D.Ü.Ziya Gökalp Eğitim Fakültesi Dergisi **7**, 95-107 (2006)

VERİ MADENCİLİĞİ UYGULAMA ALANLARI

Application Fields of Data Mining

Abdullah BAYKAL[1]

Özet

Günümüzde bilgisayar sistemleri her geçen gün ucuzluyor ve aynı zamanda güçleri de artıyor. Bilgisayar sistemlerindeki bu gelişmeyle birlikte kullanımı da bu ölçüde yaygınlaşmaktadır. Bu gelişmeyle birlikte işletmelerde üretilen sayısal bilgi miktarının arttığını buna paralel veri tabanlarının daha fazla veriyi saklayabilecek boyutlara ulaştığını,ve bilgisayar sistemlerindeki gelişme ile veriye ulaşmanın kolaylaştığını görmekteyiz. Bu sayede doğru ve daha detaylı bilgiye ulaşmamız mümkün hale gelmiş fakat başka bir sorunu ortaya çıkarmıştır. Bu sorun oluşan bu büyük sayısal veri yığınlarının yönetilmesi ve anlamlı hale getirilmesi sorunudur.

Veri kendi başına değersizdir. İstediğimiz amacımız doğrultusunda bilgidir. Bilgi bir amaca yönelik işlenmiş veridir. Veriyi bilgiye çevirmeye veri analizi denir. Bilgi de bir soruya yanıt vermek için veriden çıkardığımız olarak tanımlanabilir. Veri sadece sayılar veya harfler değildir; veri, sayı ve harfler ve onların anlamıdır. Veri hakkındaki bu veriye metaveri diyoruz. Bu veriler belli bir amaç doğrultusunda işlendiği zaman anlamlı hale gelmektedir. İşte ham veriyi bilgiye veya anlamlı hale dönüştürme işini veri madenciliği ile yapabiliriz.

Anahtar Kelimeler: veri madenciliği, veri, veri tabanı, bilgi keşfi

Abstract

Recently, computer systems are cheapening day by day and at the same time their power is increasing with this adavancement their usage spread.With this advancement, we see that the amount of numerical data produced in businesses increase and in relation to this data, bases reaching the

[1] Dr.; D.Ü.Bilgi İşlem Daire Başkanlığı, **baykal@dicle.edu.tr**

dimensions which will able to keep much more data and with advencements in computer systems, getting the data becomes easier. Accordingly, it became possible to get reliable and more elaborate data, but another problem emerged. This problem is to run numerical data mass and to get then meaningful state.
Data is worthless alone.Our wish is the in knowledge which is paralel to our aim. Knowledge is data which is operated. Turining data into knowledge is named as data analysis. Data isn't only numbers and letters. Data is numbers, letters and their meanings. We define this data as metadata . This data can become meaningful at will. We can turn crude data into knowledge or meaniful state with data mining.
Keywords : *data mining, data, database, knowledge discovery*

1.Veri Madenciliği

Veri madenciliği; önceden bilinmeyen, geçerli ve uygulanabilir bilginin veri yığınlarından dinamik bir süreç ile elde edilmesi olarak tanımlanabilir. Bu süreçte kümeleme, veri özetleme sınıflama kurallarının öğrenilmesi, bağımlılık ağlarının bulunması, değişkenlik analizi ve anomali tespiti gibi farklı birçok teknik kullanılmaktadır.

Veri madenciliği ile büyük veri yığınlarından oluşan database sistemleri içerisinde gizli kalmış bilgilerin çekilmesi sağlanır. Bu işlem, istatistik, matematik disiplinleri, modelleme teknikleri, database teknolojisi ve çeşitli bilgisayar programları kullanılarak yapılır. Veri madenciliği büyük miktarda veri inceleme amacı üzerine kurulmuş olduğu için veri tabanları ile yakından ilişkilidir. Gerekli verinin hızla ulaşılabilecek şekilde amaca uygun bir şekilde saklanması ve gerektiğinde hızla ulaşılabilmesi gerekir. Günümüzde yaygın olarak kullanılmaya başlanan veri ambarları günlük kullanılan veri tabanlarının birleştirilmiş ve işlemeye daha uygun bir özetini saklamayı amaçlar.

