çıkmaktadır. Özellikle makalelerin internet ortamında erişimine imkân sağlayan çok sayıda dijital kütüphane popülerliklerini arttırmaktadır. Bunun yanında internet kütüphaneleri olarak adlandırılabilen ve sadece internet ortamında bulunan ve yayınlara sadece internet üzerinden erişim sağlayan kütüphanelerden de söz etmek mümkündür. E-kitap kavramı ile birlikte çok fazla çeşitlenen dijital kütüphanelerden en fazla faydayı şüphesiz ki kullanıcılar sağlamaktadırlar. Gerçekleştirilen çalışmada da kütüphanelerin dijitalleşme sürecinin başında karşılaşabilecekleri zorluklar üzerinde durulmuş ve bu sürecin aslında kütüphane karar vericileri için ne gibi avantajları olduğuna vurgular yapılmıştır. Veri madenciliği ile bilgi keşfi değişen kütüphanecilik kavramını anlayabilmek için gerekli bir araçtır.

Çalışma içerisinde ilk bölümde veri madenciliği kavramı üzerinde ve dijital kütüphanelerin gelişimi üzerinde durulmuş, veri madenciliğinde kullanılan sınıflandırmalar ve algoritmalar ikinci bölümde detaylı bir şekilde açıklanmıştır. İstatistik, Karar ağacı, geri yayılım, kümeleme ve birliktelik sınıflandırmaları algoritmaları ile birlikte ortaya konulmuştur. Uygulamaların amaçlarına göre avantajları ve dezavantajları anlatılan bu bölümde algoritma davranışları açıklanmıştır. Uygulama bölümü ise Akdeniz Üniversitesi Merkez Kütüphanesi verileri üzerinde yapılan çalışmaların sonuçları üzerinde durulmuştur.

1. BÖLÜM VERI MADENCILIĞI

1.1 Veri Madenciliği Kavramı

Veri Madenciliği en basit tanımıyla bilgiyi büyük veri yığınları içerisinden çıkarmak ya da "kazmaktır" (Han & Kamber, 2001, s. 5). Yirminci yüzyılın ortalarından itibaren teknolojinin gelişmesiyle birlikte ortaya çıkan dijitalleşme süreci ile veri kaynakları artmış, ancak bilgiye ulaşmak daha kolay hale gelmemiştir. Kamu alanında, bilimsel çalışmalarda ve iş hayatında devamlı büyüyen veri yığınları kaydedilmektedir. 1990 sonrası veritabanlarının hızlı gelişimi ile birlikte bilgileri keşfetme ve analiz etme, karışık bir işlem haline dönüşmüştür. Büyük veri yığınlarının çok küçük miktarı kullanılmaktadır. Çoğu durumda veriler kontrol edilebilirlik açısından çok büyük, analiz edilebilirlik açısından düzensiz ve anlamsız bir durumdadır (Kantardzic, 2001, s. 7). Veri madenciliği bu çelişki üzerine kurulmuştur. Yöneticilerin elinde büyük miktarda veri yığınları bulunmasına rağmen yöneticiler için bir anlam ifade etmeyebilmektedir. Bu durum zamanla geliştirilen tekniklerle yeni bir uygulama alanı açmış ve veri madenciliği kavramını ortaya çıkarmıştır.

"Veritabanlarında Bilgi Keşfi" olarak da adlandırılan veri madenciliği yapay zekâ ve istatistik biliminin bir arada kullanımıdır. İstatistik biliminden faydalanarak oluşturulan algoritmalar sayesinde çoklu ve karmaşık analizler oluşturulabilmekte, geleceğe dair farklı senaryolar ortaya konabilmektedir. Bilgisayarlar ve yapay zekâ sayesinde çeşitli öngörülerde bulunulabilmektedir. Böylece veri madenciliği sadece geçmişe dayalı bir istatistiksel çıkarımdan çok karar vericiler açısından geleceğe yönelik bir yön gösterici olarak kullanılabilmektedir.

Veri madenciliğinin amacı büyük miktardaki ve büyük çoğunluğu denetlenmemiş veriyi tanım kümesi içerisinde anlamlandırmaktır (Horis, Pedrycz, Swiniarski, & Kurgan, 2007, s. 9). Veri madenciliğinde modellemeler yapılırken sadece veri yığınları arasında değil, tek bir veri ile diziler arasındaki ilişkiler de göz önüne alınır. Bu nedenle veri madenciliğinde modelleme çeşitlerini belirlenerek kurallar bütünü haline getirilmeli, veriler gruplandırılarak anlamlı ilişkiler araştırılmalıdır. Bu ilişkisel çıkarımlar veri madenciliği uzmanları tarafından yapılmakla beraber yöneticiler, veri madenciliğini yorumlaması gereken karar vericiler için anlaşılabilir düzeyde olması yine veri madenciliği uzmanlarının sorumluluğundadır. Karar

vericiler için gerekli bilgiyi yorumlayacağı şekilde sunmak, araştırmanın güvenirliliğini sağlamak, veri madenciliği uzmanlarının görevidir.

1.2 Veri Setleri

Veri setleri ham veya işlenmiş ölçümlerdir. Veri madenciliği verilerin kalitesine ve miktarına bağlı olarak değerlendirilir. Veriler birçok yapıda sınıflandırılmış ya da sınıflandırılmamış şekilde depolanabilmektedir. Tüm bu verilerin bir arada bulunmasıyla oluşan Veri setleri en basit anlamıyla çevresel ya da işlemsel süreçlerden elde edilen değerlerdir. Tek başına anlamlandırılmaya müsait olmayan veriler, veri setleri haline geldiklerinde bir anlam ifade edilebilir ve çıkarım yapılabilir hale gelebilirler. Ancak bu noktaya gelmeden önce verilerin birçok aşamadan geçmesi zorunluluktur.

Bununla birlikte günümüzde veri akışının kontrol edilebilirlikten uzak olması, veri madenciliği için verileri yapılandırılmış, yarı yapılandırılmış ve yapılandırılmamış olarak gruplandırma zorunluluğu getirmiştir. Çoklu ortam (Mültimedya) dosyaları, yapılandırılmamış veriler arasında en sık karşılaşılanıdır. Kapasite probleminin ortadan kalkmasıyla görsel ve işitsel verilerin varlığı artmakta, ancak bu verilerin içerdiği bilgiler kategorize edilmediğinden sağlıklı bir hale dönüştürülmesi zorunluluğu ortaya çıkmaktadır. Yarı yapılandırılmış verilere örnek ise yazılı dokümanlar gösterilebilir. İşletmelerde kullanılan dokümanlar, raporlar ve yazışmalar yeni nesil veritabanları sayesinde kayıt altına alınabilmekte ve işlenebilir veri olarak kullanılabilmektedir. İçerik olarak sınıflandırılmadan veri madenciliğinde kullanılmamalı, yapılandırılmamış veriler de olduğu gibi ham veri setinden, kullanılabilir veri setine dönüştürülmesi gerekmektedir (Kantardzic, 2001, s. 5).

Yapılandırılmış veriler ise veri madenciliği için kullanılmaya en uygun verilerdir. Bu veri setlerinde önemli olan nokta hangi bilginin bizim için işe yarar durumda olduğunu belirleyebilmektir (Hand, Mannilla, & Smith, 2001). Ancak dikkat edilmesi gereken husus alanların niceliksel ya da kategoriksel olup olmadığıdır. Yaş ve gelir gibi değerler herhangi bir değer almaları gerekirken, cinsiyet medeni durum gibi belirler daha önceden belirlenmiş bazı kategoriler altında toplanmalıdırlar. Bu ayrım analitik teknikler açısından önemlidir. Niceliksel ya da kategorisel verilere göre analitik teknikler uygulanmalıdır.

1.3 Veritabanları ve Veri ambarları

1.3.1 Veritabanları

Veritabanı veya Veri düzlemi düzenli bilgiler topluluğudur. Kelime anlamı bilgisayar ortamında saklanan düzenli verilerle sınırlı olmamakla birlikte, daha çok bu anlamda kullanılmaktadır. Bilgisayar terminolojisinde veritabanı, sistematik erişim imkânı olan, yönetilebilir, güncellenebilir, taşınabilir, birbirleri arasında tanımlı ilişkiler bulunabilen bilgiler kümesidir (Wikipedia, Database). Yazılımların ilk geliştirildiği yıllarda veriler basit metin dosyaları içerisinde saklanmaktayken, gelişen ve büyüyen veri yığınları her geçen gün yeni ve daha gelişmiş veritabanlarına ihtiyaç duyulmasına sebep olmaktadır. Gelişmiş veritabanları gün geçtikçe daha çok nesne odaklı ve nesne ilişkili hale gelmektedir.

Veri Madenciliğinde çoğunlukla büyük veri yığınları ile çalışılmaktadır. Büyük veri yığınları ise gelişmiş ve özelleştirilmiş yönetim sistemlerine ihtiyaç duymaktadır. Bunun dört sebebi vardır (Horis, Pedrycz, Swiniarski, & Kurgan, 2007, s. 95).

- Veri madenciliği için kullanılan veri büyüklüğünün güçlü donanımsal özelliklere ihtiyaç duyması
- Kullanılan veri madenciliği metotlarının farklı veri alt kümelerine gereksinim duyması. Buna bağlı olarak gerekli verilerin doğru şekilde kümelendirilmesi zorunluluğu
- Gerekli verilerin eş zamanlı eklenmesine ya da güncellenmesine farklı zamanlarda farklı kişilerce gereksinim olması
- İhtiyaç duyulan verilere ulaşmak için veri yığınları arasından belirli bölümlere ihtiyaç duyulması

Yukarıda belirtilen sebepler veritabanı yönetim sistemlerinin gerekliliğini ortaya koymaktadır. Veritabanı yönetim sistemi, birbiriyle ilişkili veritabanlarından meydana gelen verileri yönetmek ve verilere erişim sağlamak için kullanılan programlardır (Han & Kamber, 2001, s. 10). 1960'ların sonlarından itibaren ortaya çıkmaya başlayan veritabanı yönetim sistemleri, dosyalama sistemleri olarak oluşturulmasına rağmen günümüzde gelişmiş veritabanı sorgulama dilleri ile yönetimsel alanda kullanımı kolaylaştırılmıştır. İyi bir veritabanı yönetim sistemi aşağıdaki özelliklere sahip olmalıdır (Garcia-Molina, Ullman, & Widom, 2008):

- Yeni veritabanları ve şema oluşturabilmesini sağlaması
- Kullanıcılarına veri sorgulama şansı vermesi
- Büyük veri yığınlarını depolama imkânı sağlaması
- Verilere çoklu erişimi sağlaması

Günümüzde gelişmiş veritabanı sistemleri; uzamsal verileri(örneğin haritalar), mühendislik tasarım verileri, hipermetin ve internet ortamında oluşturulan verileri de içermektedir. Oluşturulan sistemler yapılandırılmamış veya yarı yapılandırılmış olsa dahi verileri işleyebilme ve bu verileri farklı metotlarla yeniden dinamik ortamda şekillendirmeye yardımcı olurlar.

Gelişmiş veritabanı sistemlerini yedi kategori altında toplayabiliriz (Han & Kamber, 2001);

- Nesne-yönlü veritabanı sistemleri
- Nesne-ilişkili veritabanı sistemleri
- Uzamsal veritabanı sistemleri
- Zaman serileri veritabanı sistemleri
- Metin ve çoklu ortam veritabanı sistemleri
- Heterojen Veritabanı sistemleri
- World Wide Web

(i) Nesne yönlü veritabanları

Son yıllarda bilgisayar uygulamalarının dosya bazlı veri saklama işlevleri yeterliliğini kaybetmeye başlamasıyla birlikte veri işleme dünyasında hızla yükselen veritabanları, uygulama yazılımları için de ihtiyaç haline gelmiş ve nesne yönlü veritabanlarını ortaya çıkmasını sağlamıştır (Doğaç, Özsu, Biliris, & Sellis, 1994). Yazılımlar için geliştirilen programlama dilleri ile sadece verileri değil ses, görüntü ve çeşitli dokümanlarını yönetmek ihtiyaç haline gelmiştir. C++, Visual Basic, .Net, Pyhton gibi nesne yönlü programlama dillerinin doğmasını sağlamıştır. Web tabanlı uygulamaların gelişmesi ile artık Nesne yönlü veritabanları geniş bir kullanım alanına erişmiştir.

Nesne yönlü veritabanları nesnelerin toplanması ile ilgilidir. Her Nesnenin bir durum ve davranışı mevcuttur. Nesne davranışı nesnenin durumunu günceller. Veri ve diğer veriler arasındaki ilişkiler (davranışlar) basit bir yapı olan nesnelerle tanımlanmıştır. Nesneler tek başlarına bir anlam ifade etmezler. Nesneyi açıklayan değişkenler, diğer nesnelerle iletişim kurulabilmesi için iletişim setleri ve bu iletişimi sağlayacak metotları barındırmalıdır.

Nesne yönlü veritabanları, veritabanı yönetim sistemi ve nesne-yönlü sistemlerin özellikleri taşımalıdır. Veritabanı yönetim sistemi olmasından dolayı devamlılık, ikincil depolama yönetimi, uyumluluk, geri döndürülebilirlik ve özel amaçlı sorgu araçları barındırmalıdır. Nesne yönlü veritabanı olmasından dolayı karmaşık nesneler, nesne kişilikleri, kapsülleme, sınıflandırmalar barındırmalıdır (Atkinson, Bancilhon, DeWitt, Dittrich, Maier, & Zdonik, 1989).

(ii) Nesne ilişkili veritabanları

Nesne ilişkili veritabanları ilişkisel modellerin bulunduğu veritabanlarında karmaşık nesne ve nesne yönelimlerine zengin veri kaynağı sağlamak için kullanılır. İlişkisel veritabanlarında bulunan karmaşık veri çeşitleri, sınıf hiyerarşisini ve kalıtımı geniş bir alanda kullanımını sağlar (Han & Kamber, 2001, s. 17). Bu kullanım veritabanı şemaları ve sorgulama dilleri ile desteklenir.

İlişkisel veritabanları ile Nesne yönlü veritabanlarının arasında yer alan nesne ilişkili veritabanları, ilişkisel veritabanlarında olduğu gibi sınırlı bir veri yığınına kapsamaz, geliştiricilerin kendi modelleri ve metotlarını ilişkisel veritabanlarına adapte etmesini sağlar.

(iii) Uzamsal Veritabanları

Uzamsal veritabanları bir alanı kapsayan geometrik, coğrafik ya da uzamsal verileri barındırır. Bu, dünyanın iki boyutlu yüzeysel şekilleri olabileceği gibi, tıp araştırmalarında insan beyninin yapısı ya da kimyasal araştırmalarda üç boyutlu protein molekülleri de olabilir.

Uzamsal veritabanları yukarıda belirtilen verileri barındırsa da alfasayısal verilere her zaman ihtiyaç duymaktadır. Diğer veritabanlarından ayıran nokta ise uzamsal verilerle ilgili yapabildiği ek işlemlerdir. Uzamsal veriler nokta, çizgi ve bölge şeklinde tutulmaktadır.

Uzaysal veritabanları ise bu bilgileri iki veya üç boyutlu hale getirerek birbirleri arasındaki ilişkiyi ortaya koyabilmektedir (Güting & Hartmut, 1994).

(iv) Zaman Serisi Veritabanları

Zaman serisi veritabanlarında veriler, zaman aralıklarından çıkarım vasıtasıyla elde edilen ilişkili sayılardan oluşur. Bu veriler örneğin hisse senedi borsalarında olduğu gibi zaman içerisinde değişen ardışık değerler olabilirler (Han & Kamber, 2001, s. 418). Zaman serisi veritabanları ile kurulan sistemler zaman serilerini belirlenen bir düzen içerisinde oluşturur, değiştirir ve günceller. Bu düzen genelde hiyerarşik bir düzen içerisinde gerçekleşir.

Zaman serisi veritabanı sistemlerinde, veri madenciliği özellikle strateji planlamada önemli bir araç olarak kullanılmaktadır. Özellikle yatırım araçlarında teknoloji ile birlikte gelişen anlık verilerin izlenimi zaman serilerine dayanarak tahminleme yöntemlerini önemli kılmaktadır.

(v) Metin ve Multimedya Veritabanları

Metin veritabanları anlamlı bir yapısının olduğunu bildiğimiz ancak metinden yapısı hakkında bilgi alamadığımız, bu bilgiyi elde etmek için diğer ilişkili verilere ihtiyacımız olan yapılandırılmamış verilerden oluşur. (Doedens, 1994). Yapılandırılmamış veriler salt metinlerden oluşmaktadır. İnternet ortamında kaydedilen veriler, masaüstü yazılımları tarafından saklanan veriler buna örnektir. Ancak metin veritabanları her zaman yapılandırılmamış metinlerden oluşmayabilir. E-posta içerikleri, programlama dil kodları gibi yarı-yapılandırılmış metinler, ya da dijital kütüphanelerde kullanılan yarı yapılandırılmış içerikler de metin veritabanlarında kullanılabilmektedirler.

Multimedya veritabanları ses, görüntü resim gibi verilerin kaydedilmesi için geliştirilmiştir. Multimedya veritabanlarının ortaya çıkış nedeni verilerin boyutlarının çok büyük olması, bu yüzden de özelleştirilmiş depolama ve arama tekniklerine ihtiyaç duyulmasından dolayıdır (Horis, Pedrycz, Swiniarski, & Kurgan, 2007, s. 36).

(vi) Heterojen Veritabanları

Heterojen veritabanları birbirinden bağımsız veritabanlarının kullanıcı sorgularında ilişkisel sonuçları gösterebilmek için iletişim halinde olması ile oluşur. İlişkisel veritabanlarında nesneler tamamen farklı olabilir. Bu yüzden ilişkileri belirlemek oldukça zordur.

(vii) World Wide Web (Dünya Çapında Ağ)

Günümüzde internet ile sık sık aynı anlamda kullanılmaktadır. Ancak World Wide Web interneti de içine alan daha geniş bir kavramı yansıtmaktadır. World Wide Web İnternet veya ağ ortamında bulunan dokümanların bağlantılar vasıtasıyla birbirine bağlı bir veri deposu oluşturmasıyla ortaya çıkmıştır. Tüm verilerin hiperlinkler ile birbirine bağlı olması sayesinde kullanıcıların bilgiye ulaşmasını sağlar.

Büyük miktarda verilerin birbirine bağlı olması veri madenciliği araştırmaları için de oldukça iyi fırsatlar sunmaktadır. Kullanıcıların bilgiye erişim alışkanlıkları ya da sıklıkları örneğin bir e-ticaret şirketi için daha iyi pazarlama kararları verebilmek için oldukça elverişli bir bilgidir. Ancak kullanıcılar için internet ortamındaki bilgiler anlaşılabilir düzeyde olsa da veriler yapılandırılmamış halde bulunmaktadır. Yapılandırılmamış verileri kullanıcı sorgularına göre yapılandırarak kullanıcının anlayabileceği bilgiye çevirme için oldukça gelişmiş algoritmalar kullanılmaktadır. Google, Yahoo, Bing gibi Arama motorları internet üzerinde veri madenciliğinin en gelişmiş örneklerindendir. Google'ın Pagerank algoritması oldukça gelişmiş mantıksal sorgularla internet üzerindeki yapılandırılmamış verileri yapılandırılmış ya da yarı yapılandırılmış veriler halinde sunucularında saklayarak kullanıcılarına hızlı bir şekilde bilgi akışı sağlar.

World wide web en hızlı büyüyen veritabanlarından birisidir. Mart 2010 itibariyle 20 milyar sayfaya yaklaşmaktadır (Kunder, 2010) . Google veritabanlarında bilgiler üzerinden ise yaklaşık bir trilyon linke ulaşabilmektedir (Google, 2008).

1.3.2 Veri Ambarları

Veritabanları bilgi depolama ve erişimi için kullanılırken, işletmeler daha karmaşık ve çok yönlü sistemlere ihtiyaç duyarlar. Bilgilerin veri madenciliği kullanılarak kullanımı, düzenlenmesi ve anlaşılır bir dile sahip olması veri ambarları sayesinde gerçekleşir. Şirketler için rekabetin gittikçe teknolojik alana doğru kayması, büyük şirketlerin veri ambarları için milyonlarca dolarlık yatırımlar yapmasına yol açmaktadır (Horis, Pedrycz, Swiniarski, & Kurgan, 2007, s. 27).

Veri ambarları stratejik kararların verilebilmesi için verilerin düzenlenmesi, kullanılması ve anlaşılması için yapılar ve araçlar sağlar. İşletmelerin operasyonel veritabanlarından ayrılmaktadırlar. Veri ambarları geçmiş veriler üzerinde analizler yapılabilmesini ve istenilen bilginin daha kolay elde edilebilmesini sağlar.

Veri ambarı kavramını ilk ortaya atan W.H. Inmon'a göre "Veri Ambarı yöneticilerin karar verme süreçlerinde konu-odaklı, entegre edilebilir, zaman-değişkenli ve kalıcı verilerin toplanması işlemidir" (Inmon, 2002). Veri ambarlarının konu odaklı, entegre edilebilir, zaman değişkenli ve kalıcı verileri incelemesi diğer veri depolama uygulamalarından ayıran ana özelliklerdir.

Konu Odaklı Olması: Veri ambarı günlük olarak yapılan işlemler yerine daha önceden belirlenmiş veriler üzerinde çalışır. Veri ambarının konusunun dışında kalan verileri bulundurmaz. Basit ve öz bilgileri içerir.

Entegre edilebilir: Veri ambarları çoklu heterojen kaynaklardan beslenmektedir (Örneğin; ilişkisel veritabanı sistemleri ve metin dosyaları gibi). Veri temizleme ve veri entegre teknikleri veriler arası tutarlılığı sağlama, veri yapılarını kontrol etme, davranış ölçümleri için uygulanmalıdır.

Zaman değişkeni: Veri ambarlarında zaman ilişkisel veritabanları sistemlerinden daha geniş bir aralığı kapsamalıdır. İlişkisel veritabanları anlık verileri dahil ederken veri ambarları geçmişe yönelik verilere önem verir. Veri ambarlarında her davranış bir zaman tanımlayan nesneye sahiptir.

Kalıcılık: Veri ambarları fiziksel olarak ilişkisel veritabanlarını barındıran veri depolama sistemlerinden ayrılmalıdır. Anlık kayıt güncelleme, ekleme, silme gibi işlemler veri ambarında kullanılmamaktadır. Veri ambarlarında iki önemli işlem gerçekleştirilmektedir. Bunlar sistematik veri gösterimi ve veri erişimidir.

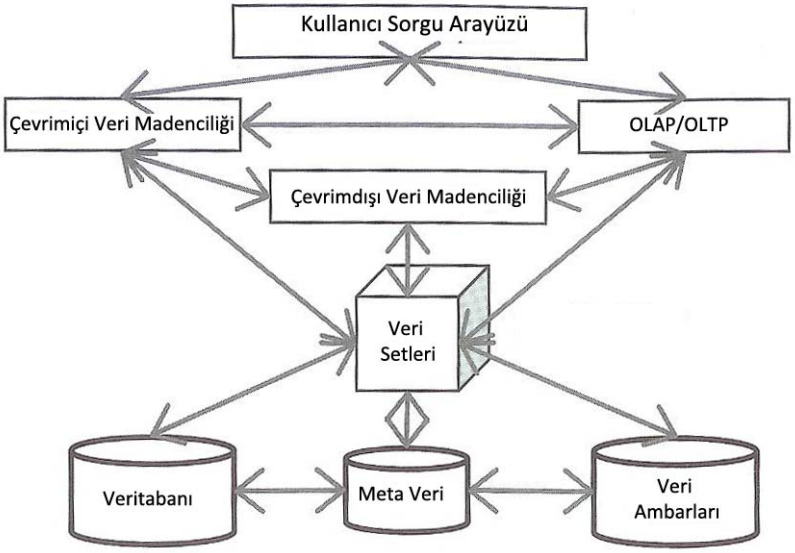
Karmaşık analizler için veri ambarları çok boyutlu bir yapıyla modellendirilmiştir. Örneğin bir satış veri ambarında, satış zamanı, personel ve ürün veri ambarının boyutlarını oluşturabilir. Genelde boyutlar sıra düzenli olarak oluşturulup zaman serileri olarak düzenlenmektedir.

Veri madenciliği veri ambarı için birçok uygulama olanağı sağlar. Veri ambarlarının oluşturulmasının amacı son kullanıcılara bilgi vermek olduğu için veri madenciliği ile birlikte çalışmalıdırlar. Veri madenciliği sayesinde gizlenmiş, önemli bilgilere ulaşılabilmektedir. Veri ambarcılığıyla elde edilen bilgiler birçok şirketin karar vericileri için önem arz etmektedir (Han & Kamber, 2001, s. 39). Şirketlerde en çok karşılaşılan kullanım alanlarından bazıları aşağıdadır;

- Günümüzde müşterilerinin satın alma alışkanlıklarını da içinde bulunduran müşteri odaklanmaları,
- Aylık, yıllık, bölgesel, satış miktarlarına göre ürün konumlandırmaları,
- Maliyet-fayda analizleri,
- Müşteri ilişkileri, şirket için çevresel düzeltmeler, varlık yönetimi

Veritabanı yönetim sistemleri ile veri ambarları arasındaki en önemli fark veritabanı yönetim sistemlerinin anlık işlem ve sorgu yapmasıdır. OLTP (online transaction processing) olarak adlandırılan yönetim sistemleri satın alma, stok tutma, üretim kontrolü, muhasebe işlemleri gibi günlük verilerin anlık olarak kaydedilmesiyle oluşmaktadır. Veri ambarı sistemleri ise farklı kullanıcıların çeşitli ihtiyaçlarını karşılamak için farklı formattaki verileri işleyerek anlamlı sonuçlar çıkarmaktadır. Bu sistemlere de OLAP (Online Analytical Processing) denmektedir.

Şekil 1-1'de OLAP ve OLTP işlemlerinin veri madenciliği sürecindeki yerleri gösterilmiştir.



Şekil 1-1 OLAP ve OLTP işlemlerinin Veri Madenciliği Sürecindeki Yeri (Horis, Pedrycz, Swiniarski, & Kurgan, 2007, s. 128)

OLAP ve OLTP arasındaki farklar şu şekildedir (Inmon, 2002) :

- Kullanıcı ve Sistem Odaklanımı: OLTP müşteri-odaklıdır. Müşterilerle ilgili işlemlerin kayda alınması ve bu kayıtlar içinde sorguların çalıştırılması için kullanılır. OLAP ise Pazar odaklıdır. Veri madenciliği uzmanları ve karar vericiler tarafından veri analizleri için kullanılır.

- Veri İçerikleri: OLTP mevcut anlık veri üzerinden işlem yapar. Karar vericilerin kolay karar vermelerini zorlaştıran çok fazla detaylı bilgiye sahiptir. OLAP ise büyük miktarda geçmiş zamanlı verilerden oluşmaktadır. Özetleme ve kümeleme seçenekleri sunar.

- Veritabanı Tasarımı: OLTP sistemleri veri ilişkisel (entity design) veri modeli ve uygulama odaklı veritabanı (application oriented) tasarımlarına sahiptir. OLAP sistemleri yıldız veya kar tanesi modelleri olarak adlandırılan konu odaklı (subject oriented) veritabanı tasarımlarından oluşurlar.

- Erişim Modelleri: OLTP sistemlerin erişim modelleri kısa ve anlık işlemlerle olmaktadır. Bu sistemler gelişmiş uyumluluk ve kontrol mekanizmasına sahip

olmalıdırlar. OLAP sistemlerinde ise yazma izni genellikle yoktur ve karmaşık sorgulama metotları içerirler

Karmaşık analizleri ve görselleştirmeyi kolaylaştırmak için veri ambarları çok boyutlu olarak tasarlanırlar (Chaudhuri & Dayal, 1997). Örneğin bir satış veri ambarında satış zamanı, satış bölgesi, satıcı ve ürün müşteri tercihlerinin çok boyutlandırılmış şekli olabilir. Oluşturulan boyutlar hiyerarşik bir düzen içinde olmalıdırlar. Bu hiyerarşi zaman alanlarında gün, hafta, ay veya yıl şeklinde olabilirken ürün sınıflandırmasında ürün, kategori veya endüstri şeklinde olabilir. Çok boyutlu veri ambarları veri küpleri olarak şekillendirilir. Bir ver seti; çok boyutlu hiper-karmaşık kavramsal modellemedir ya da kısaca veri küpüdür (Çetinyokuş & Gökçen, 2008). Veri küplerini oluşturmak için yıldız, kar tanesi, takımyıldız şemaları kullanılmaktadır.

