**Giriş**

Veri toplamada ve depolama teknolojisindeki hızlı gelişmeler şirketlerin büyük veriler biriktirmesine yol açmıştır. Ancak, faydalı bilgilerin ayrıştırılmasının çok zor olduğu da bir gerçek. Genelde, geleneksel veri analiz araçları ve teknikleri, veri kümesinin büyüklüğü yüzünden kullanılamaz. Bazen verinin geleneksel olmayan doğası, veriler küçük bile olsa geleneksel yaklaşımların uygulanacağı anlamına gelmez. Diğer durumlarda, cevaplanması gereken sorular mevcut veri analiz tekniklerini kullanarak ele alınamaz ve böylece yeni yöntemler geliştirilmeye ihtiyaç duyulur.

Veri madenciliği geleneksel veri analiz yöntemlerini büyük hacimdeki verileri işlemek için karmaşık algoritmalarla birleştiren bir teknolojidir. Ayrıca, yeni türdeki verilerin keşfedilmesi ve analizi için ve eski türdeki verilerin yeni yollarla analiz edilmesi için heyecan verici gelişmeler sunar. Giriş bölümünde veri madenciliğine genel bir bakış sunuyoruz ve bu kitapta ele alınacak kilit detayları özetliyoruz. Veri analizi için yeni teknikler gerektiren bazı iyi bilinen uygulamaların bir açıklamasıyla başlıyoruz.

İş satış noktası veri toplama (barkod tarayıcısı, radyo frekansı tanıma ve akıllı kart teknolojileri) perakendecilerin mağazalarındaki ödeme sayaçlarıyla(cihazlarıyla) müşteriye satılan ürünlerle ilgili en güncel verileri toplamalarını sağlamıştır. Perakendeciler bu bilgiden yararlanabilirler, bununla birlikte diğer iş-kritik gibi veriler Web kütüklerinde e-ticaret, web siteleri ve müşteri servisleri çağrı merkezinde kaydedilirler, böylece müşterilerin ihtiyaçlarını daha iyi anlamak ve daha bilinçli iş kararları vermek mümkün olabilir.

Tıp, Bilim ve Mühendislik, tıp araştırmacıları bilim ve mühendislik, hızla biriken önemli anahtarlı verileri topluyor. Örneğin, önemli bir adım olarak Dünya'da iklim sistemi anlayışımızı geliştirmek. NASA sürekli olarak arazi yüzeyini, okyanusları ve atmosferin küresel gözlemlerini üreten bir dizi dünya yörüngesinde uydu kullandı. Ancak, boyut ve uzay-zamansal veriden dolayı geleneksel yöntemler bu verilerin analizi için uygun değildi. Veri madenciliğinde geliştirilen teknikler dünya biliminde şu gibi soruların cevaplarına yardım edebilir "Kuraklık ve kasırgalar gibi ekosistem bozukluklarının frekansı ve yoğunluğuyla küresel ısınmanın arasında ne gibi ilişki vardır?" "Arazi yüzeyindeki sıcaklık ve yağış okyanus yüzeyine göre nasıldır?" ve "Bir bölgenin mevsim başlangıcını ve sonunu ne kadar iyi tahmin edebiliriz?"

Başka bir örnekle, moleküler biyoloji araştırmacıları büyük miktardaki genomik veriyi kullanmayı umuyorlar, halen yapı ve gen fonsiyonunu daha iyi anlamak için veri topluyorlar. Geçmişte, moleküler biyolojideki geleneksel yöntemler bilim insanlarının bir seferde sadece birkaç gen üzerinde deney yapmasına izin veriyordu. Son günlerde mikrodizin teknolojisindeki buluşlar bilim insanları binlerce gen üzerindeki çeşitli davranışları karşılaştırmasını sağlamıştır. Bu tür karşılaştırmalar her bir genin fonksiyonunu belirlemeye yardım edebilir ve bazı hastalıklar için belli genleri izole edebilir. Ancak, gürültülü ve yüksek boyutlu veriler yeni veri analiz tipleri gerektirir. İlave olarak gen dizisi verilerinin analizi veri madenciliği adreslemek içinde kullanılabilir protein yapısı gibi diğer önemli biyolojik zorluklar, çoklu dizi hizalaması, biyomedikal yollar modelleme ve filogenetik.