Veri madenciliği kendi başına bir çözüm değil çözüme ulaşmak için verilecek karar sürecini destekleyen, problemi çözmek için gerekli bilgileri sağlamaya yarayan bir araçtır. Veri madenciliği; analistin'e, iş yapma aşamasında oluşan veriler arasındaki şablonları ve ilişkileri bulması konusunda yardım etmektedir.

2. Veri Madenciliği Uygulama Alanları

- Veri tabanı analizi ve karar verme desteği
- Pazar Araştırması : Hedef pazar , müşteriler arası benzerliklerin saptanması, sepet analizi, çapraz pazar incelemesi
- Risk Analizi : Kalite kontrol, rekabet analizi, öngörü, sahtekarlıkların saptanması
- Belgeler arası benzerlik : haber kümeleri, e-posta
- Müşteri kredi risk araştırmaları
- Kurum kaynaklarının en optimal biçimde kullanımı
- Geçmiş ve mevcut yapı analiz edilerek geleceğe yönelik tahminlerde bulunma,[4]

Pazarlama

- Müşterilerin satın alma örüntülerinin belirlenmesi,
- Müşterilerin demografik özellikleri arasındaki bağlantıların bulunması,
- Posta kampanyalarında cevap verme oranının artırılması,
- Mevcut müşterilerin elde tutulması, yeni müşterilerin kazanılması,
- Pazar sepeti analizi *(Market Basket Analysis)*
- Müşteri ilişkileri yönetimi *(Customer Relationship Management)*
- Müşteri değerlendirme *(Customer Value Analysis)*
- Satış tahmini *(Sales Forecasting)*.