- Yıldız Şeması: En çok kullanılan çok boyutlu veri küpü şemasıdır. Verilerin büyük bir bölümünü barındıran merkezi bir tablo(ana tablo) ve her boyut için yardımcı verileri barındıran daha küçük tabloları(boyut tabloları) içerir. Merkezi tablonun etrafına yerleştirilen diğer tablolarla yıldız yağmuru şeklini alır.

- Kar Tanesi Şeması: Yıldız şeması modelinin farklı bir türüdür. Boyut tabloları kirli veriler temizlenerek normalleştirilmiş ve ek tablolara bölünmüştür. Böylece şema grafiği kar tanesini anımsatmaktadır. Yıldız şemasından ayrılan en büyük avantajı kirli verilerin elimine edilmesi nedeniyle boyut olarak daha küçük olmasıdır. Dezavantajı ise tabloların bölünmesi nedeniyle çok fazla ilişki bulunması, sorguları daha karmaşık hale getirmesidir.

- Takım Yıldız Şemaları: Gelişmiş veri ambarı uygulamaları boyut tablolarını kullanabilmek için birden fazla ana tabloya ihtiyaç duyabilirler. Birden fazla yıldız şemasından oluşabilmektedirler. Bir araya gelen yıldız şemaları ise takımyıldızları oluşturmaktadırlar.

(i) Veri Ambarlarının işleyişi

Veri ambarları top down (üst-alt) yaklaşımı, buttom up (aşağı-yukarı) yaklaşımları ya da her iki yaklaşımın üzerine kurulur. "top down" yaklaşımı tasarım ve planlama ile başlar. Eğer kullanılan altyapı ve uygulama iyi biliniyorsa ve sorunlar net ve anlaşılır durumdaysa "top down" yaklaşımı tercih edilir. "buttom up" yaklaşımı prototip oluşturarak deneylerle başlar. Yeni bir teknoloji (uygulama) geliştirme aşamasında oldukça faydalı bir yaklaşımdır. İşletmenin ciddi adımlar atmadan önce daha ihtiyatlı adımlar atarak sonuçları detayları incelemesi açısından faydalıdır. İki yaklaşımın birden uygulanması durumunda ise "top down" yaklaşımının planlanabilir ve stratejik unsurlarından, "buttom search" yaklaşımının hızlı uygulanabilirlik ve fırsatçı uygulamalardan yararlanabilirler (Han & Kamber, 2001, s. 41).

Yaklaşımlardan hangisini tercih edilirse edilsin veri ambarı tasarımı aşamaları aşağıdaki şekilde olmalıdır (Golfarelli & Rizzi, 1998).

- Davranış şemasını oluşturulması
- Davranış şemasının temizlenmesi ve aktarılması
- Boyutların tanımlanması
- Ölçümlerin tanımlanması
- Hiyerarşik yapının tanımlanması

Veri ambarının oluşturulması zor ve uzun bir dönemi kapsadığı için uygulamanın kapsamı iyi belirlenmelidir. Veri ambarlarının amaçları odaklanmış, gerçekleştirilebilir ve ölçümlendirilebilir olmalıdır.

(ii) Veri Ambarının Kullanım alanları

Veri ambarları her endüstriden geniş bir kullanım alanına sahiptir. Veri ambarları sayesinde işletmeler ellerindeki verileri bir arada, işlemden geçmiş ve uyumlu hale getirilmiş bir şekilde toplayabilmektedirler. Bu sayede karar vericiler için veri analizlerini gerçekleştirebilecekleri bir ortam oluşturulabilmektedir. Özellikle bankacılık ve finans sektörü, perakendecilik sektörü ve talep bazlı üretim sektörlerinde şirketler için hayati öneme sahiptir.

İşletmelerin en çok başvurduğu veri ambarı araçları şunlardır:

- Bilgi işleme araçları: sorgulama ve basit istatistiksel analizlerinden oluşur. Çıktı olarak tablolar, grafikler ve şemalar verir.

- Analiz işleme araçları: basit OLAP işlemleri(roll-up, drill-down gibi) üzerine kuruludurlar. En önemli özelliği çok boyutlu verilerin analizlerinde kolaylık sağlarlar.

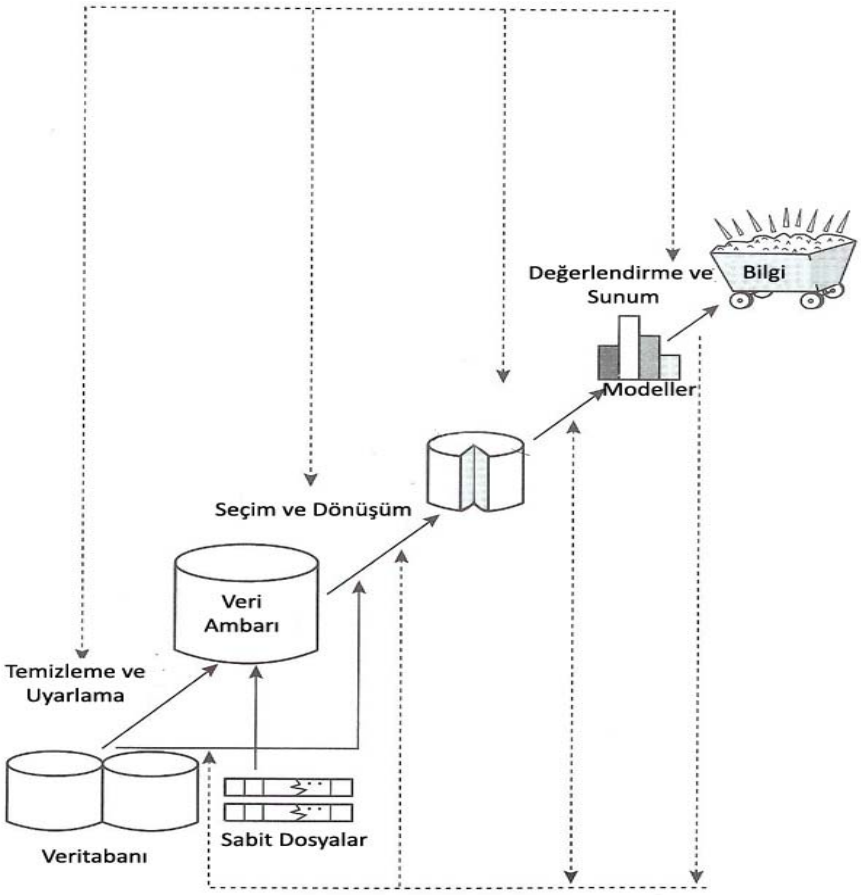
Sorgulara dayalı bilgi işleme araçları gerekli bilgilere erişebilmesine rağmen elde edilen bilgi hiçbir ilişki barındırmamaktadır. Çok yönlü modellemeleri içermemektedir. Analiz işleme araçları kullanıcı tarafından belirlenen veri ambarının segmentlerinden OLAP sayesinde bilgileri çıkarabilmektedir. Analiz işleme araçları analizlerde veri özeti sağlarken, veri madenciliği büyük veri yığınlarında ve gizli alanlarda otomatik bilgi keşfi sağlar. OLAP işlemleri kolaylaştırmak için tasarlanmasına rağmen veri madenciliği araçları olabildiğince fazla işlem üzerine yoğunlaşır. Bu yüzden geleneksel bilgi işlem araçları ve analiz işleme araçlarından(OLAP'tan) bir adım öndedir.

1.4 Verilerin Hazırlama Süreci

Bahsedildiği üzere veritabanları eksik, yanlış veya düzensiz bilgileri barındırmaya müsait bir yapıya sahiptir. Üçüncü parti yazılımlar son derece gelişmiş olmasına rağmen, günümüzde veritabanlarının boyutları terabaytlarla ifade edilmekte, verilerin kontrolü gittikçe zorlaşmaktadır. Veri Hazırlama süreci kaynak verilerin belirli madenleme uygulamaları için hazırlanma sürecidir. Veri madenciliği algoritmalarının uygulamadan önce bir takım veri hazırlama sürecinden geçirilmesi şarttır (Koutri, 2004). Bu işlemler şu şekilde sıralanabilir :

- Veri Temizleme
- Veri Uyarlama
- Veri Dönüşümü
- Veri Daraltma

Şekil 1.2'de sistematik olarak veri madenciliği süreci görülebilmektedir.



Şekil 1-2 Bilgi Keşfi Sürecinde Veri Hazırlık Aşamaları (Han & Kamber, 2001, s. 6)

1.4.1 Veri Temizleme

Veri temizleme süreci veritabanlarında eksik ve anlamsız değerleri sistemli olarak veritabanından çıkarma aşamasıdır. Bu süreç planlı bir süreç olmalı, kirli verilerin tespitinin sağlıklı yapılması gerekmektedir. Veri hazırlama süreçleri için ne yazık ki işlemi otomatik sürdürebilecek akıllı yazılımlar bulunmamaktadır. Bu yüzden kullanılabilecek veri madenciliği araçlarının iyi analiz edilerek dikkatli bir şekilde veritabanlarına uygulanmalıdır (Pyle, 1999).

Eksik Değerlerin tespiti: Veritabanında bulunan alanlardan eksikliklerin tespiti durumunda aşağıdaki adımlar izlenebilir;

- Satır iptal edilebilir: Bu yöntem genelde sınıflandırma alanı eksik olduğunda uygulanabilir. Analiz için satırda bulunan verilerden önemli bir kısmının eksikliği durumunda başvurulabilir.

- Eksik veriyi elle girmek: Genelde bu yöntem zaman alıcı ve büyük veri yığınları içerinde uygun olmayan bir yoldur.

- Eksik değerler için genel bir sabit değerin atanması: Tüm eksik değerler için "Bilinmiyor", "Eksik Değer" ya da " ∞" gibi tanımlayıcılar girilebilir. Böylece veri madenciliği programı eksik değerleri tanımlayabilir.

- Eksik değerler için ortalama değerin atanması: İlgili alanın boş olması durumunda diğer alanların ortalama değeri boş alanlara yazılabilir. Ancak bu yöntem sadece sayısal alanlarda geçerlidir.

- Eksik değerler için diğer ilgili alanların ortalama değerinin atanması: Bir kayıttaki eksik alan için benzer başka bir alanın değerinin atanması yoludur.

- Eksik alan için en uygun değerlerin hesaplanarak atanması: regresyon, karar ağaçları ya da diğer Bayesyen kuralcılığı (formalism) kullanan sonuç tabanlı araçlar kullanılarak yapılabilmektedir. Örneğin bir tabloda müşteri davranışlar göz önüne alınarak müşterilerin gelirleri hesaplanabilmektedir.

Eksik değerler için genel bir sabit değerin atanması ve eksik alan için en uygun değerlerin hesaplanarak atanması çok doğru sonuçlar vermeyebilir. Ancak Eksik alan için en uygun değerlerin hesaplanarak atanması yöntemi ve verilerin çıkartılması en çok kullanılan yöntemler arasındadır (Han & Kamber, 2001, s. 106).

Kirli verilerin temizlenmesi: Gürültülü veriler olarak da adlandırılan kirli veriler, bütününe bakıldığı zaman ilgisiz ve anlamsız kalan verilere denmektedir. En iyi veri temizleme metotlarında yeterince başarılı olamayan veri toplama işlemleri sonucunda oluşan kirli verilerin bulunması ve temizlenmesi ön plandadır. İlgisiz ve ilgisi oldukça düşük olan veriler veri analizlerine engel olduğundan kirli veri olarak ele alınmalıdır. (Xiong, Pandey, Steinbach, & Kumar, 2006)

Gürültülü/Kirli verileri temizleme yöntemlerinden bazıları aşağıdadır:

- Binning (Sınıflandırma): Küçük gözlem hatalarını ortadan kaldırmak için kullanılan bir yöntemdir (Wikipedia, 2010). Bölümleme yöntemi ile veriler arasındaki ilişkiden

yararlanarak birbirilerine yakınlık derecelerine göre gruplara ayrılmaktadırlar. Bu gruplama sonucunda belirlenen metotlarla veriler sabit bir değere dönüştürülmektedir. Örneğin, grubun ortalaması alınarak grup içerisindeki verilerin değeri olarak atanabilmekte ya da medyan değeri hesaplanarak verilerin değeri medyan olarak saptanabilmektedir.

- Kümeleme (Clustering): Benzer verilerin aynı grup içerisinde toplanarak kümelenmesidir. Kümeleme metodu ile kümeleme dışı kalan bazı uygunsuz verilerin analiz dışında tutulması sağlanarak basitleştirme yapılır. Bu metotla veri seti içerisindeki az ama öz verilerin analizlerde dikkate alınması sağlanmaya çalışılmıştır. Bu yüzden çok fazla kurala ve uygulamaya ihtiyaç duymaktadır (Berkhin, 2006).

- Bilgisayar programı ya da bireysel gözlem

- Regresyon: En sık kullanılan veri temizleme tekniklerinden biridir. Kümeleme de olduğu gibi veri seti içerisinde diğer verilerle uygunsuzluğu tespit edilen verileri analiz dışı bırakır. Regresyon analizinde kümelemeden farklı olarak bu işlemi farklı olasılık dağılımları ile gerçekleştirilir. Bu işlem sadece hatalı verileri değil, doğru girilmiş verileri bile analiz için uygun bulmayabilir (Zhu, Wu, & Chen, 2006).

1.4.2 Veri Uyarlaması

Veri Uyarlaması (Entegrasyonu) farklı kaynaklardan elde edilen verilerin elde edilmesi ve birleştirilmesi problemidir (Halevy, 2001) (Hull, 1997). Genelde çoklu bilgi sistemlerinin amacı belirlenen sistemlerden bilgileri birleştirerek tek bir bilgi sistem ağında toplamaktır. Verilerin uyarlanması gerekliliği iki sebebe dayanır: Var olan bilgi sistemlerini bilgi erişimi oluşumuna ve tekrar kullanıma uygun hale getirmek. İkincisi ise daha kapsamlı tatmin sağlanabilmesi için farklı tamamlayıcı bilgi sistemlerinden ihtiyacı gidermek içindir (Ziegler & Dittrich, 2004).

Uyarlama sırasında karşılaşılan problemlerden en önemlileri "şema uyarlaması" ve "artık bilgilerdir". Kayıt tanım problemleri de denilen şema uyarlaması esnasında çıkan hatalar birden fazla veri kaynağından gelen verilerin birbiriyle tutarsız olmasından kaynaklanmaktadır. Artik bilgiler ise kaynaktan alınan verilerin başka tablolardan elde edilen

verilerle tutarsız olmasından kaynaklanır. Örneğin toplam satış tutar verisinin satışlar tablosundan elde edilen verilerin toplam değerinden farklı olması gibi. Bu hatalar genelde korelâsyon analizleri ile bertaraf edilmektedir. Belirlenen iki değer arasında ne kadar güçlü bir ilişki olduğu kestirilebilmektedir.

1.4.3 Veri Dönüşümü

Veri dönüşüm süreci anlık-ilgili (instance-related) dönüştürme (haritalama) ve şemalama işlemlerini barındıran birçok adımdan oluşmaktadır. Veri dönüşümü sürecinde dönüşüm kodları oluşturmak ve otomatik bir şekilde gerçekleştirmek için dönüşüm işlemi belirli ve uygun yöntemlerle gerçekleştirilmelidir. Bu yöntemlerden en genel ve esnek olanı veritabanı sorgulama dillerinden olan SQL'dir. SQL ile veri dönüşümünün yanı sıra kullanıcı-tanımlı fonksiyonların oluşturulmasına ve kullanıcı uygulamaları için özelleştirilmesine olanak sağlamaktadır (Rahm & Hai, 2000). Kullanıcı tanımlı fonksiyonlar veri dönüşümünü daha verimli, hızlı bir şekilde gerçekleşmesinde en büyük yardımcılardandır. Maliyet ve zaman konusunda tasarruf sağlar.

Veri dönüşümünde kullanılan bir diğer yöntem ise veri haritalama yöntemidir. Veri haritalama veri modelleri arasında eşleştirme yapılmasıdır. Örneğin bir şirketin satış işlemleri için oluşturduğu faturaları standardize etmek için diğer şirketlerle oluşturulan ortak standartlarla karşılaştırarak fatura verilerini belirli bir sistem içinde tutmasıdır. Kullanım alanı kısıtlı olduğu için kullanıcı tanımlı fonksiyonlar kadar popüler değildir.

1.4.4 Veri Daraltma

Veri daraltma veri seti içerisinden tüm veri setini yansıtacak şekilde daha küçük bir veri seti haline dönüştürme işlemidir. Büyük veritabanlarından hızlı ve uygun cevaplar bulmak her zaman mümkün olmamaktadır. Veri madenciliği işlemlerini tüm veri üzerinde uygulamaktansa daha küçük ve veri bütünlüğünü yansıtabilecek bir veri setinde çalışmak daha etkili olacaktır.

Veri daraltma teknikleri iki ana sınıfa ayrılmaktadır. Bunlardan birincisi veri seti için bir model belirlenip bu modelin parametrelerinin hesaplandığı parametrik tekniklerdir. Lineer regresyon modelleri, log-lineer modelleri parametrik tekniklere örnektir. Diğer bir veri

daraltma tekniği ise modelleme yapmadan veri daraltma işlemini gerçekleştiren parametrik olmayan tekniklerdir. Histogram indeks ağacı en yaygın kullanılan parametrik olmayan veri daraltma tekniklerindendir.

1.5 Dijital Kütüphanelerde Veri Madenciliği ve Uygulama Örnekleri

Okullarda, üniversitelerde, derneklerde ve kurum kütüphanelerinde teknolojinin değişmesiyle birlikte hizmet anlayışı da değişmektedir. Kullanıcı çeşitliliğinin artması, yayın sirkülâsyon bilgilerinin giderek çoğalması, arşiv kaynaklarının depolanamaması gibi sorunlar günümüz kütüphanelerinin karşılaştığı büyük problemlerden birkaçıdır. Ne yazık ki sayıca oldukça az kütüphane, bu büyük veri yığınını kullanarak kullanıcı hizmetlerini artırımı, bütçe tasarrufu, stratejik karar verme gibi konularda kendilerine avantaj sağlayabilmektedirler. Kütüphane verileri üzerinde yapılan veri madenciliği çalışmaları ile dijital veriler bir araya getirildiğinde ne gibi sonuçlar elde edilebileceği üzerinde durmaktadır.

1.5.1 Dijital Kütüphane Kavramı

Dijital kütüphane kavramı, geleneksel kütüphane anlayışı ile günümüz teknolojik altyapısının birleşmesi sonucu doğmuş bir terimdir. Bu birleşim bilgi kaynaklarına erişim yöntemlerinin daha hızlı ve sağlıklı bir şekilde işlemesine yardımcı olmakta önemli bir adımdır. Üzerinde durulması gereken önemli nokta dijital kütüphane kavramının kullanıcı ihtiyaçlarını gerçekten karşılayabilecek ve kütüphane yöneticilerinin bu ihtiyaçları tatmin edebilir düzeyde bir altyapıyı kullanabilecek birikime sahip olup olmadıklarıdır.

İnternetin günümüzde vazgeçilmez bir araç olması, küçük kütüphanelerin bile yerel bir kütüphaneden küresel bir kütüphanenin parçası haline getirmiştir. Yayın sahiplerine ve yayıncılara ulaşmak isteyenlere dijitalleştirme sürecinde önemli avantajlar sağlaması, uluslararası telif hakları anlaşmalarının yaygınlaşması, süreli yayınların kısa sürede okurlarına erişebilmesi, veritabanlarının kontrollü bir şekilde üyelerine kaynaklarını anında sunabilmeleri bu sürecin önemli noktalarını oluşturmaktadır. Bu özelliklerin yaygın bir şekilde gerçekleştirilmesinde depolama maliyetlerinin hızla düşmesi en büyük engellerden birinin kalkmasına yardımcı olmuştur

2003 yılında Scott Nilson'un kütüphane verileri üzerinde yaptığı veri madenciliği çalışmalarını "bibliomining" olarak adlandırarak literatüre yeni bir terim kazandırmıştır. Kütüphane kültürüne ait olan yazarlık, alıntılama gibi bibliometrik verilerin kütüphane veri madenciliğinin temelinde yer alması bu terimin ortaya çıkmasında etkili olmuştur.

Dijitalleşme süreci içerisinde, kütüphane ve bilgi merkezleri durmaksızın veri biriktirerek bu verileri işleme, yönetme ve arşivleme yeteneklerini geliştirmektedirler. Katlanarak büyüyen bilgi yığınları zenginliği, veri madenciliğini bilgi erişim yollarının keşfi nedeniyle zorunluluk hale getirmiştir. Ancak kütüphane sistemlerinin dijitalleşmesi henüz yeni süreç iken kaynakların çok büyük olması ve bunların bilgi sistemlerine aktarılamaması ve yeterli seviyelere ulaşamaması kütüphanede veri madenciliği uygulamaları çok yaygın olmamasına neden olmuştur. Ancak Kevin Cullen'in belirttiği gibi kütüphanelerde veri madenciliği veri ambarları üzerindeki sirkülâsyon, satın alma ya da kataloglama gibi işlemsel verilerin optimize edilmesiyle gerçekleştirilmelidir. Böylece kütüphane içerisindeki modellemeler daha kolay ve başarılı bir şekilde ortaya konabilmektedir (Cullen, 2005).

1.5.2 Dijital Kütüphane Unsurları

Dijital kütüphaneler klasik anlayıştan sıyrılarak yeni bir konsept oluşturduğu için dünya genelinde belirli standartlara oturtulmaya çalışılmaktadır. Belirli sınıflandırmalara ayrılan kütüphaneler, sistematik bir bilgi sistemi odaklı ve içerden ya da dışarıdan erişim sağlamaya yönelik olmalıdır. Bu yüzden merkezi kavramlar, varsayımlar, parametreler ve dijital kütüphanelerin değişim süreci için unsurlar belirlenmiştir (Fuhr, Tsakonas, Trond, Agosti, & Hansen, 2007):

- Kullanıcı ve İçerik Etkileşimi: Dijital kütüphaneler birçok amaç için kullanılması mümkündür ancak asıl amaç bilgi erişimidir. Belirli içerikleri bulmak, özelleştirilmiş bilgilere ulaşabilmek, kütüphane kullanıcısının ihtiyaç duyduğu bilgi beklentilerini karşılamak, kısaca kullanıcının dijital kütüphane kullanımını gerçekleştirirken aklından geçenleri öngörerek, isteğine en kısa yoldan ulaşmasını sağlamak kütüphane karar vericileri için en önemli amaçtır.

- Mevcut kütüphane planlarının değişimi: Kütüphane erişimini sağlayan sistem bileşenleri çoğunlukla bilgi erişim teknolojilerine dayanmaktadır. Bu sistem sıradan bir altyapıyla değil, dijital kütüphanelerin mantığına ve planlamasına uygun olarak denenmiş ve test edilmiş yöntemsel bilimler geliştirilmelidir. Bu konuda dikkat edilmesi gereken unsur bilgi sisteminin merkezi bileşenlerinin karakteristik özelliklerinin kütüphanelerin kendilerine özgü bilgi ağını koruması ve bu ağ içerisinde kusursuz bir şekilde etkileşimi sağlamasıdır. Kullanıcı gereksinimleri, görevleri, özelleştirilebilen durumları ihtiyaçlara göre yönetebilmelidir.

- Kullanıcı-Kütüphane ilişkisi: Dijital kütüphane kavramı kullanıcı, kullanım, bilgi ihtiyacı ve bilgi arasındaki ilişkinin merkezi bir noktada oluşturulması üzerine kurulmalıdır. Bu kesişme ulaşılan metnin kullanıcı için okumaya değer bir bilginin ortaya çıkması ile gerçekleşmektedir. Daha derinde görev fonksiyonları, içerik özellikleri, kullanıcı tercihleri ve kişileştirilmesi bağları oluşturmalıdır. Bu karmaşık yapılanma kullanıcılar için basit bir ekrandan oluşmalıdır. Kullanıcı ihtiyacı olmayan bilgilere boğulmamalıdır.

- Yeni talepleri karşılayabilme: Dijital kütüphaneler içerik odaklı bir sistem olduğu için içeriğin genişletilmesi kadar sistem performansının korunması gerekliliğinin de unutulmaması gerekmektedir. Sadece klasik kütüphane kullanıcı davranışı göz önüne alınarak geliştirilen bir sistem ihtiyaçları karşılamayacaktır. İlk ve öncelikli olarak kütüphane içi yapı şeması dijital kütüphaneler için sıklıkla değerlendirilmeli ve yenilenmelidir. Diğer taraftan kullanıcı davranışları da bu değerlendirmenin içerisinde olmak zorundadır. Kullanıcı herhangi bir sebeple bir bilgi istemi gönderdiği zaman, işlemden başarılı bir cevap alana kadar tekrarlayabilir. Bir arama döngüsünü, sonuca ve bilgiye ulaşana kadar tekrar tekrar formüle edebilir. Bilgiye ulaşmasının ardından döngüyü benzer içerikler için değişik şekillerde sistemde sorgulatabilmektedir. Kullanıcının başarısız ya da birlikte kullandığı sorguları kayıt altına alarak, sorgulardan çıkarımlar yapılmalıdır. Bu çıkarımlarla oluşturulan sistem ağı geliştirilmelidir.

- Yeni tekniklerin sisteme uyarlanması: Bilgi erişiminde etkilişimsel bakış açıları, bilgi depolama araçları ve içerik odaklı sistem araçlarının internetle birlikte gelişmesiyle ortaya çıkmaya başlamıştır. Sistem içerisinde etkileşim, genellikle

kullanıcı ile birlikte gerçeleşmektedir. Bu yüzden bilgi edinme bileşenlerine bakış açısı genişleterek sistemin kendi içerisinde etkileşimin arttırılması, görev odaklı, daha kişileştirilmiş ve daha kullanışlı olması sağlanmalıdır. Bu gelişmeleri sağlayabilmek için kullanıcı tatmininin kıyaslanmamış ve tespit edilmemiş kavramları birleştirilerek yeni ölçütler oluşturulmalıdır.

1.5.3 Kütüphanelerde Veri Madenciliği Aşamaları

Kütüphanelerde veri madenciliği (bibliomining) sürecini aşağıdaki adımları takip edilerek gerçekleştirilebilir (Nicholson, 2003) :

- Odaklanacak Alanların belirlenmesi: Dijital kütüphanelerde veri madenciliğinin ilk adımı madenciliğin gerçekleştirileceği amaç ve amaca bağlı olarak alanların tespit edilmesidir. Bu odaklanma belli bir problem üzerine olabileceği gibi karar verme ve bilgi keşif için genel bir tarama da olabilir. Bu nedenle veri madenciliği işlemine karar vermeden önce veri madenciliğinin önceden yönlendirilmiş ya da yönlendirilmemiş olması belirtilmelidir. Eğer problem odaklı bir tarama söz konusu ise önceden yol haritası çizilmelidir. Örneğin bütçe sıkıntısı çeken bir kütüphane karar vericisi için hangi alanlarda kısıntıya gideceği konularda veri madenciliği tekniklerini kullanarak yolunu çizebilir. Veri madenciliğinde belirli bir soruna yönlendirilmeden de genel bir veri madenciliği taraması gerçekleştirerek kütüphanenin genel durumu hakkında bilgi sahibi olunabilmektedir. Ancak böyle bir yol izlenmesi durumunda çeşitli zorluklarla karşılaşılabilmektedir. Tüm kütüphane verilerinin veri madenciliği işleminden geçirilmesi, veri madenciliği süreci öncesinde gerçekleştirilmesi gereken veri temizleme ve uyarlama işlemleri oldukça vakit almasına neden olmaktadır. Özellikle olasılık üzerine kurulu veri madenciliği modellemelerinin kullanılması güçleşecek ve güçlü bilgisayarlara gereksinim duyulacaktır.

- Veri Kaynaklarının tespit edilmesi: Problem ya da veri madenciliğinin amacı tespit edildikten sonraki adım uygun veri kaynaklarının belirlenmesidir. Kütüphanelerde veri madenciliği süreci işlemsel, birleştirilmemiş ve düşük-seviye veriler üzerinden gerçekleştirilmektedir. Bu durum büyük veri yığınının depolanması ve depolama

maliyetlerinin yüksek bir düzeyde olması, kullanıcı gizliliğinin korunması gibi nedenlerle kütüphane karar vericilerinin bu verileri saklamakta isteksiz olmalarına sebep olmaktadır. Oysaki kütüphane kaynakları kadar işlemsel veriler de veri madenciliği için önemli bir araçtır. Dijital kütüphanelerde veri depolama sistemleri mutlak bir gereksinimdir.