**1.1 Veri madenciliği nedir?**

Veri madenciliği işlemi büyük veri ambarından yararlı bilgileri otomatik olarak bulma işlemidir. Veri madenciliği teknikleri büyük verilerden yeni bilgiyi ve kullanışlı kalıpları bulmak için kullanılır. Ayrıca gelecekteki gözlemin sonucunu tahmin yeteneğini sağlar, böylece yeni müşteri geldiğinde mağazada 100$’dan fazla harcayacağını tahmin edebiliriz. Veri madenciliğinin görevinin tüm bilgilerin keşfi olduğu düşünülmüyor. Örneğin, bir veritabanı yönetimini kullanırken bireysel kayıtları arama veya web sayfası üzerinden belirli aramaların sorgusu, İnternet arama motorundan bilgi alma alanıyla ilgili. Görevlerin böyle olmasına rağmen önemli ve karmaşık algoritmalar kullanımını içerebilir ve veri yapıları, bilgileri etkin bir şekilde düzenlemek ve almak için endeks yapılar oluşturmak için geleneksel bilgisayar bilimi tekniklerine ve verilerin açık özelliklerine güvenir. Her şeye rağmen, bilgi alma sistemlerini geliştirmek için veri madenciliği teknikleri kullanılmıştır.

Veri madenciliği teknikleri geniş bir yelpazeyle iş zekası uygulamalarını müşteri profili gibi alanları desteklemek için kullanılabilir

Önemli iş sorularını cevaplayın "En karlı müşteriler kim?","Hangi ürünler direkt veya daha az satılabilir?" ve "Şirketin gelecek yıl için gelir tahmini nedir?" Bu sorulardan bazıları birlik analizi yapmaya motive etti (başlıklar 6,7) yeni bir veri analiz tekniklerine.

Tıp, Bilim ve Mühendislik, tıp ataştırmacıları bilim ve mühendislik, hızla biriken önemli anahtarlı verileri topluyor. Örneğin, önemli bir adım olarak Dünya'da iklim sistemi anlayışımızı geliştirmek. Nasa sürekli olarak arazi yüzeyini, okyanusları ve atmosferin küresel gözlemlerini üreten bir dizi dünya yörüngesinde uydu kullandı. Ancak, boyut ve uzay-zamansal veriden dolayı geleneksel işlemler bu verilerin analizi için uygun değildir. Veri madenciliğinde geliştirilen teknikler dünya biliminde şu gibi soruların cevaplarına yardım edebilir "Kuraklık ve kasırgalar gibi ekosistem bozukluklarının frekansı ve yoğunluğuyla küresel ısınmanın arasında ne gibi ilişki vardır?" "Arazi yüzeyindeki sıcaklık ve yağış okyanus yüzeyine göre nasıldır?" ve "Bir bölgenin mevsim başlangıcını ve sonunu ne kadar iyi tahmin edebiliriz?"

**Veri Madenciliği ve Bilgi Keşfi**

Veri madenciliği, Şekil 1.1.de gösterildiği gibi ham verinin faydalı bilgiye dönüşümün genel bir süreci olan veri tabanındaki bilgi keşfinin ayrılmaz bir parçasıdır. Bu işlem,veri madenciliğinin veri önişleminden son işlemine kadar bir dizi dönüşüm adımından oluşur.

Giriş verisi çeşitli biçimlerde tutulabilir(düz dosyalar, tablolar veya ilişkisel tablolar) ve merkezi veri deposunda bulunabilir veya birden fazla siteye dağıtılmış olarak bulunabilir. Ön işlemenin amacı ham giriş verisini bir sonraki analiz için uygun bir formata çevirmektir. Veri ön işlemeye dahil olan adımlar, birden fazla kaynaktan gelen verilerin birleştirilmesi, gürültünün giderilmesi ve gözlemlerin tekrarlanması için verilerin temizlenmesi ve eldeki veri madenciliği göreviyle ilgili kayıtların ve özelliklerin seçilmesini içerir.Verilerin toplanıp saklanmasının birçok yolu nedeniyle, veri ön işleme, genel bilgi bulma sürecinde belki de en zahmetli ve zaman alıcı adımdır.