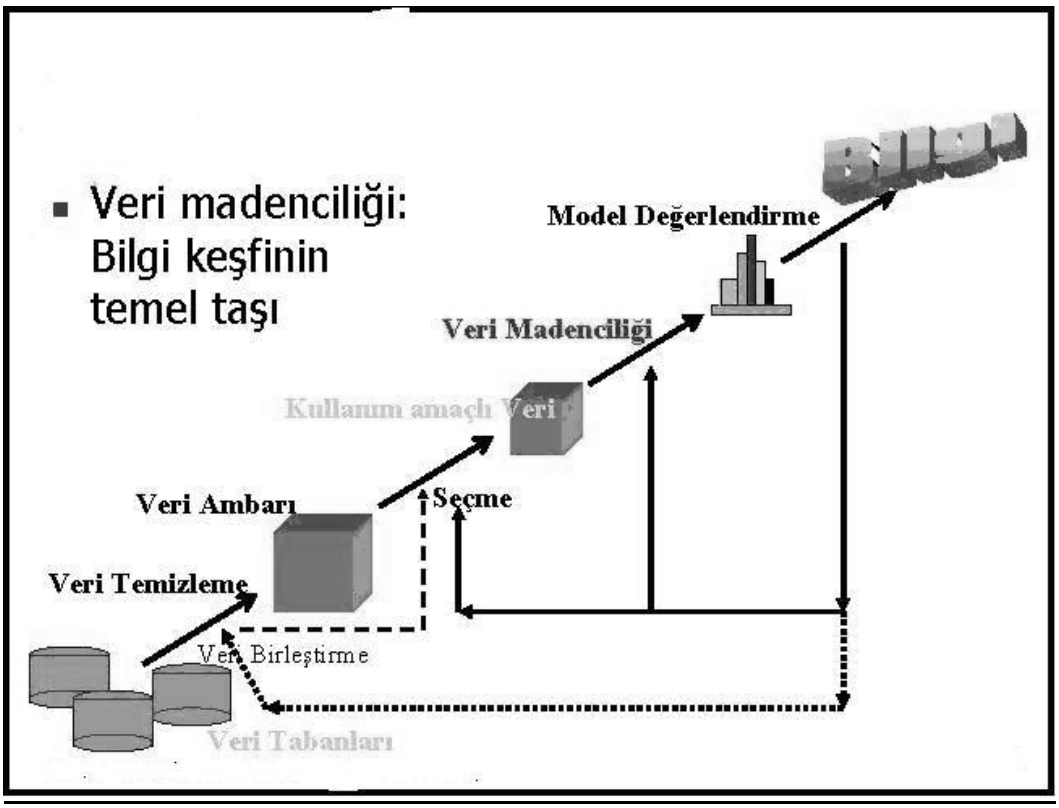
Bankacılık

- Farklı finansal göstergeler arasında gizli korelasyonların bulunması,
- Kredi kartı dolandırıcılıklarının tespiti,
- Kredi kartı harcamalarına göre müşteri gruplarının belirlenmesi,
- Kredi taleplerinin değerlendirilmesi.

Sigortacılık

- Yeni poliçe talep edecek müşterilerin tahmin edilmesi,
- Sigorta dolandırıcılıklarının tespiti,
- Riskli müşteri örüntülerinin belirlenmesi.

3. Bilgi Keşfi

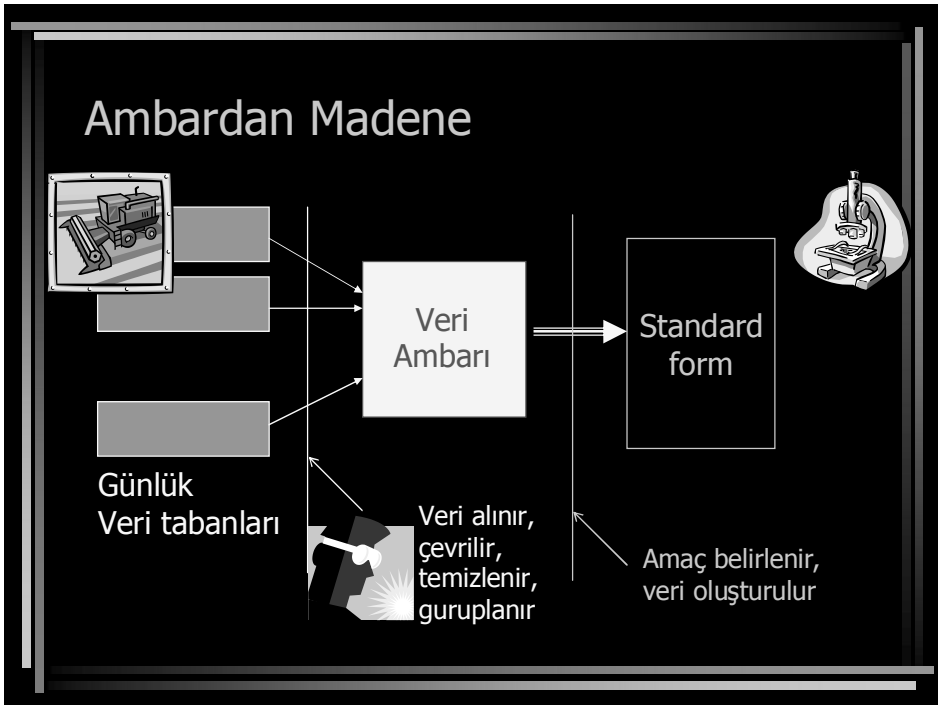


Şekil 1: Bilginin keşfi aşamaları [2]

3.1.Bilgi Keşfinin Aşamaları

- Uygulama alanını inceleme : Konuyla ilgili bilgi ve uygulama amaçları
- Amaca uygun veri kümesi oluşturma: Veri seçme
- Veri ayıklama ve önişleme : işlemin %70'lik bölümünü oluşturur
- Veri azaltma ve veri dönüşümü : incelemede gerekli boyutları (özellikleri) seçme, boyutlar arası ilişkiyi belirleme, boyut azaltma,