Dijital kütüphane veri kaynaklarında iki çeşit veri üzerinde durulmalıdır. Bunlardan ilki kütüphane sisteminde mevcut olan iç verilerdir. Bu veriler yayın bilgileri, yazar bilgileri, süreli yayınlar ya da tezler, işlemsel veriler, web arama kayıtlarından oluşabilir. Bu veriler oldukça karmaşık ve zorlayıcı bir süreç sonucunda veri madenciliğine kullanabilir şekle getirilebilmektedir. Diğer veri türü ise dışsal verilerdir. Bu veriler demografik bilgiler, yerleşim yeri gibi daha genel bilgileri dış bir kaynaktan elde edilebilmektedir

- Veri ambarının oluşturulması: Belirtildiği üzere veri madenciliği için gerekli olan birçok verinin birden fazla kaynaktan sağlanması veri ambarlarının oluşturulmasını zorunlu kılmaktadır. Veri ambarı ana kaynaktan farklı olarak diğer kaynaklarla ortak bir ilişkisel alanları dikkate alır. Veri ambarı içerisindeki veriler temizlenmiş ve dönüştürülmüş operasyonel verilerden oluşmaktadır. Veri ambarı oluşturmak için kütüphane karar vericilerinin gözleminde verilerin seçilmesi gerekmektedir. Temizlenme, düzenleme ve yeniden oluşturmadan sonra prosedür otomatik bir şekilde devam ettirilmelidir.

- Veri ambarının yapılandırılması: Veri ambarının oluşturulması ve yaplandırılıması veri madenciliği süreci içerisinde en fazla zaman ayrılması gereken bölümlerdir. Bu sürecin defalarca tekrarlanması gerekebilir. Ancak bir kez oluşturulduktan sonra, ilerideki veri madenciliği çalışmaları için bir model oluşturulabilir ve sonraki çalışmalarda zamandan tasarruf edilmesini sağlayabilmektedir. Oluşturulan algoritmalar tekrar tekrar kullanılabilmektedir.

- Uygun veri madenciliği araçlarının belirlenmesi: Ver ambarı oluşturulduktan sonra analiz aşamasına geçilebilmektedir. Veri ambarı üzerinden geleneksel istatistiksel raporlar oluşturulabileceği gibi, bu istatistiksel verilerin ardında gizlenen ilginç ve faydalı örüntüler de tespit edilebilmektedir. Bu örüntüler kütüphane karar vericileri için anlamlı veriler elde etmesinde önemli bir kaynaktır. Örneğin kullanıcıların

kütüphane kullanım yoğunlukları tespit edilerek kütüphane çalışanları için verimli çalışma saatleri belirlenebilir ya da kitap sirkülâsyonu göz önüne alınarak kitap yerleşim düzeni tekrar oluşturulabilir.

- Analiz ve Uygulama: Analizler gerçekleştirilerek raporlar ve karar verme modelleri ortaya konulduktan sonra onaylanmalıdır. Veri madenciliği sonuçlarının onaylanması kütüphane karar vericiler tarafından gerçekleştirilmelidir. Sonuçların ortaya konulması ve uygulanabilirliğinin test edilmesi modellerin geliştirilmesi için önemli bir veridir. Oluşturulan örüntüler karar verici için bir örneklem içerisinde gerçekçi bulunmadığı takdirde, uygunsuz örüntülerin neden oluştuğu derinlemesine incelenmeli ve modellemeler tekrar tekrar test edilmelidir. Son adım olarak model ve örüntüler tüm kütüphane sisteminde uygulanmasıdır. Değişkenlerin bir süre daha izlenmesi, modellerin güçlendirmesi sistemin kusursuz işlemesi için yararlı bir hareket olacaktır. Model değişkenliğini korumalıdır.

1.5.4 Dijital Kütüphane Sürecinde Gizlilik

Kütüphanelerde bu süreç yaşanırken diğer dijitalleşme süreçlerinde ortaya çıktığı gibi gizliliğin nasıl korunacağı önemli problemlerden biridir. Özellikle sunulan servislerin kişileştirilmesi ve kullanıcı dostu bir ara yüz oluşturulmaya çalışırken kullanıcıların kişisel bilgilerini üçüncü kişilerle paylaşmakta tereddüt etmesi gizlilik konusunda da ciddi önlemlerin alınması, bu önlemlerin kullanıcılara iyi bir şekilde aktarılması gerekliliği önemli bir husus haline gelmiştir. Dijital kütüphanelerde kullanıcı bilgileri sadece formalite icabı alınan kayıtlar değil, kütüphane işleyişinin nasıl işlediğine dair bilgilerin ortaya çıkartılmasında gerekli bilgilerdir.

1.5.5 Kütüphane Veri Madenciliği Uygulamaları

Kütüphane verileri üzerinde yapılan çalışmaları incelediğimizde, kütüphane sistemlerinin günümüz teknolojisine ayak uydurmaya başlamasıyla ortaya çıkan dijitalleşme sürecini henüz tamamlayamadığı gözlemlenmiştir. Bu yüzden yapılan çalışmalar oldukça kısıtlıdır. Ulusal düzeyde herhangi bir çalışmaya rastlanmamış, yazın taramasında da uluslararası çalışmalara yer verilmiştir.

Tablo 1-1 Dijital Kütüphalerde Veri Madenciliği Literatür Taraması

Çalışmanın yazarı	Yılı	Kullanılan algoritmalar	Çalışmanın Amacı
Robert Sanderson Paul Watry	2007	Metin Madenciliği	Dijital kütüphane tekniklerini Veri madenciliği ve metin madenciliği tekniklerini uygulanması ve hesaplanabilir veri çevrimiçi sınıflandırma örneği oluşturulması için çalışılmıştır.
C Battioui, KY Louisville	2007	Birliktelik Kuralları	Kütüphane dijital kaynakları içinde bulunan URL bilgileri üzerinde çalışılmıştır. Ekstrom Kütüphanesinde URL ve diğer değişken gruplar arasındaki ilişkisel düzenler araştırılmıştır.
Chi Chunjia, Mao Zhiyong	2009	Regresyon Analizleri	Üniversite kütüphanesi kitap satın almada etkinlik ve ödünç kitap dolaşımı planlaması üzerine analizler yapılıştır.
Scott Nicholson	2004	Karar Ağaçları, Sinirsel Ağlar	Dijital kütüphane ayarları üzerinde otomatik bilgi toplama sistemi oluşturulmuştur. Akademik çalışmaların yapıldığı sayfalarda öngörülebilir modellemeler oluşturularak literatür seçimi konusunda tahminler ortaya konulmuştur.
San-Yih Hwang and Ee -Peng Lim	2002	Regresyon Sinirsel Ağlar	Kütüphane içerisinde daha önce hiç ödünç alınmamış ya da puanlanmamış kitaplar üzerine yapılan veri madenciliği çalışmasıdır. Kullanıcıların Demografik özellikleri göz önüne alınarak kullanıcıların kitap ödünç alma alışkanlıkları izlenerek öneri sistemi ortaya konmuştur. Yeni bir algoritma oluşturarak sirkülâsyon veritabanı yeniden yapılandırılmıştır.
Tian Hong	2009	Birliktelik Kuralları Kümeleme	CNKI ulusal kütüphane veritabanının 1997-2005 yılları arasındaki verileri üzerinde kümeleme metoduyla
LI Mo	2008	Web madenciliği	Dijital Kütüphanelere Web üzerinden erişim metotlarının geliştirilmesi ve web madenciliği ile arama kayıtları üzerine bir çalışma
Daniel C Weaver	2004	ANN K-en yakın	Dijital Kütüphanelere Web üzerinden erişim metotlarının geliştirilmesi ve web madenciliği ile arama kayıtları üzerine bir çalışma
Aristeidis Meletiou Anthi Katsirikou	2008		Kütüphanelerden gelen günlük verilerin işlenerek hangi amaçlar için kullanılabileceği, hangi aşamalardan geçmesi gerektiği konusu işlenmiştir.

2. BÖLÜM : VERI TABANINDA KULLANILAN SINIFLANDIRMALAR VE ALGORITMALAR

2.1 İstatistiksel Sınıflandırmalar

2.1.1 Bayesyen Sınıflandırma

Bayesyen Sınıflandırma, Bayes karar teorisi ile yapılan istatistiksel işlemler ile model tanımlamada ve sınıflandırmada uygulanan temel tekniklerdendir. Modellerin sınıflandırılması olasılıklar şeklinde belirlenir. Bayesyen sınıflandırma kurallarını Bayesyen Sınıflandırmalar içerisinde en çok kullanılan "Naive Bayesyen sınıflandırma" (Basit Bayesyen sınıflandırma) karar ağaçları ya da Neural sinir ağları gibi diğer sınıflandırma araçlarına göre büyük veritabanlarında yüksek tutarlılık ve hız sağlamaktadır (Han & Kamber, 2001, s. 296).

Naive Bayesyen sınıflandırıcılar son yıllarda sıkça kullanılmakta ve başarılı sonuçlar vermektedir. Bu olasılıklı yaklaşımlar verinin nasıl oluşturulduğu hakkında kuvvetli varsayımlar oluşturmakta ve bu varsayımları içeren olasılıksal modeller ortaya koymaktadır. Sonraki aşamalarda üretken model parametrelerini hesaplamak için belirlenmiş örnek derlemeleri kullanırlar. Gerçek dünyada bu varsayımlar genelde yanlış olsa da, Naive Bayesyen sınıflandırıcı, sınıflandırmalar konusunda oldukça başarılı bir iş çıkarmaktadırlar (McCallum & Nigam, 2003).

Bayesyen sınıflandırmanın amacı nesnelerle ilgili istatistikî bilgiye göre hatalı sınıflandırmayı minimize etmektir. Yeni nesnelerle ilişkili sınıflandırma kapasitesi daha önce popülâsyon içerisinden rastgele gözlemlenen nesnelere bağlıdır. Yeni nesnelerle ilgili tahmin tutarlılık seviyesi istatistiksel ölçümlerin miktarına ve önceki deneylerden elde edilen bilgilere bağlıdır (Horis, Pedrycz, Swiniarski, & Kurgan, 2007, s. 476).

Naive Bayesyen sınıflandırıcıların çalışma şekli şu şekildedir.

1. Her veri örneklemi n-boyutlu öznitelik vektörünce, X = (x_1, x_2, ..., x_n) temsil edilmektedir. A_1, A_2,. . . , A_n *n* adet öznitelikten oluşan bir değişkenler grubunda yapılan *n* adet ölçümü belirtmektedir.

2. C_1, C_2, ... , C_m dan oluşan *m* tane sınıf olduğunu varsayalım. Bir bilinmeyen veri örneklemi olan X'de, sınıflandırıcı X'in X'e koşullu olmak üzere en yüksek ardıl olasılığa sahip olduğunu öngörür. Bu durum, Naive bayesyen sınıflandırıcının bilinmeyen örneklem X'i sadece aşağıda belirtilen durumda sınıf C_j 'ye atamaktadır

   $P(C_m \mid X) > P(C_j \mid X) \qquad 1 \le j \le m,\ j \ne a$

   Böylece $P(C_m \mid X)$ maksimize dilmiş olur. $P(C_m \mid X)$'nin maksimize edildiği sınıf C_m maksimum ardıl hipotez olarak adlandırılır. Bayes teoremine göre ise aşağıdaki şekilde formülize edilir;

   $$P(C_m \mid X) = \frac{P(X \mid C_m)\, P(C_m)}{P(X)}$$

3. Tüm sınıflar için P(X) sabit olduğu için, $P(X \mid C_m)\, P(C_m)$ maksimize edilmesi gerekmekedir. Eğer önceki sınıf olasılıkları i bilinmiyorsa, yaygın olarak kabul gören şekilde bütün sınıflar eşit olasılığa sahiptir. Bu durumda $P(C_1) = P(C_2) = \ldots = P(C_m)$ olur ve $P(X \mid C_m)$ maksimize edilmiş olur. Aksi durumda $P(X \mid C_m)\, P(C_m)$ maksimize edilmiş olacaktır. Sınıf öncül olasılılıkları s_i C_m değişkenler grubu sayısını, s'nin de toplam değişken grubu sayısını gösterdiği durumda, $P(C_m) = \frac{s_i}{s}$ şeklinde de hesaplanabilir.

4. Çok fazla özniteliği n bulunduğu veri setlerinde, $P(X \mid C_m)$ hesaplamak oldukça büyük bir kaynak kullanımını gerektirmektedir. Bu kullanımı azaltabilmek için Naive sınıf koşullu bağımsızlık değişkeni kullanılır. Böylece öznitelik değerleri birbirinden koşullu olarak bağımsız sayılmaktadır. Böylece öznitelikler arasında herhangi bir ilişki bağlılığı yok sayılır.

$$P(X \mid C_m) = \prod_{k=1}^{n} P(X_k \mid C_m)$$

$P(X_1 \mid C_m)$, $P(X_2 \mid C_m)$, . . . , $P(X_n \mid C_m)$ olasılıkları değişken grubundan hesaplanabilmektedir. Bunlar:

a) Eğer A_k koşulsuz ise, Sia 'nın C_m sınıfının değişken grubu sayısı olduğu, Si'in Ci 'ye ait değişken grup sayısı olduğu durumda $P(Xk|Cm) = \frac{Sia}{Sa}$'dir.

b) Eğer A_k sürekli bir değere sahipse, özniteliğin Gaussian dağılımı gösterdiği varsayılır.

$$P(X|\text{Cm}) = g(xk, \mu cl, \sigma cl) = \frac{1}{\sqrt{2\pi}\,\sigma cl} e^{-\frac{(xk-\mu cl)^2}{2\sigma cl^2}}$$

Gaussian yoğunluk fonksiyonu olan $g(xk, \mu cl, \sigma cl)$ ile Öznitelik A_k için ortalama ve standart sapmanın bulunmasını sağlamaktadır.

5. Bilinmeyen örneklem X'i sınıflandırmak için her sınıf Ci için $P(X|C_m)P(C_m)$ hesaplanır. Örneklem X Sınıf Ci'ye aşağıdaki şartlar içerisinde atanabilir.

$$P(X|C_m)P(C_m) > P(X|C_j)P(C_j) \quad 1 \le j \le m,\ j \ne a$$

Başka bir deyişle, $P(X|C_m)P(C_m)$ maksimum değer aldığında sınıf Ci atanabilmektedir.

2.2 Karar Ağacı Sınıflandırmaları

Karar ağaçları sınıflandırılmış verilerden tümevarım yöntemiyle ortaya çıkarılan ağaç dalları şeklinde gösterilmiş bir karar yapısı türüdür. Bir karar ağacı, basit karar verme adımları uygulanarak, büyük miktarlardaki kayıtları, daha küçük kayıp gruplarına ayırarak kullanılan bir düzendir. Her başarılı bölme işlemiyle, sonuç gruplarının üyeleri bir diğeriyle çok daha benzer hale getirilebilmektedir (Sun & Li, 2008).

Karar Ağacı modelleri veri madenciliği tanım kümeleri oluşturulmasında tutarlı ve diğer sınıflandırmalara göre hesaplamalarda daha az kaynak tüketmektedir. Karar ağacı sınıflandırıcılar iki aşama üzerinden işlemleri gerçekleştirir: Ağaç oluşturma ve ağaç budama. Ağaç oluşturma aşaması karar ağacı modeli değişken veri grubunu tekrarlanan döngülerle

belirlenen kriterlere göre küçük boyutlara bölmektedir. Bu işlem her parçada bulunan kayıtların aynı sınıf kümesi içerisinde bulunmasına kadar devam etmektedir. Karar ağacı ile genelleştirme yapabilmek için budama işlemi tek veya az sayıda bulunan veri vektörleri sınıflandırmalarını kaldırmak için gerekli bir işlemdir (Du & Zhan, 2002).

Karar ağaçlarında, en altta bulunan ağlar, yapraklar en üstte bulunan ağlar ise kök olarak adlandırılır. Kök sınıflara ayrılan tüm değişken gruplarını barındırır. Yapraklar dışında kalan ağların hepsine karar ağları denir. Karar ağları belirlenen özelliklere göre olası karar seçeneklerini ortaya çıkarırlar. Tüm karar ağacı algoritmaları kavram (konsept) öğrenme algoritması olan Hunt Temel Algoritmasına dayanır (Horis, Pedrycz, Swiniarski, & Kurgan, 2007) (Huang & Hoa, 2009). Bu algoritma insanların basit öğrenme yeteneklerinin değişken grupları içerisinde önemli ayırt edici özellikleri sınıflandırabilme yeteneğine dayanır.

Hunt algoritmasına dayanarak elde edilen algoritmalardan biri olan ID3 (Interactive Dischotomizer3) algoritması tümevarımsal bir meyil içerisindedir. Her ağdaki belirsizlik "Bilgi kazancı" (information gain) değeri hesaplanarak düşürülebilmektedir.

$$I_E(f) = -\sum_{i=1}^{m} f_i \log_2 f_i$$

Değişken grup, S, üzerinde aşağıdaki adımlar izlenerek karar ağacı oluşturulabilir (Horis, Pedrycz, Swiniarski, & Kurgan, 2007):

1. Tüm değişken grubu, S, dahil olacak şekilde kök ağ oluşturulur.
2. Eğer tüm değişkenler pozitif ya da negatifse karar ağacı tek kök ağa sahiptir. İşlem durdurulur.
3. Diğer durumda alanda bulunan her değer için önceden hesaplanan en yüksek bilgi kazanç değeri seçilir.
    a. En iyi değere göre yeni dal eklenir. Belirlenen değer örneklerinin hepsini barındıran yeni bir ağ oluşturulur
    b. Eğer örnekler sadece bir sınıfa ait ağları barındırıyorsa, yaprak ağ şekline dönüştürülür. Aksi durumda yeni alt ağaç halini alır.

Veri madenciliğinde karar ağacı algoritmalarının yanı sıra aşağıda belirtilen kategoriler de mevcuttur.

- Sınıflandırma Ağacı Analizleri (CTA)
- Regresyon Ağacı Analizleri

- Sınıflandırma ve Regresyon Ağacı (CART)
- Ki-Kare otomatik etkileşim algılayıcısı (CHi-squared Automatic Interaction Detector (CHAID))
- Random Forest

Sınıflandırma Ağaçları veri madenciliği görevlerinde çıktıların sınıflandırılabilirliklerini ve tahmin edilebilirliklerini ortaya çıkarır. Burada amaç anlaşılabilir, açıklanabilir kurallar oluşturarak SQL ya da diğer sorgulama dillerine dönüştürebilmektir. Sınıflandırma ağaçları kayıtları etiketler ve aralıklı (discrete) sınıflara atar. Sınıflandırma Ağaçları anlaması basit ve parametrik değildir. Belirli bir niteliğe sahip herhangi bir özel dağılıma (normal dağılım gibi) sahip olmayan sınıfla ilişkilendirilmiş veriye ihtiyaç duymamaktadır. Bu yüzden alışılmışın dışında karakteristik özelliklere sahip verilere uygulanabilmektedir (Clark Labs, 2008).

Regresyon Ağacı Analizleri binary recursive partitioning olarak bilinen işlemler üzerine kurulmuştur. Verileri tekrar eden periyodlarla daha küçük parçalara ayırır ve her dalın kendi içinde bölünmesini sağlar. Başlangıçta değişken gruptaki tüm kayırlar bir bütün halinde bulunur. Daha sonra çeşitli algoritmalar sayesinde her alan binary bölünme ile ayrıştırılmaktadır. Bu işlem kullanıcı tarafından belirlenen minimum ağ büyüklüğüne kadar devam etmektedir.

Sınıflandırma ve Regresyon Ağacı (CART) Analizleri en popüler karar ağacı analizlerindendir. Parametrik olmayan tekniklerle bağımlı değişkenlerin nümerik ya da kategoriksel olup olmadığına göre sınıflandırma ya da regresyon ağaçları ortaya çıkarır. Karar ağacı modellenen veri setinde kurallar bütününe göre belirli değişkenler üzerine kurulur. Kurallar değişken değerlerinin en iyi nasıl bölünmesine göre gözlemlenen bağımlı değişkenlere göre belirlenmelidir. Bir kural belirlendiğinde ve ağ üzerinde bölümleme gerçekleştiğinde, oluşan her yeni ağ için de aynı kural uygulanmak zorundadır. Algoritma tarafından hesaplanan değerlerce bölünme işlemi sonlandırılana kadar kural uygulanmalıdır.

Sınıflandırma ya da regresyon problemlerinde çözüme ulaşmak için çok sayıda metot bulunmaktadır. Ağaç sınıflandırma analizleri, tutarlı sonuçlar ortaya koyduklarında ya da eğer-sonra koşul algoritmaları ile öngörülebilir sınıflandırmalar gerçekleştirdiklerinde diğer alternatif tekniklere göre oldukça fazla avantaja sahip olurlar. Bunlardan biri sonuçların basitliğidir. CART analizlerinde ağaçların yorumlanması oldukça basittir. Bu basitlik sadece

sınıflandırma ve yeni gözlemler için değil, gözlemlerin ve öngörülerin nasıl belirli bir davranış içinde bulunduğunu keşfedilmesini sağlamaktadır.

Diğer avantajları:

- Anlaması ve yorumlanması basittir.
- Veri hazırlık aşaması kısa ve kolaydır.
- Sayısal ve kategoriksel verileri işleyebilir.
- "Beyaz kutu"(White box) modelini Kullanır.
- İstatistiksel araçlarla kullanılan modellerin geçerliliğini sorgulanabilir.
- Büyük veri yığınları ile kısa zamanda sonuçlar alınabilir.

Karar ağacı metotları parametrik ve lineer değillerdir. Sınıflandırma ve regresyon araçları ile kullanılan karar ağaçları sonuçları mantıksal koşul serileri ile özetlenebilmektedir. Bu yüzden öngörücü değişkenler ile bağımlı değişken arasında bulunan ilişkiler lineer, bazı özel bağlantıların takip edilebileceği ya da belli bir düzene sahip olduğu konusunda gizlenmiş varsayımlar bulunmamaktadır. Karar ağacı aynı gelir değişkeni üzerinde çoklu bölme işlemi ile değişkenler arasında monotonik olmayan biri ilişkiyi ortaya koyabilmektedir.

2.3 Geri Yayılım Sınıflandırmaları

Geri yayılım algoritmaları ilk olarak 1979 yılında Arthur E. Bryson ve Yu-Chi Ho tarafından ortaya atılmıştır. Geri yayılım bir sinir ağı algoritmasıdır. Sinir ağları, insan beyninin özelliklerinden olan öğrenme yolu ile yeni bilgi türetebilme, yeni bilgi oluşturabilme ve keşfedebilme gibi yetenekleri herhangi bir yardım almadan doğrudan gerçekleştirmek amacı ile geliştirilen algoritmalardır (Öztemel, 2003).

Sinirsel ağlar uzun işlem süreleri gerektiren ve ancak uygun yazılımlarla uygulanabilmektedir. Ağ topolojisi veya yapılandırması gibi deneysel yollarla belirlenmiş parametrelere ihtiyaç duyar. Sinirsel ağlar yorumlanabilirlik açısından kısıtlıdır. Bu yüzden veri madenciliğinde kullanım alanları yeterince geniş değildir. Öte yandan kirli verilere karşı oldukça toleranslıdır. Önceden tanımlanmamış sınıflandırma modellerini bile öğrenebilmektedir.

Geri yayılım algoritması çıktı değişkenlerindeki hataları hesaplayarak ve hataları her faktörün çıktı hatalarına etkisini göze alarak geriye doğru yayması işlemini gerçekleştiren denetlenebilir hata düzeltim öğrenme algoritmasıdır (Haydar, Ağdelen, & Özbeşeker, 2006). Algoritmanın işleyişi geriye doğrudur. Çıktı katmanından ilk katmana kadar olan tüm gizli katmanlara doğru düzeltmeleri gerçekleştirmektedir. Aşağıdaki adımlar izlenerek uygulanmaktadır (Han & Kamber, 2001):

1. Ağırlıkların belirlenmesi ilk adımdır. Ağ üzerindeki ağırlıklar küçük rastgele numaralarla belirlenir (-1.0 ile 1.0 aralığı gibi). Her birim atanan sayılarla bir ilişki içerisinde olacaktır.

2. Her birimin gizlenmiş ve çıktı katmalarının girdi ve çıktı değerleri hesaplanır. İlk olarak değişken örneklem, ağ içerisindeki girdi katmanlarını beslemelidir. Her birimin gizlenmiş ve çıktı katmalarının girdi ve çıktı değerlerinin hesaplanması girdilerin lineer kombinasyonuyla gerçekleştirilir. Bu işlem gizli ya da çıktı katman birimi yardımıyla açıklanabilir. Birim girdileri önceki katmanın çıktı birimleridir. Birim girdilerinin net değerini hesaplayabilmek için her birim kendilerine atanan ağırlıklarla çarpılarak toplanır. Birim *j* ve net girdileri I_j kabul edersek:

$$I_j = \sum_i w_{ij} O_i + \theta_j$$

Denklemde wj, önceki katmanda bulunan birim j'ye bağlı olan birim i'nin ağırlığını; O_i ise birim i'nin önceki katmandan gelen çıktısı; θ_j 'nin sapmasıdır. Bu sapma birim hareketlerinin değişimi sonucu çıkan limit olarak kabul edilmektedir.

3. Hataların geri yayılımı, ağ için öngörülen hataları yansıtacak şekilde, hata ağırlıkları ve eğilimleri güncellenmelidir. Birim j'deki çıktı katmanlarında oluşacak hatalara Err_j olarak ifade edersek

$Err_j = O_j(1-O_j)(T_j-O_j)$

Belirlenen değişken grup içerisindeki denklemden O_j birim j'nin gerçek çıktısı, T_j ise doğru çıktısı olarak adlandırılır. $O_j(1-O_j)$ mantıksal fonksiyonun türevini temsil etmektedir.

Birim j gizli katman hatalarını hesaplayabilmek için bir sonraki katmanda birim j ile bağlantılı olan birimlerin hatalarının ağırlık toplamları hesaplanmalıdır.

$$Err_j = O_j(1-O_j)\sum_k Errk\,Wjk$$

Wjk bir sonraki katmanda bulunan birim k ile birim j'nin bağlantı ağırlıklarıdır. Err_k ise birim k'nın hatalarını temsil etmektedir.

Geri yayılım algoritmaları yapay sinirsel ağ sınıflandırmaları için çeşitli varyasyonlar ve alternatifler tasarlanabilir. Bu özelliği sayesinde ağ topolojisi ve diğer parametrelerin öğrenme oranlarındaki hatalar anlık düzeltilebilmektedir.

Avantajları:

Doğrusalsızlık: Geri bildirim algoritmalarını barındıran sinir ağları lineer ya da lineer olmayan işlemlerden oluşur, ancak tüm ağ lineer olmayan bir sisteme sahiptir. Özellikle öğrenme görevleri boyunca oluşturulan verilerin gerçek dünya verileri ile tutarlılığı açısından doğrusalsızlık önemli bir özelliktir.

Örneklemlerden Öğrenme: Daha önce belirtildiği gibi değişken grup içerisinde birimler ağırlıkları ile ölçülmektedir. Öğrenme işleminin son aşaması parametrelerin ağ içerisinde yerleştirilmesi ve ele aldığı problemi daha önceden kaydedilmiş bilgilerle birleştirerek çözüme ulaştırmasıdır.

Uyum Sağlayabilirliği: Sinir ağları çevresel değişikliklere uyum sağlayabilecek bir yapıda geliştirilmiştir. Özellikle özel bir çevresel etkeni yöneterek, çevresel koşullarda tekrar ve tekrar değişikliklerle adaptasyon sürecini gerçekleştirir. Durağan olmayan bir çevrede gerçek zamanlı parametreleri benimseyerek sürece dahil eder.

Sonuçların tutarlılığı: Sinir ağları belirlenen örneklem içerisinde belirli sınıflar içerisinde bilgi sağlamakla kalmaz, karar verme aşamasında güvenilirlik seviyesi hakkında da bilgi vermektedir. Bu sayede ortaya çıkan belirsiz verilerin reddedilmesi düşünülebilir ya da

yeniden değerlendirmeye sokulabilir. Böylece ağ tarafından modellenen diğer görevlerde sınıflandırma performansları arttırılabilir.

Hata tolerans seviyesi: Sinir ağları performansı nöron bağlantı sorunları, ya da kirli, eksik veriler gibi ters operasyon koşullarına uyum sağlayabilmektedir.

Analiz ve Tasarımlarda İstikrarlılık: Sinir ağları temelde bilgi işlemcisi olarak istikrarlı bir seyir izlemeyi tercih etmektedir. Benzer ilkeler, formüller ve adımlar neredeyse tüm tanım kümelerinde kullanılabilmektedir.

2.4 Kümeleme Analizleri

Çok değişkenli istatistiksel tekniklerden birisi olan kümeleme analizi, grup sayısı bilinmeyen ve gruplandırılmamış verilerin benzerliklerine göre sınıflandırılması amacıyla kullanılmaktadır. Kümeleme analizi verilerin birimlere veya değişkenlere göre birbirlerine benzerlikleri bakımından ayrık kümelerde toplanmasını sağlayan bir tekniktir (Çakmak, Uzgören, & Keçek, 2005).

Ver madenciliğinde büyük veritabanlarında etkili ve verimli kümeleme analizleri konusunda odaklanılmaktadır. Kümeleme metodlarının ölçeklenebilirliği, karmaşık şekil ve veri yığınlarının kümelemede etkinliği ve sayısal ya da kategoriksel verilerin çok boyutlu ortama aktarılmasında kümeleme analizleri yardımcı olabilmektedir. Kümeleme analizlerinin mevcut uygulamalarda kullanılabilmesi için özel gereksinimlere ihtiyaç vardır. Veri madenciliğinde kümeleme analizleri aşağıdaki özellikleri barındırmalıdır (Han & Kamber, 2001, s. 338):

Ölçeklenebilirlik: Birçok kümeleme algoritması küçük veri setlerinde sağlıklı çalışmaktadır. Ancak milyonlarca veri içeren büyük veritabanlarında çalışmaları oldukça sorunlu olabilmektedir. Bu yüzden büyük veri setlerinden elde edilen örneklemler üzerinde algoritmaları uygulamak daha sağlıklı sonuçlar verebilmektedir.

Çeşitli değişken davranışları üzerinde çalışabilmesi: Çoğu algoritma sayısal veri setleri kümelemesi üzerine çalışabilmektedir. Bunun yanında bazı uygulamalar kategoriksel ya da sıralı veriler üzerinde kümeleme analizlerine ihtiyaç duyabilmektedirler.

Kümelemelerin belirlenmemiş şekiller ile keşfedilmesi: Kümeleme algoritmaları Öklid ya da Manhattan uzaklık ölçümlerini göz önüne alırlar. Bu ölçüm değerleri benzer büyüklük ve yoğunlukta küresel kümelenmeler oluşturmaktadırlar. Ancak veri madenciliğinde kümeleme belirli bir kalıp içerisinde bulunmamalıdır. Bu yüzden algoritmaların belirli bir şekil içerisinde geliştirilmemesi önemlidir.

Girdi parametrelerinin belirlenebilmesi için gerekli minimum tanım kümesi: Kümeleme algoritmaları belirli parametreleri devamlı olarak işleyebilmektedirler. Kümeleme sonuçları girdi parametrelerine hassastırlar. Çok boyutlu nesnelerde parametrelerin tahmin edilebilirliği düşük olabilmekte ve dolayısıyla da kümeleme analizi sonuçlarının geçerliliğini etkileyebilmektedir. Bu yüzden tanım kümesi veri madenciliğine uygun bir şekilde seçilmelidir.

Kirli verilerle kullanılabilirliği: Büyük veri tabanları eksik, tanımlanmamış veya veri madenciliğinde kullanılamayacak verileri barındırırlar. Bazı kümeleme algoritmaları bu verilere duyarlı olabilmektedir. Bu da verimsiz sonuçlara yol açmaktadır.

Girdi kayıtlarının sıralamasına karşı oluşan duyarsızlıklar: Bazı kümeleme algoritmaları aynı veri seti içerisinde bulunan verilerin sıralamasına karşı duyarlıdırlar. Bu duyarlılık veri seti içindeki kayıtların sıralaması değiştirildiğinde farklı kümelemelerin ortaya çıkmasına sebep olmaktadır. Veri madenciliği için sıralamaya duyarsız veri setlerinin oluşturulması önemli bir gerekliliktir.

Çok boyutsallık: Veritabanı ya da veri ambarı birden fazla boyutlu olabilmektedirler. Kümeleme algoritmaları genelde düşük boyutlu verilerle iyi sonuçlar ortaya koyar. Çok boyutlu veri ambarları ya da veritabanlarında kümeleme teknikleri asimetrik dağınık bir sonuç oluşturabilirler.

Yorumlanabilirlik ve kullanılabilirlik: Karar vericiler kümeleme sonuçlarını yorumlanabilir, anlaşılabilir ve karar aşamalarında kullanılabilir olmasını isteyebilmektedirler. Veri madenciliğinde kullanılması gereken kümeleme analizi yöntemlerinde sonuçların gözlemlenebilirliği de göz önüne alınmalıdır. Kümeleme Analizinde kullanılan algoritmaları aşağıdaki şekilde kategorileştirebiliriz (Han & Kamber, 2001, s. 348)

- Bölünme - merkezli (Partition-based) kümeleme

- Hiyerarşik kümeleme
- Yoğunluk - merkezli (density-based) kümeleme
- Grid (ızgara) - merkezli kümeleme
- Model - merkezli kümeleme

2.4.1 Bölünme-merkezli (Partition-based) Kümeleme

Bölünme-merkezli kümeleme algoritmaları, *k* giriş parametresini alarak *n* tane nesneyi *k* tane kümeye böler. Bu teknikler, dendogram gibi iç içe bir kümeleme yapısı üzerinde çalışmak yerine tek-seviyeli kümeleri bulan işlemler gerçekleştirir (Jain, 1999). Bütün teknikler merkez noktanın kümeyi temsil etmesi esasına dayanır. Bölünmeli yöntemler, hem uygulanabilirliğinin kolay hem de verimli olması nedeniyle iyi sonuçlar üretirler (Işık & Çamurcu, 2007). Bölünme merkezli kümelemede veri setinde yapısal olarak bulunan "keşif süreci" belirli nesnel fonksiyon minimizasyonu ile gerçekleştirilmektedir. Algoritmik yapı uygulanırken, nesnel fonksiyonlar hakkında öngörü tanımlanamaz. Bu yüzden nesnel fonksiyondan optimize edilmiş sonuçlar için kümeleme sayısı ve işlemleri önceden belirlenmelidir.

Genel bir optimallik varsayımı için mümkün görünen tüm bölümlemelerin kapsamlı sayımı gerçekleştirilmelidir. Bölünme-merkezli kümeleme çalışmalarında bu işlem için iki genel olgusal metot kullanılır. Bunlar k-mean ve k-medoid algoritmalarıdır. K-mean algoritması ile her kümeleme küme içerisinde bulunan nesnelerin ortalama değeri ile temsil edilir. K-medoid algoritmasında ise kümeleme içerisinde bulunan nesnelerden merkeze en yakın olan kümelemeyi temsil etmektedir. Bu metotlar özellikle küçük ve orta büyüklükteki veritabanlarında bulunan küresel şekle sahip kümelemelerde verimli sonuçlar verebilmektedir. Daha büyük veritabanlarında karmaşık şekilli kümelemeler için k-medoid algoritmasından türetilen CLARANS algoritması gibi daha gelişmiş bölünme-merkezli kümeleme algoritmaları kullanılabilmektedir.

K-mean algoritmalarının işleyişinde ilk olarak kümeleme ortalamasını temsil eden *k* parametresinin rastgele seçimini yapar. Geriye kalan nesneler için, her nesne merkeze uzaklığı ve küme ortalamasına göre benzerlik gösteren bir kümeye atanır. Bu işlemden sonra her kümeleme için tekrar yeni ortalamalar hesaplanır. Bu işlem daha önceden belirlenen kriterler

gerçekleştirilene kadar tekrarlanır. Örneğin sık kullanılan kriterlerden biri olan Karesel hata ölçütü şu şekildedir.

$$E = \sum_{i=1}^{k} \sum_{p \in C} |p - m_i|^2$$

Denklemde *E* veritabanında bulunan tüm nesnelerin karesel hata toplamlarını, *p* nesnenin kümeleme içindeki pozisyonunu ve m_i C_a kümelemesinin ortalama değerini göstermektedir. Ölçüt kümelemeleri yoğunlaştırılmış ve mümkün olduğunca birbirinden ayrık şekilde sonuçlandırma üzerine kurulmuştur.

K-mean algoritmalarında çok büyük değerler veri dağılımını bozabildikleri için algoritma kümeleme içindeki uç değerlere oldukça duyarlıdır. Kümelemelerde bulunan uç değerlerin fazla olması durumunda k-mean yerine k-medoid yöntemine başvurulabilir. K-metoid yönteminde merkezde bulunan nesne kümeleme değeri olarak belirlenir. Böylece veri ne kadar büyük olursa olsun bölünme-merkezli metotlarla her nesnenin referans noktası ile olan ayrılıkları minimize edilebilmektedir.

K-medoid kümeleme algoritmasının ana stratejisi her kümeleme için temsili nesneyi *n* nesne için *k* kümelemelerine atamaktır. Geride kalan her nesne medoid (temsili nesne) olabilecek en benzer kümeleme içinde yerini alır. Bu işlem medoid olarak belirlenemeyen nesnelerin medoid nesnelerle ilişkilendirilerek, belirlenen kriterlere uygun hale gelinceye kadar devam etmektedir.

K-mean algoritması basitliği ve etkinliği nedeniyle k-medoid algoritmasından daha popülerdir. Ancak k-medoid tabanlı algoritmalar dağınık kümelemelerde daha etkili olabilmekte, değişken çeşitlerinde kısıtlamaları göz ardı edebilmekte ve veri seti içindeki sıralamaya bağımlı olmamaktadır. Dikey dönüşümler ve aktarımlara karşı tepkisizdirler (Grandville & Peter, 2005). K-mean algoritmalarının başka bir kısıtlaması ise sadece çok-boyutlu veri ambarlarında çalışmasıdır. Çok boyutlu veri ambarlarında her koordinat için değerler herhangi bir nesne olmasa bile oluşturulabilmektedir. Bu yüzden bu algoritma ile belirlenen kümeleme merkezi veri setindeki nesnelerin değerlerini yansıtamamaktadır. Diğer taraftan ise k-medoid tabanlı algoritmalar çok boyutlu ya da tek boyutlu veri ambarlarında uzaklık fonksiyonu tanımlandığı sürece verimli sonuçlar verebilmektedir (Camila, Barioni, Razente, Traina, & Caetano, 2008).

Daha önce de belirtildiği gibi ne k-mean merkezli algoritmalar ne de k-medoid merkezli algoritmalar büyük veri setleri için uygun değildir. Bölünme-merkezli kümeleme yöntemlerinde büyük veri setleri için CLARA (Clustering Large Applications) metodu kullanılmaktadır. CLARA hesaplamalarda tüm veri setini almak yerine belirlediği daha küçük bir veri setini kullanarak medoidleri tespit etmektedir. Eğer örneklem rastgele verilerin algoritmik hesaplara dayanarak seçildiyse özgün veri setini yansıtması beklenmektedir. Medoid seçimleri özgün verideki seçimlerden farklı olmayacaktır. CLARA tek bir örneklem belirlemez. Birden fazla örneklem içinde yapılan hesaplamalardan en iyi sonucu verenler çıktı olarak gösterilmektedir.

Büyük ver setleri için kullanılan CLARA benzeri bir algoritma da CLARANS'tır (Clustering Large Applications Based Upon Randomized Search). CLARANS metodu CLARA tekniğinde bulunan sınırlandırmaları içermez. CLARA her adımda önceden belirlenmiş örneklemi kullanırken CLARANS sonuçlandırdığı adım sonrası yeni örneklem oluşturur.

2.4.2 Hiyerarşik Kümeleme

Hiyerarşik kümelenme algoritmaları iç içe geçmiş gruplardaki verileri karar ağacı ya da dendogram formunda gösterilecek şekilde düzenler (Kantardzic, 2001, s. 120). Hiyerarşik analizlerde girdi olarak belirlenen kümeleme sayısı belirlenmez. Çoğu yöntem kümeleme optimizasyonunu göz önüne almaz. Asıl amaç, yakınsaklık sağlanana dek tekrarlamalara devam ederek yaklaşık ya da ideale yakın sonuçlar ortaya koyabilmektir.

Hiyerarşi kümeleme analizinde algoritmaları iki gruba ayırabiliriz. Bunlar; Bölünebilir algoritma (divisible algorithm) ve yığmacı algoritma (agglomerative algorithm). Bölünebilir algoritma bütün bir veri örneklemini ele alarak alt veri örneklemlerine bölünmesini sağlar. Böylece daha geniş bir kümelemeden daha iyileştirilmiş kümelemelere göre bir sıralamaya imkân sağlamaktadır. Yığmacı algoritma ise her nesneyi ilk küme olarak ele alır. Daha sonra nesneler daha geniş kümelere atanır. Birleştirme işlemi tüm nesnelerin geniş ve tek bir küme içine alınmasına kadar devam etmelidir. Gerçek hayatta yığmacı algoritmalar bölünebilir algoritmalara göre daha çok kullanılmaktadır. Çoğu yığmacı hiyerarşik kümeleme algoritması tek-bağlantı (single-link) ya da tam-bağlantı (complete-link) şeklinde değişkenlerle oluşur.

Tek-bağlantı metodunda iki kümeleme arasındaki örneklem içerisindeki iki küme mesafesini minimum uzaklıkta belirlerken, tam bağlantı metodunda bu mesafe maksimum olacak şekilde ayarlanır. Tek link metodu daha basit ve pratik olmasına rağmen tam link metodu çoğu uygulamada daha iyi hiyerarşik sonuçlar ortaya koyabilmektedir.

2.4.3 Yoğunluk-merkezli Kümeleme

Yoğunluk-merkezli kümeleme yöntemi Anil K. Jains tarafından 1988 yılında yayınlanan "Algorithms For Clustering Data" adlı makalesinde k-boyutlu nokta kümelerini tanımlamak için ortaya atılmıştır (Ester, Kriegel, Sander, & Xu, 1996). Özellikle, düzenli olmayan ya da daha önceden belirlenemeyen şekillere sahip kümelemelerde kullanılmaktadır. Nesnelerin yoğun bölgelerine göre kümeleme yoluna giden bu tür kümeleme çeşitleri DBSCAN, OPTICS, DENCLUE, gibi algoritmalarla oluşturulur.

(i) DBSCAN

"Density-Based Clusternig Based of Applications with Noise" kısaltması olan DBSCAN yoğunluk-merkezli kümeleme algoritmalarından biridir. Algoritma veri setleri içerisinde yüksek yoğunluklu bölgeleri belirleyerek kümeler ile uzamsal veritabanlarında düzgün bir şekle sahip olmayan kümelemeleri tespit edilen kirli noktalar yardımıyla belirler. Kümeleme içerisinde bulunmayan her nesne kirli noktalar olarak adlandırılır.

DBSCAN veritabanındaki her noktanın ϵ- komşuluk değerini kontrol ederek kümeleme yapmaya çalışır. ϵ- komşuluk ϵ yarıçapı içinde ilişkisel yakınlık içerisinde bulunan nesneler için belirlenen değerdir. Eğer *p* noktasının ϵ- komşuluk MinPts adı verilen minimum değerden büyükse *p* noktası çekirdek nesne olarak kabul edilir. Kümeleme işlemi çekirdek nesnelerin tekrarlanan yoğunluk testleri sonucu kümelenmesi ile gerçekleşir. İşlem herhangi bir kümeye yeni bir nokta eklenememesine kadar devam eder.

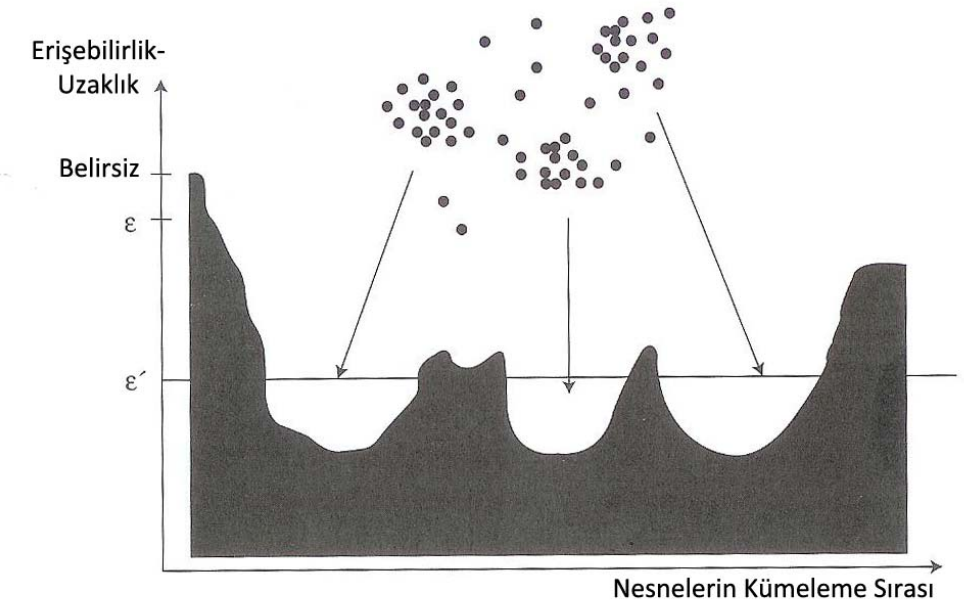
DBSCAN algoritmasının avantajlarını şu şekilde sıralayabiliriz:

- DBSCAN algoritmalarında k-means algoritmalarına göre kümeleme sayısı önceden bilinmek zorunda değildir.

- DBSCAN kümeleri herhangi bir şekil şartına bağlı değildir. Her bir kümeleme farklı bir şekle sahip olabilir.
- DBSCAN veritabanı ya da veri seti içerisindeki sıralamaya duyarsız iki parametreye ihtiyaç duyar. Böylece kümeleme herhangi bir şarta bağlı olmadan oluşturulabilir.

(ii) OPTICS

DBSCAN ϵ ya da MinPts gibi belirlenen girdi parametrelerine göre kümeleme yapmasına rağmen kullanıcı için gerekli olan diğer parametrelerin seçimine izin vermeyerek kabul edilebilir diğer kümelemeleri analiz dışı bırakmaktadır (Han & Kamber, 2001, s. 365). Ancak bu sorun gerçek dünyada çok boyutlu veri setlerinde diğer parametrelerin kolay bir şekilde belirlenememesinden de kaynaklanmaktadır. Çoğu algoritma gerçek dünya parametrelerine oldukça duyarlı, bu yüzden de kümeleme alanları küçük değişimlerden fazlaca etkilenmektedirler. OPTICS ile veri seti kümeleri belirli kalıplarla kümelemek yerine, belirli otomatik ve interaktif kümeleme analizleri ile alternatif kümeleme sıralamaları üzerine analizler gerçekleştirilebilmektedir. Bu sıralamalar yoğunluk-merkezli kümeleme yapısına göre gerçekleşmektedir. Geniş bir parametre kaynağının analizlere girmesiyle yoğunluklar tespit edilmektedir.



Şekil 2-3 OPTICS Algoritmasıyla Kümeleme (Ankerst, Breunig, & Kriegel, 1999)

OPTICS algoritması sonsuz sayıda uzaklık parametresi □ parametresinden çok daha küçük olan □′ için uygulanan genişletilmiş bir DBSCAN algoritması olarak kabul edilebilir. Ancak OPTICS algoritmalarında herhangi bir kümeleme bağı atanmıyor, onun yerine olabildiğince fazla parametre içeren algoritma için kullanılabilecek bilgiyi ve işlemden geçirilen nesnelerin sıralamasını kayıt altına alınmaktadır. Toplanan bilgiler her nesne için iki değerden meydana gelmektedir. Bunlar "merkezi uzaklık" ve "erişebilir uzaklık". (Ankerst M. , Breunig, Kriegel, & Sander, 1999).

p nesnesinin merkezi uzaklığı *p*'yi merkez nesne yapan en küçük □$_1$ değeridir Eğer *p* merkez nesne değilse, *p*'nin merkezi uzaklığı belirtisidir.
q nesnesinin erişebilir uzaklığı diğer *p* nesnesine göre *p* nesnesinin merkezi uzaklığından büyük ve Öklid uzaklığı *p* ile *q* arasındadır. Eğer *p* merkezi nesne değilse, *p* ile *q* arasındaki erişebilir uzaklık belirsizdir.
Küme sıralaması aşağıdaki şekilde daha açık bir şekilde görülebilir. Basit iki boyutlu veri seti oluşturulabilir kümelemeleri göstermektedir.

(iii) DENCLUE

DENCLUE(Density-Based Clustering) yoğunluk dağılım fonksiyonlarından oluşan bir kümeleme modelidir. DENCLUE algoritmaları iki aşamada gerçekleşir. Birinci adım ön kümeleme adımı olarak kabul edilir. Ver setinin ilgili alanları üzerine bir harita oluşturulur. Bu harita veri alanı içerisindeki çevresel alanlarına hızlı ve etkili ulaşım için gerekli olan yoğunluk fonksiyon hesaplamalarını hızlandırmak için kullanılır. İkinci adımsa gerçek kümeleme adımıdır. Algoritma yoğunluk oluşturucularını (density attractors) ve yoğunluğa göre hareket eden noktaları tanımlamaktır (Hinneburg & Keim, 1998).

DENCLUE algoritmalarında ilk aşamasında etki fonksiyonları ön plana çıkmaktadır. Etki fonksiyonu her veri noktasının etkilerinin matematiksel olarak gösterilmesidir. Bu sayede çevre üzerindeki etkisi gösterilebilmektedir. Veri alanının toplam yoğunluğu da tüm veri noktalarının etki fonksiyonları toplamına eşit olmaktadır. İkinci aşamasında ise toplam yoğunluk fonksiyonunun maksimum değerine göre yoğunluk oluşturucuların belirlenmesi ile sonuçlanmaktadır.

DENCLUE diğer algoritmalardan farklarını şu şekilde sıralayabiliriz:

- Bölünme-merkezli kümeleme ve hiyerarşik kümeleme gibi algoritma metotları da dahil çoğu kümeleme algoritmasını kapsar, daha gelişmiş matematiksel fonksiyonlara sahiptir.
- Yüksek miktarda kirli verilere sahip veri setlerinde başarılı bir algoritmadır.
- Çok boyutlu veri setlerinde belirli bir sabit şekle sahip olmayan kümelemeleri tanımlamak için matematiksel tanımlamalara izin verir.
- Veri noktalarının yanı sıra fonksiyonel bilgileri de grid hücrelerinde barındırırlar. Bu bilgiler ağaç-merkezli yapılarla yönetilir. Bu yüzden DBSCAN gibi algoritmalardan daha hızlıdır.

2.4.4 Grid (Izgara) Merkezli Yöntemler

Grid merkezli kümeleme metotları büyük veri setleri içerisinde kullanılmaktadır. Bu kümeleme metotlarında özellik alanı ızgara şeklini alan dörtgensel hücrelere bölünmüşlerdir. Tüm kümeleme işlemleri bu ızgara benzeri alanda gerçekleşmektedir (Kunttu, Lepistö, Rauhamaa, & Visa, 2004). Grid merkezli algoritmalar her girdi değişkeni için modelleme oluşturulur. Sonuçların ortaya çıkmasını sağlayan kurallar her değişken için atanan muhtemel değerler bütünü ya da değerlerden oluşan örneklem kullanılarak ortaya çıkarılır (Kantardzic, 2001, s. 268).

Bölünme-merkezli kümeleme için oluşturulan farklı geometrik şekillerde bazı problemlere neden olabilmektedir. Grid-merkezli kümeleme bu sorunları nesneleri kutular şeklinde bölerek çözebilmektedir. Kutular birleştirildiğinde geometrik şekillerin içerisinde görülebilir çeşitliliği hesaba katmaksızın kümelemenin gerçekleştiği fark edilebilmektedir.

Grid-merkezli kümeleme özelliklerini şu şekilde özetleyebiliriz (Horis, Pedrycz, Swiniarski, & Kurgan, 2007, s. 274):

- Grid-merkezli kümeleme algoritmasının verileri bir kez taraması yeterlidir. Bu özellik büyük veritabanlarının işlenmesinde kolaylık sağlar.

- Oluşturulan temel yapı kutularını göz önüne aldığımızda, çok farklı geometrik şekillerle kümeleme imkânlarını arttırmaktadır.
- Grid-merkezli kümeleme baskın olarak yoğunluğu göz önüne aldığından dolayı, belirli bir geometrik şekle sahip olmayan kümelemelerle de işlemleri gerçekleştirebilir. Benzer şekilde, nokta yoğunluğuna göre dış sınır noktalarının yerinin tespit edilebilir.

Grid-merkezli kümeleme yöntemlerinde en çok kullanılan STING, Wavecluster ve CLIQUE algoritmalarıdır.

(i) STING

STING uzamsal alanın dörtgensel şekillere bölündüğü çoklu çözünürlük sağlayan kümeleme tekniğidir. Dikdörtgen şeklinde oluşturulan hücreler farklı çözünürlük seviyelerine göre çeşitlilik gösterirler. Her hücre daha yüksek bir seviyede bulunan hücrelerden ayrılarak daha düşük bir seviyede konumlanır. Her grid hücresinde depolanan istatistiki bilgi algoritma için gerekli olan sorgusal işlemleri gerçekleştirilmesini sağlar.

STING çeşitli avantajlar sağlamaktadır. Bunlar sırasıyla aşağıdaki gibidir:

(1)Grid-merkezli kümeleme hesaplamalarında sorgulamalar birbirinden bağımsızdır. Bunun sebebi her grid hücresinde bulunan istatistiksel bilgiler verilerin özet bilgilerini içerir.

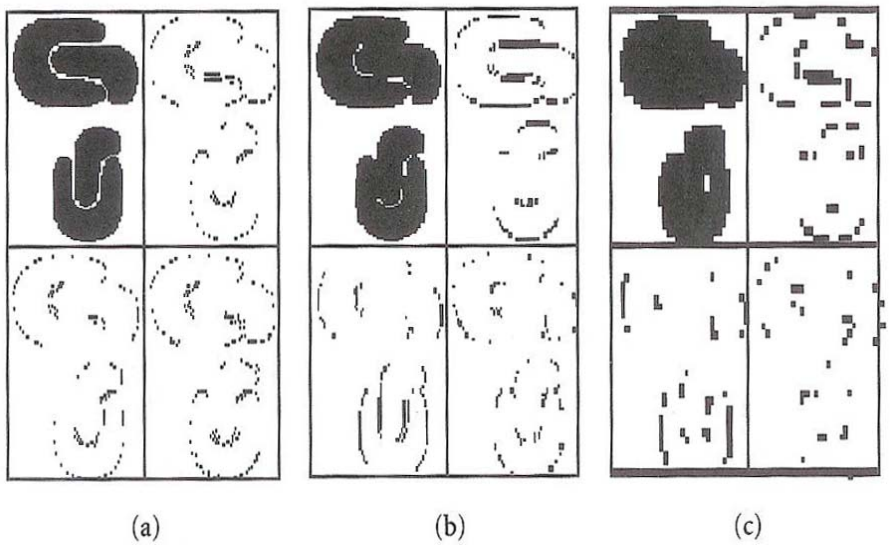
(2)Grid yapıları paralel işlemlere ve devamlılık gerektiren güncellemeler konusunda kolaylık sağlar.

(3)Metodun işleme yöntemi en büyük avantajıdır. STING hücrelerin istatistiksel parametrelerini hesaplamak için veritabanını kullanır. Bundan dolayı kümelemeler oluşturulurken ortaya çıkan zaman karışıklığı O□ (x nesne sayısı) olarak ifade ederek, hiyerarşik yapının oluşturulma süresini de O_2 (a en düşük seviyedeki grid hücre sayısı) olarak belirtilirse *a* genelde *n* değerinden çok daha küçük bir değere sahip olur.

(ii) WaveCluster (Dalga Kümeleme)

Wavecluster konkav ya da iç içe geçmiş gibi karmaşık yapıdaki kümeleme şekilleri üzerinde çalışan bir algoritmadır. Wavecluster algoritmasında gerekli kümeleme sayısı için bir ön çalışmaya gerek yoktur. Ancak beklenen kümeleme sayısını hesaplamak, kümelemeler için uygun çözünürlüğün seçiminde kolaylık sağlayacaktır (Sheikholeslami, Chatterjee, & Zhang, 1998).

Wavecluster öncelikle çok boyutlu grid yapısını uygulayarak veriyi özetleyen çoklu çözünürlük sağlayan bir algoritmadır. Algoritma daha sonra wawelet transformasyonu denilen bir teknikle özgün alanı yoğunluk bölgelerine ayırmaktadır. Wavlet transformasyon farklı frekanslara göre sinyalleri ayırmada kullanılan sinyal işleme yöntemidir. Bu model *n* boyutlu sinyallerde her bir sinyali n defa dönüştürme işlemidir. Wavelet transformasyon metodu uygulanırken farklı çözünürlük seviyelerinde nesneler arası uzaklığı koruyarak dönüştürme işlemini gerçekleştirmektedir. Böylece, veri seti içerisinde oluşan doğal kümelenmeleri düzgün bir şekilde ayıt edilebilir kılmaktadır.