- Veri madenciliği tekniği seçme
- Sınıflandırma, eğri uydurma, bağıntı kuralları, demetleme
- Veri madenciliği algoritmasını seçme
- Model değerlendirme ve bilgi sunumu
- Bulunan bilginin yorumlanması



Şekil 2 : Verinin işleme aşamaları

3.2. Bilgi Keşfi Örnek : web kayıtları

- web sitesinin yapısını inceleme
- verileri seçme: tarih aralığını belirleme

- veri ayıklama, önişleme: gereksiz kayıtları silme
- veri azaltma, veri dönüşümü: kullanıcı oturumları belirleme
- veri madenciliği tekniği seçme: demetleme
- veri madenciliği algoritması seçme: k-ortalama, EM, DBSCAN...
- Model değerlendirme/yorumlama: değişik kullanıcı grupları için sıkça izlenen yolu bulma .
- Uygulama alanları: öneri modelleri, kişiselleştirme, ön belleğe alma

4. Veri Ambarları :

1990'lardan geriye kalan bir şey varsa, o da iş yaşamında bir moda gibi hızla değişen planlar, analizler ve stratejilerdir. Bu değişime ayak uydurabilmek için ise üst seviye yöneticilerin, analistlerin ve bilgi işçilerinin ihtiyaç duydukları şey daha daha fazla enformasyondur.

Enformasyon teknolojileri ile organizasyonun dünya çapında işler yapmasına olanak tanıyan devrim niteliğindeki sistemlerin sunulması mümkündür. Ancak işin acı ama gerçek olan yanı da şudur ki masalardaki çok güçlü bilgisayarlarlara ve iletişim sistemlerine rağmen uzmanlar, karar mekanizmasını oluşturan yöneticiler ve danışmanlar, organizasyonlarında zaten mevcut olan kritik bilgilere ya da enformasyonlara ulaşamamaktadırlar.

Her gün , organizasyonlar az veya çok milyarlarca byte, müşterileri, süregiden işleri, ürünleri, çalışanları vs. hakkında data üretirler. Ancak bu datalar bilgisayar sistemlerinin içinde her geçen gün ulaşılması daha da zor bir hal alarak gömülür giderler. İşte bu tipteki verilere "data in jail" (hapishanedeki data) gözüyle bakabiliriz.

Uzmanlar, mevcut verilerden elde edilmiş, işlenmiş ve depolanmış sadece küçük bir parça veriye dayanarak değerlendirme yapabilmektedirler.

Pazardaki ani değişmelerin beklenmedik bir şekilde firmalara olan meydan okuması karşısında bilgi teknolojileri en kritik misyonu üstlenmektedirler. Bilgi Teknolojileri departmanlarının stratejik misyonları, teknolojiyi kesintisiz ve mekandan bağımsız bir halde sunarak organizasyonların performansını arttırmaktır. Sizin, organizasyonun iş stratejisinin merkezindeki yeriniz, yüksek kalitedeki veriyi doğru insana doğru zamanda sunabilmenize bağlıdır.

4.1. Bir Veri Ambarı Yapısı

Bir Veri Ambarının yapısı organizasyon içindeki bütün son kullanıcılara verileri ve işlem sonuçlarını sunan, en gelişmiş iletişimi sağlayan dizi birbiriyle bütünleşik alt bileşenlerden oluşur. Bunlar;

- *Operasyonel Veri Tabanı / Harici Veri Tabanı Katmanı,*
- *Enformasyon Ulaşım katmanı,*
- *Data Ulaşım Katmanı,*
- *Data Directory (Metadata) Katmanı,*
- *İşlem (process) Yönetim Katmanı,*
- *Uygulama Haberleşmesi Katmanı,*
- *Veri Ambarı Katmanı,*
- *Data Sunum Katmanı,*