Şekil 2-4 (a)Yüksek çözünürlükte, (b) orta çözünürlükte, (c) düşük çözünürlükte wavelet tranformasyonu kümeleme örnekleri (Sheikholeslami, Chatterjee, & Zhang, 1998)

Wavlet transformasyonunun avantajları şu şekildedir (Han & Kamber, 2001, s. 372):

- Denetlenmemiş kümeleme sağlar. Şapka şekilli filtrelerle kümeleme alanlarından eksik, düzensiz bilgileri temizlerken kümelemenin oluştuğu alanları açığa çıkarır.
- Wawelet transformasyonu çoklu çözünürlük özelliği ile kümeleme içindeki farklı seviyelerdeki tutarlığı belirleyebilmektedir.
- Wavelet kümeleme sayısal karmaşık hesaplamalarıyla hızlı bir algoritmadır. Diğer algoritma uygulamalarıyla paralel olarak çalışabilir

(iii) CLIQUE

CLIQUE yüksek yoğunluklu kümelemeler içerisinde alt uzayda otomatik olarak bulan bir algoritmadır. Girdi kayıtlarının gösterildiği, herhangi bit varsayımsal dağılımın göz önüne alınmadığı ve sıralamanın önemsiz olduğu özdeş sonuçlar ortaya koyar (Agrawa, Gehrke, Gunopulos, & Raghavan, 2005).

CLIQUE kümeleme metodunda belirlenen geniş çok boyutlu veri noktaları içerisinde veri alanı düzenli olarak veri noktalarınca oluşturulmamıştır. Algoritma alandaki seyrek ve yoğun bölgeleri tanımlar, böylece veri seti için genel dağılım modellerini ortaya koyar.

2.5 Birliktelik Kuralları

2.5.1 Pazar Sepeti Analizi

Birliktelik kuralı ve ardışık zamanlı örüntüler, "ilişki analizi" adı altında yer alır ve pazarlama amaçlı olarak pazar sepeti analizi adı altında veri madenciliğinde yaygın olarak kullanılır (Timor & Şimşek, 2008). Örneğin bir müşterinin markette ya da sanal bir mağazada yaptığı alışveriş pazar sepeti analizinin konusu içerisinde yer alır. Yapılan işlemler sonucunda kimi satıcılar milyonlarca veriye sahip bir veritabanına sahip olabilirler. Bu işlemlere karşı yapılan yaygın analizlerden biri işlem demetlerinde birlikte görülebilen ürünlerin tespiti üzerinedir. Ürün grupları kullanıcı tarafından belirlenen minimum değerinin üzerinde olmalı, böylece sıklıktan bahsedilebilmektedir.

Pazar analizlerinde ürün gruplarının sık olmasının önemi, müşteri işlemlerinin çok büyük olabilmesi ve bilgisayar işlemcilerinin yetersiz kalmasına sebep olabilmesinden kaynaklanmaktır. Bir başka nokta ise ürün grubunun gerçek sıklık sayılarının çok daha küçük olmasına rağmen ürün grubunun bulunan sıklık sayısı (frekansı) değişik ürünler tarafından üssel şekilde arttırılabilmektedir. Bu yüzden ölçeklenebilir algoritmalar (karmaşıklıkları üssel olmasına göre değil, işlem sayısına göre doğrusallığı arttıracaktır) pazar sepeti analizlerinde sıklıkla kullanılmaktadır (Kantardzic, 2001, s. 166).

Pazar sepeti analizlerinde ürün grubu içerisinde benzerlikleri keşfetmek için en sık kullanılan analizlerden biri ilginlik (affinity) analizidir. İlginlik analizi belirli bireyler ya da gruplar tarafından gerçekleştirilen satın alma aktivitelerini arasındaki birliktelik ilişkilerini ortaya çıkarmak için kullanılır. Genelde, ajanlar eşsiz olarak tanımlandığı ve yapılan işlemler hakkındaki bilgilerin kaydedilebildiği durumlarda uygulanabilmektedir. Özellikle pazarlama alanında çapraz satış (cross selling), ek satış (up selling) gibi amaçların yanında satış promosyonları, sadakat programları, mağaza tasarımı ve indirim planları gibi uygulamaların karar aşamalarında karar vericiler tarafından sıklıkla başvurulan bir metottur (Wikipedia, 2010).

2.5.2 Apriori Algoritması

Apriori algoritması birliktelik kuralları içerisinde en çok bilinen ve kullanılan algoritmadır. Sıklık değişken setleri içeriğinin aranması için kullanılan değişken gruplarının sayısının azaltılmasını temel almaktadır. Bu algoritma iki aşamada çalışır: ilk adımda Alışılmış değişken seti sınırlandırılmasıdır. Bu set belirlenen işlemler içerisinde en az yüzdeye sahip olan değişkenlerden oluşur. İkinci aşama ise birliktelik kuralı alışılmış değişken gruptan oluşturulmaktadır. İlk adım işlem sürecinde daha çok zaman alan bölüm olduğu için daha önemli bir aşamadır (Borgelt & Kruse, 2002).

Apriori seviye tespit (level-wise) araması olarak bilinen tekrarlanabilir bir yaklaşımı benimser. Bu yaklaşımda k değişken grubu $k+1$ değişken grubu üzerinde araştırma yapmak için kullanılır. Öncelikle 1. Değişken grubu tespit edilir. Bu set L_1 şeklinde gösterilir. Daha sonra L_1 2. Değişken grup olan L_2'yi bulmak için kullanılır ve bu işlem daha fazla k değişken grubu bulunamamasına kadar devam eder. Her L_k tespiti tam bir veritabanı taraması gerektirir. Sık kullanılan değişken grubu üzerinde yapılan seviye tespit aramasının verimliliğini

arttırmak için Apriori niteliği kullanılır. Bu sayede arama sırasında oluşan boşluklar azaltılabilmektedir.

Apriori niteliğinde tek bilinmeyen C_k'nın nasıl uygulanacağının bilinmemesidir. C_k , L_k içinden seçilmiş aday veri setleridir. Bu iki adımın yapılandırılması şu şekilde gerçekleşmektedir (Horis, Pedrycz, Swiniarski, & Kurgan, 2007, s. 296):

1. L_{k-1} içinden atanan her sıklık değişken grup için bu gruba ait olmayan ancak başka bir sıklıkta *(k-1)* bulunan *i* verileri tespit edilir. *k*-değişken grubunun oluşturulabilmesi için *i* gruba eklenir. Birden fazla olan *k* değişken grupları *k-1* değişken grubu oluşturulduktan sonra ortadan kaldırılır.
2. Eğer L_{k-1} den üretilmiş sıklık *(k-1)* değişkenleri ile (*k-2)* değişken arasında ortak noktalar var ise *(k-2)* değişken grubuna 2 farklı değişken eklenerek *k*-değişken grupları oluşturulmalıdır.

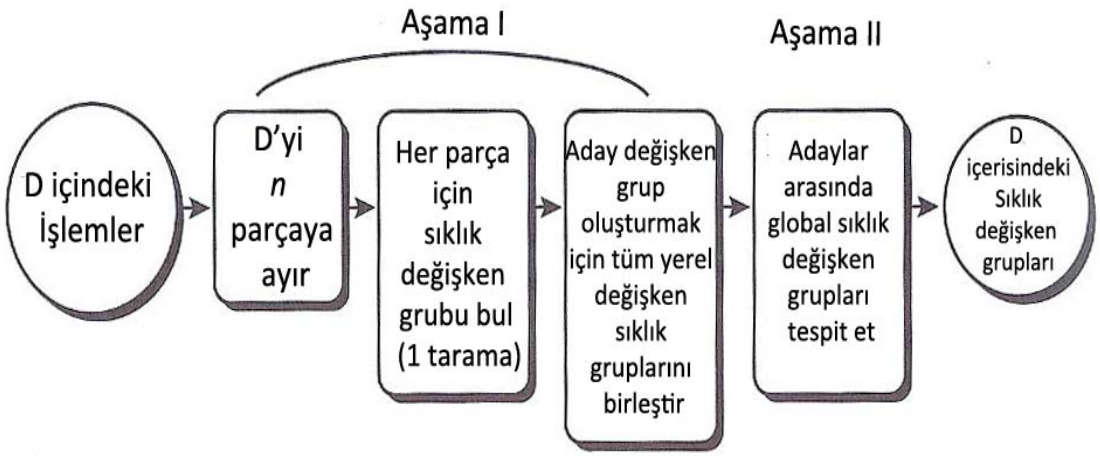
Apriori algoritması bazı metotlar kullanılarak daha etkili bir şekilde kullanılabilmektedir. Aşağıda bu metotlardan da söz edilmiştir:

Rastgeleleştirme: Bu metot C_k *k>1* koşulunda aday k-değişken grubunun boyutunu düşürmek için kullanılmaktadır. Örneğin 1. Sıklık öğe kümesi, L_1, tanımlamak için C_1 içindeki aday 1-öğe küme içerisinde veritabanında bulunan her işlem ayrı ayrı taranırken, her işlem için aynı zamanda 2-öğe kümeyi de oluşturulmuş olmaktadır. Bu öğe küme rastgele bir biçimde belli yapılara göre belli alanlarda biriktirilmekte, böylece alanların sayısı arttırılmaktadır. 2-öğe küme içerisinde bulunan alanlar içinde sayıları sıklık kurallarına göre belirlenen değerlerin altında olması durumunda alanlar ortadan kaldırılır. Bu sayede belirlenen aday k-öğe küme sayısı düşürülmüş olur.

İşlem sayısını azaltma: Sıklık öğe küme içermeyen işlemlerin kaldırılması işlemidir. Eğer bir işlem sıklık öğe küme içermiyorsa *k+1* için de sıklık öğe kümesi içerdiği söylenemez. Bu yüzden *j* > *k* koşulunda sonradan taranan j- öğe kümesi için bu işlemler dikkate alınmaz.

Veri Seti bölümleme: Bölümleme tekniği Sıklık Öğe Kümenin belirlenebilmesi için iki veritabanı üzerinde çalışmaktadır. İki aşamadan meydana gelmektedir. 1. aşamada algoritma D işlemlerini üst üste gelmeyen bölümlere ayırır. D içerisindeki minimum desteğe *min_sup* dersek, bir bölümleme için minimum öğe küme sayısını *min_sup x* bölümleme başına işlem

sayısı olarak gösterebiliriz. Her bölümleme için, bölümleme ile sıklık öğe kümeleri bulunur. Bu durumda oluşan sonuçlara yerel sıklık öğe kümesi denmektedir. Yöntem özel bir veri yapısı kullanarak öğe kümesi içindeki her veriyi işler. Bu şekilde tek bir veritabanı taraması ile k-öğe kümesi içerisindeki tüm yerel sıklık k-öğe kümelerini ortaya çıkarır. Ayrıca tüm veritabanı içerisinde belirli bir sıklık bulunmaya bilmektedir. D veritabanı içersinde herhangi bir sıklık öğe kümesi içerisinde bulunan bir veri en azından bölümleme sonucu oluşan bölgelerden birinde de mevcut olmalıdır. Tüm bu bölgelerde oluşan sıklık öğe kümelerinin bütününe ise küresel aday öğe kümeleri denmektedir. 2. aşamada D veritabanı üzerindeki ikinci taramada küresel sıklık öğe kümelerini belirlemek için her aday öğe kümesinin belirlenmesi üzerine kurulmuştur. Bölümleme boyutu ve sayısı, her bölümlemenin ana hafızaya göre belirlenmesi ve her aşamada sadece bir kez okuma izniyle gerçekleştirilmesi gerekmektedir.



Şekil 2-5 Apriori Uygulama Adımları

Örnekleme: Örnekleme yaklaşımında ana fikir belirlenen veritabanı D içerisinden rastgele örneklem S'nin çekilmesidir ve öğe kümelerinin tüm veritabanı yerine bu örneklemde araştırılmasıdır. S örneklem boyutu içerisindeki sıklık öğe kümelerinin araması sadece ana hafıza içinde ve bir kez gerçekleştirilebilmektedir. Tüm veritabanı yerine sadece örneklem üzerinde yapılan arama nedeniyle bazı küresel sıklık öğe kümelerinin gözden kaçması mümkündür. Gözden kaçırmaları en aza indirebilmek için, mümkün olduğunca düşük destek noktalarından faydalanılması gerekmektedir.

Aday öğe kümeleri oluşturmadan sıklık öğelerin tespiti: Apriori algoritmasını kısıtlayan durumlardan bir tanesi algoritmanın çok büyük sayıda aday öğe kümeleri oluşturabilmesidir. Örneğin; 1-öğe kümesi içerisinde 10000 öğe için, 2-öğe kümesi içerisinde 100000000 aday öğe kümesi tespit edebilmektedir. Başka bir kısıtlayıcı etken algoritmanın büyük miktarda adaya öğe kümesi için veritabanı üzerinde birden fazla tarama yapma gereği duyabilmesidir. En uzun öğe kümesini *n* olarak belirlersek *n+1* tane tarama gerçekleştirmesi gerekmektedir. Bu sınırlamaları kaldırmak için problemi daha küçük parçalara bölerek çözebilen böl-ve-elde et (divide-and-conquer) metodu, veya daha hızlı bir metot olan ağaç izdüşümü (tree-projection) algoritması kullanılabilmektedir.

2.6 Diğer Sınıflandırma Metotları

2.6.1 Genetik Algoritmaları

Genetik algoritmalar doğal gelişim süreçlerini kapsamaktadırlar. Süreç rastgele oluşturulmuş kurallara göre ilk populasyonun oluşturulmasıyla başlar. Her kural bir dizi parça ile temsil edilir. Örneğin bir değişken içerisinde X_1 ve X_2 öznitelikleri ve C_1 C_2 sınıfları bulunsun. " Eğer X_1 ise ve X_2 değilse, C_2'dir" kuralı "100" parça değeri olarak tanımlansın. Bu parça değerinde en soldaki değerler X_1 ve X_2 öznitelikleri, en sağdaki ise sınıfı temsil etmektedir. Eğer kural "Eğer X_1 ve X_2 değil ise C_2'dir" olsaydı parça değeri "001" olacaktı. Eğer bir değişken *k* değerine sahipse, *k>2* olması şartıyla *k* parçaları öznitelik değerlerinin kodlanması için kullanılabilir. Sınıflar benzer şekilde kodlanmalıdır.

Doğal Seleksiyon kuralına göre, yeni popülâsyon mevcut popülâsyon içerisindeki en iyilerden oluşacaktır. Kuralların yaşam döngüsü içindeki yeri değişken grubun sınıflandırabilmesine göre değer biçilmektedir. Alt kurallar genetik aktarım ve mutasyon gibi genetik operatörlerce belirlenmektedir. Genetik aktarımda kural çiftlerinin at dizileri yeni kural çiftleri ile yer değiştirmekte, mutasyonda ise rastgele seçilmiş kurallar yeni parçalara dönüşmektedir.

Genetik algoritmalar kolay bir şekilde birbirinden bağımsız şekilde çalıştırılabilir ve optimizasyon problemlerinin yanında sınıflandırma konusunda da başarılı bir şekilde

kullanılabilmektedir. Veri madenciliğinde de diğer algoritmaların veri ambarı içindeki uygunluğu konusunda yardımcı olabilmektedir.

2.6.2 Bulanık Küme Yaklaşımları (Fuzzy Set Approaches)

Kural odaklı sınıflandırmalar sürekli değişkenler için ani sonlanmalar nedeniyle dezavantaj oluşturabilmektedir. Örneğin, aşağıda bir kütüphanenin öğrenciler için kitap ödünç verme kuralı verilmiştir. Kurala göre, öğrenci eğer 1. sınıf veya 1. sınıftan büyükse ve daha önce en fazla 5 geciktirmede bulunmuşsa kitap ödünç verilebilmektedir.

EĞER(fakülte_sınıf ≥ 1) YA DA (geciktirme ≤ 5) SONRA ödünç_durum = "uygun"

Bu durumda ödünç işlemini 5 kez geciktiren rahatlıkla kitap alabilirken 6 kez geciktiren sistemden yararlanamamakta ve kısmen bir adaletsizlik ortaya çıkabilmektedir. Bu durumun önüne geçebilmek için sınırları belirlemek amacıyla "bulanık bölgeler" belirlenmelidir. Keskin sınırlar belirlemek yerine bulanık mantık 0.0 ile 1.0 arasında değerler belirleyerek belirlenen kategoride derecelendirebilmektedir. Bu durumda eğer diğer şartları sağlaması durumunda 6 kez geciktirenler de sistemden yararlanabilmektedir.

Veri madenciliğinde bulanık mantık yaklaşımı yüksek seviyede soyutlama yeteneği kazandırmakta, sınıflandırmanın başarısını arttırabilmektedir.

2.7 Web Madenciliği

Web madenciliği, Web hiperlink yapısı, sayfa içeriği ve kullanım verilerinden faydalı bilgi keşfi sürecidir (Liu, 2007, s. 6). Web madenciliği teknikleri doğrudan ya da dolaylı olarak web ortamında bulunan düzensiz bilgi problemlerini çözmek için kullanılır. Web madenciliği kendi tekniklerinin yanı sıra birçok veri madenciliği tekniğini de kullanır. Bunun sebebi verilerin heterojen, yarı yapılandırılmış ya da yapılandırılmamış olmasından kaynaklanmaktadır.

Madencilik işlemi sırasında kullanılan verilerin çeşitliliğine göre, web madenciliği görevleri üç kategoriye ayrılabilir (Kosala & Blockeel, 2000). Bunlar sırasıyla; Web yapı madenciliği, Web içerik madenciliği, Web kullanım madenciliği olarak adlandırılır.

Web yapı madenciliği: Sayfalar arası bağlantılar (hyperlink) sayesinde web yapısını ortaya çıkarılmasına olanak veren kullanılabilir bilgi keşfine denir. Web arama motorlarında oldukça yaygın kullanılan teknik ile bağlantılar vasıtasıyla belirli kriterlere göre web sayfalarının önemliliği belirlenebilmekte ya da belirli web topluluklarının ortak ilgili alanlarının tespiti için kullanılabilmektedir. Geleneksel veri madenciliği tekniklerinde ilişkisel tablolar arası bir bağlantı yapısı mevcut olmadığından bu görevler gerçekleştirilemez.

Web İçerik Madenciliği: Web sayfaları içeriğinden kullanılabilir bilgilerin çıkarılması işlemidir. Veri madenciliği teknikleri ile benzerlik gösterir. Web içerik madenciliğinin amacı kullanıcı profillerine göre bilgiye ulaşılabilmesini sağlamaktır. Bu amaç için oluşturulan veri modellemeleri ve bütünleştirmeler kelime bazlı aramalardan çok gelişmiş sorgulama yöntemleri ile oluşturulur.

Web Kullanım Madenciliği: Kullanıcı tarafından her işlemin tutulduğu Kullanıcı kayıtları (log) yardımıyla kullanıcı erişim düzeninin keşif sürecidir. Kullanıcı oturumları ve davranışları veri kaynağını oluşturmaktadır. Veri kirliliğinin yüksek olduğu bu metotta kayıtlardan doğru verilerin çekilmesi öncelikli işlemlerdendir. Web içerik ve yapı madenciliği gerçek ve ana veriyi işlerken, web kullanım madenciliği kullanıcı etkileşimi ile ortaya çıkan ikincil verileri kullanmaktadır. Bu veriler genelde erişim kayıtları, Proxy sunucu kayıtları, tarayıcı kayıtları, kullanıcı profilleri, kullanıcı kayıt verileri ve çerezlerden (cookies) oluşur.

	Web Madenciliği			
	Web İçerik Madenciliği		Web Yapı Madenciliği	Web Kullanım Madenciliği
	Bilgi içeriği	Veritabanı yapısı		
Veri Görünümü	- Yapılandırılmamış - Yarı yapılandırılmış	-Yarı yapılandırılış -Web sitesi olarak Veritabanı	-Link Yapısı	- Site içi Aktivite
Ana Veri	- Metin Dokümanları - Dinamik Dokümanlar	-Dinamik Dokümanlar	-Link Yapısı	- Sunucu kayıtları - Tarayıcı kayıtları
Sunum	- Terimler, cümle parçaları - Belirlenen Modeller - İlişkiler	- Etiketlenmiş Grafikler (OEM) - İlişkiler	-Grafik	- İlişkisel tablolar - Grafik
Metot	- Varyantlar - Yapay Sinir Ağları - İstatistiksel	- Bayesyen Algoritma - Yapay Sinir Ağları - Birliktelik Kuralları	-Bayesyen Algoritmalar	- Yapay Sinir Ağları - İstatistiksel - Birliktelik Kuralları
Uygulama Kategorileri	- Kategorizasyon - Kümeleme - Çıkarım Kuralları Oluşturma - Metin Modellemeleri - Kullanıcı Modelleri	- Alt yapılarda sıklık tespiti - Web site şema tasarımı	-Kategorizasyon -Kümeleme	- Web Sitesi Adaptasyon, Yapı ve Yönetimi - Pazarlama - Kullanıcı modelleri

Tablo 2-3 Web Madenciliği Genel İçerik, Metot ve Uygulama Kategorileri (Kosala & Blockeel, 2000)

Web madenciliği veri madenciliği ile ortak noktalara sahiptir. Ancak en farklı yaklaşım veri toplama sürecinde gerçekleşmektedir. Geleneksel veri madenciliği yöntemlerinde veri önceden veri ambarlarına toplanmış ve kullanıma hazırdır ancak web madenciliğinde veri toplama işlemi süreçle birlikte işlemelidir. Özellikle çok sayıda web sayfası ile işlem yapan web yapı madenciliği ile web içerik madenciliği metotlarında veri toplama işlemleri anlık gerçekleşebilmektedir.

Web madenciliğinde en popüler algoritmalar web yapı madenciliğinde Pagerank ve HITS, web içerik madenciliğinde ise Web Crawling dir.

2.7.1 Pagerank

1998 yılının nisan ayında gerçekleştirilen "Seventh International World Wide Web Conference" konferansında Google'ın kurucuları Sergey Brin ve Larry Page tarafından geliştirilen Pagerank algoritmasının tanıtılmasıyla web madenciliğinin en önemli günlerinden biri olmuştur.

Pagerank her sayfanın arama sorgularından bağımsız olarak çevrim-dışı hesaplanması ve sabit bir sıralamanın oluşturulduğu algoritmadır. Sosyal ağlar üzerinde en prestijli değerleme metodu olarak belirlendiğinden, günümüzde web sitelerinin değerini gösteren önemli bir kısıt olarak belirlenmiştir. Çalışma mantığı şu şekilde açıklanmaktadır (Liu, 2007, s. 246):

1. Bir sayfanın başka bir sayfaya verdiği link üstü kapalı bir şekilde ilgili bir bilgi verdiğini gösterir. Link alan sayfa aldığı link sayısına ve diğer etkenlere göre algoritma içinde yerini alır.
2. Link alan sayfaların bu linklerin hangi sayfalardan geldiği de algoritma için önemli bir unsurdur. Eğer yüksek prestijli (pagerank değerli) bir sayfadan link alıyorsa bu kendi pagerank değeri için de önemli bir etken olacaktır.

Pagerank algoritmasının en büyük avantajlarından biri web ortamındaki spam adı verilen uygun olmayan sahte içerikleri bertaraf edebilmesidir. Web sitesi sahiplerinin önemli sayfalardan link çıkışı sağlamaları çok kolay olmadığından algoritma sahte ve uygunsuz davranışlardan çok etkilenmemektedir. Başka bir avantajı ise küresel bir ölçüm olarak kabul edilmesi ve kullanıcı sorgularına bağımlı olmamasıdır. Sorgulama unsurlarına dayalı olmadan çevrim-dışı hesaplanmaktadır.

2.7.2 HITS

HITS (Hypertext Induced Topic Search) pagerank gibi statik bir sıralama algoritmasıdır. Paegarank'ten farkı ise sorgu-bağımlı olmasıdır. Kullanıcı bir sorgu oluşturduğunda HITS öncelikle alakalı sayfaları getirir ve daha sonra iki sıralama türü oluşturur. Bunlar, Yetki Sıralaması ve Merkez Sıralaması'dır. Yetki sıralaması birçok link alan sayfa üzerinden oluşur. Birden çok link alan bir site belirli bir içerik hakkında iyi ve güvenilir bilgiye sahip olduğu söylenebilir ve böylece insanlar tarafından güvenilir bir sayfa olarak addedilecektir. Merkez sıralaması ise sayfadaki link çıkışlarını göz önüne alır. Sayfanın başka sayfalara verdiği linklerin güvenilirliği, mevcut sayfanın konu ile ilgisini ve kalitesini yansıtabilmektedir. HITS algoritmasında ana fikir iyi merkez noktaları birden çok merkezi noktayla sıkı bir ilişki içinde olmalıdır. Böylece sayfanın içeriği güvenilir ve yararlı olduğu tespit edilebilir.

HITS algoritması *q* sorguları sonucunda aşağıdaki işlemler sonucu sonuçları ortaya çıkarır:

1. Sorgu q arama motoru sistemine gönderilir. Sorgu sonucu *t* tane(HITS için bu değer genelde *t=200* olarak belirlenir) en yüksek sıralamaya sahip sayfalar getirilir. Ortaya çıkarılan veri kök veri seti (*W*) olarak adlandırılır.
2. Sonuç elde edildikten sonra *W* içerisinde bulunan sayfalarda bulunan çıkış linklerine göre yeniden W sonuçları derlenir ve veri seti kümesi genişletilir. Bu daha büyük bir veri seti olan *S* şeklinde adlandırılır. *S* çok fazla sayfa içerebileceğinden, algoritma *W* içerisindeki her sayfayı sadece *k* kadar sayfanın bağlantısını göz önüne almasını sağlayabilir (HITS için bu değer genelde *k=50* olarak belirlenir.) *S* verileri ana küme olarak adlandırılır. Kullanıcılara yansıtılan bulunan ana set verileridir.

2.7.3 Web Crawling

Örümcek, robot, bot olarak da adlandırılan Web indeksleyicileri (web crawlers) web sayfalarını otomatik olarak indiren yazılımlardır. Milyonlarca sunucu içerisinde bulunan milyarlarca sayfa üzerinde sayfalar içerisindeki linkleri kullanarak birbirine bağlanmış bir ağ içerisinde bilgilere ulaşmaktadırlar. Bu işlem çevrim-içi ya da çevrim-dışı birden fazla ziyaretler şeklinde gerçekleşmektedir.

Temel indeksleme algoritması çekirdek sayfalardan başlar ve başka sayfalara geçiş için sayfa içerisindeki linkleri kullanır. Bu işlem belirlenen kriterler doğrulanana kadar devam eder. Ancak bu basit tanım ağ bağlantıları, örümcek/bot engellerini, URL evrensel kurallarını, sayfa ayrıştırmaları ve diğer indeksleme etiklerini açıklamamaktadır (Liu, 2007, s. 274). Google kurucuları Sergey Brin ve Lawrance Page arama motorları için çok yönlü ve kırılgan bir bileşen olduğunu belirtmişlerdir (Brin, 1998).

3. BÖLÜM : DIJITAL KÜTÜPHANELERDE VERI MADENCILIĞI UYGULAMASI

3.1 Araştırmanın Amacı ve Uygulama Alanı

Araştırma, kütüphanelerin dijitalleşme sürecinde oluşturdukları veri yığınlarını sağlıklı bir platform üzerinde anlamlı hale getirerek sonuçların karar vericilere yol göstermesi amacıyla yapılmıştır. Akdeniz Üniversitesi Merkez Kütüphanesi otomasyon sistemi üzerinde yapılan çalışmalar ile kütüphane uzmanları ve karar vericileri için yeni bir kaynak olması amaçlanmış, bu konuda temel ve sistematik bilgiler ortaya konulmuştur. Sürecin henüz çok yeni olmasından dolayı veriler yeterince düzenli olmayıp, temizlenerek doğru sonuca ulaşılmaya çalışılmış, bu yolda çeşitli modelleme teknikleri kullanılmıştır. Özellikle çalışmalarda, yayın ödünç sirkülâsyon bilgileri üzerinde odaklanılmış, kütüphane kullanıcılarının davranışları incelenmiştir.

Çalışmada öncelikle uygulama alanı tanıtılmış, daha sonraki aşamada uygulamada kullanılan verilerin yapıları ve veriler üzerinde gerçekleştirilen dönüştürme işlemleri üzerinde durulmuştur. Oluşturulan veri ambarı ile kümeleme ve birliktelik analizleri gerçekleştirilmiştir. Çalışmada kullanılan veriler kütüphane veritabanının yanı sıra gerekli görülen dış ortamlardan da temin edilmiştir.

Uygulama orta büyüklükte bir veritabanına sahip Akdeniz Üniversitesi Merkez Kütüphanesi'nde gerçekleştirilmiştir. Kütüphanenin dijitalleşme süreci 2001 yılından itibaren başlamış ve Türkiye genelinde yaygın bir biçimde kullanılan YORDAM kütüphane otomasyon yazılımı ile sağlanmıştır. Yazılım, detaylı bir şekilde yayın sirkülâsyon kayıtlarını, web arama sorgularını, kütüphane iç hizmet verilerini, kullanıcı bilgilerini kayıt altına alabilmektedir. Kullanılan otomasyon yazılımı Mac OS ve Windows işletim sistemleri üzerinde çalışabilen Filemaker veritabanı sistemini kullanmaktadır (Filemaker Inc., 2010). Kullanılan veritabanı sistemi küçük ve orta büyüklükte işletme/organizasyon ihtiyaçlarını, kullanıcılar için formlar ve şablonlar sağlayarak hızlı ve etkili veritabanları oluşturma olanağı sağlamaktadır. Filemaker yazılımının bazı özelliklerinin kısıtlı olması nedeniyle, temizleme ve dönüştürme işlemleri birden fazla program kullanılarak gerçekleştirilmiştir. Veritabanı dört

tablo ve bu tablolar içersinde bulunan yaklaşık 1.200.000'den fazla kayıttan oluşmaktadır. Veri madenciliği çalışması esnasında aşağıda belirtilen tabloların kullanılması uygun görülmüştür. Tablolardaki veriler herhangi bir işleme tutulmamış şekliyle yansıtılmıştır.

Tablo 3-1 Veri Madenciliği sürecinde kullanılan tablolar

Tablo Adı	Alan	Kayıt Sayısı
Kitaplar	42	100.126
Üye	45	50.167
Ödünç	14	157.225

Tablolarda yazılımdan kaynaklanan çok fazla birbirini tekrarlayan alanlar mevcuttur. Örneğin "ad" ve " soyadı" alanlarının bulunmasına rağmen ek olarak "ad_soyad" alanı da oluşturulmuştur. Kitaplar tablosunda yayınlar hakkındaki yazar, tür, yıl gibi kitaplara özgü bilgiler bulunmasının yanında, aynı tablo süreli yayınlar, tezler gibi eserleri de kapsamaktadır. Üye tablosu içerisinde kütüphaneden en az bir kere yayın sirkülâsyonunda bulunan öğrenci, akademik ya da idari personelin kayıtları bulunmaktadır. Bu kayıtlar kütüphane içi sirkülâsyon bilgilerini içermekle birlikte, kişiye özgü kimlik ve demografik bilgileri de kapsamaktadır. Ödünç tablosu ise kütüphane içi sirkülâsyon bilgilerini içermektedir. Tablo kişilerin üye numaralarına göre aldığı yayınların ödünç tarihi ve zamanı ile iade tarihi ve zamanlarını içermektedir. Diğer tablolar da olduğu gibi otomasyon yazılımı fazladan değişkenler oluşturması nedeniyle anlamsız alanlar bu tabloda da bulunmaktadır.

Filemaker veritabanı yazılımı kısıtlı seçeneklere sahip olması nedeniyle öncelikli veri temizleme işlemleri MySQL veritabanı yazılımı sistemi üzerinde yapılmasına karar verilmiştir. CentOS 5.0 işletim sistemi kurulu sunucu üzerinde MySQL veritabanı yazılımına aktarım gerçekleştirilmiştir. Filemaker veritabanını MySQL veritabanına aktarımı sırasında FmProMigrator yardımcı yazılımına başvurulmuştur. İşlem esnasında, belirlenen tabloların uygun alanlara aktarılmasında herhangi bir sınırlama veya sorun ile karşılaşılmadan gerçekleştirilmiştir.

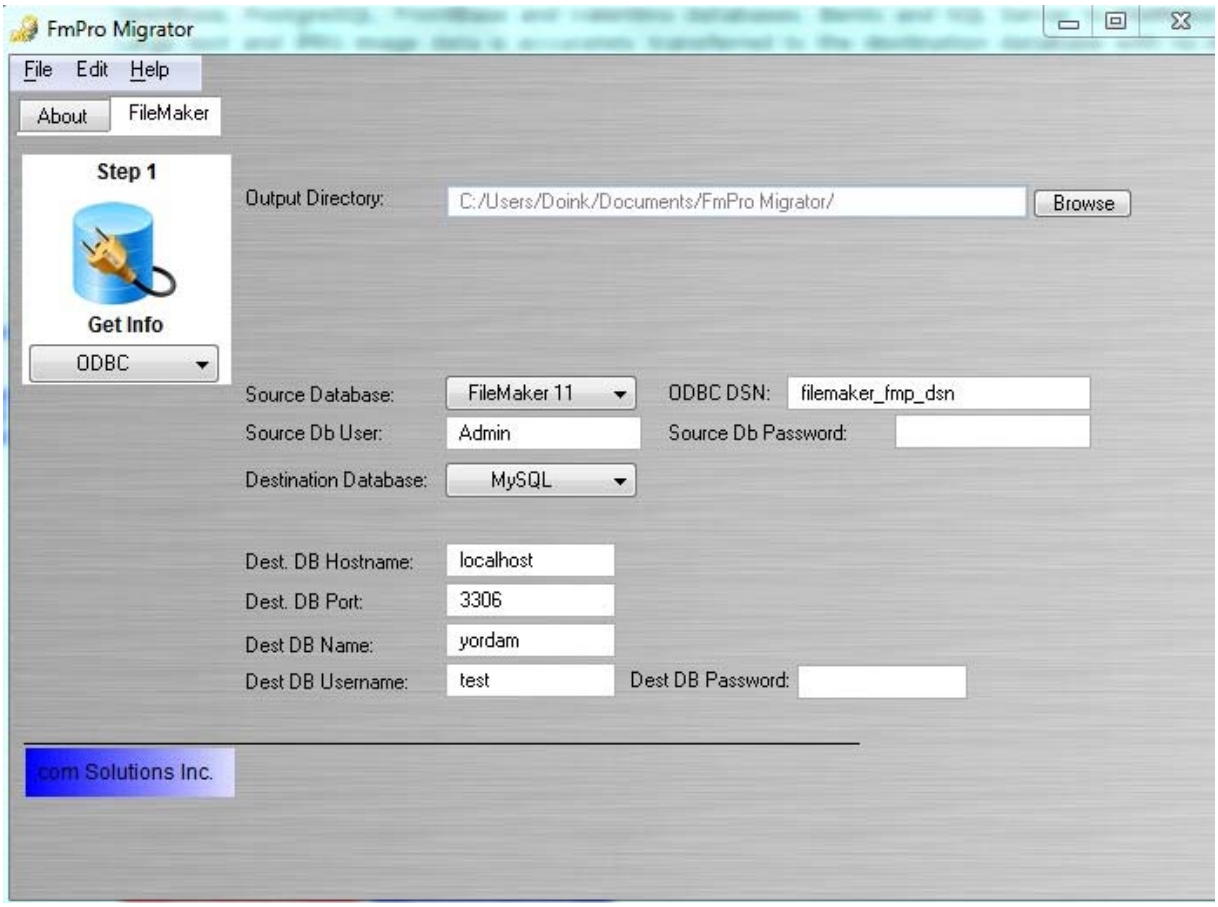
3.2 Uygulamada Kullanılan Yazılımlar

Tez çalışmasında her konu için özelleştirilmiş farklı yazılımlar tercih edilmeye çalışılmıştır. Böylece analizlerin verimliliği arttırılmış, zaman kaybına yol açacak işlemlerden kaçınılmıştır. Aşağıda çalışma esnasında kullanılan yazılımların listesi verilmiştir.

- Masaüstü İşletim Sistemi: Windows 7
- Uzak Sunucu İşletim Sistemi: Centos 5.0 64x
- Ofis Yazılımı : OpenOffice 3.1
- Veritabanı Yazılımları : Filemaker Pro 11 Advanced, MySQL 5.0.6
- Veritabanı Dönüşüm Araçları: FmPro Migrator, SQLyog
- Php Kabuk Program Geliştirme Araçları: Zend Studio 7.0
- Veri Madenciliği Analizi Yazılımları: Clementine 12.0

Veritabanı, kütüphane sunucusu üzerinde bulunan Filemaker veritabanında bulundurulduğundan ilk adımda bu veritabanı alınarak MySQL veritabanına aktarılması planlanmıştır. MySQL veritabanı sistemlerinin esnek ve hızlı olması analizlerin ve sorguların gerçekleştirilmesinde avantaj sağlamaktadır (Wikipedia, MySQL, 2010). Filemaker veritabanları sınırlı bir veri saklama kapasitesine sahip olup MySQL veritabanlarının sağladığı esnekliği gösterememektedir. MySQL Unix işletim sistemlerinde daha verimli çalıştığı için uygulama alanı olarak Unix tabanlı Centos işletim sistemi seçilmiştir.

Filemaker veritabanını MySQL veritabanına dönüştürmek için FmPro Migrator yazılımı kullanılmıştır. FmPro Migrator Yazılımı Filemaker veritabanlarını MySQL, Orcale, SQLServer, Sybase, DB2, PostgreSQL veritabanlarını tüm veritabanı yapısı ve verilerle birlikte hızlı ve doğru bir şekilde dönüştürebilmektedir. Böylece Kütüphane veritabanı yerel sunucuya eksiksiz olarak aktarılmıştır.



Şekil 3-1 FmPro Migrator Yazılım Arayüzü

MySQL veritabanı üzerindeki işlemler ise kullanışlı bir veritabanı yönetim aracı olan SQLyog yazılımı ile gerçekleştirilmiştir. Yazılım, gelişmiş sorgulama editörü ve anlık yedekleme sistemi sayesinde veri temizleme ve dönüştürme işlemlerini başarıyla gerçekleştirilmesinde yararlı olmuştur.

Veri dönüştürme ve temizleme işlemleri sadece SQL sorguları kullanılarak gerçekleştirilemeyecek kadar karmaşık ve düzensiz bir durumdadır. Bu nedenle üçüncü parti yazılımlarla gerekli düzeltmeler otomatik bir şekilde gerçekleştirilmiştir. Bu dönüşümler için ihtiyaca yönelik küçük yazılım paketleri oluşturulmuştur. Yazılım paketleri (kabuk programlama) PHP web programlama dilli kullanılmıştır.

Veri madenciliği uygulaması için ana yazılım olarak SPSS Clementine yazılımı seçilmiştir. Son sürümü ile birlikte IBM SPSS Modeler adını alan yazılım, ön tanımlı algoritmalar sayesinde sezgisel ve kolay bir şekilde modeller oluşturabilmektedir. Modeller analiz sonuçlarına göre interaktif bir şekilde görselleştirilebilme imkânı vermektedir. Öte

yandan otomatik veri hazırlama ve modelleme özellikleriyle sonuçlara hızla ulaşılmasını sağlayabilmekte, kolay bir şekilde veri ambarları yaratılabilmektedir. IBM SPSS Statistics yazılımıyla uyumlu bir şekilde çalışarak istatistiksel analizleri de ortaya koymaktadır (SPSS, 2010).

3.3 Uygulama Süreci

3.3.1 Veri Dönüştürme ve Hazırlama

Kütüphane veritabanı çok büyük olmamasına rağmen oldukça karmaşık ve kirli veriye sahiptir. Kirli verinin büyük kısmı kütüphane çalışanlarının veritabanı üzerinde yazma yetkisine sahip olmalarına rağmen silme yetkilerinin olmamasından kaynaklanmaktadır. Bu nedenle oluşan çift ve hatalı kayıtlar veritabanından çıkarılmış ya da düzeltilmiştir.

Kitaplar, üye ve sirkülâsyon tablolarında veri madenciliğinde kullanılması düşünülmeyen alanlar çıkarılmıştır. Üye tablosunda isim, soy isim, yaş, doğum tarihi gibi kişisel bilgiler, gizlilik kuralları gereği veri ambarına dahil edilmemiştir. Veritabanı içerisinde kitaplar tablosunun oluşturulmasında otomasyon yazılımında meydana gelen hatalar nedeniyle bir miktar yayın bilgisinde hata oluşmuş ve kitap adları SYS ile başlayan anlamsız verilere dönüşmüştür. Bu verilerde ayıklanarak tablodan çıkartılmıştır.

Üye tablosu içerisinde birincil anahtar üye_no alanı olarak belirlenmiş, öğrenciler, akademik ve idari personel numaraları ve kimlik bilgileri gizlenerek veri madenciliğinde kullanılmamıştır. Cinsiyet bilgileri, üniversite otomasyon ve kütüphane otomasyon sisteminde oluşan aktarım problemleri nedeniyle bazı kullanıcılar için eksik kalmıştır. Bu eksiklik hazırlanan bir web yazılımı ile Türk Dil Kurumu'nun web sitesinde bulunan kişi adları sözlüğü (http://www.tdk.gov.tr/TR/Genel/AdArama.aspx) veritabanındaki isimlerle karşılaştırılarak, isimlere göre cinsiyetler belirlenmiş, erkek ve kız ortak isimleri için "belirsiz" bilgisi girilmiştir. Tabloda kütüphane çalışanları tarafından hatalı girilen kullanıcı bilgileri de mevcuttur. Bu hatalı bilgiler çoğu kez çalışanlar tarafından tespit edilmiş ancak silme yetkileri bulunmadığı için kayıt bölümünde kaydın hatalı olduğunu belirten bir açıklama eklemişlerdir. Hatalı kayıtların ortak noktası çalışanlarca bırakılan notlarda "hatalı" veya "yanlış" kelimelerinin içermesidir. Bu nedenle bu kelimelerin içerdiği kayıtlar SQLyog yardımıyla oluşturulan sorgularla tablodan çıkartılmıştır.

Sirkülâsyon tablosu diğer tablolardan farklı olarak işlemsel verileri içermektedir. Veritabanında en fazla yeri kaplayan bu yığın verilerin temizlemesinde dönüşüm aşaması oldukça uzun sürmüştür. Tabloda tarih ve zaman bilgilerinin yanında sirkülâsyon tablosu demirbaş ve üye numaraları hakkında bilgiler yer almaktadır.

Yukarıda belirtilen temizleme işlemi SQL sorguları kullanarak gerçekleştirilmiştir. Ancak SPSS Celementine yazılımı aracılığıyla veri ambarı oluşturulurken, ihtiyaca göre dönüştürme işlemleri tüm veri madenciliği sürecince devam etmektedir. Yukarıdaki veri temizleme işlemleri sonucunda nihai olarak tabloların durumu aşağıdaki şekle dönüşmüştür.

Tablo 3-2 Veri dönüştürme ve temizleme işemleri sonucu veri madencilğinde kullanılacak tabloların durumu

Tablo Adı	Alan Sayısı	Kayıt Sayısı
Kitaplar	4	92336
Üye	8	50167
Ödünç	4	157274

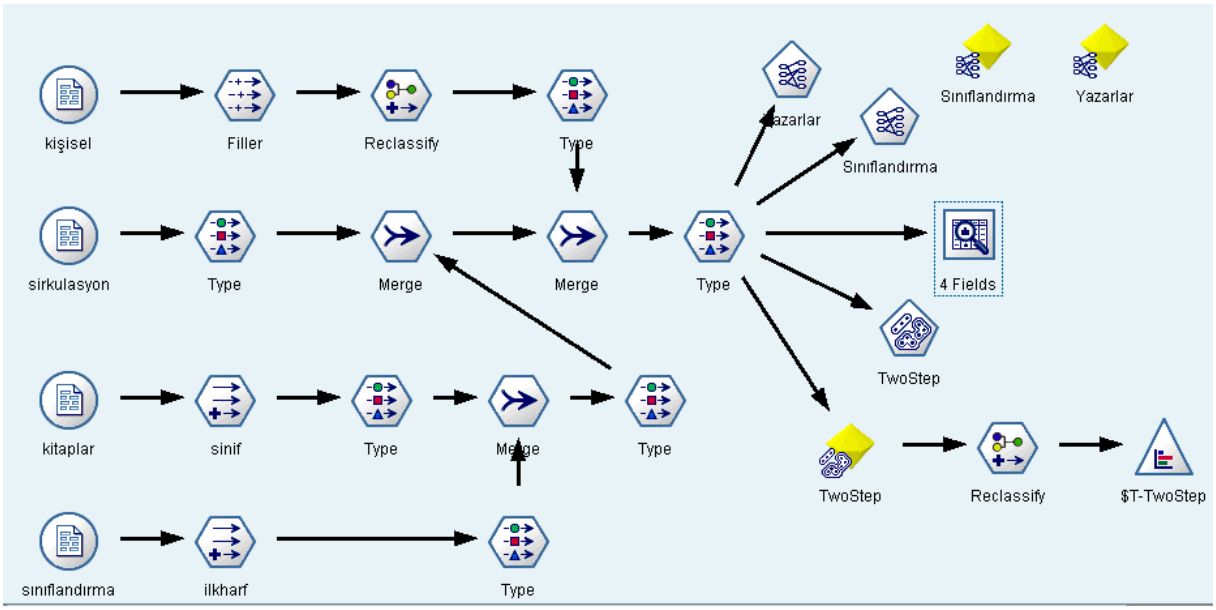
3.3.2 Veri Ambarının Oluşturulması

Veri madenciliğinde modellemelerin oluşturulması ve kullanılması için planlanan tabloların dönüştürülmesi işlemi SPSS Clementine yazılımı üzerinde gerçekleştirilmiştir. SPSS Clementine'in diğer veritabanı sistemleri ile bağlantı sağlayabilmesi veri aktarım işlemlerini kolaylaştırmaktadır. Yazılımın diğer veritabanı sistemleri ile bağlantı sağlayabilmesinin yanında metin, SPSS, SAS, Excel dosyalarını okuyabilme ve sanal olarak bir veritabanı oluşturabilme olanağı da sağlamaktadır. Bu durum Veri temizleme işlemleri için de esneklik sağlamaktadır. Tablolar içerisinde filtreleme, yeni sanal alanlar oluşturma, çoklu sonuçlarını birleştirme, alan değerlerini değiştirme, değerleri kategorize etme gibi veri temizleme ve dönüştürme işlemlerini başarıyla gerçekleştirebilmektedir.

SPSS Clementine çok farklı modellemeleri karmaşık yapısından arındırarak basit bir ara yüzle kullanımını sağlamaktadır. Yazılım, 12.0 sürümü ile birlikte otomatikleştirilmiş ikincil sınıflandırıcı, sayısal tahmin edici ve zaman serileri, sınıflandırma modellemelerinde C&R

ağacı, QUEST, CHAID, Karar Listeleri, Regresyon ve Faktör analizleri, birliktelik modellemelerinde GRI, Apriori, Carma, Sequence algoritmaları, kümeleme modellemelerinde ise K-means, Kohonen, TwoStep ve Anomaly algoritmaları kullanılabilmektedir.

Tez çalışmasında ön veri temizleme işlemleri MySQL'de gerçekleştirildiği için veriler MySQL üzerinden tablolara aktarılarak gerçekleştirilmiştir. Aşağıda SPSS Clementine üzerinde oluşturulan veri ambarı yapısı gösterilmektedir.



Şekil 3-2 SPSS Clementine üzerinde oluşturulan veri ambarı görüntüsü

Veri ambarı oluşturulurken 4 tablo dikkate alınmıştır. Bu tablolar kişisel, sirkülâsyon, kitaplar ve sınıflandırma tablolarıdır. Kişisel, sirkülâsyon ve kitaplar tablosu veritabanından elde edilebilirken, sınıflandırma tablosu uluslararası sınıflandırma sistemi olan "LC Classification System" kodlarını içermektedir. Kodlar iki harften oluşmaktadır. İlk harf genel bir kategoriyi simgelerken ikinci harf bir alt kategoriyi temsil etmektedir. Bu yüzden "Derive" özelliği kullanılarak "startstring(1, sinif)" formülü ile genel kategoriyi temsil eden harf belirlenmiştir. Yapılan düzenlemenin ardından tablo kitaplar tablosu ile birleştirilmiştir.

Kitaplar tablosu kütüphane içinde bulunan yayın bilgilerinin toplu bir şekilde bulunduğu tablodur. Veri ambarı oluşturulurken sınıflandırma sistemi yeniden yapılandırılmıştır. Her kitap için benzersiz (unique) bir kod atayan sistem, yukarıdaki sınıflandırma sistemini dikkate alarak gerçekleştirmiştir. Analizlerde kullanılması planlanan sınıflandırma öğeleri için

benzersiz sınıflandırma kodları üzerinde "Derive" özelliği kullanılarak "startstring(2, siniflandirma)" ilk iki harfi elde edilmiştir. Böylece sınıflandırma ve kitaplar tablosunu birleşimi sağlanmış sınıflandırma kodlarının açıklamaları tabloya eklenmiştir.

Sirkülâsyon tablosu kütüphane içerisinde işlemsel kayıtların tutulduğu veriyi barındırmaktadır. Ocak 2005 - Aralık 2009 tarihleri arası kullanıcı işlemlerinin ayrıntılı olarak tutulduğu tablo veri ambarına dahil edilmeden önce oldukça kirli ve düzensiz bir veriye sahipti. Ancak oluşturulan kabuk programlama yazılımları ile tekrarlanan düzensizlikler ve hatalar tespit edilmiş ve SQL sorguları ile hatalar giderilmiş veya veritabanından çıkarılmıştır. Veri ambarına dahil edilmesi ile birlikte veri yapıları düzeltilmiş ve diğer tablolardan elde edilen verilerle birleştirilmeye hazır hale getirilmiştir.

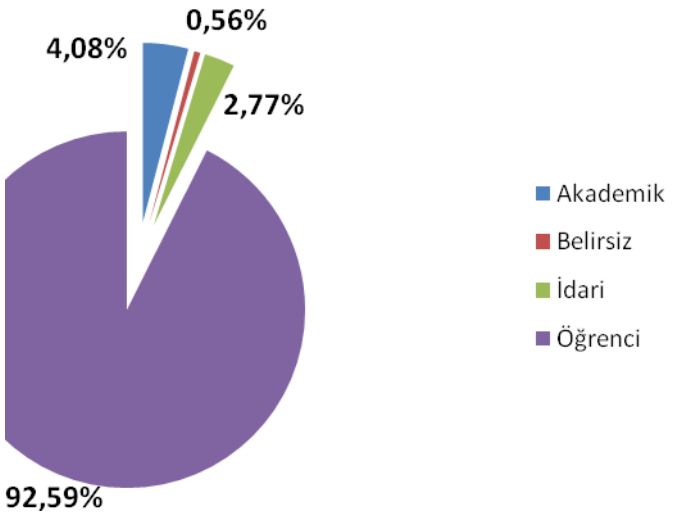
Kişisel tablosu ise kullanıcı bilgilerinin depolandığı tablodur. Daha önce belirtildiği gibi kullanıcı gizliliğine sadık kalarak kişisel bilgiler bu çalışmada tablolara yansıtılmamıştır. Veri ambarında tabloya sadece analiz için gerekli bilgiler aktarılmıştır. Veri ambarında bazı alanların gruplanmasını gerektiren analizlerde "Filler" özelliği ile veriler düzenlenmiş "reclassify" özelliği ile de yeni değerler atanmıştır. Elde edilen veriler diğer tablolarda olduğu gibi sirkülâsyon tablosu ile birleştirilmiştir.

3.4 Tanımlayıcı Bulgular

3.4.1 Kütüphane Kullanıcı İstatistikleri

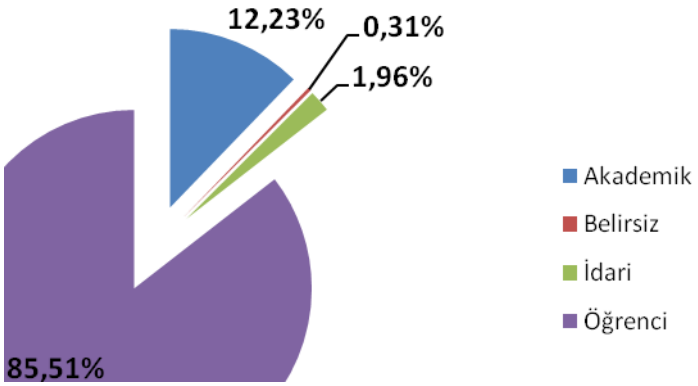
Kütüphane veritabanının dönüştürülmesi ile birlikte veri madenciliği yazılımı üzerinden frekans tekniği kullanılarak bazı bulgular grafikleştirilmiştir. Bu bölümde gerekli yerlerde Clementine yazılımında bulunan kayıt ve alan seçenekleri kullanılarak geçici alanlar oluşturulmuş, değişkenler atanmıştır.

Şekil 3-3'de kütüphaneden en az bir kere ödünç yayın almış ya da kütüphaneden yararlanmış kişilerin kullanıcı gruplarına göre dağılımını inceleyebilmekteyiz.



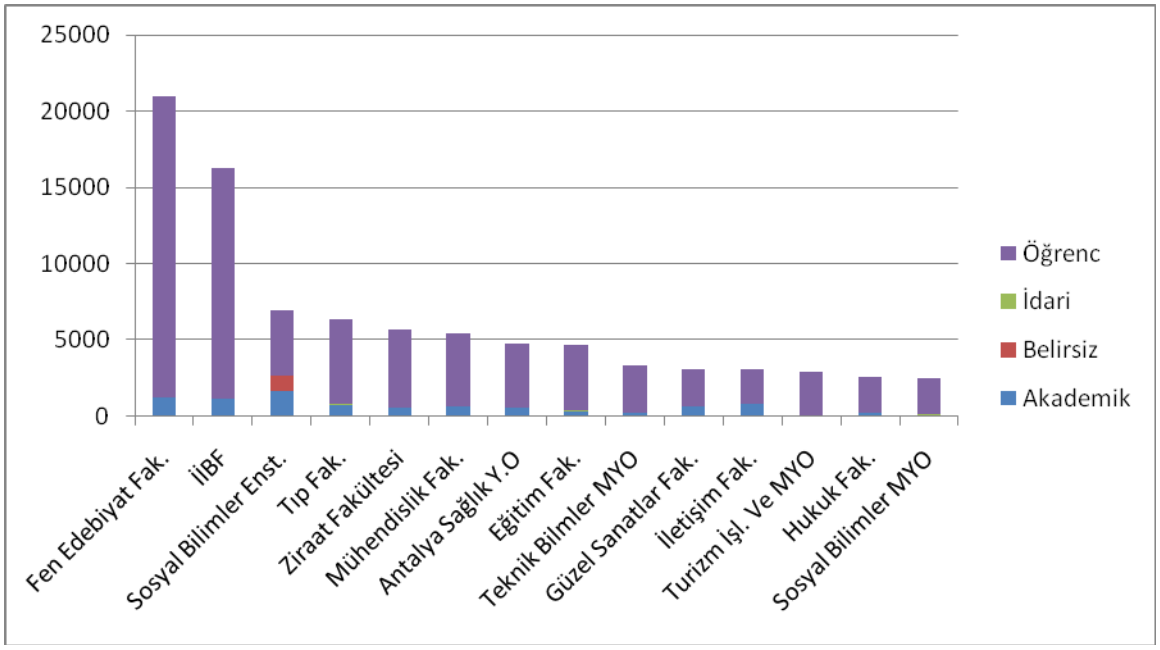
Şekil 3-3 Kullanıcı Gruplarının Dağılımı

Şekil 3-4'de kitap sirkülâsyonu içerisinde bulunan kişilerin üniversite içerisindeki görevlerine (öğrenci, personel, akademisyen, üniversite dışı, vs.) göre dağılımını göstermektedir. Gruplar veritabanında sayısal olarak tutulmakta olduğu için "Reclassify" özelliği ile etiketlendirilmiştir. En az bir kere kitap almış kullanıcılara dayanarak oluşturulan grafikte toplam sirkülâsyonun %92,59 oranına sahip olan 46035 öğrenci grubu 84106 kez, %4,08 oranına denk gelen 2044 akademik personel grubu 12026 kez ve tüm kullanıcılar içerisinde %2,77'sine kapsayan 1391 idari personel 1928 kez kütüphaneden ödünç kitap almıştır.



Şekil 3-4 Sirkülasyon Bilgilerine Göre Kullanıcı Grupları Dağılımı

Toplam sirkülâsyon verilerine dayanarak tekrar elde edilen grafikte akademisyen grubunun kullanıcı başına 5,88 yayın alması dağılımda akademisyen grubu lehine değişimlere neden olmuştur. Bu sayılar öğrenci grubu için ortalama 1,81 ve idari grup içinse 1,06'dır.