4.1.1. Operasyonel Veri Tabanı / Harici Veri Tabanı Katmanı

Operasyonal sistemler kritik operasyonel ihtiyaçlar için verileri işlerler. Bunu gerçekleştirebilmek amacıyla uzun süreçlerdeki tecrübelerden yararlanılarak en verimli iş modellerini tanımlanabilmiş ve operasyonel veri tabanları oluşturulmuştur. Bu yüzden, operasyonel sistemleri üzerindeki limitli odaklanmalar yönetim ve de enformasyonel amaçlara verilere ulaşmada zorlukların çıkmasına vesile olmuşlardır. Operasyonel verilere ulaşmadaki bu zorluklar çoğu 10-15 yıllık operasyonel sistemler tarafından da imkansız gibi görülmeye başlamıştır. Ancak bahsettiğimiz veri tabanları bu zorlukları bertaraf ederek kolay bir şekilde verilere ulaşılmasını sağlamaktadır.

Şu açıktır ki, veri ambarlarının amacı operasyonel veri tabanlarındaki verileri kullanılabilir kılmak ve diğer harici enformasyonel verilerle harmanlamaktır. Günümüzde artan bir şekilde çok büyük şirketlerin dış dünyadaki veri tabanlarından verileri kendi ihtiyaçları doğrultusunda yakalayarak bünyelerine dahil ettiğini görebiliyoruz. Bu veriler demografik, ekonometrik, rekabetçi ve satınalma trendleri ile ilgili olabiliyor. İşte bilgi otobanı diye tarif edilen şey her geçen gün daha fazla bilgiye ulaşmaya olanak sağlamaktır.

4.1.2. Enformasyon Ulaşım Katmanı

Enformasyon ulaşım katmanın bir veri ambarındaki yeri son kullanıcının direkt olarak veriye ulaşması ile ilgili kısmıdır. Bu amaca yönelik olarak enformasyon ulaşım katmanı kullanıcıların bu verilere ulaşması için araçlar sunar ; bunlar hergün kullandığımız Excel, Lotus 1-2-3, Focus, Access, vs. 'dir. Bu katman donanım ve yazılımları, grafikleri, sunuları ve sonuçları kağıt üzerine aktaran yazıcıları da kapsar. Özellikle geçtiğimiz son 10 yıl içinde bilgisayar ağlarının oldukça kolay kurulur ve kullanılabilir hale gelmesi ortak bir platformda çalışmayı da kolay hale getirmiştir.[5]

Günümüzde verileri analiz etmek ve sunmak için masalarımızda konunun çok daha fazla felsefi boyutları düşünülerek tasarlanmış araçlar mevcuttur. Bununla birlikte, ham verileri operasyonel işlemlerde kullanılabilir hale getirirken çok önemli sorunlar da ortaya çıkmıştır. Bu sorunların çözüm örneklerinden birisi olarak da verilerin sorgulanmasında ortak bir sorgulama dilinin kullanılmasını gösterebilmek mümkündür.

4.1.3. Data Ulaşım Katmanı

Bir veri ambarı yapısının data ulaşım katmanı enformasyon ulaşım katmanı ile operasyonel katmanın birbiriyle haberleşmesini sağlayan bir katmandır. Günümüzdeki network dünyasında kabul edilen ortak sorgulama dili olan SQL bu sebeple çıkmıştır. İlk olarak IBM tarafından geliştirilen bu veri tabanı sorgulama dili geçtiğimiz 20 yılda bir data sorgulama dilinden öte bir veri değişimi dili olarak endüstriyel bir standart halini almıştır.Veri Ambarcılığının en önemli stratejilerinden birisi "dataya evrensel ulaşım sağlama"dır. Dataya evrensel ulaşım sağlamak demek, son kullanıcının teorik olarak da olsa dataya herhangibir extra araç, yazılım kullanmadan ihtiyaç duyduğu her an her yerden ulaşabilmesidir.
Data ulaşım katmanı, enformasyon ulaşım araçlarıyla operasyonel veri tabanları arasında bir arabirim oluşturmaktan da sorumludur. Çoğu durumda bu son kullanıcının ihtiyaç duyduğu tek şeydir. Bunlara ek olarak, günümüzdeki birçok organizasyon veri ambarcığını desteklelyici yönde her geçen gün yeni yeni felsefi derinliği olan fikirler üzerinde çalışmaktadırlar.[5]