Şekil 3-5 Fakültelere Göre Kullanıcı Gruplarının Dağılımı

Yukarıda tabloda en çok kitap sirkülâsyonunda bulunan 15 fakültenin üniversite içi gruplara göre sayısal verileri tablolanmıştır. En fazla ödünç yayın alan fakülte olan Fen Edebiyat Fakültesi 20948 kez yayın sirkülasyonuna dahil olmuştur. Bu yayın sirkülâsyonunun sadece 1128 tanesi akademisyenler tarafından gerçekleştirilmiştir. İktisadi ve İdari Bilimler Fakültesi kullanıcıları ise kütüphaneden 15145 ödünç yayın almışladır. 1091 tanesi akademisyenlerden 23 tanesi ise idari personeldendir. Sosyal Bilimler Enstitüsünde ise 3 grup açık bir şekilde belli olmaktadır. Kütüphane otomasyonunun doktora öğrencilerini belirsiz olarak tanımlaması nedeniyle grafikte bu şekilde gösterilmiş ancak veri uyarlama ve temizleme aşamasında bu hata giderilmiştir. Bu grafikte bir başka hata ise akademik personel ile ilgilidir. Sosyal bilimler enstitüsü akademik personeli olarak yüksek lisans ya da doktora yapan araştırma görevlileri dahil edilmiş, İktisadi ve idari bilimler Fakültesi akademisyenleri dahil edilmemiştir.

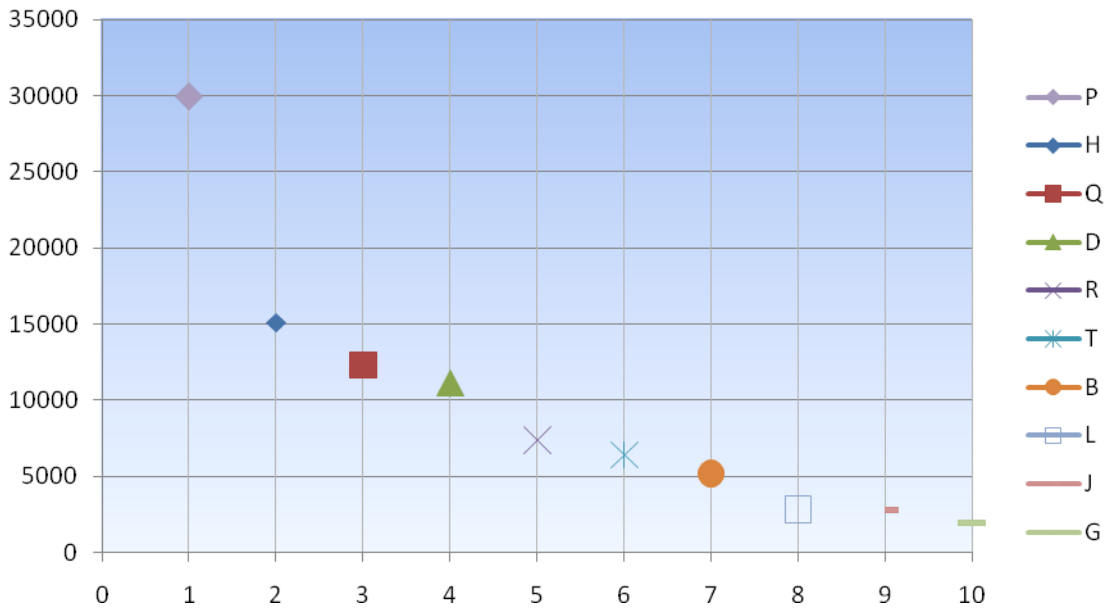
Kütüphane içi kitap sınıflandırma sistemi olarak uluslararası bir standat olan LC Classification standartı kullanılmaktadır. 1897 yılında Herbert Purham tarafından geliştirilen

bu sınıfandırrma Araştırma ve akademik kütüphanelerde yaygın bir biçimde kullanılan bu sınıflandırma sistemi birçok ülkede standartlaştırılmıştır (Library of Congress). Akdeniz Üniversitesi Merkez Kütüphanesi de LC sınıflandırma sistemine göre kütüphane düzenini oluşturmuştur. Bu oluşumda kodlama aşağıdaki gibi belirlenmiştir.

Tablo 3-3 Kütüphane İçi Yayın Sınıflandırması

A Genel yapıtlar	M **Müzik**
B Felsefe. Psikoloji. Din	**N Güzel sanatlar**
C Tarihe yardımcı bilimler	**P Dil ve edebiyat**
D Tarih	**Q Bilim**
E-F Amerika tarihi	**R Tıp**
G Coğrafya. Antropoloji. Turizm	**S Tarım**
H Sosyal bilimler	**T Teknoloji**
J Siyaset bilimi	**U Askerlik**
K Hukuk	**V Denizcilik**
L Eğitim	**Z Kaynakçalar. Kütüphanecilik**

Kütüphane verilerine göre kütüphanede bulunan kitaplar bir sınıflandırmaya dahil edilmiş ancak kategorize edilmemiştir. Oluşturulan veri ambarı sayesinde kitapların genel kategorik bilgileri ve sirkülâsyon dağılımı aşağıdadır.



Şekil 3-6 Yayın Sınıflandırmasının Kütüphane Yayınları Arasındaki Dağılımı

Şekil 3-6'da kullanıcıların kütüphaneden en fazla ödünç aldıkları 10 yayın grubu kıyaslanmıştır. Belirlenen 10 yayın grubu tüm yayınlar içerisinde %95'lik kısmı oluşturmaktadır. Grafikten de anlaşıldığı üzere P grubunun temsil ettiği Türk Dili ve Edebiyatı bölümü kütüphane içerisinde 29960 ile en fazla yayını bulundurmaktadır. Bu sayı tüm eserler arasında %30,03'e denk gelmektedir. H grubu Sosyal Bilimler Eserleri 15099 yayınla kütüphanede bulunan en fazla ikinci yayın grubudur. H grubu tüm yayınlar içerisinde % 15 oranında bulunmaktadır. Q (Bilim), D (Tarih), R(Tıp), T (Teknoloji), B (Felsefe, Psikoloji, Din), L (Eğitim), J (Siyaset Bilimi) ve G (Coğrafya. Antropoloji. Turizm) yayın grubunun tüm yayınlar içerisindeki yüzdesi sırayla %12,33, %11,15, %7,37, %6,42, %5,23, %2,86, %2,80 ve %1,92'dir.

3.5 Sirkülâsyon Verileri Üzerine Birliktelik Analizleri

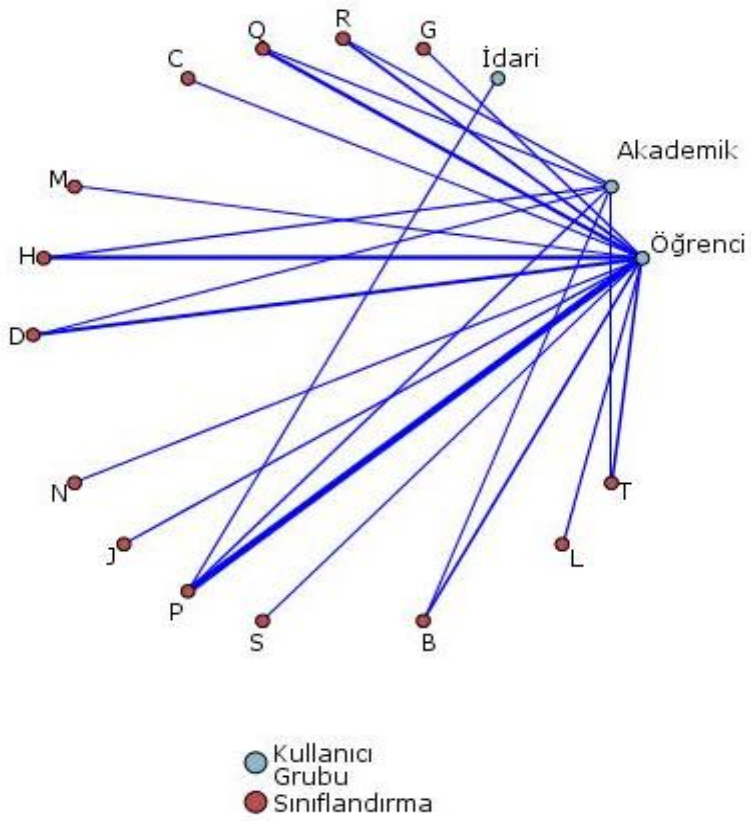
Bu çalışmada kütüphane içerisinde kitap sirkülâsyon, kitaplar ve kullanıcı bilgileri arasında mevcut olabilecek birliktelik kurallarının tespiti gerçekleştirilmiştir. Birliktelik analizi Apriori algoritması göz önüne alınarak oluşturulmuştur.

Birliktelik analizinde kullanılacak veriler daha önce oluşturulan veri ambarı üzerinden modellenmiştir. Veri ambarında ödünç, kişisel, kitaplar ve daha sonradan yardımcı tablo olarak oluşturulan sınıflandırmalar tablosu bulunmaktadır.

Tablolarda kullanılan grup, fakülte gibi bazı veriler kategoriksel alanlar sayısal olarak tutulması nedeniyle Clementine yazılımı için düzenlenmiştir.

Yapılan birliktelik analizindeki amaç, kütüphane içi sirkülâsyon bilgilerine göre kullanıcıların en çok hangi kitapları, dolayısıyla da hangi kitap türlerini/sınıflandırmalarını tercih ettiklerinin tespiti yapılmıştır. Bu analizle amaç kullanıcıların kitap seçiminde tercih ettikleri türler arasındaki ilişkiyi belirlemektir.

Şekil 3-7'de Kütüphane kullanıcıları içerisindeki kullanıcı gruplarından Akademik, Öğrenci ve İdari grubun kütüphanede bulunan yayınların kategorilerine göre belirlenmiş ana gruplara göre dağılımı ağsal grafikle tespit edilmiştir. Kuvvetli ilişkilerin koyu renkle ifade edildiği grafikte öğrenci grubu içerisindeki en büyük bağlantı ağı Türk dili ve Edebiyatı eserleri dahilindedir. %26,61 oranı ile tüm bağlantılar arasında önemli bir yer tutmaktadır. Bağlantıların %12,38'i "H" harfiyle temsil edilen Sosyal bilimler eserleri ile ilgilidir. Ardından %10,3 ile "Q" harfiyle bilimsel yayınlar ve %9,65 ile "D" Tarih eserleri gelmektedir



Şekil 3-7 Kütüphane İçi Yayın Sınıflandırmasının Kullanıcı Gruplarına Göre Ağsal Grafiği

Akademik grup içerisinde ise dağılım neredeyse homojendir. Dil ve Edebiyat (P), Sosyal Bilimler (H) , Bilimsel Yayınlar (Q), Tıp yayınları (R) bağlantıların akademik grup içerisindeki dağılımın büyük kısmını oluşturmaktadır. Toplamda ise Dil ve Edebiyat yayınları %2,28, Sosyal bilimler yayınları %2,23, Bilimsel yayınlar %1,65, Tıp yayınları ise %1,57 oranındadır. İdari kullanıcılar arasında kuvvetli bir bağlantı sadece dil ve edebiyat yayınları arasında görülmektedir. Bu bağlantı diğer tüm bağlantılar içerisinde %0,86 oranındadır.

Kullanıcı verileri ve yayın sınıflandırmaları arasında uygulanan analiz için kullanılan Apriori algoritması için minimum öncül destek % 0,1, minimum güven kuralı ise % 10 olarak belirlenmiştir. Maksimum öncül öğe sayısının ise 3 olması sağlanmıştır. Verilerin işlemsel kayıtlardan oluşması nedeniyle değerler düşük tutulmaya çalışılmıştır. Kuralı uygulamak için belirlenen 99778 sirkülâsyon verisinin yanında 20 farklı sınıflandırmaya ait 54706 kitap analize dahil edilmiştir. Analiz sonucunda ise 42072 kural ve 11005 uygun işlemsel kayıt tespit edilmiştir. Öncül destek ve güven kural değerleri, düşük sayıda bir işlemsel kayıt için

daha yüksek değerler belirlenebilirken, orta büyüklükteki bir veri yığını için bu değerlerin uygun olduğu düşünülmektedir. Analiz sonucu maksimum öncül öğe değeri destek aralığı %1.009 ile %43.507 arasında bulunmuş, güven aralığı ise minimum güven kuralı olarak belirlenen %10 ile %97,744 arasında tanımlanmıştır. İşlemsel verilerin kullanıldığı analizde ID olarak sirkülâsyon verilerinde birincil anahtar olarak bulunan üye numaraları, içerik olarak ise sınıflandırma kodları kullanılmıştır.

Uygulama sonucu oluşan model aşağıdaki tablodadır. Tablo'da kitap sınıflandırmalarına göre hangi kitap grupları arasında ilişki olduğu tespit edilmiştir. Buna göre öncül öğe grubuna ait kitapları alan kullanıcıların, ardıl öğe grubundaki kitapları tercih etmesi durumu yansıtılmaya çalışılmıştır.

Tablo 3-4 Kütüphane İçi Sınıflandırmaya Göre Birliktelik Analizi

Consequent (Öncül)	**Antecedent (Ardıl)**	**Support (Destek) %**	**Confidence (Güven) %**	**Lift (Kaldırma) %**
Türk Edebiyatı (PL)	Almanca Felemenkçe İskandinav Edebiyatı Fransızca İtalyanca İspanyolca Edebiyatı	1,172	94,574	2,174
	İslam ve Diğer İnanç Hareketleri Fransızca İtalyanca Edebiyatı	1,199	92,424	2,124
	İngiliz Edebiyatı Amerikan Edebiyatı	1,09	91,667	2,107
	Fizik Kuzey Avrupa Edebiyatı	1,081	91,597	2,105
	Almanca Felemenkçe İskandinav Edebiyatı Kuzey Avrupa Edebiyatı	1,581	91,379	2,100
	Fizik Amerikan Edebiyatı	1,408	90,968	2,091
	Aile Evlilik Kadın Amerikan Edebiyatı	1,063	90,598	2,082
	Halk Bilimleri	1,836	84,158	2,904
Ticaret	Finans Sosyoloji (Genel)	1,463	64,596	4,426
	Toplum Tarihi Sosyoloji (Genel)	1,336	58,503	4,009
	Sosyoloji (Genel) Ekonomi Teorileri	2,581	58,451	4,005
	Yerel Yönetimler Belediyeler	1,436	56,329	3,860
	Finans	2,553	55,872	3,829
İç Hastalıkları.	Halk Sağlığı	1,1	65,289	4,046

Spor Hekimliği	Fizyoloji			
	Halk Sağlığı Psikoloji	1,418	64,103	2,829
	İnsan Anatomisi Fizyoloji	1,608	58,757	3,641
	Patoloji	1,717	53,439	3,311
Sosyoloji (Genel)	Toplum Tarihi Ticaret	1,199	65,152	2,878
	Toplumlar, Sınıflar, Irklar Balkanlar ve Türkiye Tarihi	1,199	57,576	6,532
	Toplumlar, Sınıflar, Irklar Edebiyat (Genel)	1,018	56,25	6,382
	Politik Teoriler Aile. Evlilik. Kadın	1,09	55,833	6,334
	Toplumlar, Sınıflar, Irklar Doğu Asya ve Afrika Dilleri Edebiyatları	1,272	55,714	6,321
Ekonomi Teorisi	Fiziki Coğrafya	1,181	63,077	5,496
	Finans Sosyoloji (Genel)	1,463	59,006	5,141
	Finans Ticaret	2,799	58,766	5,121
	Finans Matematik	1,508	57,831	5,039
	Toplum Tarihi Ticaret	1,199	55,303	4,819
Toplumlar, Sınıflar, Irklar	Siyaset Bilimi Sosyoloji (Genel)	1,745	22,917	11,11
	Toplum Tarihi	1,199	22,727	11,018
	Siyaset Bilimi (Genel) Sosyoloji (Genel)	1,581	22,414	10,866
	Sosyal Hizmetler Sosyoloji (Genel)	1,236	22,059	10,694
	Uygarlık Tarihi Ekonomi Tarihi	1,236	22,059	10,694
Uluslararası Hukuk	Siyaset Bilimi (Genel) Siyaset Bilimi	1,699	13,904	13,662
	Siyaset Bilimi (Genel) Ekonomik Teoriler	1,699	13,904	13,662
	Avrupa Hukuku Balkanlar ve Türkiye Tarihi	1,136	13,6	13,363
	Siyaset Bilimi (Genel) Edebiyat (Genel)	2,163	11,765	11,56
Doktriner İlahiyat	Toplum Tarihi Ticaret	1,199	11,364	20,501
	Toplum Tarihi Sosyoloji (Genel)	1,336	10,884	19,636

Tabloda öncül öğe (consequent), ardıl öğe (antecedent), destek (support), güven (confidence) ve kaldırma (lift) oranları birlikte verilmiştir. Çalışmada gösterilmeye en uygun sonuçlar tabloya yansıtılmıştır. Belirlenen öncül öğeler, güven oranlarına göre en yüksek değerden en düşük değere doğru sıralanmıştır. En yüksek güven oranına sahip olan Türk Edebiyatı eserlerinin "PL" harfleriyle kategorize edildiği analiz tablosunda, bu eserlerin en fazla ilişkisel yapıya sahip olduğu keşfedilmiştir. Çok fazla yayın sınıfı ile ilişkisi tespit edilen PL kategorisi başlıca "Almanca Felemenkçe İskandinav Edebiyatı" ve "Fransızca İtalyanca İspanyolca Edebiyatı" eserleriyle ilişkilendirilmiştir. Bu kategoriler sonucu elde edilen güven değerinin %94,574 ve %94,424 olması ilişkiye dâhil olan işlemsel kayıtları içerisinde Türk Edebiyatı kategorisinde eser alan kullanıcıların %94'ünden fazlasında geçerli olduğu tespit edilmiştir. Destek değerinin %1,172 olması ise tüm analize dahil edilen kayıtlar arasındaki oranını temsil etmektedir. Kaldıraç oranını gösteren "lift" sütunu ise öncül ve ardıl değerlerinin analiz sonucu oluşturulan ilişkisel durumun kuvvetini göstermektedir. Bu değerin 1'den küçük olması analiz sonucunda kategoriler arasındaki ilişkiyi olumsuz etkilediği, pozitif bir değer taşıması durumunda ise analiz sonucu elde edilen ilişkiyi olumlu yönde desteklediği söylenebilir. Bu değerin artması ilişkinin kuvvetini gösterir. 2,174 (2,124) değeri "Türk Edebiyatı" eselerine ilgi gösterenlerin "Almanca Felemenkçe İskandinav Edebiyatı" ve "Fransızca İtalyanca İspanyolca Edebiyatı" eserlerini de birlikte ödünç aldığı sonucunun kuvvetini göstermektedir.

Analiz tablosunun son sütununda "Doktriner İlahiyat" kategorisine göz attığımızda bu kategoriden kitap alan kullanıcıların "Toplum Tarihi" ve "Ticaret" kategorisinden de kitap aldığı sonucu çıkarılabilmektedir. Bu analizde kategoriler arası ilişkinin tespiti için analize dahil edilen işlemsel kayıtlardan sadece %11,364'ü desteklenmesine rağmen kaldırma oranı oldukça bir yüksek değer sahiptir (20,501). Bu durumda doktriner ilahiyata ilgi duyan kullanıcının Toplum tarihi ve Ticaret konularına da ilgi duyduğu sonucu kuvvetli bir değere sahiptir. Aynı kategori altında kullanıcılar yine Toplum tarihi kategorisi ile birlikte Sosyoloji ve Doğu Asya Dilleri ve Edebiyatları kategorisi eserlerini Doktriner İlahiyat kategorisi kitapları ile birlikte almayı tercih etmektedirler.

Aynı birliktelik analizi kitap yazarları için de yapılmıştır. Analiz öncesi minimum destek değeri %1, minimum kural güven değeri %10 ve maksimum ardıl değeri 2 olarak öngörülmüştür. Veri ambarı üzerinden kitaplar ve sirkülâsyon tablosu kullanılarak Apriori algoritması sonucu 420 kural tespit edilmiş, 11005 sirkülâsyon kaydı dikkate alınmıştır. Minimum destek değeri %1,018 iken maksimum destek değeri %4,044 olarak belirlenmiştir.

Diğer taraftan güven değerleri %10 ile %95,161 arasında gerçekleşmiştir. Sonuç aşağıdaki tabloda gösterilmiştir.

Tablo 3-5 Yayın Sahiplerine Göre Birliktelik Analizi

Consequent (Öncül)	Antecedent (Ardıl)	Support (Destek) %	Confidence (Güven) %	Lift (Kaldırma) %
Dostoyevski, Fyodor Mihaylovic	Gorki, Maksim	1,218	31,343	7,751
Dostoyevski, Fyodor Mihaylovic	Tolstoy, Lev Nikolaevich	1,99	29,68	7,34
Parasiz, M. Ilker	Türkay, Orhan	1,027	26,549	10,861
Farkas, Hershel M.	Ongun, Ipek	1,054	25,862	94,871
Ran, Nazim Hikmet	Ilhan, Attila	1,09	25,833	8,997
Topuz, Hifzi	Mumcu, Uğur	1,018	25	42,327
Kulin, Ayşe	Uzuner, Buket	1,19	24,427	11,488
Yasar Kemal	Orhan Kemal, (Mehmet Raşit Öğütçü)	1,163	24,219	6,331
Tolstoy, Lev Nikolaevich	Gorki, Maksim	1,218	23,134	9,292
Ran, Nazım **Hikmet**	Nesin, Aziz	1,09	22,5	7,836
Birch, Beverley	Mumcu, Uğur	1,018	22,321	81,882
Dökmen, Üstün	Kulin, Ayşe	1,672	22,283	26,087
Yasar Kemal	Zola, Emile	1,072	22,034	5,76
Ran, Nazım **Hikmet**	İlhan, Attila	1,381	21,711	7,561
Yasar Kemal	Nesin, Aziz	1,09	21,667	5,664
Ümit, Ahmet	Ömer Seyfettin	1,154	21,26	10,782
Dostoyevski, Fyodor Mihaylovic	Orhan Kemal, (Mehmet Raşit Öğütçü)	1,163	21,094	5,217
Yasar Kemal	Zola, Emile	1,318	20,69	5,408
Ran, Nazım **Hikmet**	Mumcu, Uğur	1,018	20,536	7,152
Dostoyevski, Fyodor Mihaylovic	London, Jack	1,063	20,513	5,073
King, Stephen	Steel, Danielle	1,027	20,354	9,451
Tolstoy, Lev Nikolaevich	Dostoyevski, Fyodor Mihaylovic	2,917	20,249	8,133
Dostoyevski, Fyodor Mihaylovic	Zola, Emile	1,072	19,492	4,82
Bengisu, Ünal	Steel, Danielle	1,027	19,469	63,017

Auster, Paul	Pamuk, Orhan	1,881	19,324	37,974
Yasar Kemal	Orhan Kemal (Mehmet Rasit Öğütçü)	1,563	19,186	5,015
Kasgarli Mahmut	İlhan, Attila	1,09	19,167	70,31
Yasar Kemal	Uzuner, Buket	1,19	19,084	4,989
Yasar Kemal	Güntekin, Reşat Nuri	1,245	18,978	4,961
Ümit, Ahmet	Ömer Seyfettin	1,581	18,966	9,618
Auster, Paul	Pamuk, Orhan	1,545	18,824	36,992
Kulin, Ayşe	Pamuk, Orhan	1,545	18,824	8,853
Telli, Z. Kazım	Mumcu, Uğur	1,018	18,75	54,301
Mungan, Murathan	Ersöz, Cezmi	1,072	18,644	10,107
Kulin, Ayşe	Ersöz, Cezmi	1,072	18,644	8,768
Tolstoy, Lev Nikolaevich	Zola, Emile	1,072	18,644	7,488
Mungan, Murathan	Ersöz, Cezmi	1,127	18,548	10,055
Kulin, Ayşe	Ersöz, Cezmi	1,127	18,548	8,723
Kulin, Ayşe	Ümit, Ahmet	1,617	18,539	8,719

Tabloda görüldüğü gibi 420 kural arasından 38 tanesi güven değerlerine göre listelenmiştir. Türk ve dünya edebiyatının hakim olduğu kurallara göre en yüksek güven değerine sahip ilk iki kayıtta Fyodor Mihaylovic Dostoyevski eserlerini tercih eden kullanıcılar aynı zamanda Maksim Gorki veya Lev Nikolaevich Tolstoy eserlerini de ödünç almaktadırlar. Kaldırma Oranı (Lift) yüksek olan kurallar arasında bulunan arasında ünlü edebiyat romanlarını çocuklar için uyarlayan yazar Beverley Bitch ile Araştırmacı yazar Uğur Mumcu ödünç kitap sirkülâsyon ilişkisinin analiz sırasında 112 kez tekrar edildiği belirtmekte fayda vardır. Aynı durum Hershel M. Farkas ile İpek Ongun için de geçerlidir.

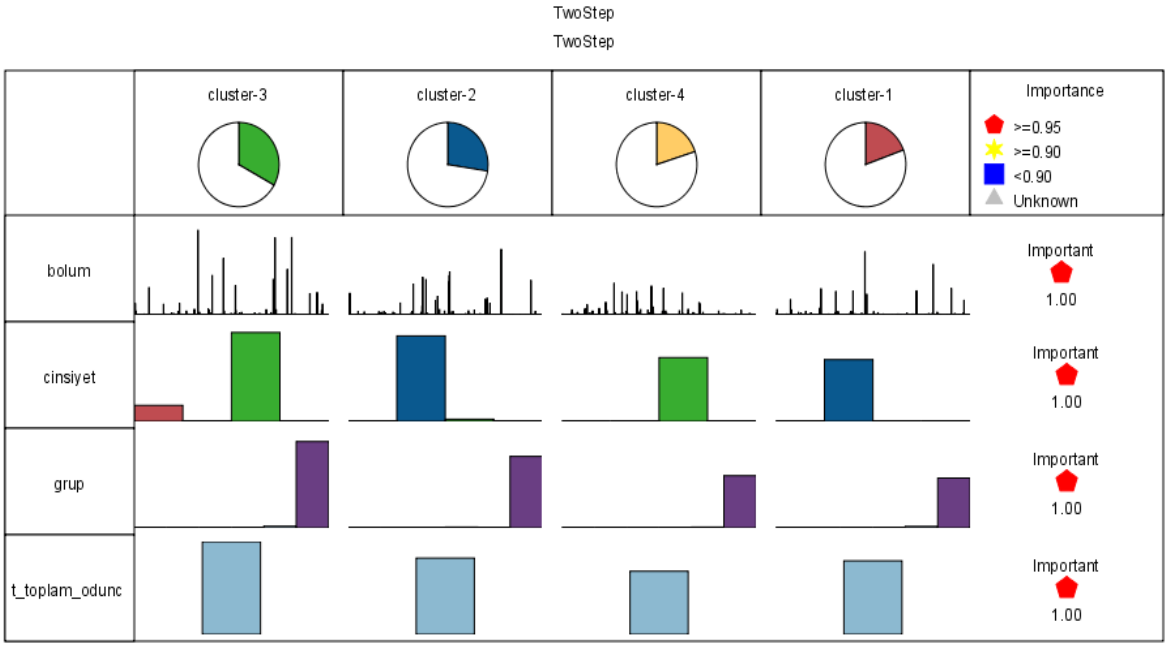
3.6 Kütüphane Kullanıcıları Üzerinde Kümeleme Analizi

Kütüphane verilerinde bulunan sirkülâsyon ve kütüphane kullanıcıları ile ilgili veriler kullanılarak kütüphane kullanıcılarının özelliklerine göre kümelendirmeye gidilebilmektedir. Kümeleme sonucu kullanıcı davranışları incelenebilir ve kütüphane içi kullanım olanakları üzerine birçok çalışma gerçekleştirilebilir.

Kümeleme modülü, kullanıcı portföyü üzerinde mevcut olabilecek gruplar gibi sonucu önceden kestirilemeyen olaylarda kullanılabilmektedir. Kümeleme modellemeleri, benzer kayıtlara sahip, belirlenen değişkenler göz önüne alınarak etiketlenebilen gruplar ortaya çıkarırlar. Bu işlem, gruplar hakkındaki bilgiler ile ilgili öncelik ya da karakteristik özellikleri dikkate alınmadan gerçekleşmesinin yanında, kümeleme modellerini diğer modelleme tekniklerinden ayıran önemli bir özelliktir. Bu özellik, tahmin örüntüleri oluşturulurken öncül olarak tanımlanmış çıktılara ya da hedeflenmiş alanlara yer vermemesidir. Model herhangi bir "doğru" ya da "yanlış" kavramını göz önünde bulundurmaz. Sonuçların önemini gruplaşmaları tespit etmek ve bu gruplar hakkında detaylı bilgiler ortaya koyabilmesidir.

Analizde TwoStep algoritması kullanılması uygun bulunmuştur. TwoStep kümeleme bileşeni, büyük veri setleri için ölçeklenebilir bir veri analiz algoritmasıdır. Devamlılığı olan ya da kategoriksel değişkenler veya davranışlarla birlikte çalışabilir, bir veri kanalı üzerinden sonuca ulaşabilmektedir. İlk adım ön kümeleme evresidir. Bu evrede veriler çok küçük alt kümelere ayrılmaktadır. Daha sonra alt kümeler birleşerek tespit edilen kümelere dönüşmektedir. Küçük veri kümelerinin daha büyük veri kümeleri haline getirilmesi esnasında hiyerarşik kümeleme metotlarından yararlanılmaktadır. TwoStep algoritmasının avantajı kümeleme sayısını otomatik olarak tespit edebilmesidir

Kütüphane kullanıcıları birçok yönden mevcut verilerle gruplandırılabilse de sağlıklı bir öngörü için birden çok veri kullanılarak yeni grupların tanımlanması gerekmektedir. Bu çalışmada kullanıcıların kitap sirkülâsyon verileri dikkate alınarak bölüm, cinsiyet, grup ve toplam ödünç yayın sayısına göre kullanıcılar kümelenecektir. Analizde kümeleme sayısı otomatik olarak belirlenmiş maksimum 15, minimum 2 kümeleme olması planlanmıştır.



Şekil 3-8 Kütüphane Verileri Üzerinde Yapılan Kümeleme Analizi Çıktısı

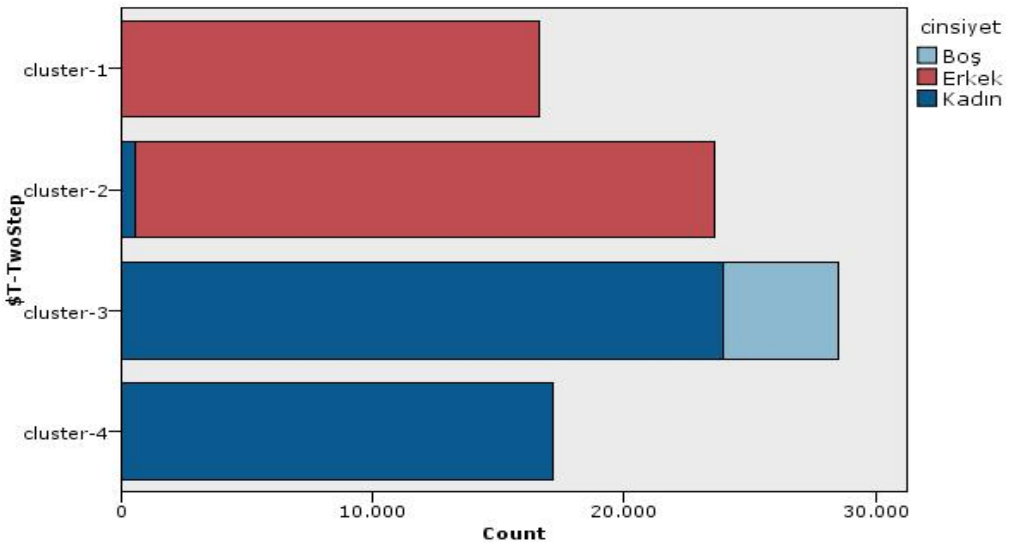
Yukarıda grafiksel olarak gösterilen TwoStep analizi sonucunda 4 küme belirlenmiştir. Bu kümeler bölüm, cinsiyet, grup ve toplam yayın ödünç sayılarına göre belirlenmiş ve kriterlerin hepsinin analizde kullanılabilecek veriler olduğu sağ tarafta bulunan önemlilik simgesi ile gösterilmiştir. Kümelerde sırasıyla 16653, 23612, 28475 ve 17192 kayıt bulunmaktadır.

1. Kümede toplam ödünç yayın ortalaması 31,656, standart sapması ise 32,39 olarak belirlenmiştir. Kullanıcıların bölümleri göz önüne alındığında ise kümeleme içerisinde en fazla paya sahip Kamu Yönetimi bölümü %16,27 değere sahip, kullanıcıların cinsiyetinde ise %99,82 oranında erkek olarak tespit edilmiştir. Kümeleme içerisinde Öğrenci yüzdesi %96,6'dır.

2. Kümede ise ortalama kitap ödünç sayısı 32,885, standart sapma ise 31,89 olarak belirlenmiştir. Bu kümede Tarih bölümünden kullanıcı sayısı küme içerisindeki kullanıcıların %11,96 'sı olarak bulunmuş, erkeklerin ve öğrencileri ağırlıklı olduğu bir küme olmuştur.

3. Kümede ortalama ödünç yayın sayısı ortalama kullanıcı başına 39,904, standart sapma 39,59 olarak bulunmuştur. Küme içerisinde Hemşirelik bölümü tüm diğer bölümler içerisinde %12,8 ile en fazla yer kaplamaktadır. Küme elemanlarının %84'ü kadınlardan oluşmakta ve %98,02 oranında öğrenci bulunmaktadır.

Son kümede ise ortalama 27,182 kullanıcı bulunmaktadır. Standart sapma 24,75 olarak hesaplanmıştır. Gıda mühendisliği kullanıcılarının ağırlıkta olduğu kümede %99,96 oranında kadın ve %98,23 oranında öğrenci bulunmaktadır. Aşağıdaki şekilde örnek olarak cinsiyetin kümelenmedeki dağılımı gösterilmiştir. 1. ve 2. Kümelerde erkek kullanıcılar bulunurken, 3. ve 4. Kümelerde kadın kullanıcılar bulunmaktadır.



Şekil 3-9 Kütüphane verileri üzerinde elde edilen kümeleme grupları

Kümeleme işleminin görsel olarak verildiği grafikte belirgin olarak ayrılabilen cinsiyet unsurunun, kümelerde baskın bir özellik olduğu görülebilmektedir. Fakülte sayısının fazla olması kümelerde belirli bir fakültenin baskınlığını ortaya çıkarmamış, ancak kümeleme içindeki fakülte unsuru homojen olarak kümelere dağılmasına neden olmuştur. Akademik ve idari personelin öğrenci grubuna göre oldukça az sayıda olması küme içerisinde belirgin bir şekilde yer almamasına neden olmuştur. Ödünç yayın sayısı anahtar olarak kullanılmış, kullanıcıların kütüphaneden ilk aldığı yayından itibaren toplam aldığı yayın sayısı kümeleme için kriter olarak belirlenmiş, belirlenen sayıya göre kümeleme içinde dağılımı sağlanmıştır.

TARTIŞMA VE SONUÇ

Dijitalleşme sürecinin gündelik hayatın vazgeçilmez bir parçası olmasıyla birlikte her hareket bilgiye çevrilebilmekte ve bu bilgiler depolanabilmektedir. Depolanan bu verilerin bilgi yığınlarına dönüştürülmesiyle de karar vericilerin karar verme yetenekleri bir yandan gelişmekte iken diğer yandan bu karmaşa içerisinde körelmektedir. Bu sürecin iyi yönetilebilmesi için veri madenciliği tekniği oldukça önemli bir yer teşkil etmektedir. Veri madenciliği ile elde edilen bilgiler işletmelerin/organizasyonların normal şartlar altında ulaşamadıkları verileri, veri yığınları arasından çeşitli algoritmik yapılar sayesinde çıkararak, karar verici için stratejik kararlar vermesine olanak sağlamaktadır.

Kütüphane alanında dijitalleşme süreci çok yeni ve henüz yeterince yaygınlaşmamış bir durumdadır. Büyük ve uluslararası kütüphanelerin sürecin öncüleri olmalarına rağmen özellikle ülkemizde bu süreci sağlıklı bir şekilde tamamlayamamışlardır. Çoğu kar amacı gütmeyen kütüphanelerde kullanıcı memnuniyetinin ön planda tutulabilmesi için dijitalleşme ve buna müteakiben veri madenciliği süreci devamlı olarak geliştirilmeli ve tekrarlanmalıdır.

Çalışmada, kütüphane eser bilgileri göz önüne alınarak oluşturulan yığın verilerin nasıl karar mekanizmasına katılacağı konusunda veri madenciliği tekniklerinden yararlanılmıştır. Çalışmanın verimli olabilmesi için en uygun modellemeler üzerinde çıkarımlar yapılmaya çalışılmıştır.

Veri madenciliğinin işlendiği birinci bölümde, kavram ve detaylar üzerinde durulmuştur. Veritabanlarında bilgi keşfi çerçevesinde verilerin nitelikleri, veritabanları ve veriler arasındaki ilişkiler, veri ambarları ve setleri ve veri işleme süreci üzerine bilgiler verilmiştir. 1980'li yıllarda büyük verilerin saklanabileceğinin ortaya çıkmasıyla beraber bu sürecin gelişimi öne çıkarılmaya çalışılmıştır. Veri madenciliğinin dayandığı temel faktörler detaylarıyla anlatılmıştır. Özellikle hazırlanma sürecinde yaşanan zorluklara dikkat çekilmiş, temizleme, uyarlama, dönüştürme ve daraltma işlemleri gösterilmiştir. Veri madenciliği süreci hakkında teorik olarak bilgiler ortaya konulduğu gibi, uygulamada karşılaşılabilecek sorunlar hakkında da açıklamalarda bulunulmuştur. Süreç literatürde kabul edilen standartlara uygun bir şekilde sıralanmıştır.

Birici bölümün son kısmında dijital kütüphanelerin oluşumu ve gelişimi konusu üzerinde durulmuştur. Kütüphanelerin dijital ortama aktarılması fikri ve bu fikrin kütüphanelerin nasıl

internet ortamında ve yerel ağ sistemlerinde çeşitlenmeye yol açtığı ortaya konulmuştur. Dünya literatüründe çok fazla çalışma olmamasına rağmen son yıllarda otaya çıkan "bibliomining" terimi açıklanmaktadır. Büyük kütüphanelerin dijitalleşme sürecini bile tam olarak tamamlayamaması, literatür çalışmalarının yeterince ortaya çıkmamasına neden olmuştur. Türkiye'de kütüphanelerin dijitalleştirilme süreci ise henüz erken bir dönemde olup kütüphaneler için geliştirilen otomasyon yazılımları ile paralel şekilde gelişme aşamasındadır. Ülkemizde kütüphaneler ilk etapta ellerinde bulunan yayınları dijital ortama sistematik olarak aktarma hedefini gerçekleştirmeye çalışmaktadır. Çalışmada da bu hedefi gerçekleştirilirken meydana gelen sıkıntılar üzerine çalışılmış ve bu hataların nedenleri üzerinde durulmuştur.

İkinci bölümde, veri madenciliği teknikleri ve sınıflandırmaları işlenmiştir. Sınıflandırmalar içerisinde algoritmik yapılar açıklanmış. Bu yapıların istatistiksel temelleri üzerinde durulmaya çalışılmıştır. Birçok model içerisinden uygun olanın seçilmesi için dikkat edilmesi gereken hususlar anlatılmış, avantaj ve dezavantajları üzerinde durulmuştur. Bu bölümde ele alınan sınıflandırmalar; istatistiksel, karar ağacı, geri yayılım, birliktelik, kümeleme sınıflandırmalarıdır.

Tez çalışmasının uygulama kısmı üçüncü bölümde yer almaktadır. Akdeniz Üniversitesi Merkez Kütüphanesi sununcularında bulunan 2005-2009 yılları arasındaki verileri kullanarak çeşitli analizler gerçekleştirilmeye çalışılmıştır. Uygulamaya başlamadan önce verilerin barındırılacağı ve veri madenciliği sürecinin gerçekleştirileceği platformlar belirlenmiştir. Süreç başlamadan hatalı ve önemli olmayan verilerin temizlenmesi SQL sorguları ve geliştirilen PHP kabuk programları ile düzeltilmiştir. Bu işlemler veri madenciliği sürecinin en fazla zaman alan bölümüdür. Veri dönüştürme işlemleri ihtiyaca göre süreç içinde tekrarlanmıştır. Sonraki aşamada ise veri ambarı oluşturulmak üzere SPSS Clementine yazılımı kullanılmıştır. Tüm gerekli verilerin toplu bir şekilde görmemizi sağlayan veri ambarı modellemeler oluşturulurken hız ve kolaylık kazandırmıştır.

Modellemeler oluşturulmadan önce tanımlayıcı bulgular tespit edilmiştir. Öncelikle analizlerde sıkça kullanılan, kütüphaneden yararlanan kullanıcıların Akademik Personel, Öğrenci ve İdari Personel olarak gruplandırıldığı "grup" alanı hakkında istatistikî bilgiler verilmiştir. Kütüphaneden en az bir kere yararlanmış kullanıcılarının yanında, tüm sirkülâsyon verileri dahil edilerek toplam ödünç yayın bilgilerine göre grup dağılımı yeniden şekillendirilmiştir. Kullanıcıların üniversite içi fakültelere göre dağılımı grafiksel olarak ele alınmış. En fazla ödünç yayın alma kıstasına göre sıralanan fakülteler içerisinde, grup

dağılımları işlenmiştir. Başka bir önemli veri olan yayın sınıflandırmaları üzerine de bilgiler verilmiştir. Kütüphane içerisinde yayın sınıflandırmaları, dünyadaki büyük kütüphanelerce yaygın olarak kullanılan uluslararası LC Sınıflandırma Sistemi temel alınarak oluşturulmuştur. Sınıflandırma hakkında bilgiler verildikten sonra kütüphanede bulunan yayınların sınıflandırmalara göre dağılımı üzerinde durulmuştur.

Çalışmada oluşturulan modellemeler iki temel kısma ayrılmaktadır. İlk etapta Yayın sirkülâsyon verilerinin bulunduğu tablo üzerinden yayın sınıfları arasında birliktelik analizi gerçekleştirilmiştir. Kullanıcıların ödünç yayın seçimindeki tercihleri göz önüne alarak ortaya konan çalışmada kütüphane kullanıcılarının ödünç yayın seçiminde hangi birliktelikleri gerçekleştirildiği üzerinde durulmuştur. Sınıflandırma sistemine göre kullanıcıların ilgilendiği yayınlar tespit edilerek, bu yayınların yanında diğer yayın sınıflarına ait hangi yayınları almayı tercih ettikleri tespit edilmiştir. Analizin gerçekleştirilmesi için Apriori algoritmasından yararlanılmış, güven ve destek değerleri veri boyutuna göre şekillendirilmiştir. Elde edilen sonuç kullanıcı tercihlerine ışık tutmuş, kütüphane karar vericileri için, özellikle kütüphanede bulundurulması gerekli olan kitapların tespiti ve kütüphane içi düzen konusunda faydalı veriler ortaya koymuştur. İleride yapılabilecek çalışmalar için başlangıç olmuştur.

İkinci modelleme kütüphane kullanıcılarının kümeleme metodu ile tespiti üzerine yapılmıştır. Yapılan kümeleme modellemesiyle kütüphane kullanıcıları bağlı oldukları bölüm, cinsiyet, üniversite içinde bulundukları konum ve aldıkları toplam yayın sayısı kıstasları temel alınarak analize dahil edilmiştir. TwoStep algoritması kullanılarak elde edilen sonuçlarda kullanıcıların dört kümeye ayrıldığı ortaya çıkarılmıştır. Birinci ve ikinci küme erkek kullanıcıların ağırlıklı olduğu, üçüncü ve dördüncü küme ise kadın kullanıcıların ağırlıklı olduğu bir yapıya sahip olduğu tespit edilmiştir. Kümelerde, akademik ve idari personel sayısının kütüphane içerisindeki kullanıcı sayısı içerisinde az bir oran kapladığı için hiçbir kümede baskın bir konum elde edememiştir. Yine bağlı oldukları bölüm göz önüne alındığında kullanıcıların aldıkları toplam ödünç yayın sayısına göre dağılım gerçekleştirilmiştir. Elde edilen bulgular kütüphane karar vericilerini kullanıcı odaklı araştırmalarında oldukça yararlı olacaktır. Kullanıcı gruplarının tercihleri tespit edilerek gelecekte kütüphane içinde bulundurmayı planladıkları yayınların tespiti ve kullanıcılarının kütüphanelerden gerçek anlamda ne istediklerinin tespiti için yararlı bir sonuç olduğu düşülmektedir.

Sonuç olarak, tüm kullanıcı odaklı işletmeler ve organizasyonlar gibi kütüphaneler de kullanıcı davranışlarını yakından takip etmek zorundadırlar. Kütüphane ortamını dijital ortama aktarmak bu yolda atılmış en büyük adımdır ancak yeterli değildir. Oluşturulan veri yığınları tek başlarına bir anlam ifade etmemekte, bu yüzden de bu verilerin işlenerek yararlı bilgilere dönüştürülmesi kaçınılmazdır. Dijitalleşme sürecinin başlarında olan Akdeniz Üniversitesi Merkez Kütüphanesi üzerinde yapılan çalışmada veri madenciliği teknikleri irdelenmeye çalışılmış, verilerin kullanılabilirliği üzerinde çıkarımlar yapılmıştır. Oluşturulan modellerle tanımlayıcı ve kestirimci bulgular ortaya konulmuştur. Veri madenciliği süreci literatürde yer verildiği gibi başından sonuna kadar gerçekleştirilmiştir. İleride yapılması muhtemel çalışmalarda veri yığınının daha fazla bir zaman dilimini kapsamasıyla farklı modellemeler kullanılabilecek duruma gelecek ve daha gelişmiş sonuçlar elde edilebilecektir.

KAYNAKLAR

Agrawa, R., Gehrke, J., Gunopulos, D., & Raghavan, P. (2005). Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. *Data Mining and Knowledge Discovery* , 5-33.

Ankerst, M., Breunig, M. M., Kriegel, H.-P., & Sander, J. (1999). OPTICS: Ordering Points To Identify the Clustering Structure. *ACM SIGMOD Record , 28* (2), 49 - 60.

Ankerst, M., Breunig, M., & Kriegel, H.-P. (1999). OPTICS: Ordering points to identify the clustering structure. *ACM-SIGMOD Int. Conf. Management of Data* (s. 49-60). Philedelphia: ACM.

Asuman Doğaç, M. T. (1994). Advances in object oriented database systems. *NATO Advanced Study institute on Object Oriented Databases* (s. 4-8). İzmir: Springer.

Berkhin, P. (2006). A Survey of Clustering Data Mining Techniques . C. N. Jacob Kogan içinde, *Grouping Multidimensional Data* (s. 25-71). Berlin : Springer Berlin Heidelberg.

Borgelt, C., & Kruse, R. (2002). Induction of Association Rules:Apriori Implementation. *Compstat: Proceedings in Computational Statistics* (s. 395). Berlin: Physica Verlag.

Brin, S. L. (1998). The Anatomy of Large-scale hypertextual web seargh engine. *Computer Networks*, (s. 107,117).

Camila, M., Barioni, N., Razente, L. H., Traina, A. J., & Caetano, T. (2008). Accelerating k-medoid-based algorithms through metric. *Journal of Systems and Software* , 343-355.

Chaudhuri, S., & Dayal, U. (1997). An overview of data warehousing and OLAP technology. *SIGMOD Rec. , 24* (1), 65--74.

Clark Labs. (2008). Classification Tree Analysis. Worcester, MA, USA.

Cullen, K. (2005). Delving into Data. *Library Journal* , 30-32.

Çakmak, Z., Uzgören, N., & Keçek, G. (2005). Kümeleme Analizi teknikleri ile illerin kültürel yapılarına göre sınırlandırılması ve değişimlerinin incelenmesi. *Dumlupınar Üniversitesi Sosyal Bilimler Enstitüsü Dergisi* , 15-37.

Çetinyokuş, T., & Gökçen, H. (2008). Bütünleşik Veri Küpü Sistemi (BVKS):Satış Küpü Uygulaması. *Gazi Üniv. Müh. Mim. Fak. Der. , 23* (2), 477-484.

David Hand, H. M. (2001). *Data Mining.* Massachusetts Institute of Technology.

Doedens, C.-J. (1994). *Text Databases: One Database Model and Several Retrieval Languages.* Amsterdam: Rodopi B.V.

Du, W., & Zhan, Z. (2002). Building decision tree classifier on private data. *ACM International Conference Proceeding Series Volume 144* (s. 1 - 8). Maebashi City, Japan: Australian Computer Society, Inc.

Ester, M., Kriegel, H., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proc. KDD* (s. 226--231). içinde

Filemaker Inc. (2010). *Filemaker - Company*. 06 2010, 16 tarihinde Filemaker: http://www.filemaker.com/company/index.html adresinden alındı

Fuhr, N., Tsakonas, G., Trond, A., Agosti, M., & Hansen, P. (2007). Evaluation of digital libraries. *International Journal on Digital Libraries* , 21-38.

Golfarelli, M., & Rizzi, S. (1998). A methodological framework for data warehouse design. *Proceedings of the 1st ACM international workshop on Data warehousing and OLAP* (s. 3 - 9). Washington, D.C., United States: ACM.

Google, O. B. (2008, 7 25). *We new the web was big*. 3 13, 2010 tarihinde Google Official Blog: http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html adresinden alındı

Grandville, S., & Peter, J. R. (2005). *Finding groups in data: An introduction to cluster analysis.* John Wiley and Sons.

Güting, R. H. (1994). An Introduction to Spatial Database Systems. *VLDB Journal , 3* (4), 357-399.

Halevy, A. (2001). Answering queries using views: A survey. *The VLDB Journal , 10* (4), 270--294.

Haydar, A., Ağdelen, Z., & Özbeşeker, P. (2006). The Use of Backproganation Algorithm in the Estimation of Firm Performance. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi Yıl: 5 Sayı:10* , 51-64.

Hector Garcia-Molina, J. D. (2008). *Database Systems: The Complete Book.* New Jersey: Prentice Hall.

Hinneburg, A., & Keim, D. (1998). An efficient approach to clustering in large multimedia databases with noise. *Knowledge Discovery and Data Mining , 5865*.

Huang, Y.-P., & Hoa, V. T. (2009). General criteria on building decision trees for data classification. *ACM International Conference Proceeding Series; Vol. 403* (s. 649-654). Seoul, Korea: ACM.

Hull, R. (1997). Managing semantic heterogeneity in databases: a theoretical prospective. *PODS '97: Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems* (s. 51--61). New York: ACM.

Inmon, W. H. (2002). *Building the Data Warehouse.* Canada: John Wiley & Sons, Inc.

Jain, A. K. (1999). Data Clustering: A Review. *ACM Computing Surveys* (3), 31.

Jiwai Han, M. K. (2001). *Data Mining Concepts and Techniques.* San Diago: Academic Press.

Kantardzic, M. (2001). *Data Mining Concepts, Models, Methods and Algorithms.* Piscataway, NJ: IEEE Press.

Kosala, R., & Blockeel, H. (2000). Web mining research: a survey. *ACM SIGKDD Explorations Newsletter* , 1-15.

Koutri, M. a. (2004). A survey on web usage mining techniques for web-based adaptive hypermedia systems. *Proceedings of the Adaptable and Adaptive Hypermedia Systems* , 125--149.

Krysztof J. Horis, W. P. (2007). *Data Mining A Knowledge Discovery Approach.* New York: Springer Sciene+Business Media.

Kunder, M. d. (2010, 3 13). *The size of the World Wide Web*. 3 13, 2010 tarihinde http://www.worldwidewebsize.com/ adresinden alındı

Kunttu, I., Lepistö, L., Rauhamaa, J., & Visa, A. (2004). Grid-based clustering in the content-based organizaton based of large image image databases.

Library of Congress. (tarih yok). *Library of Congress Classification*. 07 10, 2010 tarihinde Library of Congress Classification: www.loc.gov adresinden alındı

Liu, B. (2007). *Web data mining: exploring hyperlinks, contents, and usage data.* Chicago: Springer.

Malcolm Atkinson, F. B. (1989). The Object-Oriented Database System Manifesto. *Proceedings of the*
First International Conference on Deductive and Object-Oriented Databases, (s. 223-40). Kyoto, Japan.

McCallum, A., & Nigam, K. (2003). A Comparison of Event Models for Naive Bayes Text Classification. *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 1* (s. 307 - 314). Budapest: Association for Computational Linguistics.

Meltem IŞIK, A. Y. (2007). K-Means, K-Medoids ve Bulanık C-Means Algoritmalarının Uygulamalı Olarak Performanslarının Testleri. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi* , 31-45.

Nicholson, S. (2003). The Bibliomining Process: Data Warehousing and Data Mining for Library Decision-Making. *Information Technology and Libraries* .

Nutter, S. K. (1987). Online systems and the management of collections: Use and implications. *Advances in Library Automation Networking* , 125-149.

Öztemel, E. (2003). *Yapay Sinir Ağları.* İstanbul.

Pyle, D. (1999). *Data Prepation for Data Mining.* San Diago: Morgan Kauffman.

Rahm, E., & Hai, D. H. (2000). Data cleaning: Problems and current approaches. D. B. Lomet (Dü.), *Bulletin of the technical committee on data engineering.* içinde *23*, s. 3. Washington: IEEE Computer Society.

Sheikholeslami, Chatterjee, G., & Zhang, A. (1998). Wavecluster: A multi-resolution clustering approach for very large spatial databases. *Proceedings of the International Conference on Very Large Data Bases* (s. 428-439). Citeseer.

Sheikholeslami, G., Chatterjee, S., & Zhang, A. (1998). Wavecluster: A multiresulotion clusteirng approach for very large spatial databases. *Int. Conf. Very Large Databases*, (s. 428-439). New York.

SPSS. (2010). *Data mining, Text mining, predictive analysis*. 06 2010, 19 tarihinde SPSS: http://www.spss.com/software/modeling/modeler-pro/ adresinden alındı

Sun, J., & Li, H. (2008). Data mining method for listed companies' financial distress prediction. *Knowledge-Based Systems , 21* (1), 1-5.

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining.* Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Timor, M., & Şimşek, U. T. (2008). Veri Madenciliğinde Sepet Analizi ile Tüketici Davranışları Modellemesi. *İşletme İktisadı Enstitüsü Dergisi, Yönetim Dergisi* .

Wikipedia contributors. (2010, Nisan 14). *Affinity analysis*, 355967909 . Nisan 17, 2010 tarihinde Wikipedia: http://en.wikipedia.org/w/index.php?title=Affinity_analysis&oldid=355967909 adresinden alındı

Wikipedia. (2010, Ocak 7). *Data binning* . Mart 27, 2010 tarihinde Wikipedia, The Free Encyclopedia. : http://en.wikipedia.org/w/index.php?title=Data_binning&oldid=336389770 adresinden alındı

Wikipedia. (tarih yok). *Database*, 348574512 . Wikipedia, The Free Encyclopedia.: http://en.wikipedia.org/w/index.php?title=Database&oldid=348574512 adresinden alınmıştır

Wikipedia. (2010, 06 10). *MySQL*. 06 19, 2010 tarihinde Wikipedia: http://tr.wikipedia.org/wiki/MySQL adresinden alındı

Xiong, h., Pandey, G., Steinbach, M., & Kumar, V. (2006). Enhancing Data Analysis with Noise Removal. *IEEE Transactions on Knowledge and Data Engineering. 18*, s. 304 - 319. Piscataway, NJ, USA: IEEE Educational Activities Department.

Zhu, X., Wu, X., & Chen, Q. (2006). Bridging Local and Global Data Cleansing: Identifying Class Noise in Large, Distributed Data Datasets. *Data Mining and Knowledge Discovery , 12* (2-3), 275-308.

Ziegler, P., & Dittrich, K. R. (2004). Three Decades of Data Integration—All Problems Solved? I. I. Processing içinde, *Building the Information Society* (s. 3-12). Boston: Springer Boston.

Ö Z G E Ç M İ Ş

Adı ve SOYADI : ÖMER UÇAN

Doğum Tarihi ve Yeri : 10.06.1983 BURDUR

Medeni Durumu : BEKAR

Eğitim Durumu

Mezun Olduğu Lise : ADEM TOLUNAY FEN LİSESİ

Lisans Diploması : İSTANBUL ÜNİVERSİTESİ İNGİLİZCE İŞLETME BÖLÜMÜ

Yabancı Dil / Diller : İNGİLİZCE

İş Deneyimi

Çalıştığı Kurumlar : Maliye Bakanlığı Milli Emlak Müdürlüğü

Adres : Uncalı Mahallesi 1223. Sk. Kadeşler Sitesi B Blok K:3 D:6

Tel. no : 05052203053

E-mail : omer@ucan.gen.tr