4.1.4.Data Directory (Metadata) Katmanı

Evrensel veri ulaşımını sağlayabilmek için, verilerin yapısı da kesinlikle ve kesinlikle elden geçirilmesi gereken en gerekli işlemlerden birisidir. İdeal bir veri ambarına sahip olabilmek için kullanılabilir haldeki çeşitli meta-data tiplerine ihtiyaç vardır. Mükemmel olan, kullanıcıların veri ambarına verilerin nerede tutulduğunu bile bilmeksizin ulaşabilmesidir.

4.1.5. İşlem Yönetim Katmanı

İşlem yönetim katmanı zamanı belli olan yapılması ve devamlılığı sürekli olan bir dizi işlemleri ihtiva eder. İşlem yönetim katmanını yüksek seviyede işimizi kontrol altında tutabilmemizi sağlayan işleri planlayan ve yüksek miktarlardaki verilerin her zaman doğru ve yeni tutan bir ara birim olarak da düşünebiliriz.

4.1.6. Uygulama Haberleşme Katmanı

Uygulama Haberleşme katmanı, enformasyonun network haberleşmesi teknikleri ve teknolojileriyle bütün organizasyon içinde nakledilmesinden sorumlu olan katmandır. Uygulama haberleşmesi orta kademe katman diye de adlandırılabilir fakat sadece network haberleşmesinden öte , uygulamaların birbirinden yalıtılmasından, verilerin operasyonel ya da enformasyonel olarak ayrımına ve verilerin doğru formata çevriminden de sorumludur. Uygulama haberleşmesi katmanı, bilgisayarlar arasındaki uygulama işlemlerinin sonuçlarının tutulması ve doğru zamanda doğru kimseye en doğru verileri sağlamak amacıyla da kullanılır.

4.1.7. Veri Ambarı (Fiziksel) Katmanı

Veri Ambarları temel olarak enformasyonel datayı kullanır. Çoğu zaman insanlar veri ambarlarının fiziksel olarak dataları depolamadığını, bunun sanal olarak verilerin farklı bir formatta görünümünden ibaret olduğunu düşünürler . Aslında çoğu uygulamada veri ambarları verileri depolamak amacıyla kullanılmaz.

Fiziksel bir veri ambarında kopyalar diğer bir değişle haddinden fazla kopya ki bunlar operasyonel faaliyetleri ve müşteri bilgileridir çok kolay ulaşılabilir ve kolaylıkla diğer formatlara çevrilebilecek normlarda saklanır. Günümüzdeki veri

ambarları web tabanlı kişiselleştirilmiş platformları kullanmaya başlamıştır ancak temelde aynı yapıyı kullanırlar.

4.1.8. Data Sunum Katmanı

Bir veri ambarı yapısının sonuncu bileşeni de data sunum katmanıdır. Data sunum katmanı kopya yönetim katmanı ya da replikasyon katmanı olarak da adlandırılır fakat aslında bu katman operasyonel ve/veya harici veri tabanlarına ulaşılarak yapılan tespit, edit, sonuçlandırma, uyumlu hale getirme yükleme işlemleri için gereken bütün aşamaları da kapsar.

Data sunumu karmaşık programlama tekniklerine de gereksinim duyar. Ancak günümüzde hızla gelişen veri ambarcılığı araçları ile bu aşama her geçen gün daha da kolaylaşmaktadır.Data sunum katmanı mevcut operasyonel data içindeki kalıpları ve veri modellerini de tespit etmek amacıyla veri kalitesini analiz eden muhtelif filtreleri ve/veya programları da içerebilir.

4.2. Veri Ambar Tipleri:[2]

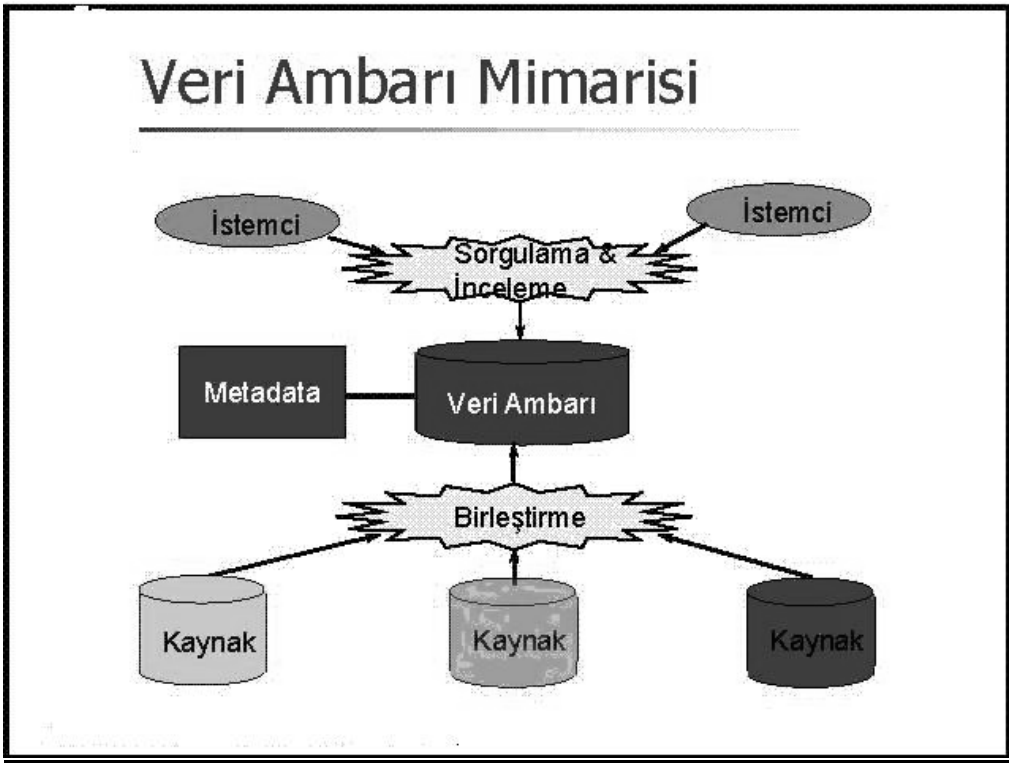
- Amaca Yönelik:
Müşteri, ürün, satış gibi belli konular için düzenlenebilir.Verinin incelenmesi ve modellenmesi için oluşturulur. Konuyla ilgili karar vermek için gerekliolmayan veriyi kullanmayarak konuya basit , özet bakış sağlanır.

- Birleştirilmiş :
Veri kaynaklarının birleştirilmesi ile oluşturulur. Veri temizleme ve birleştirme teknikleri kullanılır. Değişik veri kaynakları arasındaki tutarlılık sağlanır.

- Zaman Değişkenli:
Zaman değişkeni canlı veri tabanlarına göre daha uzundur. Canlı veri tabanları güncel veriler bulundurur.

- Değişken Değil
Canlı veri tabanından alınmış verilerin fiziksel olarak başka bir ortamda saklanması ve güncellemelerden etkilenmemesi..

4.3. Veri Ambarı Mimarisi



Şekil 3: Veri ambarı mimarisi [2]

5. Veri Madenciliğindeki Problemler

Veri madenciliği girdi olarak ham veriyi sağlamak üzere veri tabanlarına dayanır. Bu da veri tabanlarının dinamik, eksiksiz, geniş ve net veri içermemesi durumunda sorunlar doğurur. Diğer sorunlar da verinin konu ile uyumsuzluğundan doğabilir.

Sınıflandırmak gerekirse başlıca sorunlar şunlardır :

- Sınırlı Bilgi : Veri tabanları genel olarak veri madenciliği dışındaki amaçlar için tasarlanmışlardır. Bu yüzden, öğrenme görevini kolaylaştıracak bazı özellikler bulunmayabilir.

- Gürültü ve Eksik Değerler : Veri özellikleri ya da sınıflarındaki hatalara gürültü adı verilir. Veri tabanlarındaki eksik bilgi ve bu yanlışlardan dolayı veri madenciliği amacına tam olarak ulaşmayabilir. Bu bilgi yanlışlığı, ölçüm hatalarından, ya da öznel yaklaşımdan olabilir.

- Belirsizlik : Yanlışlıkların şiddeti ve verideki gürültünün derecesi ile ilgilidir. Veri tahmini bir keşif sisteminde önemli bir husustur.

- Ebat, güncellemeler ve konu dışı sahalar : Veri tabanlarındaki bilgiler, veri eklendikçe ya da silindikçe değişebilir. Veri madenciliği perspektifinden bakıldığında, kuralların hala aynı kalıp kalmadığı ve istikrarlılığı problemi ortaya çıkar. Öğrenme sistemi, kimi verilerin zamanla değişmesine ve keşif sisteminin verinin zamansızlığına karşın zaman duyarlı olmalıdır.

6. Sonuç

Yeni nesil internet, yaklaşık 155 Mbits/sn lik hatta belki de daha da üzerinde hızları kullanmamızı sağlayacak. Bu da günümüzde kullanılan bilgisayar ağlarındaki hızın 100 katından daha fazla bir sürat ve taşıma kapasitesi demektir. Böyle bir bilgisayar ağı ortamı oluştuktan sonra, dağıtık verileri analiz etmek ve farklı algoritmaları kullanmak mümkün olacaktır. Bundan 10 yıl önceki bilgisayar ağları teknolojisinde hayal edemediklerimizi artık kullanabiliyoruz. Buna bağlı olarak, veri madenciliğine uygun ağların tasarımı da yapılmaktadır. Veri madenciliği, sayısal ve istatistiksel olarak büyük veri kümeleri üzerinde yoğun işlemler yapmayı gerektirir. Gelişen bellek ve işlem hızı kapasitesi sayesinde, birkaç yıl önce madencilik yapılamayan veriler üzerinde çalışmayı mümkün hale getirmiştir. Günümüzde ticaret ve işler çok karlı olmalı, daha hızlı ilerlemeli ve daha yüksek kalitede servis ve hizmet verme yönünde olmalı, bütün bunları yaparken de minimum maliyeti ve en az insan gücünü göz önünde bulundurmalıdır. Bu tip hedef ve kısıtların yer aldığı iş dünyasında veri madenciliği, temel teknolojilerden biri haline gelmiştir. Çünkü veri madenciliği

sayesinde müşterilerin ve müşteri faaliyetlerinin yarattığı fırsatlar daha kolay tespit edilebilmekte ve riskler daha açık görülebilmektedir.

Kaynaklar

1) Tukey, J – 1973 : Exploratory Data Analysis
2) Öğüdücü Ş.: "Veri Madenciliği, Genel Bilgiler", http://www.cs.itu.edu.tr/~gunduz/courses/verimaden/
3) SPSS Inc. Chicago, Illionis - http://www.spss.com/datamine/
4) Eker H.:"Veri Madenciliği veya Bilgi Keşfi",http://www.bilgiyonetimi.org
5) Karakaş M.: "Veri Ambarları Genel Yapısı", ",http://www.bilgiyonetimi.org
6) Data Warehousing Information Center - http://www.dwinfocenter.org/
7) Information Discovery Inc. – http://www.datamining.com